Interpretable Predictive Clustering Tree for Post-Intubation Hypotension Assessment

Estefanía Žugelj Tapia
Institute of Physiology
University of Ljubljana, Medical faculty
Ljubljana, Slovenia
estefania.tapia@mf.uni-lj.si

Borut Kirn
Institute of Physiology
University of Ljubljana, Medical faculty
Ljubljana, Slovenia
borut.kirn@mf.uni-lj.si

Sašo Džeroski
Department of Knowledge Technologies
Jožef Stefan Institute
Ljubljana, Slovenia
saso.dzeroski@ijs.si

Abstract

Intraoperative hypotension following intubation is a clinically significant event associated with increased morbidity and mortality. This study presents an interpretable predictive clustering tree (PCT) model designed for multi-target prediction of hypotensive outcomes, including the prediction of minimum and maximum mean arterial pressure (MAP) values during hypotension in the post-induction period. The multi-target regression trees (MTRT) were evaluated using 10-fold cross-validation, and feature importance was assessed via a random forest model. Compared to the original tree, the pruned model demonstrated improved generalization and reduced complexity, with fewer nodes and enhanced interpretability. The pruned tree structure enabled clear decision thresholds based on modifiable variables such as MAP_after_5min, MAP_basal, and Propofol dose. While the random forest achieved the highest performance and had high complexity, its feature importance ranking analysis supported the relevance of the attributes retained in the pruned model and provided complementary insights, highlighting globally relevant features, such as SBP_after_5min, that were not prioritized in the single trees. These findings support the use of interpretable models in clinical decision-support to anticipate and potentially modify the occurrence of post-intubation hypotension.

Keywords

multi-target prediction, interpretable machine learning, decision tree pruning, feature importance, post-intubation, intraoperative hypotension

1 Introduction

Intubation is a common procedure in emergency departments and operating rooms, typically performed immediately after the administration of induction agents. These agents have been associated with hemodynamic instability and post-induction hypotension (PIH), frequently defined as mean arterial pressure (MAP) <65 mmHg[1]. Particularly, in perioperative medicine, PIH has been related to worse postoperative outcomes, increased comorbidity, and mortality[2,3]. PIH occurrence is limited to the first 30 minutes post-induction, as this period is directly affected by anesthesia effects, and is usually not related to complex factors due to surgery[4]. Regarding the risk factors, a post hoc analysis in a surgical population of patients at risk of aspiration of gastric content identified different risk factors associated with it in the multivariate analysis: age, a higher baseline heart rate, bowel occlusion requiring nasogastric tube placement before intubation, and the use of remifentanil. A prospective multicenter study found that in the group with hypotension, the dose (mg/kg) of Propofol was significantly higher at 5 and 10 minutes after induction[5]. On the other hand, the following protective factors have been described: low doses of ketamine and basal systolic blood pressure (SBP)[2].

Previous studies have employed traditional multivariate analysis to identify risk factors and have focused on predicting a single target: the presence of hypotension[2,4,5]. However, predicting multiple outcomes simultaneously can capture complex interactions and provide more informative insights, aiding clinical decision-making and support. Therefore, the hypothesis of this study is that predicting multiple outcomes of PIH simultaneously can effectively identify which variables are most influential in predicting PIH. Overall, this study contributes to the prediction of PIH, which can help anesthesiologists to make better decisions during induction, potentially improving patient outcomes.

2 Methods

Predictive clustering trees (PCT) are a machine learning framework that unifies clustering and prediction tasks. In this framework, the node at the roof (the top node or root node) corresponds to the cluster that contains all the data, and each subsequent split partitions the data to minimize intra-cluster variance. CLUS is a free software that implements this framework and supports multi-target prediction. In a multi-target regression tree (MTRT), the obtained tree is more reliable in explaining the dependencies between variables, and the prediction is a vector of values of the target attributes[6,7]. For this reason, CLUS version 2.12.8 was chosen as the software for this retrospective analysis. The documentation and latest version can be found at: https://github.com/knowledge-technologies/clus/tree/main.

Data was sourced from the subset SIS of the MOVER database (https://mover.ics.uci.edu/) —a public database of anonymized patients undergoing various types of surgery[8].

