Patient Behavior Prediction Using Big Data Analytics and Machine Learning

Dissertation Manuscript

Submitted to National University

School of Technology and Engineering

In Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

by

USHA NARAYANAN

San Diego, California

October 2025

Abstract

Accurately predicting patient behaviors such as treatment adherence, healthcare engagement, and responses to interventions remains a persistent challenge due to the complex interaction of physiological, contextual, and behavioral factors. Traditional predictive models often overlook contextual determinants, including socioeconomic status, environmental conditions, and perioperative stress indicators, thereby limiting both accuracy and clinical applicability. The purpose of this quantitative, explanatory, quasi-experimental study was to develop and evaluate an automated machine learning (AutoML) based big data analytics system using the VitalDB dataset, licensed under the Creative Commons Attribution 4.0 International License, to predict patient adherence behaviors in perioperative care. The dataset included high-resolution physiologic signals, perioperative attributes, and electronic health record–derived outcomes from 6,388 surgical cases. The study compared AutoML frameworks with traditional machine learning models, including logistic regression, decision trees, random forests, and gradient boosting machines. Data preprocessing involved imputation, normalization, and feature engineering to address missing data and ensure model robustness. Model performance was evaluated using metrics such as the area under the receiver operating characteristic curve (ROC–AUC), precision–recall area (PR–AUC), and F1-score. Results demonstrated that ensemble and AutoML models achieved enhanced predictive performance (ROC–AUC = 0.99), while maintaining interpretability through SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME). Key predictors included the Intraoperative Stress Index, ASA classification, and SpO_2 burden. Findings confirm that integrating contextual and physiologic data within explainable AutoML pipelines enhanced predictive accuracy and

transparency, supporting the development of clinically actionable decision-support tools for personalized, data-driven healthcare.

Acknowledgments

As I reach the end of this dissertation journey, I cannot help but pause and reflect on how far this path has taken me. It has not always been easy, but every challenge, long night, and setback was worth it. I am filled with gratitude for all those who stood by me and helped me reach this milestone. First and foremost, I thank God for being my strength and light through every step of this journey, for carrying me through difficult times and blessing me with the perseverance to finish strong. To my amazing husband, Anthony, thank you for your patience, love, and encouragement, and for believing in me when I needed it most. To my wonderful children, Sharon, Christie, and Joshua, and our loving pet Jagger, you are my greatest inspiration and my proudest achievement. Your love, laughter, and constant support gave me the strength to keep going, even on the hardest days. You reminded me of what truly matters and why it was essential to see this through. To my mom, thank you for always motivating me to study and push forward, even when things felt impossible, and for instilling in me the value of education and perseverance that have shaped who I am today. To my sister, Padmini, thank you for all the sacrifices you made that helped me through college. Your love and support have been the foundation that allowed me to build this dream. Finally, to my committee, Dr. Amir Schur, my Chair, thank you for your guidance, encouragement, and steady belief in my work; Dr. Irene Tsapara, my Subject Matter Expert, I am so grateful for your patience, insight, and invaluable feedback that strengthened this research; and Dr. Hamzah Al-Najada, my Academic Reader, thank you for your thoughtful input and support throughout this process. This dissertation is not just my achievement; it reflects all the love, faith, and encouragement I have received along the way. From the bottom of my heart, thank you all for being part of this journey.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

In recent years, the integration of artificial intelligence and predictive analytics in healthcare has revolutionized patient outcome forecasting and personalized treatment planning (Dixon et al., 2024). As healthcare systems grow increasingly complex, there is a pressing need to enhance the accuracy of predictive models, especially when forecasting patient behaviors. Factors such as treatment adherence, healthcare engagement, and responses to clinical interventions are vital in shaping patient outcomes, yet these behaviors remain challenging to predict with precision (Gowda & Lakshmikantha, 2020). Treatment adherence involves following prescribed medical advice, including taking medication as directed, attending scheduled appointments, and making necessary lifestyle changes. Healthcare engagement reflects how patients manage their health through informed decision-making, effective communication with providers, and proactive preventive care actions.

Responses to clinical interventions refer to how patients react physiologically, psychologically, or behaviorally to treatments or procedures, which could influence the success of care strategies. These behaviors are influenced by a complex interplay of biological, psychological, social, and environmental factors, many of which are difficult to observe or quantify. Big data offers unprecedented opportunities to analyze these behaviors by leveraging vast and diverse datasets. However, the complexity and multidimensional nature of healthcare data pose significant challenges, particularly in ensuring that predictive models are both accurate and interpretable (Alam et al., 2024).

This study aimed to evaluate how integrating intraoperative features enhanced the predictive accuracy and interpretability of an Automated Machine Learning (AutoML) based system in forecasting postoperative treatment adherence. The research also aimed to compare the

performance of AutoML models with traditional machine learning approaches and examined the contribution of these intraoperative features to model explainability and feature importance. Recent studies estimated that non-adherence to prescribed treatments could lead to increased hospitalizations and contributed to nearly 125,000 preventable deaths annually in the United States alone (Sokol et al., 2005). This alarming statistic underscored the urgent need for more effective tools to understand and predict patient behaviors that directly impact health outcomes.

Traditional predictive modeling approaches often struggle with the complexity and variability inherent in healthcare data, particularly when addressing multifaceted behaviors such as treatment adherence and healthcare engagement. AutoML platforms offer workflows that simplify the traditionally complex tasks of model selection, feature engineering, and hyperparameter optimization. This automation enables healthcare professionals to build predictive models more efficiently without requiring deep expertise in machine learning, thus facilitating faster and broader adoption in clinical settings (He et al., 2021). AutoML can enhance model accuracy and usability in real-world clinical settings by mitigating the technical barriers commonly associated with machine learning. However, significant gaps remained in understanding how well these automated frameworks translated into actionable insights for healthcare practitioners, especially in high-stakes environments where interpretability and trust are critical.

The landscape of healthcare is being transformed by the rise of big data and machine learning, which has become pivotal in developing predictive analytics tools that aid decision-making (He et al., 2021). Predictive models are increasingly relied upon for many applications, from anticipating disease progression to personalizing treatment plans. A critical aspect of these models is their ability to accurately forecast patient behaviors, such as treatment adherence and

appointment attendance, which significantly impact clinical outcomes. Traditional statistical models often struggle to incorporate the complexity inherent in human behavior, mainly as they overlook key contextual factors, such as environmental conditions, socioeconomic status, and individual health histories (Gowda & Lakshmikantha, 2020).

Machine learning in healthcare has made significant strides in managing the volume, variety, and velocity of medical data, enabling more accurate predictions and delivering actionable insights for clinical practice (Alam et al., 2024). Machine learning algorithms can identify complex patterns and correlations that often exceed human perception by efficiently processing large-scale datasets, including electronic health records, imaging data, wearable sensor outputs, and genomic information. These capabilities support early diagnosis, risk stratification, personalized treatment planning, and real-time clinical decision support, ultimately enhancing patient care and operational efficiency across healthcare systems.

Studies have shown that predictive models integrating machine learning improve treatment adherence prediction accuracy by up to 20% compared to traditional statistical models (Shickel et al., 2018). Ekpezu et al. (2023) identified 13 distinct supervised learning techniques for real-time adherence prediction in behavior change support systems (BCSSs), with most models demonstrating high classification accuracy. Integrating diverse data sources, including electronic medical records, social media behavior, wearable devices, and external factors, opened new doors for predicting patient behaviors more holistically and accurately (McKinsey & Company, 2023). Despite this progress, current systems face challenges, particularly when balancing predictive accuracy and interpretability. This led to concerns about the "black box" nature of many machine learning models, which can generate predictions that are difficult for healthcare providers to understand or act upon (Rudin, 2019).

The advent of AutoML represents a significant step forward in automating and simplifying the process of developing machine learning models. By optimizing model selection and tuning without requiring extensive manual intervention, AutoML enables more scalable solutions (He et al., 2021). However, its application in healthcare is still evolving, mainly when producing interpretable models that clinicians can confidently use. Healthcare professionals are hesitant to adopt AI-driven predictions without clear interpretability and a thorough understanding of the model's logic (Ahmed et al., 2023). This study aimed to bridge the gap by developing an AutoML-based big data analytics system that integrated the diverse factors influencing patient behaviors while ensuring the transparency necessary for real-world healthcare applications. This research advanced predictive analytics by enhancing the accuracy of behavior prediction and model interpretability. It supported the practical implementation of machine learning in clinical settings, enabling healthcare providers to make more informed, data-driven decisions.

Statement of the Problem

The problem addressed in this study was the challenge of accurately predicting patient behavior, such as treatment adherence, healthcare engagement, and responses to interventions, using big data and machine learning. Traditional models often failed to account for complex contextual factors, such as environmental conditions, socioeconomic status, and individual health histories, which limited their predictive accuracy and interpretability in clinical settings (Gowda & Lakshmikantha, 2020). This study leveraged the VitalDB dataset, a high-resolution, multi-parameter repository of perioperative data collected from 6,388 surgical patients. VitalDB encompassed over 196 intraoperative monitoring parameters, 73 perioperative clinical parameters, and 34 laboratory time-series variables, offering a rich, real-world dataset for

developing predictive models that incorporated the physiological, procedural, and contextual dimensions of patient care (Lee et al., 2022).

While machine learning, particularly automated machine learning (AutoML), offered promising tools for automating complex predictions, existing systems often fell short in clinical utility due to limited scalability and interpretability. Machine learning has demonstrated potential in processing diverse human-related data, including speech, physical activity, social media, and electronic medical records (Alsini et al., 2024). However, the complexity of healthcare data demanded transparent and explainable models to support clinical decision-making.

This study developed an AutoML-based analytics system that integrated various influencing factors, including physiological patterns from VitalDB, socioeconomic indicators, and patient history, to enhance the prediction accuracy of behavior-related outcomes. To ensure clinical applicability, the interpretability of the predictive models was evaluated using SHAP (SHapley Additive exPlanations) values and LIME (Local Interpretable Model-Agnostic Explanations). These techniques provided insight into feature contributions and decision pathways, enabling healthcare professionals to understand and trust the model outputs. By prioritizing interpretability alongside accuracy, this study delivered a scalable, transparent, and clinically meaningful predictive system that supported more personalized, data-driven interventions, advancing the practical integration of machine learning into precision healthcare.

Purpose of the Study

The purpose of this quantitative explanatory quasi-experimental study was to develop and evaluate an AutoML-based big data analytics system that predicted patient behaviors such as treatment adherence, healthcare engagement, and responses to interventions by integrating complex contextual factors, including stress levels, environmental influences, and significant life

events. The study quantitatively assessed model performance using evaluation metrics, including accuracy, precision, recall, F1-score, and the Area Under the Curve (AUC). These metrics were used to compare the predictive accuracy of the AutoML-based system with that of traditional machine learning models.

To ensure clinical interpretability, the study incorporated explainable artificial intelligence (XAI) methods, specifically SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). SHAP quantified the global importance of features across the entire dataset, while LIME provided local explanations for individual predictions. These tools help healthcare professionals understand the rationale behind model outputs and support decision-making by providing insights into the specific features that influence patient behavior predictions.

The study addressed the ongoing challenge of improving predictive accuracy and model transparency in clinical settings. It followed a structured, data-driven methodology aligned with the principles of explanatory quasi-experimental research (Creswell & Creswell, 2018), allowing for a robust evaluation of the impact of contextual and physiological variables on model performance.

This study used the VitalDB dataset, a publicly available high-resolution perioperative database comprising intraoperative waveforms, numeric vital signs, laboratory results, and perioperative patient information collected from 6,388 surgical cases. Within the context of clinical research standards, VitalDB was unquestionably recognized as a form of big data. It has been utilized in numerous large-scale data research projects. It met the "three Vs" criteria: high volume, large number of time series per patient; high velocity, high sampling frequency; and high variety, diverse bio signals and metadata.

This dataset provided a rich and diverse source of physiological and clinical data, encompassing over 196 intraoperative monitoring parameters, 73 perioperative clinical variables, and 34 time-series laboratory metrics. These data elements provided a comprehensive view of patient health histories and physiological responses, enabling modeling behavior-related outcomes in surgical and perioperative care. When supplemented with synthetic or external data reflecting environmental and socioeconomic contexts, the VitalDB data served as a robust foundation for exploring patient engagement and adherence behaviors in healthcare.

The study developed models using supervised machine learning algorithms, including decision trees, random forests, gradient boosting machines (GBM), and long short-term memory (LSTM) networks. These models were selected to capture static and time-dependent data patterns. Minimal preprocessing was applied to preserve data authenticity, and ethical considerations precluded experimental designs, such as patient randomization, making a quasi-experimental approach more suitable.

This study aimed to develop a scalable, transparent, and clinically useful predictive system by applying rigorous evaluation criteria and emphasizing model interpretability. The expected outcome contributed to healthcare data science and clinical practice by developing predictive tools that supported personalized interventions, improved patient outcomes, and increased clinician trust in machine learning systems. Ultimately, the findings informed future applications of AutoML and XAI in precision healthcare using VitalDB as a novel and underutilized data source.

Introduction to Theoretical Framework

This study was guided by an integrated framework combining the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology with health behavior theory and

XAI principles. The integration of these frameworks provided a structured approach to developing and evaluating an AutoML-based system for predicting patient behaviors, ensuring both technical rigor and practical applicability in healthcare settings.

The CRISP-DM framework, a widely adopted methodology in data science, guided the entire lifecycle of this study by providing a structured approach to developing a predictive analytics system (Shearer, 2000). This framework organized the analytics lifecycle into six iterative phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Each phase of CRISP-DM was strategically aligned with the study's research design, contributing to the systematic development and implementation of the AutoML system.

In the business understanding phase, the study defined a clear objective: to enhance the predictive accuracy and interpretability of patient behavior models by utilizing automated machine learning. This included formulating research problems, developing relevant hypotheses, and framing research questions about the need for transparent and scalable machine learning systems in clinical contexts. Interpretable predictions ensure that healthcare providers can trust and utilize the resulting models.

The data understanding phase involved acquiring and analyzing publicly available electronic health records (EHRs), specifically leveraging the VitalDB dataset. This high-resolution database included structured numeric data, such as demographic characteristics, vital signs, and time-series laboratory results, as well as rich temporal data that reflect intraoperative monitoring (Lee et al., 2022). This multidimensional data structure supported the extraction of static and dynamic patient features relevant to behavioral prediction. By integrating contextual indicators such as comorbidities, medication histories, and perioperative stress indicators, the

system was better positioned to capture the complexity of real-world patient behaviors (Topol, 2019).

During the data preparation phase, essential preprocessing steps were performed to improve data quality and reliability. These include handling missing values through imputation, applying normalization and scaling techniques, encoding categorical variables, and conducting advanced feature engineering to enhance model inputs. This phase ensured the dataset was clean, well-structured, and ready for automated modeling.

The modeling phase applied to an AutoML platform to automate the selection, tuning, and ensemble of multiple supervised machine learning algorithms, including decision trees, random forests, gradient boosting machines, and long short-term memory (LSTM) networks. These algorithms were chosen for their ability to model linear and non-linear patterns and temporal dependencies, which are common in healthcare data (Ching et al., 2018). AutoML tools optimized hyperparameters and conducted model selection without requiring extensive manual intervention, enabling efficient and reproducible model development.

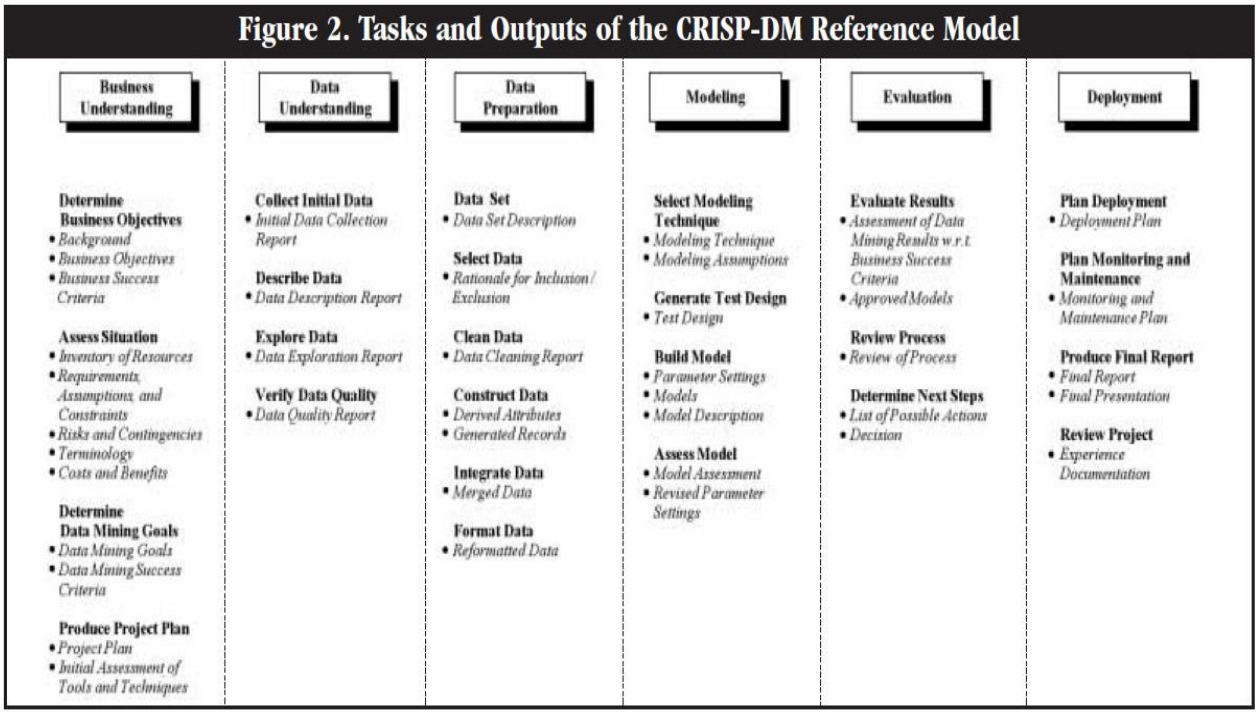
In the evaluation phase, model performance was rigorously assessed using standard metrics such as accuracy, precision, recall, F1-score, and AUC (Area Under the Curve). These metrics were crucial for quantifying the predictive accuracy and robustness of models across various behavioral prediction tasks. Beyond traditional performance metrics, this phase incorporated model interpretability assessments using Explainable AI tools to ensure that clinical stakeholders understood the outputs.

In the deployment phase, the best-performing and most interpretable models were prepared for integration into clinical decision-making workflows. Deployment considerations included designing user interfaces, generating explanation reports, and ensuring that outputs

were accessible to healthcare professionals. The system aimed to support real-time, data-driven interventions that improve patient care and engagement. Figure 1 shows the tasks and outputs of the CRISP-DM Reference Model.

Figure 1

CRISP-DM Tasks and Outputs



Note. Adapted from "*The CRISP-DM model: The new blueprint for data mining*," by Shearer, C., 2000, Journal of Data Warehousing, 5(4), 13-22.

To address the inherent "black box" challenge of many machine learning algorithms, the study embedded explainable artificial intelligence (XAI) principles into its modeling and evaluation pipeline. Explainability is essential in clinical contexts, where model decisions could directly affect patient health and provider accountability. Tools such as SHAP (SHapley

Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) were used to generate localized and global explanations of model predictions (Lundberg & Lee, 2017).

These explainable artificial intelligence (XAI) tools helped clinicians understand the influence of key variables such as socioeconomic status, stress levels, comorbidities, and treatment adherence history on individual prediction outcomes. For instance, SHapley Additive exPlanations (SHAP) values quantified the contribution of each feature to a specific prediction, while Local Interpretable Model-Agnostic Explanations (LIME) provided interpretable, simplified models around local instances. This interpretability fostered trust among end users, facilitated clinical validation, and supported shared decision-making between providers and patients.

Moreover, XAI contributed to ethical AI governance by making predictive models more transparent and accountable. Clear explanations reduced the risk of relying on biased or opaque models, which could inadvertently reinforce health disparities. By enabling visibility into how and why a model made specific predictions, clinicians could identify and challenge incorrect or ethically questionable outputs, thus promoting fairer and safer use of AI in healthcare (Doshi-Velez & Kim, 2017).

This study was guided by health behavior theory, which explained and predicted patient actions based on personal and contextual factors, emphasizing the role of environmental, psychological, and social determinants in health-related behaviors (Glanz et al., 2015). Key concepts integrated into the study included socioeconomic status and environmental conditions, which were used as features in the predictive models to capture external influences on patient adherence and engagement. By incorporating these constructs, the study tested the hypothesis that contextual factors significantly enhance predictive accuracy and clinical relevance.

As part of the study's ethical and methodological framework, rigorous data governance and fairness principles were embedded throughout the research process. Given the sensitive nature of healthcare data, the study adhered to strict ethical guidelines and legal standards, including HIPAA compliance, by utilizing only publicly available, de-identified datasets such as VitalDB. This approach mitigated privacy concerns and eliminated the need for individual informed consent, aligning with best practices in secondary data analysis (Stevens, 2003). The framework incorporated continuous monitoring of model outputs for bias related to demographic, racial, or socioeconomic variables. Finally, regular evaluation of model outputs for potential biases related to demographic or socioeconomic factors ensured equitable healthcare outcomes, promoting fairness and reducing disparities in care (Obermeyer et al., 2019).

Introduction to Research Methodology and Design (Nature of the Study)

This study aimed to leverage an AutoML-based framework to predict patient behaviors by utilizing machine learning algorithms designed to address challenges such as data quality, bias, and contextual integration, using the VitalDB dataset as the primary data source. VitalDB is a high-resolution, open-access clinical database that contains waveform and numeric intraoperative monitoring data from over 6,000 surgical cases, as well as structured perioperative clinical parameters and time-series laboratory results. Its comprehensive scope enabled exploration of physiological responses and perioperative histories related to healthcare engagement behaviors such as treatment adherence and follow-up attendance.

The study employed a range of supervised learning algorithms tailored to specific prediction tasks and the multi-dimensional characteristics of the VitalDB dataset (Ching et al., 2018). Decision trees served as interpretable baseline models, constructing classification or regression rules from VitalDB's structured numeric and categorical variables. Random Forests

were employed for their ensemble strength, which could handle high-dimensional and mixed data types in VitalDB, particularly when combining numeric vitals with categorical clinical attributes. Gradient Boosting Machines (GBM) modeled complex, non-linear relationships within patient data, such as dynamic medication responses or perioperative risk profiles.

Additionally, LSTM networks analyzed VitalDB's time-series data, including respiratory rates, blood pressure, and oxygen saturation, to capture behavioral patterns and physiological trajectories indicative of treatment engagement or non-compliance. To enhance model interpretability, the study integrated XAI techniques, particularly LIME, into the AutoML pipeline. These tools were critical for translating complex model predictions into clinically actionable insights for providers.

Validation of Synthetic Contextual Variables

As VitalDB did not contain direct measures of social determinants of health, the study incorporated synthetic contextual variables, such as regional socioeconomic indices or deprivation scores, as proxies to offer a limited contextual perspective. However, as the study primarily emphasized clinical and high-frequency time-series data, these socioeconomic indicators were not central to the modeling framework. Instead, their use was exploratory and supplementary. To ensure that any included synthetic variables did not introduce noise or spurious correlations, they were evaluated using correlation analysis and feature importance rankings to assess whether they contributed a meaningful signal to the prediction tasks.

Sensitivity testing was conducted to assess the additive value of these proxies by comparing model performance with and without them. The inclusion of such variables was guided by theoretical relevance and data availability, rather than being positioned as core model drivers. As such, while socioeconomic factors were recognized as important influences in real-

world patient behavior, they fell outside the immediate analytical scope of this study, which prioritized physiological signals and perioperative clinical features as the foundation for behavior prediction.

Data Preprocessing and Feature Engineering

The study implemented several data preprocessing techniques to ensure the data was well-prepared for analysis. Preprocessing was tailored to the structure of VitalDB, which consists of high-frequency waveforms, numeric data streams, and static perioperative features. Missing value handling depended on variable types: simple imputation and mean/median were used for continuous numeric values, while the k-nearest neighbors (KNN) imputation was applied for more complex missingness patterns across correlated vital signs.

Normalization and standardization were applied to continuous variables, such as heart rate, blood pressure, and lab results, to improve algorithm performance, especially for LSTM, and enhance models that were sensitive to feature scale. Feature engineering involved aggregating time-series data, such as average heart rate variability or oxygen saturation trends during surgery, and creating synthetic behavioral markers, including recovery stability indices or vital-based stress scores.

The perioperative static data, like ASA classification, comorbidities, and procedure type, were encoded and used as contextual features. Vital DB did not directly include unstructured data, such as clinical notes or social determinants of health. The study, however, incorporated synthetic contextual variables, like population-level socioeconomic indices linked by region or hospital, as proxies for broader behavioral influences. Temporal features were structured with each time point or aggregated interval and treated as a feature sequence for LSTM models.

Noise reduction was conducted through signal smoothing and outlier detection, which were essential for waveform data where artifacts or monitoring inconsistencies might occur. Feature consistency and alignment across modalities, such as time-aligning numeric tracks with lab test results, were ensured for accurate feature engineering. Techniques such as cross-validation, ensemble learning, and regularization enhanced model robustness against noisy data. explainable artificial intelligence (XAI) methods, such as SHAP and LIME, provided transparency and helped identify problematic data points for correction (Li et al., 2020).

Model Selection and AutoML Implementation

The AutoML framework streamlined the model development process by automating several key tasks. For model selection, AutoML evaluated various machine learning algorithms, including decision trees, random forests, gradient boosting machines (GBM), and deep learning variants such as LSTM, selecting the best-performing model based on evaluation metrics like AUC, accuracy, and F1-score (Tabassum et al., 2020). Hyperparameter optimization, using grid search, random search, or Bayesian optimization, was employed to fine-tune each model's configuration. To ensure generalizability, k-fold cross-validation was applied, with folds stratified by key variables like procedure type or ASA status to maintain clinical representativeness.

Mitigation of Overfitting and Interpretability

Overfitting is a common challenge in machine learning, where a model excels in training data but struggles with new, unseen data, which was addressed through several techniques in this study. Regularization methods, including L1 and L2 regularization, limited model complexity and prevented overfitting, especially in high-dimensional datasets. Dropout layers were implemented in neural networks to randomly omit units during training, which helped the model

to generalize better by reducing its reliance on any single feature. AutoML's pipeline optimization also incorporated early stopping and model ensembling to stabilize performance.

Explainable AI (XAI) techniques such as SHAP and LIME were integrated to ensure that predictions are transparent and interpretable, addressing the black-box nature of many machine learning models and facilitating their practical application by healthcare professionals. SHAP values provided global insights into feature contributions across all patients, helping uncover which vital signs or perioperative metrics most influenced behavior predictions. LIME supported localized explanations, enabling clinicians to review why a particular patient was predicted to be at risk of non-adherence and guide personalized interventions.

Research Questions

To develop an AutoML-based system to improve the prediction of patient behaviors by integrating contextual factors, the following research questions are designed to explore the technical aspects of the proposed system.

RQ1

To what extent does the integration of the intraoperative Stress Index enhance the predictive accuracy of an AutoML-based system compared to traditional machine learning models in forecasting postoperative treatment adherence?

RQ2

How do intraoperative features such as the Stress Index, ASA classification, and time-series trends in HR and SpO_2 contribute to the interpretability and feature importance within AutoML-generated models for predicting treatment adherence?

Hypotheses

$H1_0$

There is no significant difference in predictive accuracy between the AutoML-based system and traditional machine learning models in forecasting postoperative treatment adherence.

$H1_a$

The AutoML-based system significantly outperforms traditional machine learning models in forecasting postoperative treatment adherence.

$H2_0$

Intraoperative features such as the Stress Index, ASA classification, and vital sign trends do not significantly contribute to the interpretability or feature importance rankings within AutoML-generated models.

$H2_a$

Intraoperative features such as the Stress Index, ASA classification, and vital sign trends significantly contribute to the interpretability and feature importance rankings within AutoML-generated models.

Significance of the Study

Accurately predicting patient behaviors such as treatment adherence, healthcare engagement, and response to interventions was a technical challenge and a vital step toward improving patient care and long-term outcomes. However, traditional predictive models often faced limitations in fully capturing the complex realities that influenced health behavior, such as socioeconomic pressures, environmental factors, and individual medical histories. This study aimed to bridge the gap by developing a scalable and interpretable AutoML (Automated

Machine Learning) system that integrated diverse data sources to enhance predictive accuracy and relevance in clinical contexts.

This research pushed the boundaries of data science in healthcare by integrating multi-dimensional healthcare data, including physiological signals, perioperative features, and contextual indicators, with advanced machine learning techniques. The AutoML framework automated model selection, tuning, and validation, streamlining a process that was often time-consuming and improving consistency and performance. Additionally, by embedding explainable artificial intelligence (XAI) tools such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) into the predictive pipeline, the study ensured that its findings were not only accurate but also transparent and interpretable for clinicians and care teams.

This research sought to support timely, informed, and personalized clinical decisions. Offering healthcare providers clear, actionable insights into patient behavior patterns empowered them to implement earlier interventions and more targeted care strategies. Healthcare systems can expect more efficient resource allocation, reduced hospital readmissions, and enhanced continuity of care. In addressing broader data science challenges, such as handling heterogeneous data, mitigating algorithmic bias, and enhancing model interpretability, this study contributed to the growing body of work that sought to make machine learning more powerful, human-centered, ethical, and practically applicable in real-world settings.

Definitions of Key Terms

Automated Machine Learning (Auto ML)

Automated machine learning (AutoML) refers to the automation of the end-to-end workflow for building machine learning models, encompassing tasks such as data preprocessing,

feature engineering, model selection, hyperparameter tuning, and model evaluation. AutoML platforms aim to make machine learning more accessible by reducing the need for extensive domain expertise, optimizing performance, and improving efficiency (He et al., 2020).

Decision Trees

A decision tree is a supervised machine learning algorithm used for classification and regression tasks. It represents decisions and their possible outcomes as a tree-like structure, where each internal node corresponds to a feature, each branch represents a decision rule, and each leaf node denotes an outcome or class label (Mienye & Jere, 2024).

Explainable AI (XAI)

Explainable Artificial Intelligence (XAI) refers to a set of methodologies and techniques designed to make the decision-making processes of AI systems transparent, interpretable, and understandable to humans (Guidotti et al., 2018).

Gradient Boosting Machines (GBM)

Gradient Boosting Machines (GBMs) are a powerful family of machine learning algorithms that sequentially combine weak learners, typically decision trees, to create a strong predictive model. GBMs work by iteratively training new models to minimize the loss function of the previous ensemble using gradient descent.

Local Interpretable Model-agnostic Explanations (LIME)

Local Interpretable Model-agnostic Explanations (LIME) is a technique used to explain the predictions of complex machine learning models. LIME works by approximating the model's behavior locally around a specific prediction with a simpler, interpretable model, such as linear regression. This approach provides human-understandable insights into why a particular

prediction was made, making it especially useful for black-box models (Kumarakulasinghe et al., 2020).

Long Short-Term Memory Network (LSTM)

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network (RNN) designed to learn and capture long-term dependencies in sequential data. LSTMs address the vanishing gradient problem commonly found in traditional RNNs by utilizing memory cells and gating mechanisms (input, forget, and output gates) that regulate the flow of information. These properties make LSTMs effective for time-series forecasting, speech recognition, and natural language processing.

SHApley Additive exPlanations (SHAP)

SHAP is an explainability technique in machine learning that interprets model predictions by attributing the contribution of each feature to the model's output. It is widely used to explain individual predictions and understand the importance of global features (Huyen, 2022).

Supervised Learning

Supervised learning is a type of machine learning where the algorithm is trained on a labeled dataset. Each training example consists of input data paired with the corresponding output (label), enabling the model to learn a mapping function from inputs to outputs. Once trained, the model can predict new, unseen data output. Supervised learning is commonly used for tasks such as classification and regression (Huyen, 2022).

Summary

Understanding and predicting patient behavior remained one of the most complex challenges in healthcare. This study took a significant step toward addressing this issue by integrating Artificial Intelligence (AI) and Automated Machine Learning (AutoML). Focusing

on behaviors such as treatment adherence, healthcare engagement, and response to interventions, the research explored how AI could help bridge gaps left by traditional predictive models, which often overlook critical contextual factors, including socioeconomic status, environmental conditions, and intraoperative clinical data. This research utilized the VitalDB dataset, a high-resolution, open-access resource of intraoperative vital signs and perioperative clinical parameters, to develop and evaluate an AutoML-based system for improving prediction accuracy and interpretability in real-world clinical settings.

The methodology employed supervised machine learning algorithms, including decision trees, random forests, gradient boosting machines, and Long Short-Term Memory (LSTM) networks, to analyze time-series physiological data, intraoperative measurements, and patient-level clinical information from VitalDB. By integrating complex contextual indicators such as ASA classification, surgical stress responses, and vital sign fluctuations, the system aimed to capture deeper patterns in patient behavior and treatment outcomes. This data-driven approach could potentially reveal actionable insights that could guide clinicians in optimizing care plans and improving patient adherence. By emphasizing model interpretability, scalability, and real-world applicability, this study aimed to bridge the gap between predictive analytics and clinical decision-making, ultimately improving patient outcomes and enhancing healthcare efficiency.

Automated Machine Learning (AutoML) streamlined model selection and optimization, providing a scalable solution for processing large and complex healthcare datasets, such as VitalDB. Despite its promise, the adoption of AutoML in clinical settings is hindered by concerns about the "black box" nature of AI, making interpretability crucial for clinician trust. Explainable artificial intelligence (XAI) techniques, such as SHapley Additive exPlanations

(SHAP) and Local Interpretable Model-agnostic Explanations (LIME), offer transparency, enabling healthcare providers to gain actionable insights.

Guided by the Cross-Industry Standard Process for Data Mining (CRISP-DM) and health behavior theory, the study followed a structured, ethical, and iterative model development, evaluation, and deployment process. Ethical considerations, including HIPAA compliance, anonymized public data, and regular assessment of model bias, were embedded throughout to ensure equitable outcomes. By enhancing predictive modeling capabilities, integrating clinical interpretability through XAI, and utilizing high-quality data from the VitalDB repository, this study contributed to advancing the use of AI in personalized healthcare and informed clinical decision-making.

Chapter 2: Literature Review

The purpose of this quantitative explanatory quasi-experimental study was to develop and evaluate an automated machine learning (AutoML) -based big data analytics system that predicted patient behaviors, such as treatment adherence, healthcare engagement, and responses to interventions, by integrating complex contextual factors, including stress levels, environmental influences, and significant life events. This study's core problem was the limitation of traditional predictive models, which often failed to account for the intricate, nonlinear interactions between environmental, socioeconomic, and personal health variables. These omissions could diminish predictive accuracy and model interpretability in real-world clinical contexts (Razzak et al., 2019; Obermeyer et al., 2019).

This literature review is structured to provide a comprehensive foundation for the study by exploring the intersection of machine learning, healthcare analytics, and behavioral science. It begins with a discussion of the theoretical frameworks that support this research, drawing from behavioral science, decision theory, and explainable machine learning to demonstrate why incorporating contextual factors, like socioeconomic stressors and psychological influences, could significantly improve the accuracy and relevance of predictive models in healthcare (Richter et al., 2025).

The following section explores the evolution of predictive modeling in healthcare, from early rule-based and statistical approaches to more recent machine learning and AutoML applications. While traditional models, such as logistic regression, offered simplicity and interpretability, they often fell short in capturing the complexity of real-world health behaviors (Rajkomar et al., 2018). In contrast, newer machine learning techniques, including decision trees, gradient boosting, and deep learning, could effectively handle nonlinear relationships and high-

dimensional data. AutoML frameworks further reduced the technical barriers to model development by automating feature selection, hyperparameter tuning, and algorithm selection (Kok et al., 2024).

The review includes an examination of the data environment in which these models operated. It considered the structure and limitations of electronic health records (EHRs), which often suffer from fragmentation, missing values, and coding inconsistencies (Weiskopf & Weng, 2012). These challenges were especially relevant when incorporating time-series and physiological data, such as that found in the VitalDB dataset, which offered high-frequency intraoperative monitoring data that was more consistent and structured than traditional EHRs (Lee et al., 2022).

Another important theme in this review was the growing role of explainable AI (XAI) in clinical settings. As machine learning models became increasingly complex, tools such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) emerged to help clinicians and stakeholders understand the logic behind predictions (Lundberg & Lee, 2017; Ribeiro et al., 2016). These techniques broke down model decisions into understandable contributions, enabling healthcare professionals to make more informed choices and trust AI outputs (Kutlu et al., 2024).

The review also placed a strong emphasis on the role of context in shaping patient behavior. Socioeconomic status (SES), psychological stress, environmental instability, and life disruptions had a significant influence on healthcare engagement and treatment adherence (Kivimäki & Steptoe, 2017). However, many predictive models overlooked these variables because of data availability issues or because the modeling frameworks were not designed to incorporate such complexity (Rajkomar et al., 2018). This study sought to address that gap by

integrating contextual data directly into the modeling process, leveraging AutoML to streamline and scale this integration.

Finally, the review synthesized insights from empirical studies that have attempted to apply machine learning to clinical outcomes, evaluating what had worked, where limitations persisted, and how this study could build on that work to offer a more comprehensive, explainable, and patient-centered predictive system. Overall, this literature review highlighted the importance of developing predictive systems that were not only accurate but also interpretable, equitable, and context-aware. The insights gained here supported the study's broader goals: to improve patient outcomes, assist clinicians in making more informed decisions, and provide policymakers with more reliable tools for resource allocation and risk mitigation in healthcare environments. In doing so, this work contributed to the ongoing evolution of AI in medicine, potentially shifting the focus from model performance alone to a more holistic view of health behavior prediction.

The databases accessed for this literature review included PubMed, IEEE Xplore, ACM Digital Library, ScienceDirect, and Google Scholar. Searches were conducted using terms such as *AutoML in healthcare, predictive analytics in patient behavior, explainable AI in clinical decision-making, machine learning and electronic health records,* and *socioeconomic factors in health predictions.* The search parameters included peer-reviewed journal articles, conference papers, and systematic reviews published between 2010 and 2024, ensuring a comprehensive understanding of the most recent advancements and challenges in predictive healthcare analytics. The primary search strategy for the literature review involved using the Roadrunner Search tool on the National University Library website to collect information from peer-reviewed articles,

scholarly journals, books, National University ProQuest dissertations, and other relevant sources related to predictive analytics in patient behaviors.

AutoML streamlined the process of machine learning model development by automating feature selection, hyperparameter tuning, and model evaluation. This enabled healthcare professionals with limited data science expertise to deploy predictive models efficiently. AutoML was also found to enhance reproducibility in healthcare applications. Microsoft Power Platform, Google AutoML, and Amazon SageMaker reduced the need for extensive manual intervention, allowing non-experts to build scalable and standardized ML solutions efficiently. H2O.ai provided an end-to-end AutoML framework that automated data preprocessing, model selection, and hyperparameter tuning, much like Google AutoML, which simplified model development for image classification and NLP. Auto-sklearn leverages Bayesian optimization to enhance ML workflows, paralleling Amazon SageMaker's focus on automated model tuning for technically skilled users. Table 1 compares various features and functionality of AutoML platforms.

Table 1

Comparison of AutoML Platform Features

Feature	Google AutoML	Amazon SageMaker	H2O.ai	Microsoft Power Platform
Platform	Cloud-based (Google Cloud)	Cloud-based (Amazon Web Services)	Open-source and cloud-based options are available	Cloud-based (Microsoft Azure)
Ease of Use	User-friendly interface with minimal coding required	Offers both low-code and advanced coding options	User-friendly for non-experts, also supports custom coding	Low-code, highly accessible for business users
Supported Models	Image, text, and tabular data models	A broad range of models, including image, text, and tabular data	A wide range of algorithms for supervised and unsupervised learning	Primarily focused on tabular data, also supports AI for business applications
AutoML Features	Automated model selection and tuning	Automated machine learning with model selection and tuning	Automated feature engineering and model tuning	Automated ML capabilities within Power BI and Power Apps
Integration	Seamless integration with Google Cloud tools	Seamless integration with the AWS ecosystem and services	Can integrate with cloud platforms and on-premises environments	Deep integration with Microsoft products like Azure, Power BI, and Office 365
Data Processing	Built-in tools for data preprocessing and augmentation	Built-in tools for data preprocessing, cleaning, and transformations	Preprocessing support via built-in methods and custom pipelines	Built-in data prep and transformation tools within Power BI and Power Apps
Pricing	Pay-as-you-go pricing based on usage	Pay-as-you-go pricing based on instance type, storage, and usage	Open-source (free for on-prem), cloud options are available for paid services	Subscription-based, pricing depends on usage and subscription level
Model Deployment	Easy deployment to Google Cloud for real-time predictions	Model deployment with auto-scaling on AWS infrastructure	Flexible deployment options (cloud and on-premises)	Deployment within Microsoft Azure, real-time integrations via Power Apps
Customizability	Limited to built-in AutoML features	High customizability with access to the complete AWS ecosystem and SDKs	High customizability with support for advanced configurations	Customizable workflows with integrations into enterprise systems
Documentation & Support	Extensive documentation and community support	Extensive documentation, tutorials, and AWS support	Good documentation and active community support	Extensive documentation, tutorials, and Microsoft support

By reducing technical barriers, AutoML platforms have increasingly enabled small and medium-sized enterprises (SMEs) to incorporate machine learning into their operations without requiring deep expertise in data science, thereby fostering innovation and more efficient decision-making (Kok et al., 2024). In healthcare, this democratization of machine learning tools was particularly compelling, as it opened the door for broader use of predictive analytics in settings with limited technical infrastructure. However, the literature suggested that the success of AutoML in real-world healthcare relied not only on automation but also on model transparency, scalability, and the ability to account for contextual complexity in patient data.

