Protocol for the development and validation of machine-learning models for predicting the risk of hypertriglyceridemia in critically ill patients receiving propofol sedation using retrospective data

Jiawen Deng[1], Kiyan Heybati[2], Hemang Yadav[3]

[1]Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada
[2]Mayo Clinic Alix School of Medicine, Mayo Clinic, Jacksonville, FL, USA
[3]Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, Mayo Clinic, Rochester, MN, USA

Corresponding author
Hemang Yadav, MBBS, Department of Pulmonary and Critical Care Medicine, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA. **E-mail:** yadav.hemang@mayo.edu.

ABSTRACT

Introduction

Propofol is a widely used sedative-hypnotic agent for critically-ill patients requiring invasive mechanical ventilation (IMV). Despite its clinical benefits, propofol is associated with increased risks of hypertriglyceridemia. Early identification of patients at risk for propofol-associated hypertriglyceridemia is crucial for optimizing sedation strategies and preventing adverse outcomes. Machine learning (ML) models offer a promising approach for predicting individualized patient risks of propofol-associated hypertriglyceridemia.

Methods and analysis

We propose the development of a ML model aimed at predicting the risk of propofol-associated hypertriglyceridemia in ICU patients receiving IMV. The study will utilize retrospective data from four Mayo Clinic sites. Nested cross-validation (CV) will be employed, with a 10-fold inner CV loop for model tuning and selection as well as an outer loop using leave-one-site-out CV for external validation. Feature selection will be conducted using Boruta and LASSO-penalized logistic regression. Data preprocessing steps include missing data imputation, feature scaling, and dimensionality reduction techniques. Six ML algorithms will be tuned and evaluated. Bayesian optimization will be used for hyperparameter selection. Global model explainability will be assessed using permutation importance, and local model explainability will be assessed using SHapley Additive exPlanations (SHAP).

Ethics and dissemination

The proposed ML model aims to provide a reliable and interpretable tool for clinicians to predict the risk of propofol-associated hypertriglyceridemia in ICU patients. The final model will be deployed in a web-based clinical risk calculator. The model development process and performance measures obtained during nested cross-validation will be described in a study publication to be disseminated in a peer-reviewed journal. The proposed study has received ethics approval from the Mayo Clinic Institutional Review Board (IRB #23-007416).

KEYWORDS

Critical Care, Sedation, Mechanical Ventilation, Propofol, Machine-Learning, Hypertriglyceridemia

Strengths and limitations of this study

- Robust external validation using a nested cross-validation (CV) framework will help assess the generalizability of models produced from the modeling pipeline across different hospital settings.
- A diverse set of machine learning (ML) algorithms and advanced hyperparameter tuning techniques will be employed to identify the most optimal model configuration.
- Integration of feature explainability will enhance the clinical applicability of the ML models by providing transparency in predictions, which can improve clinician trust and encourage adoption.
- Reliance on retrospective data may introduce biases due to inconsistent or erroneous data collection, and the computational intensity of the validation approach may limit replication and future model expansion in resource-constrained settings.

INTRODUCTION

Propofol is a sedative-hypnotic agent commonly used for sedation in critically ill adults requiring invasive mechanical ventilation (IMV) [1]. It is recommended as one of the first-line regimens for this indication by the 2018 Pain, Agitation/sedation, Delirium, Immobility, and Sleep (PADIS) guidelines [2] due to its rapid onset and short duration of action. As a highly lipophilic drug, propofol is formulated in a 10% fat emulsion, typically using soybean oil [1]. However, this formulation has the disadvantage of predisposing patients to hypertriglyceridemia [1,3]. Up to 10% of patients who develop propofol-associated hypertriglyceridemia may progress to pancreatitis [4], which substantially increases these patients' risks of morbidity and mortality in the ICU [5].

Thus, it is crucial to assess potential risk factors associated with the development of propofol-associated hypertriglyceridemia when selecting sedative regimens. Several retrospective cohort studies have identified important predictors of hypertriglyceridemia following propofol sedation, including advanced age [4], propofol dose and duration, body mass index (BMI), illness severity, and concomitant medications [6,7]. However, the question of how this knowledge can be applied systematically for guiding clinical practice remains unanswered.

Machine-learning (ML), which is a modeling paradigm that can identify complex and non-linear patterns in large datasets [8,9], has proven effective for providing individualized patient risk stratification in medical diagnoses and prognoses [10]. ML models could potentially help identify recently intubated ICU patients who are at higher risks for developing propofol-associated hypertriglyceridemia, allowing for these patients to be switched to alternative sedation regimens. Thus, the objective of the proposed ML development study is to create a ML-powered clinical calculator to aid in selecting sedating regimens based on personalized patient risk predictions of developing propofol-associated hypertriglyceridemia in ICU settings.

METHODS AND ANALYSIS

The proposed study will be conducted and reported in accordance with the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis+AI (TRIPOD+AI) Checklist for Prediction Model Development and Evaluation [11].

Data sources

The proposed study involves a secondary analysis of data from a multi-centered retrospective cohort investigation conducted at Mayo Clinic sites in the United States. The original retrospective study included consecutive adults (≥18 years of age) admitted to one of 11 ICUs across four Mayo Clinic sites: 1) Mayo Clinic Rochester, 2) Mayo Clinic Phoenix, 3) Mayo Clinic Jacksonville, and 4) Mayo Clinic Health System (MCHS) community sites in Mankato, Minnesota and Eau Claire, Wisconsin. Patient inclusion criteria for the proposed model development study is the same as the eligibility criteria in the original retrospective study, which included: 1) admission to one of the study ICUs between May 5th 2018, and June 30th 2023, 2) required IMV for greater than 24 hours, and 3) received continuous propofol infusion for >24 hours. Exclusion criteria included: 1) development of hypertriglyceridemia (defined as serum triglyceride levels >400 mg/dL) [5] prior to propofol infusion, and 2) lack of prior authorization for medical records to be assessed for research purposes. The index ICU admission for eligible patients were identified through institutional data warehouses (Mayo Clinic ICU DataMart and Unified Data Platform).