The inclusion and exclusion criteria were the following:

Information Society 2025, 6–10 October 2025, Ljubljana, Slovenia

http://doi.org/DOI 10.70314/is.2025.skui.9188

- *Inclusion criteria:* 1) patients who underwent major surgery procedures with documented application and dose of one of the next medications during induction of general anesthesia: 'Midazolam', 'Propofol', 'Fentanyl', 'Succinylcholine', 'Ketamine', 'Cisatracurium', 'Etomidate', 'Vecuronium', and 'Rocuronium', 2) high temporal resolution vital signs of systolic blood pressure (SBP), diastolic blood pressure (DBP), and mean arterial pressure (MAP) measured from the radial arterial line, registered before the time of intubation and 30 minutes after it, with at least one measure of MAP < 65mmHg during the post-intubation period.

-*Exclusion criteria:* Patients with vital signs out of physiological ranges (MAP <30mmHg and MAP > 200mmHg), and patients who do not meet the inclusion criteria.

As descriptive and target attributes of the learning problem (see Table 1), the following variables were calculated:
1) MAP_basal: average of MAP measures before intubation, 2) SBP_basal: average of SBP measures before intubation, 3) DBP_basal: average of DBP measures before intubation, 4) MAP_after_5min: average of MAP measurements taken after intubation, over a 5-minute period, 5) SBP_after_5min: average of SBP measurements taken after intubation, over a 5-minute period, 6) DBP_after_5min: average of DBP measurements taken after intubation, over a 5-minute period, 7) Min_MAP<65: Minimum MAP <65 mmHg registered from the intubation up to 30 minutes after, 8) Max_MAP<65: Maximum MAP <65 mmHg registered from the intubation up to 30 minutes after, 9) MAP<65_count: Counts of registered measurements <65mmHg over the 30 minutes interval after intubation, 10) MAP_mean_after_30min: average of MAP measures over 30 minutes interval after intubation, 11) SBP_mean_after_30min: average of SBP measures over 30 minutes interval after intubation, 12) MAP<65_mean_after_30_min: average of MAP measures <65 mmHg over 30 minutes interval after intubation, and 13) Body mass index (BMI): weight / ((height / 100)^2).

During data preparation, missing values of the height attribute were replaced with the mean value of the attribute.

Table 1: Descriptive and target attributes

Descriptive attributes (20)	Target attributes (6)
MAP_basal	Min_MAP<65
SBP_basal	Max_MAP<65
DBP_basal	MAP<65_count
MAP_after_5min	MAP<65_mean_after_30_min
SBP_after_5min	MAP_mean_after_30min
DBP_after_5min	SBP_mean_after_30min
Age	
Gender	
Height	
Weight	
BMI	
Midazolam (cumulative dose)	
Propofol (cum. dose)	
Fentanyl (cum. dose)	
Succinylcholine (cum. dose)	
Ketamine (cum. dose)	
Cisatracurium (cum. dose)	
Etomidate (cum. dose)	
Vecuronium (cum. dose)	
Rocuronium (cum. dose)	

After defining the descriptive and target attributes, the entire dataset of 340 patients was split into training and test sets using the sklearn library and the train_test_split function: 80% of the dataset was used for training (272 patients) and 20% for testing (68 patients). To run CLUS, the training and test sets were converted to ARFF format. Corresponding settings file (.s) were created to define the model parameters for the MTRT tasks. Both single-tree and ensemble models were trained, as summarized in Table 2.

Table 2: Tree and ensemble specifications for each respective MTRT.

Model	Predictive clustering tree (PCT)	Random forest
Heuristic	Variance Reduction	Variance Reduction
Pruning Method	M5Multi	-
Ensemble Method	-	RForest
Feature Ranking	-	Genie3

As an alternative to the train/test split, when running CLUS, the -xval command-line option was used to perform cross-validation on all 340 examples. The number of folds (n = 10) was previously specified in the settings file.

Model performance was evaluated using the following metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Relative Squared Error (RRMSE), and Pearson correlation coefficient (r^2), computed on both training and test sets.

3 Results

After applying the exclusion criteria, we were left with 340 patients. Figure 1 illustrates the flow chart for patient selection, and Table 3 shows their demographic characteristics.

Table 3: Data set population characteristics

Age, years, mean (SD)	58.9 (18.9)
Gender (male), count	201
Weight, kg, mean (SD)	78.6 (23.1)
Height, cm, mean (SD)	168.4 (11.1)
BMI, kg/m2, mean (SD)	1.5 (6.8)

3.1 Complexity of the Models and Structure

The induction time for the original model was significantly shorter (0.03 sec, pruning time 0 sec) compared to the random forest model (1.62 sec), reflecting its reduced complexity. Structurally, the original tree consisted of 241 nodes, 121 leaves, and a depth of 17, whereas the pruned tree was noticeably simpler, with only 19 nodes, 10 leaves, and a depth of 6.