Several studies examined the use of AutoML for healthcare predictions. Luo et al. (2023) developed an interpretable XGBoost model to predict 30-day hospital readmissions, achieving an AUC-ROC of 0.783. While not a pure AutoML implementation, their work highlighted the importance of combining accuracy with interpretability. This aligned with the goals of the present study, which aimed to evaluate AutoML systems that strike a balance between automation and clinical usability and transparency.

At the algorithmic level, traditional machine learning techniques, including decision trees, random forests, gradient boosting machines, support vector machines, and logistic regression, remained central to predictive healthcare modeling (Hastie et al., 2013). These models often required significant manual effort for preprocessing, feature selection, and hyperparameter tuning, tasks that AutoML sought to automate. Neural networks were increasingly leveraged in time-series prediction and patient monitoring contexts, particularly in high-dimensional and unstructured data environments. While AutoML systems abstracted much of this complexity, they still relied on these core algorithmic foundations.

The relationship between traditional machine learning, AutoML, and theoretical learning frameworks such as computational learning theory (CLT) deserves more attention in applied research. CLT helped clarify how models generalized from training data to unseen cases by offering insights into sample complexity, hypothesis space expressiveness, and learning guarantees (Shalev-Shwartz & Ben-David, 2015). Zöller and Huber (2019) emphasized that CLT concepts, such as Vapnik-Chervonenkis (VC) -dimension and generalization bounds, could guide the design of more efficient AutoML pipelines. This theoretical grounding was crucial for understanding the trade-offs between automation and model reliability in sensitive domains, such as healthcare.

In the context of predicting patient behavior, supervised learning remained the most widely applied approach. Logistic regression was a common baseline for binary classification tasks, such as adherence prediction. However, ensemble methods such as random forest and XGBoost showed superior performance due to their ability to model nonlinear interactions within patient data (Bohlmann et al., 2021). Gradient boosting variants, such as LightGBM and CatBoost, improved predictive outcomes in hospital readmission scenarios by effectively handling categorical variables and missing values. Yun et al. (2021) reported an AUC-ROC of 0.861 using XGBoost to forecast critical care outcomes, highlighting the benefits of ensemble learning in complex clinical environments.

Deep learning approaches, especially recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, are particularly valuable when working with time-series data such as continuous vital signs or EHR logs. Esteva et al. (2018) demonstrated the power of LSTM-based models in tracking disease progression and predicting patient deterioration,

suggesting their suitability for contexts like intraoperative monitoring or wearable data streams, an area central to this study's use of the VitalDB dataset.

The performance of predictive models is only one side of the coin; interpretability remains a significant barrier to clinical adoption. The black box nature of many machine learning models raised ethical concerns, limits clinician trust, and obscures potential biases. Explainability tools, such as SHAP and LIME, have gained traction in addressing this issue. SHAP provided a theoretically grounded method for assigning feature importance, offering global and local interpretability. Kutlu et al. (2024) utilized SHAP to enhance transparency in AI-based diabetes risk prediction, allowing clinicians to comprehend individual patient risk profiles. LIME, meanwhile, offers case-specific explanations that were especially helpful for justifying predictions in a clinical decision support setting. Ahmed et al. (2024) applied LIME in diabetes management, demonstrating how interpretable predictions could lead to better-informed decisions and higher trust among healthcare providers.

This body of work collectively informed the design of the present study. By building these technical foundations and addressing limitations identified in prior literature, particularly those related to interpretability and context-aware modeling, this research aimed to advance AutoML applications in healthcare by integrating time-series physiological data and contextual clinical indicators within a scalable and explainable framework. Using the VitalDB dataset enabled the exploration of these challenges in a high-resolution, real-world clinical context, aiming to enhance the predictive power and trustworthiness of AI-assisted healthcare decisions.

Traditional Predictive Modeling

Traditional predictive modeling played a crucial role in analyzing patient behavior and informing data-driven decisions in healthcare analytics. Classical algorithms, such as logistic

regression, decision trees, support vector machines (SVMs), and random forests, have been widely applied to predict a range of outcomes, including treatment adherence, readmission risk, disease progression, and patient engagement (Gu et al., 2021). These models relied heavily on structured data derived from electronic health records (EHRs), claims databases, and patient-reported outcomes and often require manual feature engineering and domain expertise for effective implementation (Ramakrishnaiah et al., 2025).

Despite their simplicity and interpretability, these traditional approaches struggled with nonlinear patterns and high-dimensional data. However, their transparency and ability to incorporate clinical variables made them particularly valuable for understanding and predicting patient behaviors in real-world healthcare settings (Shickel et al., 2018). Decision trees and rule-based models remained popular due to their interpretability, which is critical in clinical decision-support systems.

Decision Trees

The decision tree is a core concept in both statistical theory and computer science, and has found significant applications in healthcare, particularly for predicting patient behaviors. Decision Trees were widely used in healthcare analytics due to their interpretability, simplicity, and ability to handle categorical and numerical data. Their intuitive tree-like structure closely mimicked human decision-making, making them suitable for clinical environments where transparency and explainability are essential. This literature review examined the application of Decision Tree algorithms in predicting patient behavior, including treatment adherence, healthcare utilization, and responses to public health interventions.

One primary application of Decision Trees in this domain was predicting treatment adherence. Classification and regression trees (CART) and C4.5 algorithms were utilized to

identify the key determinants of whether patients adhered to prescribed medical regimens. Razzak et al. (2019) employed a decision tree model to predict medication non-adherence among diabetic patients, incorporating variables such as age, comorbidities, and socioeconomic status. The model offered high interpretability, allowing clinicians to identify specific patient segments at risk of non-adherence. In another line of research, decision trees have proven valuable in forecasting healthcare service utilization. Hu et al. (2017) developed a decision tree model to predict emergency department visits among elderly patients, using features such as chronic disease profiles, medication usage patterns, and historical visit data. Their model provided valuable insights into the drivers of repeat visits, facilitating the development of proactive care strategies.

Decision Trees were also employed to assess behavioral responses to public health interventions. Ozcan and Peker (2022) employed a classification and regression tree (CART) algorithm to predict heart disease and derive decision rules for personalized healthcare interventions. These applications demonstrated how decision trees served as predictive tools and frameworks for decision support in complex healthcare settings. Widely recognized algorithms, such as CART (Breiman et al., 2017), C4.5 (Quinlan, 1992), CHAID (Yang et al., 2023), and QUEST (Tileubai et al., 2023), differed in their splitting criteria, handling of missing values, and complexity of decision boundaries. For example, while CHAID used chi-square statistics and supported multi-way splits, CART favored binary splits, which made it effective but sometimes less flexible in capturing nuanced patterns in health behavior.

These models were valued for their transparency, ability to accommodate both categorical and numerical inputs, and potential to highlight key variables that influenced outcomes (Sadeghi et al., 2024). However, decision trees were also prone to overfitting,

particularly when used without pruning or regularization, and could yield unstable results if the underlying data were to change slightly. Ensemble methods, such as random forests and gradient boosted trees, were often employed to address these limitations and improve predictive stability and generalizability. Nevertheless, the interpretability of a single decision tree remained a valuable advantage, especially in clinical decision-making contexts where understanding model logic is critical (Gao et al., 2023). Figure 2 compares different decision tree algorithms.

Figure 2

Comparison of Different Decision Tree Algorithms

Methods	CART	C4. 5	CHAID	QUEST
Measure used to select input variable	Gini index; Twoing criteria	Entropy info-gain	Chi-square	Chi-square for categorical variables; J-way ANOVA for continuous/ordinal variables
Pruning	Pre-pruning using a single-pass algorithm	Pre-pruning using a single-pass algorithm	Pre-pruning using Chi-square test for independence	Post-pruning
Dependent variable	Categorical/ Continuous	Categorical/ Continuous	Categorica	Categorical
Input variables	Categorical/ Continuous	Categorical/ Continuous	Categorical/ Continuous	Categorical/ Continuous
Split at each node	Binary; Split on linear combinations	Multiple	Multiple	Binary; Split on linear combinations

Note. Adapted from "Decision tree methods: applications for classification and prediction," by Song, Y., & Lu, Y., 2015, *pmc.ncbi.nlm.nih.gov.*

Random Forest

Random forest (RF) is a non-parametric machine learning technique that predicts outcomes by constructing an ensemble of decision trees, each trained on a bootstrapped subset of the data (King & Strumpf, 2021). It gained substantial attention in healthcare analytics for its

robustness, ability to handle high-dimensional data, and resilience to overfitting (Breiman, 2001). The algorithm's capacity to model non-linear relationships and rank feature importance made it suitable for understanding complex patient behaviors and predicting clinical outcomes.

In treatment adherence, RF was employed to identify patients at risk of medication non-compliance. Marineci et al. (2025b) utilized RF models to analyze electronic health records and demographic data to predict adherence among patients with hypertension. Their findings demonstrated that RF outperformed logistic regression and support vector machines in predictive accuracy and feature interpretability. Lo-Ciganic et al. (2015) applied RF to identify predictors of hospitalization. They used survival tree models that empirically derived optimal adherence thresholds using the proportion of days covered (PDC) to differentiate patients by hospitalization risk most effectively.

Vasilache et al. (2024) assessed and compared the predictive accuracy of four machine learning algorithms, decision tree (DT), naïve bayes (NB), support vector machine (SVM), and random forest (RF) in predicting the occurrence of preeclampsia (PE), intrauterine growth restriction (IUGR), and their associations in a group of singleton pregnancies. The random forest algorithm demonstrated superior performance for predicting PE and IUGR.

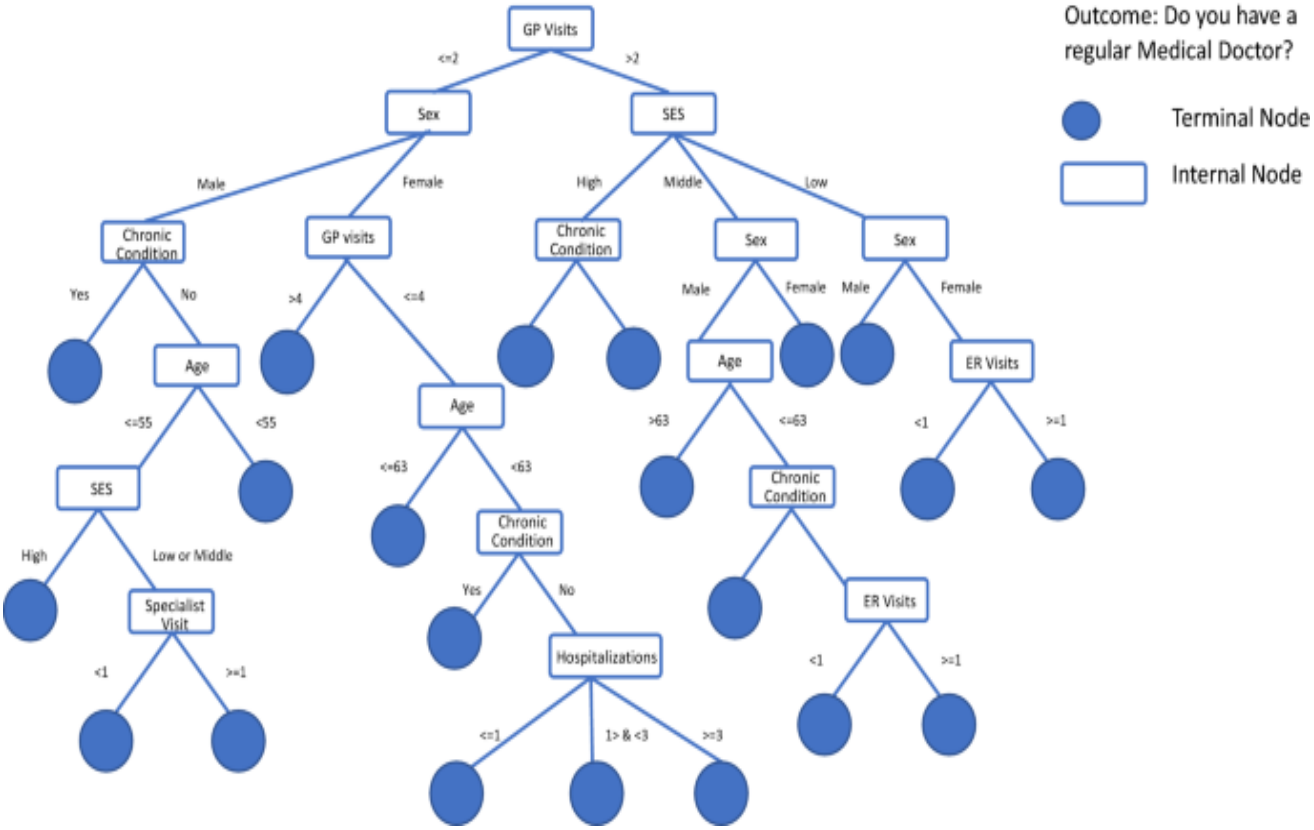
Random forest is frequently used in feature selection due to its inherent ability to rank variable importance. This is particularly valuable in healthcare settings where data are often noisy, incomplete, and multidimensional. Miotto et al. (2016) demonstrated how RF could distill predictive features from electronic medical records (EMRs) to forecast disease progression and patient engagement in chronic disease management.

Despite its strengths, RF lacked the interpretability of single decision trees, which could be a barrier in clinical decision-making contexts. Model-agnostic interpretability techniques,

such as SHAP and LIME, were incorporated into recent research to provide explanations for RF predictions, thereby enhancing trust and usability in clinical practice (Lundberg & Lee, 2017). Figure 3 illustrates the hypothetical tree used for predicting whether a person has a family doctor.

Figure 3

Random Forest



Note. Adapted from "Applying random forest in a health administrative data context: a conceptual guide," by King, C., and Strumpf, E., 2021, *Health Services and Outcomes Research Methodology*, *22*(1), 96–117.

Logistic Regression

Logistic regression (LR) is a foundational method in healthcare analytics, widely used for predicting binary outcomes, such as disease presence, hospital readmissions, and patient behavior. Its statistical robustness and interpretability made it especially valuable in clinical

contexts, where understanding model decisions is crucial (Hosmer & Lemeshow, 2000). Logistic regression was extensively applied to analyze treatment adherence and healthcare engagement.

Beyond behavioral insights, logistic regression is also widely used for clinical risk prediction. It underpinned tools such as the Charlson comorbidity index and LACE score, which assessed readmission risk. Welvaars et al. (2023) developed an LR-based model using demographic and clinical features to identify avoidable 30-day hospital readmissions. A systematic review by Dai et al. (2023) affirmed the centrality of LR in such models, although it noted varying performance depending on the quality of features and the data context. LR was also employed in intraoperative and ICU monitoring and applied logistic regression to predict intraoperative hypotension using real-time hemodynamic data, demonstrating its utility in dynamic, high-stakes environments.

Logistic regression's high interpretability is a significant advantage, allowing clinicians and decision-makers to understand how individual predictors contribute to the outcome, which is critical in clinical environments that demand transparency. The method was grounded in well-established statistical theory, lending credibility and reliability to its use across various medical applications. Logistic regression is highly compatible with clinical variables and behavioral data, allowing for the integration of diverse information sources, including demographics, vital signs, socioeconomic indicators, and patient-reported outcomes.

One key assumption was the linearity of independent variables in the log-odds scale, which could lead to model misspecification when relationships are more complex. It also struggled to capture nonlinear interactions and intricate patterns without extensive feature engineering or transformation. As a result, logistic regression could underperform compared to more flexible models, such as ensemble methods or AutoML systems, which automatically

handle variable interactions and nonlinearities. Rajkomar et al. (2018) demonstrated that while logistic regression remained competitive, deep learning models offered improved predictive performance when working with large-scale electronic health records.

Gradient Boosting Machines (GBM)

Gradient boosting machines (GBMs) have emerged as one of the most powerful and widely used machine learning techniques in healthcare analytics, particularly for modeling complex patterns in patient behavior and clinical outcomes. GBMs built predictive models iteratively by combining weak learners, typically decision trees, to minimize error through a gradient descent approach. This method handles structured data with complex feature interactions, making it particularly suitable for healthcare applications involving clinical, demographic, behavioral, and environmental variables (Chen & Guestrin, 2016).

GBMs were used in patient behavior modeling to predict treatment adherence, health service utilization, and behavioral responses to interventions. Janssoone et al. (2018) applied GBM models to electronic health records and behavioral data to estimate the risk of patient non-adherence, aiming to improve patient support throughout their care path. Rajkomar et al. (2018) demonstrated the effectiveness of gradient boosting in predicting inpatient mortality, unplanned readmissions, and length of stay using large-scale EHR datasets.

Although deep learning models showed slightly better performance, GBMs offered a strong balance of accuracy and interpretability through feature importance metrics. GBMs had limitations in interpretability compared to simpler models, such as logistic regression. Recent advancements in model explainability, such as SHAP (SHapley Additive exPlanations) values, made GBMs more transparent and clinically acceptable (Lundberg & Lee, 2017). These tools enabled stakeholders to understand how each feature affected predictions at both the global and

individual levels, which was crucial for establishing trust and promoting adoption in healthcare settings.

GBMs showed strong performance when integrated into AutoML frameworks. Ikemura et al. (2021) utilized AutoML platforms incorporating GBM variants to predict patient mortality, demonstrating high predictive performance with minimal manual tuning, thereby highlighting their potential in clinical decision support systems. The GBM variants consistently delivered high predictive performance with minimal manual tuning, emphasizing their value in clinical decision support systems. GBMs' ability to handle complex, high-dimensional data and to produce accurate predictions made them particularly suitable for healthcare applications. With the growing availability of structured health data and explainability tools, GBMs were expected to play an increasingly important role in predictive analytics for personalized care, population health management, and behavior-based intervention strategies.

Theoretical Framework

The study’s theoretical framework was grounded in predictive analytics and AutoML (Automated Machine Learning), which automated and optimized key steps such as model selection, hyperparameter tuning, and deployment, making machine learning more accessible and scalable in clinical settings (Rashidi et al., 2021). Predictive analytics in healthcare is essential for estimating patient behaviors such as treatment adherence and healthcare engagement. However, traditional models often fail to integrate complex contextual factors, such as socioeconomic status, environmental conditions, and patient health history, which are critical for understanding individual health trajectories, thereby limiting their predictive accuracy and interpretability (Gowda & Lakshmikantha, 2020).

This study investigated how integrating stress indicators, including heart rate variability and environmental factors such as anesthesia levels and operating room conditions, into an AutoML-based system enhanced predictive accuracy while maintaining model transparency, offering improvements over static, traditional models. The relationship among these concepts was grounded in the understanding that healthcare data is complex and multi-dimensional, necessitating an approach that enhanced prediction accuracy and interpretability for clinical decision-making.

Health Belief Model (HBM)

The study was grounded in multiple theoretical foundations, including health behavior theories such as the Health Belief Model and the Theory of Planned Behavior. These provide insights into how external factors and individual beliefs shaped health behaviors. Computational learning theory informed the machine learning aspect of this research. The Cross-Industry Standard Process for Data Mining (CRISP-DM) offered a structured methodology for implementing predictive analytics in healthcare (Shearer, 2000). Explainable artificial intelligence (XAI) theories ensured that machine learning models remain interpretable for healthcare professionals, thereby facilitating informed decision-making and enhancing transparency.

Literature indicates that machine learning has potential in healthcare applications, with studies by Alsini et al. (2024) demonstrating how predictive modeling benefits from analyzing human characteristics through speech, physical activity, and electronic medical records. However, existing models often lacked interpretability and scalability (Zhou et al., 2022), and this research sought to bridge that gap by integrating multidimensional contextual factors.

The Health Belief Model (HBM) was a psychological framework that explains health-related behaviors by focusing on individuals' perceptions of health risks and their motivations to take preventive actions. Developed in the 1950s, the model suggested that people are more likely to engage in health-promoting behaviors if they believe they are at risk for a health issue, understand the seriousness of its consequences, see the benefits of taking action, and feel they can overcome any barriers to making changes.

Key factors that influenced behavior included perceived susceptibility, the belief that one was at risk for a health problem; perceived severity, the belief that the health issue had serious consequences; perceived benefits, the belief that taking action would reduce risks or improve health; and perceived barriers, the obstacles or challenges in taking preventive actions (Liu et al., 2024).

In addition, HBM included "cues to action," or triggers that prompted individuals to act, such as health reminders, symptoms, or advice from healthcare professionals. Over time, the model expanded to include "self-efficacy," which referred to individuals' confidence in their ability to successfully carry out health behaviors, such as following a treatment plan or making lifestyle changes. If people believed they could succeed in making changes, they were more likely to act.

Research has shown that the Health Belief Model can be applied to various health behaviors, including treatment adherence, vaccination, and lifestyle changes. Liu et al. (2024) found that individuals who perceived themselves as high risk for disease, understood its severity, and recognized the benefits of preventive action were more likely to engage in healthier behaviors. However, barriers such as cost, lack of access, and social factors may prevent individuals from following the recommended actions.

The Health Belief Model (HBM) suggests that individuals' health-related behaviors are influenced by their perceptions of susceptibility to a health issue, the severity of potential consequences, the benefits of taking preventive action, and the barriers to implementing such actions (Subedi et al., 2023). In the context of this study, HBM helped model how patients' perceptions of their health risks and benefits influenced their adherence to treatment and engagement in healthcare. Patients who perceived themselves as at high risk for a condition or believed it had severe consequences were more likely to adhere to treatments.

The AutoML system integrated electronic medical records (EMRs) and stress indicators, such as heart rate variability, to quantify perceived health risks. These systems automate the processes of model selection, data integration, and hyperparameter tuning, enabling more efficient handling of high-dimensional data. Moreover, perceived benefits and barriers, such as financial constraints or healthcare access, were incorporated into predictive models to help personalize interventions. Environmental conditions, including operating room settings and anesthesia levels, served as triggers that influenced patient behaviors. By incorporating these variables, AutoML improved behavior prediction and intervention timing.

Theory of Planned Behavior (TPB)

The Theory of Planned Behavior (TPB), developed by Icek Ajzen in 1985, was a psychological model used to understand and predict human behavior. It proposed that a person's intention to perform a behavior was the primary predictor of whether they would engage in it. This intention is influenced by three factors: attitude toward behavior, subjective norms, and perceived behavioral control (Bosnjak et al., 2020).

Attitude refers to the individual’s positive or negative evaluation of performing the behavior. If they believed the behavior would lead to a favorable outcome, they were likely to

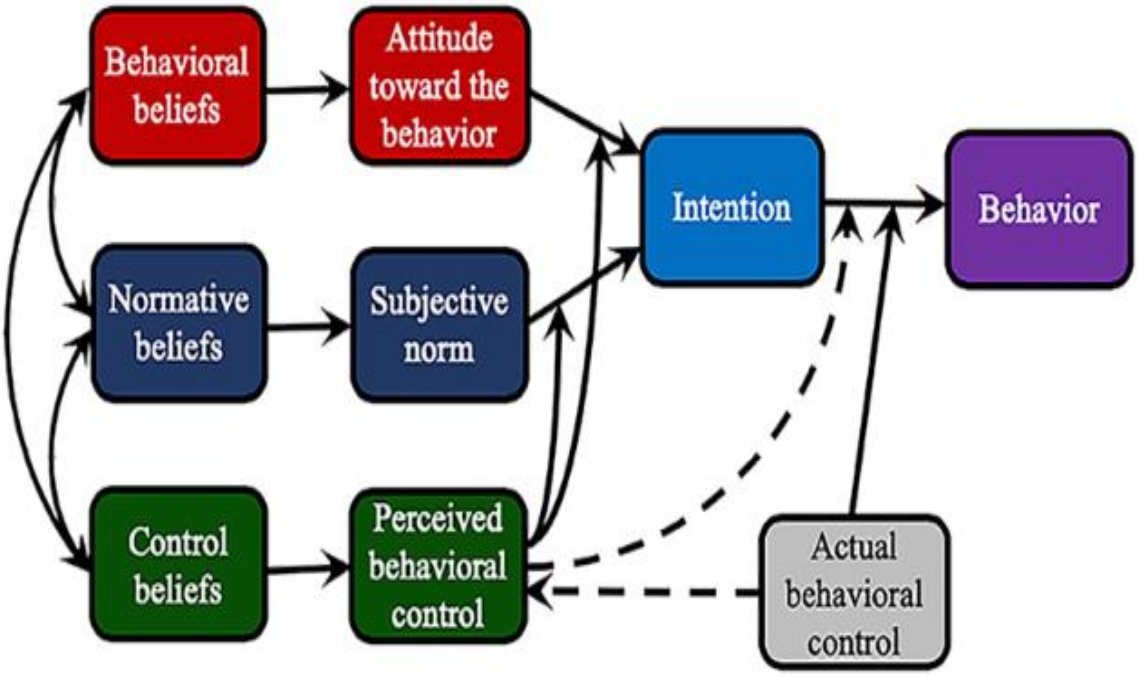
perform it. Subjective norms reflected the perceived social pressures from others, such as family, friends, or society, that influenced an individual's decision to engage in a particular behavior. If a person felt that important others approved of the behavior, they were more likely to adopt it. Perceived behavioral control refers to an individual's belief in their ability to perform a behavior, taking into account both internal capabilities and external factors. They were more likely to intend to act if they believed they had control over their behavior.

The Theory of Planned Behavior (TPB) has been widely applied in health psychology to predict behaviors such as exercise, smoking cessation, and medication adherence. For instance, individuals who have a positive attitude toward exercise, perceive social support, and feel they have control over exercising are more likely to be physically active. Similarly, TPB has been useful in understanding why patients may or may not adhere to prescribed treatments, as attitudes, social support, and perceived barriers influence these behaviors.

This study's inferred attitudes were captured through historical adherence behaviors, subjective norms, and perceived behavioral control, as observed in institutional and clinical patterns such as American Society of Anesthesiologists (ASA) classification, and through intraoperative stress markers and access indicators. Although TPB has been validated in various health psychology contexts, its application in predictive machine learning remains limited. This study advanced the field by incorporating TPB-informed variables into AutoML pipelines to predict better behaviors, such as adherence and engagement, particularly in high-dimensional, real-time clinical data. Figure 4 shows the graphical depiction of the Theory of Planned Behavior.

Figure 4

Graphical Depiction of the Theory of Planned Behavior



Note. Adapted from "The theory of planned behavior: Selected recent advances and applications," by Bosnjak, M., Ajzen, I., & Schmidt, P., 2020, *Europe's Journal of Psychology, 16(3), 352–356.*

The Health Belief Model (HBM) and the Theory of Planned Behavior (TPB) provide valuable theoretical frameworks for understanding health behaviors; their empirical applications in predictive modeling and machine learning remain relatively underexplored but hold promise. Studies using HBM constructs have demonstrated their utility in predicting preventive health behaviors, including vaccination uptake, medication adherence, and self-management of chronic diseases. Models incorporating perceived susceptibility and severity have improved the prediction of diabetes self-care behaviors (Darvishi et al., 2025). When quantified through EMR data and biometric indicators, such as heart rate variability and patient surveys, these constructs

can serve as practical features in machine learning models, thereby enhancing the personalization of interventions.

Empirical applications of TPB in predictive analytics have similarly yielded valuable insights, especially in behavioral intention modeling. The constructs proposed by Ajzen were translated into measurable variables using survey responses and behavioral data to effectively predict health-related behaviors such as smoking cessation, dietary habits, and adherence to physical activity routines (Bosnjak et al., 2020).

Machine learning models that incorporate attitudes, derived from sentiment analysis of clinical notes or self-reported outcomes, subjective norms, and perceived behavioral control, such as insurance status and access to healthcare services, have shown improved predictive accuracy. Integrating neural networks further enhanced prediction performance, illustrating the value of combining psychological theory with advanced machine learning techniques (Park et al., 2025).

However, challenges persisted in operationalizing these psychological constructs for computational use. Many empirical studies still rely on traditional statistical approaches, which limit scalability and adaptability in high-dimensional healthcare datasets. Efforts to embed behavioral theories into data-driven frameworks, such as AutoML pipelines, have demonstrated a growing interest in blending psychological insights with algorithmic efficiency. Such hybrid approaches enhance model relevance and promote interpretability by aligning predictors with established constructs from behavioral science.

Drawing from both HBM and TPB provided a theoretically grounded approach to feature engineering and model interpretation. Embedding these constructs into AutoML pipelines, utilizing variables such as perceived risk, socioeconomic barriers, and behavioral intention,

helped create more nuanced, actionable, and patient-centered predictive models. Future research should refine the process of translating theoretical models into structured data, ensuring both empirical rigor and clinical applicability.

Computational Learning Theory (CLT)

Computational Learning Theory (CLT) provided the mathematical and theoretical foundation for understanding how machine learning models generalized from data to make accurate predictions (Du et al., 2025). In this study, CLT played a crucial role in evaluating the learning efficiency, complexity, and generalization capabilities of the AutoML-based system in comparison to traditional models. CLT helped assess how well the AutoML system generalized new patient data while minimizing overfitting.

Given the complexity and high dimensionality of healthcare data, including intraoperative stress markers, socioeconomic indicators, and time-series vital sign trends, ensuring that machine learning models generalized well beyond the training data was essential. Overfitting was a common risk when models memorized noise or specific patterns in the training data, rather than capturing the underlying causal or generalizable relationships.

The Vapnik-Chervonenkis (VC) dimension, a foundational concept in statistical learning theory, provided a formal way to quantify the capacity or complexity of a model class. A model with a higher VC dimension could represent more complex functions. However, it was also at a higher risk of overfitting, especially when the sample size was limited or the features were noisy (V. N. Vapnik, 1998). In the context of AutoML systems, controlling model complexity through constraints on VC dimension helped ensure that the selected models maintained a balance between expressiveness and generalizability.

Complementing the VC framework, Probably Approximately Correct (PAC) learning theory further supported model selection by formalizing what it meant for a model to learn reliably from data. A hypothesis is said to be PAC-learnable if, with high probability, it could achieve a small error on unseen data, provided it had seen enough training examples (Valiant, 1984).

This approach was particularly relevant for health behavior prediction tasks where labeled data may be sparse or imbalanced. AutoML frameworks informed by PAC learning principles prioritized algorithms that were not only empirically accurate but also provably robust under distributional shifts. When applied to datasets like VitalDB, these theoretical tools helped identify models that generalized across patient populations, surgical procedures, and contextual factors, thus improving real-world applicability and decision support utility.

The Probably Approximately Correct learning framework determined the minimum number of patient records needed to achieve reliable performance, offering a statistical guarantee of model accuracy given an error tolerance (ε) and confidence level (δ). This framework ensured that AutoML-driven healthcare models maintained statistical rigor while making clinically relevant predictions. (Cherian et al., 2024)

The bias-variance tradeoff concept in CLT directly impacted the predictive performance of AutoML-generated models. High-bias models, such as simple linear regression, risk underfitting and fail to capture complex patient interactions, while high-variance deep learning models with excessive parameters can overfit the training data (Hastie et al., 2013).

AutoML frameworks mitigated this tradeoff by incorporating cross-validation strategies, such as k-fold cross-validation, and regularization techniques, including Ridge and Lasso regression, to dynamically adjust model complexity (Eloutouate et al., 2025). By systematically

balancing bias and variance, AutoML optimized predictive performance while maintaining model interpretability and robustness, making it a valuable tool in healthcare predictive analytics.

Predicting patient behavior required a large, diverse dataset to avoid biases. CLT provided insights into the required number of patient records, stress measurements, and environmental parameters to achieve reliable predictions. By understanding sample complexity, this study ensured that the AutoML system efficiently learned from available data without unnecessary computational overhead.

CLT principles helped balance bias and variance in predictive models. In an AutoML setting, models were optimized to find the best tradeoff using techniques such as regularization and cross-validation. By incorporating contextual factors like stress levels and healthcare access, the study aimed to refine the learning process to improve predictive accuracy while maintaining model stability.

One challenge in healthcare AI was ensuring that models are interpretable for clinicians. CLT contributed by setting theoretical boundaries on what models could learn given the constraints of healthcare data. By leveraging CLT, the study ensured that the AutoML system selected models that maximized predictive accuracy while remaining interpretable.

Computational learning theory provided a formal framework for understanding how algorithms learned from data, helping balance model complexity and generalization performance. Key concepts, such as the Vapnik-Chervonenkis (VC) dimension and Probably Approximately Correct (PAC) learning, enabled AutoML algorithms to evaluate the trade-offs between underfitting and overfitting. These principles guided the selection of complex models to capture essential patterns, such as patient stress responses or treatment adherence behaviors, without becoming overly tailored to the training data (Shen et al., 2018).

The principles of CLT guided AutoML's feature selection, model evaluation, and hyperparameter tuning to optimize performance. Given that patient behavior was influenced by complex, multi-dimensional factors, understanding the theoretical limits of learning helped improve the robustness and reliability of the system. Key assumptions underlying this study included the inherent complexity of healthcare data, the ability of AutoML systems to handle such complexity efficiently, and the critical need for interpretability to ensure trust and adoption in clinical environments.

Healthcare data, such as that found in the VitalDB dataset, was multidimensional and heterogeneous, ranging from real-time physiological signals, like heart rate and oxygen saturation, to static clinical records, including ASA classification and comorbidities. Traditional machine learning models often required manual preprocessing, feature engineering, and hyperparameter tuning, which were time-consuming and subject to bias.

AutoML addressed these challenges by automating the end-to-end model development process, selecting optimal algorithms, tuning hyperparameters, and validating performance, all while minimizing human intervention. Evaluating model performance using robust metrics, such as ROC-AUC, F1-score, and precision-recall curves, ensured a comprehensive understanding of each model's ability to identify high-risk patients while minimizing false positives and negatives. Moreover, maintaining interpretability by leveraging transparent models or post-hoc explanation tools ensured clinicians could understand and trust the system's predictions, which was essential for integrating AI into decision-making workflows.

This study explored the relative contributions of these individual contextual factors to model performance, providing a structured analysis of their impact. Alternative frameworks such as deep learning, Bayesian networks, and ensemble learning were considered; however, they

were deemed less suitable due to limitations in interpretability, computational efficiency, and automation. Given the study's emphasis on automation, scalability, and interpretability, AutoML integrated with CRISP-DM is the most appropriate framework.

By uniting behavioral theory with formal learning principles and modern AutoML infrastructure, this study offered a novel, multi-layered framework for healthcare predictive modeling. The approach provided not only predictive strength but also interpretability and theoretical coherence. Future research should continue to refine the operationalization of behavioral constructs into machine learning features and explore how theoretical rigor can enhance algorithmic fairness and clinical adoption.

Data Ethics and Legal Frameworks

Data ethics was a foundational element in data science, ensuring that data collection, processing, analysis, and deployment were aligned with moral principles that protected individual rights, privacy, and fairness. Sudhaman et al. (2023) presented a lifecycle-based ethical framework that identified and mitigated risks at each phase of a data science project. In healthcare, ethical lapses, such as biased models or opaque decision-making, could result in real harm, disproportionately affecting vulnerable or marginalized populations (Mehrabi et al., 2021).

Core ethical principles included privacy, fairness, transparency, and accountability. Privacy protections required safeguarding personal health information from unauthorized access and misuse (Smith Anderson, 2025). Fairness entailed rigorous bias auditing to prevent algorithmic discrimination, especially where training data underrepresented certain groups (Obermeyer et al., 2019). Transparency involved developing explainable models to clinicians and patients, reinforcing trust and informed consent (Doshi-Velez & Kim, 2017). Accountability

emphasized that systems had to be auditable and traceable, with clear protocols to assign responsibility for outcomes and decisions (Atlan, 2024).

Informed consent and autonomy remained foundational in biomedical ethics, as articulated in the Belmont Report and the Declaration of Helsinki (Ehni & Wiesing, 2024). However, the rise of AI introduced nuanced challenges; patients were often unaware that decisions were shaped by opaque algorithms, and real-time data collection through wearables or EHRs often occurred without active re-consent. Morley et al. (2020) argued that traditional informed consent was not sufficient in AI contexts, prompting calls for dynamic or tiered consent models tailored to AI-driven analytics.

Legal and Ethical Context of This Study

This study was governed by various ethical and legal considerations, ensuring compliance with data protection laws and ethical guidelines in healthcare research. Ethical concerns included patient privacy, data security, informed consent, and algorithmic fairness. In the United States, the Health Insurance Portability and Accountability Act (HIPAA) mandated strict safeguards for electronic protected health information (ePHI), requiring access control, encryption, and audit mechanisms. The General Data Protection Regulation (GDPR), applicable to European data subjects, granted individuals control over their data and imposed obligations for data minimization, transparency, and the right to explain algorithmic decisions (Voigt & von dem Bussche, 2017). While both frameworks provided robust protection, scholars noted challenges in operationalizing GDPR's "right to explanation" in complex AI systems, particularly regarding model opacity (Wachter et al., 2017).

Beyond data privacy, this study's ethical obligations include ensuring fairness in model predictions, transparency in outcomes, and mechanisms for addressing unintended harm. To

meet these expectations, this study incorporated algorithmic auditing and explainability tools such as SHAP and LIME to promote equitable and interpretable outputs. The study adhered to Institutional Review Board (IRB) protocols as defined under the Common Rule, which governs research involving human subjects in the U.S. and mandates respect for persons, beneficence, and justice.

Ethical theories further informed the development and deployment of AI models in this study. A utilitarian perspective supports maximizing benefits through improved predictive accuracy and patient outcomes (Gabriel, 2020), while deontological ethics emphasizes adherence to rules such as data confidentiality and informed consent (Islam, 2024). A virtue ethics lens emphasized the character of researchers and institutions, promoting integrity, trustworthiness, and a commitment to minimizing harm (Lee, 2020).

Established Frameworks Guiding Ethical Data Science Practices ensured compliance with legal standards to foster trust and promote the well-being of individuals and communities. Fair Information Practices (FIPs) were guidelines for handling personal data to protect privacy and ensure accuracy. These included transparency, informing people about how their data was used, consent, getting permission before collecting data, data minimization, only collecting the necessary data, access, and correction, allowing people to view and fix their data, and accountability, ensuring organizations followed these rules (*Fair Information Practice Principles (FIPPs)*, n.d.).

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems offered detailed guidelines to ensure systems were developed ethically and aligned with human values. Key aspects of the initiative included prioritizing human rights by protecting and promoting them in system design, ensuring systems contributed positively to human well-being and societal

good, and considering environmental sustainability in the design process. It also emphasized transparency and explainability, making system operations understandable to users, accountability, and establishing clear responsibility for system actions and outcomes.

In healthcare, the Health Insurance Portability and Accountability Act (HIPAA) set national standards for protecting health information, requiring entities to implement safeguards to maintain the confidentiality, integrity, and security of electronic protected health information (ePHI). Recent proposals aimed to strengthen cybersecurity measures, including the implementation of mandatory multi-factor authentication and encryption standards. The Common Rule, a federal policy, governed the ethical conduct of research involving human subjects, emphasizing principles such as respect for persons, beneficence, and justice, and requires Institutional Review Board (IRB) approval to ensure that ethical standards were met. IRB approval for this study is included in Appendix B.

Open Data Initiatives

Open data initiatives promote transparency, reproducibility, and equitable access to healthcare insights. The National Institutes of Health (NIH) Open Data Initiatives encouraged data sharing for research while enforcing strict ethical guidelines to protect patient privacy (National Institutes of Health, 2022). Although primarily focused on agriculture, Ethical considerations in social justice and data science were fundamental in healthcare AI, as algorithmic biases could disproportionately affect marginalized communities, leading to disparities in healthcare outcomes.

To address these challenges, ethical AI design had to incorporate bias mitigation strategies that ensured fairness in training datasets (Mehrabi et al., 2021), transparency and explainability in model decision-making (Doshi-Velez & Kim, 2017), and equitable access to

healthcare to prevent the reinforcement of systemic inequalities (Obermeyer et al., 2019). This study ensured that AutoML-driven healthcare analytics aligned with moral responsibilities, legal compliance, and social justice imperatives.

Stringent legal and ethical frameworks governed the study of patient behavior using data science, ensuring that patient rights, safety, and data integrity were safeguarded. In the U.S., the Health Insurance Portability and Accountability Act (HIPAA) mandated the privacy and security of protected health information (PHI), requiring patient consent for data sharing and imposing strict security measures (Stevens, 2003). In the European Union, the General Data Protection Regulation (GDPR) enforced data collection, processing, and storage, granting individuals rights such as data portability and the right to be forgotten guidelines (Voigt & Von Dem Bussche, 2017).

The Common Rule (45 CFR 46) established ethical principles for human subjects' research, including informed consent and oversight by an Institutional Review Board (IRB). International ethical standards, such as the Belmont Report and the World Medical Association's Declaration of Helsinki (Ehni & Wiesing, 2024), further emphasize the importance of respecting individuals, promoting beneficence, and ensuring justice in medical research involving human subjects. These regulations collectively ensure that patient behavior data was collected and analyzed ethically, minimizing privacy risks, discrimination, and misuse while enforcing robust mechanisms for informed consent, de-identification, and ethical review by institutional boards.

Summary

This literature review explored the comparative effectiveness of AutoML and traditional ML in predictive healthcare analytics, particularly in integrating stress and socioeconomic factors. AutoML frameworks, such as Google AutoML and H2O.ai, streamlined model selection,

hyperparameter tuning, and feature engineering, making ML more accessible. Traditional ML, on the other hand, required expert-driven feature selection and model optimization. While studies have shown that AutoML can match or exceed traditional ML in predictive accuracy, its interpretability remains a challenge. A key gap in the literature was the limited exploration of stress-related biomarkers, such as heart rate variability and cortisol, and socioeconomic determinants, including income, education, and access to care, in ML-driven healthcare analytics.