Outcomes of interest

The goal of model development is to predict the probability of hypertriglyceridemia within 10 days following the start of propofol infusion. Hypertriglyceridemia is defined as serum triglyceride levels exceeding 400 mg/dL [5]. The predicted probability estimates will be transformed into a binary classification to categorize patients into high

risk and low risk groups for developing hypertriglyceridemia based on a decision threshold. The anticipated usage setting for the model is during the ICU admission process or shortly before/after intubation and sedation for IMV.

Nested cross-validation

A nested cross-validation (CV) methodology [12] will be used to evaluate the performance and consistency of the overall ML modeling process. Nested CV involves two layers of CV (**Figure 1**):

1. **Inner CV loop:** The inner CV loop is used to tune each ML algorithm and to select the best-performing model. This is similar to regular, non-nested "flat" CV.
2. **Outer CV loop:** The dataset is divided into *k* outer folds. In each iteration of the outer loop, one fold is held out as the test set while the remaining *k-1* folds are used for model tuning and selection using the inner CV loop. The main function of the outer loop is to estimate the performance of the best-performing model selected by the inner loop [13].

For our proposed study, we will use stratified 10-fold CV as our inner CV loop for hyperparameter tuning and model selection. For the outer CV loop, we will use leave-one-site-out CV (LOSO-CV). In our study, LOSO-CV involves tuning, selecting, and training models on data from three of the four included Mayo Clinic sites and externally validating the best performing model on data from the remaining site. This is repeated four times so that each Mayo Clinic site serves as the external validation set at least once. In essence, we are externally validating our modeling process four times to better assess how our models will perform on new, unseen data [14]. Following the nested CV process, we will run the inner CV loop on the entire dataset to generate the final production model for deployment into a clinical calculator.

Feature selection and sample size assessment

Candidate features will be first filtered based on availability (≤10% of missing data) and expert domain (see **Table 1** for the list of candidate features). For each training set in the outer CV loop, the dimension of the feature set will be further reduced using random-forest-based Boruta [15] and Least Absolute Shrinkage and Selection Operator (LASSO) penalized logistic regression [16]. During feature selection, the regularization hyperparameter α of the LASSO penalized logistic regression model will be determined via a grid-search of 1000 α values along the regularization path to minimize the LASSO objective function across 10-fold CV. A union of features selected by the two methods will be chosen as the final feature set used for predictive modeling.

The goal of the feature selection process is to follow the "one-in-ten" rule—which states that there should be 10 patients with the target outcome in the dataset to support every one feature included to ensure stable ML model predictions [17,18]—as closely as possible. Given that our previous analyses involving our dataset showed that there were 851 patients who developed hypertriglyceridemia following propofol administration (without applying the 10 day post-propofol-initiation restriction) [19], we will likely have capacity for 70-80 features. This is far greater than the number of candidate features that we expect to include in the model, thus we anticipate that our dataset will contain enough case samples for modeling purposes.

Data preprocessing

Within each iteration of the outer CV loop, the training set will be split into 10 folds of different training and testing subsets for the inner CV loop. A data preprocessing pipeline will be constructed to identify and impute missing data, correct for class imbalance, and improve compatibility with ML algorithms before the data is entered into the ML models (**Figure 2**).

Categorical data encoding. As all of the candidate categorical features are non-ordinal, they will be transformed via one-hot encoding into multiple binary features [20].

Missing data imputation. Because we expect the amount of missing data in the dataset to be low (≤10%), we will assume missing data to be missing at random and perform missing data imputation. Imputation will be conducted using Multivariate Imputation by Chained Equations (MICE) [21]. As previous research had suggested that the number of MICE iterations should be determined based on the highest percentage of data missing (e.g., if 30% of data needs to be imputed for one feature, then MICE will be run for 30 iterations) [22,23], we will run MICE for 10 iterations to account for the maximum missing data percentage of 10%. The training and testing subsets will be imputed separately to avoid data leakage [24]. When assessing and tuning ML algorithms with native missing data handling methods, we will treat the use of MICE as a hyperparameter to trial both MICE and the algorithms' native missing data handling approaches.

Data imbalance and resampling. Given that we aim to develop a model that can perform probability predictions, we will not resample the dataset even if the dataset is imbalanced. However, we will tune ML algorithms' weight scaling factors as a hyperparameter during model tuning to help reduce bias towards the majority class.

Feature scaling. Before the dataset is entered into dimensionality reduction steps and ML algorithms, the normality of each continuous feature within the dataset will be assessed using the Shapiro–Wilk test [25] and by visual inspection of the features' histograms [26]. If the distribution of all continuous features are determined to be close to normal, the continuous features will be Z-score normalized to center the feature data around the mean and scale the data according to its standard deviation [27]. If any of the continuous features were not close to normal, the continuous features will be transformed using robust normalization, which centers the data around the median and scales the data around its interquartile range [27]. Compared to Z-score normalization, robust normalization is more robust towards the presence of outlier values in the dataset [28].

Dimensionality reduction. Potential multicollinearity in the final feature set will be assessed using variance inflation factors (VIFs) [29]. If the VIF for any feature exceeds 5, we will trial the use of principal component analysis (PCA) as a dimensionality and multicollinearity reduction approach in our data preprocessing pipeline. The lowest number of principal components needed to explain at least 95% of the variance will be kept [30].

