

ORIGINAL ARTICLE

Machine learning analysis of bleeding status in venous thromboembolism patients

Soroush Shahyari Fard¹  | Theodore J. Perkins^{1,2}  | Philip S. Wells^{1,3} 

¹The Ottawa Hospital Research Institute, the Ottawa Hospital, Ottawa, Ontario, Canada

²Department of Biochemistry, Microbiology and Immunology, University of Ottawa, Ottawa, Ontario, Canada

³Department of Medicine, University of Ottawa, Ottawa, Ontario, Canada

Correspondence

Philip S. Wells, The Ottawa Hospital, 501 Smyth Rd, PO Box 206, Ottawa, Ontario K1H 8L6, Canada.
Email: pwells@toh.ca

Handling Editor: Kristen Sanfilippo

Abstract

Background: Anticoagulation therapy is the mainstay of therapy for patients with venous thromboembolism (VTE). However, continuing or stopping anticoagulants after the first 3 to 6 months is a difficult decision that requires ascertainment of the risk of bleeding and recurrent VTE. Despite the development of several statistical models to predict bleeding, the benefit of machine learning (ML) models has not been investigated in depth.

Objectives: To assess the benefits of ML algorithms in bleeding risk evaluation in VTE patients and gain insight into their baseline information.

Methods: The baseline clinical, demographic, and genotype information was collected for 2542 patients with VTE who were on extended anticoagulation therapy. Six unsupervised dimensionality reduction and clustering ML algorithms were used to visualize and cluster the data for patients with major bleeding (118 patients) and nonbleeders. Eight supervised ML algorithms were trained and compared with the previously derived clinical models using a 5-fold nested cross-validation scheme.

Results: The baseline dataset for bleeders and nonbleeders showed a high degree of similarity. Two novel clusters were discovered within the dataset for bleeders based on the presence of isolated pulmonary embolism or isolated deep vein thrombosis, though the difference in bleeding risks was not statistically significant ($P = .32$). The supervised analysis showed that the ML and clinical models have similar discrimination (c -statistics, $\sim 62\%$) and calibration performance (Brier score, ~ 0.045).

Conclusion: The clinical variables recorded at baseline are not distinctive enough to improve bleeding prediction beyond the performance of the existing models, and other strategies or data modalities should be considered.

KEYWORDS

anticoagulants, calibration, hemorrhage, machine learning, venous thromboembolism

Essentials

- People with blood clots are treated with anticoagulants, which increases their bleeding risk.
- We developed machine learning algorithms to predict bleeding risk in these patients.
- The machine learning algorithms offered no prediction benefits compared with the available clinical tools.
- Different strategies should be investigated to improve bleeding risk prediction models.

1 | INTRODUCTION

Venous thromboembolism (VTE) is one of the leading causes of mortality [1]. According to the 2020 American Society of Hematology guidelines, after a brief acute management phase, patients with VTE should be treated with anticoagulants for at least 3 months, and then physicians need to decide if a secondary prevention phase is required to prevent the recurrence of VTE [2]. Unfortunately, making such a decision is difficult since physicians need to balance the risk of recurrence and the side effects of continuing anticoagulation therapy, most notably bleeding [3]. To date, more than 10 clinical predictive models have been developed to help physicians group patients into high and low bleeding risk groups; however, these models are based on simple statistical models and a small number of clinical predictor variables to make them easy to use and implement. As such, the peak performance of these models, as measured by their c-statistics, seems to have reached a limit of ~70% [4]. Meanwhile, the advances in machine learning (ML) and the wide availability of suitable hardware to run these models have raised the need to investigate if ML can improve bleeding prediction.

In addition to modeling the input features to predict a response, ML algorithms can also help explore, visualize, and find interesting relationships in a dataset. In particular, unsupervised ML algorithms have been developed for these purposes, and they can guide the downstream supervised ML analysis and help interpret their results [5]. In this study, we first attempted to explore a previously published [6] baseline dataset of patients with VTE who were on the extended phase of anticoagulation treatment using several unsupervised ML algorithms including dimensionality reduction algorithms and clustering analysis and then trained 8 supervised ML models and compared them with the previously developed clinical models including CHAP [6], HAS-BLED [7], VTE-BLEED [8], RIETE [9], ACCP [10], and OBRI [11]. Finally, based on both the unsupervised and supervised analyses, we provided some suggestions and next steps that could lead to better prediction models.

2 | METHODS

2.1 | Study design and dataset

The dataset consists of the baseline clinical, demographic, and genotype data collected for 2542 patients enrolled in a prospective cohort study (ClinicalTrials.gov: NCT00788736) over 8 years starting from September 2008. The patients were diagnosed with VTE (provoked

with minor transient risk factors or unprovoked) and were on anti-coagulant therapy for at least 3 months after diagnosis. Every 6 months, routine follow-up visits/phone calls were conducted to monitor their bleeding status. The detailed methodologies for data collection, selection process, and exclusion criteria were described previously [6]. We referred to the original patients' case report form or original medical record to revise the missing or corrupted data. They were kept as missing values when no informative data could be retrieved, but features containing more than 10% missing values were subsequently removed (Supplementary Table S2). The median and mode of the continuous and categorical features were used to impute the missing values in the remaining columns, respectively. The categorical features were one-hot encoded. Supplementary Table S1 lists all the baseline features, and Supplementary Methods include all the special preprocessing applied to the raw data.

2.2 | ML

Generally, ML algorithms can be divided into supervised and unsupervised algorithms. The supervised models are provided with the input variables and their associated labels, and their task is to model the relationship between the inputs and their associated labels. For example, features like age, sex, and diabetes status recorded at baseline could be the input variables, and the output label is whether or not the patients had a bleeding event at any time while enrolled in the study. The supervised models can ultimately be used to predict the bleeding risk for new patients. On the other hand, the unsupervised ML algorithms are only provided with the input variables, and their task is to identify patterns and relationships in the dataset without accessing the labels. Unsupervised learning methods are used in an exploratory data analysis phase [5]. In this paper, we used both unsupervised and supervised ML algorithms on the baseline dataset, which are explained below briefly. All the ML models were developed using scikit-learn [12] version 1.2.2, and all the 2-dimensional data manipulations were performed with pandas [13] version 1.5.3. All the figures were generated with matplotlib [14] version 3.7.1 and seaborn [15] version 0.12.2.

