Signal Processing Systems
Mekelweg 4,
2628 CD Delft
The Netherlands
https://sps.ewi.tudelft.nl/

CAS-2023-00

M.Sc. Thesis

Prediction of Post-induction Hypotension by Machine Learning

Shuoyan Zhao

Faculty of Electrical Engineering, Mathematics and Computer Science **Delft University of Technology**

Prediction of Post-induction Hypotension by Machine Learning

Thesis

submitted in partial fulfillment of the
requirements for the degree of

Master of Science

in

Electrical Engineering

by

Shuoyan Zhao
born in Hefei, China

This work was performed in:

Signal Processing Systems Group
Department of Microelectronics
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology

Delft University of Technology

Delft University of Technology
Department of
Microelectronics

The undersigned hereby certify that they have read and recommend to the Faculty of Electrical Engineering, Mathematics and Computer Science for acceptance a thesis entitled **"Prediction of Post-induction Hypotension by Machine Learning"** by **Shuoyan Zhao** in partial fulfillment of the requirements for the degree of **Master of Science**.

Dated: August 15, 2023

Chairman: _____
prof.dr.ir. J.Dauwels

Advisor: _____
prof.dr.ir. J.Dauwels

Committee Members: _____
dr. M.Gürel

_____

Abstract

Anesthesia-related hypotension is a significant concern during surgery, occurring shortly after induction and potentially leading to severe complications. Since the anesthetic drug is believed to have an important role in the occurrence of post-induction hypotension (PIH), anesthesiologists now advocate for the appropriate selection of anesthetics dosage to avoid PIH. To facilitate such decision-making, an accurate prediction of PIH associated with a certain dosage of anesthetics is necessary. This thesis presents a high-accuracy prediction model for PIH that supports anesthesia decision-making. The model is trained on data from the VitalDB database of 320 patients undergoing general anesthesia. The target output of this classification model is the occurrence of PIH, as defined through comprehensive analysis that incorporates clinical operations. Besides demographic data and vital signs, our model incorporates the dosage of propofol administered during the induction period as an input variable, mimicking real-world anesthetic plans. By employing the model in the target control infusion system of anesthesia, the anesthetics dosage can be varied as input, providing outcome predictions as security suggestions. An ensemble algorithm is employed to balance the prediction performance and the ability to elucidate the positive relationship between propofol and PIH risk, forming an anesthetics advice model. Compared to previous PIH prediction studies, our prediction model is validated in more reliable nested cross-validation approach and achieves a higher performance (precision of 0.83 and recall of 0.84). We believe utilizing demographic and dynamic vital signs to predict HIP can be useful in determining the appropriate anesthetic dosage plan, offering potential improvements in patient care and safety.

Acknowledgments

At this moment of completing this paper, I would like to express my deepest gratitude to many people whose support and encouragement have been the driving force behind the accomplishment of this research. I would like to thank my advisor prof.dr.ir. J.Dauwels for his assistance during the writing of this thesis. Your patient guidance and professional insight have been invaluable, enabling me to grow academically. I am grateful for your support and encouragement, especially during the times when I needed it the most. Next, I want to thank my friends Wenrui Yu, Ruiyu Shen, and Xuan Gao. You have been an indispensable part of my life, and your companionship and encouragement have made the writing process more enjoyable and light-hearted. Thank you for being by my side, sharing both the joys and sorrows, even amidst my busy academic life. I would also like to extend my special thanks to my parents. Thank you for your understanding, support, and encouragement throughout. Special thanks to Dr. Ottenhof and Dr. Korstanje from Erasmus Medical Center for providing this fascinating topic and their expert medical guidance. Their valuable insights have played a crucial role in shaping our research.

Shuoyan Zhao
Delft, The Netherlands
August 15, 2023

Contents

List of Figures

List of Tables

1 Introduction

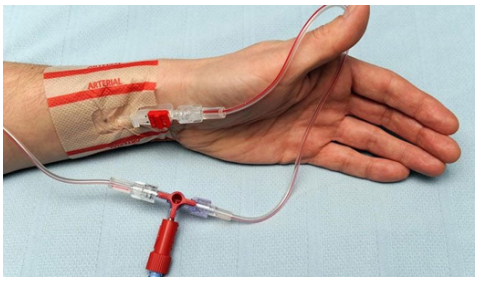
1.1 Motivation

Our project aims to improve the safety of anesthesia in surgeries by offering personalized advice on anesthetic dosage. Presently, the dosage is determined by algorithms within the target-controlled infusion (TCI) system. The TCI system automatically calculates the appropriate dosage and infusion rate of anesthetics to achieve the desired depth of anesthesia, using basic patient information such as gender, weight, and age and employing built-in pharmacokinetic models. However, the current system lacks accuracy and personalization, frequently resulting in adverse outcomes during operations, especially for patients with compromised health conditions. Traditional modeling techniques are inadequate in addressing the complex factors contributing to these adverse events. To tackle this issue, researchers are exploring the integration of machine learning (ML) models capable of analyzing extensive sets of vital signs and demographic data to offer safer anesthesia plans for patients [2][3].

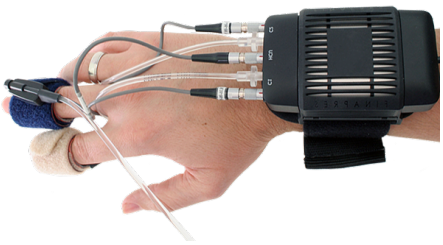
As an initial step in this project, our work focuses on predicting one of the most common adverse events related to perioperative and anesthesia care: hypotension. Specifically, we aim to predict acute hypotension that often occurs shortly after the induction of anesthesia, known as post-induction hypotension (PIH). Providing feedback on the predictive risk of PIH based on anesthetic dosage could facilitate safer decision-making for individual patients.

Before we proceed with the technical details, it is essential to clarify the current background related to clinical matters. Firstly, the prediction of hypotension is not an entirely new topic in anesthesia. Traditional methods such as blood pressure (BP) monitoring and electrocardiogram measurements have offered references for anesthesiologists to make timely adjustments to anesthesia. Some commercial devices also provide risk warnings by monitoring changes in blood vessels. While these devices are effective at early warning and enabling early healthcare, they do not assist in decision-making before anesthesia, which has the potential to directly prevent adverse events from occurring. Additionally, relying solely on the hemodynamic data from these devices might overlook other personal variations among patients. Researchers suggest that integrating demographic data could improve the accuracy of predictions [4].

Another drawback of the current hemodynamic monitoring approaches is their cost and associated risks. In most surgical scenarios, BP monitoring is achieved through either invasive arterial lines or non-invasive inflatable cuffs. Arterial lines involve inserting a small catheter into a patient's artery to provide continuous BP



(a) Arterial line [10]



(b) Volume clamp BP monitor [11]

(c) Upper arm BP monitor [12]

Figure 1.1: Three types of clinical-use BP measurement device. (a) Arterial line for ICU environment BP monitoring. (b) Volume clamp BP monitor provides continuous BP data but is less accurate and stable. (c) Upper arm BP monitor is a traditional and intermittent approach.

measurements. While this method offers continuous and high-fidelity BP waveforms, it is expensive, technically demanding, and carries some risk of complications. Arterial lines are typically reserved for high-risk patients undergoing surgery or in intensive care units, which may introduce a bias in the health situation data. Therefore, non-invasive BP (NIBP) measurements are commonly preferred in clinical settings whenever possible, especially in pre-operative environments, despite their relatively lower accuracy [5]. Non-invasive devices utilizing cuff-based oscillometric methods automatically inflate the cuff and calculate BP. While some products offer continuous NIBP through additional volume clamping [6][7], this approach is costly and not widely utilized. Generally, NIBP measurements are intermittent, with a frequency of approximately every 3 minutes in the operating room at Erasmus Medical Center (EMC). Although the intermittent nature of these measurements may lead to missed detection of hypotension [8], a study on hypotension prediction in ICUs (Intensive Care Units) argues that even with a measurement interval of 5 minutes, the predictive ability is not substantially impaired [9]. Therefore, using NIBP data in ML-based hypotension prediction is preferred in clinical practice. Consequently, the prediction tool is expected to utilize intermittent NIBP measurements to ensure its applicability across a broader range of cases.

1.2 Problem Formulation

In light of the stated motivation, the primary objective is to develop a ML model for predicting PIH using common pre-operation medical data. This predictive model aims to offer valuable decision-making suggestions on anesthesia dosage plan to medical professionals. To accomplish this, we must address the following key objectives:

1. Prediction Based on Pre-Induction Data: The ML model should solely rely on pre-induction data to enable proactive decision-making. This will allow us to forecast the likelihood of PIH before any medical interventions are initiated.

2. Intuitive Output for Anesthesiologists: The model's output should be presented in a comprehensible manner for anesthesiologists, providing them with dose-related information. We might illustrate the PIH risk with respect to changes in anesthetic dosages, making the results easily explainable and actionable.

3. Ensuring Prediction Accuracy: We should emphasize reliable ML techniques to achieve a high level of prediction accuracy. This accuracy is crucial to instill confidence in the model's results and pave the way for further research on this topic.

4. Practicality with Limited Data Types: Given practical constraints, our model should utilize a limited set of data types, specifically intermittent and NIBP measurements and other commonly available data in the operation room.

1.3 Contribution

This work collaborated with Dr. Niki Ottenhof and Dr. Jan Wiebe Korstanje from the Department of Anesthesiology in EMC, who provided professional criteria from an anesthesia perspective. In this study, we aimed to improve the prediction of PIH using ML techniques. We proposed a set of advancements that enhance the accuracy and practicality of the prediction model, enabling more effective clinical use. By combining insights from previous medicational studies on PIH and binary ML prediction knowledge, we presented the following improvements:

1. We introduced a novel labeling algorithm that offers a more practical definition of HIP based on two types of intermittent NIBP, thereby facilitating its clinical application in real-world scenarios.

2. Through the implementation of advanced ML tools in feature engineering and model training, we enhanced the performance of our PIH prediction compared to previous ML-based approaches.

3. We demonstrated the feasibility of predicting PIH using intermittent BP measurements instead of relying on costly continuous and invasive data, which may lead to a more practical and convenient approach for monitoring and predicting PIH.

4. To validate our approach effectively despite having a single small-sized dataset of 320 cases, we employed a leave-one-out nested cross-validation methodology, providing a more robust and persuasive validation technique.

5. Additionally, we proposed an ensemble model capable of accurately and intuitively illustrating the PIH risk corresponding to propofol dosage. By leveraging data from before induction and medication information during induction, the TCI system could proactively adjust dosage decisions prior to administration, enhancing patient safety and care.

1.4 Outline

The rest part is organized as follows:

- Chapter 2 reviews previous studies on PIH, ranging from medical statistical approaches to ML-assisted methods, including discussion on clinically validated ML approaches. Additionally, it presents explorations of ML-based dosage recommendation algorithms.

  Chapter 3 to 5 follows the pipeline of a binary-classification prediction problem.

- Chapter 3 discusses the dataset and our preprocessing strategies. The outcome labeling and feature extraction methods incorporate up-to-date medical background knowledge. Furthermore, we combine three feature selection approaches to identify the most valuable input for the ML models. We also provide medical explanations for the selected factors.

- Chapter 4 presents several strategies used to improve the performance of the predictive and dosage advice models.

- Chapter 5 elaborates on the assessment algorithm, along with the dosage advice model capable of providing risk warnings correlated to propofol dosage.

- Chapter 6 summarizes the contributions of this thesis and suggests ideas for future research from different aspects.

2 Background and Related Work

2.1 Medical Background

This section aims to provide an overview of the academic background behind the HIP prediction study. It starts by defining PIH in a clinical context and then explores its relative predictors, which have been validated as highly correlated to PIH through medical research or adopted in machine learning studies. In addition to presenting machine learning algorithms, we will also delve into ML-based decision-making approaches. The interest lies in understanding how such algorithms validate their generated cases, which may lack ground truth.

2.1.1 Overview of HIP

Hypotension is a condition characterized by abnormally low BP in a patient. Unlike hypertension, which is usually a long-term health issue, hypotension often occurs as a result of specific events such as blood loss, sudden posture changes, or using of medication. Prolonged hypotension can lead to serious complications or even mortality due to insufficient blood supply to organs and tissues [13]. While clinical experience plays a vital role in determining hypotension, there is currently no consensus on the precise definition of hypotension, resulting in varying rates of hypotension during operation being reported [14]. The definitions are based on BP thresholds and duration. In our work, we adopt the definition provided by EMC, which is further elaborated in the primary outcome statement in Chapter 3. In the context of surgical procedures, intraoperative hypotension (IOH) refers explicitly to occurrences of low BP during surgery in the operating room, while PIH refers to those that happen just after the "induction" period or also be explained as "hypotension before intubation" in some literature. Generally, an anesthesiologist will take care of patients during the entire duration of surgery. The anesthesia period could be divided into three stages: induction, maintenance, and emergence. The first stage, "induction," involves rapidly sedating the patient to induce unconsciousness. This is accomplished through the administration of anesthetics at a high rate of injection. Following the induction, the "maintenance" phase begins. A low infusion rate of anesthetics helps to sustain the patient's deep sleep state. This stage ensures that the patient remains unconscious and stable throughout the surgery. The end of the "emergence" stage marks the end of the anesthesia period, when the infusion of anesthetics is ceased, allowing the patient to gradually recover from their unconscious state.

PIH is a significant concern that often occurs within the first 15 minutes after induction due to hemodynamic instability during this period [15][16]. To mitigate the occurrence of PIH, various clinical strategies are commonly employed, such as the use

Table 2.1: Predictors and the performance in related literature.

Predictors/Risk Factors	OR (95% CI) *	Supporting Literature
Age	1.03 (1.02–1.04)	[15]
Baseline BP		
SBP 70 mmHg	5.00 (2.78-9.02)	[17]
SBP increment	0.97 (0.97–0.98)	[15]
MAP increment	1.05 (1.01-1.11)	[18]
Gender		
Male	1.41 (1.12–1.79)	[15]
Medication Plan		
propofol (v.s. thiopental or etomidate)	3.94 (2.42–6.43)	[17]
Weight	0.85 (0.79-0.91)	[19]
Shock Index (SI)		
$0.7 < \text{SI} \leq 1.0$	1.8 (1.1–2.8)	[20]
$1.0 < \text{SI} \leq 1.3$	2.9 (1.3–6.1)	[20]
ASA III-V	1.55 (1.22–1.99)	[17]

* OR: Odds Ratio, CI: Confidence Interval.

of more moderate anesthetics, administration of vasopressor drugs to high-risk patients or when hypotension occurs, or adjusting the depth of anesthesia (DoA). Ideally, an optimal anesthesia plan should be designed to efficiently achieve the target DoA while minimizing the potential risk of PIH, which is the motivation for our project.

2.1.2 Predictors Studies of PIH

The data collection process involves deciding which kind of data to collect, and it largely depends on the application context. The choice is especially deliberate-careful in medical applications where the clinical measurements are costly or even harmful. Previous studies thus explored deeply to identify the most valuable predictors of hypotension. Some medical works attempt to prove the relationship between predictors and hypotension through statistical approaches in either prospective or retrospective experiments. In Table 2.1, we survey the medical literature and present the predictive factors which are considered highly associated with PIH and their performance. The predictors under investigation include not only the raw and statistical features but also some artificial features. Shock index (SI), for example, is a valuable factor which is the ratio of blood pressure to heart rate, which does not has physical meaning.

Medication is the key factor directly associated with PIH during anesthesia. Therefore, the knowledge of anesthetics and anesthesia devices can provide insights into predicting hypotension. The current dosage plan for anesthetics in TCI products is calculated based on input information such as weight, age, gender, and height,

which is manually input before induction. Manual adjustments can be made by setting a lower target concentration for induction in weaker patients diagnosed with pre-existing hypertension or other cardiovascular conditions. Some TCI products simplify the anesthesia process by automatically filling in standard data, saving time for anesthesia doctors. TCI controls the infusion of two anesthetic drugs, propofol, which is primarily used to induce intoxication, and remifentanil, which is used for analgesia, in combination to achieve the depth of hypnosis required for surgery. Vasodilators such as ephedrine, norepinephrine, and epinephrine are also administered during surgery to treat acute hypotension.

However, only the relationship between drug regimens and hypotension is generally studied in the research, and the effect of a particular drug at a particular dose is not widely studied, although the improper dosage is considered the direct cause. It is probably due to the difficulty of controlling variables in a clinical setting and complex potential pathophysiologic mechanisms [21].

2.2 ML-based Hypotension Prediction

In this section, we will discuss the literature concerning the prediction of acute intra-operative hypotension. Additionally, we will explore the level of trust that clinicians place in these prediction methods and how engineers validate their performance.