Hyperparameter tuning

Within each iteration of the outer CV loop, we will tune and evaluate:

1. Two classical ML algorithms: 1) logistic regression (with LASSO, ridge, or elastic net penalization) and 2) support vector machines (SVMs, with linear, polynomial, radial basis function, or sigmoid kernels).
2. Three ensemble ML algorithms: 1) random decision forest, 2) Light Gradient-Boosting Machine [31] (LightGBM, with either regular Gradient Boosting Decision Trees [GBDT] algorithm [32] or Dropouts meet Multiple Additive Regression Trees [DART] algorithm [33]), and 3) eXtreme Gradient Boosting (XGBoost, with either regular gradient boosting decision trees [GBDT] algorithm [32] or Dropouts meet Multiple Additive Regression Trees [DART] algorithm [33]) [34].
3. A multilayer perceptron (MLP) neural network with 2, 3, or 4 hidden layers. Each hidden layer will use a rectified linear unit (ReLU) activation function with Kaiming kernel initialization [35]. Each ReLU layer will be followed by a dropout layer for regularization [36] and a batch normalization layer to improve training stability and speed [37]. All hidden layers will use the same number of neurons determined via hyperparameter tuning. The output layer will use a sigmoid function to produce the predicted probability. An AdamW optimizer will be used [38]. An example of the proposed network architecture with 2 hidden layers is shown in **Figure 3**.

Each algorithm will be tuned to minimize cross-entropy loss across stratified 10-fold CV. The full list of hyperparameters that will be tuned for each model is tabulated in **Supplementary Table S1 to S17**. The optimal hyperparameters will be selected using Bayesian optimization [39].

Bayesian optimization starts with several initial rounds of random hyperparameter searches to gather data points for building a probabilistic model that predicts the performance of different hyperparameter combinations. An acquisition function then uses this model to identify the most promising hyperparameter combinations for the next round of evaluations. The results are used to update the probabilistic model, and the process is repeated until a pre-established performance budget is exhausted [40]. In essence, Bayesian optimization can be thought of as a more intelligent and "guided" version of random hyperparameter searching, and it is empirically considered to outperform traditional grid-search and random search approaches [41]. Because random search can reliably identify hyperparameter combinations from the top 5% of the most performant combinations with 60 iterations [39], we aim to perform Bayesian optimization at or beyond this performance budget. With computational limitations, we estimate that Bayesian optimization will be performed for at least 200 iterations per model.

Calibration

As the objective of our hyperparameter tuning process is to minimize cross-entropy loss, we can reasonably assume that our models will be well-calibrated. Nevertheless, we will trial Platt/sigmoidal scaling [42,43], isotonic regression [42], or no further calibration to assess if our calibration performance can be further improved. The calibration approach with the lowest average cross-entropy loss on stratified 10-fold cross-validation will be selected.

Threshold selection

Following calibration, each model will undergo threshold tuning to maximize their binary classification performance. Youden's Index [44] will be calculated for every classification threshold between 0.01 and 0.99 at an interval of 0.01, and the threshold with the highest average Youden's Index across stratified 10-fold cross-validation will be selected as the most optimal threshold.

Model selection and evaluation

We will follow a previously published framework proposed for the evaluation of clinical prediction models [45] to perform model selection and external validation. The classification performance of each tuned and calibrated model will be assessed using the area under the curve of the receiver operating characteristic curve (AUC-ROC), accuracy, sensitivity, specificity, and Youden's Index. Calibration performance will be assessed using Brier score.

During model selection within the training set of each outer CV loop, all metrics will be produced and averaged across 10-fold cross-validation, and 95% confidence intervals (CIs) will be used to assess performance variance. The model configuration with the best classification and calibration performance will be selected as the most optimal configuration.

For external validation within each iteration of the outer CV loop, the same set of classification and calibration metrics will be generated using the most optimal model configuration and data from the held-out site not involved in model tuning and selection. The external validation metrics will be averaged across all 4 iterations of the outer CV loop to assess the generalization performance of our modeling process, and 95% CI will be used to assess the variance of the generalization performance.

Model explainability and fairness

To understand how the final production model makes its predictions and to assess model bias arising from over-reliance on patient demographics, we will use the permutation importance method to assess which features have the greatest impact on the performance of our final production model [46]. Permutation importance is a global explainability method that involves repeatedly shuffling individual predictors in the training dataset, which effectively renders the predictor useless to the model. It then observes how this operation affects model performance, thereby identifying the most influential features. To understand how the production model makes individual predictions, we will use permutation-based SHapley Additive exPlanations (SHAP) [47], which is a local explainability method that produces an estimate of the direction and magnitude of each features' effect on individual predictions. A bee swarm summary plot will be used to illustrate SHAP feature attributions for each patient in our dataset.

Model deployment

The final production model will be integrated into a web-based clinical risk calculator based on *shiny*. SHAP waterfall plots will be provided for each individual prediction generated through the calculator to allow users to understand how the model is generating its predictions and to assess whether the predictions make clinical and biological sense.

Implementation of permutation SHAP in the clinical risk calculator requires access to the full training dataset as the user is using the calculator in order to allow SHAP to generate perturbations around a prediction to understand the local behaviors of the model. Involvement of training data in a web-based context represents a data privacy concern. To facilitate SHAP implementation, we will generate a synthetic dataset that mimics the distribution and characteristics of our training dataset for SHAP perturbations using a Differentially Private Conditional Tabular Generative Adversarial Network (DP-CTGAN) [48]. DP-CTGAN uses random noise during network training and data generation to offer a privacy guarantee. Hyperparameters of the DP-CTGAN will be tuned manually to minimize generator and discriminator loss. The privacy budget ε will be set to between 0 and 10. We aim to generate a synthetic dataset that yields similar SHAP value predictions and SHAP base values for use with the clinical risk calculator.