2.2.1 | Unsupervised learning algorithms

Dimensionality reduction algorithms can be used to visualize a high-dimensional dataset by creating lower-dimensional projections of the data [16]. For example, principal component analysis (PCA)

decomposes the data into its linear principal components (PCs) that explain the most variance [5]. In addition to PCA, we used kernel PCA [17], *t*-distributed stochastic neighboring embedding [18], and isometric mapping [19] to visualize the baseline dataset.

Clustering analysis is another type of unsupervised learning that is commonly used to group similar points into clusters. Herein, we used K-means clustering [20] and agglomerative clustering [5] to cluster the patients into 2 clusters representing patients with and without bleeding, and then we measured the quality of their clustering. Finally, agglomerative clustering was used to identify any notable subclusters within the dataset for patients with bleeding, and the risk of bleeding over time was estimated for each subcluster according to the Kaplan–Meier method [21]. Subsequently, a logistic regression model with the least absolute shrinkage and selection operator regularization was trained to determine the most relevant features that distinguish each subcluster. The optimal value of the regularization strength *C* was determined using 3-fold cross-validation (CV) to maximize accuracy. The implementation details of the dimensionality reduction and clustering algorithms are described in [Supplementary Methods](#).

2.2.2 | Supervised learning algorithms

Eight supervised learning algorithms were compared with the previously developed clinical models: logistic regression [5], linear discriminant analysis [5], quadratic discriminant analysis (QDA) [5], Gaussian Naïve Bayes [5], support vector machine [22], random forest [23], adaptive boosting (AdaBoost) [24], and gradient boosting [25]. The details for these models are briefly described in [Supplementary Methods](#). Also, a “Dummy” classifier was trained as a standard control, which either returned the most frequent class or returned predictions based on the class distributions in the training dataset. Finally, we compared the above ML algorithms with the modified versions of the previously derived clinical models: CHAP [6], HAS-BLED [7], VTE-BLEED [8], RIETE [9], ACCP [10], and OBRI [11]. The list of the predictor variables in the original and the modified versions of the clinical models had been published previously [6].

CV was used to estimate the generalization performance of the models on new data. Thus, the supervised ML and the clinical models were compared using a 5-fold nested CV scheme, which consists of an inner loop and an outer loop, as shown in [Figure 1](#); the outer loop is a 5-fold CV in which the dataset is first divided into 5 stratified sets where all the sets contain the same proportions of bleeders and nonbleeders as the original dataset. Then, 1 set is kept aside for calibration and testing, while the model is trained on the remaining sets. Note that the calibration and testing sets are disjoint, but they have similar sizes. After training, Platt scaling [26] was used to calibrate each uncalibrated model on the calibration set; therefore, each model is paired with a calibrator that maps its output into a calibrated number between 0 and 1 (note that the same process was used to calibrate the clinical models but without training them on the training set). This process was repeated 5 times with each set, producing an estimate of generalization performance on

new data along with a CI. However, the ML models have some hyperparameters that must be specified before training, and since the choice of hyperparameters strongly affects their performance, hyperparameter optimization was performed using 3-fold CV on the training sets in the inner loop via grid search method; ie, the best combination of hyperparameters (with highest mean of area under receiver operating characteristic curve [AUROC]) was used to train the model on the training set and then was calibrated and tested as shown in the right panel of [Figure 1](#). [Supplementary Table S3](#) lists the hyperparameter space for the supervised ML models.

When there is a relatively high number of features compared with the number of samples, the supervised ML algorithms face a problem formally known as the curse of dimensionality; ie, as the number of features (dimensions) increases, it becomes harder for the ML algorithms to generalize to meaningful data, and instead, they are likely to be influenced by noise [27]. As such, 5 strategies were used to select the best feature sets for each iteration of 5-fold nested CV during the hyperparameter tuning process; PCA was used to obtain 5 or 10 PCs that explain the most variance; another technique was sequential forward feature selection, where the models select the best 5 or 10 features through an iterative process using 3-fold CV where optimizing AUROC is the CV’s goal. Finally, the models had the option to not use any of the above techniques and use all the features.

2.3 | Performance metrics

The unsupervised methods used in this paper are purely for visual and exploratory purposes and do not require any objective performance measure. However, the goodness of clustering was calculated using homogeneity and completeness metrics [28]. Both metrics range from 0 to 1, where 1 is the best clustering, and 0 is the worst clustering quality.

Discrimination and calibration metrics were measured to compare the models. In particular, AUROC and area under precision-recall curve (AUPRC) were used to measure the discriminative abilities of the models. All the calibrated models produce values from 0 to 1, which can be interpreted as the probability of bleeding for the patients, and by changing the threshold that defines bleeders and nonbleeders, the precision and sensitivity (or recall) of the models can be modified to create precision-recall curves, and the AUPRC could be calculated to compare the models.

In addition, the calibration performance of the models was measured using 4 metrics; Brier score [29] measures the mean squared difference between the predicted probabilities and the actual outcome, and it ranges between 0 and 1, with a smaller value indicating a better score. Brier score can be decomposed into its discrimination and calibration components, also known as reliability and resolution, respectively, where lower reliability and higher resolution values are better [30]. Finally, Cox’s slope and intercept were calculated by regressing a linear model to the probability outputs and the binary outcomes; a perfectly calibrated model will have a slope of 1 and an intercept of 0, and a deviation from the perfect line can be interpreted as a lower calibration.

