PharmaSUG 2025 - Paper PO-252

Methodology for AI-driven Outcome Prediction for Patients with Atrial Fibrillation After Transcatheter Aortic Valve Implantation (TAVI)

Felix Just, Daiichi Sankyo Europe, GmbH, Munich, Germany;
Krishna Padmanabhan, Cytel, Cambridge, MA, USA;
Parth Jinger, Cytel, Cambridge, MA, USA;
Amanda Borrow, Daiichi Sankyo, Inc., Basking Ridge, NJ, USA;
Rüdiger Smolnik, Daiichi Sankyo Europe, GmbH, Munich, Germany;
Eva-Maria Fronk, Daiichi Sankyo Europe, GmbH, Munich, Germany;
Neelam Yadav, Daiichi Sankyo Europe, GmbH, Munich, Germany;
Sergei Krivtcov, Daiichi Sankyo Europe, GmbH, Munich, Germany;
Stas Zakharkin, Daiichi Sankyo, Inc., Basking Ridge, NJ, USA;
Martin Unverdorben, Daiichi Sankyo, Inc., Basking Ridge, NJ, USA;
Asmir Vodencarevic, Daiichi Sankyo Europe, GmbH, Munich, Germany;

ABSTRACT

The identification of risk factors for adverse clinical outcomes using traditional statistical methods (e.g., Cox regression) has provided physicians with valuable insights for treatment decision-making. However, these methods are often restricted to capturing only linear associations within the data. Artificial Intelligence (AI) and Machine Learning (ML) overcome this limitation, offering the potential to uncover complex, non-linear relationships, thereby providing powerful new insights at the patient level.

We introduce an ML approach to develop predictive models for four clinical key endpoints from a randomized controlled trial: ischemic stroke, major gastrointestinal bleeding, major or non-major clinically relevant bleeding, and net adverse clinical events. Our methodology involves a pipeline where 10 ML algorithms are trained, optimized, and evaluated using nested cross-validation. The best model for each endpoint is chosen by its F1-score and validated on a separate hold-out set. The selected models are analyzed using the SHAP (SHapley Additive exPlanations) explainable AI framework to extract insights on the predictive factors of outcomes both at the cohort and patient levels.

On data from a randomized controlled trial, final models showed low to moderate F1 scores in the range from 0.08 to 0.39 and identified predictive factors that are largely consistent with previous clinical knowledge as well as potentially new ones. Due to the extreme imbalanced data, this performance is to be expected and in line with established medical risk scores. The applied methodology makes use of readily available Python packages and can easily be adapted to similar scenarios with minimal adjustments and represents a potential step towards patient-tailored treatment strategies.

INTRODUCTION

Patients with atrial fibrillation (AF) after a successful transcatheter aortic valve implantation (TAVI) procedure are at increased risk of ischemic stroke (IS) and major bleeding. A patient-tailored ML approach has the potential to support clinicians in their decision-making process to optimize anticoagulation treatment by providing better risk estimates and identifying new predictors of risk. Moreover, proposed methodology can identify in which direction identified predictors impact predicted risk.

In this work, we present a framework for training and evaluating predictive ML models for clinical endpoints (IS, major gastrointestinal bleeding [MGIB], net adverse clinical events [NACE], and major or clinically relevant nonmajor bleeding [CRNMB]) in a randomized controlled trial investigating Direct Oral Anticoagulants.

Our objective was to systematically assess a variety of ML algorithms, ranging in complexity from logistic regression to deep neural networks, for each clinical outcome. This approach acknowledges the inherent uncertainty in predicting which algorithm will perform best on a given dataset [1]. Nested cross-validation

was used for hyperparameter optimization and model selection. The best model for each outcome is then validated on a separate hold-out dataset. The methodology is designed to achieve good predictive power while remaining broadly applicable and reusable.

To identify the key predictors for each outcome and how they impact it, we employed the SHAP (SHapley Additive exPlanations) framework [2]. Inspired by game theory, SHAP provides a model-agnostic means of determining the contribution of input features to model predictions, potentially offering new clinical insights into the factors driving adverse events.

