Prediction of Postinduction Hypotension by Machine Learning

Zhao, Shuoyan; Hamo, Alan; Ottenhof, Niki; Korstanje, Jan Wiebe H.; Dauwels, Justin

DOI
10.1109/EMBC53108.2024.10782974

Publication date
2024

Document Version
Final published version

Published in
46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2024 - Proceedings

2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) | 979-8-3503-7149-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/EMBC53108.2024.10782974

Prediction of Postinduction Hypotension by Machine Learning

1st Shuoyan Zhao
Department of Microelectronics
Delft University of Technology
Delft, The Netherlands
S.Zhao-2@tudelft.nl

2nd Alan Hamo
Department of Microelectronics
Delft University of Technology
Delft, The Netherlands
Alan.hamo1991@hotmail.com

3rd Niki Ottenhof
Department of Anaesthesia
Erasmus University Medical Center
Rotterdam, The Netherlands
n.ottenhof@erasmusmc.nl

4th Jan-Wiebe H Korstanje
Department of Anaesthesia
Erasmus University Medical Center
Rotterdam, The Netherlands
j.korstanje@erasmusmc.nl

5th Justin Dauwels
Department of Microelectronics
Delft University of Technology
Delft, The Netherlands
J.H.G.Dauwels@tudelft.nl

***Abstract*—Post-induction hypotension (PIH) occurs shortly after anesthesia induction and is related to several post-operative complications. Medications delivered during induction and maintenance of anesthesia are significantly related to PIH occurrence, which remains common due to the intricate nature of clinical factors. To enhance decision-making on anesthestic dosing, machine learning (ML) is proposed to predict the risk of PIH associated with specific anesthetic dosages. This study focuses on the development of a prediction model for PIH to support anesthesia decision-making. Trained on 320 cases from the VitalDB database, the model incorporates demographic data, vital signs, and medication dosing information. By including the dosage of propofol administered during the induction period as an input variable, the algorithm predicts PIH risk before induction, providing valuable insights into the safety of propofol dosage plans. The results were validated using nested cross-validation, achieving high performance (precision of 0.83 and recall of 0.84). Moreover, an advisory model demonstrates the potential for personalizing a safe propofol anesthetics range for an individual patient.**

***Index Terms*—Postinduction hypotension, Anesthesia, Machine learning**

I. Introduction

Effective management of arterial hypotension is crucial throughout surgical procedures, as the resulting organ ischemia may lead to serious complications or even mortality [1]. The induction period commands special attention in hypotension care. Induction refers to the early stage of anesthesia when anesthetics and other medications are administrated at a high infusion rate to rapidly induce unconsciousness in patients. Hypotension occurring shortly after induction is termed post-induction hypotension (PIH) and is directly associated with anesthetic drugs such as propofol or remifentanil [2] [3]. Throughout the history of anesthesia, continuous efforts have been made to prevent PIH by adjusting the anesthesia plan. For example, modifying the types and dosage of the anesthetic is thought to be helpful [4]. Presently, anesthesia plans are determined based on characteristics of patients and experience-driven preference of anesthesiologists. Unfortunately, such plans suffer from shortcomings in accuracy and personalization, particularly in planning for patients with compromised health conditions [5].

In response to these limitations, various machine learning studies leveraging insights from big data have been proposed. A logistic regression model has been developed to predict hypotension during intensive care unit stays [6]. Although the model demonstrates an impressive performance with an AUC (area under the receiver operating characteristic curve) value of 0.95, it relies on high-fidelity waveforms of arterial pressure, introducing a bias in the dataset toward patients' healthy status. Contrastingly, non-invasive signals of arterial pressure have been analyzed in studies of PIH prediction [7] [8]. Both works utilize features extracted from demographic data, intraoperative medications, and vital signs. However, due to the emphasized raw nature of the data, minimal data processing is employed in [7], resulting in limited performance with an AUC value of 0.76. Although the AUC improves to 0.84 in [8], the work lacks a convincing validation approach when incorporating feature selection. Furthermore, these studies focus on analyzing data late after induction, targeting early warning of PIH but offering little information for anesthesiologists on how to proactively prevent PIH.