Additionally, the ensemble random forest model, composed of 100 trees, contained a total of 21,050 nodes and 10,575 leaves,

with an average tree depth of 154, indicating a significantly higher complexity and capacity for capturing intricate patterns in the data.

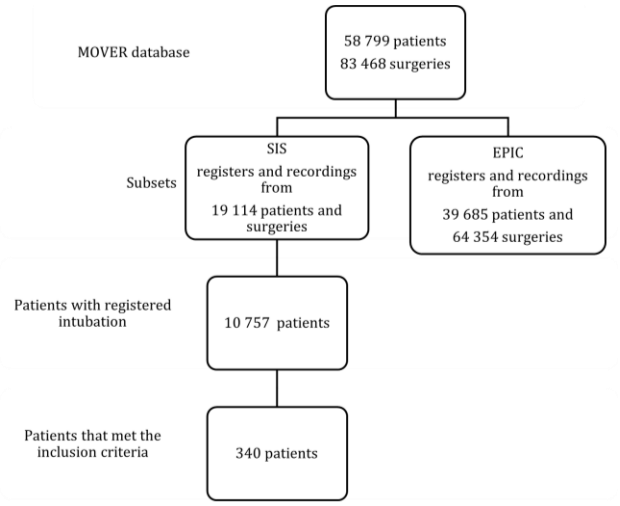

Figure 1: Overview of sample population included in this study.

3.2 Model Performance

The forest with 100 trees exhibits the best performance overall. However, pruning significantly simplified the original model while retaining, and even improving, its predictive power, with lower testing errors for MAE, MSE, RMSE, and RRMSE compared to the original tree (See Table 4).

Table 4: Metrics for training and testing errors (Train/Test)

Metric	Default	Original (Unpruned)	Pruned	Forest (100 trees)
MAE	7.27 / 7.30	2.58 / 7.55	5.41 / 6.18	2.93 / 5.22
MSE	109.35 / 110.12	17.77 / 120.54	61.39 / 83.03	18.41 / 51.05
RMSE	9.41 / 9.45	3.81 / 10.15	7.09 / 8.33	3.96 / 6.76
RRMSE	1.00 / 1.00	0.42 / 1.13	0.76 / 0.91	0.44 / 0.86
Pearson r²	– / 0.04	0.82 / 0.14	0.42 / 0.21	0.89 / 0.26

3.3 Cross-Validation Results

The 10-fold cross-validation was conducted using all 340 examples, with an induction time of 0.26 sec for the single tree and of 9.75 sec for the ensemble random forest. The mean number of tests for the original model was 267, for the pruned model 39.2, and for the random forest 100.

As shown in Table 5, the absolute error metrics (MAE, MSE, RMSE) were higher than when using a train/test split, however the cross-validation approach yielded lower testing errors for RRMSE and higher Pearson r² values.

Table 5: Cross-validation metrics for training and testing errors (Train / Test)

Metric	Default	Original (Unpruned)	Pruned	Forest (100 trees)
MAE	13.62 / 13.69	1.80 / 10.49	5.78 / 9.28	2.82 / 5.6
MSE	300.15 / 302.3	9.22 / 193.3	64.2 / 150.5	16.83 / 63.15
RMSE	17.32 / 17.39	3.04 / 13.90	8.01 / 12.27	3.81 / 7.41
RRMSE	1.00 / 1.00	0.18 / 0.80	0.46 / 0.70	0.43 / 0.84
Pearson r²	0.0003 / 0.02	0.97 / 0.45	0.79 / 0.52	0.89 / 0.28

Note that cross-validation yields more realistic estimates of error on unseen examples as compared to a single train-test split.

3.4 Original Model

As stated in section 3.1, the original model contains 241 nodes and 121 leaves. MAP_after_5min is at the root node, followed by MAP_basal, these two variables repeat along the tree on more than one occasion. Except for cisatracurium, ketamine, and etomidate, in the remaining nodes, the rest of the descriptive attributes appear at least once, showing different thresholds.

3.5 Pruned Model

In the pruned model, the descriptive attributes retained for multi-target prediction were MAP_after_5min, MAP_basal, BMI, SBP_basal, DBP_after_5min, and Propofol dose. Compared to the original tree, the pruned model demonstrated improved generalization and interpretability, with a significantly reduced number of nodes, as illustrated in Figure 2.

The highest predicted values for the target attributes—97.9 mmHg for MAP_mean_after_30min and 149.8 mmHg for SBP_mean_after_30min—were observed when MAP_after_5min exceeded 93 mmHg and SBP_basal was greater than 181 mmHg.