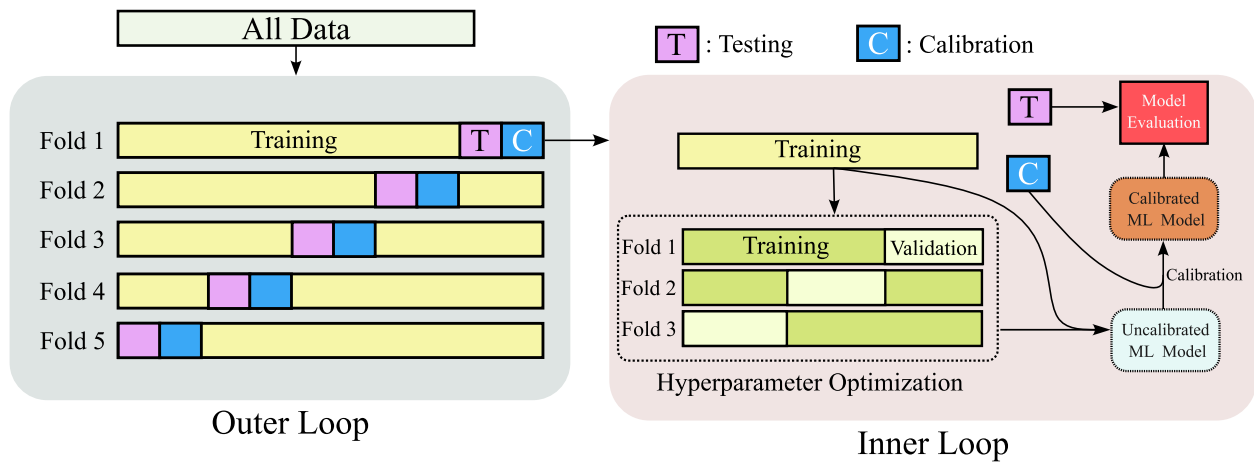


FIGURE 1 The 5-fold nested cross-validation (CV) scheme. The outer loop represents the 5-fold CV, and the inner loop represents the hyperparameter tuning via 3-fold CV and grid search along with the calibration and testing process. ML, machine learning.

2.4 | Statistical analysis

The Kaplan–Meier estimates of bleeding risk for the 2 detected clusters were compared using the log-rank test [21]. The clinical relevance of each of the selected features was compared between the 2 clusters using the chi-squared test for the categorical variables and the Student’s *t*-test for the continuous variables. The resulting *P* values were corrected with the Benjamini–Hochberg [31] correction for multiple testing, and the significance level was .05.

The models were compared across each metric using Friedman’s test [32], which tests the null hypothesis that the mean performances of all the models are similar. When the null hypothesis is rejected, there is at least 1 pair of classifiers with statistically significantly different performance. Therefore, multiple post hoc pairwise Wilcoxon signed-rank tests [33] were performed to detect the pairs of classifiers with performances that were statistically significantly different. Benjamini–Hochberg method was used to correct for multiple hypothesis testing, and the significance level .05 was used for all the tests. Friedman’s and Wilcoxon signed-rank tests were implemented in SciPy [34] version 1.10.1, the Benjamini–Hochberg method was implemented in statsmodels [35] version 0.14.0, and Kaplan–Meier estimates with the log-rank test were implemented in scikit-survival [36] version 0.21.0.

3 | RESULTS

3.1 | Data visualization

Overall, 4.6% of the patients (118 individuals) had major bleeding. Figure 2 shows the resulting 2-dimensional plots of the baseline dataset as projected by the dimensionality reduction algorithms, where the orange and black circles represent the bleeders and non-bleeders, respectively. Figure 2A shows a high degree of overlap between the bleeders and nonbleeders in the first 2 dimensions of the

PCA plot, which explains 7.9% and 7.4% of variability in the baseline dataset, respectively. Additionally, it shows that the first 40 PCs can explain >95% of the variability in the dataset, while the remaining 57 PCs account for 5% of the variance. Figure 2B shows the resulting scatter plot for kernel PCA, isometric mapping, and *t*-distributed stochastic neighboring embedding algorithms, which also show a high degree of overlap between the bleeders and nonbleeders. Although these results do not rule out that certain select patient characteristics may be predictive of bleeding outcomes, they suggest that most features and/or the overall patterns of variability seen in patients’ features are not correlated to bleeding.

3.2 | Unsupervised clustering of the baseline dataset

Figure 3 shows the clusters obtained from the K-means and agglomerative clustering algorithms, which are visualized on the PCA plot. Both algorithms were used to group the points into 2 clusters without having access to their labels. If the baseline information for the bleeders and nonbleeders was completely distinct in the high-dimensional feature space, the clustering algorithms would be able to discern the 2 classes; however, the resulting clusters did not correspond to the bleeding status given their small homogeneity and completeness metrics. Despite the low performance of both algorithms, the agglomerative clustering algorithm performed more than 10 times better than the K-means algorithm as it had 10 times higher homogeneity and completeness values.