2.2.1 ML Applications on Acute Hypotension Prediction

In Table 2.2, we summarize a comprehensive overview of machine-learning-based acute studies of hypotension prediction, categorized into two groups: one group studies the prediction of PIH, and the other one focuses on the prediction of IOH and hypotension occurrences during ICU stay. It is important to note that the studies in this table exclude long-term prediction approaches which predict hypotension 30 minutes later or longer. There are several distinctions between the two types of hypotension predictions. Firstly, PIH is a symptom directly associated with anesthesia induction, whereas IOH and ICU hypotension are more related to the temporary health condition of patients during the maintenance stage of anesthesia. This fundamental difference is also why IOH and ICU hypotension can sometimes be accurately predicted solely through vital signs while that is not the case for PIH [24]. Secondly, PIH prediction is limited to detecting outcomes within a narrow and fixed time range, specifically shortly after anesthesia induction. This restriction results in only one case being available from all recordings for a single patient. On the other hand, IOH retrospective labeling can generate multiple cases by analyzing monitoring data throughout the entire surgery, significantly enriching the training data. Lastly, PIH predictions face challenges due to limited access to comprehensive databases in terms of both duration and fidelity. During the pre-operative stage, when the patient has just entered the operating room, few monitoring devices are available. Moreover, unlike ICU studies that benefit from solid support from the MIMIC (The Medical Information Mart for Intensive Care) database [25], which provides large and de-identified public ICU data, machine

Table 2.2: Summary of important literature of the predictions of hypotension.

Source	**Prediction Outcome**	**Data TYPE**	**Algorithm**	**Datasize**	**Performance**
Post-induction Hypotension (PIH)					
2018 Kendale [22]	PIH in 10 min after induction start	non-invasive biosignals + EHR + drug	XGBoost	13,323 cases (1185 PIH)	$precision$ = 0.85, $recall$ =0.84, ROC_AUC =0.
2020 Kang [16]	PIH between intubation and incision	non-invasive biosignals + EHR + drug	Random Forest	222 cases (126 PIH)	$precision$ =0.85, $recall$ =0.83
2020 Lee [23]	PIH between intubation and incision	non-invasive biosignals + EHR + vasoactive drug	Random Forest	282 cases (151 PIH)	$precision$ =0.76, $recall$ =0.78
Intra-operative Hypotension (IOH) or ICU Hypotension					
2018 Hatib [24]	ICU hypotension 5 min ahead	(invasive) high-fidelity arterial pressure waveform	Logistic Regression	1,334 patients	$recall$ =0.87, $specificity$ =0.89
2021 Lee [4]	IOH 5 min ahead	invasive multichannel biosignals	Multichannel DL	3301 patients	$recall$ =0.86, $specificity$ =0.86
2021 Moghadam [9]	ICU hypotension 30 min ahead	noninvasive physiological signals	LR, SVM	1000 patients	$recall$ =0.84, $specificity$ =0.94, F_1score =0.78

learning studies on hypotension in surgical settings have primarily relied on private or commercial data until the publication of VitalDB [26]. Unfortunately, VitalDB still lacks enriched pre-operative data. In the following paragraphs, we will explore these two groups of studies in detail, discussing their methodologies and performance.

Kendale et al. [22] predicted PIH within 10 minutes after the start of induction, strictly defining PIH as a single value of mean arterial pressure (MAP) < 55 mmHg since MAP is measured intermittently every minute. In their study of 13,323 cases (1185 of which experience PIH), minimal preprocessing was performed on raw data to mimic real-world conditions. They extracted features from demographic data, intraoperative medications, and vital signs within the same 10-minute window. Notably, medication information is more detailed and plays a crucial role in the final model compared to other similar studies. However, there is no clarification on how they avoid potential data leakage, and there might be a slight concern about predictions using data collected after the event "happens". Furthermore, a repeated 10-fold cross-validation strategy was employed during the training process, but the training-testing-set splitting was performed only once without cross-validation. The work also lacks detailed evaluation metrics except for AUC (Area under the receiver operating characteristic curve), and there is no indication of how the model handles the data imbalance problem in the dataset, which only contains 8.89% positive events. Although it acknowledges that a model capable of dealing with imbalance performs better, this aspect remains unexplored. Furthermore, among the eight machine learning classification models tested in the tenfold cross-validation, the GBM model achieves the highest AUC of 0.76, which is far from perfect accuracy and has been criticized for potential misclassification issues and the inherent weaknesses of boosting models related to this misclassification.

Kang et al. [16] employed a different definition of the post-induction period, specifically referring to the short interval between tracheal intubation and incision. They labeled hypotension occurring after intubation as positive when a singular measurement of systolic blood pressure (SBP) < 90 mmHg or MAP < 65 mmHg is observed. However, it should be noted that the definition provided in the article appears somewhat ambiguous. Additionally, they collected data on the frequency and duration of early PIH, which takes place between induction and intubation, as input features. This approach essentially divides the PIH labeling period into training and labeling stages, a distinction that can significantly impact classification results, particularly when working with limited data for which a clear and direct reference in the PIH topic is lacking. By selecting 15 features from electronic health records (EHR), medications, and vital signs collected before intubation, the random-forest model achieves an impressive AUC of 0.84. However, a statistical analysis of the 222 cases reveals a PIH ratio of 56.8%, considerably higher than the clinically occurring ratio of 20% reported in [27]. The article attributes this difference to more frequent BP measurements, but it is essential to consider the small size of the database and the potential impact of variations in the PIH definition as other contributing factors.

Hatib et al. [24] introduced the HPI (Hypotension Prediction Index) algorithm for predicting hypotension events in the ICU. The classification outcome, "HPI", represents the probability of a hypotension event occurring 5 minutes later, generated by a logistic regression model. The algorithm leverages high-fidelity waveforms of arterial pressure from the records of 293 patients and employs 3022 extracted features from waveform featurization, achieving an impressive performance with an AUC of 0.95. The results are further validated using an external database of 204 patients. However, the performance of model diminishes when attempting to predict hypotension events happening 15 minutes later or when setting a higher classification threshold. It is important to note that the applicability of this work to PIH prediction is limited for several reasons. Firstly, the labeling of long-term hypotension events is done retrospectively. In the record of a patient, several positive and negative events are identified and then matched with data 5 minutes ahead to form the training and testing sections, creating specific distinctions in time. This can not be generalized to PIH prediction, which predicts in a short induction period. Secondly, the use of invasive measurements of BP, which is not common in standard surgeries, may introduce data bias concerning the health conditions of patients. Additionally, some subtle features can only be reliably captured within high-fidelity signals, potentially limiting the generalizability of algorithm to scenarios with less precise or invasive measurements.

2.2.2 Clinical Validation

As mentioned earlier, what anesthetists desire from the system is the detection of hypotension risks that might be overlooked with basic monitoring alone. Therefore, clinical validation holds more significance and persuasiveness than the accuracy of the predictive models. Some recent studies have focused precisely on this aspect. For instance, Palla et al. [28] conducted a study on the prediction of post-anesthesia care unit (PACU) hypotension using binary classification. In their research, nine anesthesiologists reviewed real data and made anesthesia-related decisions, simulating a real clinical environment. Comparative experiments involving decisions made with and without the assistance of AI models demonstrated that the predictive models indeed had a positive impact on decision-making. Furthermore, there have been several clinical experiments on HPI-guided care [29][30]. The study in [30] showed that early intervention implemented when HPI<80 (indicating a hypotension risk of more than 80% as calculated by the HPI algorithm) led to a reduction in the duration of hypotension in patients, thus validating the effectiveness of the HPI algorithm. However, it should be noted that the originally suggested threshold given by HPI is 43, largely different from 80, and as a result, retrospective validation may not be as persuasive.

In conclusion, it is evident that there is still a long road ahead before ML-based hypotension prediction can be practically employed, regardless of the type of acute hypotension being addressed. Further research and thorough clinical validation are necessary to ensure the reliability and applicability of such predictive models in real-world healthcare settings.

2.2.3 ML Algorithms

In this section, we will offer a concise overview of the two primary groups of basic machine learning models employed in the literature and this project: logistic regression and decision tree. Additionally, there is also a discussion of the explainability and reliability of ML models in the context of medical prediction.

2.2.3.1 Logistic regression

Logistic regression (LR) is a popular statistical model for binary classification tasks. It uses the sigmoid function, also known as the logistic function, to generate probabilities for binary events, with the outcome interpreted as the probability of predicting an event. The sigmoid function ensures that the predicted probabilities lie between 0 and 1, as shown in Figure 2.1, allowing the model to classify data into two classes based on a chosen threshold. LR assumes independence among variables, which can be a limitation in certain scenarios. This assumption, however, may not hold true when dealing with correlated variables. Another limitation of LR is its linearity assumption, assuming a linear relationship between independent variables and the log-odd of the dependent variable. This may not accurately capture complex data relationships. LR is sensitive to outliers, as extreme values can distort the predictions of the model and affect its performance. Additionally, LR assumes the proportional odds assumption, which assumes the relationship between variables and the log-odd remains constant. Violations of this assumption can lead to biased results. Despite these limitations, LR remains valuable for its simplicity and interpretability. It is often used when the data characteristics align with the assumptions behind the model.

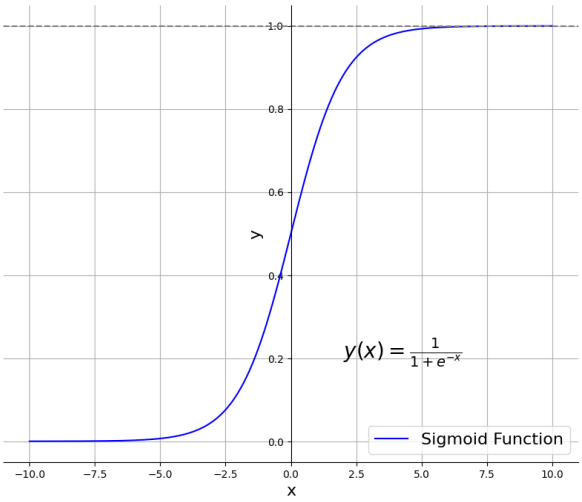


Figure 2.1: The sigmoid function.

2.2.3.2 Decision Tree

Decision tree is a popular type of machine learning model, adept at handling both classification and regression problems. The construction of the big tree is built by recursively partitioning the input space based on the values or scores of the selected features. A decision rule is applied at each internal node to determine the branch to follow, leading to further partitioning. The process continues until the terminal nodes, or leaves, are reached, which hold the final prediction or regression value. The main drawback of the decision tree is its vulnerability to overfitting, where the model captures noise in the data, leading to poor generalization.

Random forest (RF) models address the limitations of decision trees by employing an ensemble learning technique [31]. They consist of an ensemble of decision trees, each of which is trained on a different subset of the training set. During the training process, a random subset of features is considered for splitting at each node. The final prediction of a random forest is obtained by averaging the predictions of all individual trees. Random forests mitigate overfitting by reducing the variance associated with a single decision tree, resulting in improved generalization performance and the ability to handle large feature spaces.

XGBoost (Extreme Gradient Boosting) further leverages gradient boosting to create a powerful ensemble model [32]. It builds decision trees sequentially, where each subsequent tree corrects the mistakes made by the previous trees. XGBoost incorporates regularization techniques such as shrinkage, feature subsampling, and tree pruning to prevent overfitting. It uses a differentiable loss function to optimize the model's performance, enabling efficient training and scalability. XGBoost is known for its ability to handle complex tasks, high-dimensional datasets, and imbalanced data. It has gained popularity in both academia and industry due to its exceptional predictive accuracy and flexibility.

2.2.3.3 Explainability of ML models

Medical applications of AI require high interpretability before implementation, as the medical industry seeks to understand the reasons behind predictions, especially when it affects decision-making. While some ML models may be considered simple and easy to interpret traditionally, the field of Explainable AI (XAI) aims to find a certain set of features within complex predictive models to achieve both accuracy and interpretability. XAI research in the medical domain began in 2018, and many studies employ methods like SHapley Additive exPlanations [33] in XGBoost or RF algorithms, which are traditionally considered black-box models. For instance, Lundberg et al. [34] applied a predictive system to warn of hypoxemia during surgery while visualizing the weighting of features over time to explain the model's predictions. This model also underwent clinical validation, where doctors made decisions based on the prediction results and explanations. The results showed that anesthesiologists, with the help of machine learning, could increase their anticipation by 15%.

2.3 ML-based Dosage Recommendation

Data-driven treatment recommendation studies have gained significant attention in the medical field for quite some time. However, the application of machine learning algorithms to build medical recommendations is relatively new. In this section of the review, we will focus on the model-building and result-accessing approaches of those studies rather than the training process of their models. Specifically, we are interested in the validation process without clinical prospective experiments. The results are typically compared to baseline values, which can either be the manual diagnoses provided by doctors or specific target scores or detection.

Javad et al. [35] employ Q-learning to recommend personalized insulin dosage levels for controlling the symptoms of type 1 diabetes mellitus (T1DM). The reward in this reinforcement learning model is determined by the change in glycated hemoglobin levels, which is considered positive when levels decrease. The evaluation is based on a comparison between the generated Lantus dose and the actual dose prescribed by the patient's physician. The results show an 88% agreement with the physician's decision, demonstrating the effectiveness of the model in dosage recommendations that physicians accept. In a word, the way the model calculates an optimal dosage is based on the patient's health condition and validates it through manual comparison. Such a strategy also is effective for recommending in-time or long-term dosages if a persuasive score (reward) can be established. For example, the optimization of warfarin is based on a reward related to the International Normalized Ratio (INR), which indicates bleeding risk [36].

Bertsimas et al. [37][38] explore the prediction of adverse events from the time of diagnosis to a potential adverse event (TAE). The study benefits from a large dataset that includes diverse cases of various treatments under consideration. In their 2020 work, the model is trained on the entire dataset, while in their later work, separate models are trained for each treatment, and an ensemble method is used to determine the final result. Multiple models "vote" to predict whether the health condition will improve or worsen under such treatment. During testing, a case is evaluated using all individual models for different treatments, and the optimal treatment choice is selected based on the treatment that receives the maximum "will improve" judgments from the machine learning models. Bertsimas designs prescription effectiveness (PE) and prescription robustness (PR) metrics to evaluate the prediction. For each testing case, the ground truth is the actual treatment the patient undergoes and the real change in the antihypertensive situation. Therefore, PE compares the predictive outcome in the actual treatment model with the real outcome, while PR compares the predictive outcome with other outcomes from different treatments.

In summary, while there is limited research on anesthesia dosage recommendations using ML, the idea of using ML to assist in dosage decisions has been explored, and

cautious and conservative attempts have been made. However, the lack of a definitive "correct answer" for medication dosages makes validation challenging. Clinical and prospective experiments could be persuasive, but they are expensive and difficult to conduct. New validation metrics, like PE and PR, show promise but require further theoretical validation and development to establish their effectiveness and reliability. As the field progresses, finding suitable and robust validation methods will be crucial for advancing ML-based medical recommendation systems.

3 Dataset and Preprocessing

This study employs a classification pipeline to predict medical outcomes using various features. The pipeline includes several interconnected steps that systematically process and analyze the data, yielding the desired classification results. Specifically, this chapter will focus on the first three steps of the classical classification pipeline:

1. Data Collection: The raw data used in this study were obtained from an open-source surgical database. To facilitate further analysis, various methods for information management were employed to ensure data integrity and enhance the quality of the dataset for further analysis.

2. Data Pre-processing: The data underwent a preprocessing procedure to ensure its quality and suitability, preparing for a high-quality of further analysis. This involves addressing any errors or inconsistencies present in the data.

3. Feature Extraction and Selection: The raw data was transformed into meaningful features that capture relevant information for the predictive model. Feature Selection then identified the most relevant and informative features to enhance the model's accuracy and efficiency.

Detailed methodologies and techniques related to data collection and pre-processing will be explored, laying the foundation for a robust and data-driven prediction process.

3.1 Primary Outcome

The primary outcome of our work is to identify PIH, defined as an SBP of less than 75 mmHg or a relative SBP drop of more than 30% from the baseline. It is essential to note that there is no academic consensus regarding the exact definition of hypotension [39]. The literature varies in the different thresholds, with some adjustments based on recorded physician activity. To ensure a comprehensive analysis, we considered these factors and defined PIH as follows:

- For the NIBP recordings measured every 2 seconds, a case was labeled with PIH when, within a one-minute measurement window, at least 90% of the measurements are below the threshold. However, if the mean value of the measuring window is higher than 100 mmHg, further detection was performed, as this range is disputed in the medical field [40]. Additionally, if SBP increases by more than 20 mmHg within 3 minutes, it was considered a healthcare intervention and confirms the presence of PIH. On the other hand, if the situation persists for the rest of the operation without any significant increase, the PIH label was considered erroneously marked due to the initial SBP value, which is common due to patients' stress.

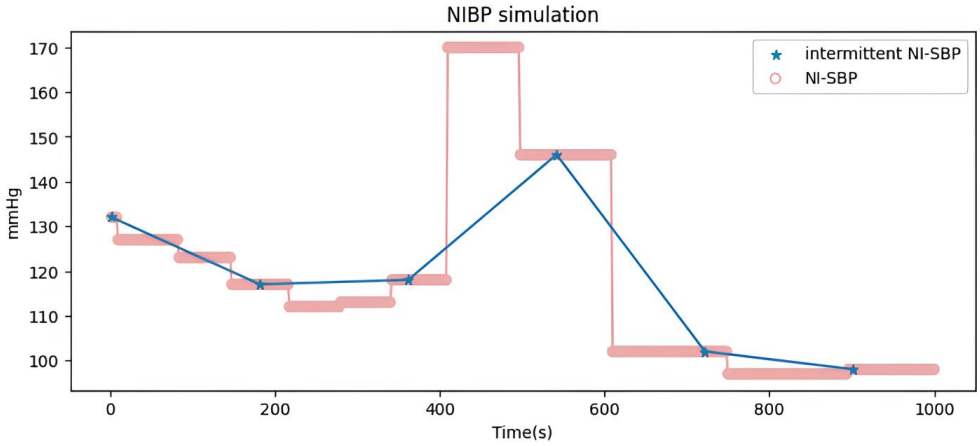


Figure 3.1: A sample recording of frequent SBP measurements vs. intermittent SBP measurements at the beginning of anesthesia.