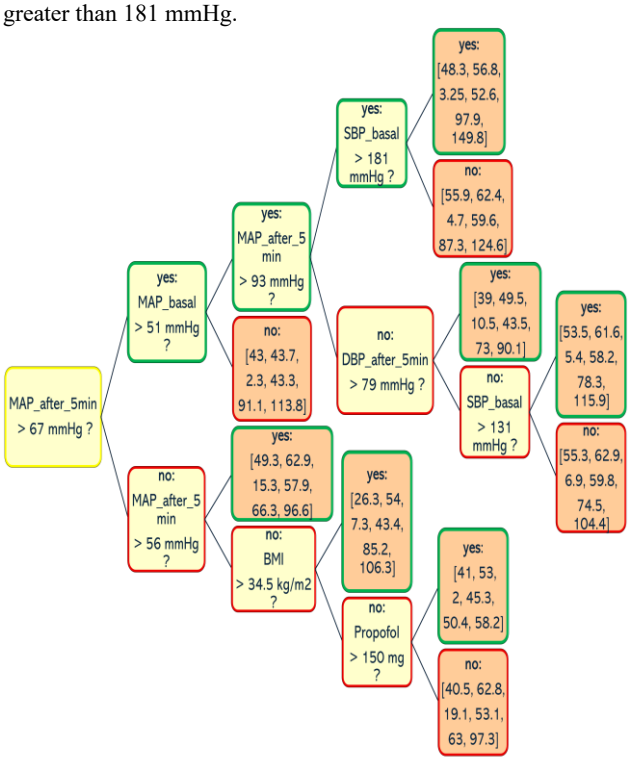

Figure 2: Pruned tree, predicting min_MAP<65, max_MAP<65, MAP<65_count, MAP<65_mean_after_30_min, MAP_mean_after_30min, and SBP_mean_after_30min. Leaves display predictions in orange.

On the other hand, the lowest predicted values of these target variables—50.4 and 58.2 mmHg— were derived from the following nodes: MAP_after_5min below 56 mmHg, BMI < 34.5 kg/m^2, and the Propofol dose >150 mg. Additionally, the leaf node corresponding to BMI >34.5 kg/m^2 yielded the deepest value for min_MAP <65, at 26.3 mmHg.

Other notably low predictions related to hypotension included max_MAP<65 at 43.7 mmHg and MAP<65_mean_after_30_min at 43.3 mmHg, both derived from the node where MAP_basal was below 51 mmHg.

3.6 Forest and Feature Ranking

Despite the complexity of the forest with 100 trees, the feature ranking, where feature importance was assessed using the Genie3 score, helps to understand the descriptive attributes that mainly contributed to the final multi-target prediction. Figure 3 lists the first eleven descriptive attributes, ranked by their corresponding importance score. MAP_after_5min and SBP_after_5min are clearly the most influential features in the model; MAP_basal and SBP_basal also contribute significantly, closely following in importance.

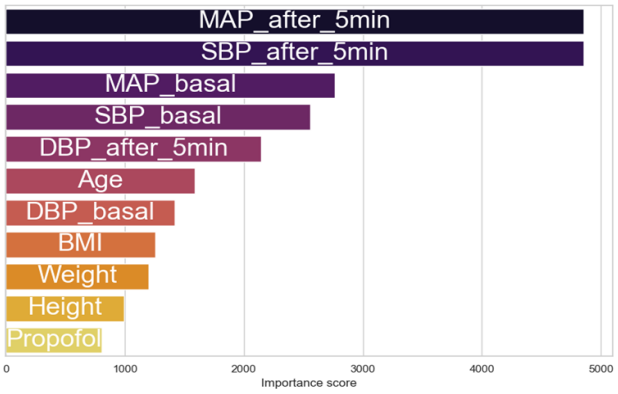

Figure 3: Descriptive attributes contributing most to the random forest's prediction, sorted by importance score.

4 Discussion & Conclusion

The advantages of using a predictive clustering method for multi-target prediction include the ability to capture complex interactions between descriptive attributes and the simultaneous prediction of multiple outcomes [6,7]. A key novelty of this study is its focus on predicting multiple outcomes related to hypotension. This multi-target approach provides a more comprehensive overview and enhances clinical decision-support. In clinical practice, anesthesiologists need to anticipate and often ask themselves: How low will MAP values drop? How will MAP evolve throughout the procedure? This is highly relevant because deeper and longer hypotensive episodes increase the presence of adverse events associated with intraoperative hypotension[3,4].