3.3 | Identifying novel clusters in patients with bleeding

Agglomerative clustering was used to explore any useful pattern within the dataset for patients with bleeding, and Figure 4A shows the

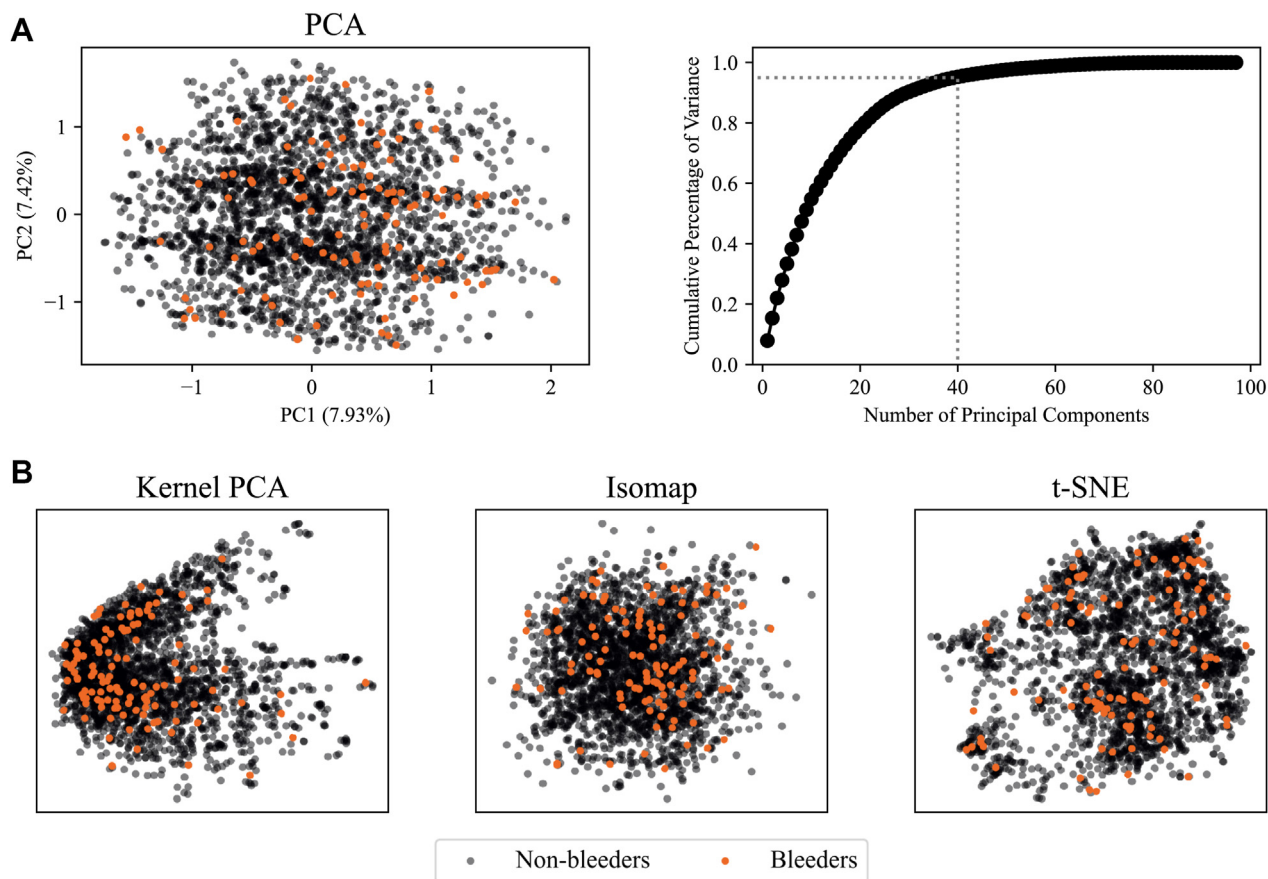


FIGURE 2 Scatter plots of the dimensionality reduction algorithm projections applied on the baseline dataset. (A) The principal component analysis (PCA) scatter plot of the bleeding dataset and the cumulative percentage of variance as a function of the number of principal components (PCs), and (B) the resulting scatter plot for kernel PCA, isometric mapping (Isomap), and *t*-distributed stochastic neighboring embedding (*t*-SNE). The dotted line in panel (A) indicates that the >95% explained variance occurs after 40 PCs. Bleeders are plotted on top of the nonbleeders to make them visually discernable.

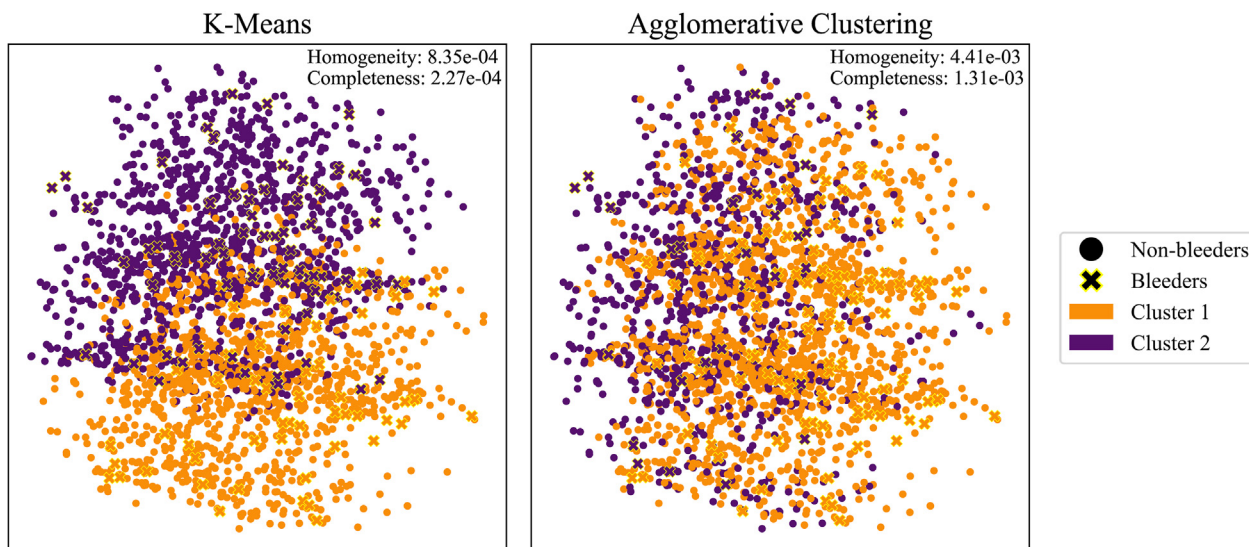


FIGURE 3 Scatter plots of the K-means and agglomerative clustering on the baseline dataset plotted on the principal component analysis plot. For the K-means algorithm, *K* (number of clusters) was set to 2, and for the agglomerative clustering, 2 clusters were selected. On the principal component analysis plots, different colors represent the clusters identified by the algorithms, and the actual nonbleeders and bleeders are drawn using filled circles and “x” markers, respectively.