- The threshold was determined by anesthesiologists following the clinical regulations of EMC. To match this, a 3-minute intermittent simulation of BP measurements should be built based on the 2-second NIBP records from VitalDB. Thus, for the EMC external database, the decision was made based on at least one single value of SBP of NIBP recorded within the first 15 minutes after anesthesia induction.

We performed a simulation on the 2-second BP data to discuss the possibility and performance of using more intermittent BP measurements in PIH prediction. For convenience, we will call the original BP data, which is measured every 2 seconds, as "frequent BP", and call the 3-minute simulated BP "intermittent BP" in the following content. While both measurements are intermittent in theory, the latter aligns better with the clinical definition as BP monitoring during anesthesia is usually performed in minute-based intervals. Figure 3.1 illustrates the comparison between the original frequent SBP measurement and the generated intermittent measurements. The intermittent SBP values were sampled from the first value of a 3-minute interval due to the auto-collecting interval in practice. The different measurements of BP also influence the way that a PIH event is detected. The results in Table 3.1 demonstrate that more frequent detection leads to more acute identification of hypotension events, supporting the possibility proposed by Kang et al. [16]. However, in the following experiments, we only use the labels generated by frequent SBP. Although the simulation of more intermittent data leads to different labeling results, it should not affect the ground truth of whether a patient has undergone PIH or not.

Table 3.1: Distribution of PIH under different definitions.

Data Size	Determine of PIH	Measurement	PIH	Non-PIH	Positive Rate
320 sujects	SBP < 75 mmHg or ΔSBP $> 30\%$	frequent	199	121	62.2%
		intermittent	194	124	60.6%

3.2 Data Collection and Processing

We considered data from VitalDB [26] to be used for model development and internal validation. VitalDB is an open-source surgical database containing records of 6388 cases that underwent non-cardiac routine or emergency surgery at Seoul National University Hospital, Seoul, Republic of Korea. As illustrated in Figure 3.2, specific criteria were employed for data selection. The study focuses only on general anesthesia cases, requiring patients to receive propofol during anesthesia for inclusion in the propofol dosage analysis later. NIBP recordings, which contribute to generalizability, are required. To enable pre-anesthesia dosage suggestions, data must be captured at least one minute before induction. Moreover, instances of missing or extremely erroneous data were excluded from the dataset, which will be explained in the next paragraph. In the end, data from 320 patients were selected for this study.

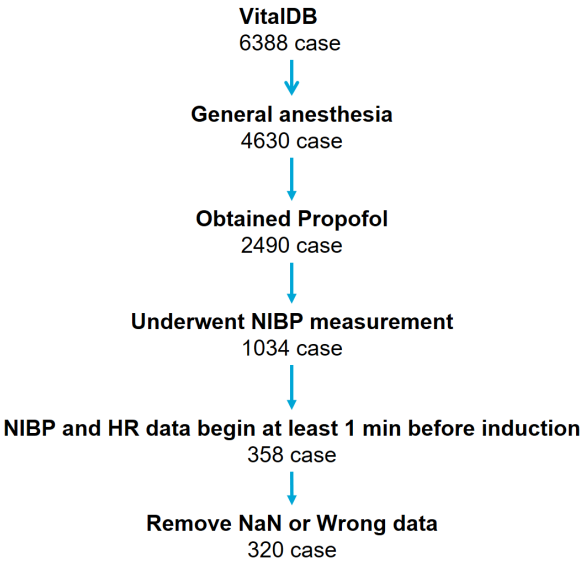


Figure 3.2: Cases filtering process based on the availability of time-series data.

We cleaned the data as follows. Firstly, we removed cases containing NaN (Not a Number) values in the demographic data. While decision-tree models can effectively handle the dataset containing NaN values, others, namely the logistic regression, may struggle to manage them. Next, we excluded cases in which vital signs had 1-minute successive missing values or more than 15% missing values in total, as well as cases with incorrect values. As a result, a total of 32 cases were excluded due to missing values, and 6 cases were filtered out due to the presence of incorrect data. The incorrect data tends to concentrate on the TCI system. There are two types of erroneous values: infusion rates higher than 2000 mL/hr and decreases in the accumulated infused volume of the anesthetics. Throughout our experiments, we did not identify any significant error in the vital signs. Data

cleaning ensures the data is reliable and free from artifacts that could negatively impact the analysis. While filling null values with current or mean data is a popular method, analysis of available data samples from VitalDB showed that there is a significant difference between the lost data and the mean or nearby values. This disparity can be attributed to the hemodynamic instability during the induction period.

The VitalDB dataset lacks precise timing information for the induction phase. To overcome this limitation, we developed a method to identify the onset of the induction phase by analyzing the infusion data collected through the TCI system. We defined the onset of induction phase as the point when the infusion rate of propofol or remifentanil first exceeds 100 mg/hr. This approach to determining the induction phase has been proposed by the Department of Anesthesia at EMC.

However, in clinical practice, there is a lack of a clear distinction between the end of the induction phase and the beginning of the maintenance period. Additionally, it is common to administer additional doses after a certain monitoring period during the maintenance phase. Moreover, the absence of intubation and incision information makes it difficult to define the post-induction period as the post-intubation period precisely. To address these challenges, we established the following criteria for determining infusion medication information. During the period prior to the start of the operation, we divided the time based on the propofol infusion rate series. The period when the rate exceeds 100 mg/hr was considered the induction phase, and the accumulated injected volume was calculated as the induction dosage during these phases. A similar process was applied to the remifentanil information.

3.3 Feature Extraction

In order to identify and capture relevant patterns and characteristics in the data, various data engineering techniques were employed. However, research on PIH faces a significant limitation in terms of data availability compared to other popular medical studies like ECG (Electrocardiogram) or EEG (Electroencephalogram). The pre-operative studies of anesthesia suffer from challenges such as the busy anesthesia environment, cost considerations, and risk factors. Collecting high-fidelity data in real clinical conditions is seldom achieved, even in clinical prospective experiments. As mentioned above, without invasive devices, only intermittent vital signals are available. The original measurements in VitalDB are recorded every 2 seconds, but for intermittent simulation, we consider BP measurements to be taken every 3 minutes. Therefore, it is reasonable to focus on time-domain features and statistical features. Additionally, we can extract extra-temporal trend features for frequent BP and other vital sign signals. The following tables provide a comprehensive overview of the different types of features. In summary, the feature groups consist of the following:

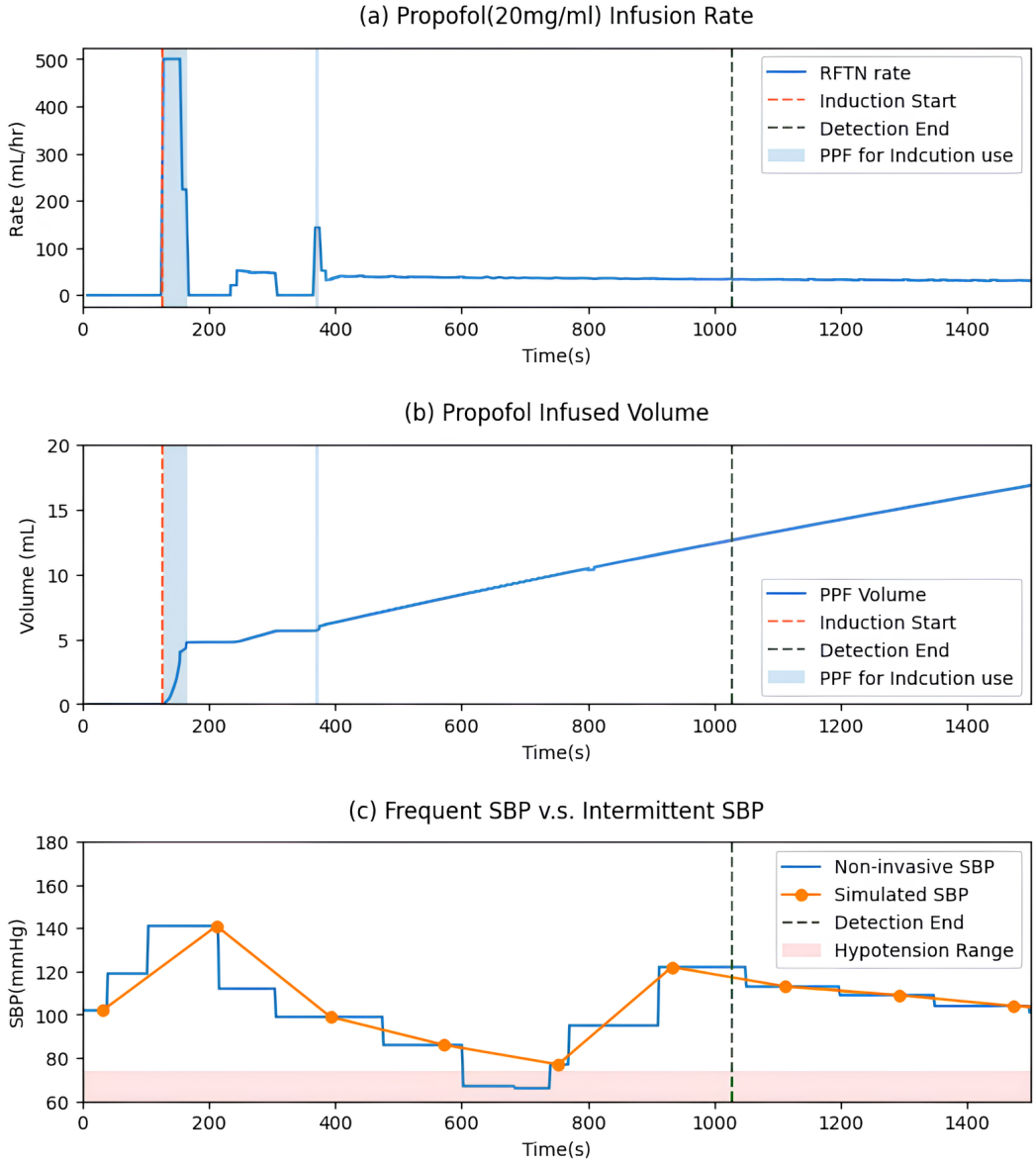


Figure 3.3: Information extraction from time-series data. (a) and (b) demonstrate the extraction process of induction dosage information from propofol infusion data. Specifically, the effective induction dosage is determined by considering only the dosage administered during the recognized induction period, depicted by the blue rectangle. (c) illustrates the detection of PIH using both frequent NIBP and intermittent NIBP approaches. The monitoring of PIH continues until the end of the first 15 minutes after induction, indicated by the green vertical dotted line. However, due to the presence of large intermittency in the simulated SBP data, there is a detection omission marked by the red area.

1. Demographic features:
   These features are directly provided in VitalDB in CSV (Comma Separated Values) format and are ensured to be available in all cases.

Table 3.2: List of demographic features.

Features	**Description**
Age	-
Gender	-
BMI (Body Mass Index)	Ratio of weight to height
Height	-
Weight	-
Preoperative disease records	Records of diseases before the operation
Preoperative hypertension	Preoperative hypertension status
Preoperative diabetes	Preoperative diabetes status
Preoperative ECG diagnosis	Cardiac diagnosis from preoperative ECG
Operation Type	-

2. Medication features:
   We included features related to medication. In VitalDB, anesthetics in TCI have concentration, volume, and rate information recorded, allowing us to calculate time-domain features about volume and rate. However, for other drugs used for vasopressor and analgesia, only the whole intra-operative dosage data is collected, which may potentially lead to some minor data leakage issues in this context. Target concentration refers to the concentration of anesthetics in order to keep certain DoA and was calculated by the embedded algorithm inside TCI. Since target concentration is not a reliable measure of actual dosage, we rely instead only on induction time and dosage.

Table 3.3: List of medication features.

Features	**Description**
Intraoperative anesthetics	
Propofol Dosage	Propofol administrated during induction
Remifentanil dosage	Remifentanil administrated during induction
Time	Start time of induction
Phenylephrine dosage	Phenylephrine intraoperative usage
Ephedrine dosage	Intra-operative usage of ephedrine
Epinephrine dosage	Intra-operative usage of ephinephrine

3. Time-domain features of vital signs:
   These features encompass statistical measures such as mean, standard deviation, minimum, maximum, variance, slope, intercept, and delta changes. Specific features like pulse pressure (pp) for BP are also included.

Table 3.4: List of features about blood pressure.

Features	**Description**
Blood Pressure	
Systolic Blood Pressure	
first/baseline	Average of first 5 seconds measurement
SBP before induction	A single value of SBP right before induction
mean, min, max	-
standard deviation	-
delta-change from baseline	A single value before induction v.s. baseline
pulse pressure	(pp)
pulse pressure variation	(PPV)
Mean Arterial Pressure	
first/baseline	-
MAP before induction	-
mean, min, max	-
standard deviation	-
delta-change from baseline	-
pulse pressure	(pp)
pulse pressure variation	(PPV)
Diastolic Blood Pressure	
first/baseline	-
DBP before induction	-
mean, min, max	-
standard deviation	-
delta-change from baseline	-
pulse pressure	(pp)
pulse pressure variation	(PPV)

Table 3.5: List of features from other vital signs.

Features	**Description**
Peripheral capillary oxygen saturation (SpO2)	
delta change from baseline	-
SpO2 before induction	-
moving mean, min, max	-
moving standard deviation	-
rate in low	The rate of saturation that under the threshold in a 1 min window

rate in high	-
Heart rate	
mean, min, max	-
standard deviation	-
delta change from baseline	-
Heart rate before induction	-
moving mean, min, max	-
moving standard deviation	-
ECG (AVF, II, III)	
mean, min, max	-
standard deviation	-
moving mean, min, max	-
moving standard deviation	-
elevation or depression compared to baseline	”eleORdep”

4. Combinatorial features:
   This group comprises features derived from 2-degree polynomial combinations and ratios of one demographic feature and one feature from the previous three groups. Additionally, it includes ratios of two vital sign features that have been validated in other literature.

Table 3.6: List of combinatorial features.

Features	**Description**
Shock Index	(SI) HR/SBP
age Shock Index	age*SI
Modified shock index	(editSI) HR/MAP
Propofol to weight	$propofol.dose/weight$
Propofol to bmi	$propofol.dose/bmi$
Remifentanil to weight	$Remifentanil.dose/weight$
Remifentanil to bmi	$Remifentanil.dose/bmi$
Polynomial or Ratio with age	-
aged propofol dosage	$age * propofol.dose$
aged remifentanil dosage	$age * remifentanil.dose$
Polynomial or Ratio with weight or BMI	-
SBP.mean to weight	-
MBP.mean to weight	-
SBP.mean to bmi	-
MBP.mean to bmi	-
Other Polynomial or Ratio	
SBP.std to Heart rate	-
MBP.std to Heart rate	-
SBP.mean to Heart rate	-
MBP.mean to Heart rate	-

The combinatorial strategy was integral for enhancing the representation of raw features, while techniques like encoding also contributed to this enrichment of the feature set. For example, textual demographic and medication history data were transformed into numerical representations. Categorical variables such as gender were encoded as 1 or 0, improving compatibility between the features and the modeling process.

We extracted a total of 88 individual features from three groups of data. Additionally, we synthesized 17 combinatorial features, resulting in a combined set of 105 features. However, for intermittent experiments, the total number of features decreased to 90 due to certain features, such as pulse pressure (pp) of BP, not being available in the intermittent data. Moreover, these features were normalized after being extracted. The normalization process aimed to eliminate the magnitude differences among the various features. For instance, in our dataset, while BMI values ranged between 12.9 and 34.6, induction time values fell within the range of 9.215 to 1034.013. As a result, even minor fluctuations in the BMI values could significantly impact the predicted outcomes, thereby complicating the search for stable optimization results. Scaling both

these features to a range between 0 and 1 will benefit the classifier's optimization performance.

3.4 Feature Selection

Features in a dataset may not always be helpful for a given prediction task. Instead, they may add unnecessary complexity or introduce noise that jeopardizes the prediction. To create accurate predictions, it is important to identify the most relevant and informative features from the original feature sets. Various algorithms have been proposed for feature selection, which can be grouped into three categories [41][42]:

- Filter methods [43]: These methods use statistical tools to assign a score to each feature. Features with scores below a certain threshold are considered unqualified and filtered out. The filter methods do the evaluation on each individual feature independently, therefore efficient in handling high-dimensional datasets. However, they may ignore feature interactions, thus discarding potentially useful features in some cases.

- Wrapper methods [44]: These methods search for an optimal combination of features based on an evaluation score. This evaluation is performed by iteratively training the model with different feature subsets, which can be done randomly or by following certain search strategies (such as forward selection and backward elimination). Wrapper methods tend to yield better results than filter methods since they account for feature interactions [41]. Their disadvantages are the higher computational cost due to multiple model training iterations and the increased risk of overfitting, as they may select features that perform well on the training data but generalize poorly to new, unseen data. The choice of the evaluation metric during the feature selection process can also influence the performance of model and may introduce bias in the final feature subset selected.