Statistical softwares

The modeling and evaluation processes will be completed using Python 3.11. Boruta-based feature selection will be conducted using the *BorutaPy* package, and LASSO-based feature selection will be conducted using *scikit-learn*. ML models will be fitted, tuned, calibrated, and evaluated using *scikit-learn*, *lightgbm*, *xgboost*, *tensorflow*, *keras*, and *scikit-optimize*. Permutation importance will be assessed using *eli5*, and permutation SHAP will be implemented using *shap*.

Patient and public involvement

None.

ETHICS AND DISSEMINATION

The proposed study has received ethics approval from the Mayo Clinic Institutional Review Board (IRB #23-007416). The requirement for informed consent is waived by institutional review. The development process of the final model and performance measures yielded by nested CV will be described in a research article to be disseminated in a peer-reviewed academic journal.

DISCUSSION

While traditional statistical modeling is useful for identifying predictors of hypertriglyceridemia in mechanically ventilated patients receiving continuous propofol infusion, it is difficult to systematically translate these findings to actual clinical practice. By leveraging ML-based modeling approaches and by incorporating ML models into an accessible clinical risk calculator, we aim to develop a tool that could guide clinicians in making informed decisions surrounding the choice of sedation regimens and potentially reduce the incidence of propofol-related hypertriglyceridemia and associated adverse events such as acute pancreatitis.

A notable use case for our proposed clinical risk calculator is to identify patients who may benefit from increased triglyceride level monitoring. In the absence of widely-accepted guidelines and protocols, routine measurements of triglycerides among patients receiving IMV with propofol sedation is often *ad hoc* and remains care-team dependent. Previous studies have shown that only 15-24% of intubated patients with propofol sedation receive routine measurements of triglyceride levels, with about a fifth of these patients having elevated triglyceride levels >400 mg/dL [5]. It is possible that a ML model can identify specific patient characteristics at the time of ICU admission or intubation that can predict patient risks of developing hypertriglyceridemia, so that triglyceride levels in these patients can be monitored and used to tailor the patients' sedation regimen accordingly.

The proposed ML modeling study has several notable strengths. First, we propose the use of multiple well-established feature selection methods to reduce the dimensionality of our training dataset. This helps us to keep only the most important predictors of hypertriglyceridemia in our models, and to avoid the "curse of dimensionality", including overfitting the models on noise in the data. Secondly, we aim to use robust external validation methods involving a nested CV framework with LOSO-CV. This approach allows the generalizability of our modeling process to be rigorously and repeatedly evaluated across different hospital settings, which improves the external validity of our performance metrics. Third, we aim to trial a diverse set of popular ML algorithms, including classical, ensemble, and neural network models, along with state-of-the-art hyperparameter tuning techniques using Bayesian optimization. This exhaustive approach will allow us to identify the most optimal ML model configuration for our dataset and target outcome. Lastly, we considered the clinical applicability of our ML models by proposing to integrate both global and local feature explainability to assess how our final model makes its predictions, as well as by integrating our final model into a web-based clinical risk calculator.

Despite these strengths, we foresee several potential challenges associated with the proposed study. First, the study relies on retrospective data, which can introduce biases relating to misclassification from potential errors in charting and missing data. The retrospective nature of the data also results in limited granularity on the exact severity of comorbid conditions such as liver disease and prevents us from capturing baseline triglyceride values which precludes their inclusion in the model. These shortcomings in our dataset highlights the need for additional prospective work to incorporate and further assess these features in future works. Secondly, while the proposed nested CV approach is thorough, it will be computationally intensive and time-consuming. This limits the feasibility of future replications of the proposed study methodology, especially in resource-constrained settings or when using larger datasets. Lastly, while the generalizability of our dataset is improved by including information from multiple Mayo Clinic sites, these sites may share similar treatment or data recording protocols that reduce the external validity of our developed models.

Overall, our protocol outlines a comprehensive approach to developing and validating a clinical tool designed to provide a personalized risk estimate of developing hypertriglyceridemia when sedated with continuous propofol infusion while on IMV. In the absence of routine triglyceride monitoring, it remains important to prevent the development of hypertriglyceridemia in an effort to avoid a sequela of adverse events.

Competing interest statement

Jiawen Deng is a member of the OpenAI Researcher Access Program and receives grants in the form of API credits for purposes of research involving large language models from OpenAI. Kiyan Heybati and Hemang Yadav report no relevant conflicts of interest.

Funding statement

This work is supported by the Mayo Clinic Critical Care Research Committee, as well as the National Heart, Lung, and Blood Institute (NHLBI) of the National Institute of Health (NIH) Grant Number K23HL151671 (Recipient: Hemang Yadav). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

Authors' contributions

Jiawen Deng, Kiyan Heybati, and Hemang Yadav contributed equally to the conception and methodological design of the proposed study, and all three authors contributed equally to the drafting and review/editing of the protocol manuscript.