METHODOLOGY

DATA PREPROCESSING AND FEATURE ENGINEERING FOR MACHINE LEARNING ANALYSIS

In this study, we conducted a comprehensive data preprocessing pipeline to prepare a clinical dataset for our machine learning modeling and analysis. The raw dataset was stored in SAS format and contained 1377 patients with 180 variables of mixed data types. The variables were pre-selected from the study data by medical experts and required systematic cleaning and transformation before model development. A particular challenge were the low event rates for all outcomes (IS: 3%, MGIB: 6%, NACE: 18%, and MCRNB: 27%).

First, all object-type columns were UTF-decoded to standardize character encoding. We then filtered the dataset to only include patients receiving one of two different anticoagulation treatment options. Variables containing invalid entries were converted to missing values (NA) and subsequently cast to numeric format.

To ensure numerical consistency we converted several score-based columns into numeric data types and applied automated type conversion to optimize memory usage. Furthermore, all Yes/No categorical variables were standardized to a binary Y/N format to facilitate downstream encoding.

Feature selection was performed by removing homogeneous columns, free-text variables, and additional redundant features. We also treated missing values for categorical variables, following a systematic imputation approach where missing values were replaced based on domain-specific logic.

Subsequently, we implemented a missing data imputation strategy based on the data type. For numerical variables, we applied k-Nearest Neighbors Imputation (with k=3) followed by Min-Max scaling to normalize feature distributions. Categorical features underwent a label encoding transformation, imputation using k-NN (k=1), and inverse transformation back to categorical labels.

To mitigate multicollinearity, we identified highly correlated variables (Pearson correlation coefficient $|r| \geq 0.8$) and removed selected features based on domain expert recommendations while retaining clinically relevant ones. Finally, categorical variables were encoded using label encoding for binary features and one-hot encoding for nominal variables without inherent order.

This structured data preprocessing workflow resulted in a refined dataset with 155 numerical features, ensuring suitability for subsequent machine learning modeling. By implementing rigorous feature engineering, standardization, and imputation strategies, we aimed to enhance the robustness and interpretability of our predictive models.

MACHINE LEARNING-BASED PREDICTION OF CLINICAL ADVERSE EVENTS: METHODOLOGICAL FRAMEWORK

This study presents an ML-based approach to predicting the risk of major adverse events (MGIB, IS, MCRNB, and NACE) in a clinical trial setting, leveraging a comprehensive suite of classification models and rigorous evaluation strategies. The analysis pipeline was implemented after the previously described data preprocessing. Model training with nested cross-validation, hyperparameter optimization, and performance assessment using clinically relevant metrics were implemented. These are described in this section.

Data Splitting and Stratification

Initially, the dataset was randomly split into a 75% training and 25% holdout test set, ensuring stratification on the respective outcome to maintain class balance.

Model Development and Hyperparameter Tuning

A diverse set of ten classification models was implemented, ranging from traditional statistical methods to advanced ensemble techniques:

1. Logistic regression
2. L1-regularized logistic regression (Lasso)
3. L2-regularized logistic regression (Ridge)
4. Elastic Net
5. LDA
6. Naïve Bayes
7. Decision Tree
8. Random Forest
9. XGBoost
10. Multi-layer perceptron (MLP)

Given the imbalanced nature of the outcome (low % positive cases), several techniques were applied to mitigate bias, including class-weight adjustments in logistic regression-based models, prior probability adjustments in Naïve Bayes and Linear Discriminant Analysis (LDA), and scale weighting for XGBoost.

To ensure robust model selection, a nested cross-validation (CV) approach was applied, addressing the risk of data leakage and overly optimistic performance estimates that can arise from improper hyperparameter tuning. This methodology consists of two levels: an inner loop, responsible for hyperparameter optimization, and an outer loop, which provides an unbiased estimate of the model's generalization performance. By employing this strategy, model selection is based on the most reliable estimate of predictive performance rather than being influenced by fluctuations in training data splits.