- Embedded methods [43]: Feature selection in these methods are integrated into the model training process itself rather than being performed as a separate step. It thus reduces the computation cost of reclassifying subsets in the wrapper methods [44]. Particularly, it benefits classification with a large number of features and also reduces the training complexity by removing the selection procedure.

For our study, we employed multiple approaches, including Recursive Feature Elimination (RFE), which is a representative algorithm of the wrapper method, and elastic net, which is an embedded method. We have also used feature importance and correlation analysis, which are filter methods. Before conducting feature selection, we split the dataset into training and testing sets, which will be discussed further in chapter 4.

3.4.1 Correlation Analysis

A correlation analysis is conducted to examine the relationships between features and the target variable, thus identifying any strong associations indicating feature relevance.

To do so, we computed the Pearson correlation coefficient, which determines the linear dependence between two variables as:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}. \tag{3.1}$$

Here x_i and y_i represent the individual observations of the variables X and Y, respectively. $\bar{x}$ and $\bar{y}$ represent the means of X and Y, respectively. The sums are taken over all the observations in the dataset.

This coefficient helps assess the strength and direction of the linear relationship between each feature and the target variable. When the correlation coefficient is close to 1, it indicates a strong positive linear relationship. Conversely, when the correlation coefficient is close to -1, it indicates a strong negative linear relationship. A correlation coefficient near 0 suggests a weak or no linear relationship between the variables. However, it is important to note that the Pearson correlation coefficient only captures linear relationships and may not detect complex non-linear associations. This means that significant relationships might be missed if they exist in a non-linear form. Additionally, correlation does not imply causation, so even if two variables are highly correlated, it does not necessarily mean that one variable causes the other.

In Figure 3.4, we show the correlation coefficients and the corresponding p-values for each feature-target pair. We observed that several features exhibited a strong correlation with the target variable, surpassing a significance threshold of $p < 0.05$, which suggests that the observed results are less than 5% to be due to chance alone. These highly correlated features were considered potential candidates for inclusion in our predictive model.

3.4.2 Feature Importance

We also applied feature importance techniques to evaluate the relevance of each feature in predicting the target variable. Decision trees and ensemble methods based on decision trees, such as Random Forest, Gradient Boosting Machines, and XGBoost, naturally provide feature importance as part of their output. The decision tree model uses a criterion, often Gini impurity or entropy, to identify the optimal feature and split point that reduces the impurity of the target variable most effectively. This reduction in impurity is determined by comparing the impurity of the parent node with the weighted impurity of the child nodes after the split [45]. The feature importance is calculated as the mean impurity reduction when each feature is selected for splitting during the construction of the ensemble. In our project, we utilized the Random Forest model, a widely used ensemble-based decision tree model, to estimate the importance of each feature based on its contribution.

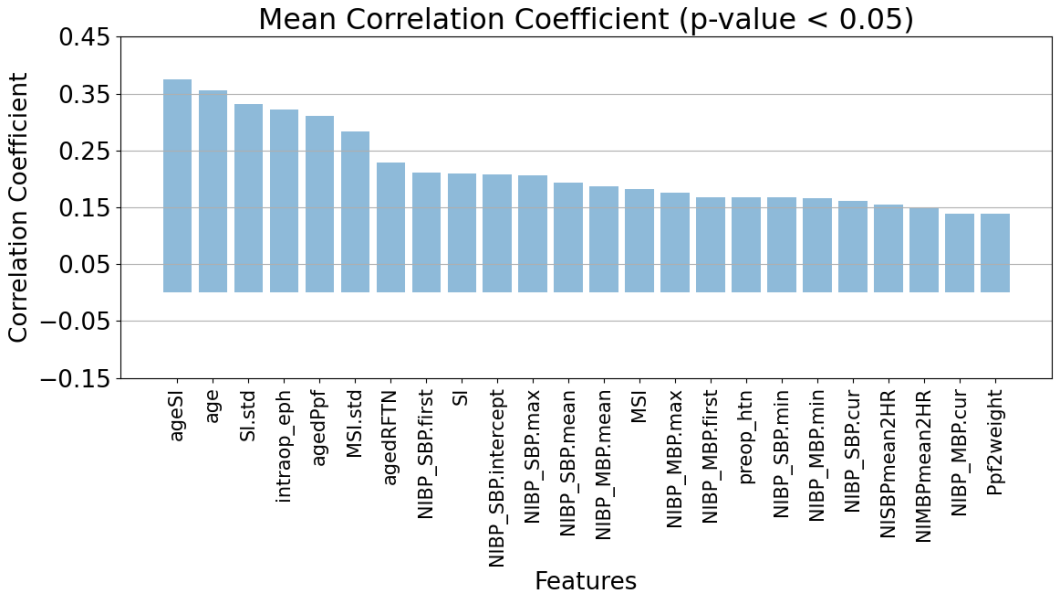


Figure 3.4: Correlation analysis of dataset within a sample loop of leave-one-out cross-validation (LOOCV).

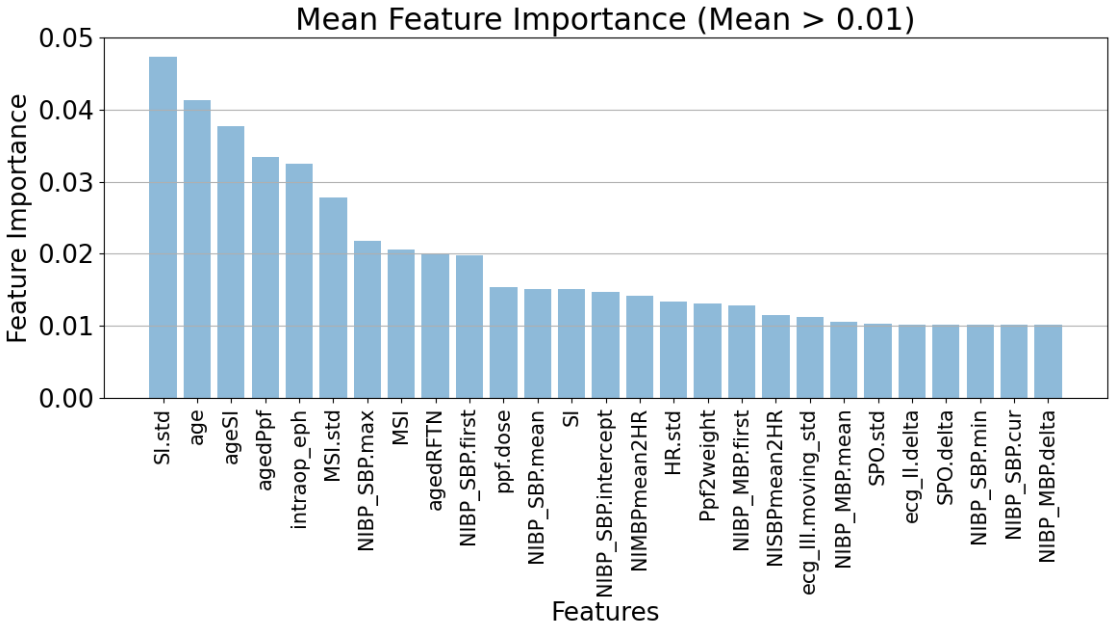


Figure 3.5: The features with a mean score larger than 1% in one loop of LOOCV, generated by a 5-fold cross-validation.

The weighted impurity at a splitting node is computed as follows:

$$\text{Weighted Impurity} = \sum_{\text{child}} \left(\frac{\text{num\_samples\_child}}{\text{num\_samples\_parent}} \times \text{impurity\_child} \right). \tag{3.2}$$

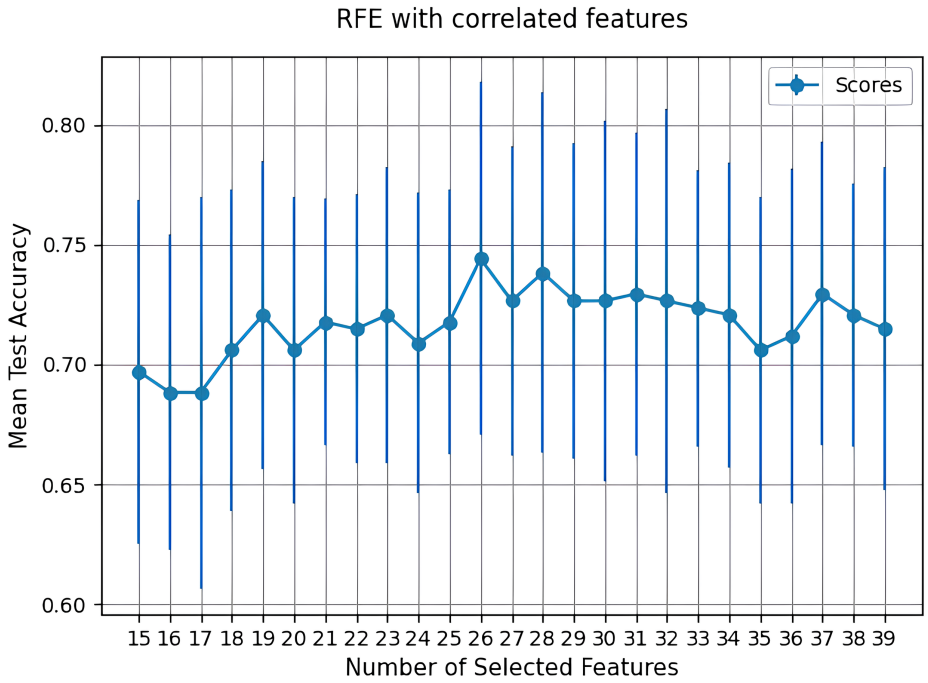


Figure 3.6: The mean accuracy result of the different selecting number of features in one loop of LOOCV. The measurement is tested by a 5-fold cross-validation.

Here num_samples_child represents the number of samples in a specific child node and num_samples_parent means the total number of samples in the parent node. The impurity_child is the impurity measure of a specific child node belonging to the splitting root. This function sums the weighted impurities, weighted by the fraction of samples it contains relative to the total number of samples in the parent node of all child nodes resulting from the split.

The feature importance scores obtained from the Random Forest model are illustrated in Figure 3.5. Certain features exhibited notably higher importance scores compared to others. These highly ranked features were considered strong indicators for the target variable in our predictive model. We chose to focus on the feature importance of Random Forest for further selection to prioritize simplicity and resource efficiency over using other bagging algorithms like the XGBoost model, which also provides built-in importance ranking.

3.4.3 Recursive Feature Elimination

RFE is a powerful feature selection technique used to systematically eliminate less important features from a given dataset to improve the model's predictive ability while reducing the complexity. The RFE algorithm starts by training the model on the entire set of features and then iteratively removes the least significant feature(s) based on a predefined criterion, such as feature importance scores or coefficients. This recursive process continues until a predetermined number of features remains.

Figure 3.6 presents the outcomes of RFE, illustrating the count of retained features on the x-axis and the corresponding performance metric (e.g., accuracy, F1 score) on the y-axis. To ensure consistent results and mitigate random bias, a repeated k-fold validation was employed during testing. Notably, the accuracy validation attains its peak at the 26-feature point. This number is chosen as the optimal feature subset to be incorporated in this iteration. By employing RFE, we aimed to enhance the explainability and generalizability of our models by focusing on the most relevant features. The resulting feature subset will be used as the input for subsequent modeling and analysis tasks.

3.4.4 Elastic Net

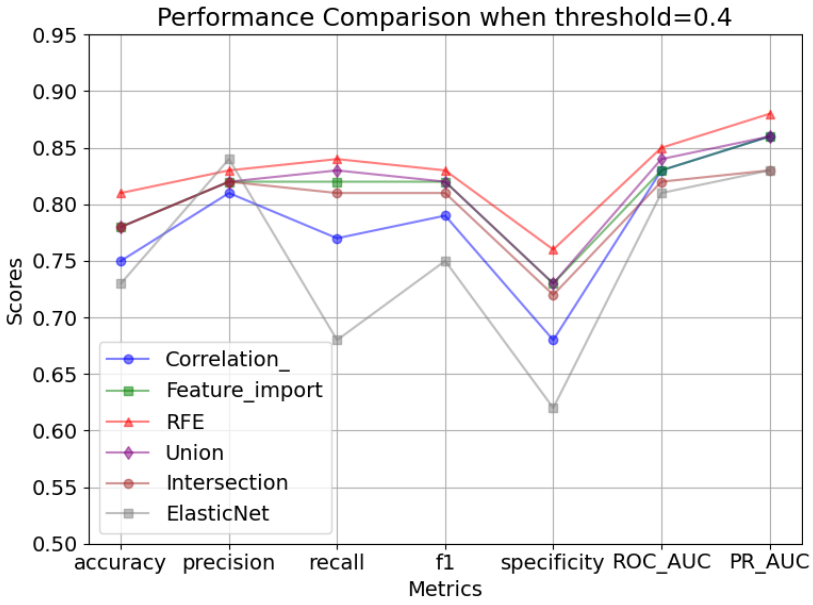
The Elastic Net method serves as both a feature selection and regularization technique, combining L1 (Lasso) and L2 (Ridge) regularization within the regression. By adjusting the alpha hyper-parameter—used to control the level of regularization—between 0 (Ridge) and 1 (Lasso), the Elastic Net adapts to achieve the desired degree of feature selection [46][47].

In the context of binary classification problems, the embedded feature selection procedure of Elastic Net involves an iterative process that minimizes the logistic loss function alongside L1 and L2 regularization terms. This iterative process leads the model to recognize and emphasize significant features while reducing the influence of less relevant ones. Over the course of optimization iterations, the L1 penalty effectively pushes certain coefficients to zero, facilitating feature elimination [47]. The selected features are those associated with non-zero coefficients, indicating their significance in predicting binary outcomes. After training, the model makes predictions by applying a threshold to the predicted probabilities.

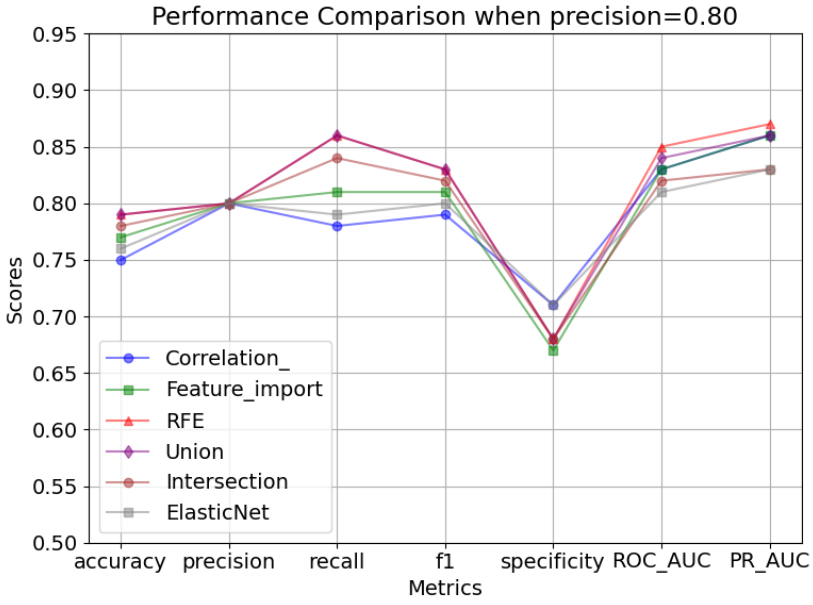
3.4.5 Hybrid Approach

After employing the three distinct non-embedded selection techniques, we obtained three unique sets of chosen features. To enable more potentially optimal feature combinations, we adopted an ensemble approach that intersects or unites these three sets. This process yields two new hybrid feature sets known as the "intersection" and "union" methods. Generating both the intersection and union feature sets holds several possible advantages. The ensemble strategy enhances stability by mitigating sensitivity to small data variations. Moreover, it reduces the risk of overfitting by restraining the complexity of the model. However, this approach should be carefully assessed for its appropriateness to the specific problem and dataset. We hereby compared the prediction performance of models that use the previously-mentioned feature selection methods respectively, as shown in Figure 3.7.

The comparison depicted in 3.7 presents the classification metrics observed during a LOOCV. As per the outcomes, when we established a threshold of 0.4 for distinguishing binary outcomes, the RFE method demonstrates the most favorable overall



(a) When the threshold of the classification is 0.4.



(b) The precision is fixed as 0.80 when varying the threshold of classification.

Figure 3.7: Comparison of performance evaluation metrics across multiple feature selection approaches. The initial five models utilized the XGBoost algorithm, whereas the last one model employed Elastic Net.

performance. On the other hand, when we maintained a precision of 80%, both the union method and RFE exhibited commendable results. Hence, within the context of our dataset and the specific problem configuration, the model employing the RFE method stood out as the top performer. Consequently, our subsequent analyses were conducted based on the feature set selected by the RFE method. Nevertheless, it is worth highlighting that these results could potentially be influenced by variables such as the validation methodologies employed and the choices made for hyper-parameters.

3.5 Features Analysis

In this section, we will summarize the explanation of the selected features obtained through the above approaches. Our expectation was that these explanations would strongly support the selection process and demonstrate the reasonability of the model.