REFERENCES

1 Sahinovic MM, Struys MMRF, Absalom AR. Clinical Pharmacokinetics and Pharmacodynamics of Propofol. *Clin Pharmacokinet*. 2018;57:1539–58. doi: 10.1007/s40262-018-0672-3

2 Devlin JW, Skrobik Y, Gélinas C, *et al.* Clinical Practice Guidelines for the Prevention and Management of Pain, Agitation/Sedation, Delirium, Immobility, and Sleep Disruption in Adult Patients in the ICU. *Crit Care Med*. 2018;46:e825–73. doi: 10.1097/CCM.0000000000003299

3 Kotani Y, Shimazawa M, Yoshimura S, *et al.* The experimental and clinical pharmacology of propofol, an anesthetic agent with neuroprotective properties. *CNS Neurosci Ther*. 2008;14:95–106. doi: 10.1111/j.1527-3458.2008.00043.x

4 Devlin JW, Lau AK, Tanios MA. Propofol-associated hypertriglyceridemia and pancreatitis in the intensive care unit: an analysis of frequency and risk factors. *Pharmacotherapy*. 2005;25:1348–52. doi: 10.1592/phco.2005.25.10.1348

5 Pancholi P, Wu J, Lessen S, *et al.* Triglyceride Concentrations and Their Relationship to Sedation Choice and Outcomes in Mechanically Ventilated Patients Receiving Propofol. *Ann Am Thorac Soc*. 2023;20:94–101. doi: 10.1513/AnnalsATS.202205-403OC

6 Devaud J-C, Berger MM, Pannatier A, *et al.* Hypertriglyceridemia: a potential side effect of propofol sedation in critical illness. *Intensive Care Med*. 2012;38:1990–8. doi: 10.1007/s00134-012-2688-8

7 Dube KM, Szumita PM, Rocchio MA, *et al.* The Effect of Concomitant Sirolimus and Propofol Therapy on Triglyceride Concentrations in Critically Ill Patients. *Am J Ther*. 2019;26:e103–9. doi: 10.1097/MJT.0000000000000461

8 Howell MD, Corrado GS, DeSalvo KB. Three Epochs of Artificial Intelligence in Health Care. *JAMA*. 2024;331:242–4. doi: 10.1001/jama.2023.25057

9 Deng J, Heybati K, Park Y-J, *et al.* Artificial intelligence in clinical practice: A look at ChatGPT. *Cleve Clin J Med*. 2024;91:173–80. doi: 10.3949/ccjm.91a.23070

10 Ching T, Himmelstein DS, Beaulieu-Jones BK, *et al.* Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface*. 2018;15. doi: 10.1098/rsif.2017.0387

11 Collins GS, Moons KGM, Dhiman P, *et al.* TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. doi: 10.1136/bmj-2023-078378

12 Wilimitis D, Walsh CG. Practical Considerations and Applied Examples of Cross-Validation for Model Development and Evaluation in Health Care: Tutorial. *JMIR AI*. 2023;2:e49023. doi: 10.2196/49023

13 Wainer J, Cawley G. Nested cross-validation when selecting classifiers is overzealous for most practical applications. *Expert Syst Appl*. 2021;182:115222. doi: 10.1016/j.eswa.2021.115222

14 Bradshaw TJ, Huemann Z, Hu J, *et al.* A Guide to Cross-Validation for Artificial Intelligence in Medical Imaging. *Radiol Artif Intell*. 2023;5:e220232. doi: 10.1148/ryai.220232

15 Kursa MB, Rudnicki WR. Feature Selection with theBorutaPackage. *J Stat Softw*. 2010;36. doi: 10.18637/jss.v036.i11

16 Muthukrishnan R, Rohini R. LASSO: A feature selection technique in predictive modeling for machine learning. *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*. IEEE 2016.

17 Harrell FE Jr, Lee KL, Califf RM, *et al.* Regression modelling strategies for improved prognostic prediction. *Stat Med*. 1984;3:143–52. doi: 10.1002/sim.4780030207

18 Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15:361–87. doi: 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4

19 Heybati K, Deng J, Xie G, *et al.* Propofol, Triglycerides, and Acute Pancreatitis: A Multi-Center Epidemiologic Analysis. *Ann Am Thorac Soc*. Published Online First: 11 October 2024. doi: 10.1513/AnnalsATS.202407-781OC

20 Ashenden SK, Bartosik A, Agapow P-M, *et al.* Chapter 2 - Introduction to artificial intelligence and machine learning. In: Ashenden SK, ed. *The Era of Artificial Intelligence, Machine Learning, and Data Science in the Pharmaceutical Industry*. Cambridge, MA, USA: Academic Press 2021:15–26.

21 van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations inR. *J Stat Softw*. 2011;45. doi: 10.18637/jss.v045.i03

22 Bodner TE. What improves with increased missing data imputations? *Struct Equ Modeling*. 2008;15:651–75. doi: 10.1080/10705510802339072

23 White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med*. 2011;30:377–99. doi: 10.1002/sim.4067

24 Apicella A, Isgrò F, Prevete R. Don't push the button! Exploring data leakage risks in machine Learning and Transfer Learning. Published Online First: 2024. doi: 10.48550/ARXIV.2401.13796

25 Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). *Biometrika*. 1965;52:591–611. doi: 10.1093/biomet/52.3-4.591

26 Ghasemi A, Zahediasl S. Normality tests for statistical analysis: a guide for non-statisticians. *Int J Endocrinol Metab*. 2012;10:486–9. doi: 10.5812/ijem.3505

27 de Amorim LBV, Cavalcanti GDC, Cruz RMO. The choice of scaling technique matters for classification performance. *Appl Soft Comput*. 2023;133:109924. doi: 10.1016/j.asoc.2022.109924

28 Cabello-Solorzano K, Ortigosa de Araujo I, Peña M, *et al.* The impact of data normalization on the accuracy of machine learning algorithms: A comparative analysis. *Lecture Notes in Networks and Systems*. Cham: Springer Nature Switzerland 2023:344–53.

29 James G, Witten D, Hastie T, *et al. An Introduction to Statistical Learning: with Applications in R*. New York, NY: Springer Science & Business Media 2013.