In this study, our main objective is to develop an ML-based predictive model for PIH that aligns with clinical requirements previously overlooked or unaddressed. To enhance practicability, we redefine the prediction target PIH, making it more commonly applicable within the anesthesia context. To improve predictive capabilities beyond previous PIH predictions, we apply advanced ML techniques, addressing challenges such as dataset imbalance and feature extraction. Specifically, we adopt nested validation for a more robust evaluation while identifying efficient features. For actionable suggestions for preventing PIH, we design a novel implementation workflow centered around the ML-based prediction model. The algo-

rithm enhances result reliability and applicability, facilitating safer anesthesia planning advice for anesthesiologists.

II. Materials and Methods

A. Data Collection

The dataset for this study is sourced from VitalDB [9], an open-source surgical database that comprises records from non-cardiac routine or emergency surgeries conducted at Seoul National University Hospital in the Republic of Korea. The study focuses on a population of 320 adult patients who underwent general anesthesia and obtained propofol during the surgery. The recorded data encompasses three main categories:

1) Demographic Data: The demographic information is collected from electrical health records, including age, gender, height, weight, BMI (Body Mass Index), and preoperative disease records.

2) Vital Signs Recordings: Recordings include intermittent measurements of blood pressure, heart rate, oxygen saturation, and electrocardiogram signals. Each vital sign is measured every 2 seconds.

3) Medication Dosing Data: Data on the anesthetics, vasopressor drug, and analgesia medication are acquired at the same sampling rate as the vital signs. They include infusion details (rate and volume) of propofol, remifentanil, phenylephrine, ephedrine, and epinephrine.

To ensure dosage suggestions could be provided before propofol administration, the vital signs as input features are collected before the induction starts. However, medication data are collected until the end of induction. During model application, anesthesiologists can manually input medication dosing data based on their anesthesia plan, allowing the predictive model to assess the safety of the proposed plan. The workflow of the model is depicted in Fig. 1, illustrating its functionality in real-world use.

B. Primary Outcome

The primary outcome of the predictive model is the PIH occurrence. Detection of PIH events is achieved through a one-minute measurement window going along the initial 15 minutes after induction. Specifically, an event is classified as PIH if over 90% of the Systolic Blood Pressure (SBP) values within the measurement window are below 75 mmHg or showed a relative drop of more than 30% from the baseline [10]. Employing the relative definition adds a more personalized touch to the results, aligning with clinical experiences. The distribution of the binary outcomes reveals 191 instances of positive PIH events and 121 instances of negative non-PIH events, indicating a slight data imbalance (9.7%).

C. Features Engineering

We extract a total of 88 individual statistical features from the dataset. Additionally, 17 combinatorial features are generated by calculating ratios or polynomials of demographic features and vital signs. This combinatorial approach helps to enrich the representation of raw features. Certain features, such as the shock index (the ratio of heart rate to SBP), have been previously established as highly relevant to the occurrence of hypotension during surgery [11] [12].

In an ML model, not all features contribute effectively to a given prediction task; some may even introduce unnecessary complexity or noise that hampers the model. Identifying the most relevant and informative features from the original set is crucial for prediction accuracy. There are various techniques for feature selection, categorized into filter methods, wrapper methods, and embedded methods [13]. We apply Recursive Feature Elimination (RFE) [14], a powerful wrapper method for small-size classification problems. RFE starts by training the model on the entire set of features and then iteratively removes the least significant feature based on a predefined criterion, such as accuracy or precision. This recursive process continues until a predetermined number of features remains. Importantly, feature selection in our model is integrated during cross-validation to prevent data leakage, ensuring an unbiased and automated process.