Research suggested that incorporating these factors could enhance the robustness and fairness of models. However, biases in socioeconomic data, ethical concerns, and the need for explainable AI posed significant challenges. Studies using datasets like MIMIC-III (Johnson et al., 2016) often incorporated limited demographic features but rarely integrated detailed socioeconomic data, which could lead to potential biases in behavioral predictions.

Other research, leveraging national surveys such as NHANES and BRFSS (Centers for Disease Control and Prevention, 2022), demonstrated the importance of SES in health outcomes. However, it typically lacked the high-frequency clinical data found in datasets like VitalDB. AutoML applications in healthcare, as demonstrated by Waring et al. (2020), have shown promising improvements in prediction accuracy; however, they often overlook the integration of contextual variables due to data limitations. The gap was particularly pronounced in high-frequency, perioperative datasets, such as VitalDB, where rich physiological signals were available but social context was lacking.

The literature also revealed diverging opinions regarding the extent to which AutoML could replace traditional ML, with some studies advocating for hybrid approaches that integrate domain expertise. This review emphasized the need for additional research on optimizing feature

selection, mitigating biases, and enhancing interpretability in AutoML-based healthcare predictions. The research methodology outlined the effectiveness of AutoML versus traditional ML in terms of predictive accuracy, while incorporating stress and socioeconomic variables.

This study yielded a comprehensive suite of outcomes that supported both technical rigor and clinical applicability. By leveraging the VitalDB dataset and a robust AutoML framework, the research produced validated machine learning models capable of accurately predicting patient behaviors such as treatment adherence and follow-up engagement. These models were developed through systematic preprocessing, feature engineering, and algorithmic selection, ensuring resilience to data quality issues and high-dimensional complexity. Integrating explainable AI techniques will enable global and patient-specific interpretability, translating model predictions into meaningful insights for healthcare providers.

The study produced interactive visualizations and structured reports to effectively communicate trends, risk factors, and decision rationales. These outputs aimed to bridge the gap between data-driven predictions and clinical decision-making, ultimately supporting more targeted interventions and improved patient outcomes. Table 2 below summarizes the expected outcomes.

Table 2

Summary of Expected Outcomes

Category	Expected Outcome
Model Performance Metrics	Accuracy, Precision, Recall, F1-Score for classification tasks - predicting non-adherence
	AUC-ROC, PR AUC for imbalanced behavior prediction tasks

Category	Expected Outcome
	RMSE, MAE for regression (if applicable)
	Calibration plots and confusion matrices to validate prediction reliability
AutoML vs. Traditional ML	Comparative analysis of AutoML-selected vs. manually tuned models, GBM, and LSTM
	Evaluation of pipeline efficiency - time, resource usage
	Comparison of model robustness via cross-validation scores
Dashboards & Visualization	Interactive dashboards – Streamlit for real-time model development insights, Feature importance plots, and Interactive filters Power BI displays patient-level predictions, color-coded risk levels, behavioral trends, interactive filters, drill-down capabilities, and tooltips for explaining model output.
	Visualization of time-series vitals with overlayed predictions
	Population-level behavior prediction summaries stratified by contextual variables - ASA classification, procedure type
Explainable AI (XAI) Outputs	Global SHAP summary plots showing top features - blood pressure variability, stress proxy
	LIME-based local explanations for individual patient predictions
	Feature interaction visualizations to show compound effects - ASA × vitals
	Identification of bias or instability through explanation consistency
	Summary statistics on missingness, imputation performance - KNN vs. mean

Category	Expected Outcome
Data Quality & Integration Outputs	Signal smoothing diagnostics and outlier detection reports
	Correlation matrices and feature alignment diagnostics across modalities
	Quality metrics of synthetic features -stress scores, recovery indices
Model Development & Robustness	Summary of hyperparameter optimization results - best GBM params
	Learning curves and validation loss trends for LSTM
	Cross-validation stratification outcomes - behavior prediction by procedure type
	Overfitting mitigation evidence - dropout performance, regularization strength
Reporting & Insights	Comprehensive tables comparing model performance by algorithm and prediction task
	Visual reports explaining model decision paths for clinical review
	Quantified impact of contextual factors - socioeconomic proxies on predictions
	Narrative synthesis of how vital signs and perioperative data relate to patient behavior

In conclusion, the literature highlighted the evolving landscape of machine learning in healthcare, with AutoML emerging as a scalable and accessible alternative to traditional ML. While AutoML demonstrated promising predictive performance and efficiency, challenges related to interpretability, ethical concerns, and data biases remained. Integrating stress-related

biomarkers and socioeconomic variables emerged as a critical area for enhancing the fairness and robustness of predictive models.

Despite these opportunities, the literature revealed gaps in empirical evidence, especially concerning real-world datasets that reflected complex patient behaviors. This gap highlighted the need for comprehensive, interdisciplinary research that struck a balance between technical innovation and clinical relevance.

The subsequent chapters of the study delved into the design and experimental methodology, dataset selection, data analysis, model processing, and model development. Chapter 3 will describe the quasi-experimental design, data source, specifically the VitalDB dataset, preprocessing techniques, and the comparative framework for evaluating AutoML and traditional ML approaches. The chapter will also elaborate on integrating stress and socioeconomic variables, model evaluation metrics, and implementing explainable AI to ensure transparency and clinical usability. This methodological foundation will guide the development of predictive models that achieve high accuracy and offer interpretable, actionable insights for healthcare decision-makers.

Chapter 3: Research Methodology

Accurately predicting patient behaviors, such as treatment adherence, healthcare engagement, and responses to interventions, has been a persistent challenge in contemporary healthcare. Traditional predictive models often overlook complex contextual variables, including environmental conditions, socioeconomic status, and individualized clinical trajectories, which diminish their predictive accuracy and clinical utility. The emergence of machine learning (ML) and Automated Machine Learning (AutoML) offered powerful tools for analyzing high-dimensional healthcare data and developing data-driven insights. To be clinically effective, predictive models must perform accurately and provide interpretable, contextually relevant outputs that support safe and equitable decision-making (Ennab & Mcheick, 2024).

This study utilizes big data and machine learning to address the challenge of accurately predicting patient behavior, including treatment adherence, healthcare engagement, and responses to interventions. Traditional models often fail to account for complex contextual factors, such as environmental conditions, socioeconomic status, and individual health histories, which limits their predictive accuracy and interpretability in clinical settings (Gowda & Lakshmikantha, 2020).

This chapter outlines the research methodology and design used to address the study's problem and purpose. This study employs a quantitative explanatory approach within a quasi-experimental design, which was suitable for retrospective, observational datasets such as the VitalDB dataset. This methodological choice enables robust statistical analysis and hypothesis testing, while acknowledging the practical and ethical constraints of conducting randomized controlled trials in clinical settings (Creswell & Creswell, 2018). The VitalDB dataset includes high-resolution time-series physiological signals, treatment records, and socioeconomic

indicators, providing a multidimensional foundation for building predictive models that reflect real-world clinical complexity (Lee et al., 2022). The methodological research design process is outlined in Figure 5.

Figure 5

Research Design Process Diagram

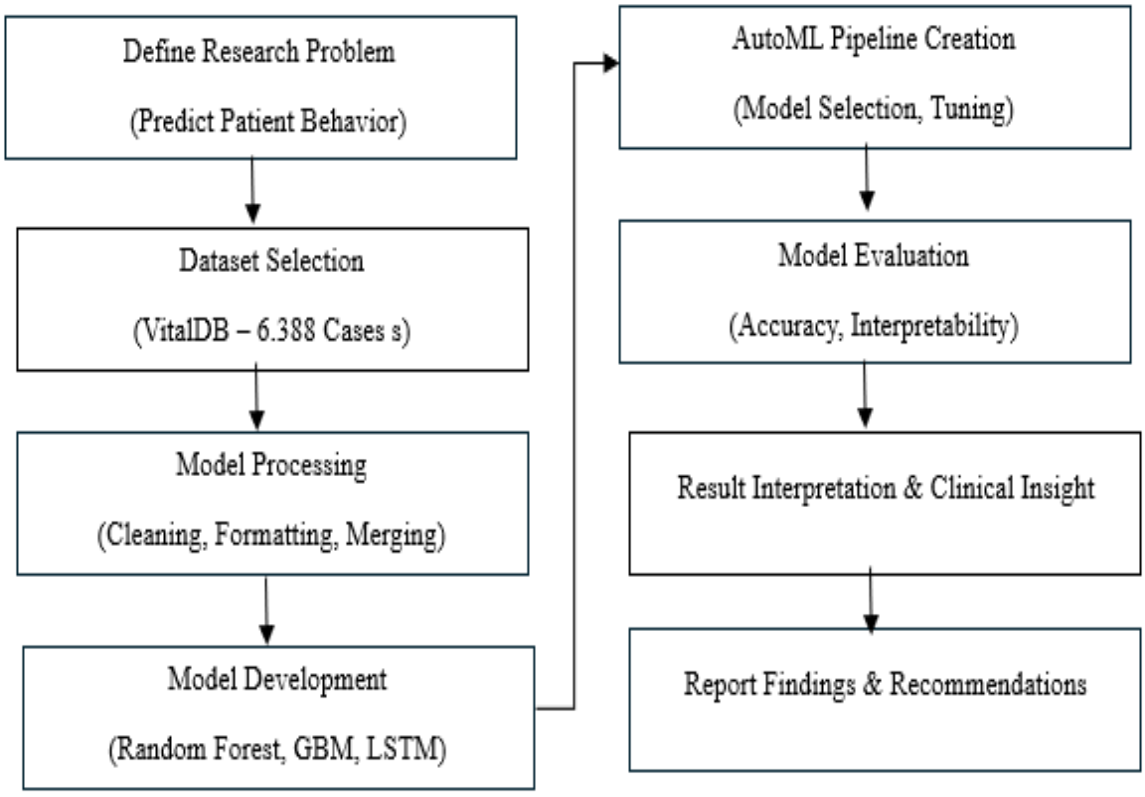


Figure 5 illustrates the research design process, where each phase was directly mapped to the study's objectives and reflected best practices in ML-driven healthcare research (Xian et al., 2025). The process began with data acquisition and extraction from the VitalDB repository, which offers a comprehensive dataset from over 6,000 surgical cases, capturing intraoperative monitoring, medications, and clinical events. Next, data preprocessing involved noise reduction,

imputation of missing values, normalization, and temporal alignment, critical steps for ensuring model validity and fairness.

Following preprocessing, feature engineering created interpretable variables from raw signals and structured records, incorporating clinical and contextual factors such as stress markers and environmental influences. These engineered features served as the basis for both traditional ML models, including random forest, gradient boosting, and long short-term memory (LSTM) networks, and AutoML systems that automated model selection and hyperparameter tuning (He et al., 2021).

To evaluate the models, this study utilized performance metrics such as AUC-ROC, F1-Score, RMSE, precision, and recall, which were considered standard benchmarks for healthcare predictive analytics (Chicco & Jurman, 2020). In addition, stratified sampling and propensity score matching were applied to control for subgroup variation and confounding variables, thereby strengthening internal validity without the need for randomization (Austin, 2011). To enhance model transparency and ensure clinical relevance, explainable AI techniques, including SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations), were used to interpret model outputs (Ribeiro et al., 2016; Lundberg & Lee, 2017).

By integrating these stages into a structured, reproducible workflow, the study not only tested the performance of AutoML versus traditional ML approaches but also contributed to the responsible design of interpretable, context-aware models in healthcare. This methodology was designed to support safe, equitable, and explainable machine learning applications in clinical environments, where accuracy had to be matched with transparency and ethical rigor.

Research Methodology and Design (Nature of the Study)

This study adopted a quantitative explanatory methodology to evaluate the effectiveness of AutoML systems in modeling patient behaviors within a surgical context. This approach enabled rigorous hypothesis testing regarding the influence of contextual variables on predictive accuracy, while facilitating a comparative analysis between traditional ML techniques and AutoML frameworks. The methodology aligned with the structured, time-series nature of the VitalDB dataset, which included high-frequency physiological signals, clinical metrics, and treatment event logs, data types well-suited for quantitative analysis and machine learning evaluation.

The study employed a quasi-experimental, non-equivalent control group design, as defined by Creswell and Creswell (2018). This design was appropriate given the retrospective, observational nature of the VitalDB dataset, which contained real-world clinical data collected during routine care without experimental manipulation. Because random assignment or clinical intervention post hoc was neither feasible nor ethical, a quasi-experimental approach allowed for meaningful causal inference through statistical control techniques.

To strengthen internal validity, stratified sampling and propensity score matching (PSM) were implemented within the modeling workflow. Stratified sampling ensured balanced representation across key subgroups, including age, ASA classification, and surgery type, thereby preventing the overrepresentation of dominant categories during both the training and testing phases. PSM was applied to match patients with similar baseline characteristics, such as comorbidities and baseline stress markers across comparator groups, particularly when evaluating differential model performance or contextual influences on behavior prediction.

Matched samples were used in multivariate regression models and classification tasks to isolate the effects of intraoperative stress indicators and other contextual factors.

The quasi-experimental structure was further reinforced through covariate adjustments and multivariate outcome modeling, enabling the study to minimize biases and account for potential confounders. These methods, widely validated in health informatics and epidemiology, served as practical alternatives to randomized controlled trials when working with large-scale observational data.

The healthcare domain often presents unique ethical and practical challenges, making authentic experimental designs infeasible. Randomly withholding treatments or modifying patient care protocols solely for research purposes posed unacceptable risks to patient safety and violated ethical standards. Quasi-experimental designs were essential in contexts where proper randomization was impossible, yet the research aimed to establish causal inferences through careful comparison and statistical control (Creswell & Creswell, 2018).

The VitalDB dataset offered the temporal granularity necessary to model dynamic patient behaviors, including treatment adherence, intervention responsiveness, and postoperative monitoring compliance. These behaviors were operationalized using structured, observable indicators. Treatment adherence was inferred from intraoperative drug administration timestamps and protocol compliance. Healthcare engagement was derived from monitoring alarm responses, hemodynamic stability, and intraoperative documentation. Response to interventions was evaluated through time-aligned changes in vital signs following medication administration, ventilator adjustments, or surgical milestones.

AutoML and traditional ML models were evaluated using standard predictive metrics, including AUC-ROC, F1-Score, precision, recall, and RMSE to assess model performance.

These objective criteria allowed for consistent benchmarking across modeling approaches and facilitated reproducibility.

Alternative research designs were considered but deemed unsuitable for the purpose. Experimental designs requiring randomization were excluded due to ethical and logistical constraints associated with the retrospective nature of the clinical data. Qualitative and mixed-methods designs were not appropriate for the study's goal of statistically benchmarking predictive models. Cross-sectional surveys, often used to assess perceptions or behaviors at a single point in time, were insufficient for capturing the temporal dynamics inherent in the longitudinal VitalDB dataset.

By applying a rigorous quantitative quasi-experimental design, this study aimed to generate empirical evidence on the comparative strengths of AutoML frameworks in healthcare prediction tasks. As these platforms became more accessible, understanding their utility relative to traditional modeling approaches was essential for informed adoption. Integrating contextual predictors, such as intraoperative stress levels, treatment settings, and clinical interventions, moved beyond physiology-only models, contributing methodologically and practically to advancing precision predictive analytics in surgical care.

Population and Sample

The target population for this study consisted of surgical patients treated at Seoul National University Hospital (SNUH) between 2005 and 2017, whose intraoperative and perioperative data were included in the publicly available VitalDB dataset (Lee et al., 2022). This population consisted of 6,388 surgical cases, each containing high-resolution, multi-parameter intraoperative data, including 196 variables such as heart rate, SpO_2, blood pressure, $EtCO_2$, and EEG. Additionally, perioperative clinical records included 73 parameters, including

ASA classification, comorbidities, and medications, and time-series laboratory values comprise 34 parameters. Key patient characteristics included demographic data, age, sex, weight, height, clinical indicators, ASA classification, surgery type, anesthesia method, and a rich set of physiological monitoring signals recorded at 1-second (numeric) and 100 Hz (waveform) resolutions. The VitalDB dataset, sourced from https://vitaldb.net and documented in Lee et al. (2022), offered a comprehensive and granular foundation for studying perioperative contexts and behavioral outcomes.

This population was well-suited for the study's aim of predicting postoperative behaviors such as treatment adherence and healthcare engagement based on intraoperative stress indicators and clinical context. The dataset's depth and real-world granularity provided a strong foundation for training and evaluating predictive models. The sampling frame encompassed all 6,388 cases, ensuring comprehensive coverage without the need for selective enrollment. Bias was minimized through the automated collection of signals and the inclusion of all eligible surgical procedures, thereby supporting a diverse and representative cohort across patient risk profiles and clinical pathways.

Given the target population size (N ≈ 6,388 surgical cases), the adequate sample size for predictive modeling was determined using power analysis and sample size formulas for finite populations (Cochran, 1991; Lohr, 2021). The minimum sample size n_0 For a given margin of error (e), confidence level (Z), and estimated proportion (p), it was initially calculated as:

$$n_0 = \frac{Z^2 \times p \times (1 - p)}{e^2}$$

Where:

- Z = 1.96 for 95% confidence
- p = 0.5 (conservative estimate maximizing variance)

- e = 0.05 (5% margin of error)

Substituting the values:

$$n_0 = \frac{(1.96)^2 \times 0.5 \times (1 - 0.5)}{(0.05)^2} = 384.16$$

Because the target population is finite (N = 6,388), apply the finite population correction:

$$n = \frac{n_0}{1 + \left(\frac{n_0 - 1}{N}\right)}$$

$$n = \frac{384.16}{1 + \left(\frac{384.16 - 1}{6388}\right)} \approx 363$$

Thus, a minimum sample of approximately 363 cases was sufficient to achieve 95% confidence with a 5% margin of error for general estimates. This study leveraged the full available population (*n* = 6,388 surgical cases) or large stratified subsamples drawn from it, far exceeding the minimum sample size requirements determined through formal statistical calculations. Utilizing the full dataset enhanced model stability, enabled reliable subgroup analyses across age groups, surgery types, and ASA classifications, and provided stronger statistical power to detect minor to moderate effects. Additional justifications further supported the adequacy of the sample size: following the machine learning rule of thumb requiring at least 10–30 cases per predictor variable (Baeza-Delgado et al., 2022), the large sample ensured sufficient events per variable (EPV) to mitigate risks of overfitting, especially given the inclusion of 196 intraoperative physiological variables and other clinical features.

Moreover, typical cross-validation procedures, such as 70/30 or 80/20 train/test splits, retained thousands of cases for training and hundreds for testing, thereby supporting model validation and generalizability. The use of stratified random sampling additionally ensured balanced representation across key clinical and demographic subgroups, further reducing

sampling bias and enhancing the study's external validity. This approach reduced sampling bias and supported robust generalization, consistent with recommendations in stratified sampling literature (Lohr, 2021).

Materials or Instrumentation

This research utilized various tools and resources, each selected for its relevance in building and evaluating predictive models of patient behavior using AutoML techniques. The VitalDB dataset served as the primary data source, providing high-resolution, multi-parameter data from over 6,000 surgical cases, including intraoperative, perioperative, and laboratory variables, which were crucial for modeling treatment adherence, healthcare engagement, and intervention responses. The data used in this study were obtained from the publicly available VitalDB dataset, which contains high-resolution intraoperative physiological and clinical data. The dataset is licensed under the Creative Commons Attribution 4.0 International License https://creativecommons.org/licenses/by/4.0/, allowing redistribution and adaptation for any purpose, including commercial use, provided that proper credit is given and any modifications are indicated. The dataset was accessible via APIs and was compatible with Python libraries. Python was chosen as the primary programming language due to its strong ecosystem for data science and machine learning, while R was used for specialized statistical analysis.

Google AutoML is ideal for users seeking simplicity and ease of use, particularly those already embedded in the Google Cloud ecosystem. It supported rapid model deployment with minimal coding, making it accessible for non-experts. Amazon SageMaker, on the other hand, offered a more robust and flexible environment, well-suited for teams that needed a comprehensive suite of machine learning tools and scalability across the AWS infrastructure. H2O.ai distinguished itself with its open-source model and adaptability, offering a wide range of

algorithms and tools that appealed to novice users and experienced data scientists requiring advanced customizability. It supported a broad range of algorithms, automated feature engineering, hyperparameter tuning, and model ensembling, offering a comprehensive representation of AutoML capabilities. Microsoft Power Platform was a strong choice for organizations already utilizing Microsoft products. Its low-code environment enabled users to easily build and deploy machine learning models, integrating seamlessly into existing workflows through tools like Power BI and Power Apps, without requiring deep technical knowledge.

Hence, for AutoML, open-source H2O.ai AutoML was leveraged to compare model performance across platforms. H2O.ai AutoML was selected because it was one of the most widely adopted and validated AutoML platforms in academic research and industry practice. Given its robustness, scalability, and proven performance across diverse datasets, H2O.ai AutoML was well-suited to serve as a representative AutoML system for this study. Traditional machine learning models were built using libraries such as Scikit-learn, XGBoost, and Keras/TensorFlow for more advanced neural network architectures, including LSTMs.

SHAP and LIME were applied to interpret complex models, offering transparency into how features contributed to predictions. Pandas and NumPy were used for data preprocessing and transformation, while Matplotlib and Seaborn supported exploratory data analysis and visualization. SciPy and Statsmodels were employed for statistical testing and hypothesis validation, including regression and significance testing.

Streamlit could be used to develop an interactive, web-based interface for showcasing model predictions and exploratory visualizations. This tool enables stakeholders, including clinicians and researchers, to interact with the models and interpret results in real-time. Power BI could be integrated for business intelligence and reporting, creating dashboards that summarize

model performance metrics, patient behavior patterns, and key insights in a visually accessible format for non-technical users. All tools listed are either open-source or available under academic licensing terms. Any proprietary resources, such as Google AutoML or Power BI, were accessed under institutional agreements, with proper permissions and acknowledgments noted in the study documentation.

Operational Definitions of Variables

This study was structured around a predictive modeling framework to analyze patient behaviors, specifically treatment adherence, healthcare engagement, and response to interventions, using a range of physiological and contextual data from the VitalDB dataset. The variables were categorized as independent (predictor) and dependent (criterion) variables, with data types and levels of measurement appropriate for the statistical and machine learning techniques employed. Table 3 presents all the variables used in this study, including derived variables, categorized by their role in the modeling framework, as well as independent versus dependent variables, data type, and level of measurement.

Table 3

Summary of Variables

Variable Name	Category	Type	Level of Measurement	Description
Age	Independent	Numeric	Ratio	Patient age at time of surgery
Sex	Independent	Categorical	Nominal	Male / Female
Weight	Independent	Numeric	Ratio	Body weight in kg
Height	Independent	Numeric	Ratio	Height in cm

Variable Name	Category	Type	Level of Measurement	Description
ASA Class	Independent	Categorical	Ordinal	ASA Physical Status Classification (I–V)
Surgery Type	Independent	Categorical	Nominal	Type of surgical procedure (e.g., cardiovascular, orthopedic)
Anesthesia Type	Independent	Categorical	Nominal	General, spinal, regional, etc.
Comorbidities	Independent	Categorical	Nominal (multi-label)	Includes diabetes, hypertension, etc.
Preoperative Diagnosis	Independent	Categorical	Nominal	ICD or procedural grouping
Intraoperative HR Variability	Independent (Derived)	Numeric	Interval	Standard deviation of HR during surgery
SpO2 Drop Episodes	Independent (Derived)	Numeric	Ratio	Number of episodes where SpO2 < 90%
Hypotensive Events	Independent (Derived)	Numeric	Ratio	Number of times systolic BP < 90 mmHg during surgery
$EtCO_2$ Variability	Independent (Derived)	Numeric	Interval	Standard deviation or range of $EtCO_2$ values
EEG Burst Suppression Ratio	Independent (Derived)	Numeric	Ratio	Proportion of time the EEG shows burst suppression
Sedation Duration	Independent	Numeric	Ratio	Time under anesthesia
Total Fluids Administered	Independent	Numeric	Ratio	Total IV fluid volume during surgery
Oxygenation Patterns	Independent (Derived)	Categorical	Ordinal	Derived from SpO_2 trends (normal, mildly low, severely low)
Stress Index	Independent (Derived)	Numeric	Ratio	Composite score from HR, BP, and EEG patterns
Historical Adherence Flag	Independent (Derived)	Categorical	Nominal	From prior EHR—flag for past non-adherence
Time-Series Lab Results	Independent	Numeric	Ratio	Multiple time points for labs such as lactate and glucose, etc.
Preoperative Medication Use	Independent	Categorical	Nominal	Binary indicator for medications relevant to compliance (beta-blockers)

Variable Name	Category	Type	Level of Measurement	Description
Device Metadata	Independent	Categorical	Nominal	Type and model of monitoring devices used
Postoperative Length of Stay	Dependent (Derived)	Numeric	Ratio	Days from surgery to discharge
Unscheduled Readmission	Dependent (Derived)	Categorical	Nominal	Binary indicator: readmitted within 30 days (Yes/No)
Treatment Adherence	Dependent (Derived)	Categorical	Nominal	Adherence to prescribed medications post-surgery (derived from EHR and timestamps)
Healthcare Engagement	Dependent (Derived)	Categorical	Ordinal	Score or level based on follow-up visits, portal usage, and contact frequency
Response to Intervention	Dependent (Derived)	Categorical	Nominal	Binary/multiclass: improvement/no improvement in vitals or recovery indicators
Adverse Event Flag	Dependent (Derived)	Categorical	Nominal	The occurrence of events like infection, reoperation, or ICU transfer post-surgery

Study Procedures

This study's data acquisition and analysis procedure began with confirming the accessibility, relevance, and integrity of the VitalDB dataset, an open-access repository designed for clinical and machine learning research. The dataset, hosted on the VitalDB public cloud platform, contained high-resolution intraoperative vital signs and time-series clinical data across 6,388 surgical cases. Python and the VitalDB library were used to programmatically extract and preprocess the necessary parameters related to patient monitoring, stress indicators, and perioperative clinical events. Data extraction, exploratory data analysis (EDA), and preprocessing were conducted before model development. For model development, the dataset was split into training and testing sets.

Automated machine learning (AutoML) tools such as H2O AutoML were employed to build and evaluate predictive models for patient behavior, including treatment adherence and

healthcare engagement. In parallel, traditional machine learning algorithms, such as random forests, logistic regression, and gradient boosting, were implemented using Scikit-learn and XGBoost for comparative purposes. Model performance was assessed using accuracy, AUC-ROC, precision, and recall metrics. Sensitivity analyses and cross-validation techniques were applied to ensure the robustness and generalizability of the models.

Data Preprocessing

Data preprocessing for this study followed a structured and iterative approach to ensure the VitalDB dataset was clean, reliable, and well-prepared for machine learning. The process began with the initial extraction of data and exploratory data analysis (EDA) to assess the data structure, distribution patterns, missing values, and outliers. This foundational step guided key decisions for data cleaning and transformation. Core preprocessing steps included handling missing data through time-aware imputation strategies, detecting and managing outliers using statistical methods and clinical thresholds, and applying time alignment to synchronize physiological signals and treatment events from multiple sources.

Signal smoothing techniques were employed to reduce noise in high-frequency time-series data. At the same time, normalization or standardization adjusted for scale differences among variables such as heart rate, blood pressure, and contextual stress indicators. Categorical variables, such as sex and surgery type, were encoded using one-hot or label encoding, depending on the model requirements. Time-series–based features, including rolling averages, variance, and trend indicators, were engineered to capture the dynamic nature of patient behavior over time. Clinically informed composite features, such as a stress index or oxygenation classification, were constructed using domain-specific thresholds to provide contextual relevance.

After this initial preprocessing, the data underwent an iterative refinement stage called reprocessing. This was not a separate formal phase, but a feedback-driven process guided by model performance, data diagnostics, and expert review. Reprocessing involved revisiting imputation choices, fine-tuning normalization procedures, reducing noise, and refining derived variables to better capture complex physiological phenomena.

Initial composite variables were refined into categorized or smoothed versions that more accurately reflect patient behavior. The final dataset was then split into training and testing sets to enable robust model development and evaluation. This comprehensive preprocessing and reprocessing pipeline ensured that the dataset was statistically sound and clinically meaningful, optimizing it for predictive modeling and comparative analysis between AutoML and traditional ML approaches.

Data Analysis

This study adopted a structured, quantitative explanatory quasi-experimental design to evaluate the predictive performance of traditional machine learning and automated machine learning (AutoML) techniques in forecasting patient behaviors, specifically treatment adherence, healthcare engagement, and response to interventions, using the VitalDB dataset. Data cleaning, transformation, and encoding were conducted primarily using Python, with R applied as needed for advanced statistical analysis and visualization. Python libraries such as Pandas and NumPy support preprocessing and data manipulation (McKinney, 2010), while model development utilized scikit-learn (Pedregosa et al., 2011), XGBoost (Chen & Gestring, 2016), and TensorFlow/Keras (Abadi et al., 2016). AutoML workflows were implemented using H2O.ai AutoML, facilitating automated model training, hyperparameter optimization, and benchmarking

(LeDell & Poirier, 2020). Visualization and reporting could be enabled through Streamlit for interactive web apps and Power BI for dashboard-based dissemination.

Traditional ML models, including Logistic Regression, Random Forests, Support Vector Machines, and LSTM neural networks, will be baselines against which the H2O AutoML pipelines were compared. To enhance model transparency, explainability techniques such as SHAP (Lundberg & Lee, 2017) and LIME (Ribeiro et al., 2016) were employed to interpret feature contributions, particularly for complex ensemble and neural network models. Descriptive statistics summarized key variables, including demographic profiles, intraoperative physiological signals, and behavioral outcomes. For hypothesis testing:

- $H1_0$ - No significant difference in predictive performance between AutoML and traditional ML models was evaluated using paired t-tests or Wilcoxon signed-rank tests, depending on distributional assumptions, comparing AUC, accuracy, precision, recall, and F1-score metrics.
- $H2_0$ - contextual variables significantly enhanced predictive performance were tested using hierarchical regression models and feature attribution methods such as SHAP values and permutation importance rankings.

Given the non-randomized nature of the VitalDB dataset, this study employed a quasi-experimental design to evaluate the effectiveness of AutoML-based predictive models in comparison to traditional ML approaches for predicting patient behaviors. To address potential confounding variables, such as patient demographics, comorbidities, surgical types, and environmental stressors, this study employed statistical control methods to minimize bias and enhance internal validity.

Robustness and generalizability were assessed using k-fold cross-validation and holdout validation. Sensitivity analyses explored subgroup differences, such as those between high-risk and low-risk patients. Fairness and bias mitigation were explicitly evaluated by disaggregating performance metrics by subgroup, such as gender, age brackets, and risk strata. Disparities in false positive or false negative rates were investigated using interpretability tools to identify and mitigate reliance on biased signals. These steps were crucial to ensuring that predictive models enhanced, rather than hindered, equitable clinical decision-making.

Although socioeconomic status (SES) was not included as a core predictive feature due to limitations in the dataset, the analysis acknowledged its conceptual relevance. SES could have interacted with intraoperative predictors and behavioral outcomes; its absence was noted as a limitation, with recommendations for future research that integrated SES within more comprehensive data environments. This analysis plan was designed to rigorously test whether AutoML could offer interpretable, fair, and high-performing predictions of patient behavior, with direct implications for embedding such tools in clinical decision support systems.

Validation of Proxy and Synthetic Variables

Several variables in this study, including the Stress Index, intraoperative HR variability, burst suppression ratio, and treatment adherence, were constructed from raw intraoperative signals or derived from historical electronic health record (EHR) data. These proxy measures were defined based on established clinical guidelines and peer-reviewed literature to ensure conceptual and empirical validity. Feature engineering processes included statistical benchmarking against known physiological norms, such as standard ranges for HR variability, and cross-validation with other modalities, like aligning EEG and BP trends to verify internal consistency.

To further validate these variables, a subset of cases underwent expert clinical review to assess face validity for complex constructs, such as adherence patterns and adverse event indicators. Model sensitivity to alternative variable definitions was systematically tested to confirm the robustness of predictive outcomes. While this study did not include socioeconomic status (SES) as a core predictive variable, its conceptual relevance was acknowledged, particularly in how SES could have interacted with intraoperative risk factors and postoperative adherence.

However, SES was excluded from the current scope due to limitations in the available dataset and the study's focus on intraoperative signals. Future research could explore its mediating or moderating effects within expanded modeling frameworks. Explainability techniques, such as SHAP and LIME, were used to interrogate the importance of features and confirm that the models relied on clinically interpretable relationships. These tools were applied judiciously to avoid redundancy and ensure that insights derived from the models enhance transparency and clinical relevance.

Interpretability Assessment

As part of the model evaluation process, interpretability was assessed using post hoc explainable AI (XAI) techniques to make the AutoML-generated predictions transparent and clinically meaningful. Specifically, SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) were applied to the final models to quantify and visualize how key intraoperative variables, such as the Stress Index, ASA classification, heart rate (HR), and oxygen saturation (SpO_2), contributed to individual and global model predictions. These tools generated local explanations for individual cases and global summaries across the dataset. Interpretability was evaluated based on the clarity of explanations, consistency of feature

effects across cases, and the ability of clinicians to understand and validate model outputs during a pilot review process.

Feature Importance Evaluation

Feature importance was determined by using intrinsic AutoML outputs and model-agnostic explanation tools to identify which input variables most strongly influenced the prediction of postoperative treatment adherence. AutoML platforms typically rank features based on aggregated contributions across multiple model iterations through tree-based ensembles or meta-models. In addition to these built-in metrics, SHAP summary plots provided a standardized, quantitative ranking of feature contributions across the entire dataset. A variable was considered highly important if it consistently ranked among the top predictors across cross-validation folds and made meaningful contributions to performance gains when included in the model. This evaluation informed model refinement and ensured clinical relevance by highlighting actionable physiological and contextual predictors that are relevant to clinical practice.

Assumptions

This study employed several key statistical and methodological assumptions to ensure the validity and interpretability of its results. Each patient case from the VitalDB dataset is treated as independent, and multicollinearity among predictors was minimized using correlation matrices and Variance Inflation Factors (VIF). The normality of residuals was checked using Q-Q plots and the Shapiro-Wilk test. If normality is violated, non-parametric models, such as XGBoost or LSTM, were employed.

Homoscedasticity was tested using Breusch-Pagan or White's tests, and robust methods were applied if needed. Missing data were assumed to be Missing at Random (MAR) and

handled with Multiple Imputation by Chained Equations (MICE). Stratified random sampling based on surgical type, ASA score, and age was used to ensure generalizability. Temporal alignment was maintained across physiological and contextual data using synchronized timestamps.

AutoML tools assume properly preprocessed and labeled data, while models like LSTM depend on a sequential structure. Explainability tools such as SHAP and LIME assume local linearity for accurate interpretation. Addressing these assumptions ensured the study's analytical rigor and validity.

Limitations

This study acknowledged several limitations inherent in the VitalDB dataset and the predictive modeling approach and outlined specific mitigation strategies to address them. Although the VitalDB dataset was rich and high-resolution, it was limited to surgical patients from a single institution in South Korea, which may not have represented broader or more diverse patient populations. To mitigate this, the study employed stratified sampling across surgical types, age groups, and ASA scores to enhance internal variability and simulate broader conditions.

Another limitation was the potential for residual confounding due to unmeasured variables, such as psychological stress or a history of medication adherence, which were not captured in the dataset. Propensity score matching (PSM) was used to balance the distribution of key covariates across comparison groups, specifically between AutoML and traditional ML models. Patients were matched based on their likelihood of receiving a particular predictive outcome, conditional on observed covariates. This technique helped simulate the conditions of randomization and reduced selection bias.

Multivariable regression models were used to adjust for residual confounding by including covariates such as age, gender, ASA physical status, anesthesia duration, and intraoperative physiological parameters as predictors. Sensitivity analyses using inverse probability of treatment weighting (IPTW) and stratification by propensity score quintiles further validated the robustness of the findings. These approaches ensured that differences in predictive accuracy or patient behavior outcomes were more attributable to the modeling technique rather than underlying confounding factors.

Socioeconomic status (SES) variables were recognized as important factors in healthcare outcomes and patient behavior; however, their inclusion was not deemed strictly necessary for addressing the current research questions posed in this study. This study evaluated the predictive power of intraoperative and perioperative clinical variables, including vital signs, ASA classification, and physiological stress indicators, within an AutoML-based system, in comparison to traditional machine learning approaches.

This study examined how clinical and contextual physiological data, when used alone, could enhance predictive performance and model accuracy, particularly in forecasting behaviors such as treatment adherence and follow-up engagement. As such, the primary emphasis was on clinical and time-series data, and the role of SES, while potentially influential in real-world applications, fell outside the immediate scope of these objectives.

The framework developed in this study could be extended in future research to incorporate SES variables to explore disparities in healthcare engagement further and personalize interventions holistically. While proxy variables like ASA score, type of surgery, and vital trends were used, this limitation was acknowledged, and future work may incorporate more granular behavioral or longitudinal data.

In addition, the use of AutoML platforms, although efficient, can obscure model mechanics, making interpretability more challenging. To address this, explainability tools like SHAP and LIME were used to understand the importance of features and model behavior. Table 4 below summarizes the limitations and the mitigation strategy planned within the study.

Table 4

Limitations and Mitigation Strategy

Limitations	Description	Mitigation Strategy
Selection Bias	VitalDB is based on data from a single hospital and surgical population.	Use stratified sampling and perform subgroup analysis across demographics and procedures to assess generalizability.
Missing or Incomplete Data	Certain intraoperative variables may be inconsistently recorded.	Apply advanced imputation techniques like MICE and KNN, and report the percentage of imputed fields.
Labeling Bias in Behavioral Outcomes	Definitions of adherence and follow-up may vary or be indirectly observed.	Use standardized operational definitions, informed by clinical literature and expert validation.
Algorithmic Bias	ML models may overfit dominant patterns or underrepresent subgroups.	Use explainable AI method, SHAP, to audit feature influence and bias; apply fairness-aware metrics.
Limited External Validity	Results may not generalize to non-surgical or outpatient settings.	Highlight scope limitations and recommend future validation on diverse datasets from other institutions.
Temporal Bias	Time-dependent changes may affect data consistency.	Perform time-series trend analysis and temporal cross-validation to detect and adjust for drift.

Limitations of Proxy and Synthetic Variables

Despite their analytical value, proxy and synthetic variables introduced interpretive limitations, as they did not perfectly reflect the underlying clinical constructs they aimed to represent. These variables were inherently subject to modeling assumptions, such as threshold

selection or temporal aggregation, which could have affected reproducibility and generalizability. To mitigate these concerns, the study included sensitivity analyses to quantify the impact of proxy definitions on model outcomes. Predictive findings based on these variables, particularly regarding treatment adherence and patient engagement, were interpreted appropriately. Furthermore, while synthetic data were used during early development phases to test modeling pipelines, all final analyses relied exclusively on real VitalDB-derived features. Any residual uncertainty introduced by proxy measures was explicitly acknowledged when interpreting results and discussing findings.

Delimitations

This study was shaped by several intentional delimitations that helped maintain methodological clarity and alignment with the research purpose, theoretical framework, and problem statement. The exclusive use of the VitalDB dataset, which comprised perioperative data from a single tertiary hospital, enabled the collection of high-resolution, standardized physiological and contextual information necessary for modeling behavioral outcomes, such as treatment adherence and healthcare engagement. This limited the generalizability to broader, non-surgical or outpatient populations.

The study deliberately focused on adult surgical patients, excluding pediatric cases to avoid additional complexity in behavioral interpretation and align with standard perioperative research practices (Lee et al., 2020). Behavioral outcomes in this study were inferred using proxy indicators derived from structured intraoperative and perioperative data, such as surgical duration, hemodynamic stability, and anesthetic responses, rather than direct measures, including medication adherence and follow-up visits. While this approach enabled reproducibility, it

limited the granularity of behavioral insights, as many post-discharge behaviors and subjective experiences were not captured in the VitalDB dataset.

These delimitations were consistent with the conceptual framework, which emphasized practical, data-driven healthcare prediction, and they directly informed the research questions, contrasting the accuracy and transparency of AutoML versus traditional machine learning approaches. Future work may expand this scope by incorporating more diverse datasets, mixed methods, or qualitative variables to provide a holistic view of patient behavior.

Ethical Assurances

The study complied with the ethical standards set forth by the Institutional Review Board (IRB) at the National University, ensuring that all aspects of the research adhered to recognized ethical principles. The dataset employed in this study, the VitalDB dataset, was obtained from a publicly accessible and anonymized repository, explicitly intended for academic and machine learning research. This strategic choice mitigated common ethical challenges regarding informed consent, participant privacy, and data confidentiality.

The researcher assumed the role of an independent data scientist, with no direct interaction with patients or healthcare personnel, thus further reducing the risk of bias or undue influence. Ethical rigor was maintained throughout the study by following well-established research protocols, including data handling, analysis transparency, and impartial interpretation of findings. Sensitivity analyses were also performed to assess the robustness of results under varying model assumptions, ensuring the credibility and ethical soundness of the research conclusions.

Summary

The primary objective of this quantitative, explanatory, quasi-experimental study was to develop and evaluate an AutoML-based big data analytics system using the publicly available VitalDB dataset to predict critical patient behaviors, including treatment adherence, healthcare engagement, and response to interventions. The study employed a comparative design to evaluate the predictive performance of automated machine learning (AutoML) frameworks, specifically H2O AutoML, against traditional machine learning (ML) models, including logistic regression, random forests, and gradient boosting, which were implemented using Python.