- **Demographic group**: Among all the demographic features, age ranked first as a significant predictor of PIH. Age is highly related to patients' cardiovascular health. Older patients, for example, tend to have reduced vascular compliance and decreased baroreceptor sensitivity, making them more susceptible to fluctuations in BP during anesthesia induction. Additionally, age-related comorbidities, such as atherosclerosis or hypertension, can further exacerbate the risk of PIH. Other demographic factors like height, weight, and BMI also held importance. These variables can impact drug distribution and metabolism during anesthesia. They are also required input for a TCI system to calculate dosage plans.
- **Medical history**: Only the medical history of preoperative hypertension mattered here. Preoperative hypertension can cause changes in the cardiovascular system, such as increased vascular resistance and decreased vascular compliance. Hypertension also indicates that the patient already has elevated baseline BP, which the body's compensatory mechanisms may regulate. Anesthesia induction and surgical stress can disrupt these compensatory mechanisms, leading to a drop in BP, especially in patients with poorly controlled hypertension. Besides, patients with preoperative hypertension may be on antihypertensive medications. The interaction between these medications and anesthesia drugs can influence BP regulation.
- **BP group**: The BP features were driven from all three waveforms (SBP, MAP, and DBP). The "first" feature, namely the first value of a recording, represents the baseline BP, which is used to determine the PIH. In other literature that used a single threshold of BP as the determinant of PIH, the first value of BP usually dominated the prediction [16]. They were very intuitive. The setting of a single threshold means that their determination of PIH never considers the basic condition of the patient. That means a patient with a higher baseline MAP will, of course, be less likely to undergo a PIH, even if the patient has already suffered a sudden steep drop in BP because her or his MAP is still more than 65 mmHg. Such influence is more several when prediction is done in a small database, which

is also the case in our project. Additionally, the delta (change over time) and standard deviation features of BP reflect the patient's current hemodynamic system stability. Large fluctuations in BP or higher standard deviation may suggest impaired cardiovascular compensatory mechanisms and increase the likelihood of PIH.

- **SI group**: The SI, calculated as the ratio of heart rate to SBP, and its derived features are strong predictors of hypotension during anesthesia induction. SI has been validated in clinical settings as a reliable indicator of hypotension, as mentioned in chapter 2. An elevated SI suggests an imbalance between oxygen demand and supply to the tissues, which can lead to hypotension and other adverse events.

- **Medication group**: This group includes vasoconstrictors and anesthetics, both of which play significant roles in PIH. Vasoconstrictors lead to an increase in vascular resistance and BP. On the other hand, anesthetics, namely propofol and remifentanil, can lead to vasodilation and reduced vascular resistance. When the body's compensatory mechanisms fail to maintain BP, hypotension may occur.

- **Heart rate group**: Anesthesia induction can cause a drop in sympathetic outflow, leading to vasodilation and reduced vascular resistance, resulting in PIH. In response, the body may activate the baroreceptor reflex, leading to an increase in heart rate as an attempt to compensate for the decreased BP and maintain cardiac output [15].

- **ECG group**: During perioperative periods, patients may experience fluctuations in BP, including episodes of hypotension. Hypotension can lead to reduced coronary blood flow, compromising oxygen supply to the heart muscle, and subsequently manifest as ST segment changes on the ECG. Therefore, continuous monitoring of the ST segment can be an invaluable tool in the detection of PIH and its potential adverse effects on cardiac function. However, it is important to note that while ST segment changes can raise suspicion for PIH, they are not specific to this condition and can also be influenced by other factors, such as myocardial ischemia or medication effects.

- **Oxygen Saturation**: SpO2 (Oxygen Saturation) is not a predictor of hypotension; rather, it is a monitoring factor. Hypotension can lead to reduced perfusion to various organs, including the lungs. As a result, the oxygen saturation levels may drop, indicating inadequate tissue oxygenation. Continuous monitoring of SPO2 allows prompt detection of hypoxemia, enabling timely intervention to prevent adverse outcomes.

- **Other combinational features**: The ratio of propofol to weight is an additional important predictor in PIH. Propofol is commonly used for anesthesia induction, and its dosage is typically calculated based on the patient's weight to ensure safe and effective administration. However, individuals with different body weights may metabolize, distribute, and clear drugs at varying rates. Therefore, considering the propofol-to-weight ratio can improve the accuracy of hypotension

prediction during anesthesia induction. Traditionally, a dosage ratio of 1.5 to 2.5 is selected, with the specific value varying depending on the patient's health condition and individual characteristics.

3.6 Summary

Chapter 3 dealt with data preparation for analysis and modeling. It began with a discussion of the primary outcome (Section 3.1) and moved on to data collection and processing (Section 3.2). Feature extraction (Section 3.3) was explored, followed by feature selection (Section 3.4), which included methods like correlation analysis, feature importance, recursive feature elimination, Elastic Net, and a hybrid approach. Section 3.5 analyzed the selected features to gain insights into the relationships of features with the target variable. This chapter formed the foundation for building accurate and effective predictive models.

4 Model Development

At this stage, we have gathered all the essential components required to train a machine-learning model and enable it to generate predictions. However, as with any ML application, several crucial tasks need to be addressed to ensure the development of a high-quality model. In this chapter, we comprehensively discuss our strategies for addressing data imbalance, the tuning tool to optimize the model, and the implementation of ensemble learning to fit an explainable model for the decision-suggestion of anesthetics.

4.1 Dealing with Dataset Imbalance

Dataset imbalance refers to a situation where the distribution of the dataset will tilt to certain classes. The severity of the imbalance can vary, depending on the specific problem and dataset. Generally speaking, in a binary classification, when the minority class only takes less than 40% of the whole dataset could lead to a mild imbalance problem and less than 20% to a moderate one [48]. Therefore a slight imbalance is present in our dataset. It is important to address this issue since PIH or hypotension data suffer from data imbalance in most cases, although our project is less vulnerable due to the relatively small dataset size. To tackle dataset imbalance, there are two directions of solution:

1. Assigning different weights to individual classes during training.
   It is commonly known as weighting training. By assigning higher weights to the minority class and lower weights to the majority class, the model will thus pay more attention to the minority class, mitigating the impact of dataset imbalance.

2. Under-sampling the major class or over-sampling the minor class.
   Under-sampling that removing some cases from the majority class will effectively reduce its dominance in the dataset, allowing the model to pay equal attention to classes during training. On the other hand, over-sampling will increase cases of minority classes to improve their representation in the dataset. Duplicating some samples, called naive random over-sampling, is one approach to doing so, but it may also lead to over-fitting problems. Another way is using interpolation instead of duplication. Synthetic samples through algorithms such as SMOTE (Synthetic Minority Over-sampling Technique) [49], ADASYN (Adaptive Synthetic Sampling) [50], etc. They differ in the selection of samples for interpolation.

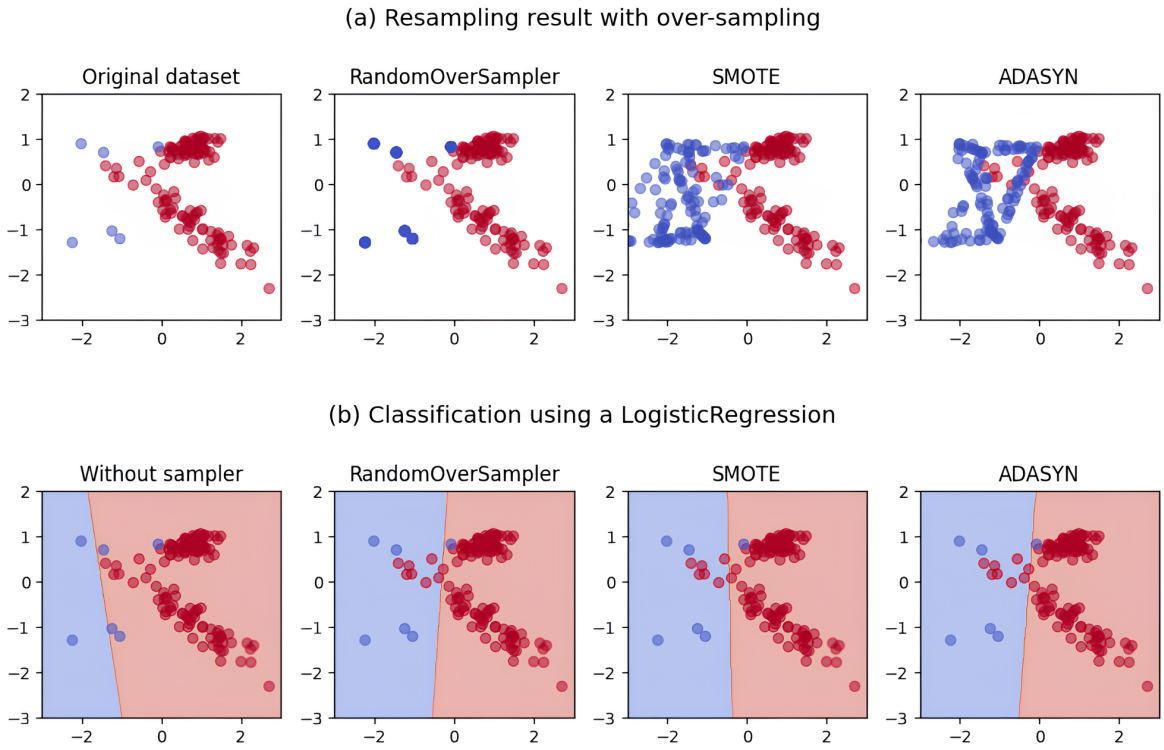


Figure 4.1: An example imbalance dataset with different oversampling approaches, generating with the 0.11 version of imbalanced-learn library [1]. The imbalance ratio here is 0.05:0.95.

SMOTE randomly selects a minority class instance and finds its k nearest neighbors. Next, it generates new samples by connecting the selected instance to its neighbors in the feature space, effectively expanding the minority class. SMOTE assumes that all minority class samples are equally important and does not consider the distribution of the minority class in the feature space, which may lead to an unrealistic or irrelevant generation.

ADASYN attempts to create samples in regions where the class imbalance is more pronounced. It assigns a weight to each sample based on the density distribution of classes and then randomly selects a minority instance and determines the number of synthetic samples to be generated based on the weight of sample.

However, it should be noted that oversampling technique does not always guarantee improved performance. As previously discussed, the algorithm relies on the generated synthetic samples to make accurate inferences. When working with insufficient data, oversampling can potentially introduce additional noise during the training process. When applying the oversampling techniques using the XGBoost model, the SMOTE approach improved the average accuracy by approximately 1%, while ADASYN showed an improvement of less than 1%. Similarly, in the Random Forest model, the improvement was around 2%, which, although not significant, indicates a modest enhancement. Nonetheless, despite the marginal improvements and potential noise risk, the utiliza-

tion of oversampling techniques is still promising for future research and may yield significant improvements when applied to larger and more diverse datasets.

4.2 Hyper-parameter Tuning

During the ML model training process, certain parameters are automatically learned from the data and learning algorithm, representing the model's behavior. These are known as model parameters. In contrast, model hyper-parameters are external settings that define the structural limitations of the model and are set before training begins. They include variables like learning rate, iteration times, depth of trees, and regularization strength. Selecting appropriate hyper-parameters is crucial as they significantly influence the performance of the model, balancing computational complexity, accuracy, and generalization to new data. To accomplish this, nested cross-validation was used, which involves dividing the data into multiple folds. In each iteration of the outer loop, one partition serves as the validation set, while the rest are used for training. Within each outer loop iteration, the inner loop of cross-validation is performed, further dividing the training set into training and validation subsets. It is within this inner loop that feature selections are performed and evaluated.

For hyperparameter tuning within the inner loop, the Python package Hyperopt [51] is employed. Hyperopt is a powerful and distributed asynchronous hyperparameter optimization tool that utilizes the Bayesian optimization algorithm. This approach allows for a more efficient hyper-parameter space search than traditional grid search methods. Hyperopt also leverages Tree of Parzen Estimators (TPE) algorithms, which enhance computation speed and effectiveness [51]. By using Hyperopt within each inner loop of nested cross-validation, researchers can systematically explore the hyperparameter space and identify optimal combinations that yield the best performance for the model on the given task. This approach is more efficient and effective, especially when dealing with large datasets or complex models, as it avoids the need to exhaustively try out all possible hyperparameter combinations through grid search.

4.3 Ensemble Learning

The introduction of ensemble learning [52] into our pipeline aims to address the following problems:

1. **Explainability:** Explainability is particularly important in our work to explore the relationship between anesthesia drugs and risks. Models that are highly explainable, such as decision trees and LR, provide intuitive interpretations of input factors as coefficients or weights. This not only enhances the credibility of the model for medical applications but also facilitates a clearer understanding of the associations between variables and outcomes due to their simpler structures. XGBoost, being an ensemble of decision trees, is generally less interpretable compared to simpler models like LR, which directly models the linear relationship between features and the target variable. By combining XGBoost with an LR model, you

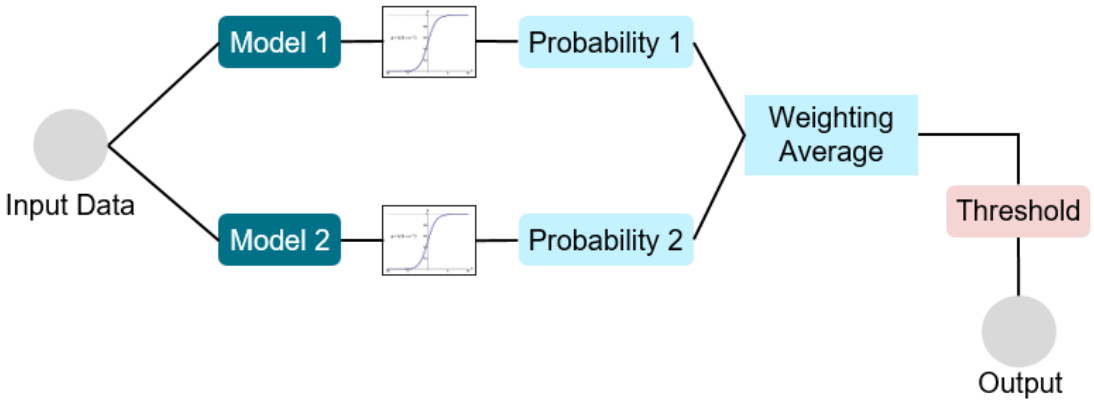


Figure 4.2: Illustration of ensemble learning with an averaging strategy.

can retain some level of interpretability from the LR part, as it allows you to understand the impact of each feature on the predictions more explicitly.

2. **Overfitting Mitigation:** By incorporating an LR model in the ensemble, which inherently has a more straightforward and linear structure, the combined model can become more resistant to overfitting. The LR model is less likely to overfit due to its simplicity and its focus on the linear relationship between features and the target variable. The ensemble leverages the strengths of both models, striking a balance between complexity and simplicity. XGBoost can capture intricate relationships and non-linear interactions, while LR helps provide a smoother approximation of the true underlying patterns. This enhances the generalization performance and makes the ensemble more capable of handling unseen data.

3. **Improvement of Accuracy:** Ensemble learning can also lead to improved prediction accuracy. By combining multiple models with different strengths and weaknesses, the ensemble can outperform individual models, especially when the base models are diverse and complementary to each other. This can be particularly valuable in our work, where accurate predictions regarding anesthesia drug risks are crucial for medical decision-making.

Ensemble methods can be categorized into two main groups: averaging and boosting. Boosting methods, including Gradient Boosting [53] and XGBoost, build models in a stage-wise manner, minimizing the loss function of the previous residuals of models. Averaging methods, such as Bagging and Random Forests, train multiple base models independently, and their predictions are combined through averaging or majority voting. In our approach, we are currently averaging the outputs of the XGBoost model and LR in proportion. Since both models produce probability values between 0 and 1, their weighted average can yield a new probability. By applying a certain threshold to this probability, we can obtain a new binary classification result. Ensemble methods can also take into account the importance of different features in making predictions. For example, if the propofol dosage is considered highly significant,

an ensemble model can assign more weight or emphasis to the predictions of models that have demonstrated a good understanding of the dosage-result relationship.

5 Results

In this chapter, the focus is on evaluating the algorithm's performance. The evaluation begins with an analysis of the binary classification metrics, followed by the implementation of cross-validation techniques. Additionally, the dosage advice model is introduced. This chapter aims to provide a comprehensive understanding of the algorithm's effectiveness and its potential application in practical scenarios.

5.1 Evaluation

We employed leave-one-out validation, a particular case of k-fold cross-validation, which can better estimate model performance based on the small-size dataset. Besides, we also applied nested cross-validation in order to include hyper-parameter optimization inside the cross-validation.

5.1.1 Evaluation of Binary Classifier

To evaluate the performance of a binary classifier, various scores were calculated, each with its own preferences in different areas. These scores provide insights into different aspects of the classifier's performance. In this work, we aimed to provide an overall evaluation of the performance by reporting popular scores, namely precision, sensitivity, specificity, accuracy, and the F1 score. We illustrate the classification result metrics along with these scores in Figure 5.1.

1. Precision: Precision measures the proportion of correctly predicted positive instances out of the total instances predicted as positive. It assesses the model's ability to avoid false positives. A classifier aiming at predicting positive outcomes may concentrate more on precision.