D. Model Development

1) ML Predictive Model: The prediction problem of our work is in fact a binary classification problem, categorizing the predictive results into either PIH events or non-PIH events. The ML-based prediction model is built using the established algorithms, including the Logistic Regression (LR), Random Forest (RF) [15], and Extreme Gradient Boosting (XGBoost) [16] models. LR holds a relatively simple structure, facilitating interpretation at the cost of potential compromise in prediction performance. Both RF and XGBoost are representations of ensemble learning methods, with specific emphasis on decision trees in the ensemble. While RF and XGBoost excel in capturing non-linear relationships, they may be susceptible to noise which is common in clinical data. Hence, beyond the three individual models, we explore an additive simple averaging ensemble method. The scheme averages the generated probabilities of binary results (between 0 and 1) from XGBoost and RF models and determines the final outcome based on a chosen threshold, striking a balance between accuracy and stability. To address the slight imbalance in the dataset, the Synthetic Minority Over-sampling Technique (SMOTE) method [17], an oversampling algorithm that generates new data points in the feature space, is applied. Moreover, each model is fine-tuned on hyper-parameters, with the tuning process embedded into the validation. This procedure is known as nested cross-validation, as depicted in Fig 2. The step is emphasized due to the fact that the XGboost algorithm is sensitive to parameters that its performance largely depends on the selection of hyper-parameters. Without the nesting step, there is a risk of data leakage, potentially leading to over-fitting issues and bias in the result, thus impairing prediction [18].

2) Dosage Advisory Model: The dosage advisory model is essentially an application of the predictive model, with certain input features manually planned instead of being derived from recordings. It processes vital signs and demographic information from the patients, along with the anesthesia plan (propofol dosage) provided by the anesthesiologist as input.

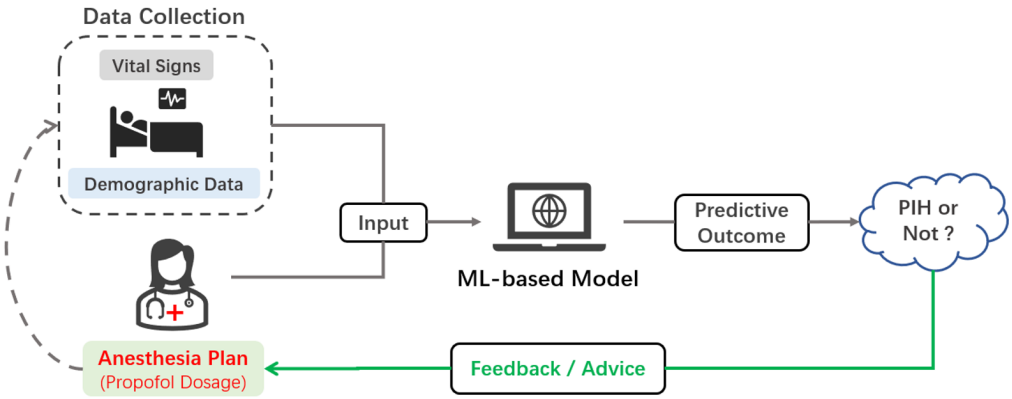

Fig. 1. The designed workflow of model implementation in clinical practice.

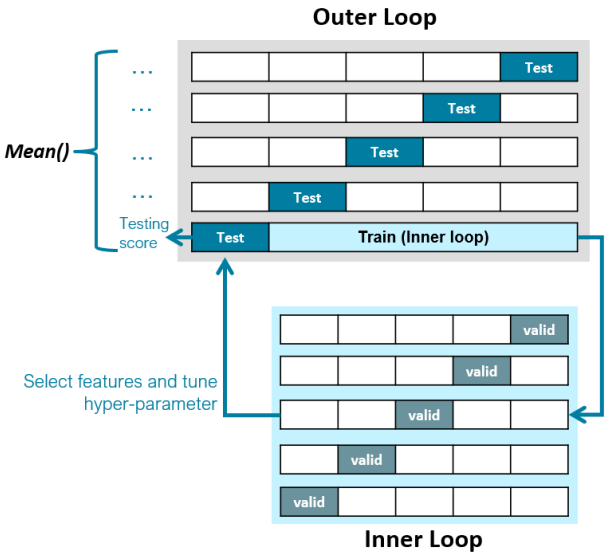

Fig. 2. The structure of nested cross-validation.