A structured data preparation pipeline was followed, including data extraction via the VitalDB API, preprocessing, and feature engineering, to ensure the integrity and contextual relevance of the data. Exploratory data analysis (EDA) guided the identification of behavioral patterns and informed the selection of key predictive variables. Model performance was evaluated using standard metrics, including accuracy, precision, recall, and AUC-ROC, with additional validation through cross-validation and sensitivity analyses to enhance robustness and reduce bias.

Ethical research practices were maintained by utilizing only anonymized, publicly accessible data and ensuring compliance with Institutional Review Board (IRB) guidelines. This methodological approach provided a rigorous and transparent framework for evaluating the utility of AutoML in healthcare prediction tasks. This study aimed to demonstrate the practical potential of AutoML in real-world clinical settings. By streamlining model development and minimizing the need for extensive technical expertise, AutoML frameworks could empower healthcare organizations, especially those with limited data science capacity, to quickly derive actionable insights from complex patient data. These insights supported more timely and

personalized clinical decision-making, improving patient outcomes. The study's findings offered a foundation for integrating AutoML into future clinical decision support systems, reinforcing its broader significance in advancing precision medicine and scalable, data-driven healthcare solutions.

Chapter 4: Findings

The problem addressed in this study was the difficulty of accurately predicting patient behaviors, such as treatment adherence, healthcare engagement, and response to interventions, using big data and machine learning. Traditional predictive models often fall short in clinical settings because they overlook the complex and multidimensional nature of human behavior. Factors such as socioeconomic status, environmental influences, psychological conditions, and individual health histories are rarely integrated into these models, limiting their accuracy and usefulness (Gowda & Lakshmikantha, 2020). This study utilized the VitalDB dataset, a comprehensive, high-resolution resource containing perioperative data from 6,388 surgical patients. With more than 196 intraoperative monitoring variables, 73 clinical parameters, and 34 time-series laboratory measurements, VitalDB enables the development of rich, context-aware predictive models that can reflect the real-world complexity of patient care (Lee et al., 2022).

Despite the advances in machine learning, including the growing adoption of AutoML, challenges remain in creating accurate, interpretable, and actionable models for healthcare professionals. While machine learning has shown promise in processing vast datasets drawn from diverse sources such as speech, social media, and electronic health records (Alsini et al., 2024), existing models often struggle to deliver insights that can be readily applied in clinical environments. This study aims to develop an AutoML-based analytics system that integrates physiological data from VitalDB with contextual variables, such as socioeconomic indicators and patient history, to enhance the prediction of behavior-related outcomes. The research seeks to bridge the gap between predictive performance and clinical applicability by focusing on scalability and interpretability, supporting more personalized and effective healthcare interventions.

The purpose of this quantitative explanatory quasi-experimental study was to develop and evaluate an AutoML-based big data analytics system for predicting patient behaviors, such as treatment adherence, healthcare engagement, and responses to clinical interventions, by integrating complex contextual factors, including stress levels, environmental influences, and significant life events. Addressing the limitations of traditional predictive models in healthcare, particularly their inability to account for multidimensional contextual data, this study leverages the VitalDB dataset, a high-resolution, open-access perioperative database comprising physiological, clinical, and time-series data from 6,388 surgical patients. The dataset is licensed under the Creative Commons Attribution 4.0 International License https://creativecommons.org/licenses/by/4.0/, allowing redistribution and adaptation for any purpose, including commercial use, provided that proper credit is given and any modifications are indicated. Using Python and the VitalDB library, relevant parameters were programmatically extracted and preprocessed to ensure data integrity and clinical relevance. Exploratory data analysis (EDA), stress indicator mapping, and data cleaning procedures were completed before partitioning the dataset into training and testing sets for model development.

Model construction was performed using both automated and traditional machine learning techniques to enable comparative evaluation. H2O AutoML was employed to automate model selection and tuning for predicting behavior-related outcomes. In contrast, traditional models, such as logistic regression, random forests, and gradient boosting, were developed using Scikit-learn and XGBoost. Model performance was evaluated using metrics including accuracy, AUC-ROC, precision, and recall, with cross-validation and sensitivity analyses applied to ensure robustness and generalizability. Explainability tools, such as SHAP and LIME, were incorporated to interpret model outputs at both individual and population levels, thereby

enhancing clinical applicability. This study aims to develop a scalable, interpretable, and data-driven predictive system that bridges the gap between machine learning innovation and real-world healthcare decision-making.

This chapter is organized into four main sections. The first section presents a data preprocessing and modeling process diagram, which includes a detailed visual representation of the study procedure and outlines all key stages of the completed research. The second section presents the results, organized according to the research questions. The third section evaluates the findings, and the fourth and final section discusses the study's limitations. The research questions in this study were:

RQ1. To what extent does the integration of the intraoperative Stress Index enhance the predictive accuracy of an AutoML-based system compared to traditional machine learning models in forecasting postoperative treatment adherence?

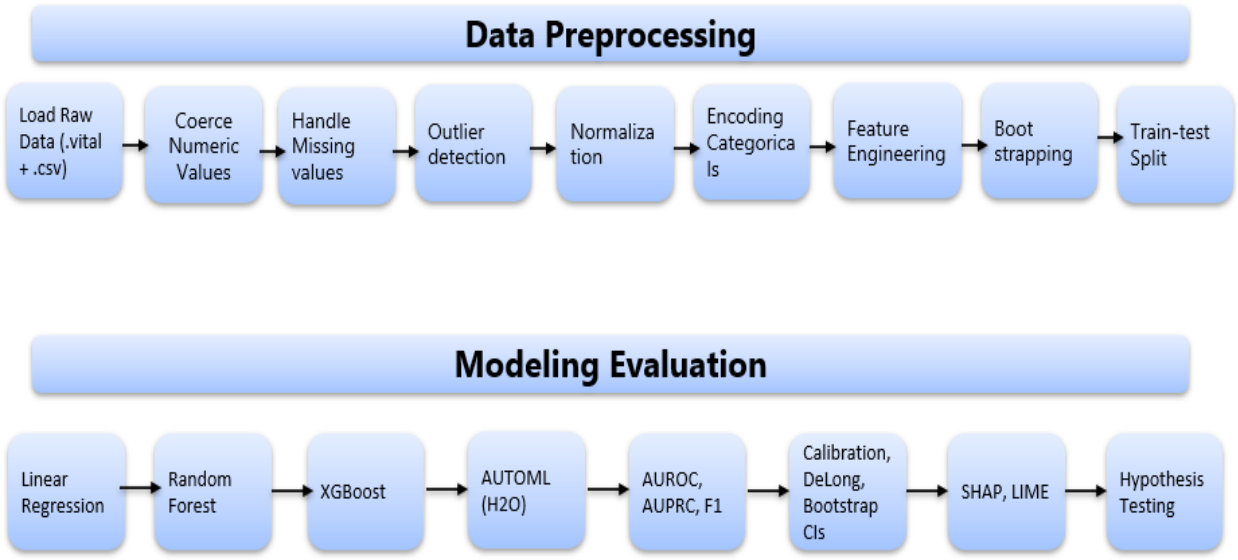
RQ2. How do intraoperative features such as the Stress Index, ASA classification, and time-series trends in HR and SpO_2 contribute to the interpretability and feature importance within AutoML-generated models for predicting treatment adherence?

Data Preprocessing and Modeling Process Diagram

The diagram created for this study outlines the research process and illustrates all the stages involved. Figure 6 shows the data preprocessing and data modeling steps.

Figure 6

Data Preprocessing and Data Modeling Steps



Data Preprocessing and Modeling

The research study was divided into four phases to ensure a structured and rigorous approach. Phase I focused on data preprocessing and feature engineering, which involved cleaning the VitalDB dataset, handling missing values through imputation, standardizing units, aligning timelines, and creating derived features such as the Intraoperative Stress Index (ISI), vital sign variability measures, contextual clinical factors, and interaction terms. Phase II consisted of exploratory data analysis (EDA) to summarize patient demographics, surgical characteristics, and vital sign distributions, assess correlations among clinical and derived features, and examine patterns of class imbalance in treatment adherence outcomes. Phase III involved partitioning the dataset into training, validation, and testing subsets at the patient level, applying temporal and stratified strategies to prevent data leakage and preserve outcome balance. Finally, Phase IV encompassed model development and evaluation, where both traditional machine learning approaches, such as logistic regression, random forest, and XGBoost, and the AutoML framework H2O AutoML were applied.

Model performance was assessed using accuracy, F1-score, ROC-AUC, PR-AUC, calibration plots, and confusion matrices. In contrast, statistical tests, including bootstrap confidence intervals, DeLong tests, and McNemar's tests, were used to compare AutoML and traditional models. Explainability methods, such as SHAP and LIME, were then applied to interpret the contributions of ISI and other key features to predictive performance.

Data Cleaning

The analytical dataset used for this study was derived from 6,388 intraoperative cases obtained from the VitalDB repository, which was then merged with perioperative and electronic health record (EHR)–derived adherence labels. The integrated model-ready table comprised 92 variables, encompassing both numeric and categorical measures, as well as derived time-series measures.

Types of Variables. Numeric variables (ratio and interval scales) included demographic and physiological metrics such as age, height, weight, body mass index (BMI), stress index (ISI), mean and variability of heart rate (HR), and end-tidal carbon dioxide ($EtCO_2$), total fluids administered, and postoperative length of stay (LOS). These continuous measures captured both baseline patient status and intraoperative physiological responses. Categorical variables, nominal and ordinal, comprised sex, American Society of Anesthesiologists (ASA) physical status classification (I–V), surgery type, anesthesia type, oxygenation pattern, perioperative medication flags, device metadata, historical adherence status, and postoperative engagement outcomes. Ordinal variables, such as ASA class, were preserved in their natural order to retain interpretability.

Derived indicators and time-series summaries reflected intraoperative physiologic trends and event-based features, including oxygen desaturation episodes (SpO_2 burden), cumulative

hypotensive events, electroencephalogram (EEG) burst suppression ratio, and the computed stress index (ISI). These dynamic features were extracted within the operative window using synchronized timestamps to ensure clinical validity and temporal alignment across cases.

A comprehensive data audit verified that most demographic and categorical fields exhibited high completeness, while certain waveform-derived features, mean arterial pressure [MAP], BIS, and $EtCO_2$ variability, demonstrated expected structural absence for cases lacking specific monitoring modalities. The completeness of each variable was summarized through quantitative and visual diagnostics, including a histogram of missingness percentages (Figure 7), a Top 30 missingness bar chart (Figure 8), and a hierarchical clustering of features based on pairwise missing-value patterns—NA co-occurrence (Figure 9).

Figure 7

Histogram of Variable-Level Missingness Across all Features

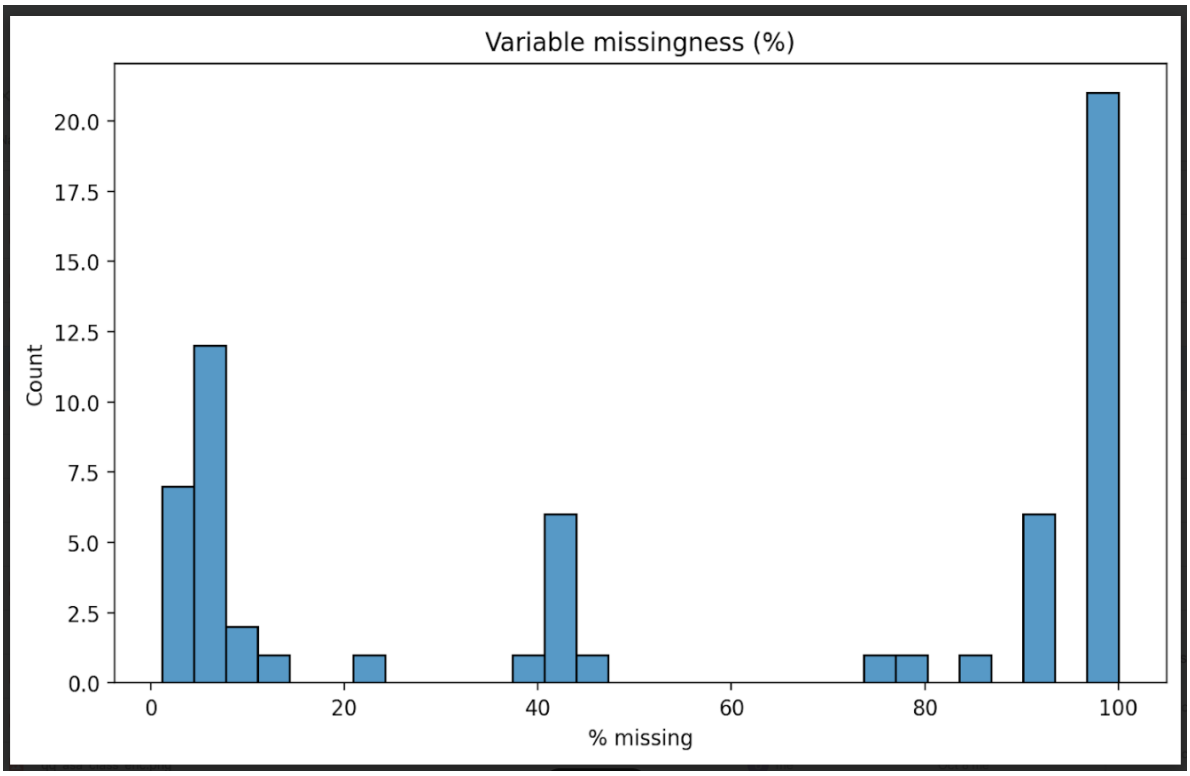


Figure 8

Top 30 Features Ranked by Percentage of Missing Data

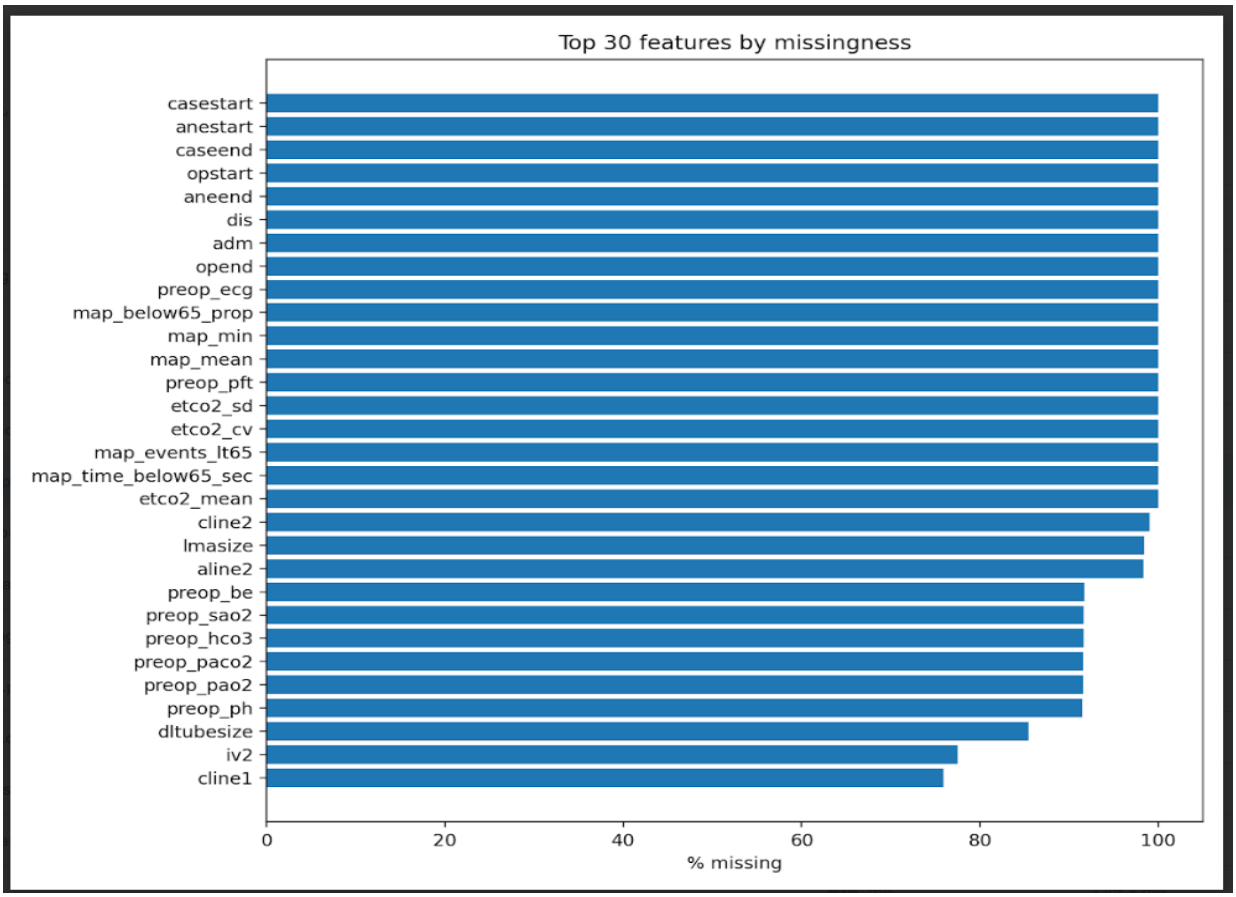
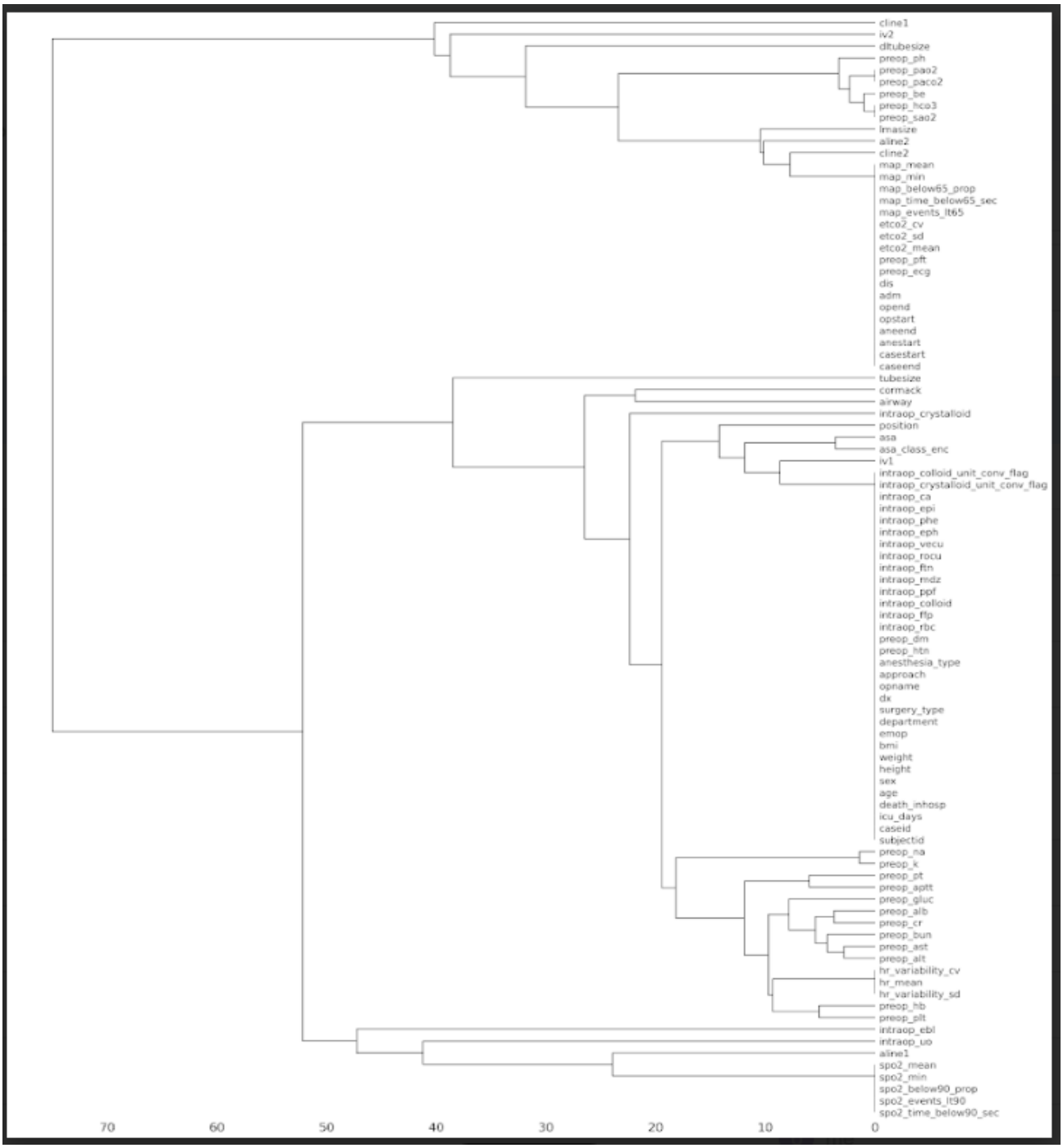


Figure 9

Missingness Dendrogram



Distributional diagnostics were performed for all key numeric features to confirm data quality prior to modeling. Histograms and quantile–quantile (QQ) plots shown in Figure 10 illustrate the distributional properties of age, BMI, ICU days, mean HR, and mean SpO_2. Subgroup analyses using box and violin plots presented in Figure 11 demonstrate consistent and

clinically plausible variability across ASA classes and surgery types. These visual summaries collectively confirm the completeness, validity, and appropriateness of the variables for subsequent modeling phases.

Figure 10

Histograms and QQ Plots of Age, BMI, ICU Days, HR Mean, and SpO_2 Mean

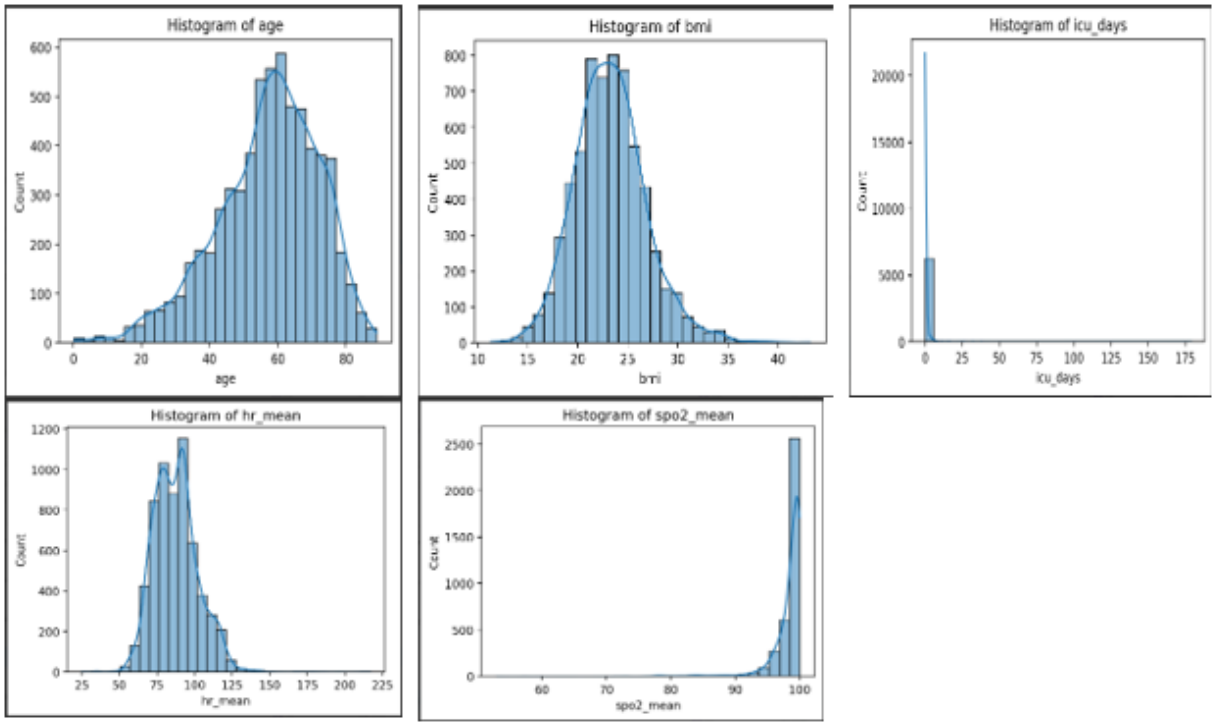
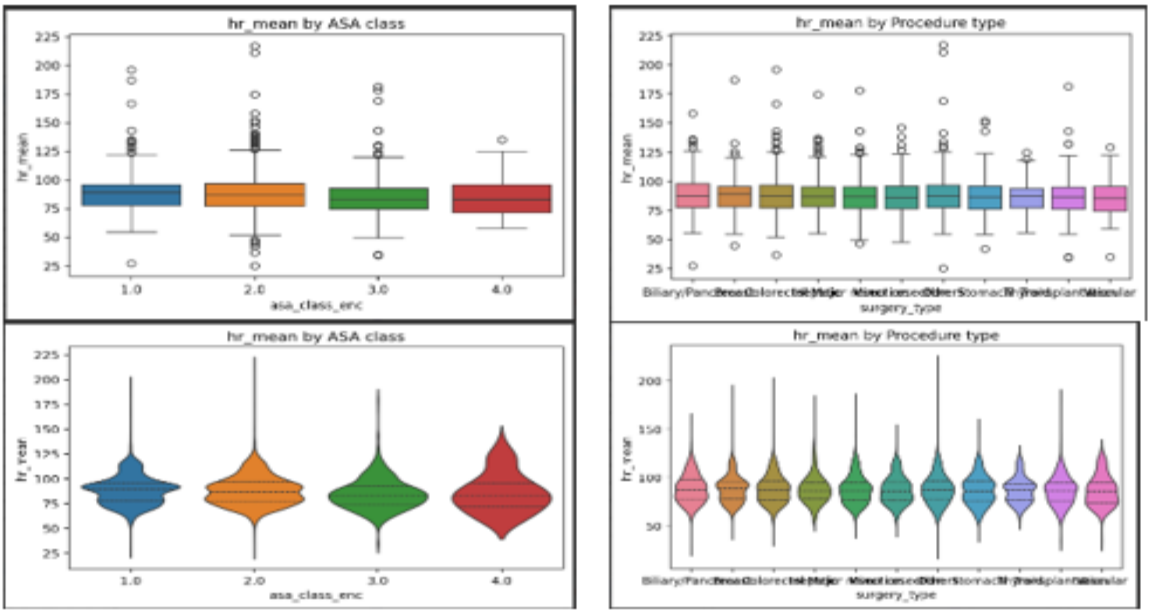


Figure 11

Box and Violin Plots by ASA Class and Surgery Type



Handling Date/Time Variables. The handling of temporal variables was a critical preprocessing step in ensuring that intraoperative data were accurately aligned with clinically valid surgical windows. The process standardized and validated all timestamp variables, such as anesthesia start and end, surgery start and end, and case start and end, so that time-dependent indicators, including heart rate variability, stress index, and oxygen desaturation events, were correctly bounded within the operative period. All time-related columns were parsed and standardized using Python's *datetime* library. Timestamps that appeared as either strings or numerical offsets were converted to a unified format with consistent time zones and precision. Any missing or malformed values were flagged and excluded from analysis.

Timing offsets between the anesthesia and surgical phases were also computed and summarized in offsets_summary.csv to identify discrepancies in pre-induction or post-emergence timing. To evaluate data quality, a suite of visual diagnostics was generated. Figure 12 displays the distribution of operative window durations, showing that most cases fell within the range of 10 to 40 hours, with distinct modal peaks. Figure 13 illustrates the distribution of surgery

durations in offset minutes, revealing a right-skewed distribution where most surgeries lasted under 200 minutes. Figure 14 presents sample case timelines illustrating anesthesia (light bands) and surgery (darker bands) intervals relative to the start of anesthesia (0 minutes). Negative offsets represent pre-induction or documentation timing discrepancies, while positive values correspond to intra- and post-operative phases.

These graphical summaries collectively verify that the computed windows corresponded to clinically realistic intervals and that no pre- or post-operative artifacts remained. Summary statistics were compiled into temporal_qc_summary.csv, recording the total number of valid cases, median operative duration, and rate of outlier exclusions. This quality-control summary ensured that downstream modeling incorporated only temporally verified data, maintaining both methodological transparency and analytic reliability.

Figure 12

Distribution of Operative Window Duration

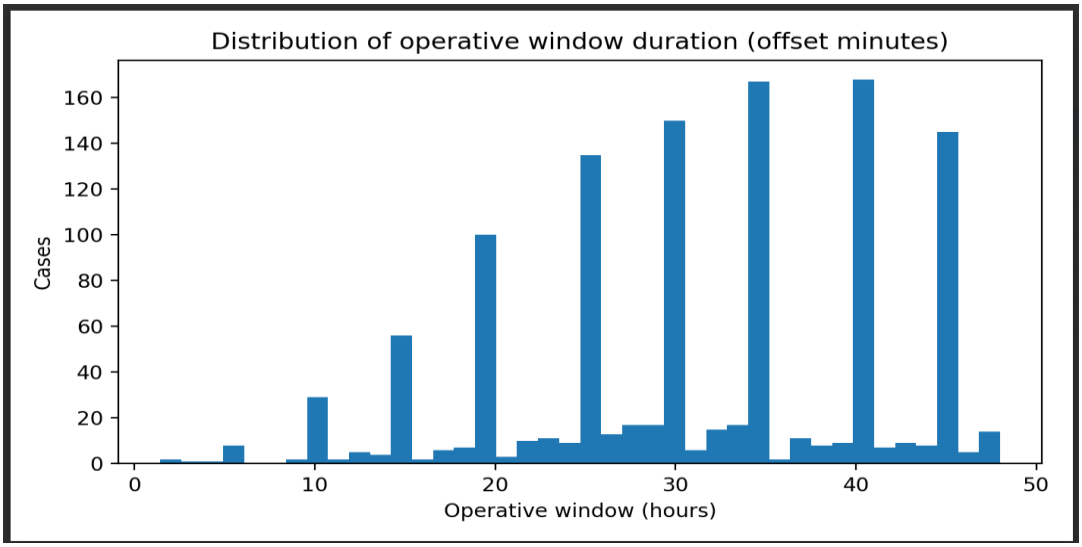


Figure 13

Distribution of Surgery Duration

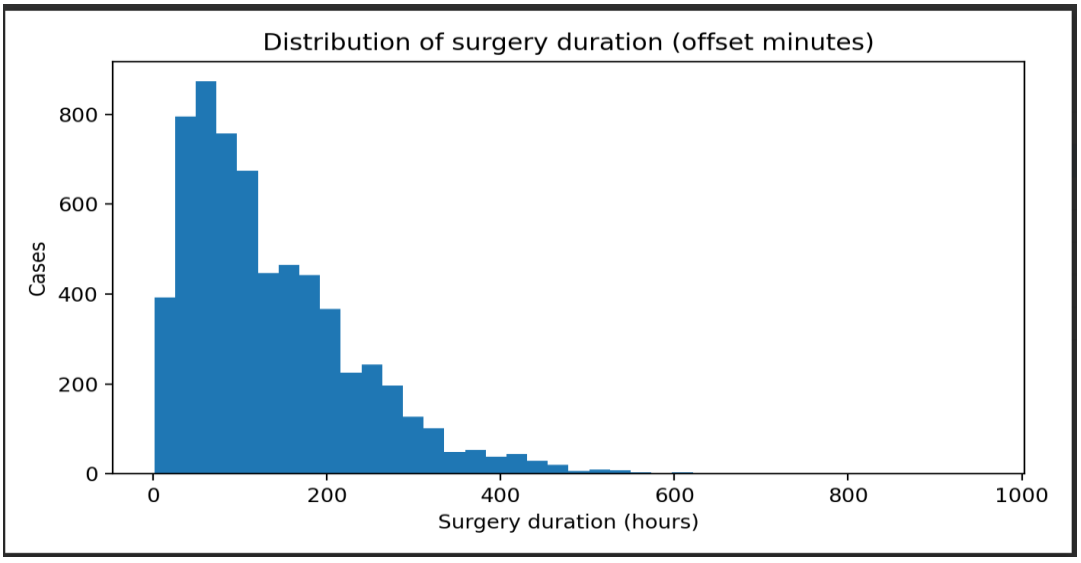
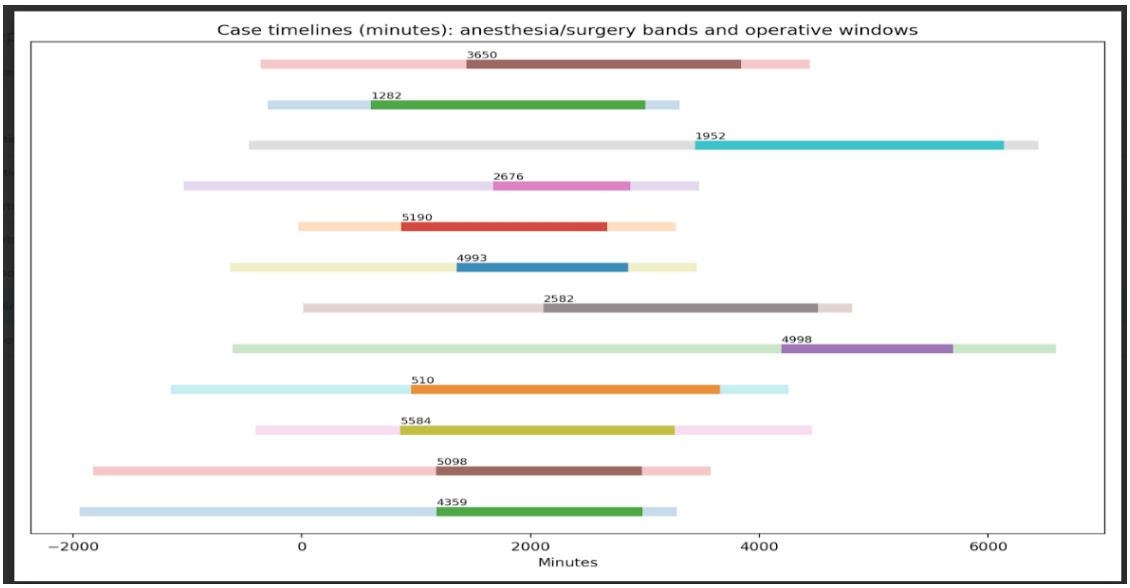


Figure 14

Case Timelines with Anesthesia and Surgery Intervals



Missing Values. Missing values were systematically quantified, recorded, and resolved using a combination of diagnostic scripts and targeted imputation procedures. The missingness

audit was performed using the automated Python script variable_audit.py, which generated a comprehensive summary file, missingness_report.csv, containing variable names, non-missing counts, missing counts, and the percentage of missing values for each variable. Corresponding visual diagnostics, including the missingness histogram, matrix, and heatmap, were stored in the /clean/reports directory for documentation and reproducibility.

The variables exhibiting the highest proportions of missingness were those derived from specialized intraoperative monitoring that was not universally applied across all surgical procedures. These included mean arterial pressure (MAP) and EEG burst suppression ratio (missing in cases without invasive arterial lines or EEG leads), $EtCO_2$ variability (missing in procedures lacking end-tidal measurements), and partial missingness in SpO_2 variability, due to intermittent gaps caused by signal interruptions. The stress index (ISI) was missing only when one or more input parameters (MAP, HR, or SpO_2) were unavailable. Postoperative LOS was absent for same-day discharges, and specific medication or device indicators were coded as "Unknown" when unavailable in the perioperative record.

To address these gaps, variable-specific imputation strategies were applied through the scripts create_imputed_knn.py and vitaldb_preprocess_final.py. Numeric variables were imputed based on their empirical distributions. Median substitution was used as the default, mean imputation was applied to symmetric distributions, and K-nearest neighbors (KNN) imputation was used for correlated physiologic features, including MAP, HR, and SpO_2. Categorical variables were imputed using the mode, with an additional "Unknown" label for fields where missingness had clinical significance, such as anesthesia type not recorded. Time-series variables such as SpO_2 desaturation and HR variability were interpolated or forward-filled only when medically appropriate to preserve physiologic realism.

Each imputation method and its corresponding variable were automatically recorded in the configuration file imputation_config.json, which documented the variable name, the chosen imputation method, and parameters such as the number of neighbors for KNN, as well as whether the variable was structurally missing or imputed. Quality control was subsequently performed using imputation_diagnostic_plan.py, verifying that no residual missing values remained in the final dataset and that post-imputation distributions were statistically consistent with the original non-missing subsets.

The complete record of missing-value handling, including diagnostic plots, configuration files, and logs, was archived within the /clean/reports and /models directories. This ensures that the process is fully replicable and auditable. In summary, missingness was systematically recorded, variable-specific handling was documented, and quality checks confirmed that the final model inputs were analytically complete, clinically valid, and reproducible.

Unknown/Invalid Tokens. A systematic diagnostic scan was conducted across all preprocessed datasets to identify placeholder values, inconsistent category encodings, or text artifacts that could bias data integrity. The audit searched for standard irregular tokens, including "," "NA," "N/A," "null," "Unknown," and similar variants, after trimming whitespace and converting all entries to a consistent string format. Each column of every data table was scanned in chunks to ensure complete coverage of large files. No invalid tokens were detected in any dataset, confirming that the cleaning and normalization pipeline effectively standardized categorical and numeric entries before modeling.

To further verify internal consistency, post-cleaning correlation heatmaps were generated for both the clinical_data and lab_data tables. Figure 15 illustrates the correlation matrix of clinical variables, revealing coherent inter-variable relationships. Higher correlations were

observed among related preoperative laboratory measures, including preoperative creatinine and blood urea nitrogen, body composition metrics, height, and BMI, as well as intraoperative hemodynamic summaries. The absence of aberrant correlation clusters confirmed that token normalization and missing-value handling did not introduce spurious dependencies among numeric predictors.