resulting dendrogram obtained from this analysis where the dendrogram was cut at distance 9 from the base to separate the bleeders into 2 clusters: cluster 1 with 62 patients and cluster 2 with 56 patients. To understand the underlying patient characteristics that gave rise to the 2 clusters, a logistic regression with least absolute shrinkage and selection operator regularization was trained to distinguish the 2 clusters, and [Figure 4B](#) shows the non-zero coefficients for the logistic regression model; the features with positive coefficients are positively correlated to the patients in cluster 1 and negatively correlated to the patients in cluster 2, and vice versa.

The table in [Figure 4B](#) lists the frequencies and mean of categorical and continuous variables, respectively. Most patients (59.7%) in cluster 1 were females, and 83.9% of them had isolated deep vein thrombosis (DVT). On the other hand, cluster 2 consisted mostly of male patients (62.5%), and most of them (62.5%) had isolated pulmonary embolism (PE). However, statistical tests showed that there were only 5 features that were significantly different between the 2 clusters: isolated DVT, isolated PE, postthrombotic syndrome, heterozygous *CYP4F2* mutation, and wild-type *VKORC1639*. The Kaplan–Meier estimates of the probability of bleeding for each cluster are shown in [Figure 4C](#), which shows that for the patients in cluster 2, the probability of bleeding is higher than for patients in cluster 1; however, the log-rank test showed that the difference is not statistically significant ($P = .32$).

3.4 | Predicting bleeding in VTE patients using supervised ML algorithms

The [Table](#) summarizes the mean and SD of each metric for the models obtained from the 5-fold nested CV experiment. QDA had the highest mean AUROC and AUPRC, but its AUPRC was similar to that of the CHAP. Furthermore, all the models had similar Brier scores, but HAS-BLED had the lowest and the best score. CHAP and random forest had the 2 highest (and best) resolutions, while gradient boosting and Dummy had the lowest (and worst) resolutions. However, OBRI and Dummy classifiers had the lowest (and best) reliability scores, while CHAP and Gaussian Naïve Bayes had the highest (and worst) reliabilities. VTE-BLEED and OBRI had the best slopes (closest to 1), while random forest, AdaBoost, logistic regression, and Dummy had negative (and worst) slopes. Furthermore, OBRI and VTE-BLEED had the best intercept (closest to zero), while random forest and AdaBoost had the worst intercepts.

Friedman's P values suggest that P values for resolution and reliability scores are statistically significantly different among at least 1 pair of classifiers. However, pairwise Wilcoxon signed-rank tests did not find any significant difference between any pairs of classifiers for these 2 metrics as shown in [Supplementary Figure S1](#).

As explained in section 2.2.2, the sequential forward feature selection method can be used to choose the most distinctive features for classification. For each ML model, features that were chosen at least twice during the 5-fold nested CV experiment were pooled together, and [Figure 5](#) shows the features that were deemed important by at

least 3 ML algorithms. Recent provoked VTE due to hospitalization was the most useful feature chosen by all the 8 ML models. Furthermore, the number of concomitant medications, *CYP2C9* polymorphism, and use of antiplatelet agents were the next best features.

4 | DISCUSSION

We explored the potential advantages of ML algorithms to analyze the baseline dataset of VTE patients. The unsupervised dimensionality reduction algorithms suggested that the overall information content within the baseline dataset does not have a strong correlation to the bleeding status, and there is a high degree of overlap between bleeders and nonbleeders. Furthermore, we were able to identify 2 clusters within the patients with major bleeding, which differed mainly based on the type of their VTE. We were not able to see any benefit in using supervised ML algorithms compared with the conventional statistical models in predicting major bleeding from patients' baseline information.

4.1 | High overlap of baseline information for bleeders and nonbleeders

Visualizing the high-dimensional baseline information could be insightful. For instance, when the points from each group are well separated on the plots, one could expect a good performance from the supervised ML models. However, the lack of such clear distinction does not necessarily mean a challenging classification task, as there might still be one or several sets of distinctive features that were not captured by the unsupervised models. Nonetheless, the lack of separation in the plots ([Figure 2](#)) suggests that the overall variabilities of the features in the dataset are not strongly correlated to the bleeding status. Furthermore, despite the categorical nature of most features, the correlation heatmap ([Supplementary Figure S2](#)) and the high cumulative explained variance by 40 PCs of PCA suggest the existence of a few strongly correlated features in the dataset. Both clustering analyses had poor performance, indicating the difficulty of the algorithms to distinguish between bleeders and nonbleeders, which is likely due to large class imbalance, especially for the K-means clustering algorithm, which often generates clusters with uniform sizes [37], as shown in [Figure 3](#).