30 Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. *Philos Trans A Math Phys Eng Sci*. 2016;374:20150202. doi: 10.1098/rsta.2015.0202

31 Ke G, Meng Q, Finley T, *et al.* LightGBM: A highly efficient gradient boosting decision tree. In: von Luxburg U, Guyon I, Bengio S, *et al.*, eds. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc. 2017:3149–57.

32 Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Stat*. 2001;29:1189–232. doi: 10.1214/aos/1013203451

33 Rashmi KV, Gilad-Bachrach R. DART: Dropouts meet Multiple Additive Regression Trees. Published Online First: 2015. doi: 10.48550/ARXIV.1505.01866

34 Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM 2016.

35 He K, Zhang X, Ren S, *et al.* Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. Published Online First: 2015. doi: 10.48550/ARXIV.1502.01852

36 Srivastava N, Hinton G, Krizhevsky A, *et al.* Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J Mach Learn Res*. 2014;15:1929–58.

37 Ioffe S, Szegedy C. Batch Normalization: Accelerating deep network training by reducing internal covariate shift. Published Online First: 2015. doi: 10.48550/ARXIV.1502.03167

38 Loshchilov I, Hutter F. Decoupled weight decay regularization. Published Online First: 2017. doi: 10.48550/ARXIV.1711.05101

39 Jones DR, Schonlau M, Welch WJ. Efficient Global Optimization of Expensive Black-Box Functions. *J Global Optimiz*. 1998;13:455–92. doi: 10.1023/a:1008306431147

40 Snoek J, Larochelle H, Adams RP. Practical Bayesian optimization of machine learning algorithms. Published Online First: 2012. doi: 10.48550/ARXIV.1206.2944

41 Turner R, Eriksson D, McCourt M, *et al.* Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020. Published Online First: 2021. doi: 10.48550/ARXIV.2104.10201

42 Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. *Proceedings of the 22nd international conference on Machine learning - ICML '05*. New York, New York, USA: ACM Press 2005.

43 Platt JC. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *ADVANCES IN LARGE MARGIN CLASSIFIERS*. 1999;10:61–74.

44 Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3:32–5. doi: 10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3

45 Steyerberg EW, Vickers AJ, Cook NR, *et al.* Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21:128–38. doi: 10.1097/EDE.0b013e3181c30fb2

46 Breiman L. Random Forests. *Mach Learn*. 2001;45:5–32. doi: 10.1023/a:1010933404324

47 Lundberg S, Lee S-I. A unified approach to interpreting model predictions. Published Online First: 2017. doi: 10.48550/ARXIV.1705.07874

48 Fang ML, Dhami DS, Kersting K. DP-CTGAN: Differentially private medical data generation using CTGANs. *Artificial Intelligence in Medicine*. Cham: Springer International Publishing 2022:178–88.

FIGURE LEGENDS

Figure 1: Illustration of the proposed nested cross-validation approach. The outer loop consists of a leave-one-site-out cross-validation approach for repeated external validation of models produced by the inner loop. The inner loop involves a 10-fold cross-validation approach for model tuning and selection.

Abbreviations: MCHS Mayo Clinic Health System (community sites).

Figure 2: Flowchart illustrating the flow of training data through our proposed modeling pipeline. Because principal component analysis requires a complete dataset with no missing data, model configurations that did not use imputation (and relied on machine-learning algorithms' native missing data handling methods) cannot use principal component analysis. Flow of data through configurations without imputation is shown using the dotted arrows.

Abbreviations: LASSO Least Absolute Shrinkage and Selection Operator (penalized logistic regression), **MICE** Multivariate Imputation by Chained Equations, **PCA** Principal Component Analysis, **ML** Machine-Learning.

Figure 3: Example schema of the proposed network architecture, with dropout and batch normalization components in each hidden layer. An example with 2 hidden layers is shown.

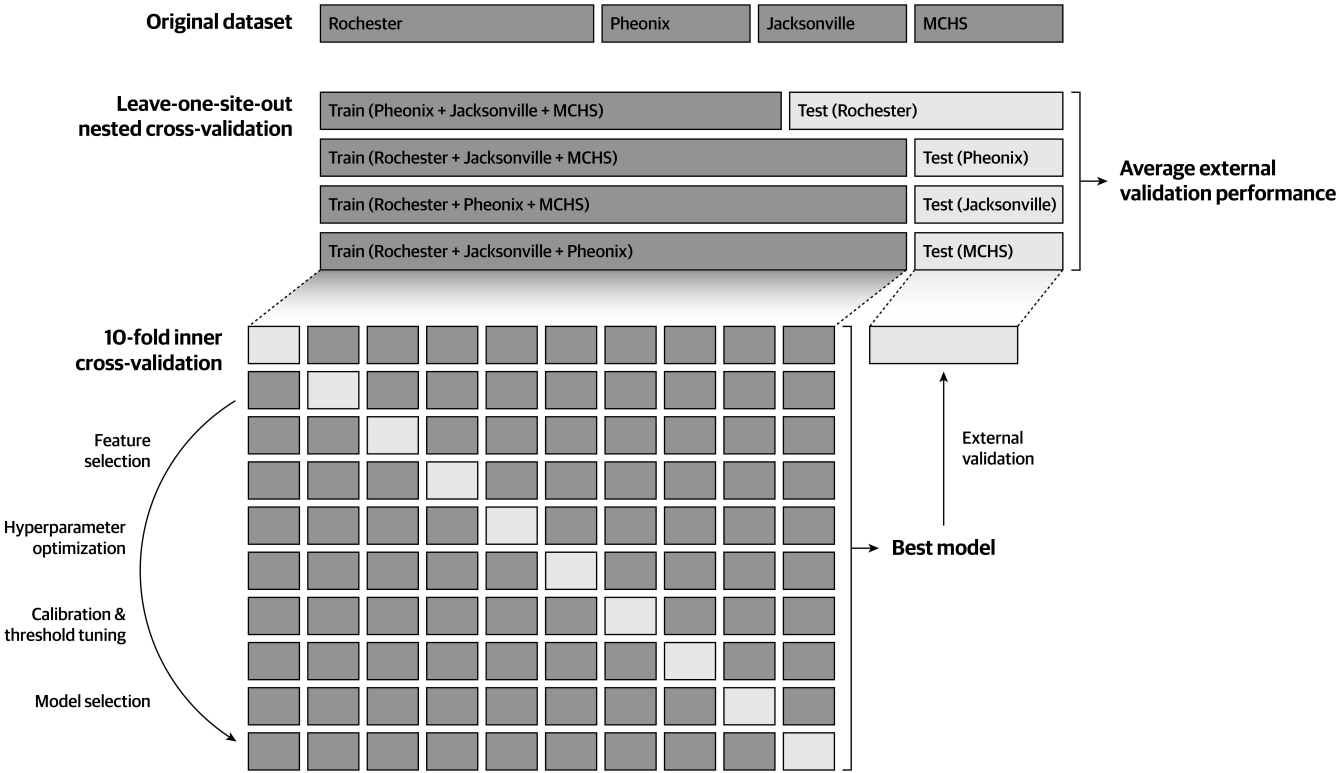
Abbreviations: ReLU Rectified Linear Unit.