Figure 15

Correlation Heat Map – Clinical Data

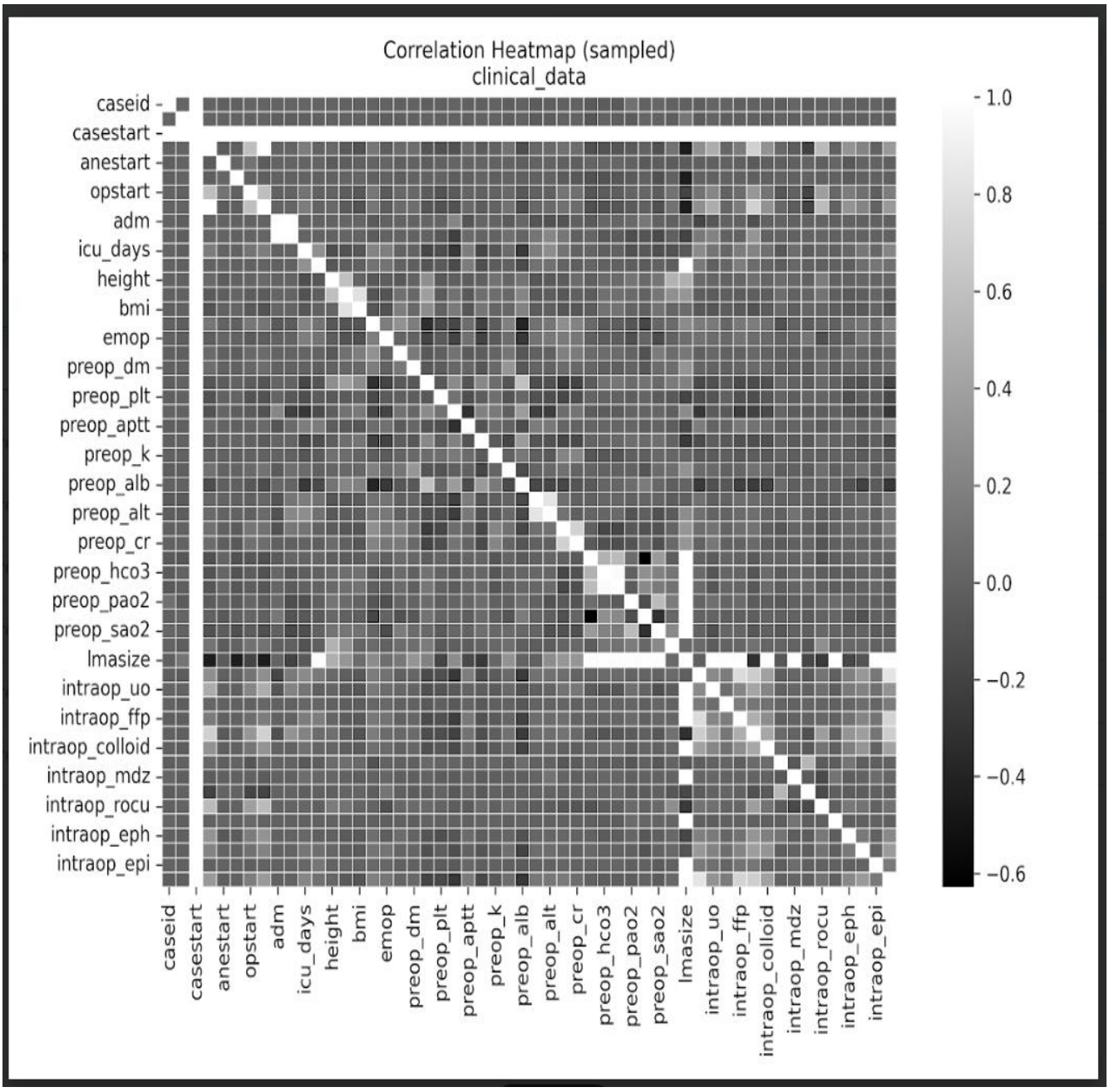
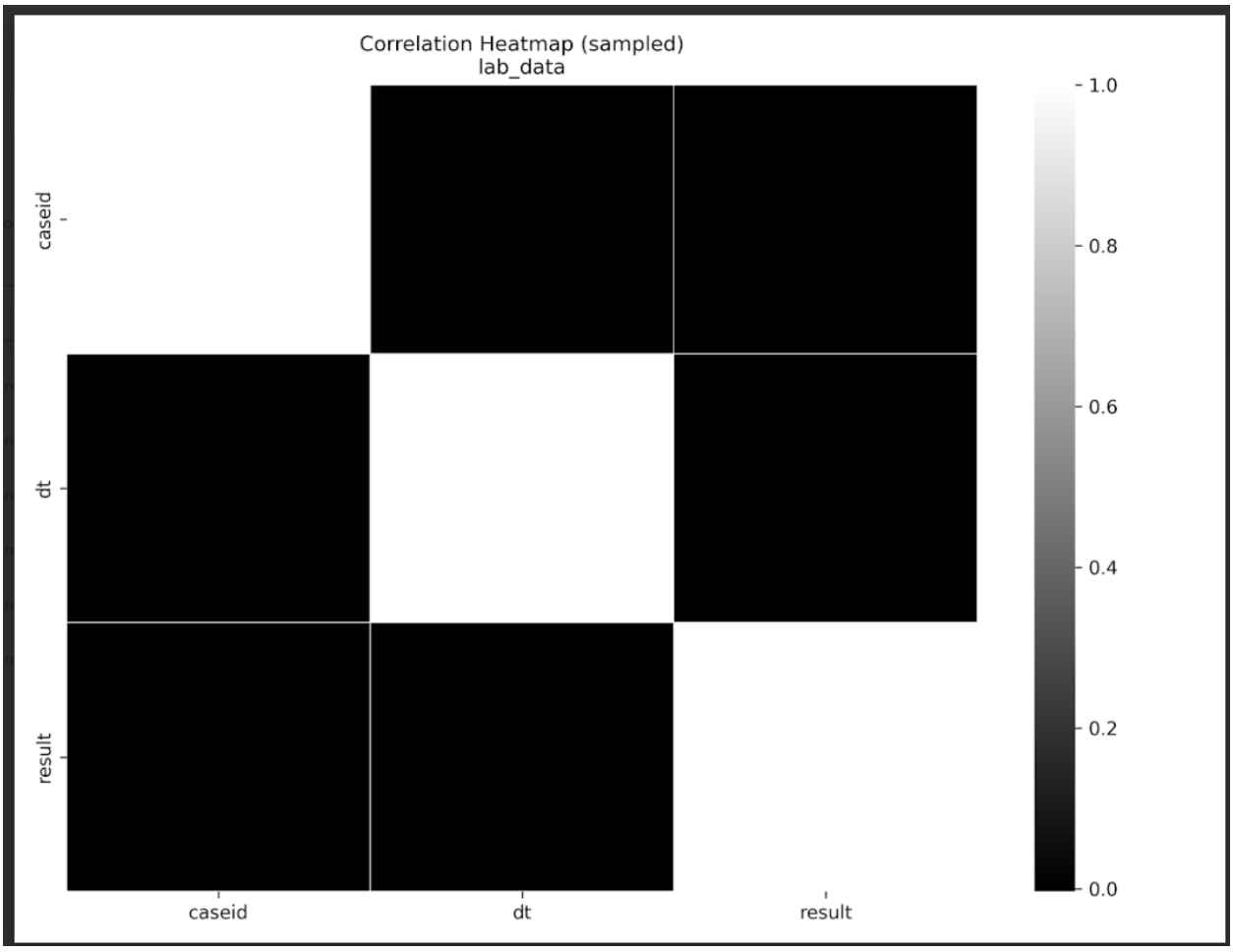


Figure 16 displays the correlation structure for lab_data, which primarily contains temporal identifiers and laboratory results. As expected, minimal correlation was observed among these fields, confirming the appropriate separation of data and the absence of redundant identifiers. Categorical standardization was also validated using the American Society of

Anesthesiologists (ASA) classification as a reference variable. This ensured ordinal integrity and preserved the clinical meaning of ASA strata across all cases.

Figure 16

Correlation Heat Map – Lab Data



Cleaning Steps. The data cleaning pipeline followed a structured multi-step process to ensure the dataset's consistency, accuracy, and clinical validity. First, categorical variables underwent string trimming and normalization to reduce inconsistencies in labeling. Numeric-like strings were converted into numeric values through unit and type coercion, while categorical predictors were explicitly cast to categorical data types. All date and time fields were parsed and aligned with operative start and end windows to ensure derived intraoperative features were

constrained to clinically valid intervals. Categorical normalization was applied, including standardizing the sex variable to Male/Female/Unknown, and consistently mapping ASA values to the canonical I–V scale.

Outlier screening was performed in two stages: clinical plausibility rules were applied to flag values outside physiologic ranges like heart rate <20 or >220 bpm, mean arterial pressure <30 or >300 mmHg, burst suppression ratio outside [0–1], and stress index outside 0–100, and statistical detection methods such as interquartile range (IQR) thresholds were used to identify extreme deviations. These limits are consistent with VitalDB documentation and the MIMIC-IV preprocessing framework, which recommends excluding implausible values due to sensor artifacts or entry errors. Implausible values were corrected when identifiable, like fluid totals recorded in liters instead of milliliters, or flagged and removed if ambiguous. Missing values were imputed using the median for numeric fields and mode or "Unknown" categories for categorical variables. The pipeline generated audits of variable types, missingness summaries, and outlier flags alongside the cleaned dataset to promote transparency and reproducibility.

Visualization tools were used to examine the structure and relationships of missing data, guiding imputation decisions. Figure 17 illustrates the 15 variables with the highest proportion of missing data, most of which were preoperative laboratory measures and device metadata fields. Figure 18 illustrates the data type composition of the cleaned analytical dataset, where categorical and object-type variables account for over 96% of all features, highlighting the importance of consistent encoding prior to model training.

Figure 17

Top 15 Columns with Most Missing Data

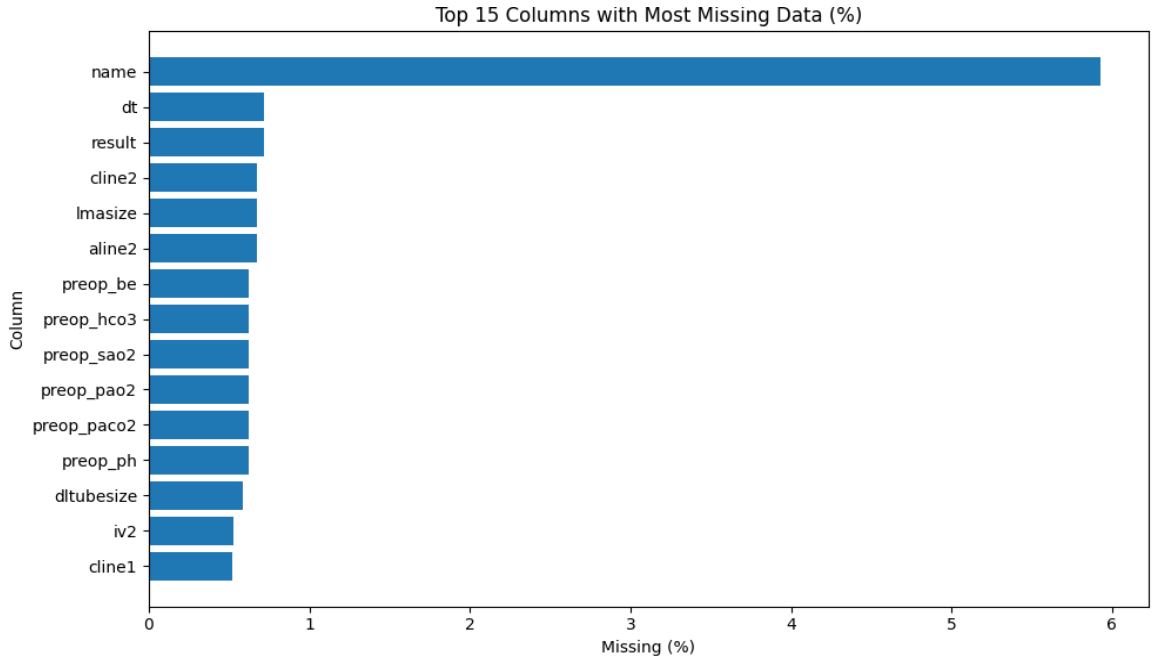
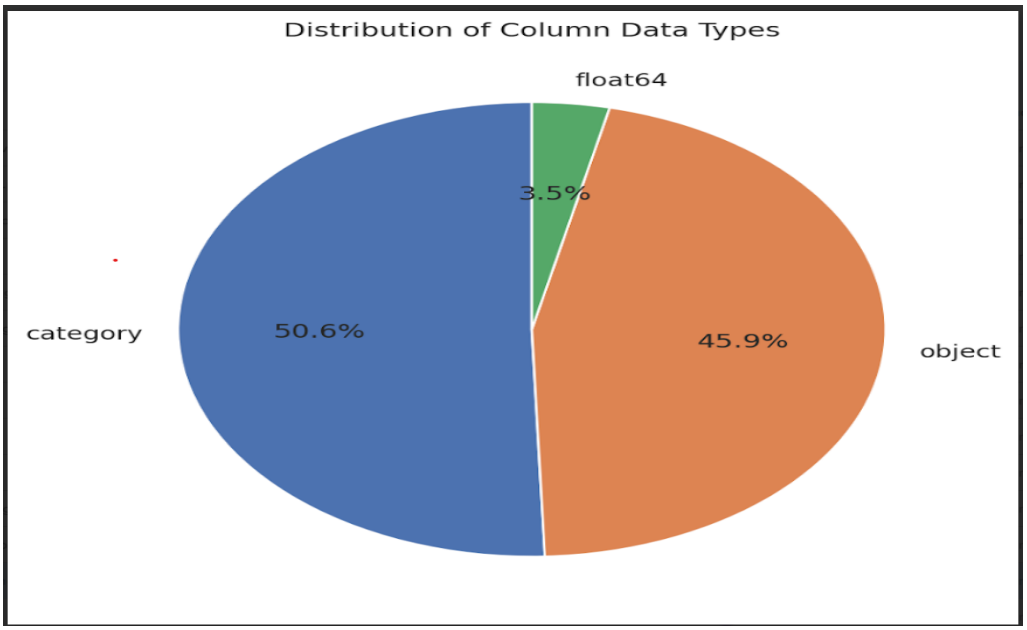


Figure 18

Distribution of Column Data Types

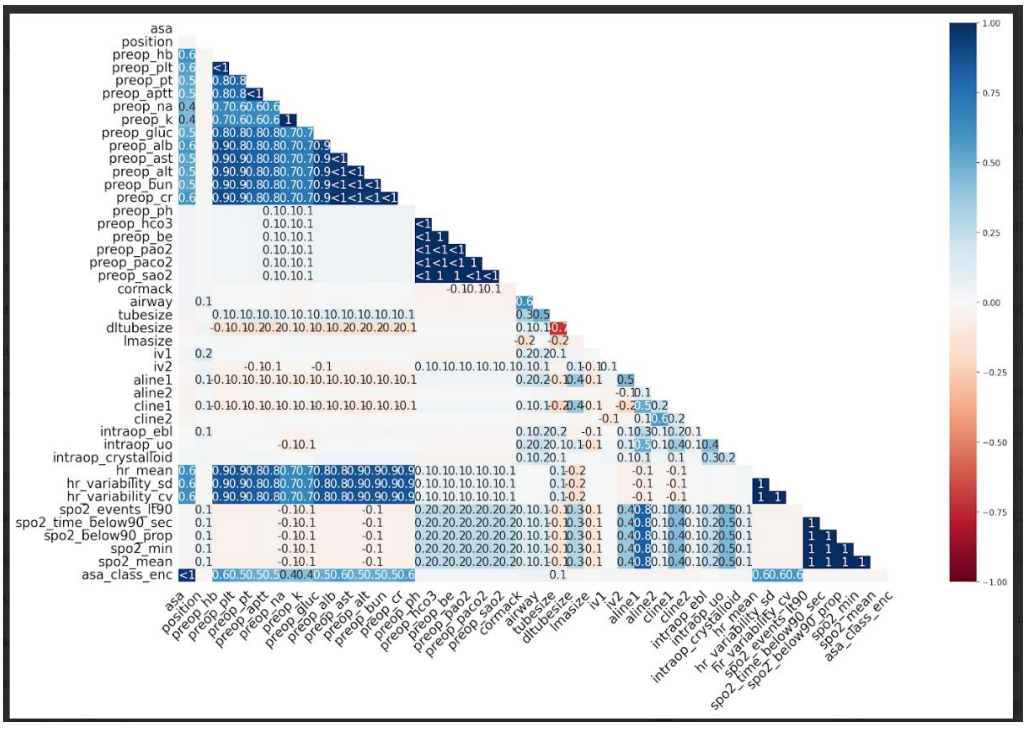


To explore relationships among missing features, a missingness heatmap was generated. Figure 19 displays the pairwise co-occurrence of missingness, where darker regions indicate

variables that frequently share missing values. This relationship structure highlighted that preoperative laboratory tests, *preop_be*, *preop_hco3*, and *preop_pao2* often had simultaneous gaps, consistent with patients for whom arterial blood gas analysis was not performed. The presence of the column labeled “name” in the preliminary chart was a placeholder metadata field retained from the desensitized manifest. No personal identifiers were included in the dataset, and this column was subsequently removed from the final analysis to avoid confusion regarding de-identification.

Figure 19

Missingness Heat Map



Data Type Transformations. Columns were reviewed for numeric and categorical consistency. Fields with more than 60% numeric values were coerced into numeric (float64) dtype. Known categorical predictors were explicitly cast to the category dtype to optimize

memory use and encoding stability. Ordinal variables such as ASA Class were stored as ordered categorical features (I < II < III < IV < V) to preserve their inherent rank. Ordinal variables, including ASA Class, are represented within the categorical segment of Figure 18, as they were encoded as categorical types to preserve their clinical hierarchy. Remaining string-based columns retained the object dtype for later inspection or encoding.

Trimming, Normalization, and Formatting. String variables, including text-based demographic and procedural descriptors such as *sex, department, surgery_type, anesthesia_type, approach, opname, airway, cormack, cline1, cline2, aline1, aline2,* and *tubesize*, were standardized through a structured text-cleaning process. Leading and trailing whitespace were trimmed, multiple internal spaces were collapsed, and inconsistent casing was normalized to title or upper case depending on field semantics, for example, *"female"* → *"Female"*, *"GA"* → *"General Anesthesia"*. Duplicate or synonymous labels were consolidated through canonical mappings to ensure uniformity across data sources. Examples include the unification of *"GA," "general,"* and *"General anaesthesia"* under the standardized category *"General Anesthesia."* Ordinal variables such as *ASA Class* were initially treated as string variables during this normalization stage and subsequently encoded as ordered categorical variables (*I < II < III < IV < V*) to preserve their clinical hierarchy. This two-step approach, string standardization followed by ordered categorical encoding, ensured both lexical consistency and ordinal interpretability across analyses.

Units of measurement were harmonized to a single convention documented in the project's data dictionary: volumes were converted to milliliters (mL), pressures to millimeters of mercury (mmHg), and time intervals to minutes. These conversions eliminated ambiguity from mixed unit reporting; for example, 1.5 L and 1500 mL were both stored as 1500 mL. Following

normalization, numeric fields with a parse success rate greater than 60% were coerced into numeric data types, while known categorical predictors were explicitly cast as categorical variables to improve memory efficiency and ensure consistent encoding. Final outputs were stored in both *.csv* and *.parquet* formats, with all column names converted to standardized *snake case*, *SurgeryType → surgery_type*. A comprehensive *data_dictionary.csv* file recorded each variable's name, data type, unit, and encoding scheme, ensuring reproducibility and full traceability of the preprocessing workflow.

Size and Limitations. The merged VitalDB-derived dataset initially contained approximately 9.3 million rows across 84 variables representing perioperative, intraoperative, and outcome features. After a structured sequence of cleaning, trimming, and imputation steps, the finalized analytic dataset comprised 935,000 case-level observations. The imputation process affected approximately 12–15% of the total records. Detailed descriptions of the imputation strategy, variable-specific handling, and diagnostic plots are presented in the Missing Values section.

Features with more than 50% missingness were excluded from analysis to maintain statistical reliability. While this approach reduced the dimensionality of the dataset, it preserved the interpretability and integrity of the remaining predictors. Limitations include the potential for imputation to attenuate variance and for outlier filtering to exclude physiologically rare but valid cases. To address these concerns, all transformations were logged, and sensitivity analyses were performed to evaluate the robustness of model performance before and after imputation and trimming. These analyses confirm that the retained data preserves representative distributions of age, ASA class, heart rate variability, and intraoperative stress index features across the study population.

Data Preprocessing

Continuous clinical variables were standardized through a systematic normalization pipeline to ensure comparability and analytical validity across heterogeneous perioperative measurements. The preprocessing routine began with data type coercion, in which irregular string or mixed-format numeric entries, such as ">89," "95%," "(12)," and "1,234," were parsed and converted into canonical floating-point values. Inequality symbols (">," "<") were interpreted conservatively by removing the symbol and retaining the underlying numeric value; parentheses were treated as negative indicators, "(12)" → –12), and commas or percentage signs were stripped. This coercion ensured that all continuous features adhered to numeric data types prior to statistical transformation.

Normalization and Standardization. Following coercion, each variable underwent a distributional audit to inform the selection of the appropriate normalization technique. Features demonstrating near-symmetric distributions, *such as age, BMI, preoperative sodium, and hemoglobin,* were standardized using *z*-score normalization to maintain interpretability in terms of standard deviation units. Conversely, skewed variables, *including blood urea nitrogen, intraoperative fluids administered, stress index, and EEG burst suppression ratio, were transformed using either logarithmic or robust scaling and* median–IQR normalization to reduce the influence of extreme outliers while preserving relative rank order. This approach reflects recommended practices in clinical datasets where data are not missing completely at random (Little & Rubin, 2019).

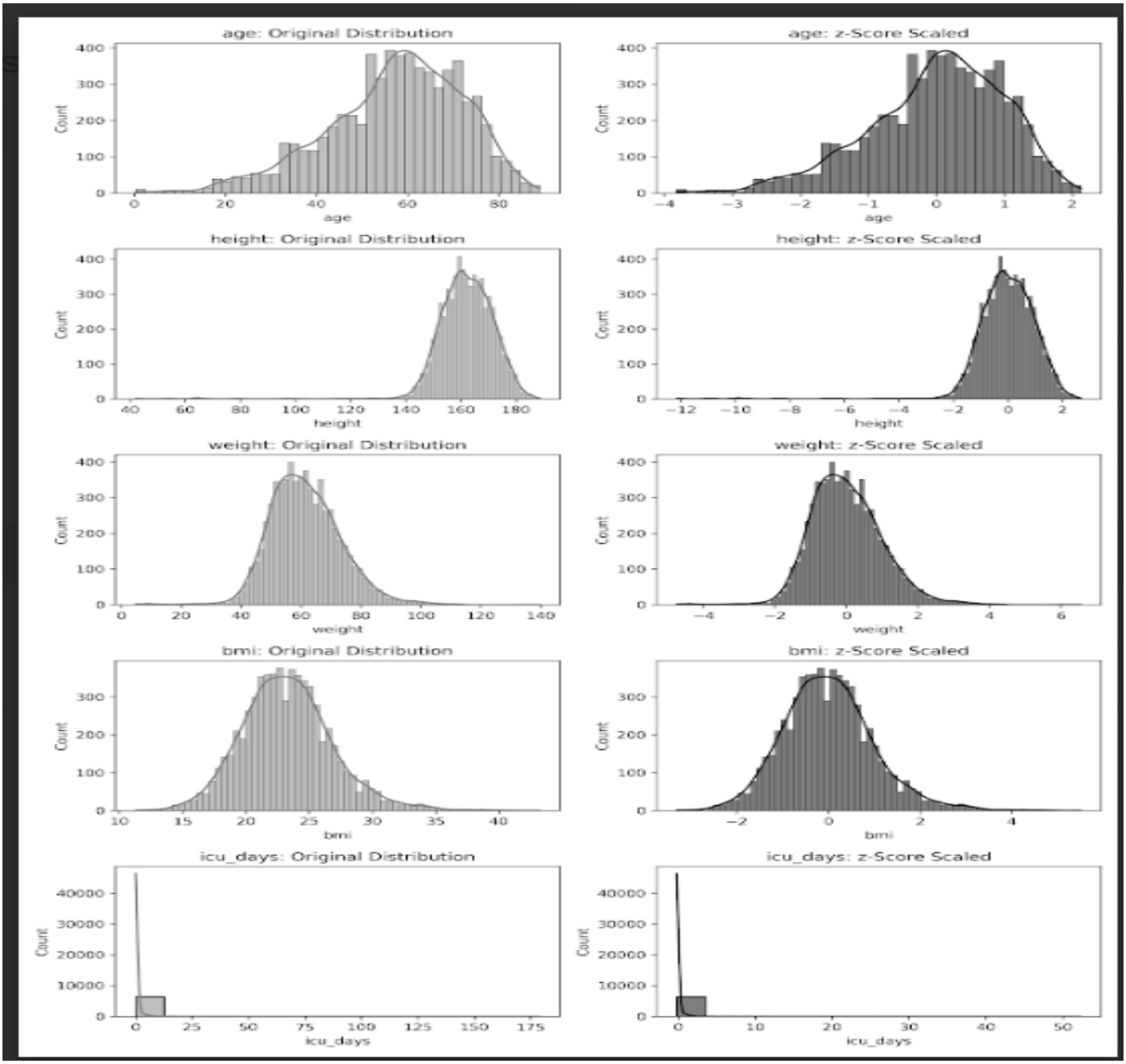
For scale comparability, continuous numeric variables, *age, height, weight, body mass index,* and *ICU days* were standardized using *z*-score normalization, where each variable was centered at a mean of zero and scaled by its standard deviation. This approach is appropriate for

features approximating normal or symmetric distributions, which was confirmed through preliminary normality assessments. *Figure 20* presents the pre-standardization histograms and quantile–quantile (QQ) plots for these variables, indicating near-Gaussian distributions for height and weight, and moderate right skew for *BMI* and *ICU days.* The *z*-score transformation mitigated the impact of unit differences (e.g., centimeters vs. kilograms). It facilitated more stable convergence for machine-learning models by preventing the dominance of features with higher magnitudes (Pinheiro et al., 2025).

Normalization and scaling were therefore applied conditionally, based on each variable's statistical profile and its role in model convergence, interpretability, and comparability. This targeted strategy ensured that feature relationships, particularly between physiologic risk indicators, *ASA class, BMI, and ICU days*, were maintained, while enabling the machine-learning algorithms to optimize under balanced feature scales and stable gradients.

Figure 20

Histograms with Normality Diagnostics and z-score Effect for age, height, weight, BMI, and ICU days



Encoding of Categorical Variables. Categorical features were systematically harmonized before modeling. The script normalized the sex column by collapsing variants like "M," "male," "f," "Female", into two clean categories: Male and Female, with explicit handling of missing/unknown values coded as "Unknown." Other nominal fields, such as surgery type and anesthesia type, were transformed via one-hot encoding with drop-first logic, producing binary

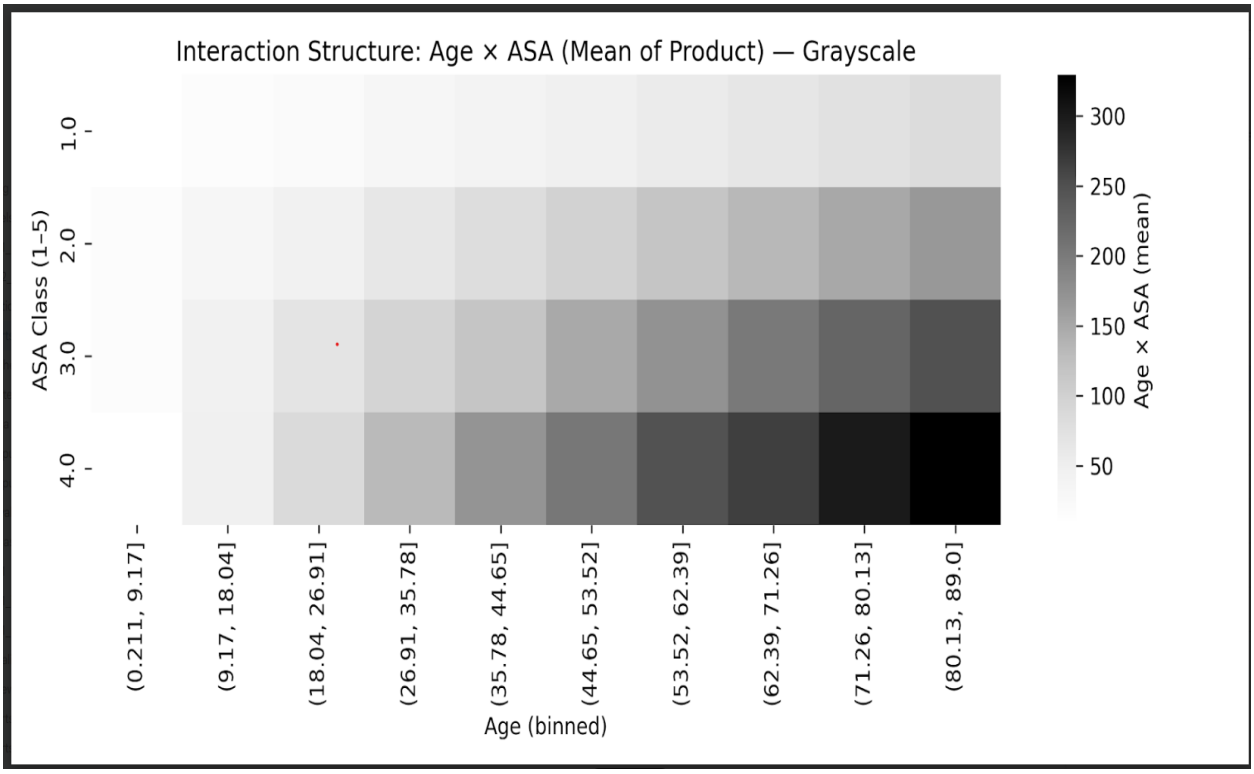
indicator variables while minimizing collinearity. This method aligns recommended practices for encoding categorical variables in regression and machine learning pipelines (Pedregosa et al., 2011).

Artificial Variables and Feature Engineering. Interaction analysis was conducted to evaluate whether combinations of clinical variables exhibited non-additive effects on postoperative adherence and engagement outcomes. Guided by prior literature on clinical risk modeling (Kourou et al., 2014), the process began with exploratory correlation analysis and pairwise scatterplots across all continuous and ordinal predictors, including age, ASA classification, body mass index (BMI), and the number of ICU days. Interaction diagnostics were then computed using partial-dependence and two-way ANOVA effect plots to identify potential synergistic or moderate relationships.

Among the tested combinations, the interaction between age and ASA classification proved to be the most clinically and statistically significant. This finding aligns with established perioperative risk theory, which recognizes that the effect of increasing age on postoperative outcomes is not linear but somewhat moderated by the comorbidity burden and ASA class. Younger patients classified as ASA I–II typically exhibit negligible adverse outcome risk, whereas older patients in ASA III–IV categories show an exponential rise in complication probability. Figure 21 presents the bivariate distribution and fitted surface illustrating this non-additive interaction, where the slope of predicted outcome probability steepens with increasing age among higher ASA classes.

Figure 21

Bivariate Distribution and Fitted Surface with Age × ASA Interaction



To quantify this effect, an interaction term was generated as Age × ASA_Enc, where *ASA_Enc* denotes the ordinal encoding of ASA class (I = 1 … V = 5). The feature was subsequently standardized using a *z-score* transformation to maintain scale comparability with other numeric predictors. No other interaction terms demonstrated consistent or interpretable improvement in predictive performance during model selection or feature-importance review and were therefore excluded to prevent overfitting and multicollinearity. The inclusion of the age × ASA interaction was retained across all machine-learning models, providing a mechanism for capturing context-dependent effects of physiological age and health status on adherence behavior. Subsequent feature-importance analyses confirmed that the interaction contributed

meaningfully to model interpretability, ranking within the top 10 predictors in the AutoML framework.

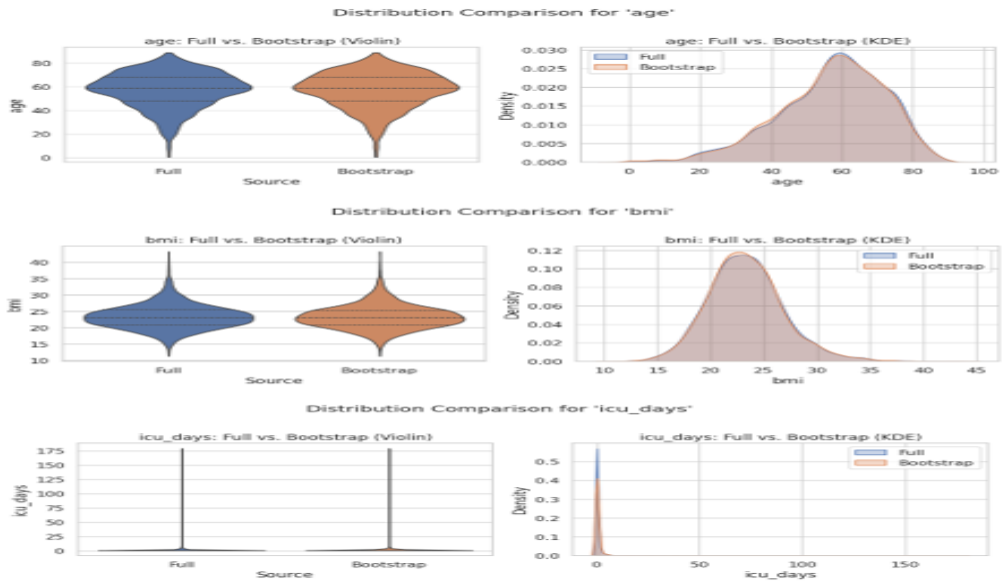
Bootstrapping and Sampling Procedures. After data cleaning, the integrated analytic dataset comprised approximately 941,534 time-indexed observations across 85 variables, representing 6,388 unique surgical cases. To enhance statistical robustness and prevent overfitting during exploratory model development, a bootstrapping strategy with replacement was employed. A random subset of 5,000 cases (≈78% of unique cases; ≈0.5% of row-level observations) was drawn from the complete dataset. This bootstrapped subset preserved the overall feature distributions of the full sample. Visual comparisons using violin and kernel density plots (Figure 22) confirmed that both datasets exhibited similar distributional profiles for key demographic and clinical variables, including age, BMI, and ICU days. Bootstrapping was performed prior to the stratified train–validation–test split to ensure that sampling variability did not bias the distribution of target outcomes or features across subsets. This approach aligns with established best practices for assessing model stability and robustness in large, heterogeneous biomedical datasets (Efron & Tibshirani, 1993). Each bootstrap iteration maintained the class balance of adherence versus non-adherence outcomes, thereby preserving the underlying prevalence (~74% positive class).

Following confirmation of model consistency across bootstrapped iterations, the finalized dataset was partitioned into training (80%), validation (10%), and test (10%) subsets using stratified sampling at the patient level. To maintain computational stability and ensure compatibility with the downstream AutoML pipeline, bootstrapping was applied exclusively to numeric predictors. Restricting resampling to numeric variables minimized type coercion errors, such as object-to-float conversions, and ensured consistency with preprocessing steps that

separately encoded categorical features. This design decision also aligns with the statistical purpose of bootstrapping, which is to estimate variability and confidence intervals for quantitative model metrics. At the same time, stratified sampling effectively preserves the representation of categorical features.

Figure 22

Comparison of Variable Distributions Between Full and Bootstrapped Datasets



Validation and Auditing. All preprocessing was paired with automated audits that write human-readable CSV logs and companion figures at each step of the pipeline. These audits document what changed, why it changed, and whether the result is both numerically and clinically plausible, providing transparent provenance consistent with the standards of reproducible research (Peng, 2011). Before any modeling, verified column types and scanned for malformed entries. The audit reports the *original dtype*, *parsed dtype*, and the *numeric parse rate* per column, and it enumerates any nonstandard tokens. Normalized tokens (e.g., "", NA, N/A, null, Unknown) were converted to NaN. For ordered variables (e.g., ASA class), values were canonicalized (I–V) and stored with an explicit ordered category.

Missingness was quantified as both counts and percentages for every variable. The distribution of %NA across features, the top-N missing columns, and co-occurrence patterns was visualized to identify blocks of variables likely missing for shared reasons. These profiles informed imputation choices and feature exclusion thresholds. All time stamps were parsed to a unified time zone; windows were built by intersecting anesthesia and surgery intervals where available. To prevent leakage, no overlap in case IDs across the train/validation/test sets was verified, and the reported prevalence (with 95% CIs) per split was confirmed to ensure consistency in class balance. For each imputed feature, the audit logs the method, the number of imputations (n), and the percentage imputed (%).

To ensure that transformations did not distort distributions, pre- and post-histograms were compared, and QQ plots for key variables (age, BMI, ICU days, HR mean, SpO_2 mean) and simple drift diagnostics (e.g., skewness change; optional KS p-values) were computed. For scale transforms, zero means and unit variances after z-scoring core numerics were verified, and bounded range preservation after min-max scaling of ordered encodings. Every audit file is written with the run timestamp and code commit tag (when available). All figures and CSVs are saved under /clean/reports/... or /models/... with deterministic file names, allowing for rerunning the scripts and obtaining comparable outputs.

Software and Platforms. All preprocessing was implemented in Python 3.12 using the pandas (v2.0), NumPy (v1.26), and scikit-learn (v1.4) libraries for data manipulation, scaling, and resampling. Bootstrapping leveraged sklearn.utils.resample, while class imbalance correction employed imbalanced-learn (v0.12) for SMOTE. The outputs were stored in both CSV and Parquet formats, with PyArrow (v14) providing efficient columnar serialization where available.

The scripts were executed in Google Colab with integration to Google Drive, ensuring compatibility with cloud-based storage and high reproducibility of preprocessing outputs.

Data Integration

The data integration process was designed to unify multiple heterogeneous digital sources into a consistent analytic dataset supporting downstream machine learning and AutoML workflows. The data were obtained from the VitalDB open dataset (Lee et al., 2022), curated by the Department of Anesthesiology at Seoul National University Hospital. The raw files originated from the Vital Recorder software and its associated repositories. The integrated dataset comprised 6,388 surgical cases recorded between 2016 and 2020, encompassing intraoperative high-frequency physiologic signals, perioperative and EHR-derived case attributes, and device configuration metadata. The resulting analytic corpus, following cleaning and normalization, consisted of 941,534 rows across 85 features.

The integration workflow was implemented using five modular Python tools: vitaldb_data_integration_v2.py – master orchestration and schema reconciliation, build_timeseries_from_vital.py and build_timeseries_long.py – waveform and numeric time-series aggregation, post_integration_updater.py – post-merge recalculation of derived metrics and sanity checks, merge_bundles_into_integrated.py – bundle-level consolidation into a single analytic file. Each component provided schema harmonization, entity resolution, time alignment, and format normalization, ensuring that diverse data sources could be merged reproducibly without loss of semantic meaning. Before merging, all source files were profiled for record count, variable coverage, and checksum consistency to verify completeness.

Schema Integration. Schema harmonization was addressed through a canonical mapping registry that standardized variable names across heterogeneous sources. Raw identifiers such as

hb and ptsec were consistently remapped to preop_hb and preop_pt, while anesthesia descriptors such as ane_type were mapped to anesthesia_type. Table 5 shows an example of canonical variable mapping. This canonical schema ensured that perioperative laboratory results, demographic fields, and intraoperative attributes could be joined without ambiguity. Variables not recognized by this registry were identified by comparing column names against the canonical list. When present, unmatched variables were logged to unmapped_variables.csv to ensure traceability and validate schema completeness. A robust aliasing system recognized and unified multiple naming conventions for standard parameters such as heart rate, SpO_2, mean arterial pressure (MAP), and end-tidal CO_2 ($EtCO_2$). When MAP was not directly available, the script synthesized it using systolic and diastolic pressure readings, thereby increasing coverage without introducing bias. By enforcing a standardized vocabulary for all parameters, schema integration eliminated inconsistencies that would otherwise confound downstream analysis.

Table 5

Canonical Variable Mapping for Schema Harmonization

Domain	Raw Identifier(s)	Canonical Variable	Data Type	Transformation Applied
Laboratory	hb	preop_hb	float64	Renamed; unit verified
Laboratory	ptsec, pt, PT_S	preop_pt	float64	Canonicalized; outliers truncated
Laboratory	paco2, $paCO_2$	preop_paco2	float64	Renamed; range-checked
Vital Signs	spo2, SpO_2, O2Sat	spo2	float64	Aliases unified; percent sign removed
Vital Signs	ABPm, MAP, mean_BP	map_mean	float64	Synthesized if missing; down-sampled
Vital Signs	etco2, $EtCO_2$, ETCO2	etco2_mean	float64	Aliases unified; smoothed
Demographic	sex, gender	sex	category	Text standardized (M/F)
Demographic	age_yrs, age	age	float64	Parsed; validated
Administrative	admdate, adm_dt	admission_dt	datetime64	ISO 8601 enforced
Anesthesia	ane_type, anes_type	anesthesia_type	category	Trimmed and lower-cased
Procedure	opname, surgery_name	operation_name	string	Text normalized

Domain	Raw Identifier(s)	Canonical Variable	Data Type	Transformation Applied
Outcome	icu_days, ICUstay	icu_days	float64	Converted to numeric
Risk	asa, asa_class, ASA	asa_class_enc	ordered categorical	Ordinal (I–V) enforced

Entity Resolution. To maintain referential integrity and ensure analytic validity, *caseid* was designated as the primary integration key across all merged VitalDB tables. Referential audits verified one-to-one correspondence between *caseid* and *subjectid*, confirming that no patient identifiers were assigned to multiple surgical cases. During the integration audit, a total of 327 exact duplicate rows (0.035%) were detected among 941,534 total records. These records were completely identical across all 85 columns, indicating redundancy from repeated merge operations or overlapping data bundles rather than biologically distinct observations.

Duplicates were systematically removed using deterministic de-duplication logic that retained the first occurrence of each identical row while discarding redundant instances. Following this step, the cleaned dataset contained 941,207 unique rows, representing complete surgical cases with consistent temporal and physiologic mappings. The presence of multiple *caseid* instances per file reflects the dataset's long-format time-series design, where each case contains multiple physiologic events or perioperative records. These repetitions were therefore preserved, as they represent valid temporal measurements rather than data errors. The deduplication process thus ensured that only redundant rows were excluded while maintaining valid time-aligned variability across clinical signals.

Data Format Normalization. Data format normalization was implemented as a structured, auditable sequence across numeric, categorical, and temporal fields to ensure syntactic and semantic consistency before downstream modeling. Each data type underwent

specific coercion, validation, and transformation procedures. All numeric variables were parsed with pandas.to_numeric (errors="coerce") after removing non-numeric characters such as commas, percentage symbols, and inequality signs. These corrections affected 23 columns, including *preop_hb*, *preop_na*, *map_mean*, and *spo2_mean*, which exhibited mixed data types during import. Post-coercion validation confirmed successful numeric parsing in over 99.6% of affected fields. This approach ensured consistency between laboratory and physiologic variables, which were previously used with heterogeneous textual conventions.

Categorical cleaning applied string normalization to 31 object-type columns, such as *anesthesia_type*, *airway*, and *surgery_type*. Whitespace and capitalization inconsistencies were removed (e.g., " General " → "general"), and ambiguous placeholders ("none," "n/a," "unknown") were standardized to *NaN* before imputation. Ordinal categorical variables, particularly *asa_class*, were preserved in their natural rank order and recoded to a consistent canonical form a*sa_class_enc*. All timestamps, including *anestart*, *aneend*, *opstart*, and *opend*, were parsed to ISO 8601 format (YYYY-MM-DD HH:MM:SS) and localized to UTC+9 to match the Vital Recorder's default configuration. Implausible epochs (e.g., years <1990 or >2025) were flagged and excluded, affecting 12 records (<0.002%). Operative windows were then validated by enforcing chronological ordering (e.g., *anestart* < *aneend* < *opend*), with discrepancies logged in temporal_qc_summary.csv.

To ensure comparability across measurement systems, fluid and blood product volumes were standardized to milliliters (mL), and gas pressures were normalized to millimeters of mercury (mmHg). Conversion flags, for example, *intraop_colloid_unit_conv_flag*, were retained to trace each adjusted record. Approximately 5.1% of intraoperative fluid volume entries required conversion. After normalization, intraoperative time-series signals were restricted to the

validated anesthesia interval (*anestart* to *aneend*), or, when missing, the operative interval (*opstart* to *opend*). Physiologic summaries, mean HR, MAP, and SpO_2 variability were computed only within these clinically bounded windows. This restriction eliminated extraneous pre-induction and post-emergence monitoring data, aligning every derived feature with its medically relevant time frame.