When the logistic-core models predict the possibility of PIH, anesthesiologists can hereby assess the risk and then make reasonable decisions to limit the dosage to a safe range.

III. Evaluation and Results

We apply leave-one-out cross-validation (LOOCV) for model evaluation. LOOCV is a special case of k-fold cross-validation where k equals the dataset size. In each round of iteration, we train the model on all but one sample, and the left-out sample was used for testing. LOOCV ensures that every sample has an equal chance to be tested individually and also benefits the small-size training due to the full usage of data.

Table I presents a summary of the performance of the prediction models. All values of criteria are the mean values across the whole rounds of LOOCV. Fig. 3 shows the PR (Precision-Recall) curve and the ROC (Receiver Operating Characteristic) curve of the three original models. Notably, the XGBoost model showcases outstanding performance, achieving an accuracy of 0.81 and a precision of 0.83. The AUC values of ROC and PR further underscore the good ability of the model to maintain high precision while effectively capturing positive instances and distinguishing between the two classes. Evidently, the ensemble learning strategy improves the performance compared to the LR model. However, it does not guarantee the enhancement across all criteria.

TABLE I
Evaluation Comparison of Machine Learning Models

Criteria	XGBoost	Logistic Regression	Random Forest	Ensemble Model
Accuracy	0.81	0.77	0.78	0.79
Precision	0.83	0.75	0.74	0.77
Recall	0.84	0.88	0.91	0.89
F1 Score	0.83	0.81	0.82	0.83
Specificity	0.79	0.81	0.85	0.83

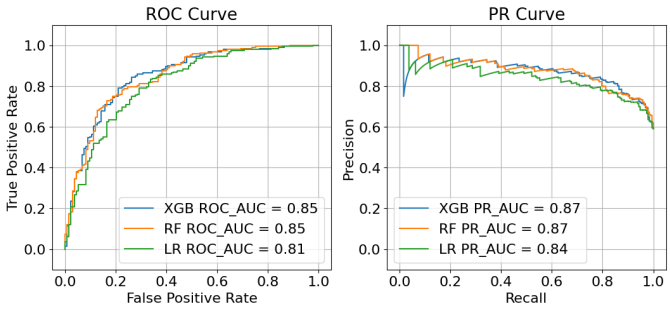

Fig. 3. The PR curve and the ROC curve of different models.

During LOOCV, each individual case is tested in one iteration, allowing for the illustration of the dosage advisory model. In each testing case, we vary the propofol dosage to different values, akin to how an anesthesiologist might adjust the plan during clinical practice. The PIH possibilities at various propofol dosage levels are then generated and plotted in Fig. 4. However, stability is not consistently observed. In some cases of the XGBoost model, the risk decreases as the propofol dosage increases, contrary to clinical experience. Although this anomaly does not necessarily imply a flawed model, we