2. Sensitivity (or recall and true positive rate): Sensitivity calculates the percentage of accurately predicted positive instances among all actual positive instances. It evaluates the model's ability to identify positive cases correctly. A high sensitivity indicates that the model is effective at capturing positive instances.

3. Specificity (or true negative rate): Specificity measures the proportion of correctly predicted negative instances out of the total actual negative instances. It evaluates the model's ability to identify negative cases correctly. Opposite to precision, specificity may be more important to the negative-aiming classifier.

4. Accuracy: Accuracy measures the overall correctness of the classifier by calculating the proportion of correctly predicted instances (both positive and negative) out of the total instances. It provides a general assessment of the model's performance.

5. F1 score: The F1 score is a harmonic mean of precision and sensitivity. It provides a balanced assessment of the model's performance by considering both false positives and false negatives. The F1 score is particularly useful when dealing with imbalanced datasets.

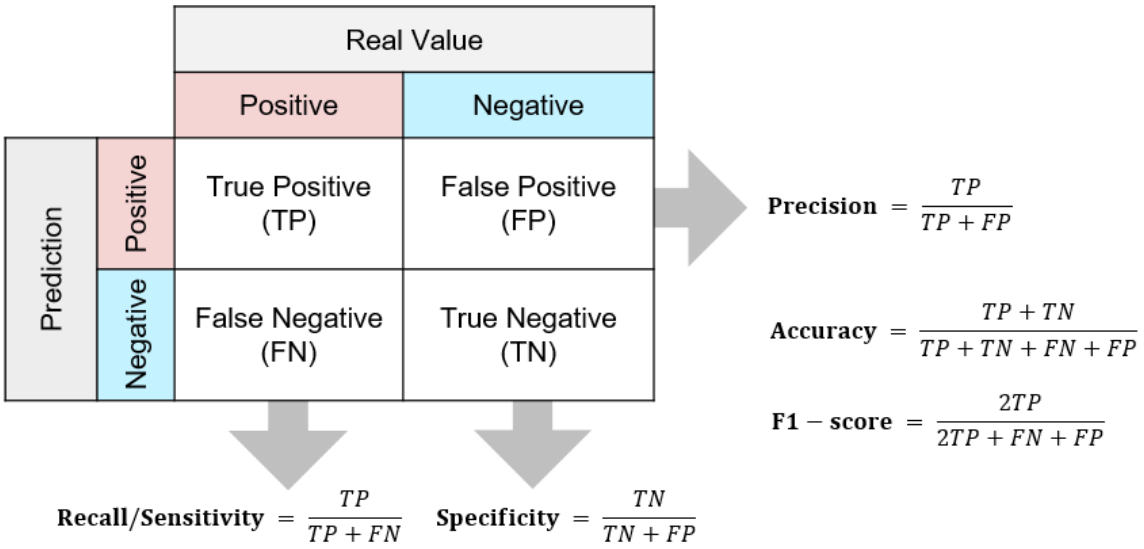


Figure 5.1: Evaluation metrics of the binary classification problem.

In addition to the previously mentioned scores, two other commonly used metrics are the PR curve (Precision-Recall curve) and the ROC curve (Receiver Operating Characteristic curve). Including these curves will provide a comprehensive evaluation of the classifier's performance.

1. PR Curve: The PR curve represents the precision-recall trade-off of the classifier at different probability thresholds. It plots precision (positive predictive value) on the y-axis against recall (sensitivity) on the x-axis. The PR curve provides insights into the classifier's ability to maintain high precision while capturing positive instances. A curve that is closer to the top-right corner indicates better performance.

2. ROC Curve: Similarly, the ROC curve shows the true positive rate (sensitivity) against the false positive rate (1 - specificity) at various thresholds. It shows how well the classifier distinguishes between positive and negative instances. The ROC curve enables the analysis of the trade-off between sensitivity and specificity. ROC AUC (Receiver Operating Characteristic Area Under the Curve) is a famous measurement in statistics, with values closer to 1 indicating better performance.

Although a classification threshold of 50% may seem intuitive for determining the possibility of an event happening, the interpretation changes when it is viewed as a warning for risk. In the context of risk assessment, however, instead of merely predicting an event, we view the output as a warning sign indicating the likelihood of a risk. For example, the HPI, proposed by Hatib et al. in their work on ML for risk prediction [24], suggests varying the threshold depending on the specific application. The researchers found that the optimal threshold for predicting hypotension differed based on the time horizon. For instance, when predicting hypotension five minutes ahead, the best-performing threshold was identified as 41. However, for a 15-minute prediction horizon, the optimal threshold was 36. These findings highlighted the importance of discussing the threshold choice instead of treating it solely as an evaluation score. The selection of an appropriate threshold can significantly impact the performance and effectiveness of the risk prediction system.

To visually represent the trade-offs associated with threshold adjustment in binary classification, we present Figure 5.2. The graph illustrates how various performance metrics change as the classification threshold is varied. It is important to note that we did not assert the existence of an optimal threshold since the model's performance can be influenced by factors such as class distribution, feature selection, or hyperparameter tuning. The threshold selection process requires careful consideration of overall performance metrics and the underlying data characteristics. Considering our concern about positive hypotension risk and the evaluation metrics, a threshold between 0.4 and 0.55 appears to be reasonable, as it achieves a balanced and high performance at these levels. In the subsequent sections of this chapter, we will also present a comparison of setting the threshold at 0.4 and 0.5.

5.1.2 Cross-validation

When dealing with a limited dataset, k-fold cross-validation (CV) is a good approach to assess the overall performance of the prediction model. The whole dataset is randomly or proportionally divided into k folds of approximately equal size. In one validation, $k-1$ folds are used for training, with the remaining fold for validation. Such validation iterates k times on each different sub-fold. This strategy promises that the validation considers each component of the dataset. The final performance is calculated on the average of every iteration. A stratified k-fold cross-validator is an edited improvement of the k-fold that it will make sure all folds will contain the approximately same distribution of every class. An illustration of the k-fold CV approach is provided in Figure 5.3.

The selection of k relies on comprehensive factors, such as the size of the dataset, computational resources, and the desired trade-off between bias and variance. Generally, a higher k value (larger than 10) lead to a more reliable estimate of the model. However, the dataset is small enough to better take the leave-one-out cross-validation (LOOCV) into consideration. LOOCV is a special case of k-fold cross-validation where k equals the size of the whole training set. In this approach, we trained the model on all but one sample, and the left-out sample was used for testing. This process was

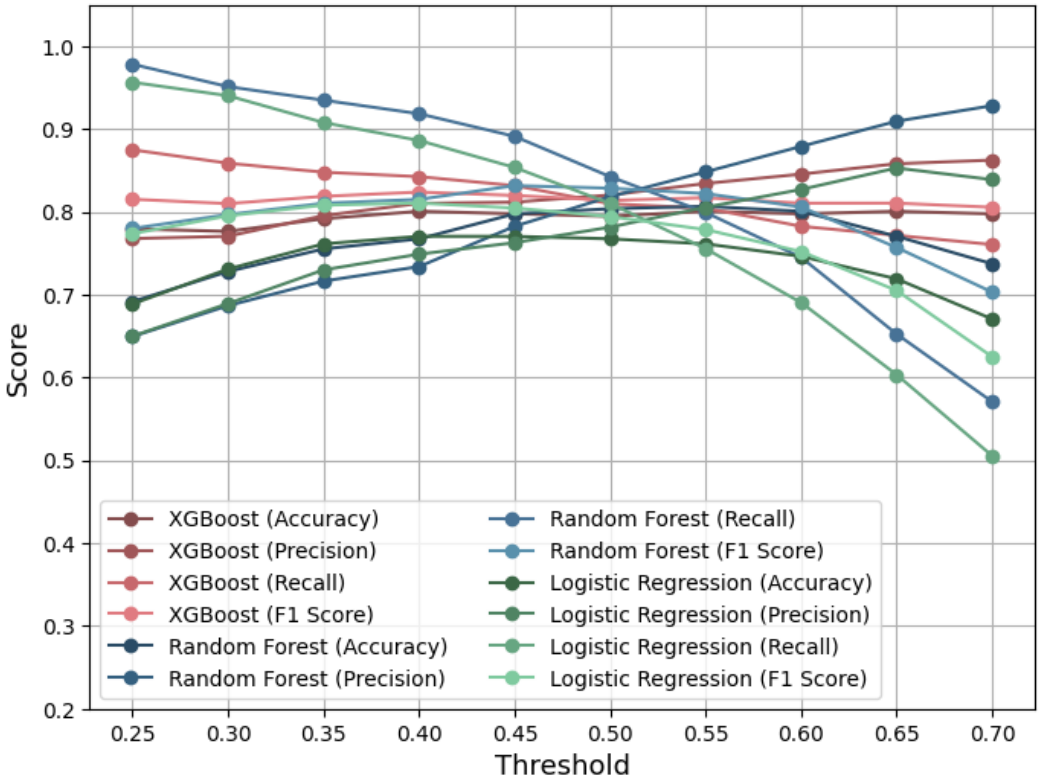


Figure 5.2: The illustration of how the decision threshold of binary result influence evaluation scores.

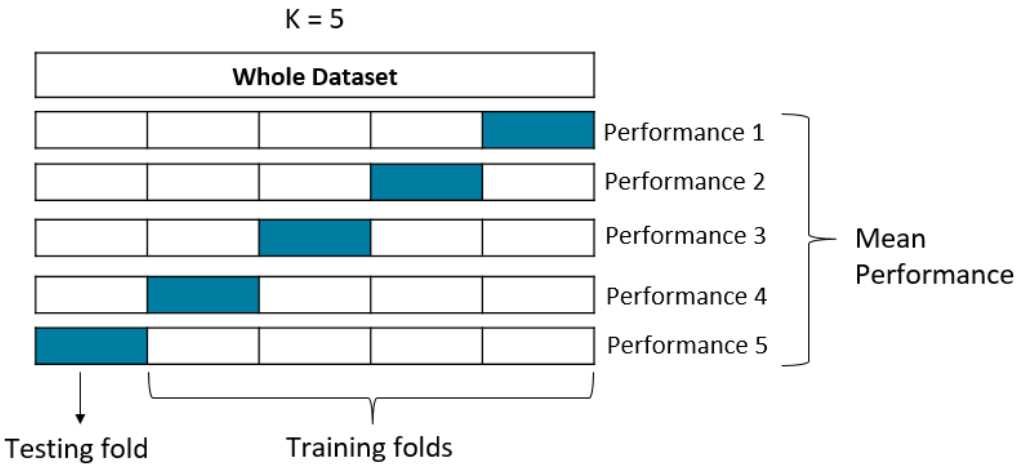


Figure 5.3: K-fold CV illustration.

repeated for each sample in the dataset, ensuring that every sample had a chance to be tested individually.

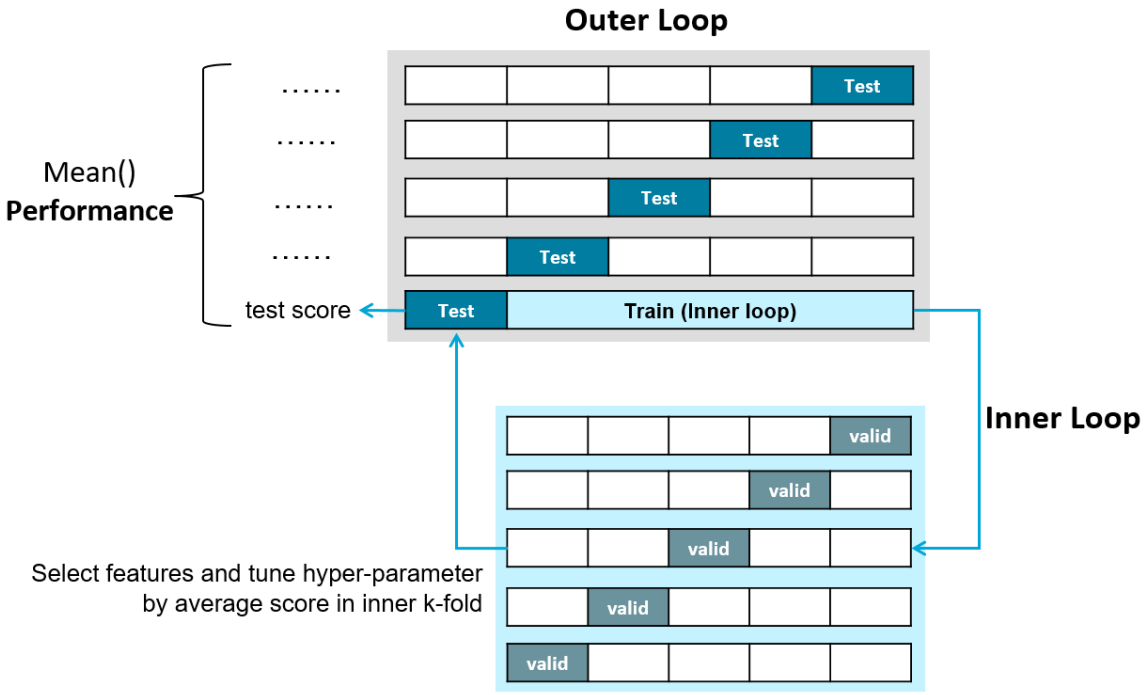


Figure 5.4: Neated k-fold cross-validation approach.

The average performance across all folds in k-fold CV provide an estimate of the model's overall performance, while LOOCV helped to assess the model's performance on individual samples, taking into account dataset imbalance. The overall results of the LOOCV approach with different classification thresholds are presented in Table Table 5.1.

5.1.3 Nested Cross Validation

In the classical k-fold CV way, the model is trained and evaluated within a single cross-validation loop. However, literature has pointed out that the classical cross-validation (usually k-fold CV) tunes the model on the same data causing overfitting and obtaining overly optimistic performance estimates [54]. It thus proposes the nested cross-validation approach, which is claimed to have near accuracy as the truth. Nested cross-validation introduces the outer and inner loops into the structure. The outer loop divides the data into training and testing sets, as in the k-fold CV we mentioned before. The inner loop, which is nested within each outer loop iteration, is responsible for training and evaluating the model with a specific feature selection and a hyper-parameter combination. Then in each iteration of the outer loop, the model is trained on the training set using the optimal hyper-parameter combination generated by inner loops and then evaluated on the corresponding test set. The outer loop is responsible for assessing the performance of the model with the selected hyper-parameters. It provides the training data for the inner loop and keeps a separate

test set to evaluate the performance of the model obtained from the inner loop. By using this approach, information leakage is prevented, and a more unbiased estimate of the model's performance can be obtained. This process is illustrated in the Figure 5.4.

The nested cross-validation is not intuitive because it generates k sets of hyper-parameters (and may also include feature selection) during the validation process. A simpler way to understand it is that the nested approach treats cross-validation as an evaluation of the procedure or training process rather than only evaluating the model itself. Some tutorials suggest setting aside a separate testing set before performing cross-validation, but this can introduce new bias into the methodology.

An alternative approach to validation could be to train the model on the entire dataset and introduce an external validation set during the evaluation process. By incorporating an external validation set that is independent of the training and testing data, we can gain a more comprehensive understanding of the model's performance in real-world scenarios while making full use of data collected for training. However, currently, we do not have an external validation set available. Therefore, we have made the decision to keep this question open for future research.

5.1.4 Result

Table 5.1 and Figure 5.5 present a summary of our predictive model's performance under different validation approaches.

Firstly, we compared the performance of models under different classification threshold settings. Notably, the XGBoost model demonstrated superior performance when the threshold was set to 0.4. Specifically, it achieved an accuracy of 0.81, a precision of 0.83, and a recall of 0.84. These metrics are crucial since the positive PIH event's correct detection is essential for an effective risk assessment project. Furthermore, all models exhibited high AUC values of PR curves, indicating their excellent ability to distinguish the positive class. The theoretical strength of XGBoost aligns with our experimental results, owing to its powerful ensemble strategy and boosting structure. However, in some cases during k-fold validation, XGBoost's superiority is compromised. Several reasons could explain this phenomenon, such as misclassified data affecting boosting algorithms and the susceptibility of the model to overfitting. In contrast, simpler LR performs better in such scenarios. Additionally, Random Forest's averaging strategy is more adept at handling noise. Another factor that affects the predictive ability of XGBoost is its intricate parameter tuning. Moreover, the choice of threshold slightly impacted the model's performance. Although the differences were not significant between thresholds of 0.4 and 0.5, medical considerations should guide threshold selection rather than solely relying on the best metrics.