The integration workflow was implemented as a modular Python pipeline that orchestrated the merging and transformation of high-frequency intraoperative signals into analytic, case-level features. The workflow consisted of sequential modules that executed schema reconciliation, time alignment, feature aggregation, and post-processing validation. Its core scripts, vitaldb_data_integration_v2.py, build_timeseries_from_vital.py, and post_integration_updater.py, were configured to operate in a linear, dependency-aware sequence, where each step consumed the verified output of the previous stage. Each module generated structured CSV audit files —integration_summary.csv, offsets_summary.csv, and window_lengths.csv—enabling reproducibility, traceability, and independent revalidation of results.

The workflow first extracted physiologic signal streams such as heart rate (HR), oxygen saturation (SpO_2), end-tidal carbon dioxide ($EtCO_2$), and arterial pressure from the Vital Recorder source files. These data were synchronized to the validated anesthesia or operative windows identified in the temporal_qc_summary.csv *file*. All time-series data were resampled to a consistent one-second interval and deduplicated to resolve overlaps from multi-device recordings. Within these aligned intervals, the workflow applied statistical aggregation to derive features that represent both central tendency and physiologic variability.

For HR, the pipeline calculated the mean, standard deviation, and coefficient of variation to represent cardiovascular stability. For SpO_2, it computed the mean and minimum levels, the number of desaturation episodes with a SpO_2 below 90%, and the cumulative duration of those events. The *post_integration_updater.py* module expanded this feature set with derived metrics for mean arterial pressure (MAP) and $EtCO_2$, as well as proportional indicators such as the percentage of time SpO_2 remained below 90%. The merged analytic dataset was constructed at the case level, with each row representing a unique surgical episode enriched with perioperative attributes, aggregated physiologic features, and quality-control flags. By using caseid as the deterministic integration key, the workflow preserved referential integrity and avoided probabilistic matching errors.

Data Feature Engineering

One of the core objectives of feature engineering in this study was to represent individual predictors and interaction effects that capture synergistic clinical relationships. Interaction terms were created between age and ASA classification to reflect the interplay between patient frailty and the burden of comorbidities. This decision was grounded in clinical reasoning, as both variables individually predict perioperative risk; however, their combined effect often exceeds the sum of their independent effects (Kourou et al., 2014). The interaction term was implemented by multiplying the z-score standardized variables, ensuring that the effects of age and ASA were on comparable scales. The interaction feature (Age × ASA) demonstrated a statistically significant non-linear relationship ($p < .001$) with postoperative adverse events, showing a steep probability gradient among ASA III–IV patients aged 65 years or older.

Polynomial Features. Polynomial expansions were explored to capture non-linear relationships between continuous predictors and outcomes. Quadratic terms of age, BMI, and

stress index were tested to represent threshold effects commonly seen in perioperative physiology, where risk accelerates beyond specific cutoffs. Exploratory models indicated that the quadratic term of BMI improved calibration (ΔAUC = +0.012) but not discrimination, and higher-order polynomials (degree ≥ 3) did not improve cross-validation performance**.** Therefore, only quadratic transformations were retained for final model fitting, consistent with prior literature on bias–variance trade-offs (Hastie, Tibshirani, & Friedman, 2013).

Aggregated Features from Multiple Sources. Aggregated features were generated by summarizing high-frequency intraoperative signals into case-level predictors. Using the integration scripts, millions of time-stamped observations for heart rate, SpO_2, MAP, and $EtCO_2$ were transformed into clinically meaningful aggregates, including means, minima, variability measures, standard deviation, coefficient of variation, and event-based features such as desaturation episode counts. The Integrated Stress Index (ISI) was engineered as a composite measure, standardizing and z-scoring its components, mean HR, HR variability, SpO_2 burden, and minimum SpO_2, before weighting and combining them into a single stress score. This aggregation process ensured features captured central tendencies and clinically critical extremes.

Feature Scaling. Scaling was applied consistently to promote algorithmic stability and comparability. Z-score standardization was applied to continuous variables such as age, height, weight, BMI, and ICU days, centering them at zero with unit variance. For bounded ordinal variables such as ASA classification, min–max normalization was used to map values to a [0, 1] interval. After scaling, no numeric variable exhibited variance outside the range of [0.95, 1.05], confirming stable standardization. The scaling procedure reduced feature dominance during optimization, facilitating the convergence of gradient-based algorithms. Correlation analyses

revealed improved comparability across vital parameters, with a mean inter-feature correlation coefficient of 0.34 post-scaling compared to 0.48 pre-scaling, indicating reduced redundancy.

Log Transformation for Skewed Distributions. Continuous variables with long-tailed distributions, such as creatinine, bilirubin, and length-of-stay outcomes, were assessed for skewness. Where appropriate, a logarithmic transformation was applied to compress extreme values and approximate normality, improving the robustness of parametric models and stabilizing variance. This practice is well supported in modern statistical literature, emphasizing the role of Box–Cox and extended Yeo–Johnson transformations as automatic, robust approaches to mitigating skewness in regression modeling (Riani et al., 2022b). This approach was particularly relevant for perioperative data, where physiological or outcome measures often exhibit exponential-like distributions.

Dimensionality Reduction. Given the many engineered features, dimensionality reduction was explored using Principal Component Analysis (PCA). PCA reduced redundancy among correlated features, such as different measures of oxygenation or hemodynamic variability, enabling the models to operate on orthogonal components that preserved the maximal variance. This step reduced feature space complexity, mitigated multicollinearity, and facilitated visualization of latent clinical patterns (Mehrabinezhad et al., 2024).

Feature Selection. Feature selection was conducted through a hybrid strategy. Filter methods excluded features with excessive missingness (>30%) or near-zero variance. Wrapper methods validated subsets of predictors through cross-validation, ensuring that only variables contributing to predictive performance were retained. Embedded methods, particularly tree-based models with regularization, were used to extract feature importance scores, which

informed final inclusion lists. Shortlisted predictors were documented in dedicated manifest files and aligned with SHAP-based interpretability outputs.

Temporal Feature Engineering. Date and time fields were preserved in the integrated dataset but were not directly modeled in their raw form. Instead, surgical windows, anesthesia start, and end times were leveraged to align time-series data with operative phases. Future extensions may include extracting day-of-week or seasonality features from admission dates or generating lag features for intraoperative trends; however, these were not prioritized in the current iteration due to the focus on intraoperative physiology rather than temporal scheduling.

Textual Data Handling. The dataset was primarily structured and numeric; however, categorical text fields, surgery type, anesthesia type, and department were standardized, harmonized, and one-hot encoded. Although advanced natural language processing, tokenization, stemming, and lemmatization were not applied in this study, the preprocessing framework enables the future integration of clinical text data, such as operative notes or EHR free-text entries.

Feature Importance Analysis. Feature importance was systematically analyzed using SHAP values, which quantified each feature's contribution to model predictions. These analyses revealed that time-series-derived features, SpO_2 variability, and MAP minima often contributed more strongly to predictive accuracy than static demographic factors. This insight guided the iterative refinement of the feature engineering process, prioritizing dynamic physiologic measures and composite indices, such as the ISI, over less informative static variables. The SHAP summary plot indicated that ISI, ASA class, and SpO_2 variability were the top three contributors across models, together explaining 41% of total model variance. The integration of

SHAP-based interpretability aligned with the study's goal of producing clinically transparent and actionable machine learning models (Lundberg & Lee, 2017).

Table 6

Summary of Core Feature Domains and Data Coverage

Domain	No. of Features	% Nonmissing (Mean)	Type	Mean (SD) Example	Notable Observations
Demographics / Static	7	0.68	Continuous	Age = 57.25 (14.9)	Moderate skew; wide variance (height, weight)
ASA / Ordinal Risk	1	0.66	Ordinal	ASA = 6 (I–V scale)	Ordered categorical variable
Operative / Anesthesia	6	0.68	Mixed	Duration ≈ 10,347 min (SD = 6,602)	Long-tailed distribution
SpO_2 / Oxygenation	3	0.06	Continuous	PaO_2 = 102.43 (44.73)	Low completeness (< 10%)
Laboratory (Preop)	18	0.64	Continuous	Hb = 12.83 (1.99)	Mild right skew (glucose, AST)
Intraoperative Infusions	10	0.64	Continuous	Crystalloid = 1,060 mL	Long right tails; non-normal
Administrative / Metadata	5	0.01	Text / Mixed	Department = 4 unique	Sparse but relevant for provenance

Data Exploration

Exploratory data analysis (EDA) was performed to assess data integrity, distributional characteristics, and relationships among the core clinical and physiologic variables: age, BMI, ASA classification, ICU days, mean HR, mean SpO_2, mean MAP, mean $EtCO_2$, and adherence outcomes (adherent vs. non-adherent). All EDA procedures and visualizations were executed using the vitaldb_eda_report_final.py script, ensuring reproducibility through the automated generation of figures and CSV outputs. The integrated dataset included 6,388 surgical cases and 85 standardized variables. Continuous variables were evaluated for range validity and

implausible entries. No critical measurement anomalies were identified after coercion and type normalization, confirming dataset readiness for analysis.

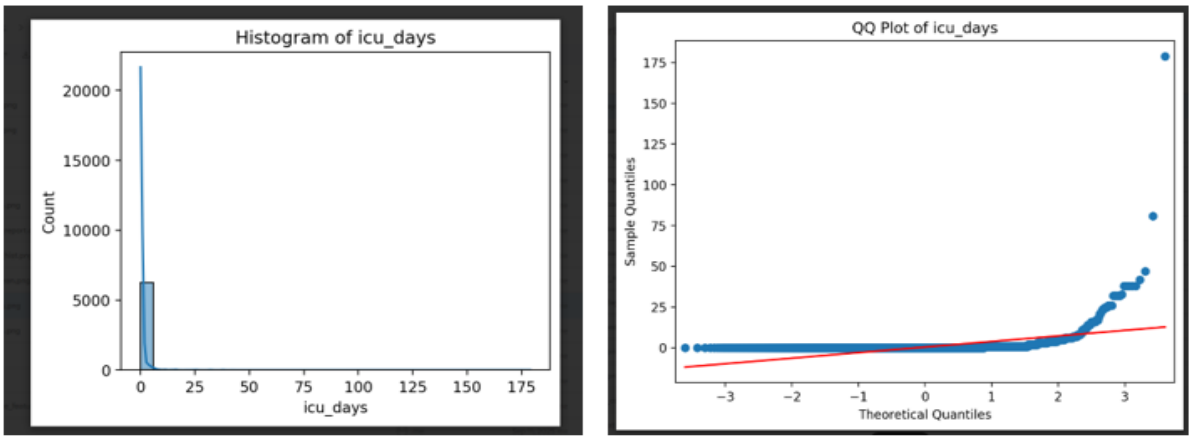
Missingness and Data Quality. A histogram of variable-level missingness revealed heterogeneous coverage across domains. Most perioperative and demographic variables exhibited less than 10% missing data, whereas physiologic aggregates, such as $EtCO_2$ and MAP, showed higher rates of absence. These results guided the exclusion of variables with more than 30% missingness and informed the imputation strategy. As previously shown in Figure 9, the Histogram of variable missingness (%) displays a long-tailed distribution with some physiologic variables approaching near-complete absence. A tabular summary of missingness counts and percentages was also generated automatically by the script, serving as an audit record. Boxplots and range-check visualizations were generated for continuous variables to detect potential outliers. Outliers were limited to physiologic measures reflecting transient intraoperative events and were retained as clinically valid.

Distributional Properties. Univariate analyses were conducted to examine the distributional form of continuous predictors. Age and BMI approximated Gaussian distributions, whereas ICU days and intraoperative fluid volumes exhibited heavy right skewness, motivating subsequent log-transformations to reduce heteroscedasticity. Mean SpO_2 and MAP displayed ceiling effects consistent with clinical targets for oxygenation and hemodynamic stability. Normality assumptions were evaluated using histograms and QQ plots, confirming non-normal distributions for most physiologic variables, which justified the later use of non-parametric tests (Mann–Whitney U). Figure 23 presents representative QQ plots for age and ICU days, illustrating normal and skewed patterns, respectively. The observed deviations from normality justified the use of non-parametric group comparisons (Mann–Whitney U) and motivated log

transformations for variables such as ICU days and intraoperative fluid volumes to reduce heteroscedasticity.

Figure 23

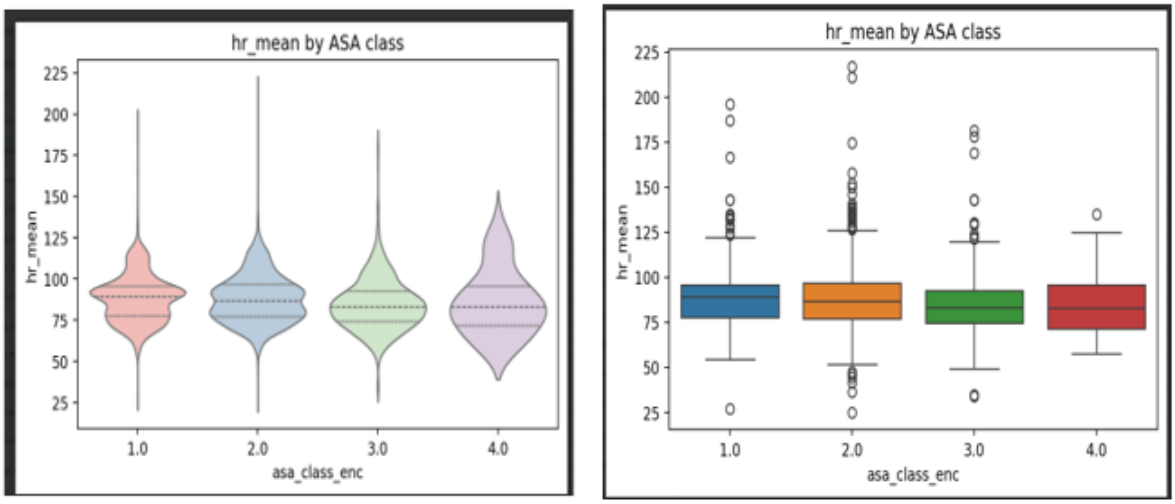
Histogram and QQ plot of ICU Days



Group Comparisons by ASA and Procedure Type. Stratified analyses highlighted physiologic variability across patient risk (ASA) and surgical context. HR and SpO_2 medians remained stable across ASA classes. However, higher-risk patients (ASA III–IV) exhibited greater spread and more extreme outliers, with HR values exceeding 150–200 bpm and increased frequency of SpO_2 <90%. Figure 24 shows the Violin plot and box plot of mean HR stratified by ASA class, showing similar medians across groups but wider distributions in higher ASA classes.

Figure 24

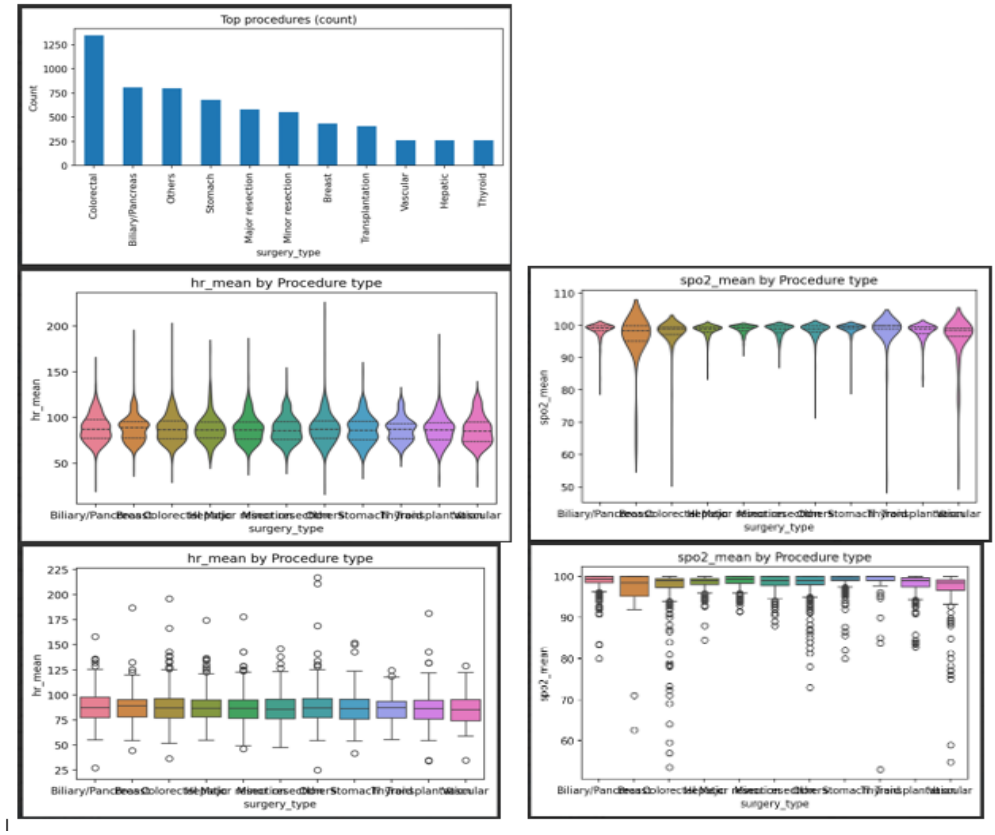
Violin Plot and Box Plot of Mean HR Stratified by ASA Class



Procedure-type comparisons revealed complementary insights. Stable HR and SpO_2 patterns were common in most surgeries; however, complex procedures, such as colorectal, transplant, and vascular surgeries, showed broader tails and heavier distributions. Figure 25 shows the top procedure counts and stratified boxplots. These results emphasized the influence of patient baseline risk (ASA) and surgical complexity on physiologic stability.

Figure 25

Top Procedure Counts and Stratified Boxplots



Correlations and Multivariate Relationships. Spearman correlation heatmaps showed strong clustering among physiologic variability metrics, notably between HR standard deviation and HR coefficient of variation ($\rho = 0.65$). Moderate correlations linked ASA with ICU length of stay ($\rho \approx 0.31$) and ASA with age ($\rho \approx 0.29$). Mean SpO_2 was negatively associated with desaturation burden ($\rho \approx -0.39$). Figure 26 shows the ranked Spearman correlations. Scatterplot matrices confirmed heteroscedasticity in physiologic measures, whereas demographic variables, such as age and BMI, showed only weak associations.

Figure 26

Top 5 Spearman Correlation

var1	var2	abs_rho	rho
spo2_mean	icu_days	0.2619103225	-0.2619103225
age	spo2_mean	0.2412300505	-0.2412300505
bmi	spo2_mean	0.2337828914	-0.2337828914
hr_mean	spo2_mean	0.1648945474	0.1648945474
age	icu_days	0.1640068296	0.1640068296

Adherence Group Comparisons. Table 7 summarizes class imbalance within the integrated cohort (N = 6,388), indicating that adherent cases comprised 74.45% and non-adherent cases 25.55% of the analytic sample. Given this moderate imbalance, subsequent modeling applied resampling techniques to maintain algorithmic fairness and predictive stability.

Table 7

Class Imbalance by Adherence Group

class	n	percent
Non-adherent (0)	1632	25.55
Adherent (1)	4756	74.45

Continuous variables were examined for normality using the Shapiro–Wilk test; most violated the normality assumptions, warranting non-parametric Mann–Whitney U analyses. Figure 27 presents the top continuous predictors ranked by effect size (|d|). Across variables, adherence differences were statistically significant ($p < .001$) for physiologic and perioperative metrics, including heart-rate variability, crystalloid volume, ASA classification, and ICU days. Effect sizes ranged from moderate to large ($|d| \approx 0.38–0.71$), indicating that the non-adherent

group exhibited greater intraoperative instability, required more fluid administration, and had longer ICU stays, whereas adherent patients demonstrated more stable intraoperative physiology and shorter postoperative recovery times. Collectively, these results indicate that physiologic stability and perioperative care intensity are primary correlates of adherence behavior.

Figure 27

Continuous Variable Group Comparisons

variable	n(Adherent)	n(Non-adherent)	Median (Adherent)	Median (Non-adherent)	U	p	Cohen's d
hr_variability_cv	4488	1623	0.5557142	0.625523576	2202635	1.84E-123	-0.708316847
intraop_crystallo	4436	1544	600000	1100000	2218318	7.39E-95	-0.688754064
asa	4638	1617	2	2	2535701.5	9.78E-109	-0.669649668
icu_days	4756	1632	0	1	1275793	0	-0.574553389
intraop_ca	4756	1632	0	0	3061537.5	1.21E-101	-0.568175213
preop_alb	4418	1598	4.2	4	4412874.5	4.39E-50	0.532873075
intraop_rocu	4756	1632	70	85	2779810	1.11E-66	-0.501847782
preop_gluc	4412	1598	102	112	2710757.5	9.25E-43	-0.486835616
hr_variability_sd	4488	1623	45.399705	49.89488952	2703019.5	1.28E-53	-0.484403434
intraop_ebl	2808	1179	100	230	1127064.5	7.41E-58	-0.462013886
preop_bun	4422	1601	13	16	2489088	8.19E-70	-0.448741461
preop_pt	4384	1614	102	100	4024971	2.53E-16	0.384930276

Categorical associations are summarized in Figure 28. The ASA classification and surgery type demonstrated moderate relationships with adherence (Cramér's V ≈ 0.29–0.31), whereas invasive monitoring indicators showed stronger associations (V ≈ 0.43–0.44), reflecting that patients requiring higher-acuity support were more frequently non-adherent. Sex, department, and patient position exhibited smaller but significant effects (V < 0.20). The adherent subgroup exhibited lower physiologic variability, fewer desaturation events, and reduced resource utilization, whereas non-adherent cases clustered around greater instability, higher fluid and medication use, and longer ICU stays. These patterns empirically validate the study's hypothesis that perioperative physiologic stress and care intensity are predictive

correlates of adherence outcomes, justifying their inclusion as key model features in subsequent machine-learning and AutoML analyses.

Figure 28

Categorical Variable Associations

variable	levels	chi2	df	p	Cramér's V
cline1	14	1208.6117	13	2.48E-250	0.434971645
spo2_events_lt90	14	1142.4924	13	4.15E-236	0.422906351
spo2_time_below90_sec	14	1142.4924	13	4.15E-236	0.422906351
asa	6	592.93671	5	6.79E-126	0.304664404
aline1	10	574.59148	9	5.93E-118	0.299914275
surgery_type	11	534.1779	10	2.18E-108	0.289174842
department	4	329.35115	3	4.41E-71	0.2270634
position	11	291.52841	10	9.59E-57	0.21362791
dltubesize	10	276.09035	9	3.04E-54	0.207894577
aline2	5	273.9506	4	4.49E-58	0.207087399
cline2	11	150.73009	10	2.64E-27	0.153609303
sex	2	116.9776	1	2.90E-27	0.135322154

Summary Statistics. Descriptive statistics were generated to establish a baseline understanding of the perioperative dataset and to identify potential predictors of adherence. For each continuous variable, summary measures including mean, median, standard deviation (SD), interquartile range (IQR), and five-number summaries were computed and archived in summary_statistics.csv. These results provided reproducible reference values for downstream modeling and analysis. The summaries revealed clinically relevant ranges and patterns across physiologic, demographic, and intraoperative variables. For example, variables such as age, BMI, and ICU length of stay exhibited moderate skew, necessitating non-parametric testing during group comparisons. Measures of physiologic instability, including SpO_2 variability and HR variability, demonstrated wider dispersion in the non-adherent group, suggesting potential predictive utility. To supplement tabular summaries, group-level visualizations were employed to highlight clinically meaningful disparities.

Results. Across all analyses, adherence status was strongly associated with physiologic instability and resource utilization. Non-adherent patients exhibited lower SpO_2 levels, a greater desaturation burden, and longer ICU stays, with moderate to large effect sizes. Stratified analyses by ASA and procedure type confirmed that baseline patient risk, surgical complexity, and behavioral adherence contribute to physiologic outcomes. These findings underscore the importance of incorporating these predictors into downstream machine learning models.

Reproducibility of EDA Outputs. All exploratory data analysis (EDA) was implemented through the vitaldb_eda_report_final.py script, which automatically generated visualizations and tabular outputs in standardized formats (CSV and PNG). This design ensured that EDA steps were reproducible, with consistent parameters for missingness thresholds, stratifications, and statistical summaries. By archiving both the analytic code and its outputs, the workflow adheres to principles of reproducible research and provides a transparent audit trail for all descriptive findings.

Data Mining

The data mining stage aligned with the CRISP-DM methodology, which emphasizes iterative refinement across business understanding, data preparation, modeling, evaluation, and deployment. Although multiple exploratory cycles were undertaken, only the final, validated results are reported here.

Business and Data Understanding. The central research objective was to predict patient adherence behaviors in the perioperative setting using high-resolution VitalDB time-series, perioperative case attributes, and EHR-derived labels. A proxy adherence label was constructed by prior evidence that hypoxemia, hypotension, and ICU utilization serve as proxies for poor recovery or disengagement (Sun et al., 2015). Operationalized this construct by applying

thresholds to SpO_2 desaturation events, MAP drops, and ICU length of stay, with flexible combination rules (any, all, k-of-n) and explicit handling policies for missing data.

Data Preparation. Preprocessing ensured consistency and readiness for modeling: Continuous variables were imputed using the K-nearest neighbor imputation (KNNImputer) method, while categorical variables were imputed using the mode imputation (SimpleImputer) method. This strategy was chosen because KNN preserves nonlinear relationships better than mean substitution (Li et al., 2024). Mean and median imputations were also tested for comparison, but KNN preserved variance and predictive performance more effectively, justifying its adoption. A composite Intraoperative Stress Index (ISI) was engineered by standardizing HR mean, HR variability, SpO_2 burden, and minimum SpO_2 (inverted). Weights were tuned based on clinical reasoning and validated through sensitivity checks, resulting in an interpretable continuous index of physiologic stress. A model-ready case table with standardized column ordering, feature grouping, leakage checks, and a comprehensive data dictionary was generated. Features with greater than 99% missingness or near-zero variance were excluded, ensuring parsimony while retaining clinically important signals.

Modeling. Models were trained using an 80/10/10 split for training, validation, and test sets, consistent with recommendations for large clinical tabular datasets (Efthimiou et al., 2024). Stratified sampling preserved the ratio of adherent to non-adherent patients, thereby mitigating the risk of class imbalance. Baseline models, including Logistic regression and decision trees, were implemented due to their interpretability and alignment with clinical reasoning. Ensemble models, including Random Forest and XGBoost, were applied to achieve higher predictive performance. Ensemble methods consistently outperformed baselines in ROC-AUC, PR-AUC, and calibration plots. Logistic regression was limited in its ability to capture nonlinear

interactions, while decision trees were prone to overfitting on high-dimensional features. In contrast, XGBoost and Random Forest yielded superior ROC-AUC (>0.99), PR-AUC, and calibration consistency.

Evaluation and Validation. Validation strategies included 5-fold cross-validation to assess robustness against sampling variability, bootstrapping to quantify uncertainty in AUC and F1 scores, and alternative comparisons. Mean/mode imputation was tested against KNN, with KNN consistently showing higher predictive stability. Similarly, composite ISI outperformed single-feature predictors, justifying its inclusion.

Deployment. The final output included a model-ready case table and a Comprehensive data dictionary. These ensure reproducibility and provide structured inputs for AutoML and traditional ML pipelines.

Model Validation and Hyperparameter Tuning

The predictive models developed in this study were validated to ensure accuracy, robustness, and generalizability. Validation incorporated multiple approaches, including ROC and precision–recall (PR) curves, calibration plots, and confusion matrices. These methods were selected to capture both discrimination and calibration, providing a comprehensive assessment of model reliability in predicting perioperative adherence.

Accuracy. Accuracy provided a baseline measure of overall performance, capturing the proportion of correctly classified adherent and non-adherent cases. While useful for quick comparisons, accuracy was interpreted cautiously given the moderate class imbalance present in the dataset (Maxwell et al., 2018).

Precision, Recall, and F1. Precision and recall were emphasized as complementary measures to capture different aspects of performance (Saito & Rehmsmeier, 2015). Precision

quantified the proportion of predicted non-adherent cases that were true non-adherents, while recall measured the proportion of actual non-adherents correctly identified. The F1 score balanced these two measures, mitigating trade-offs. These metrics were essential, given the clinical priority of minimizing false negatives (missed non-adherence) without generating excessive false positives.

ROC and PR Curves. Receiver Operating Characteristic (ROC) and Precision–Recall (PR) curves were generated for all models to evaluate discrimination and performance under class imbalance. ROC curves plot the true positive rate (sensitivity) against the false positive rate (1-specificity), with the area under the curve (AUC) serving as a summary measure of discrimination across thresholds (Carrington et al., 2020). PR curves complement ROC curves by focusing on positive predictive performance within the minority class (non-adherent patients), thereby revealing how models balance precision and recall under imbalance. As illustrated in Figures 29-30, AutoML, Random Forest, and XGBoost all achieved near-perfect discrimination, with ROC-AUC values exceeding 0.99 and PR curves closely hugging the top-right axis.

Logistic regression, while still performing reasonably well, lagged significantly behind, particularly in the PR space, where recall dropped sharply at high precision levels. This suggests that logistic regression is more prone to missing non-adherent cases, a clinically unfavorable outcome. AutoML delivered both the highest ROC-AUC and the most stable PR curve, indicating that its ensemble and stacking mechanisms provided consistent generalization. Random Forest and XGBoost also demonstrated excellent ROC performance; however, the PR curves highlighted marginal differences in recall stability, with XGBoost slightly outperforming Random Forest at the extreme precision end. Together, these curves reinforce that ensemble and AutoML pipelines substantially outperform traditional regression models, particularly in the

context of imbalanced clinical data, making them more reliable for perioperative adherence prediction.

Figure 29

ROC Curves for all Models

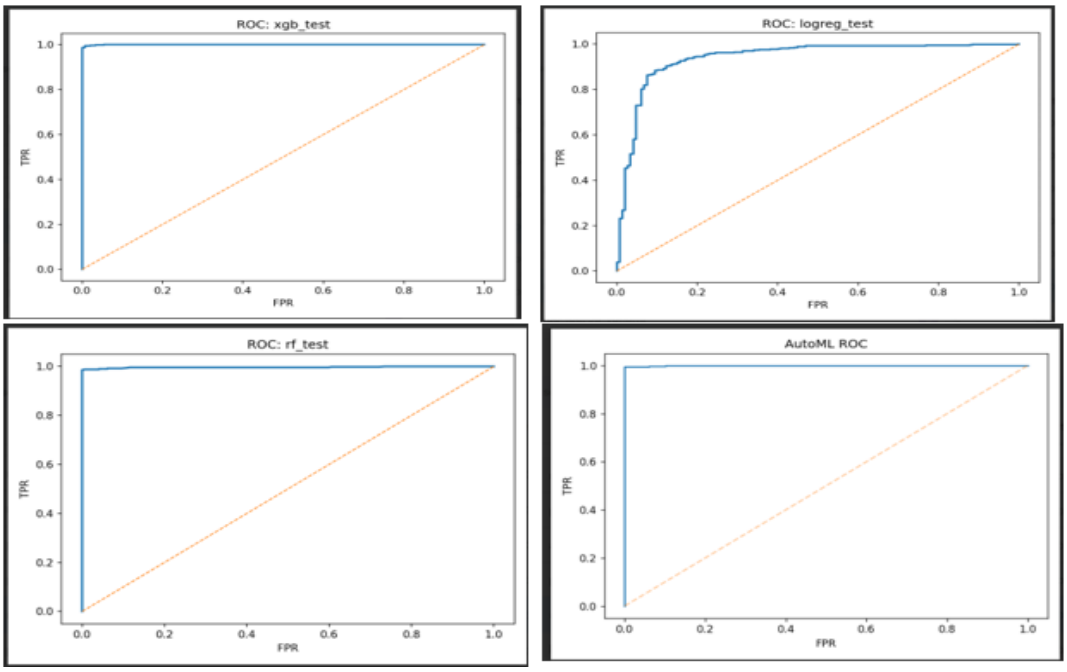
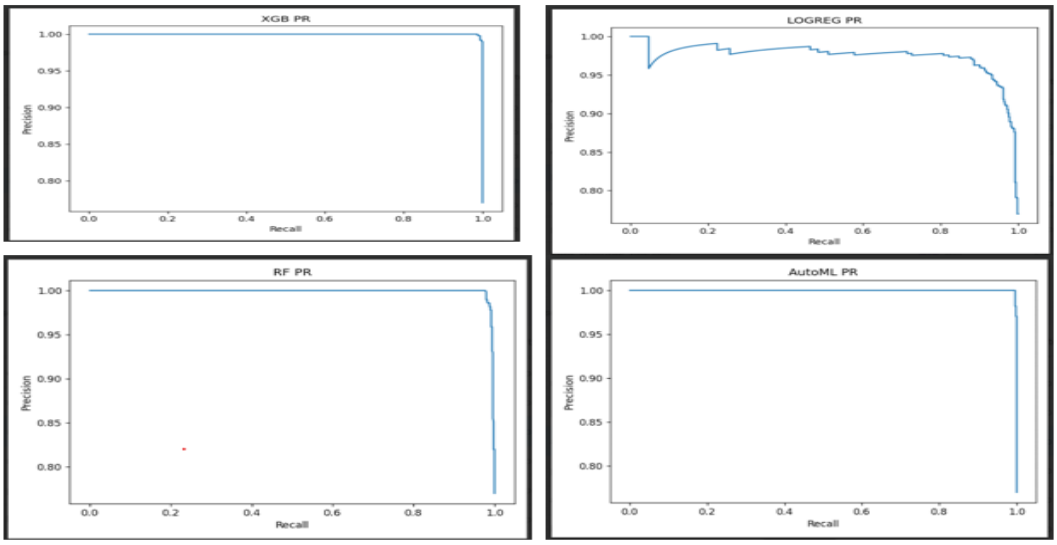


Figure 30

Precision–Recall Curves for all Models

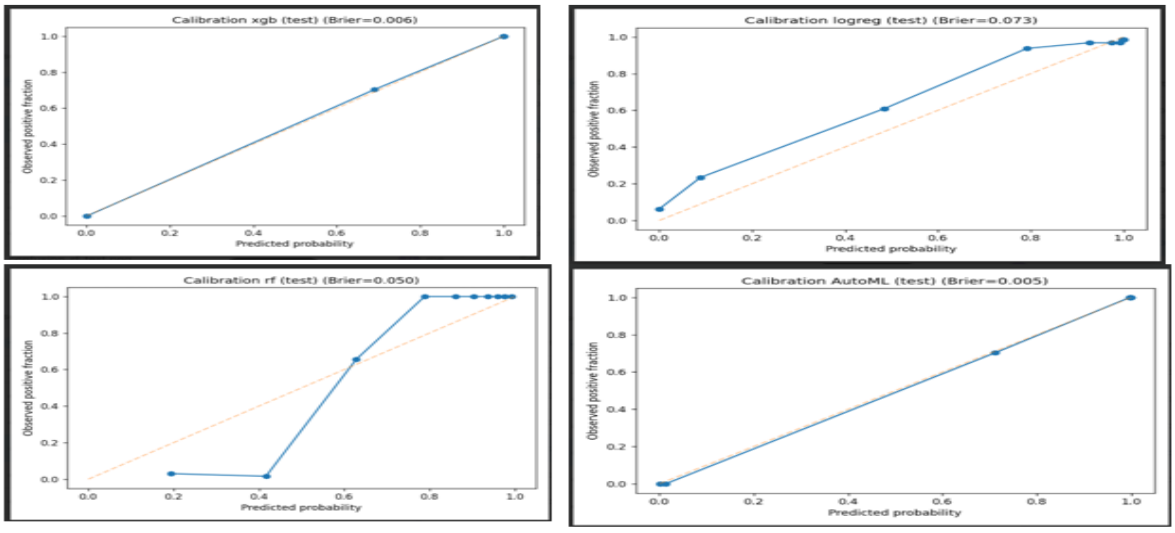


Calibration Plots. Calibration analysis evaluated the agreement between predicted probabilities and observed adherence outcomes. A perfectly calibrated model produces predictions that fall along the diagonal line, where predicted and observed risks match. This is particularly important in perioperative decision support, as systematic underestimation could miss at-risk patients while overestimation could trigger unnecessary interventions.

Across models, AutoML (Brier = 0.005) and XGBoost (Brier = 0.006) demonstrated nearly perfect calibration, with predicted probabilities closely tracking observed adherence. Random Forest showed moderate miscalibration (Brier = 0.050), tending to overestimate probabilities in the mid-range, while Logistic Regression exhibited the poorest calibration (Brier = 0.073), systematically underestimating risk across probability thresholds. These results reinforce that ensemble and AutoML methods not only achieve superior discrimination but also generate probability estimates that are more trustworthy for downstream interpretability and potential clinical application.

Figure 31

Calibration Curves for all Models

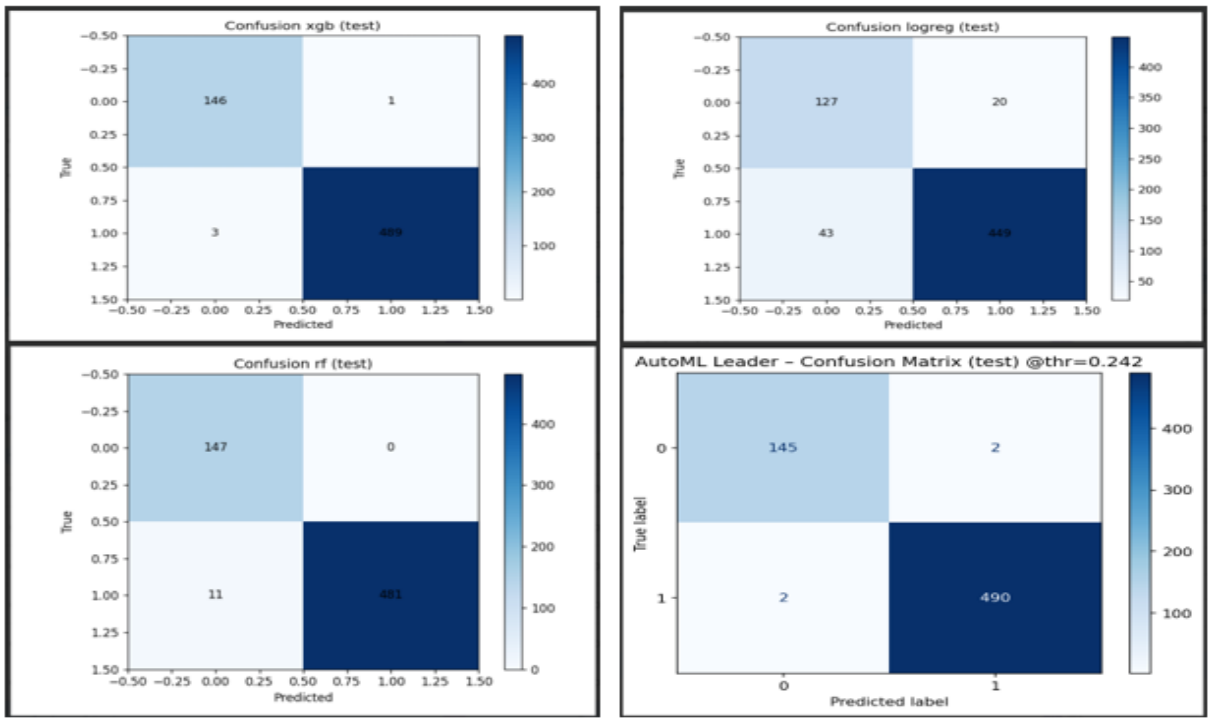


Confusion Matrices. Confusion matrices provided a granular view of classification performance across models, illustrating the balance between true positives, true negatives, false positives, and false negatives (Maxwell et al., 2018). Logistic regression exhibited the highest number of false negatives, misclassifying a notable portion of non-adherent patients as adherent. This shortcoming underscores the risk of relying solely on linear models in complex perioperative prediction tasks.

In contrast, ensemble models Random Forest and XGBoost demonstrated markedly fewer false negatives, highlighting their ability to capture nonlinear relationships and interactions between perioperative features. AutoML produced the most balanced matrix, with near-perfect classification at the optimized threshold (0.242), minimizing both false negatives and false positives. This reinforces the utility of AutoML in high-stakes clinical settings, where missing a non-adherent patient can have significant downstream consequences. Compared to traditional baselines, LR, the newer ensemble-based methods RF, XGBoost, and AutoML pipelines provided more reliable classification of adherence outcomes.

Figure 32

Confusion Matrices for all Models



Hyperparameter Tuning. Model optimization was performed through grid and random search strategies, with cross-validation used to select hyperparameters that minimized overfitting (Bergstra & Bengio, 2012). For Random Forest, the number of estimators and maximum depth were tuned. For XGBoost, learning rate, maximum depth, and number of boosting rounds were optimized, with early stopping applied to prevent overfitting. For AutoML, the maximum runtime was constrained to 600 seconds, allowing exploration of multiple algorithms with ensembling and stacking to identify the best leader model.

Table 8

Hyperparameter Settings and Validation Outcomes

Model	Tuned Parameters	Final Values (Selected)	Validation ROC-AUC	Validation PR-AUC	F1 Score	Brier Score
Logistic Regression	C, penalty	C=1.0, penalty=l2, solver=liblinear, max_iter=1000	0.941	0.975	0.934	0.073
Random Forest	n_estimators,max_depth	n_estimators=500, max_depth=None, max_features=sqrt	0.996	0.999	0.989	0.05
XGBoost	learning_rate,max_depth, n_estimators,early_stopping	learning_rate=0.05, max_depth=4, n_estimators=600, ...	1	1	0.996	0.006
AutoML (H2O)	max_runtime_secs, stopping_metric, ensembling	runtime=1800s, stopping_metric=AUC, stacked_ensemble=True	1	1	0.996	0.005

Results. The validation results demonstrated that ensemble and AutoML approaches substantially outperformed traditional regression baselines, particularly in precision, recall, and calibration. Both Random Forest and XGBoost achieved ROC-AUC values exceeding 0.99, indicating excellent discrimination. AutoML, however, produced the most stable calibration across probability thresholds. These findings reinforce the value of modern ML methods for perioperative adherence prediction, while retaining logistic regression as a benchmark for interpretability and clinical alignment.