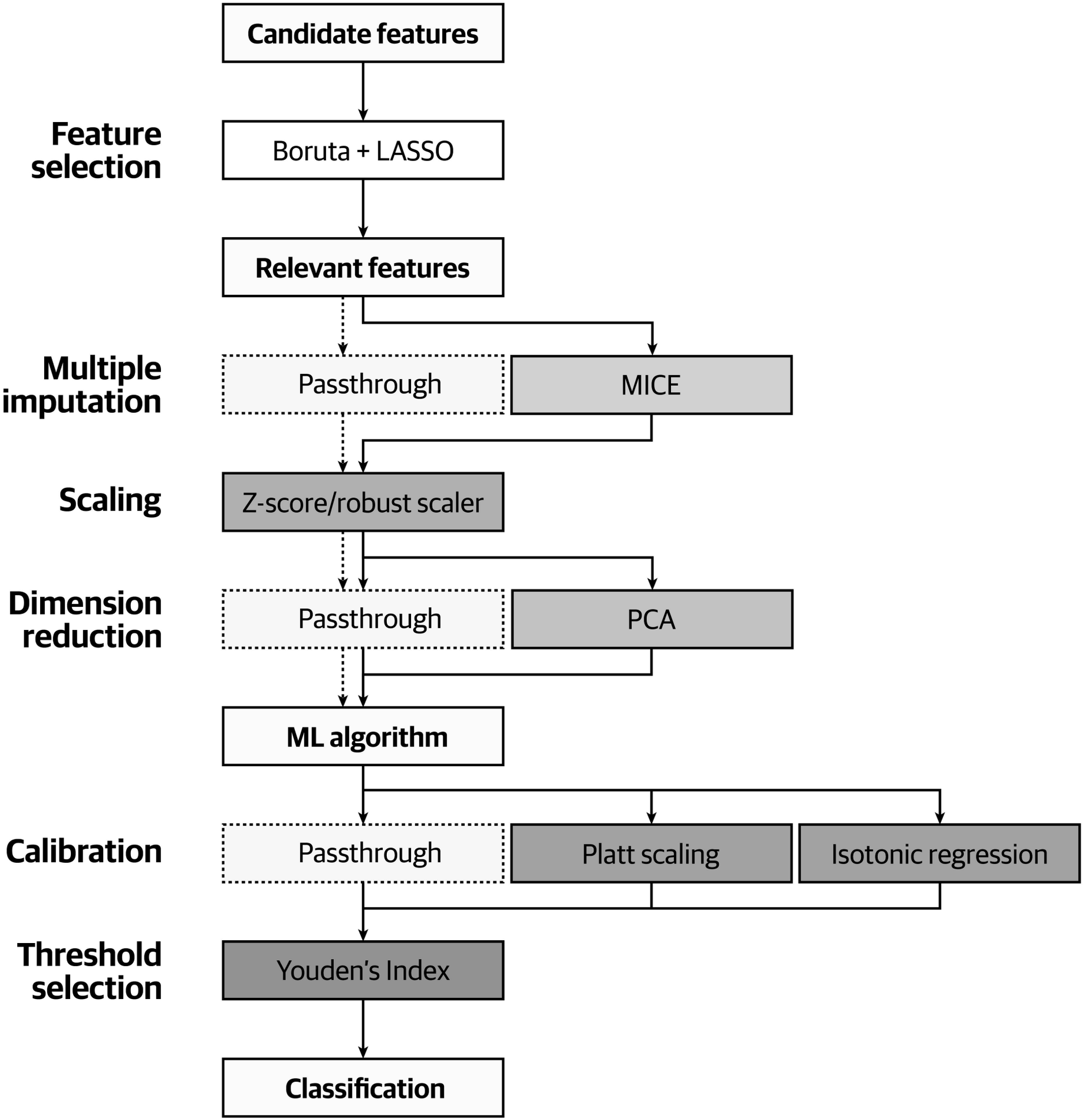
Table 1: Candidate features selected based on data availability and expert domain.