Secondly, we also tested a k-fold CV with the results from LOOCV. It showed that LOOCV consistently outperformed the latter in our small-sized model validation. This suggests that LOOCV benefits from more training data, exhibits lower variance, and

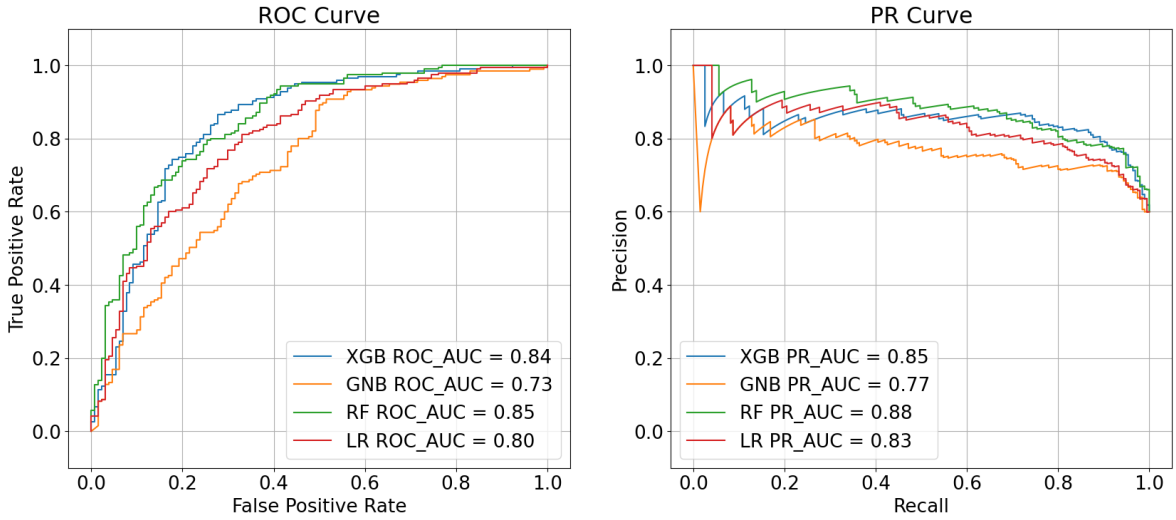
Table 5.1: Performance of the models of PIH prediction under different settings.

Model	Threshold	Accuracy	Precision	Recall	f1-score	Specificity
Frequent-LOO						
XGBoost	40%	0.81	0.83	0.84	0.83	0.79
Random Forest		0.78	0.74	0.91	0.82	0.85
Logistic Regression		0.77	0.75	0.88	0.81	0.81
Naive Bayesian		0.71	0.82	0.61	0.70	0.63
Ensemble		0.79	0.77	0.89	0.83	0.83
XGBoost	50%	0.80	0.81	0.83	0.82	0.78
Random Forest		0.79	0.79	0.85	0.82	0.82
Logistic Regression		0.75	0.77	0.77	0.77	0.72
Naive Bayesian		0.70	0.82	0.58	0.68	0.62
Ensemble		0.79	0.87	0.73	0.80	0.72
Frequent-Kfold						
XGBoost	40%	0.76	0.73	0.91	0.81	0.83
Random Forest		0.79	0.75	0.93	0.83	0.87
Logistic Regression		0.77	0.74	0.90	0.81	0.83
Naive Bayesian		0.70	0.80	0.61	0.69	0.62
Intermittent-LOO						
XGBoost	50%	0.78	0.77	0.76	0.76	0.75
RF intermittent		0.75	0.75	0.75	0.75	0.72
LR intermittent		0.76	0.77	0.74	0.76	0.72

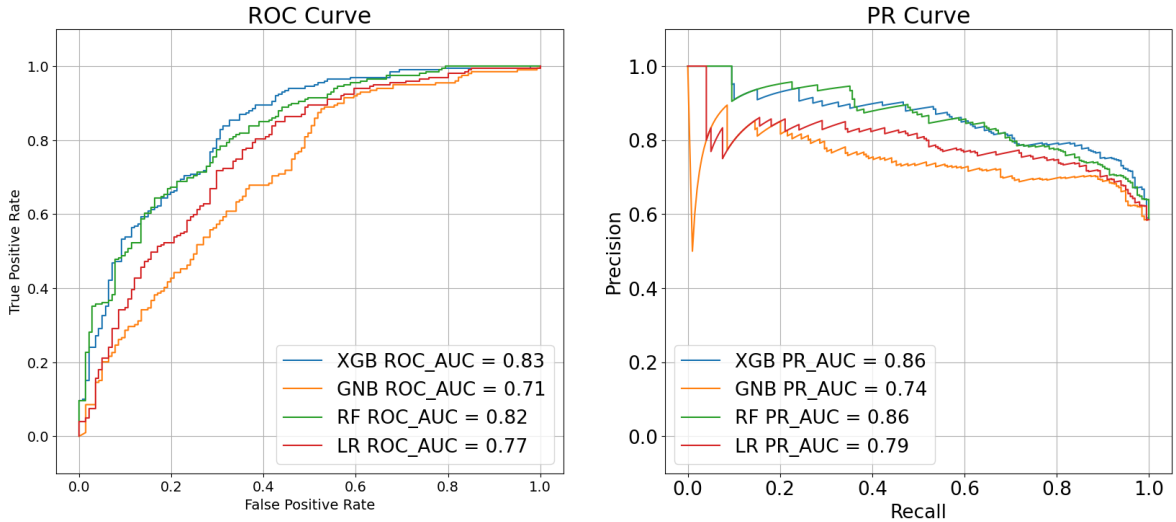
may handle imbalanced datasets better. The lower variance in LOOCV is attributed to its training set remaining consistent across iterations. Nevertheless, LOOCV's reliance on a large training set could make it computationally expensive and prone to capturing dataset noise, resulting in over-optimistic outcomes.

Thirdly, while ensemble learning generally improved performance, it did not guarantee enhanced results in all cases. We performed ensemble learning on models that used frequent BP measurements in order to balance between accuracy and robustness by combining XGBoost and LR. In the 40%-threshold experiment, while this improved performance in recall, f1-score, and specificity, it did not always lead to improvements. Factors such as metric sensitivity, model disagreement, and loss of model-specific information could result in decreased performance.

Finally, our experiment comparing intermittent and frequent BP measurements for predicting PIH revealed that intermittent data did not perform as accurately as frequent data. The intermittent approach decreased the precision of all models by less than 6% and reduced recall (sensitivity) by more than 10%. This indicates that less frequent measurements impair the prediction's ability to detect positive events,



(a) Performance of the model built on frequent (every 2 seconds) BP measurement.



(b) Performance of the model built on intermittent (every 3 minutes) BP measurement.

Figure 5.5: ROC and PR curve as well as their AUC of different models.

which is understandable since intermittent data are less representative. However, using intermittent measurements ensures practicality and safety in a clinical setting, even though they may not be as effective for the model as frequent measurements.

5.2 Dosage Advice Model

The dosage advice model aims to predict the risk of PIH based on changes in input features related to propofol dosage. The ultimate goal is to develop a reliable and safe propofol dosage warning system. This section explores different modeling approaches

and their implications for predicting PIH risk.

Initially, an LR model is considered practical due to its ability to provide valuable insights into the relationship between features and outcomes. The LR output serves as the foundation for building the prediction model. But it is clear that the success of this approach will rely heavily on the accuracy of the prediction model. While XGBoost and other decision tree models boast high accuracy, they may lack the ability to present a relationship between individual input and the outcome, which LR offers. A comparative analysis, as depicted in Figure 5.6, reveals that XGBoost exhibits some instability despite accurately distributing testing samples. This instability may be attributed to overfitting, particularly when dealing with limited data. In some instances, as the dosage increases, the likelihood of a PIH event decreases, seemingly contrary to established medical knowledge. Conversely, LR consistently demonstrates a positive relationship between dose-related features and the prediction value, aligning with the medical understanding. To address these issues, ensemble learning is introduced to strike a balance between the performance of LR and XGBoost. The ensemble model, as shown in the third column of Figure 5.6, corrects the negative slope observed in some cases of the XGBoost model and marginally improves accuracy compared to standalone LR as in Table 5.1.

The model also considers the impact of another anesthetic used in the TCI system, namely remifentanil. However, the combined effect of propofol and remifentanil is not simply additive [55], introducing further complexities in the model. Although the feature related to remifentanil did not play a significant role in the model compared to propofol, as shown in the figures, this observation may be influenced by potential biases in the distribution of the anesthesia plan.

In conclusion, the dosage advice model built through our approach holds promise for predicting the risk of PIH by leveraging LR, XGBoost, and ensemble learning. Further refinements and adjustments may be necessary to address the complexities arising from multiple anesthetics interactions and potential biases. Nonetheless, this model represents a step forward in developing a robust and reliable propofol dosage warning system to enhance patient safety during medical procedures.

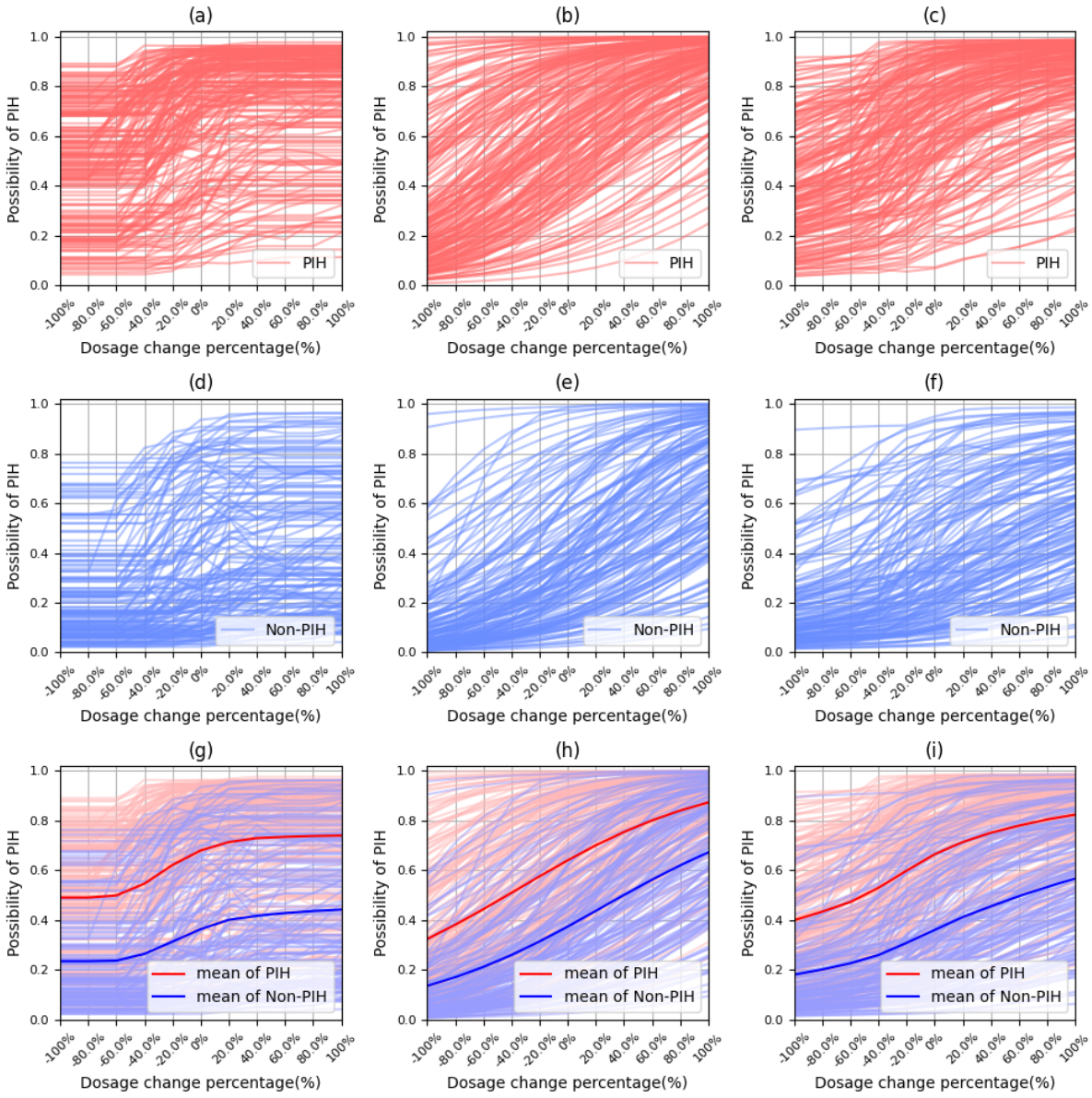


Figure 5.6: The change of PIH probabilities corresponding to the change of propofol dosage. Each line in the plot presents an individual test case in a LOOCV. Along the x-axis, "0%" represents no change in the propofol dosage-related features, which is the prediction made by original data, while "100%" means all the dosage-related features are risen to 200% compared to the ground truth.

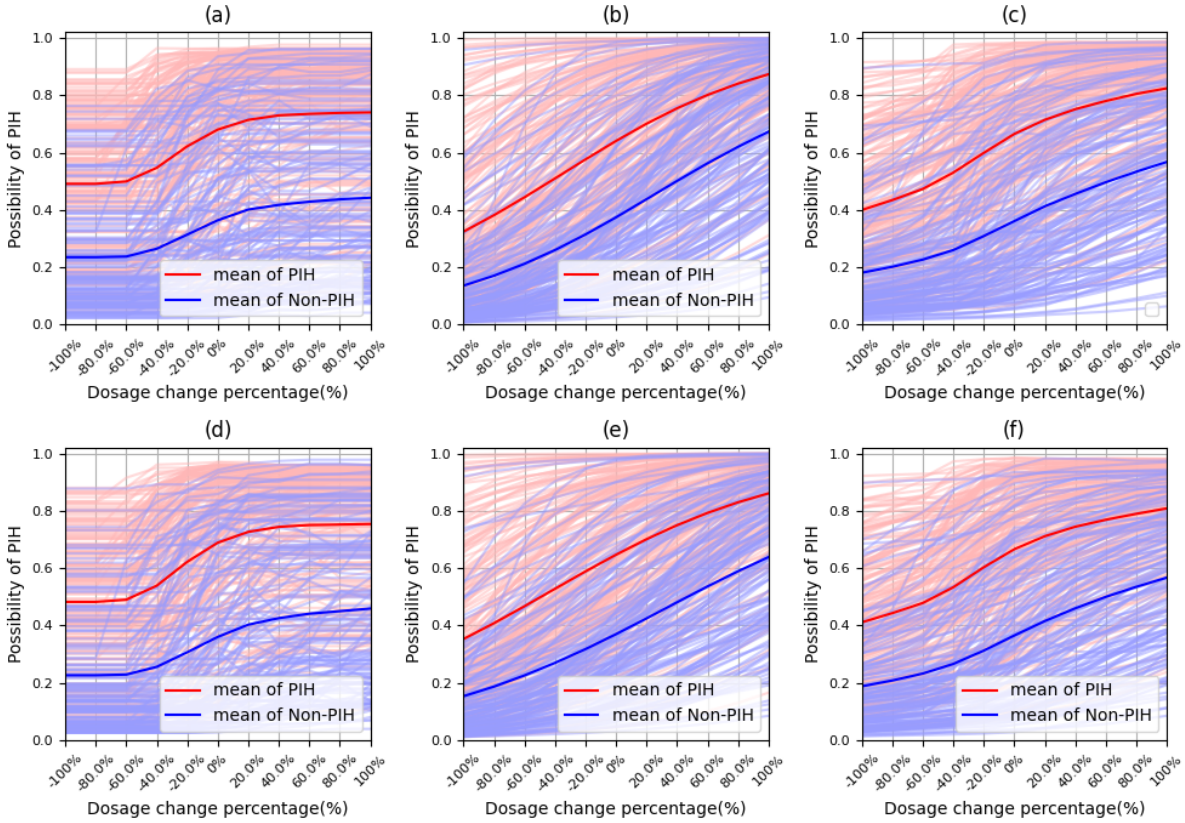


Figure 5.7: PIH probability v.s. propofol and remifentanil dosage. Along the x-axis, "0%" represents no change in the propofol and remifentanil dosage, while "100%" means both of them have risen to 200%.

6 Conclusion and Future Work

6.1 Conclusion

In this work, we investigated pre-operative demographic data, vital signs, and medication information to predict PIH. Compared to previous PIH studies, our data-driven approach has been validated using nested cross-validation, which incorporates feature selection and hyper-parameter optimization into the evaluation process. This enhancement not only improves the performance of complex models but also prevents data leakage. Additionally, we have employed a leave-one-out validation technique to maximize the utility of our limited dataset. By utilizing ensemble learning, the final model effectively balances explainability and prediction accuracy. Consequently, we have developed an intuitive dosage-risk model that suggests a safe dosage of propofol to prevent PIH before induction. Importantly, it is the first attempt to combine probability prediction of hypotension with pre-operative input of anesthetics dosage to aid decision-making in anesthesia plans.

6.2 Limitation

This section provides an overview of the challenges that remain to be addressed within the scope of our project, along with potential resolutions and future works. Our examination of these challenges is approached from three angles: a data-driven perspective, an ML aspect, and a medical model-building aspect.

6.2.1 Data Limitation

As previously discussed, PIH studies suffer from limitations on experimental data in terms of length, type, fidelity, and size. The main reasons behind these limitations are the historical neglect of data-driven methods in PIH research and the scarcity of datasets that meet specific requirements. Furthermore, the lack of data sharing among researchers and organizations contributes to the clustering of studies within certain groups.

To address these challenges and improve future research in PIH, we propose the following suggestions:

1. **Increase Database Availability:** One way to overcome data limitation is by encouraging the creation and sharing of more databases. A practical approach to research could involve collaborating with existing databases like VitalDB and

combining them with private databases within research organizations. For example, our project is collaborating with EMC to access its more extensive and diverse dataset in the future.