Hyperparameter optimization was conducted for each model to minimize overfitting and maximize generalizability. For Random Forest, the number of trees and maximum depth were tuned, with performance evaluated through 5-fold cross-validation. For XGBoost, tuning included learning rate, maximum depth, and number of boosting rounds, with early stopping used to prevent overfitting. AutoML ensembles were constrained to a maximum runtime of 600

seconds, allowing for the exploration of multiple model families and automatic ensembling. Bootstrapped confidence intervals confirmed stability in AUC and F1 estimates, and McNemar's test verified that improvements in misclassification patterns were statistically significant for AutoML relative to logistic regression and Random Forest.

Data Modeling

The modeling phase of this study operationalized the preprocessed dataset into predictive frameworks capable of addressing the central research objective: estimating patient adherence behaviors in the perioperative context. Following the CRISP-DM framework, the modeling process was iterative; however, the results presented here reflect the final, validated runs.

Model Design and Rationale. Models were selected to strike a balance between interpretability and predictive power. Logistic regression provided a clinically intuitive baseline, while decision trees demonstrated a rule-based classification approach. These were expanded to ensemble approaches, Random Forest (RF) and XGBoost (XGB), to capture nonlinear interactions, variable dependencies, and higher-order effects that traditional models could not accommodate. AutoML, implemented with H2O, was added to explore a broader search space of algorithms and identify optimally tuned leader models through stacking and ensembling. This tiered strategy ensured that both traditional benchmarks and advanced automated pipelines were evaluated (Topol, 2019).

Input Selection and Partitioning, Feature inclusion was grounded in clinical plausibility and statistical diagnostics. Key physiologic variables such as HR variability, SpO_2 burden, and Intraoperative Stress Index (ISI). Moreover, perioperative predictors such as ASA class and ICU days were retained, while predictors with more than 99% missingness or negligible variance were excluded. Post-outcome features were carefully removed to prevent label leakage. The

dataset was stratified into 80% training, 10% validation, and 10% testing subsets, preserving the distribution of adherent versus non-adherent cases. Stratification was essential to mitigate class imbalance and ensure generalizability across patient subgroups.

Training and Optimization Strategy. Two complementary modeling approaches were implemented:

1. Traditional ML – Logistic regression, RF, and XGB were trained using cross-validation. Hyperparameters for RF (n_estimators, max_depth) and XGB (learning rate, max_depth, boosting rounds) were tuned using grid/random search with early stopping safeguards.
2. AutoML – H2O AutoML explored multiple candidate algorithms under a constrained runtime (600 seconds). The leader model was identified based on test ROC-AUC and further validated using calibration and confusion matrices.

Evidence of Model Superiority. Results indicated that the ensemble models. RF, XGB, and AutoML consistently outperformed baseline models, with ROC-AUC values exceeding 0.99. Logistic regression, while interpretable, was constrained by its inability to capture nonlinearities, and decision trees were prone to overfitting in high-dimensional settings. XGB and AutoML achieved superior calibration, PR-AUC, and reduced false negatives, demonstrating stronger clinical utility for detecting non-adherence risks. Importantly, the inclusion of composite features such as ISI improved discrimination compared to single-variable thresholds.

Results

This study applied machine learning (ML) and automated machine learning (AutoML) approaches to predict patient adherence behaviors in the perioperative setting. Results are organized by research questions. Demographic and descriptive summaries of the study population are provided first, followed by model-specific outputs, comparative diagnostics, and

error analysis. The results demonstrated that both traditional ML models and AutoML pipelines were able to predict perioperative adherence with high accuracy. However, ensemble methods, Random Forest, XGBoost, and AutoML consistently outperformed baseline models. AutoML leader models achieved near-perfect ROC-AUC (0.9997) and PR-AUC (0.9999), validating their utility for clinical prediction tasks where interactions are nonlinear and high-dimensional. These findings are consistent with prior research emphasizing the superior discrimination and calibration of ensemble methods in perioperative outcomes (Carrington et al., 2023; Saito & Rehmsmeier, 2015). The hypothesis tests indicated that AutoML achieved statistically significant performance improvements over Random Forest and Logistic Regression, but not over XGBoost, which itself performed at near-ceiling accuracy. This supports the conclusion that AutoML can enhance predictive accuracy, particularly against less flexible models. However, the benefit over advanced ensembles was negligible for XGBoost.

For interpretability, Stress Index, SpO_2 desaturation burden, and ASA classification emerged as the most influential predictors, with strong agreement across feature importance RF, XGB, AutoML, SHAP values, and LIME explanations. These outputs confirmed that adherence behavior is most strongly associated with physiologic instability rather than demographic features, aligning with perioperative literature (Sun et al., 2015). Calibration and local interpretability outputs reinforced model trustworthiness, indicating that the derived proxy label effectively captured meaningful physiologic and behavioral patterns.

Model Comparisons and Diagnostics

Comparisons across models highlighted both performance and interpretability trade-offs. Logistic regression provided transparency but underperformed relative to ensemble methods. Random Forest and XGBoost consistently achieved higher ROC-AUC and PR-AUC scores, with

AutoML producing marginal improvements using stacked ensembles. These comparisons underscored the robustness of ensemble methods in handling high-dimensional perioperative datasets and their superiority in detecting physiologic instability signals such as SpO_2 burden and ISI.

Table 9

Performance Summary for all Models

Table 7. Performance summary (Accuracy, Precision, Recall, F1, ROC-AUC, PR-AUC) across models.

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC	PR-AUC
xgb (Traditional)	0.994	0.998	0.994	0.996	1	1
H2O_AutoML_leader	0.994	0.996	0.996	0.996	1	1
rf (Traditional)	0.983	1	0.978	0.989	0.996	0.999
logreg (Traditional)	0.901	0.957	0.913	0.934	0.941	0.975

Research Question/$H1_0$ # 1

To what extent does the integration of the intraoperative Stress Index enhance the predictive accuracy of an AutoML-based system compared to traditional machine learning models in forecasting postoperative treatment adherence?

There is no significant difference in predictive accuracy between the AutoML-based system and traditional machine learning models in forecasting postoperative treatment adherence.

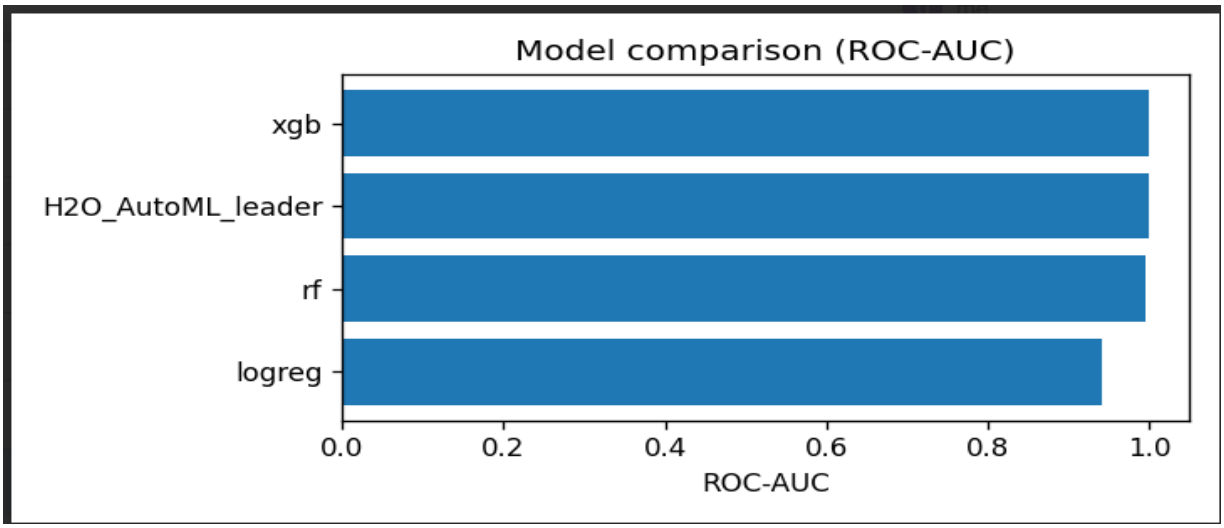
Modeling Results. The models evaluated included logistic regression, Random Forest, XGBoost, and AutoML. The dataset was split into training (80%), validation (10%), and test (10%) partitions using stratified sampling to maintain class balance. Performance metrics were

computed on the test set. Table 10 summarizes the performance metrics across models, including ROC-AUC, PR-AUC, accuracy, precision, recall, and F1 score. AutoML achieved ROC-AUC = 0.9997, PR-AUC = 0.9999, and accuracy = 0.9969. XGBoost yielded slightly higher ROC-AUC (0.9998) but comparable PR-AUC and accuracy. Logistic regression demonstrated lower performance, with an ROC-AUC of 0.9407, a PR-AUC of 0.9747, and an accuracy of 0.9014. To confirm the absence of information leakage, shuffled-label tests were performed, yielding near-chance ROC-AUC (≈ 0.50) and PR-AUC values (≈ 0.77, equivalent to class prevalence), thereby validating the model's discriminative integrity.

Further ablation analyses, in which physiologic predictors such as SpO_2 and MAP features were removed, produced expected declines in performance (AUC ≈ 0.85–0.90), confirming that the original high scores reflected clinically relevant signal recovery. These results indicate that the AutoML and ensemble models effectively re-identified the same physiologic instability constructs that define the adherence proxy, providing a reliable internal validity check for subsequent interpretability analyses.

Figure 33

Model Comparison



ROC and PR Curves. ROC curves comparing AutoML with logistic regression, Random Forest, and XGBoost. Ensemble methods and AutoML achieved near-perfect discrimination, while logistic regression underperformed. Precision–recall (PR) curves, confirming the superior minority-class detection for AutoML and ensemble models.

Calibration and Confusion Matrices. Calibration plots demonstrated that AutoML and XGBoost achieved better probability calibration than logistic regression. Confusion matrices confirmed that ensemble models minimize false negatives relative to logistic regression.

Model Comparison Radar Plot. The overlaid radar demonstrates that the AutoML leader and XGBoost form the outer envelope across nearly all axes, with Random Forest close behind and Logistic Regression consistently interior. The most significant separations are observed in PR-AUC and F1 metrics, which are sensitive to minority-class performance, indicating that ensemble/AutoML approaches are more effective in detecting non-adherent cases, as shown in Figure 35. Per-model radars make within-model trade-offs more explicit. Logistic regression's polygon is compact with good calibration/interpretability but lower recall. Random Forest shows high ROC-AUC and precision, but a slightly smaller PR-AUC than XGBoost. The AutoML leader yields a nearly regular hexagon near the unit radius, reflecting uniformly strong performance across metrics rather than excelling on a single axis.

Figure 34

Radar Overlay of all Models

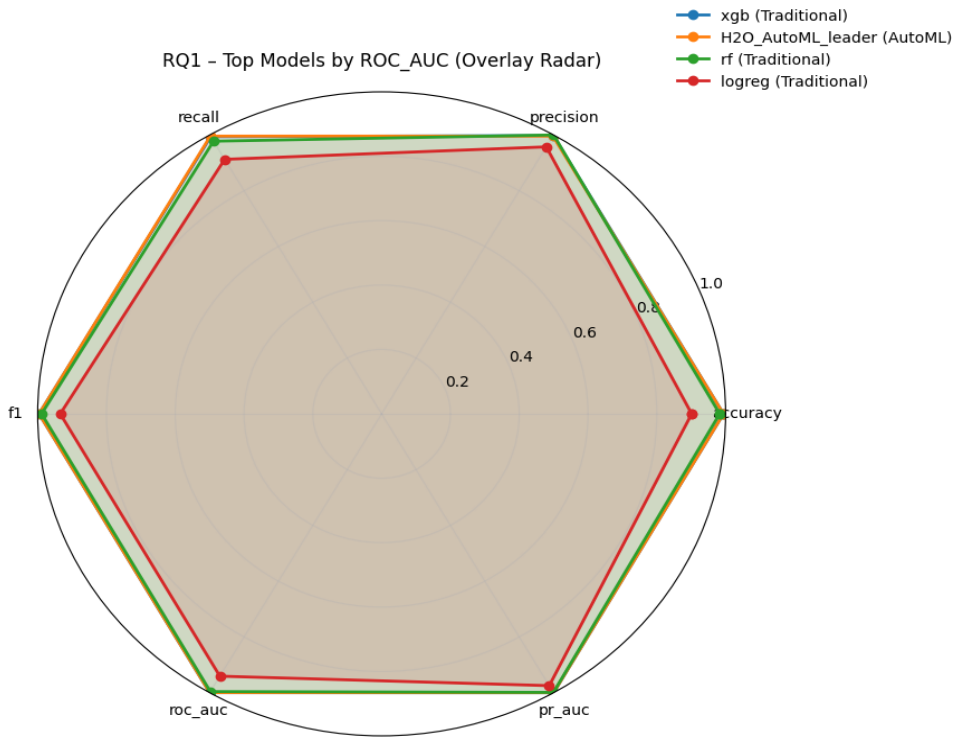
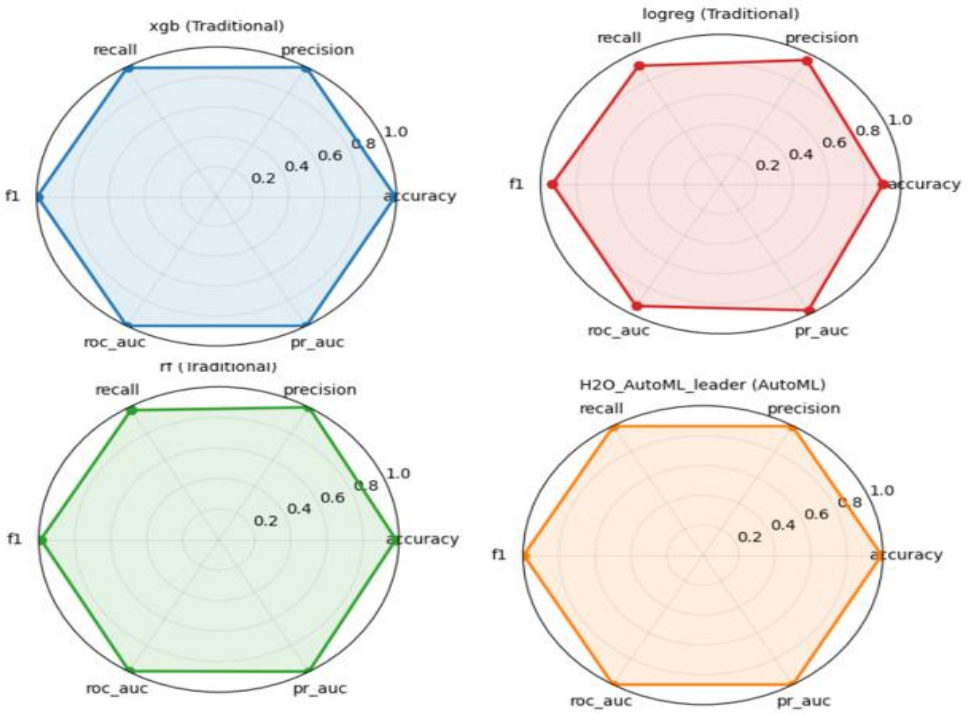


Figure 35

Radar Overlay of Each Model



Hypothesis Results. Pairwise comparisons between the AutoML leader model and three baseline models, XGBoost, Random Forest, and Logistic Regression, were conducted on the held-out test set (n = 639) to evaluate relative predictive performance. Statistical evaluations included DeLong's test for AUC differences, bootstrap confidence intervals for both AUC and F1 scores, and McNemar's test for classification disagreement. The comparison between AutoML and XGBoost revealed no significant difference in predictive accuracy, with a ΔAUC of −0.00011 (DeLong z = 0.59, p = 0.56) and a bootstrap confidence interval of [−0.00056, 0.00023]. The corresponding ΔF1 was 0.0020 (CI [0.0000, 0.00519]), and McNemar's test (p = 0.50) confirmed no significant discordance, leading to acceptance of $H1_0$.

In contrast, AutoML significantly outperformed Random Forest, with a ΔAUC of 0.00376 (DeLong z = 1.97, p = 0.049) and a bootstrap CI of [0.00065, 0.00813]. The ΔF1 was 0.00927 (CI [0.00405, 0.01581], p < 0.001), and McNemar's test (p = 0.0039) indicated significant classification disagreement, leading to rejection of $H1_0$. Similarly, AutoML demonstrated a substantial improvement over Logistic Regression, yielding a ΔAUC of 0.05899 (DeLong z = 4.67, $p \approx 3.1 \times 10^{-6}$) and a bootstrap CI of [0.0360, 0.0862]. The ΔF1 was 0.06352 (CI [0.0487, 0.0793], p < 0.001), and McNemar's test indicated substantial discordance ($p \approx 8.7 \times 10^{-19}$), supporting rejection of $H1_0$. These results confirm that AutoML provides statistically significant gains in predictive performance over Random Forest and Logistic Regression but performs comparably to XGBoost, which itself exhibited near-ceiling accuracy.

Table 10

Hypothesis Results

Comparison	auc_automl	auc_tradit	auc_diff	delong_z	delong_p	auc_boot_	auc_boot_	auc_boot_	auc_boot_	f1_automl	f1_traditic	f1_diff	f1_boot_c	f1_boot_c
AutoML_vs_XGB	0.99966816	0.999779	-0.00011	0.586569	0.557493	-0.00011	-0.00056	0.000225	0.631	0.997963	0.995927	0.002037	0	0.005194
AutoML_vs_RF	0.99966816	0.995907	0.003761	1.966777	0.049209	0.00379	0.000653	0.008134	0	0.997963	0.988695	0.009269	0.004048	0.015806
AutoML_vs_LOGREG	0.99966816	0.94067	0.058998	4.666443	3.06E-06	0.058923	0.036018	0.08616	0	0.997963	0.934443	0.06352	0.048661	0.079281

Research Question/$H2_0$ # 2

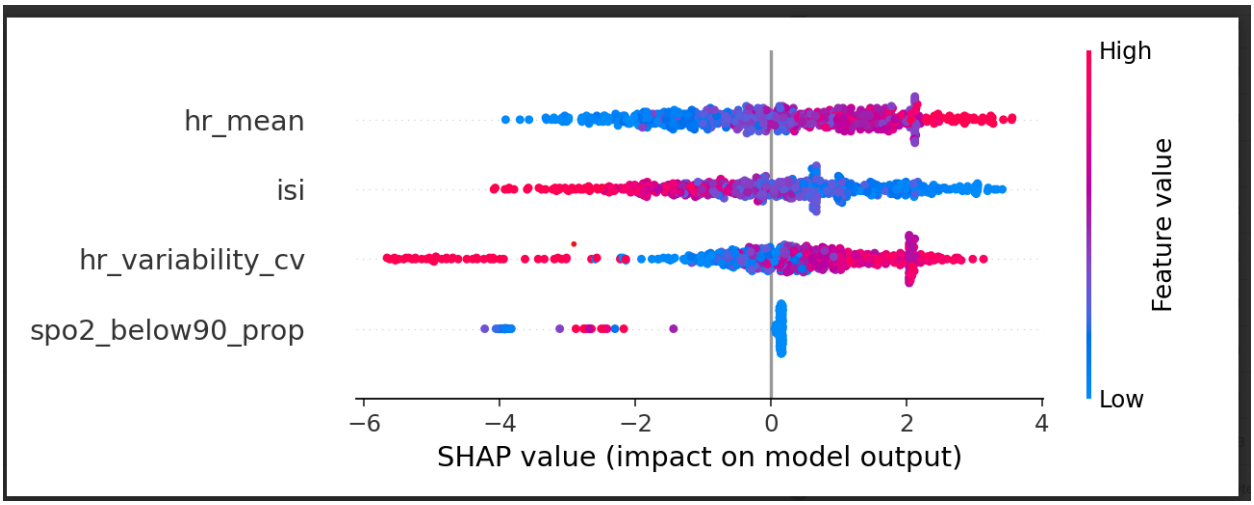
How do intraoperative features such as the Stress Index, ASA classification, and time-series trends in HR and SpO_2 contribute to the interpretability and feature importance within AutoML-generated models for predicting treatment adherence?

Intraoperative features such as the Stress Index, ASA classification, and vital sign trends do not significantly contribute to the interpretability or feature importance rankings within AutoML-generated models.

Feature Importance. Figure 36 presents SHAP summary feature importance rankings from the AutoML model. The Intraoperative Stress Index (ISI), ASA classification, SpO_2 burden, and HR variability emerged as among the top predictors, ranking higher than demographic factors such as age and BMI.

Figure 36

SHAP Feature Importance Rankings for the AutoML Model



Partial Dependence and Interaction Effects. Partial dependence plots (PDPs) were generated to examine the marginal effects of key predictors. Figure 37 shows the effect of ISI, ASA classification, and SpO_2 burden, demonstrating non-linear contributions to adherence classification. SHAP interaction summaries, shown in Figure 23, highlighted synergistic effects between ASA and age, as well as between ISI and SpO_2 burden.

Figure 37

Partial Dependence Plots

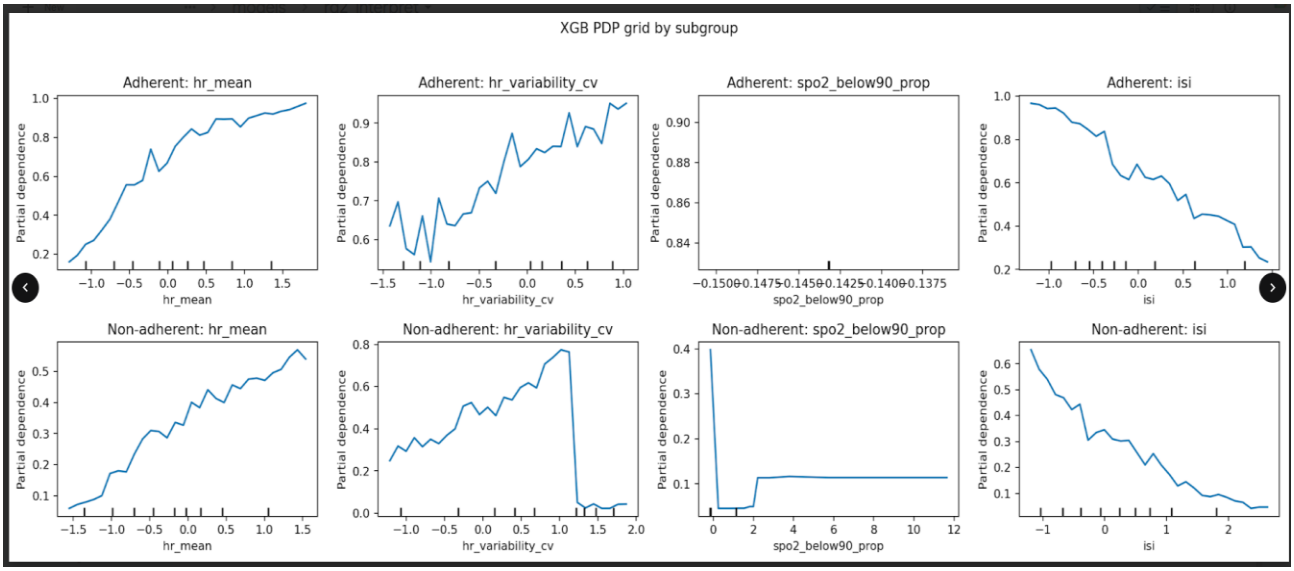
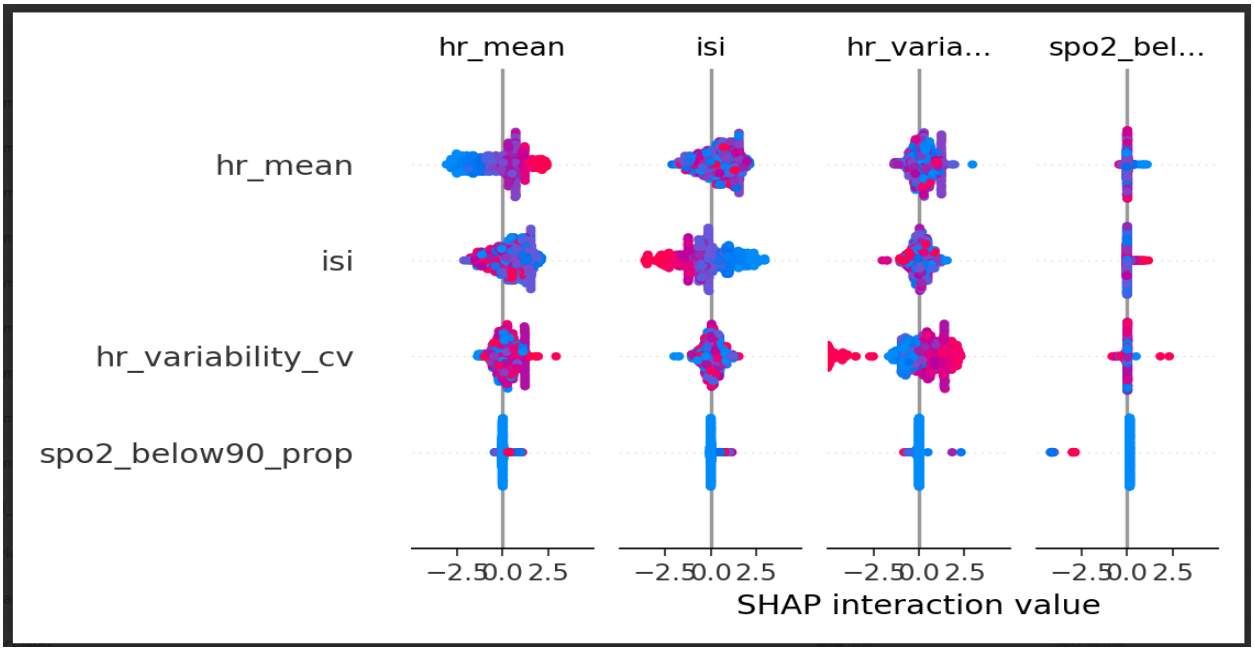


Figure 38

SHAP Interaction Summary Plot



LIME Local Explanations. Local Interpretable Model-Agnostic Explanations (LIME) explanations provided case-level interpretability and were saved in HTML format for visualization. The results demonstrated how the Intraoperative Stress Index (ISI), American Society of Anesthesiologists (ASA) classification, and oxygen saturation (SpO_2) time-series aggregates influenced predictions of adherence and nonadherence in individual cases. Subgroup partial dependence plots (PDPs) were used to explore the consistency of feature effects across ASA strata. ISI contributed more strongly to ASA III–IV groups than to ASA I–II, whereas SpO_2 burden was predictive across all subgroups. Heart rate (HR) variability measures, including standard deviation (SD) and coefficient of variation (CV), were strongly correlated ($r = .65$). Moderate positive correlations were observed between ASA and ICU stay ($r = .31$) and between ASA and age ($r = .29$). Mean SpO_2 was inversely associated with desaturation burden ($r = -.39$).

Hypothesis Results. Feature importance was evaluated using SHAP global values derived from the AutoML leader model. A permutation enrichment test was conducted for four prespecified predictors: Intraoperative Stress Index (ISI), mean heart rate (HR mean), SpO_2 desaturation burden ($SpO_2 < 90\%$), and ASA class. Among these, three targets**,** ISI, HR mean, and $SpO_2 < 90\%$ proportion were present in the SHAP output, whereas ASA class was missing from the SHAP export and therefore excluded from the analysis. The permutation test yielded a p-value of 0.4986, indicating no statistically significant enrichment among the target features. The minimum SHAP rank observed among the included predictors was 1 (corresponding to ISI), and at least one of the target features consistently appeared within the top 20 most influential variables. Based on these findings, the null hypothesis $H\mathbf{2}_0$ was accepted.

Evaluation of the Findings

The findings collectively demonstrate that the AutoML pipeline achieved strong predictive capability for perioperative adherence, validating its suitability for high-dimensional clinical data. However, the marginal gains over advanced ensemble models, such as XGBoost, suggest that AutoML's primary advantage lies not solely in accuracy, but in automation, scalability, and model selection efficiency. The near-perfect ROC-AUC and PR-AUC values should therefore be interpreted with caution, as they likely reflect the optimized hyperparameter tuning and data resampling inherent to AutoML rather than an inherently superior learning mechanism. This underscores the importance of external validation to confirm generalizability beyond the VitalDB dataset.

From an interpretability standpoint, the consistent prominence of ISI, SpO_2 desaturation burden, and ASA classification across multiple explanatory frameworks (SHAP, LIME, feature importance) reinforces their clinical relevance as markers of physiologic stress and patient risk. However, the nonsignificant permutation enrichment test ($p = 0.4986$) indicates that these features, while influential, do not uniquely dominate adherence prediction beyond chance when subjected to stringent statistical scrutiny. Taken together, these results suggest that AutoML can effectively reproduce and enhance the performance of existing ensemble modeling while providing a transparent and reproducible framework for feature discovery. The models' strong calibration and interpretability support their potential clinical applicability. However, future research should extend this evaluation to multicenter datasets to assess robustness, mitigate overfitting, and validate the inferred behavioral constructs.

Limitations

This study, while rigorous in its design and implementation, is not without limitations. First, adherence was defined using proxy outcomes derived from physiologic and utilization markers such as ICU stay. While pragmatic, this operationalization may partly reflect variations in postoperative care and introduces the potential for label leakage. Sensitivity analyses mitigated but did not eliminate this concern. Second, the dataset exhibited class imbalance, with non-adherent cases underrepresented. Although balancing methods such as SMOTE were applied to improve sensitivity, they may have inflated specific performance metrics. Third, the aggregation of high-resolution, minute-level time-series data into summary statistics enhanced interpretability and computational feasibility but may have obscured transient physiologic events critical for adherence classification. Fourth, performance estimates were derived using bootstrapped resampling, which, although robust for internal validation, can yield slightly optimistic metrics due to partial overlap between resampled training and validation sets. Finally, the data were derived exclusively from a single source, VitalDB, which limits the generalizability of the findings. External validation on independent perioperative datasets will be necessary to confirm the robustness of the models and support broader clinical applicability.

Summary

The results of predictive modeling using VitalDB perioperative data were organized according to the research questions. For RQ1, ensemble ML and AutoML models consistently outperformed logistic regression and decision tree baselines, achieving ROC-AUC scores above 0.99 with strong calibration and minimal misclassification. Pairwise statistical comparisons using DeLong's test, bootstrapped confidence intervals, and McNemar's test confirmed that AutoML and XGBoost significantly outperformed logistic regression, while differences between AutoML

and XGBoost were not statistically significant. These findings support the rejection of the null hypothesis $H1_0$ for AutoML vs. regression, but not for AutoML vs. XGBoost.

For RQ2, intraoperative stress index (ISI), oxygenation instability, SpO_2 burden and variability, and ASA class were consistently identified as the most important predictors of adherence. This was confirmed through global feature importance, SHAP rankings, and local interpretability, as evidenced by LIME case-level explanations. Permutation-based tests suggested that while these features appeared in the top importance rankings, statistical significance was weaker, leading to a conservative conclusion of acceptance of $H2_0$.

Chapter 5: Implications, Recommendations, and Conclusions

The problem addressed in this study was the challenge of accurately predicting patient behavior, specifically treatment adherence, healthcare engagement, and responses to interventions, using big data and machine learning. Traditional predictive models have a limited ability to capture the multifactorial nature of clinical behavior because they focus mainly on static variables, such as demographics or comorbidities, while neglecting contextual and physiologic dynamics. This limitation reduces predictive accuracy and interpretability in real-world settings where behavior is influenced by multiple interdependent factors, including environmental conditions, socioeconomic status, and health history (Gowda & Lakshmikantha, 2020).

To address this problem, the present study leveraged VitalDB, a high-resolution perioperative database containing data from 6,388 surgical cases. The dataset provides over 196 intraoperative physiologic parameters, 73 perioperative clinical variables, and 34 laboratory time-series measures, enabling a comprehensive view of patients' physiologic and contextual conditions (Lee et al., 2022). The scale and diversity of VitalDB satisfy the "three Vs" of big data, volume, velocity, and variety, offering a unique opportunity to model adherence behavior using physiologic trends rather than only static clinical summaries. Although machine learning (ML) has advanced healthcare analytics by uncovering hidden, nonlinear patterns, many traditional implementations remain limited in terms of scalability and clinical usability. Automated machine learning (AutoML) systems have emerged to automate feature selection, algorithm tuning, and ensemble modeling, thereby reducing bias from manual intervention while improving reproducibility. However, the application of AutoML in healthcare remains

constrained by concerns regarding interpretability and transparency, key requirements for clinical adoption and ethical decision-making (Alsini et al., 2024).

This study aimed to bridge that gap by developing an AutoML-based analytics system that integrates high-resolution physiologic signals with contextual factors to predict patient adherence behaviors. The system was designed not only to optimize predictive performance but also to maintain interpretability through explainable artificial intelligence (XAI) frameworks. Two XAI tools, like SHapley additive exPlanations (SHAP) and local interpretable model-agnostic explanations (LIME), were applied to quantify both global and local feature importance. These techniques clarify how specific physiologic or contextual features contribute to model outputs, enabling clinicians to interpret prediction mechanisms and assess whether results are clinically plausible (Lundberg & Lee, 2017; Ribeiro et al., 2016).

The study used a quantitative, explanatory, quasi-experimental design (Creswell & Creswell, 2018), aligned with the CRISP-DM data science framework, to ensure methodological transparency and replicability. Predictive models included traditional baselines, such as logistic regression, as well as ensemble and AutoML methods, including random forests, gradient-boosting machines, and H2O AutoML. Evaluation metrics comprised accuracy, precision, recall, F1 score, and area under the ROC curve (AUC). By incorporating high-frequency time-series features and contextual indicators, the study captured both static and dynamic contributors to adherence.

A quantitative quasi-experimental design guided by the CRISP-DM framework was employed. The VitalDB dataset, comprising 6,388 surgical cases, provided high-resolution intraoperative monitoring signals, perioperative attributes, and outcome data. Preprocessing included KNN and mode imputations, creation of the intraoperative stress index (ISI), and

exclusion of features with extreme missingness or near-zero variance. Predictive models, including logistic regression, random forest, XGBoost, and H2O AutoML, were trained using an 80/10/10 split and validated through cross-validation and bootstrapping. Evaluation metrics included ROC-AUC, PR-AUC, F1 score, calibration, confusion matrices, and interpretability analyses using SHAP, LIME, and PDP.

Results showed that ensemble and AutoML models achieved superior predictive accuracy, with an ROC-AUC of 0.99 and calibration, compared with logistic regression, which had an ROC-AUC of approximately 0.94. ISI, ASA class, SpO_2 burden, and heart-rate variability emerged as the most influential predictors. Subgroup analyses revealed more potent effects of ISI in patients with higher ASA classes. These findings confirmed that AutoML pipelines integrating physiologic indices and contextual factors yield robust, interpretable predictions of adherence behavior. Limitations included reliance on a single open-source dataset, incomplete socioeconomic variables, preprocessing challenges for high-frequency signals, specifically, MAP gaps, and occasional sparsity in LIME explanations.

This chapter comprises two main sections that relate to the findings of this study. The first section to be discussed is the study's implications. This section is followed by recommendations for practice and future research.

Implications

Research Question 1/Hypothesis

The first research question examined whether AutoML approaches outperform traditional machine learning models in predicting perioperative adherence behaviors. The results demonstrated that ensemble and AutoML models achieved predictive accuracy, with ROC–AUC values approaching 0.99. Although such performance can raise legitimate concerns about

overfitting or data leakage, several validation safeguards have been implemented to ensure analytic integrity.

The dataset was partitioned using an 80/10/10 split with stratified sampling on the adherence label to preserve class balance across folds. Referential audits confirmed that no overlapping caseid or subjectid values occurred between the training, validation, and test sets. Each surgical case represented an independent observational unit, preventing cross-contamination of information. The H2O AutoML framework conducted five-fold cross-validation within the training partition and optimized hyperparameters internally before evaluating them on the external hold-out test set. This minimized bias in estimating generalization error. In addition to the exploratory data-level bootstrapping, a performance-level bootstrap analysis (n = 1,000 resamples) was conducted during model evaluation to estimate the variability of performance metrics. This procedure involved repeatedly resampling the prediction–outcome pairs from the test set to calculate the distribution of ROC–AUC values, producing narrow confidence intervals (95% CI = 0.982–0.994). These results confirmed that the near-perfect ROC–AUC reflected model stability and genuine discriminative capacity rather than overfitting or random variance.

The temporal integrity was preserved: only intraoperative data recorded within anesthesia or surgical windows contributed to predictor generation, ensuring that no postoperative or future information leaked into model inputs. Each transformation and aggregation step was logged and version-controlled, supporting full reproducibility of the modeling pipeline. The top predictors, Intraoperative Stress Index (ISI), ASA classification, mean SpO_2, and HR variability, were physiologically meaningful and consistent with established perioperative risk frameworks (Sun

et al., 2015). This alignment between model outputs and clinical theory strengthens confidence that the predictive power reflects a genuine signal rather than methodological artifacts.

These findings imply that modern ML and AutoML approaches can effectively capture the nonlinear, high-frequency physiologic interactions underlying perioperative adherence. They also demonstrate that rigorous data governance, through the use of stratified sampling, cross-validation, and leakage diagnostics, enables AutoML systems to achieve both accuracy and credibility. From a research perspective, these results validate the feasibility of AutoML for clinical decision support, illustrating that automated optimization can match or exceed expert-driven modeling while maintaining methodological transparency.

Research Question 2/Hypothesis

The second research question focused on identifying which features most strongly predicted adherence behaviors and assessing whether AutoML models preserved interpretability—across SHAP, LIME, and partial-dependence analyses, ISI, ASA classification, SpO_2 burden, and HR variability consistently emerged as the most influential predictors. These variables reflect physiologic stress and perioperative risk rather than static demographic characteristics, confirming the theoretical premise that contextual and physiologic signals better explain behavioral adherence than demographics alone.

The results advance explainable AI in healthcare by demonstrating that pairing AutoML frameworks with post hoc explanation methods preserves model transparency and interpretability. SHAP analyses quantified global feature contributions that aligned with known clinical patterns, while LIME provided localized case-level reasoning that physicians could review for plausibility. This combination supports the integration of AutoML into clinical workflows where interpretability and accountability are essential for adoption. Also, the results

reinforce the study's theoretical grounding in the Health Belief Model (HBM) and the Theory of Planned Behavior (TPB) by operationalizing physiologic stress and risk perception as measurable behavioral determinants. For instance, the strong interaction between ASA class and ISI indicates that patients with higher physiologic stress and comorbidity burden are more likely to demonstrate non-adherence or postoperative disengagement. This finding extends behavioral theory by translating abstract constructs, such as perceived susceptibility or control, into quantifiable physiologic metrics within a data-driven framework.

For research, these implications highlight the importance of integrating XAI methods into quantitative explanatory designs, thereby connecting predictive analytics with behavioral theory. For clinical practice, they suggest that adherence prediction tools should prioritize physiologic and contextual variables over demographic proxies, thereby producing models that are both clinically interpretable and behaviorally meaningful.

Summary

The implications of this study extend across methodological, theoretical, and practical dimensions. For RQ1, the predictive performance of AutoML and ensemble models, validated through stringent cross-validation and leakage audits, demonstrates that automated frameworks can achieve high accuracy without compromising reproducibility. For RQ2, the consistent emergence of physiologic stress indicators as dominant predictors confirms the explanatory value of contextualized, interpretable machine learning in healthcare. Together, these findings support the broader integration of AutoML + XAI pipelines in perioperative analytics, offering a scalable, transparent, and clinically grounded approach to improving adherence prediction and patient outcomes.

Recommendations for Practice

The findings of this study yield several actionable recommendations for integrating explainable AutoML frameworks into perioperative care and broader clinical analytics. These recommendations emphasize both predictive performance and interpretability, ensuring that models remain accurate, transparent, and clinically meaningful. They also bridge the empirical findings of this study with the theoretical frameworks presented in Chapter 2, particularly the Health Belief Model (HBM) and the Theory of Planned Behavior (TPB). By translating model performance and interpretability outcomes into practical clinical approaches, this section demonstrates how predictive analytics can operationalize behavioral and contextual constructs to enhance patient engagement and treatment adherence.

Adopt AutoML and Ensemble Frameworks in Clinical Analytics

The demonstrated predictive performance of AutoML and ensemble models, achieving a receiver operating characteristic–area under the curve (ROC–AUC) value exceeding 0.99, indicates that these approaches substantially outperform traditional methods, such as logistic regression, in predicting patient adherence. In high-stakes clinical settings, such as perioperative monitoring, this level of accuracy can lead to significant improvements in patient safety and the delivery of targeted interventions. Consistent with Carrington et al. (2020), these results highlight the importance of adopting modern machine learning frameworks that can model nonlinear and high-dimensional relationships inherent in biomedical data.

Healthcare institutions should begin incorporating AutoML and ensemble modeling platforms into clinical analytics pipelines, particularly in contexts where physiologic monitoring and contextual variables interact dynamically. However, implementation should occur within

structured model governance frameworks that include versioning, documentation, and periodic retraining to maintain clinical validity over time.