4.2 | Clustering the patients with bleeding into 2 distinct clusters

The agglomerative clustering revealed 2 clusters in the baseline dataset for patients with bleeding: cluster 1, which mostly consists of females who had isolated DVT and experienced postthrombotic syndrome, and cluster 2, which consists of males who had isolated PE, wild-type *VKORC1639*, and heterozygous *CYP4F2* mutation. However, the 2 most distinctive features of cluster 1 and cluster 2 were

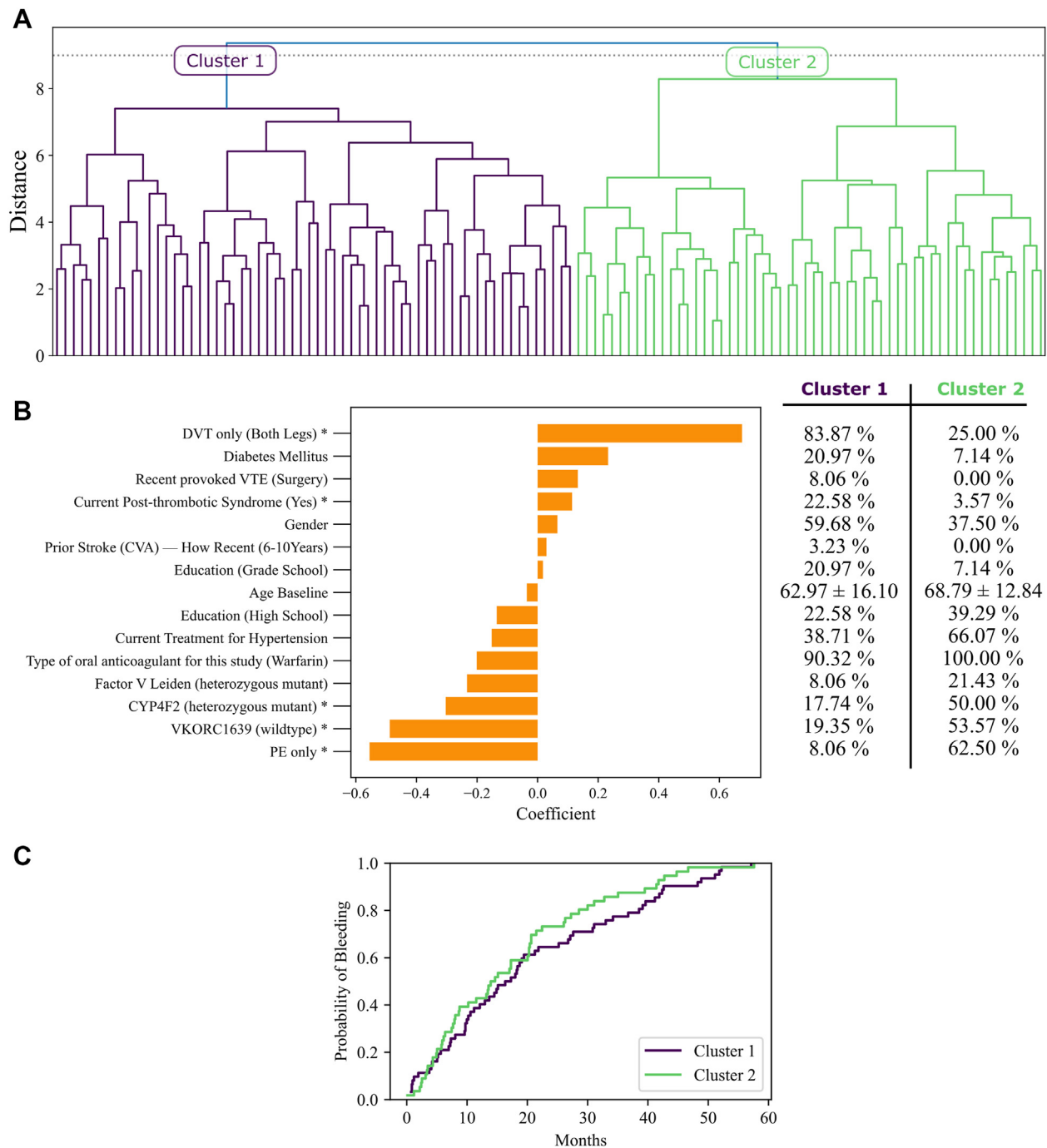


FIGURE 4 Cluster analysis of the baseline dataset for patients with bleeding. (A) The dendrogram from agglomerative clustering, which was cut at distance 9 to create 2 clusters, (B) the coefficients of a least absolute shrinkage and selection operator-regularized logistic regression model trained to separate cluster 1 and cluster 2 along with the frequencies for categorical variables and mean with SD for continuous variables, and (C) Kaplan-Meier estimates of the probability of bleeding for patients in cluster 1 and cluster 2. There are 62 patients in cluster 1 and 56 patients in cluster 2. The features with statistical significance difference (adjusted $P < .05$) are indicated with an asterisk (*). CVA, cerebral vascular accident; DVT, deep vein thrombosis; PE, pulmonary embolism; VTE, venous thromboembolism.

isolated DVT and isolated PE, respectively. To our knowledge, no study has yet investigated the difference in bleeding risk between patients with isolated PE and isolated DVT who are in the extended phase of anticoagulation therapy; the Kaplan-Meier curves illustrated that the patients in cluster 1 had a lower probability of bleeding

compared with the patients in cluster 2, but the log-rank test showed that the difference was not statistically significant. Moreover, the literature does not support the association of VKORC1 and CYP4F2 variants, which are the defining features of cluster 2, with the risk of major bleeding [38].

TABLE Summary of the metrics measured to compare machine learning and clinical models.