In the inner CV loop, hyperparameter tuning was conducted using the hyperopt framework, which efficiently explores the hyperparameter space using Bayesian optimization. For each model, key parameters were optimized to achieve the best trade-off between bias and variance. For instance, logistic regression-based models tuned the regularization strength parameter (C), XGBoost and decision trees adjusted the depth of the trees (max_depth), and multi-layer perceptron (MLP) models explored different numbers of hidden layers to balance complexity and generalization. The best-performing hyperparameters were identified based on the F1 score, a suitable metric given the class imbalance in the dataset.

The outer CV loop then evaluated the best model configurations obtained from the inner loop. This evaluation involved training the selected model on the entire training set (excluding the holdout test set) and validating its performance on unseen folds to ensure robustness. This hierarchical cross-validation framework not only reduces the risk of overfitting to specific hyperparameter values but also provides a more realistic estimate of model performance on truly unseen data. By integrating nested CV, the study ensured that hyperparameter tuning did not inflate performance metrics and that the final model selection was rigorous and reproducible.

Model Evaluation and Interpretation

The performance of the best model for each of the outcomes was assessed using multiple clinically relevant metrics such as F1 score (primary metric, balancing precision and recall), Precision, Recall, and Specificity, Brier score (calibration measure) Area Under the Precision-Recall Curve (AUPRC) and Confusion matrices (with thresholds at 0.5 and at the optimized Youden index)

Beyond classification performance, interpretability was enhanced using SHAP values for model explanation, along with permutation importance to quantify the contribution of each feature. Additionally, learning curves were generated to assess the relationship between sample size and model performance, providing insights into potential data limitations and overfitting risks.

The pipeline was implemented using Python 3.12. and open-source packages, most notably scikit-learn 1.5.2. This methodological framework integrates robust ML techniques tailored to an imbalanced clinical dataset. The combination of nested cross-validation, hyperparameter tuning, and advanced model interpretation ensures a clinically meaningful and reproducible approach for predicting adverse events in clinical trials. The methodology is scalable and adaptable to other binary classification tasks in healthcare analytics.

RESULTS

Choice of model performance metric

The F1 score is a metric of choice for evaluating model performance when working with unbalanced outcomes since it provides a single measure that balances precision and recall and ensures that both false positives and false negatives are accounted for.

The best model was selected based on the F1 score for the hold-out independent test set to ensure robustness of the results. Overall, no single classifier consistently outperformed all others across all outcomes. Instead, the best-performing model varied depending on the specific outcome being evaluated. For MGIB the Logistic Lasso achieved the highest performance with the F1 score equal to 0.11. For Ischemic Stroke LDA (F1 = 0.08) performed the best. For CRNMB, Naïve Bayes (F1 = 0.39) was the top performer. For Net NACE, Logistic Regression (F1 = 0.33) achieved the best performance. These results highlight the importance of selecting the appropriate model based on the specific outcome rather than assuming a single model will work best in all scenarios.