In this study, the pruned model included among the most important variables for the multi-target prediction MAP_after_5min and MAP_basal. Previous studies have significantly associated PIH with the basal or pre-induction MAP[2,4,5], and our results confirm this observation: In the node root, the MAP value was the most relevant when calculated immediately 5 minutes after intubation, specifically with a decisive threshold of 67 mmHg. To diminish the impact of the basal blood pressure values in the occurrence of PIH episodes, some proposals include discontinuing renin–angiotensin–aldosterone system antagonists the day of the surgery and proactive measures to elevate preoperative values to relieve the effect of the anesthetic medications, which could prevent the appearance of PIH [3,4].

The obtained pruned predictive clustering tree model showed lower testing errors across all metrics compared to the original tree, with improved performance, interpretability, and generalization. Nevertheless, the random forest model performed the best. Regardless of the complexity of the ensemble model, the feature ranking provided valuable insights into the contribution of each attribute to the final prediction; some of these top-ranked features also appear along the nodes of the unpruned and pruned trees. By aggregating importance across multiple trees, random forests can highlight globally relevant features that may not dominate early decision paths in a single tree. For example, SBP_after_5min was ranked second in importance, but it did not appear in the top splits of the unpruned tree. In the pruned tree, BMI and Propofol dose are included, but SBP_after_5min, age, and DBP_basal, which ranked higher than BMI and Propofol dose, are not incorporated in the pruned tree. The association between higher age and PIH has been noted in the past [2,5], and it is a variable usually considered during risk evaluation; however, it is not a modifiable attribute.

In sum, this study demonstrates that interpretable models, such as pruned trees, when supported by feature importance from high-performing models, can validate and offer clear, decisive thresholds of modifiable and actionable variables that impact MAP values in the post-induction period, thereby reducing PIH-related comorbidity and mortality. This highlights its potential utility as a decision support tool in clinical settings.

References

[1] Salmasi V, Maheshwari K, Yang D, Mascha EJ, Singh A, Sessler DI, et al. Relationship between Intraoperative Hypotension, Defined by Either Reduction from Baseline or Absolute Thresholds, and Acute Kidney and Myocardial Injury after Noncardiac Surgery. Anesthesiology 2017;126(1):47–65. DOI: https://doi.org/10.1097/ALN.0000000000001432

[2] Grillot N, Gonzalez V, Deransy R, Rouhani A, Cintrat G, Rooze P, et al. Post-induction hypotension during rapid sequence intubation in the operating room: A post hoc analysis of the randomized controlled REMICRUSH trial. Anaesth Crit Care Pain Med 2025; 44(3):101502. DOI: https://doi.org/10.1016/j.accpm.2025.101502

[3] Sessler DI, Bloomstone JA, Aronson S, Berry C, Gan TJ, Kellum JA, et al. Perioperative Quality Initiative consensus statement on intraoperative blood pressure, risk and outcomes for elective surgery. Br J Anaesth 2019;122(5):563–74. DOI: https://doi.org/10.1016/j.bja.2019.01.013

[4] Südfeld S, Brechnitz S, Wagner JY, Reese PC, Pinnschmidt HO, Reuter DA, et al. Post-induction hypotension and early intraoperative hypotension associated with general anaesthesia. Br J Anaesth 2017;119(1):57–64. DOI: https://doi.org/10.1093/bja/aex127

[5] Jor O, Maca J, Koutna J, Gemrotova M, Vymazal T, Litschmannova M, et al. Hypotension after induction of general anesthesia: occurrence, risk factors, and therapy. A prospective multicentre observational study. J Anesth 2018; 32(5):673–80. DOI: https://doi.org/10.1007/s00540-018-2532-6

[6] Kocev D, Vens C, Struyf J, Džeroski S. Tree ensembles for predicting structured outputs. Pattern Recognit 2012;46:817–33. DOI: https://doi.org/10.1016/j.patcog.2012.09.023

[7] Petković M, Levatić J, Kocev D, Breskvar M, Džeroski S. CLUSplus: A decision tree-based framework for predicting structured outputs. SoftwareX 2023; 24:101526. DOI: https://doi.org/10.1016/j.softx.2023.101526

[8] Samad M, Angel M, Rinehart J, Kanomata Y, Baldi P, Cannesson M. Medical Informatics Operating Room Vitals and Events Repository (MOVER): a public-access operating room database. JAMIA Open 2023;6(4). DOI: https://doi.org/10.1093/jamiaopen/ooad084