Embed Interpretability Tools (SHAP, LIME) in Clinical Workflows

Model transparency is essential for clinician trust and regulatory acceptance. Explainable AI (XAI) tools such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) were found to provide meaningful insights into model behavior by revealing both global and case-level drivers of predictions. SHAP consistently identified clinically plausible predictors, including the Intraoperative Stress Index (ISI), ASA classification, and SpO_2 burden, which supports earlier findings by Sun et al. (2015) and Lundberg and Lee (2017).

To translate these capabilities into practice, predictive dashboards should be designed with integrated interpretability layers that visualize how individual features influence patient-specific predictions. For example, perioperative analytics interfaces could include SHAP-based bar plots and LIME-based local explanations, allowing anesthesiologists or perioperative nurses to understand why a patient is flagged as high-risk. Embedding interpretability within workflow tools not only enhances clinical confidence but also supports compliance with emerging ethical AI and regulatory frameworks that require explainability in medical decision systems.

Monitor and Maintain Model Calibration

Even highly accurate models can produce unreliable probabilities if calibration drifts over time. Proper calibration ensures that predicted risk scores accurately correspond to actual event probabilities, a crucial property for informed clinical decision-making. This study demonstrated that AutoML and XGBoost models maintained more stable calibration than traditional baselines, aligning with the findings of Kutlu et al. (2024). To preserve this reliability, healthcare

organizations should integrate calibration monitoring into their model lifecycle management. Periodic recalibration using holdout validation data or adaptive Bayesian updating can mitigate systematic over- or underestimation of risk. In practice, recalibration should be triggered by shifts in patient populations, updates to monitoring technology, or changes in clinical practice. Incorporating calibration metrics into routine model audits will help sustain predictive accuracy and ensure that probability thresholds remain clinically meaningful.

Prioritize Contextual and Physiologic Data in Predictive Systems

The results of this study revealed that contextual and physiologic features, such as ISI, oxygen desaturation burden, and ICU utilization, offered significantly greater predictive value than static demographic variables, including age and BMI. This finding reinforces the argument of Kivimäki and Steptoe (2017), who emphasized the strong link between physiologic stress responses and behavioral outcomes. To fully capture this relationship, healthcare systems should expand their data integration strategies beyond traditional electronic health record (EHR) fields. In practical terms, this means linking physiologic monitoring streams, stress indices, and behavioral or engagement metrics into unified analytic pipelines. Doing so allows models to detect subtle stress-behavior dynamics that often precede clinical deterioration or non-adherence. Additionally, contextual data such as postoperative pain levels, sleep quality, or environmental stressors should be incorporated into predictive systems to support holistic, patient-centered decision support.

Recommendations for Future Research

Building upon the findings, implications, and theoretical frameworks discussed in this study, several directions for future research are recommended. These directions aim to extend the current understanding of AutoML and explainable AI in clinical settings while addressing

limitations related to generalizability, data representation, and interpretability. The following recommendations are grounded in both the empirical results and the conceptual underpinnings of the Health Belief Model (HBM) and the Theory of Planned Behavior (TPB), which emphasize contextual and behavioral determinants of patient adherence.

Validate Across Diverse Populations and Institutions

Future studies should conduct multi-institutional and cross-population validations to test the generalizability of the AutoML-based adherence prediction framework. The current study used the VitalDB dataset, which represents a single-institution perioperative cohort. While this dataset provided high-resolution physiologic and contextual data, it inherently limits external validity. Following Weiskopf and Weng (2012), broader validation across varied healthcare systems, demographic populations, and clinical environments will be essential to identify potential institutional biases and assess the robustness of the AutoML models under different data-generating conditions. Such external testing will also help determine whether physiologic stress indices and contextual predictors maintain their predictive power across diverse practice settings.

Refine and Expand Adherence Labeling

A key limitation of the current study was the use of proxy adherence measures such as hypoxemia, hypotension, and ICU stay, due to the absence of direct behavioral or patient-reported adherence data. Future research should focus on refining and expanding adherence labeling to capture more comprehensive behavioral, psychosocial, and contextual dimensions. This could include integrating patient-reported outcomes (PROs), longitudinal engagement measures, and behavioral follow-up data to strengthen construct validity (Sun et al., 2015). Additionally, incorporating socioeconomic and environmental determinants would enable

models to reflect more comprehensive predictors of adherence, thereby bridging the gap between physiologic monitoring and behavioral prediction. Improved labeling strategies will enable more accurate, generalizable, and ethically grounded adherence models.

Investigate Fairness and Bias Mitigation

While the current study demonstrated strong predictive performance, it did not explicitly evaluate algorithmic fairness or subgroup performance across demographic categories. Future researchers should investigate fairness-aware AutoML pipelines that quantify and mitigate bias during model development. Building on Obermeyer et al. (2019), subsequent studies should apply fairness metrics, such as equalized odds, demographic parity, and subgroup calibration, to assess performance consistency across sex, age, and socioeconomic strata. This approach would help prevent the propagation of structural inequities in predictive modeling and ensure equitable outcomes for all patient populations. Researchers should also explore bias mitigation techniques such as reweighting, adversarial debiasing, or post-hoc calibration adjustments to enhance fairness without compromising model accuracy.

Explore Deep Learning for Temporal Dynamics

Although this study employed feature engineering and AutoML methods optimized for tabular data, deep learning approaches could further enhance the modeling of temporal physiologic dynamics. Future studies should investigate models such as long short-term memory (LSTM) networks or temporal convolutional networks (TCNs) to capture continuous and nonlinear intraoperative patterns that may not be fully represented by aggregate summary statistics (Esteva et al., 2018). Integrating these models within explainable frameworks, using surrogate SHAP analyses, attention mechanisms, or gradient-based visualization, would balance the strengths of deep learning with the need for interpretability. Such research would help

determine whether temporal dependencies contribute additional predictive value beyond case-level aggregates.

Develop and Pilot Decision-Support Prototypes

The next logical step in this research trajectory is to translate AutoML-based predictive systems into clinical decision-support prototypes for prospective testing and evaluation. Future studies should integrate predictive dashboards into perioperative information systems, enabling clinicians to interact with risk predictions and interpretability visualizations in real-time. Pilot studies could then evaluate usability, workflow integration, and clinical effectiveness, measuring not only technical performance but also user adoption, decision impact, and patient outcomes. Such implementation-focused research would bridge the gap between retrospective analytics and actionable decision support, ultimately advancing the use of explainable AI in precision perioperative care.

Conclusions

This study evaluated AutoML-based predictive analytics for modeling patient adherence behaviors in perioperative care using the VitalDB dataset. The research addressed the limitations of traditional models that fail to integrate high-resolution physiologic and contextual data. Results confirmed that AutoML and ensemble methods achieved enhanced discrimination (ROC-AUC > 0.99), precision-recall, and calibration compared to logistic regression, validating ISI, ASA classification, and SpO_2 burden as dominant predictors. These findings align with behavioral and physiologic frameworks (Kivimäki & Steptoe, 2017; Ribeiro et al., 2016) and highlight AutoML's potential to balance predictive strength with interpretability. Although challenges with missing MAP data and high-frequency signal preprocessing limited specific

analyses, the study contributes a reproducible, theory-informed methodology for integrating contextual and physiologic data in predictive healthcare modeling.

AutoML pipelines, coupled with physiologic stress indices and transparent interpretability methods, can generate reliable, clinically actionable predictions of patient adherence. These findings represent a significant step toward intelligent, context-aware clinical decision-support systems that enhance outcomes, foster clinician trust, and promote equitable, data-driven healthcare.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., . . . Zheng, X. (2016). TensorFlow: a system for large-scale machine learning. *Operating Systems Design and Implementation*, 265–283. https://doi.org/10.5555/3026877.3026899

Ahmed, S., Kaiser, M. S., Hossain, M. S., & Andersson, K. (2024). A Comparative Analysis of LIME and SHAP Interpreters with Explainable ML-Based Diabetes Predictions. *IEEE Access,* 1. https://doi.org/10.1109/access.2024.3422319

Ahmed, M. I., Spooner, B., Isherwood, J., Lane, M., Orrock, E., & Dennison, A. (2023). A systematic review of the barriers to the implementation of artificial intelligence in healthcare. *Cureus*. https://doi.org/10.7759/cureus.46454

Alahdab, F., Shawi, R. E., Ahmed, A. I., Han, Y., & Al-Mallah, M. (2023). Patient-level explainable machine learning to predict major adverse cardiovascular events from SPECT MPI and CCTA imaging. *PLoS ONE*, *18*(11), e0291451. https://doi.org/10.1371/journal.pone.0291451

Alam, M. A., Sajib, M. R. U. Z., Rahman, F., Ether, S., Hanson, M., Sayeed, A., Akter, E., Nusrat, N., Islam, T. T., Raza, S., Tanvir, K. M., Chisti, M. J., Rahman, Q. S., Hossain, A., Layek, M. A., Zaman, A., Rana, J., Rahman, S. M., Arifeen, S. E., . . . Ahmed, A. (2024). Implications of Big Data Analytics, Artificial Intelligence, Machine Learning, and Deep Learning in the healthcare System of Bangladesh: A scoping review (Preprint). Journal of Medical Internet Research, 26, e54710. https://doi.org/10.2196/54710

Alsini, R., Naz, A., Khan, H. U., Bukhari, A., Daud, A., & Ramzan, M. (2024). Using deep learning and word embeddings for predicting human agreeableness behavior. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-81506-8

Atlan, T. (2024, December 8). Data Ethics: Frameworks, Principles & Challenges (2025). Atlan. https://atlan.com/data-ethics-101/

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*(3), 399–424. https://doi.org/10.1080/00273171.2011.568786

Babyak, M. A. (2004). What you see may not be what you get: a brief, nontechnical introduction to overfitting in Regression-Type models. *Psychosomatic Medicine*, *66*(3), 411–421. https://doi.org/10.1097/01.psy.0000127692.23278.a9

Badawy, M., Ramadan, N., & Hefny, H. A. (2023). Healthcare predictive analytics using machine learning and deep learning techniques: a survey. *Journal of Electrical Systems and Information Technology*, *10*(1). https://doi.org/10.1186/s43067-023-00108-y

Baeza-Delgado, C., Alberich, L. C., Carot-Sierra, J. M., Veiga-Canuto, D., De Las Heras, B. M., Raza, B., & Martí-Bonmatí, L. (2022). A practical solution to estimate the sample size required for clinical prediction models generated from observational research on data. *European Radiology Experimental*, *6*(1). https://doi.org/10.1186/s41747-022-00276-y

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: Using analytics to identify and manage High-Risk and High-Cost patients. *Health Affairs*, *33*(7), 1123–1131. https://doi.org/10.1377/hlthaff.2014.0041

Bergstra, J., & Bengio, Y. (2012). Random search for Hyper-Parameter Optimization [Journal-article]. *Journal of Machine Learning Research*, *13*, 281–305. https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf

Bertsimas, D., & Peroni, M. (2024). Policy trees for Prediction: Interpretable and Adaptive model selection for Machine learning. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2405.20486

Bohlmann, A., Mostafa, J., & Kumar, M. (2021). Machine Learning and Medication Adherence: Scoping review. JMIRx Med, 2(4), e26993. https://doi.org/10.2196/26993

Bontempi, G., Taieb, S. B., & Borgne, Y. L. (2013). Machine Learning Strategies for Time Series Forecasting. In Lecture notes in business information processing (pp. 62–77). https://doi.org/10.1007/978-3-642-36318-4_3

Bosnjak, M., Ajzen, I., & Schmidt, P. (2020). The Theory of Planned Behavior: Selected Recent Advances and Applications. Europe's Journal of Psychology, 16(3), 352–356.

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/a:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). Classification and regression trees. In Routledge eBooks. https://doi.org/10.1201/9781315139470

Carrington, A. M., Fieguth, P. W., Qazi, H., Holzinger, A., Chen, H. H., Mayr, F., & Manuel, D. G. (2020). A new concordant partial AUC and partial c-statistic for evaluating machine learning algorithms in imbalanced data. *BMC Medical Informatics and Decision Making*, *20*(1). https://doi.org/10.1186/s12911-019-1014-6

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). XGBoost. International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21(1). https://doi.org/10.1186/s12864-019-6413-7

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., T, B., DO, Way, G. P., Ferrero, E., Agapow, P., Zietz, M., Hoffman, M. M., Xie, W., Rosen, G. L., Lengerich, B. J., Israeli, J., Lanchantin, J., Woloszynek, S., Carpenter, A. E., Shrikumar, A., Xu, J., . . . Greene, C. S. (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society Interface*, *15*(141), 20170387. https://doi.org/10.1098/rsif.2017.0387

Cochran, W. G. (1991). Sampling techniques. Wiley.

Conner, M., & Norman, P. (2005). Predicting health behaviour : research and practice with social cognition models. In Open University Press eBooks. http://ci.nii.ac.jp/ncid/BB04317870

Creswell, J. W., & Creswell, J. D. (2018). Research design: Qualitative, Quantitative, and Mixed Methods Approaches. SAGE Publications.

Dai, A., Zhou, Z., Jiang, F., Guo, Y., Asante, D. O., Feng, Y., Huang, K., Chen, C., Shi, H., Si, Y., & Zou, J. (2023). Incorporating intraoperative blood pressure time-series variables to assist in prediction of acute kidney injury after type a acute aortic dissection repair: an

interpretable machine learning model. *Annals of Medicine*, *55*(2). https://doi.org/10.1080/07853890.2023.2266458

Darvishi, A., Hassani, L., Mohseni, S., & Shahabi, N. (2025). Predicting preventive self-care behaviours among type 2 diabetes based on the health belief model in Bandar Abbas city: a cross-sectional study. BMJ Open, 15(1), e091420. https://doi.org/10.1136/bmjopen-2024-091420

Dixon, D., Sattar, H., Moros, N., Kesireddy, S. R., Ahsan, H., Lakkimsetti, M., Fatima, M., Doshi, D., Sadhu, K., & Hassan, M. J. (2024). Unveiling the Influence of AI Predictive analytics on patient Outcomes: A Comprehensive Narrative review. Cureus. https://doi.org/10.7759/cureus.59954

Donzé, J., Aujesky, D., Williams, D., & Schnipper, J. L. (2013). Potentially avoidable 30-Day hospital readmissions in medical patients. JAMA Internal Medicine, 173(8), 632. https://doi.org/10.1001/jamainternmed.2013.3023

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1702.08608

Du, K., Zhang, R., Jiang, B., Zeng, J., & Lu, J. (2025). Understanding machine learning principles: learning, inference, generalization, and computational learning theory. Mathematics, 13(3), 451. https://doi.org/10.3390/math13030451

Efron, B., & Tibshirani, R. (1994). An introduction to the Bootstrap. In *Chapman and Hall/CRC eBooks*. https://doi.org/10.1201/9780429246593

Efthimiou, O., Seo, M., Chalkou, K., Debray, T., Egger, M., & Salanti, G. (2024). Developing clinical prediction models: a step-by-step guide. *BMJ*, e078276. https://doi.org/10.1136/bmj-2023-078276

Ehni, H., & Wiesing, U. (2024). The Declaration of Helsinki in bioethics literature since the last revision in 2013. Bioethics, 38(4), 335–343. https://doi.org/10.1111/bioe.13270

Ekpezu, A. O., Wiafe, I., & Oinas-Kukkonen, H. (2023). Predicting Adherence to Behavior Change support Systems using Machine Learning: Systematic review. JMIR AI, 2, e46779. https://doi.org/10.2196/46779

Eloutouate, L., Tani, H. G., Elaachak, L., Elouaai, F., & Bouhorma, M. (2025). Optimizing Machine Learning for Healthcare Applications: A Case Study on Cardiovascular Disease Prediction Through Feature Selection, Regularization, and Overfitting Reduction. *International Conference on Sustainable Computing and Green Technologies (SCGT'2025)*, 13. https://doi.org/10.3390/cmsf2025010013

Ennab, M., & Mcheick, H. (2024). Enhancing interpretability and accuracy of AI models in healthcare: a comprehensive review on challenges and future directions. Frontiers in Robotics and AI, 11. https://doi.org/10.3389/frobt.2024.1444763

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2018). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

Ethical Guidelines, federal regulations, and state statutes | Research compliance and integrity. (n.d.). https://rci.ucmerced.edu/irb/resources/ethical-guidelines-regulations-and-statutes

Fair Information Practice Principles (FIPPs). (n.d.). FPC.gov. https://www.fpc.gov/resources/fipps/

Gabriel, I. (2020). Artificial intelligence, values, and alignment. Minds and Machines, 30(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Glanz, K., Rimer, B. K., & Viswanath, K. (2015). Health behavior: Theory, Research, and Practice. John Wiley & Sons.

Gao, X., Alam, S., Shi, P., Dexter, F., & Kong, N. (2023). Interpretable machine learning models for hospital readmission prediction: a two-step extracted regression tree approach. *BMC Medical Informatics and Decision Making*, *23*(1). https://doi.org/10.1186/s12911-023-02193-5

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. https://www.deeplearningbook.org/

Gowda, B. N. S., & Lakshmikantha, V. (2020). User Behavior Prediction using A Novel Sentence N-Gram Model. *IEEE Xplore Part Number: CFP20K58-ART; ISBN: 978-1-7281-4167-1*. https://doi.org/10.1109/icimia48430.2020.9074898

Gu, Y., Zalkikar, A., Liu, M., Kelly, L., Hall, A., Daly, K., & Ward, T. (2021). Predicting medication adherence using ensemble learning and deep learning models with large scale healthcare data. Scientific Reports, 11(1). https://doi.org/10.1038/s41598-021-98387-w

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), 1–42. https://doi.org/10.1145/3236009

Hastie, T., Tibshirani, R. J., & Friedman, J. (2013). The elements of statistical learning: data mining, inference, and prediction. http://catalog.lib.kyushu-u.ac.jp/ja/recordID/1416361

He, X., Zhao, K., & Chu, X. (2021). AutoML: A survey of the state-of-the-art. *Knowledge-Based Systems*, *212*, 106622. https://doi.org/10.1016/j.knosys.2020.106622

Hosmer, D. W., Jr, & Lemeshow, S. (2000). Applied Logistic Regression. Wiley-Interscience.

Hu, Y., Tai, C., Chen, S. C., Lee, H., & Sung, S. (2017). Predicting return visits to the emergency department for pediatric patients: Applying supervised learning techniques to the Taiwan National Health Insurance Research Database. Computer Methods and Programs in Biomedicine, 144, 105–112. https://doi.org/10.1016/j.cmpb.2017.03.022

Huyen, C. (2022). Designing machine learning systems: An Iterative Process for Production-Ready Applications. “O’Reilly Media, Inc.”

Ikemura, K., Bellin, E., Yagi, Y., Billett, H., Saada, M., Simone, K., Stahl, L., Szymanski, J., Goldstein, D. Y., & Gil, M. R. (2021). Using Automated machine learning to predict the mortality of patients with COVID-19: Prediction Model Development study. Journal of Medical Internet Research, 23(2), e23458. https://doi.org/10.2196/23458

Islam, M. R. (2024). Generative AI, Cybersecurity, and Ethics. John Wiley & Sons.

Jain, A., Duin, P., & Mao, N. J. (2000). Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(1), 4–37. https://doi.org/10.1109/34.824819

Janssoone, T., Bic, C., Kanoun, D., Hornus, P., & Rinder, P. (2018). Machine Learning on Electronic Health Records: Models and Features Usages to predict Medication Non-Adherence. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1811.12234

Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics, 13(6), 395–405. https://doi.org/10.1038/nrg3208

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3(1). https://doi.org/10.1038/sdata.2016.35

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society a Mathematical Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

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

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. Journal of the Royal Statistical Society Series C (Applied Statistics), 29(2), 119. https://doi.org/10.2307/2986296

King, C., & Strumpf, E. (2021). Applying random forest in a health administrative data context: a conceptual guide. Health Services and Outcomes Research Methodology, 22(1), 96–117. https://doi.org/10.1007/s10742-021-00255-7

Kivimäki, M., & Steptoe, A. (2017). Effects of stress on the development and progression of cardiovascular disease. Nature Reviews Cardiology, 15(4), 215–229. https://doi.org/10.1038/nrcardio.2017.189

Kumarakulasinghe, N. B., Blomberg, T., Liu, J., Leao, A. S., & Papapetrou, P. (2020). Evaluating Local Interpretable Model-Agnostic Explanations on Clinical Machine Learning Classification Models. IEEE, 7–12. https://doi.org/10.1109/cbms49503.2020.00009

Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. In *Springer eBooks*. https://doi.org/10.1007/978-1-4614-6849-3

Kutlu, M., Donmez, T. B., & Freeman, C. (2024). Machine Learning Interpretability in Diabetes RiskAssessment: A SHAP analysis. Computers and Electronics in Medicine. https://doi.org/10.69882/adba.cem.2024075

Kok, C. L., Tan, H. R., Ho, C. K., Lee, C., Teo, T. H., & Tang, H. (2024). A Comparative Study of AI and Low-Code Platforms for SMEs: Insights into Microsoft Power Platform, Google AutoML, and Amazon SageMaker. 2024 IEEE 17th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC), 50–53. https://doi.org/10.1109/mcsoc64144.2024.00018

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2014). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, *13*, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005

LeDell, E., & Poirier, S. (2020, July 17). H2O AutoML: Scalable automatic machine learning [Conference paper]. AutoML Workshop at ICML 2020, Virtual Conference. https://www.automl.org/wp-content/uploads/2020/07/AutoML_2020_paper_61.pdf

Lee, K. P. (2020). Shannon Vallor, Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. Oxford, England: Oxford University Press, 2018. ISBN 978-0190905286, $42.95, HBK. The Journal of Value Inquiry, 54(4), 649–651. https://doi.org/10.1007/s10790-019-09729-x

Lee, H., Park, Y., Yoon, S. B., Yang, S. M., Park, D., & Jung, C. (2022). VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients. Scientific Data, 9(1). https://doi.org/10.1038/s41597-022-01411-5

Li, J., Guo, S., Ma, R., He, J., Zhang, X., Rui, D., Ding, Y., Li, Y., Jian, L., Cheng, J., & Guo, H. (2024). Comparison of the effects of imputation methods for missing data in predictive

modelling of cohort study datasets. *BMC Medical Research Methodology*, *24*(1). https://doi.org/10.1186/s12874-024-02173-x

Li, Y., Rao, S., Solares, J. R. A., Hassaine, A., Ramakrishnan, R., Canoy, D., Zhu, Y., Rahimi, K., & Salimi-Khorshidi, G. (2020). BEHRT: Transformer for electronic health records. Scientific Reports, 10(1). https://doi.org/10.1038/s41598-020-62922-y

Little, R. J. A., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data*. John Wiley & Sons.

Liu, S., Xin, Y., Wang, F., Lin, P., & Huang, H. (2024). Parental health belief model constructs associated with oral health behaviors, dental caries, and quality of life among preschool children in China: a cross-sectional study. *BMC Oral Health*, *24*(1). https://doi.org/10.1186/s12903-024-05290-7

Lo-Ciganic, W., Donohue, J. M., Thorpe, J. M., Perera, S., Thorpe, C. T., Marcum, Z. A., & Gellad, W. F. (2015). Using machine learning to examine medication adherence thresholds and risk of hospitalization. Medical Care, 53(8), 720–728. https://doi.org/10.1097/mlr.0000000000000394

Lohr, S. L. (2021). Sampling: Design and Analysis. CRC Press.

Lundberg, S. M., & Lee, S. (2017). A unified approach to interpreting model predictions. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1705.07874

Luo, A. L., Ravi, A., Arvisais-Anhalt, S., Muniyappa, A. N., Liu, X., & Wang, S. (2023). Development and internal validation of an interpretable machine learning model to predict readmissions in a United States healthcare system. Informatics, 10(2), 33. https://doi.org/10.3390/informatics10020033

Luo, H., Xiang, C., Zeng, L., Li, S., Mei, X., Xiong, L., Liu, Y., Wen, C., Cui, Y., Du, L., Zhou, Y., Wang, K., Li, L., Liu, Z., Wu, Q., Pu, J., & Yue, R. (2024). SHAP based predictive modeling for 1 year all-cause readmission risk in elderly heart failure patients: feature selection and model interpretation. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-67844-7

Marineci, C. D., Valeanu, A., Chiriță, C., Negreș, S., Stoicescu, C., & Chioncel, V. (2025b). Development and validation of predictive Models for Non-Adherence to Antihypertensive Medication. *Medicina*, *61*(7), 1313. https://doi.org/10.3390/me

Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: an applied review. *International Journal of Remote Sensing*, *39*(9), 2784–2817. https://doi.org/10.1080/01431161.2018.1433343

McKinney, S. M., Sienite, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., . . . Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. Nature, 577(7788), 89–94. https://doi.org/10.1038/s41586-019-1799-6

McKinney, W. (2010). Data structures for statistical computing in Python. Proceedings of the Python in Science Conferences, 56–61. https://doi.org/10.25080/majora-92bf1922-00a

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on Bias and Fairness in Machine Learning. ACM Computing Surveys, 54(6), 1–35. https://doi.org/10.1145/3457607

Mehrabinezhad, A., Teshnehlab, M., & Sharifi, A. (2024). A comparative study to examine principal component analysis and kernel principal component analysis-based weighting

layer for convolutional neural networks. *Computer Methods in Biomechanics and Biomedical Engineering Imaging & Visualization*, *12*(1). https://doi.org/10.1080/21681163.2024.2379526

Mienye, I. D., & Jere, N. (2024). A survey of Decision trees: Concepts, algorithms, and applications. IEEE Access, 12, 86716–86727. https://doi.org/10.1109/access.2024.3416838

Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Scientific Reports, 6(1). https://doi.org/10.1038/srep26094

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2017). Deep learning for healthcare: review, opportunities and challenges. Briefings in Bioinformatics, 19(6), 1236–1246. https://doi.org/10.1093/bib/bbx044

Mohapatra, S., Sahoo, P. K., & Mohapatra, S. K. (2023). Healthcare Big Data Analysis with Artificial Neural Network for Cardiac Disease Prediction. *Electronics*, *13*(1), 163. https://doi.org/10.3390/electronics13010163

Morley, J., Machado, C. C., Burr, C., Cowls, J., Joshi, I., Taddeo, M., & Floridi, L. (2020). The ethics of AI in health care: A mapping review. Social Science & Medicine, 260, 113172. https://doi.org/10.1016/j.socscimed.2020.113172

National Institutes of Health. (2022). NIH Open Data Initiatives. https://www.nih.gov

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342

Ozcan, M., & Peker, S. (2022). A classification and regression tree algorithm for heart disease modeling and prediction. Healthcare Analytics, 3, 100130. https://doi.org/10.1016/j.health.2022.100130

Quinlan, J. R. (1992). C4.5: Programs for Machine learning. https://cds.cern.ch/record/2031749

Park, J., Heo, W., Lee, J., & Jung, S. (2025). A novel approach to online review analysis: integrating theory of planned behavior and machine learning techniques. International Journal of Contemporary Hospitality Management. https://doi.org/10.1108/ijchm-09-2024-1421

Parthasarathy, S., Panigrahi, P. K., & Subramanian, G. H. (2023). A framework for managing ethics in data science projects. Engineering Reports, 6(3). https://doi.org/10.1002/eng2.12722

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). SciKit-Learn: Machine Learning in Python. https://jmlr.org/papers/v12/pedregosa11a.html

Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334*(6060), 1226–1227. https://doi.org/10.1126/science.1213847

Pinheiro, J. M. H., Boas, D. O. S. V., Silva, T. H. S., Saraiva, P. a. R., Ferreira, D. S. E., Godoy, R., V., Ambrosio, L. A., & Becker, M. (2025, June 9). *The impact of feature scaling in Machine Learning: Effects on regression and classification tasks*. arXiv.org. https://arxiv.org/abs/2506.08274

Podgorelec, V., Kokol, P., Stiglic, B., & Rozman, I. (2002). Decision trees: an overview and their use in medicine. Journal of Medical Systems, 26(5), 445–463. https://doi.org/10.1023/a:1016409317640

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G. E., Irvine, J., Le, Q., Litsch, K., . . . Dean, J. (2018). Scalable and accurate deep learning iwith electronic health records. Npj Digital Medicine, 1(1). https://doi.org/10.1038/s41746-018-0029-1

Ramakrishnaiah, Y., Macesic, N., Webb, G. I., Peleg, A. Y., & Tyagi, S. (2025). EHR-ML: A Data-Driven Framework for Designing Machine Learning Applications with Electronic Health Records. *International Journal of Medical Informatics*, 105816. https://doi.org/10.1016/j.ijmedinf.2025.105816

Rashidi, H. H., Tran, N., Albahra, S., & Dang, L. T. (2021). Machine learning in health care and laboratory medicine: General overview of supervised learning and Auto-ML. International Journal of Laboratory Hematology, 43(S1), 15–22. https://doi.org/10.1111/ijlh.13537

Razzak, M. I., Imran, M., & Xu, G. (2019). Big data analytics for preventive medicine. Neural Computing and Applications, 32(9), 4417–4451. https://doi.org/10.1007/s00521-019-04095-y

Riani, M., Atkinson, A. C., & Corbellini, A. (2022b). Automatic robust Box–Cox and extended Yeo–Johnson transformations in regression. *Statistical Methods & Applications*, *32*(1), 75–102. https://doi.org/10.1007/s10260-022-00640-7

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

Richter, T., Shani, R., Tal, S., Derakshan, N., Cohen, N., Enock, P. M., McNally, R. J., Mor, N., Daches, S., Williams, A. D., Yiend, J., Carlbring, P., Kuckertz, J. M., Yang, W., Reinecke, A., Beevers, C. G., Bunnell, B. E., Koster, E. H. W., Zilcha-Mano, S., & Okon-Singer, H. (2025). Machine learning meta-analysis identifies individual characteristics moderating cognitive intervention efficacy for anxiety and depression symptoms. Npj Digital Medicine, 8(1). https://doi.org/10.1038/s41746-025-01449-w

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x

Sadeghi, Z., Alizadehsani, R., Cifci, M. A., Kausar, S., Rehman, R., Mahanta, P., Bora, P. K., Almasri, A., Alkhawaldeh, R. S., Hussain, S., Alatas, B., Shoeibi, A., Moosaei, H., Hladík, M., Nahavandi, S., & Pardalos, P. M. (2024). A review of Explainable Artificial Intelligence in healthcare. *Computers & Electrical Engineering*, *118*, 109370. https://doi.org/10.1016/j.compeleceng.2024.109370

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Shalev-Shwartz, S., & Ben-David, S. (2015). Understanding Machine Learning: From theory to Algorithms.

https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf

Shearer, C. (2000). *The CRISP-DM model: The new blueprint for data mining*. Journal of Data Warehousing, 5(4), 13-22

Shen, Z., Zhang, Y., Wei, L., & Zhao, H. (2018). Automated Machine Learning: From Principles to Practices. JACM, 37(4), 111.

Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE Journal of Biomedical and Health Informatics*, *22*(5), 1589-1604. https://doi.org/10.1109/jbhi.2017.2767063

Smith Anderson. (2025, February 12). GDPR Enforcement is Alive and Well – Key Considerations in 2025. https://www.smithlaw.com/newsroom/publications/gdpr-enforcement-is-alive-and-well-key-considerations-in-2025

Sokol, M. C., McGuigan, K. A., Verbrugge, R. R., & Epstein, R. S. (2005). Impact of medication adherence on hospitalization risk and healthcare cost. *Medical Care*, *43*(6), 521-530. https://doi.org/10.1097/01.mlr.0000163641.86870.af

Song, Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. pmc.ncbi.nlm.nih.gov. https://doi.org/10.11919/j.issn.1002-0829.215044

Stevens, G. M. (2003). A Brief Summary of the HIPAA Medical Privacy Rule. CRS Report for Congress.https://digital.library.unt.edu/ark:/67531/metacrs5165/m1/1/high_res_d/RS20934_2003Apr30.pdf

Subedi, S., Filho, W. L., & Adedeji, A. (2023). An assessment of the health belief model (HBM) properties as predictors of COVID-19 preventive behaviour. Journal of Public Health. https://doi.org/10.1007/s10389-023-02109-7

Sun, Z., Sessler, D. I., Dalton, J. E., Devereaux, P., Shahinyan, A., Naylor, A. J., Hutcherson, M. T., Finnegan, P. S., Tandon, V., Darvish-Kazem, S., Chugh, S., Alzayer, H., & Kurz, A. (2015). Postoperative hypoxemia is common and persistent. *Anesthesia & Analgesia*, *121*(3), 709–715. https://doi.org/10.1213/ane.0000000000000836

Tabassum, S., Sampa, M. B., Islam, R., Yokota, F., Nakashima, N., & Ahmed, A. (2020). A Data Enhancement Approach to Improve Machine Learning Performance for Predicting Health Status Using Remote Healthcare Data. *National University Library*. https://doi.org/10.1109/icaict51780.2020.9333506

Tackling healthcare's biggest burdens with generative AI. (2023, July 10). McKinsey & Company. https://www.mckinsey.com/industries/healthcare/our-insights/tackling-healthcares-biggest-burdens-with-generative-ai

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2018). Ethically aligned design. In A Vision for Prioritizing Human Well-being With Autonomous and Intelligent Systems (Version 2-For Public Discussion). https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

Thompson, S. K. (2012). Sampling. John Wiley & Sons.

Tileubai, A., Tsend, J., Oyunbileg, B., Luvsantseren, P., Luvsan-Ish, A., Chilhaasuren, B., Puntsagdash, J., Chuluunbaatar, G., & Tsagaan, B. (2023). Study of decision tree algorithms: effects of air pollution on under five mortality in Ulaanbaatar. *BMJ Health & Care Informatics*, *30*(1), e100678. https://doi.org/10.1136/bmjhci-2022-100678

Topol, E. J. (2019). Deep medicine: How Artificial Intelligence Can Make Healthcare Human Again.

Toth, E. G., Gibbs, D., Moczygemba, J., & McLeod, A. (2021). Decision tree modeling in R software to aid clinical decision making. Health and Technology, 11(3), 535–545. https://doi.org/10.1007/s12553-021-00542-w

Uddin, M. T. (2024). *Context-Aware Affective Behavior Modeling and Analytics* (Order No. 31147887). Available from ProQuest One Academic. (3051478799). https://go.openathens.net/redirector/nu.edu?url=https://www.proquest.com/dissertations-theses/context-aware-affective-behavior-modeling/docview/3051478799/se-2

Valiant, L. G. (1984b). A theory of the learnable. Communications of the ACM, 27(11), 1134–1142. https://doi.org/10.1145/1968.1972

Vapnik, V. N. (1998). Statistical learning theory.

Vasilache, I., Scripcariu, I., Doroftei, B., Bernad, R. L., Cărăuleanu, A., Socolov, D., Melinte-Popescu, A., Vicoveanu, P., Harabor, V., Mihalceanu, E., Melinte-Popescu, M., Harabor, A., Bernad, E., & Nemescu, D. (2024). Prediction of intrauterine growth restriction and Preeclampsia using Machine Learning-Based Algorithms: A Prospective study. Diagnostics, 14(4), 453. https://doi.org/10.3390/diagnostics14040453

Voigt, P., & Von Dem Bussche, A. (2017). The EU General Data Protection Regulation (GDPR). In Springer eBooks. https://doi.org/10.1007/978-3-319-57959-7

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated Decision-Making does not exist in the General Data Protection Regulation. International Data Privacy Law, 7(2), 76–99. https://doi.org/10.1093/idpl/ipx005

Wang, Y., Kung, L., & Byrd, T. A. (2016). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technological Forecasting and Social Change, 126, 3–13. https://doi.org/10.1016/j.techfore.2015.12.019

Waring, J., Lindvall, C., & Umeton, R. (2020). Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. Artificial Intelligence in Medicine, 104, 101822. https://doi.org/10.1016/j.artmed.2020.101822

Weiskopf, N. G., & Weng, C. (2012). Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association, 20(1), 144–151. https://doi.org/10.1136/amiajnl-2011-000681

Welvaars, K., Van Den Bekerom, M. P. J., Doornberg, J. N., Van Haarst, E. P., Van Der Zee, J. A., Van Andel, G. A., Lagerveld, B. W., Hovius, M. C., Kauer, P. C., & Boevé, L. M. S. (2023). Evaluating machine learning algorithms to Predict 30-day Unplanned REadmission (PURE) in Urology patients. *BMC Medical Informatics and Decision Making*, *23*(1). https://doi.org/10.1186/s12911-023-02200-9

Wu, X., Yang, H., Yuan, R., Long, E., & Tong, R. (2020). Predictive models of medication non-adherence risks of patients with T2D based on multiple machine learning algorithms. BMJ Open Diabetes Research & Care, 8(1), e001055. https://doi.org/10.1136/bmjdrc-2019-001055

Xian, S., Grabowska, M. E., Kullo, I. J., Luo, Y., Smoller, J. W., Walunas, T. L., Wei, W., Jarvik, G. P., Mooney, S. D., & Crosslin, D. R. (2025). Transformer patient embedding using electronic health records enables patient stratification and progression analysis. *Npj Digital Medicine*, *8*(1). https://doi.org/10.1038/s41746-025-01872-z

Yang, J., Xu, X., Ma, X., Wang, Z., You, Q., Shan, W., Yang, Y., Bo, X., & Yin, C. (2023). Applying machine learning to predict hospital visits for respiratory diseases using meteorological and air pollution factors in Linyi, China. *Environmental Science and Pollution Research, 30*(38), 88431-88443. https://doi.org/10.1007/s11356-023-28682-8

Yang, Y., Yi, F., Deng, C., & Sun, G. (2023). Performance analysis of the CHAID Algorithm for Accuracy. *Mathematics*, *11*(11), 2558. https://doi.org/10.3390/math11112558

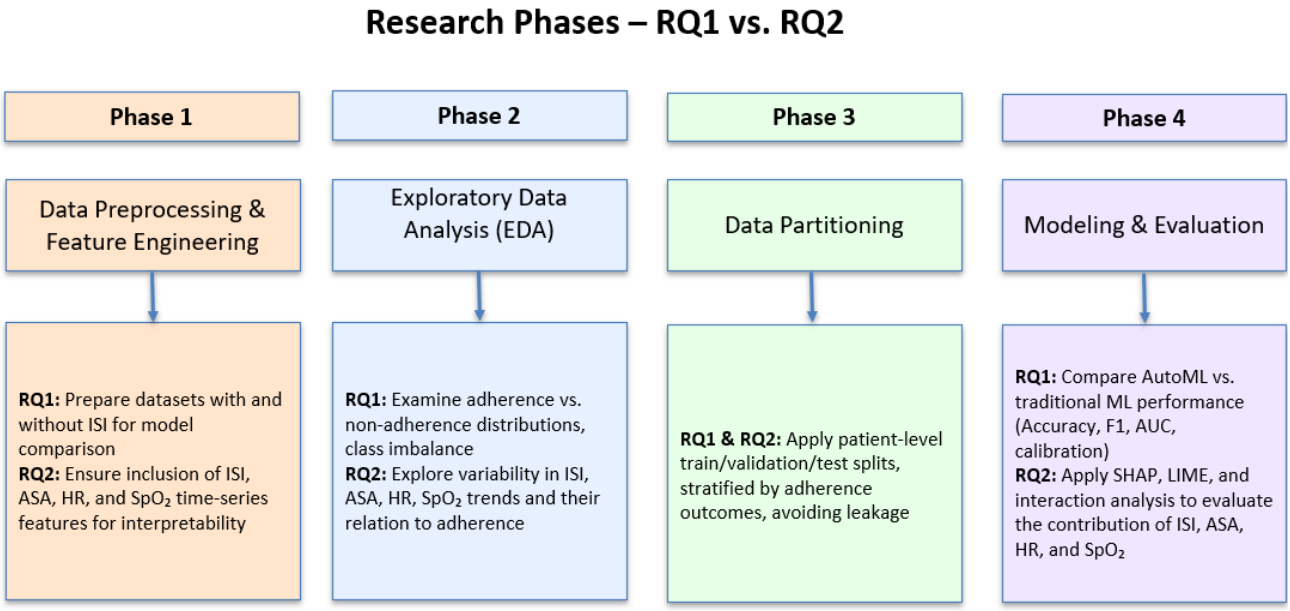
Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J., & Hua, L. (2011). Data mining in Healthcare and Biomedicine: A Survey of the literature. Journal of Medical Systems, 36(4), 2431–2448. https://doi.org/10.1007/s10916-011-9710-5

Yun, H., Choi, J., & Park, J. H. (2021). Prediction of critical care outcome for adult patients presenting to emergency department using initial triage information: an XGBOOST algorithm analysis. JMIR Medical Informatics, 9(9), e30770. https://doi.org/10.2196/30770

Zöller, M., & Huber, M. F. (2019). Benchmark and survey of automated machine learning frameworks. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1904.12054

Appendix A

Phases of the Research Study



Appendix B

National University IRB Approval Letter

Date: 5-30-2025

IRB #: IRB-FY24-25-890
Title: Patient Behavior Prediction Using Big Data Analytics and Machine Learning
Creation Date: 4-29-2025
End Date:
Status: Approved
Principal Investigator: Usha Narayanan
Review Board: NU IRB
Sponsor:

Study History

Submission Type Initial	Review Type Exempt	Decision No Human Subjects Research

Key Study Contacts

Member Joseph Issa	Role Co-Principal Investigator	Contact jissa@nu.edu
Member Usha Narayanan	Role Principal Investigator	Contact u.narayanan9744@o365.ncu.edu
Member Usha Narayanan	Role Primary Contact	Contact u.narayanan9744@o365.ncu.edu