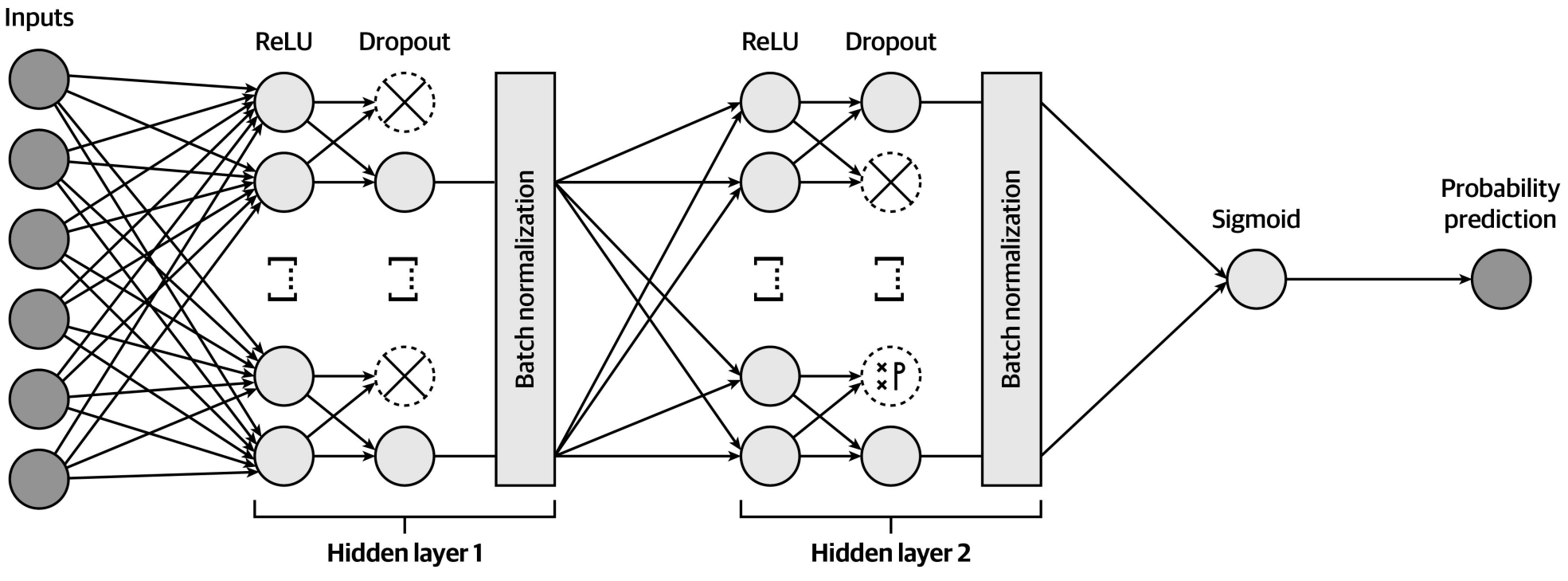
Feature	Description
Patient demographics and comorbidities	
Age at ICU admission	Continuous feature, in the unit of years
Sex	Binary feature: yes, no (male)
BMI	Continuous feature, in the unit of kg/m^2
History of hypertension	Binary feature: yes, no
History of diabetes mellitus	Binary feature: yes, no
Charlson Comorbidity Index	Continuous feature, unitless
Clinical characteristics at ICU admission	
ICU admission 24-hour respiratory SOFA score	Categorical feature: 0 to 4
ICU admission 24-hour CNS SOFA score	Categorical feature: 0 to 4
ICU admission 24-hour cardiovascular SOFA score	Categorical feature: 0 to 4
ICU admission 24-hour hepatic SOFA score	Categorical feature: 0 to 4
ICU admission 24-hour coagulation SOFA score	Categorical feature: 0 to 4
ICU admission 24-hour renal SOFA score	Categorical feature: 0 to 4
COVID-19 positive at ICU admission	Binary feature: yes, no
First lab values during ICU admission	
First WBC count in ICU	Continuous feature, in the unit of $x10^9/L$
First blood glucose in ICU	Continuous feature, in the unit of mmol/L
First serum lactate in ICU	Continuous feature, in the unit of mmol/L

Propofol and ventilation characteristics	
Initial propofol dose	Continuous feature, in the unit of mg/kg/hour, represents the average hourly propofol dose over the first 24 hours of infusion
Number of days from ICU admission to onset of IMV	Continuous feature, in the unit of days

Abbreviations: SOFA Sequential Organ Failure Assessment (Score), **CNS** Central Nervous System, **WBC** White Blood Cell (Count), **IMV** Invasive Mechanical Ventilation.


Original dataset
Rochester
Pheonix
Jacksonville
MCHS
Leave-one-site-out nested cross-validation
Train (Pheonix + Jacksonville + MCHS)
Test (Rochester)
Train (Rochester + Jacksonville + MCHS)
Test (Pheonix)
Train (Rochester + Pheonix + MCHS)
Test (Jacksonville)
Train (Rochester + Jacksonville + Pheonix)
Test (MCHS)
Average external validation performance
10-fold inner cross-validation
Feature selection
Hyperparameter optimization
Calibration & threshold tuning
Model selection
Best model
External validation


Candidate features
Feature selection
Boruta + LASSO
Relevant features
Multiple imputation
Passthrough
MICE
Scaling
Z-score/robust scaler
Dimension reduction
Passthrough
PCA
ML algorithm
Calibration
Passthrough
Platt scaling
Isotonic regression
Threshold selection
Youden's Index
Classification


Inputs
ReLU
Dropout
Batch normalization
ReLU
Dropout
Batch normalization
Sigmoid
Probability prediction
Hidden layer 1
Hidden layer 2