2. **Enhance Data Collection Methods:** To obtain more comprehensive data, it is essential to improve data collection methods. Despite the medical history and demographic information, additional data from anesthesia machines and other clinical operations which provide valuable insights are hardly recorded. While automation has limits in medical settings due to the importance of experience-based diagnosis, finding more flexible and efficient ways to collect data beyond vital signs can be beneficial. For instance, involving doctors and nurses in data collection could lead to a more accurate and practical diagnosis of hypotension.

3. **Prospective Experimentation:** Conducting prospective experiments can offer several advantages. Not only would this approach solve the problems associated with limited data and data collection, but it would also enable the inclusion of various data types that may otherwise be challenging to obtain. For example, information related to induction time and dosage could be easily and accurately captured. However, it is crucial to ensure the safety and ethical compliance of such experiments, especially when using invasive devices which introduce the risk of incision.

With an enriched database, promising possibilities can be explored, leading to substantial improvements in the predictive model. The incorporation of multiple high-quality data enhances the learning process and generality while increasing reliability. External databases provide valuable additional context, such as patient demographics, comorbidities, or medication history, further enhancing the model's understanding of factors influencing hypotension. However, when integrating external data, it is essential to address potential challenges that may arise:

1. **Data Compatibility:** Merging and aligning different healthcare databases can be complex due to variations in data formats, structures, and terminologies. Addressing these issues will require thorough data preprocessing and normalization to ensure meaningful integration.

2. **Privacy Concerns:** When incorporating external data, it is crucial to handle sensitive patient information with the utmost care to protect patient privacy. Strict adherence to data anonymization and de-identification protocols should be maintained to uphold ethical standards in the model development process.

3. **Data Quality Assurance:** External databases may have varying levels of data quality and reliability. Rigorous data quality assurance processes are necessary to identify and mitigate potential biases or inaccuracies that could impact the model's performance.

By thoughtfully considering these considerations and leveraging external databases, hypotension prediction models can be significantly improved, providing more accurate and clinically relevant predictions to aid in patient care and decision-making.

6.2.2 Machine Learning Improvement

While data quality undoubtedly plays a critical role in predictive performance, there are several issues in the ML pipeline that, if addressed, could further enhance the accuracy of PIH prediction. Currently, our PIH prediction results need to be more accurate and reliable. Despite having built a comprehensive pipeline based on prior-induction data, there is room for improvement in our methods. To address this, we can explore several avenues:

1. **Feature Engineering:** Our current approach to feature generation heavily relies on previous studies of hypotension predictors. For instance, we calculate polynomial features only if they have been documented in prior literature. However, there is a possibility that novel combinations could provide more representative features. By exploring diverse feature engineering techniques and integrating domain knowledge into feature selection, we can achieve a more effective representation of patient profiles and improve the identification of hypotensive events.

2. **Ensemble Strategies:** While the current ensemble strategy has been applied, its implementation is relatively simple, relying on a straightforward weighting average. More sophisticated ensemble techniques, such as voting or stacking, could be explored to leverage the diversity of multiple models and enhance the overall predictive power. Although these approaches might require more computation, the potential gain in accuracy justifies the investment.

3. **Model Selection and Hyperparameter Tuning:** Careful selection of appropriate ML algorithms and thorough hyperparameter tuning can significantly impact model performance. Exploring various algorithms and tuning their respective hyperparameters can optimize predictive capabilities for PIH.

4. **Interpretability:** Interpretability is crucial in medicine, but our project currently ignores it. Since TCI systems calculate anesthetic dosage using embedded equations, it would be highly beneficial if the prediction model could explain which factors affect the PIH and how they do so. This would enable advanced auto-decision-making. Moreover, it would also foster trust among clinicians and validate the model's decisions.

6.2.3 Medical Perspective

From a medical perspective, although anesthesiologists from EMC acknowledge the promising results generated by our dosage advice model, several challenges still need to be addressed before the model can be practically applied. Further improvements and considerations are essential for the model's effectiveness and safety.

Firstly, the dosage curve demonstrates a clear relationship between dosage and risk, but it is evident that minimizing propofol dosage does not necessarily translate to improved anesthesia quality. To enhance the control system, additional factors like the DoA should be incorporated. DoA provides an estimation of the patient's anesthesia

level and can somehow reflect the overall risk. Moreover, determining the threshold of a "safe enough" dosage requires thorough discussions that balance prediction performance with medical plausibility. Finding the right balance is crucial to ensure patient safety and effective anesthesia. Secondly, the result still lacks essential validation. It is clear that currently, we can only validate the prediction with the ground truth. It is the same with the PE validation proposed in [37], but the PR performance could only provide an evaluation of the robustness rather than accuracy. Also, the treatment recommendation method it proposed is not easy to replant since there is no conclusion on which kind of strategy the original cases use. Although the TCI system always follows a calculation function, the process usually includes the intervention of anesthesiologists who make decisions based on their experience and will not record the "strategy" they used. Therefore, validation study, or even the problem formulation, needs further studying. Thirdly, our project builds a pipeline of predicting intra-operative adverse events. Such a framework provides good guidance for other events which also need binary predictions. Hypoxemia, for example, is a reasonable object which also plays a vital role in anesthesia safety. Other events, such as delirium and acute kidney injury, have also been studied in Ml-based prediction. However, They can be enabled when more data, such as text diagnostic logs, are available.

Bibliography

[1] imbalanced-learn development team, "Imbalanced-learn: Over-sampling." https://imbalanced-learn.org/dev/over_sampling.html#smote-adasyn, Year. Accessed: August 15, 2023.

[2] T. Sidiropoulou, M. Tsoumpa, P. Griva, V. Galarioti, and P. Matsota, "Prediction and prevention of intraoperative hypotension with the hypotension prediction index: A narrative review," *Journal of Clinical Medicine*, vol. 11, no. 19, p. 5551, 2022.

[3] N. Asai, C. Doi, K. Iwai, S. Ideno, H. Seki, J. Kato, T. Yamada, H. Morisaki, and H. Shigeno, "Proposal of anesthetic dose prediction model to avoid post-induction hypotension using electronic anesthesia records," in *2019 Twelfth International Conference on Mobile Computing and Ubiquitous Network (ICMU)*, pp. 1–4, IEEE, 2019.

[4] S. Lee, H.-C. Lee, Y. S. Chu, S. W. Song, G. J. Ahn, H. Lee, S. Yang, and S. B. Koh, "Deep learning models for the prediction of intraoperative hypotension," *British journal of anaesthesia*, vol. 126, no. 4, pp. 808–817, 2021.

[5] D. B. Wax, H.-M. Lin, and A. B. Leibowitz, "Invasive and concomitant noninvasive intraoperative blood pressure monitoring: observed differences in measurements and associated therapeutic interventions," *The Journal of the American Society of Anesthesiologists*, vol. 115, no. 5, pp. 973–978, 2011.

[6] J. R. Martina, B. E. Westerhof, J. van Goudoever, E. M. H. de Beaumont, J. Truijen, Y.-S. Kim, R. V. Immink, D. A. Jöbsis, M. W. Hollmann, J. R. Lahpor, *et al.*, "Noninvasive continuous arterial blood pressure monitoring with nexfin®," *The Journal of the American Society of Anesthesiologists*, vol. 116, no. 5, pp. 1092–1103, 2012.

[7] B. P. Imholz, G. A. V. MONTFRANS, J. J. Settels, G. M. V. D. HOEVEN, J. M. Karemaker, and W. Wieling, "Continuous non-invasive blood pressure monitoring: reliability of finapres device during the valsalva manoeuvre," *Cardiovascular research*, vol. 22, no. 6, pp. 390–397, 1988.

[8] M. Wijnberge, B. van der Ster, A. P. Vlaar, M. W. Hollmann, B. F. Geerts, and D. P. Veelo, "The effect of intermittent versus continuous non-invasive blood pressure monitoring on the detection of intraoperative hypotension, a sub-study," *Journal of Clinical Medicine*, vol. 11, no. 14, p. 4083, 2022.

[9] M. C. Moghadam, E. Masoumi, S. Kendale, and N. Bagherzadeh, "Predicting hypotension in the icu using noninvasive physiological signals," *Computers in Biology and Medicine*, vol. 129, p. 104120, 2021.

[10] Y. Mariyaselvam, Blunt, "The non-injectable arterial connector eahsn final report," 2017.

[11] ADInstruments, "Nibp recording analysis machines." 2023.7.29.

[12] Depositphotos, "Blood pressure monitor," 2017.

[13] J.-L. Vincent, N. D. Nielsen, N. I. Shapiro, M. E. Gerbasi, A. Grossman, R. Doroff, F. Zeng, P. J. Young, and J. A. Russell, "Mean arterial pressure and mortality in patients with distributive shock: a retrospective analysis of the mimic-iii database," *Annals of intensive care*, vol. 8, pp. 1–10, 2018.

[14] J. B. Bijker, W. A. Van Klei, T. H. Kappen, L. Van Wolfswinkel, K. G. Moons, and C. J. Kalkman, "Incidence of intraoperative hypotension as a function of the chosen definition: literature definitions applied to a retrospective cohort using automated data collection," *The Journal of the American Society of Anesthesiologists*, vol. 107, no. 2, pp. 213–220, 2007.

[15] S. Südfeld, S. Brechnitz, J. Wagner, P. Reese, H. Pinnschmidt, D. Reuter, and B. Saugel, "Post-induction hypotension and early intraoperative hypotension associated with general anaesthesia," *BJA: British Journal of Anaesthesia*, vol. 119, no. 1, pp. 57–64, 2017.

[16] A. R. Kang, J. Lee, W. Jung, M. Lee, S. Y. Park, J. Woo, and S. H. Kim, "Development of a prediction model for hypotension after induction of anesthesia using machine learning," *PloS one*, vol. 15, no. 4, p. e0231172, 2020.

[17] D. L. Reich, S. Hossain, M. Krol, B. Baez, P. Patel, A. Bernstein, and C. A. Bodian, "Predictors of hypotension after induction of general anesthesia," *Anesthesia & Analgesia*, vol. 101, no. 3, pp. 622–628, 2005.

[18] J. Zhang and L. A. H. Critchley, "Inferior vena cava ultrasonography before general anesthesia can predict hypotension after induction," *Anesthesiology*, vol. 124, no. 3, pp. 580–589, 2016.

[19] A. P. Morley, B. P. Nalla, S. Vamadevan, G. Strandvik, A. Natarajan, A. T. Prevost, and C. M. Lewis, "The influence of duration of fluid abstinence on hypotension during propofol induction," *Anesthesia & Analgesia*, vol. 111, no. 6, pp. 1373–1377, 2010.

[20] K. Lee, J. S. Jang, J. Kim, and Y. J. Suh, "Age shock index, shock index, and modified shock index for predicting postintubation hypotension in the emergency department," *The American journal of emergency medicine*, vol. 38, no. 5, pp. 911–915, 2020.

[21] B. Saugel, E.-J. Bebert, L. Briesenick, P. Hoppe, G. Greiwe, D. Yang, C. Ma, E. J. Mascha, D. I. Sessler, and D. E. Rogge, "Mechanisms contributing to hypotension after anesthetic induction with sufentanil, propofol, and rocuronium: a prospective observational study," *Journal of Clinical Monitoring and Computing*, pp. 1–7, 2021.

[22] S. Kendale, P. Kulkarni, A. D. Rosenberg, and J. Wang, "Supervised machine-learning predictive analytics for prediction of postinduction hypotension," *Anesthesiology*, vol. 129, no. 4, pp. 675–688, 2018.

[23] J. Lee, J. Woo, A. R. Kang, Y.-S. Jeong, W. Jung, M. Lee, and S. H. Kim, "Comparative analysis on machine learning and deep learning to predict post-induction hypotension," *Sensors*, vol. 20, no. 16, p. 4575, 2020.

[24] F. Hatib, Z. Jian, S. Buddi, C. Lee, J. Settels, K. Sibert, J. Rinehart, and M. Cannesson, "Machine-learning algorithm to predict hypotension based on high-fidelity arterial pressure waveform analysis," *Anesthesiology*, vol. 129, no. 4, pp. 663–674, 2018.

[25] A. Johnson, T. Pollard, and R. Mark, "Mimic-iii clinical database (version 1.4)." https://doi.org/10.13026/C2XW26, 2016.

[26] H.-C. Lee, Y. Park, S. B. Yoon, S. M. Yang, D. Park, and C.-W. Jung, "Vitaldb, a high-fidelity multi-parameter vital signs database in surgical patients," *Scientific Data*, vol. 9, no. 1, pp. 1–9, 2022.

[27] M. Walsh, P. J. Devereaux, A. X. Garg, A. Kurz, A. Turan, R. N. Rodseth, J. Cywinski, L. Thabane, and D. I. Sessler, "Relationship between intraoperative mean arterial pressure and clinical outcomes after noncardiac surgery: toward an empirical definition of hypotension," *Anesthesiology*, vol. 119, no. 3, pp. 507–515, 2013.

[28] K. Palla, S. L. Hyland, K. Posner, P. Ghosh, B. Nair, M. Bristow, Y. Paleva, B. Williams, C. Fong, W. Van Cleve, *et al.*, "Intraoperative prediction of postanaesthesia care unit hypotension," *British Journal of Anaesthesia*, vol. 128, no. 4, pp. 623–635, 2022.

[29] J. Schenk, M. Wijnberge, J. M. Maaskant, M. W. Hollmann, L. Hol, R. V. Immink, A. P. Vlaar, B. J. van der Ster, B. F. Geerts, and D. P. Veelo, "Effect of hypotension prediction index-guided intraoperative haemodynamic care on depth and duration of postoperative hypotension: a sub-study of the hypotension prediction trial," *British Journal of Anaesthesia*, vol. 127, no. 5, pp. 681–688, 2021.

[30] W. H. van der Ven, D. P. Veelo, M. Wijnberge, B. J. van der Ster, A. P. Vlaar, and B. F. Geerts, "One of the first validations of an artificial intelligence algorithm for clinical use: The impact on intraoperative hypotension prediction and clinical decision-making," *Surgery*, vol. 169, no. 6, pp. 1300–1303, 2021.

[31] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[32] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[33] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U. R. Acharya, "Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)," *Computer Methods and Programs in Biomedicine*, p. 107161, 2022.

[34] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim, *et al.*, "Explainable machine learning predictions to help anesthesiologists prevent hypoxemia during surgery," *BioRxiv*, p. 206540, 2017.

[35] M. O. M. Javad, S. O. Agboola, K. Jethwani, A. Zeid, S. Kamarthi, *et al.*, "A reinforcement learning–based method for management of type 1 diabetes: exploratory study," *JMIR diabetes*, vol. 4, no. 3, p. e12905, 2019.

[36] S. A. Zadeh, W. N. Street, and B. W. Thomas, "Optimizing warfarin dosing using deep reinforcement learning," *Journal of biomedical informatics*, vol. 137, p. 104267, 2023.

[37] D. Bertsimas, A. Orfanoudaki, and R. B. Weiner, "Personalized treatment for coronary artery disease patients: a machine learning approach," *Health care management science*, vol. 23, pp. 482–506, 2020.

[38] D. Bertsimas, A. R. A. Borenstein, A. Dauvin, and A. Orfanoudaki, "Ensemble machine learning for personalized antihypertensive treatment," *Naval Research Logistics (NRL)*, vol. 69, no. 5, pp. 669–688, 2022.

[39] V. Salmasi, K. Maheshwari, D. Yang, E. J. Mascha, A. Singh, D. I. Sessler, and A. Kurz, "Relationship between intraoperative hypotension, defined by either reduction from baseline or absolute thresholds, and acute kidney and myocardial injury after noncardiac surgery: a retrospective cohort analysis," *Anesthesiology*, vol. 126, no. 1, pp. 47–65, 2017.

[40] B. J. Eastridge, J. Salinas, J. G. McManus, L. Blackburn, E. M. Bugler, W. H. Cooke, V. A. Concertino, C. E. Wade, and J. B. Holcomb, "Hypotension begins at 110 mm hg: redefining "hypotension" with data," *Journal of Trauma and Acute Care Surgery*, vol. 63, no. 2, pp. 291–299, 2007.

[41] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)*, pp. 1200–1205, Ieee, 2015.

[42] J. Brownlee, *Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python*. Machine Learning Mastery, 2020.

[43] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[44] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial intelligence*, vol. 97, no. 1-2, pp. 245–271, 1997.

[45] L. Breiman, *Classification and regression trees*. Routledge, 2017.

[46] P. Waldmann, G. Mészáros, B. Gredler, C. Fuerst, and J. Sölkner, "Evaluation of the lasso and the elastic net in genome-wide association studies," *Frontiers in genetics*, vol. 4, p. 270, 2013.

[47] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, 2005.

[48] Google Machine Learning, "Imbalanced Data Sampling and Splitting." https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data, 2023.

[49] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[50] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pp. 1322–1328, IEEE, 2008.

[51] J. Bergstra, D. Yamins, and D. Cox, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," in *International conference on machine learning*, pp. 115–123, PMLR.

[52] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*, pp. 1–15, Springer, 2000.

[53] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in neurorobotics*, vol. 7, p. 21, 2013.

[54] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *The Journal of Machine Learning Research*, vol. 11, pp. 2079–2107, 2010.

[55] V. Billard, F. Moulla, J. Bourgain, A. Megnigbeto, and D. Stanski, "Hemodynamic response to induction and intubation. propofol/fentanyl interaction.," *Anesthesiology*, vol. 81, no. 6, pp. 1384–1393, 1994.