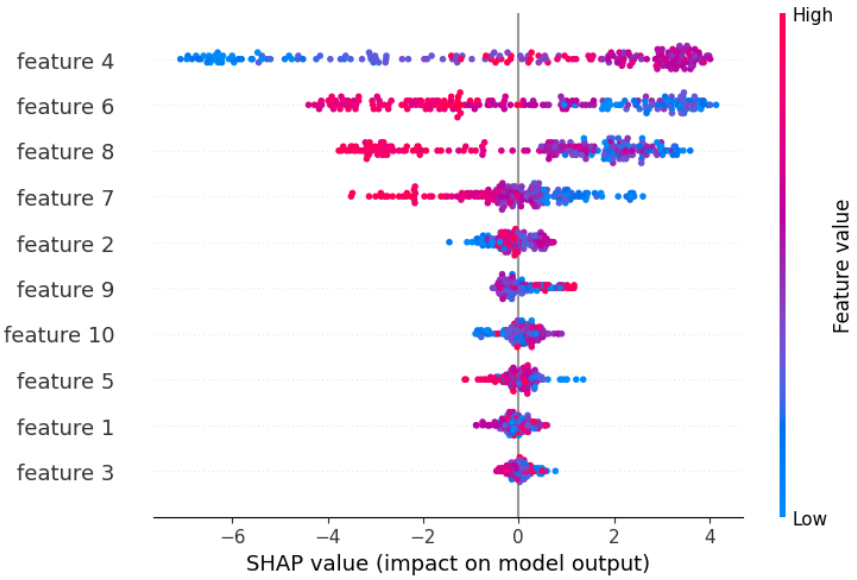
SHAP values

SHAP (SHapley Additive exPlanations) values are a game-theoretic approach to explainable AI, quantifying each feature's impact on model predictions [2]. They fairly distribute predictions among input features, offering both global and local insights into model behavior. SHAP values enhance transparency, aid feature selection, and support decision-making, making them particularly valuable in healthcare applications.

Below is a code example of how to calculate SHAP values and generate a beeswarm plot.

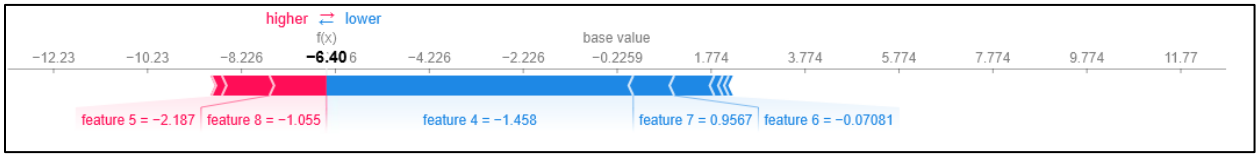
```
import shap

shap.initjs()

# take 100 points from the training set as background data
sampled_background_data = shap.sample(X_insample, 100)
explainer = shap.KernelExplainer(clf.predict_proba, sampled_background_data)
# shap values for positive class
shap_values = explainer.shap_values(X_holdout)[:, :, 1]
# beeswarm plot

shap.summary_plot(shap_values, X_holdout, feature_names = X.columns)
```

Figure 1 Beeswarm Plot example (artificial data)



A SHAP beeswarm plot visualizes feature importance, their impact on a model's predictions, and their variability across data points. Features are listed on the y-axis, ranked by their influence, while the x-axis represents the magnitude and direction of their impact—positive values increase predictions, and negative values decrease them. Each point represents a SHAP value for an individual observation, with color indicating feature magnitude (red for high values, blue for low values).

Figure 2. SHAP (Shapley Additive Explanations) Force Plot example (artificial data)



The plot explains how individual features influence a model's prediction by shifting it from a baseline output logit of -0.2259 to a final output of -6.40. By illustrating the cumulative effect of feature contributions, the SHAP Force Plot enhances model interpretability, providing valuable insights into the decision-making process. Features that enhance the prediction probability (favoring the positive class) are indicated by red arrows, while those that diminish it are shown with blue arrows. The length of each arrow reflects the magnitude of a feature's impact, with longer arrows indicating a stronger influence.

CONCLUSION

In conclusion, we have successfully demonstrated a comprehensive framework for the training, evaluation, and interpretation of machine learning models in predicting clinical outcomes. The observed F1 scores were low to moderate which can be easily explained by the relatively small sample size and, more importantly, by the low frequency of occurrences for the events of interest. While the identified key predictors offer potential for new clinical insights, it is crucial to consider the predictive power of the models when interpreting SHAP values.

Looking forward, it is essential to validate the preliminary results obtained from our data on a larger dataset to ensure robustness and reliability. Additionally, exploring the integration of synthetically generated data could enhance the pipeline and address current limitations.

Overall, we have illustrated a robust methodology for training ML models to predict clinical outcomes in a clinical trial. This adaptable framework holds promise for application in similar contexts, paving the way for advancements in clinical predictive modeling.

REFERENCES

[1] Wolpert, David H. 1996. “The Lack of A Priori Distinctions Between Learning Algorithms.” Neural Computation, 8(7):1341-1390.

[2] Lundberg, Scott M and Lee, Su-In. December 4, 2017. “A unified approach to interpreting model predictions.” NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. 4768 – 4777

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Felix Just
Daiichi Sankyo Europe, GmbH, Munich, Germany
felix.just@daiichisankyo.com