Bleeding risk score	AUROC	AUPRC	Brier score	Resolution	Reliability	Slope	Intercept
CHAP	0.61 (0.15)	0.13 (0.07)	4.52×10^{-2} (1.26×10^{-3})	4.49×10^{-3} (2.53×10^{-3})	4.69×10^{-3} (1.38×10^{-3})	-0.61 (3.98)	0.08 (0.19)
ACCP	0.66 (0.04)	0.10 (0.03)	4.43×10^{-2} (5.48×10^{-4})	2.64×10^{-3} (1.56×10^{-3})	1.90×10^{-3} (1.11×10^{-3})	1.31 (0.55)	-0.01 (0.02)
RIETE	0.63 (0.11)	0.08 (0.03)	4.50×10^{-2} (1.24×10^{-3})	2.80×10^{-3} (1.39×10^{-3})	2.83×10^{-3} (1.55×10^{-3})	1.82 (1.80)	-0.04 (0.08)
VTE-BLEED	0.65 (0.05)	0.08 (0.02)	4.45×10^{-2} (4.80×10^{-4})	1.85×10^{-3} (1.16×10^{-3})	1.43×10^{-3} (7.13×10^{-4})	1.15 (0.42)	-0.01 (0.02)
HAS-BLED	0.66 (0.06)	0.11 (0.05)	4.41×10^{-2} (8.96×10^{-4})	2.77×10^{-3} (2.38×10^{-3})	1.87×10^{-3} (1.53×10^{-3})	3.36 (2.08)	-0.10 (0.09)
OBRI	0.65 (0.03)	0.08 (0.02)	4.45×10^{-2} (6.76×10^{-4})	1.05×10^{-3} (8.96×10^{-4})	5.66×10^{-4} (3.91×10^{-4})	0.85 (0.32)	0.01 (0.02)
Logistic regression	0.58 (0.14)	0.08 (0.04)	4.47×10^{-2} (3.82×10^{-4})	2.24×10^{-3} (1.02×10^{-3})	1.98×10^{-3} (8.47×10^{-4})	-2.07 (6.04)	0.15 (0.29)
LDA	0.66 (0.06)	0.09 (0.02)	4.45×10^{-2} (4.64×10^{-4})	2.40×10^{-3} (1.48×10^{-3})	1.91×10^{-3} (1.07×10^{-3})	2.22 (1.76)	-0.06 (0.08)
QDA	0.67 (0.10)	0.13 (0.07)	4.44×10^{-2} (5.11×10^{-4})	3.33×10^{-3} (2.01×10^{-3})	2.70×10^{-3} (1.59×10^{-3})	2.56 (2.15)	-0.07 (0.10)
SVC	0.65 (0.05)	0.11 (0.05)	4.47×10^{-2} (2.91×10^{-4})	2.35×10^{-3} (1.30×10^{-3})	2.04×10^{-3} (1.06×10^{-3})	1.85 (1.03)	-0.04 (0.05)
Gaussian NB	0.66 (0.09)	0.11 (0.03)	4.56×10^{-2} (1.89×10^{-3})	3.31×10^{-3} (2.61×10^{-3})	4.21×10^{-3} (2.13×10^{-3})	3.40 (3.75)	-0.11 (0.18)
Random forest	0.60 (0.14)	0.11 (0.07)	4.46×10^{-2} (8.34×10^{-4})	3.50×10^{-3} (2.41×10^{-3})	3.00×10^{-3} (1.57×10^{-3})	-8.10 (18.78)	0.40 (0.82)
AdaBoost	0.54 (0.14)	0.07 (0.02)	4.49×10^{-2} (2.31×10^{-4})	2.36×10^{-3} (1.65×10^{-3})	2.25×10^{-3} (1.74×10^{-3})	-3.61 (9.21)	0.22 (0.44)
Gradient boosting	0.58 (0.08)	0.08 (0.03)	4.49×10^{-2} (2.15×10^{-4})	9.60×10^{-4} (8.79×10^{-4})	8.57×10^{-4} (6.90×10^{-4})	1.70 (1.20)	-0.03 (0.06)
Dummy	0.49 (0.02)	0.05 (0.00)	4.51×10^{-2} (1.54×10^{-4})	1.00×10^{-4} (1.28×10^{-4})	1.99×10^{-4} (2.58×10^{-4})	-0.65 (1.30)	0.08 (0.06)
Friedman's <i>P</i> value	.320	.065	.197	.021	.002	.110	.146

Mean scores (SD) are written for each metric.

AdaBoost, adaptive boosting; AUPRC, area under precision-recall curve; AUROC, area under receiver operating characteristic curve; Gaussian NB, Gaussian Naïve Bayes; LDA, linear discriminant analysis; QDA, quadratic discriminant analysis; SVC, support vector machine.

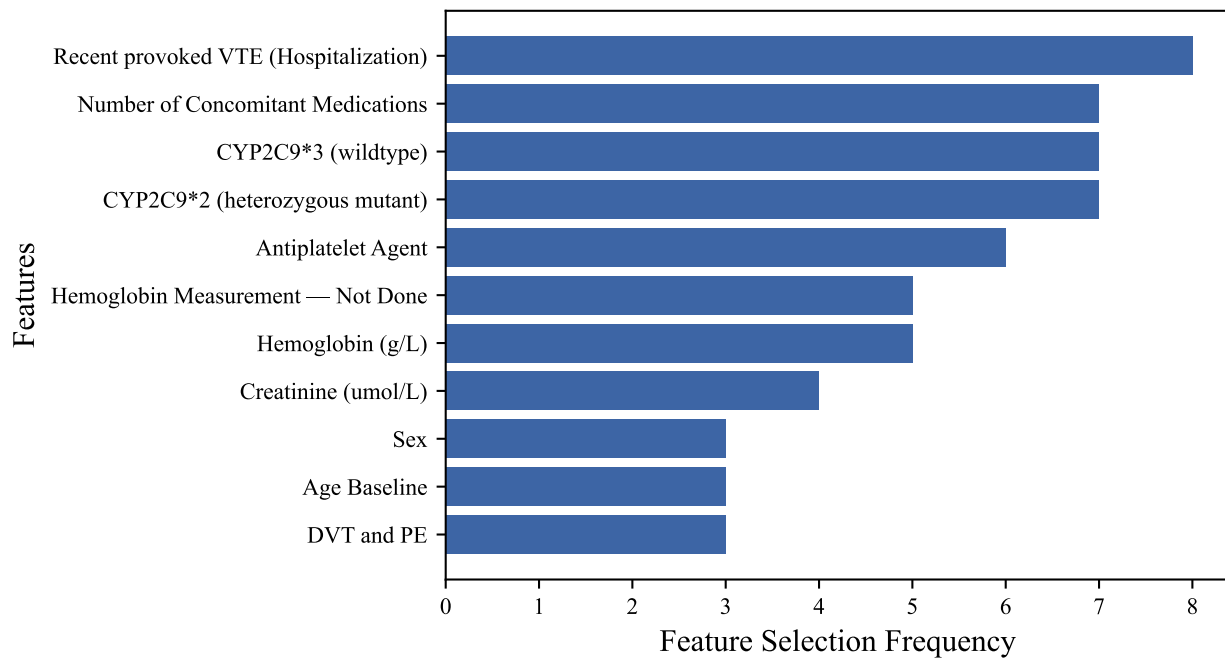


FIGURE 5 The most distinctive features in the baseline dataset for bleeding risk prediction. Frequency indicates the number of machine learning models that chose a given feature at least twice during the 5-fold nested cross-validation experiment. DVT, deep vein thrombosis; PE, pulmonary embolism; VTE, venous thromboembolism.