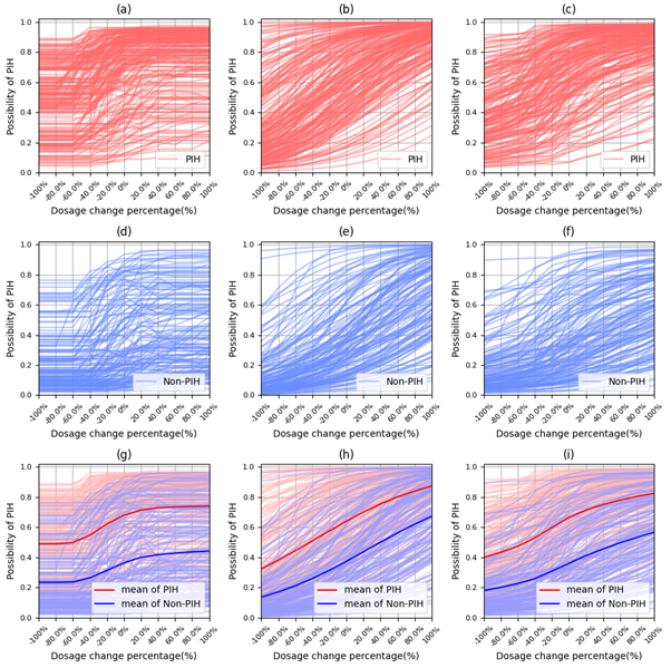

Fig. 4. Advisory model for propofol dosage. Each line represents an individual test case in LOOCV. From left to right, columns correspond to the XGBoost model, LR model, and the ensemble model. From top to bottom, rows depict cases labeled with PIH (red lines), cases labeled with Non-PIH (blue lines), and the mean values of each class across all cases (thick lines), respectively. The thick lines in the last row illustrate the general trend of how the probability varies corresponding to the increase in propofol dosage. In each subplot, the input feature of propofol dosage varies from zero (-100%) to double the amount (+100%) along the x-axis.

remain cautious about the result from the XGBoost method. In contrast, by incorporating the results from LR, the ensemble model corrects the trend of most cases, despite its relatively inferior performance compared to the XGBoost model.

IV. Conclusion and Discussion

In this study, we explored demographic data, vital signs, and medication dosing information to predict PIH. Our approach leveraged various feature engineering techniques to enhance the representative of input features. Through the integration of feature selection and hyper-parameter optimization, we not only maximized the performance of the ML models but also mitigated data leakage. The LOOCV technique maximized the utility of a limited dataset of small size. The XGBoost model exhibited high performance in PIH prediction, and through ensemble learning, the final model effectively balanced the explainability and the prediction accuracy. In the end, we developed an intuitive dosage advice workflow that recommends a safe propofol dosage to prevent PIH before induction for anesthesiologists. Notably, our work is the first effort to combine probability prediction of hypotension with pre-operative anesthetics dosage input, providing support for decision-making in anesthesia care.

In future work, we will validate the proposed ML algorithms on external datasets. To validate the dosage advisory model, additional statistical analyses under the supervision of anesthesiologists are recommended. We plan to tailor the prediction tool by further designing and optimizing the algorithm developed in this work for practical use by anesthesiologists.

References

[1] T. G. Monk, M. R. Bronsert, W. G. Henderson, M. P. Mangione, S. J. Sum-Ping, D. R. Bentt, J. D. Nguyen, J. S. Richman, R. A. Meguid, and K. E. Hammermeister, "Association between intraoperative hypotension and hypertension and 30-day postoperative mortality in noncardiac surgery," *Anesthesiology*, vol. 123, no. 2, pp. 307–319, 2015.

[2] 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.

[3] J. R. Sneyd, A. R. Absalom, C. R. Barends, and J. B. Jones, "Hypotension during propofol sedation for colonoscopy: a retrospective exploratory analysis and meta-analysis," *British Journal of Anaesthesia*, vol. 128, no. 4, pp. 610–622, 2022.

[4] R. B. Schonberger, F. Dai, G. Michel, M. T. Vaughn, M. M. Burg, M. Mathis, S. Kheterpal, S. Akhtar, N. Shah, and A. Bardia, "Association of propofol induction dose and severe pre-incision hypotension among surgical patients over age 65," *Journal of Clinical Anesthesia*, vol. 80, p. 110846, 2022.

[5] T. W. Schnider, C. F. Minto, M. M. Struys, and A. R. Absalom, "The safety of target-controlled infusions," *Anesthesia & Analgesia*, vol. 122, no. 1, pp. 79–85, 2016.

[6] 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.

[7] 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.

[8] 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.

[9] 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, p. 279, 2022.

[10] 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.

[11] S. Trivedi, O. Demirci, G. Arteaga, R. Kashyap, and N. J. Smischney, "Evaluation of preintubation shock index and modified shock index as predictors of postintubation hypotension and other short-term outcomes," *Journal of critical care*, vol. 30, no. 4, pp. 861–e1, 2015.

[12] 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.

[13] 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.

[14] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, pp. 389–422, 2002.

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

[16] 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.

[17] 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.

[18] 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.