The high prevalence of postthrombotic syndrome in cluster 1 is likely due to the presence of isolated DVT compared with cluster 2 with isolated PE [39,40]. A recent meta-analysis has shown that patients with isolated PE had a lower incidence of prothrombin gene and factor V Leiden mutation, but they were more likely to be female, have diabetes mellitus, and have recent invasive surgery [41]. However, in our study, the patients in cluster 2 who were mainly characterized by having isolated PE, despite lack of statistical significance, had a higher prevalence of heterozygous factor V Leiden mutation, had a lower incidence of diabetes mellitus compared with patients in cluster 1, and were mostly males, and none of them had recent surgery, consistent with postsurgical patients not being enrolled in our study. Thus, the genotypic and phenotypic differences for patients characterized in each cluster warrant further meta-analysis as the evidence suggests multivariable characteristics of each cluster and the presence of the correlated features (Supplementary Table S4).

4.3 | Similar predictive performance of ML algorithms and clinical models

C-statistic or AUROC is commonly used in clinical studies to evaluate the discrimination power of the models; however, some argue that AUPRC is a better predictor of discrimination performance, especially when the data are imbalanced [42], as it can capture the precision that represents the difficulty in correctly identifying the bleeders without making false positive predictions. Despite the lack of statistical significance difference, QDA had the highest AUPRC and AUROC compared with the other ML models, indicating its superior discrimination

performance, while tree-based models, such as gradient boosting and AdaBoost, had the worst performance. A nested CV scheme was chosen to compare the models since it is less biased than the regular CV scheme [43]. However, the 5 different heterogeneous test sets with small sizes (~11 bleeders and 240 nonbleeders) led to high variance in the performance of the ML models and limited us to use less powerful nonparametric tests, which resulted in no significant *P* value.

The low and optimistic Brier scores for the models were due to their strong calibration components (small reliability values) rather than their weak discrimination components (small resolution values); therefore, the Brier score is not a good metric when there is a class imbalance such as in our dataset [44]. The calibration slope and intercept showed that, overall, the clinical models are better calibrated than the ML models. The ML models such as logistic regression and random forest with negative regression slopes were poorly calibrated as the bleeding probabilities were inversely related to patients' risks. Furthermore, their regression intercept, which is the measure of mean calibration or calibration-in-the-large, was larger than 0, indicating their overall tendency to underestimate the risks. On the other hand, models such as QDA, linear discriminant analysis, and support vector machine with regression slopes >1 overestimated the risk for the low-risk patients and underestimated the risk for high-risk patients, and they had intercepts smaller than 0, which indicates their proclivities to overestimate the risks. The poor calibration of the ML models is likely due to our small, imbalanced dataset [45]. Overall, QDA had the best discrimination and calibration performance compared with the other models, likely because QDA was less prone to overfitting as it had fewer parameters and made less stringent assumptions about the data.

The feature selection strategy revealed that recently provoked VTE associated with hospitalization is the most important feature in predicting bleeding risk. Furthermore, CYP2C9 polymorphism is the second most important feature that is known to affect warfarin dosing and bleeding risk [46]. None of the currently used clinical models rely on CYP2C9 polymorphism, which could be investigated in the future. Furthermore, features such as the use of antiplatelet agents, hemoglobin level, creatinine level, sex, and age were commonly used features between the clinical and ML models. However, other features used in the clinical models such as prior gastrointestinal bleeding, prior stroke, and hypertension medications were not deemed important by the ML models. Nonetheless, pinpointing the important features should be preferably carried out in an external validation dataset to prevent biased results that are not generalizable; thus, these results should be further investigated.

4.4 | Limitations

There are some limitations in this paper. Firstly, the relatively small number of major bleeds in the dataset lowered the capacity of the ML models to generalize and may have contributed to their similar performance and hindered the use of independent test sets and the utilization of more powerful parametric statistical testing. In addition, the presence of many categorical features has introduced sparsity in the dataset, which increased models' training time and hindered generalization. Although these are the inevitable limitations of any clinical dataset, data augmentation strategies could be used to balance the dataset before ML analysis, and their benefits need to be investigated [47]. Furthermore, the modified versions of the clinical models were used to compensate for the missing features in our dataset that could also affect their performance, but this is unavoidable given our dataset. Finally, we imputed the missing values using the median and mode of the values as this is the most straightforward approach. However, imputation using regression or classification models for the important features can reduce noise.

4.5 | Next steps and suggestions

The overall results from the literature [4] and our attempt at using a baseline dataset to predict bleeding risk have illustrated its challenging nature as the best models can achieve the AUROC, or c-statistics, of around 70%. As far as we are aware, the only other study [48] that developed ML algorithms for bleeding risk prediction was able to achieve an estimated AUROC of around 63%, which is on par with the performance of the models we developed. We tried to understand why there has been no significant improvement in the prediction models beyond the current state of the art, and we have hypothesized 3 main reasons for such stagnation in the performance based on our findings. First, the variables that are recorded and measured at baseline may not be predictive of the bleeding status. Therefore, we suggest expanding the variety of the predictor variables

that are measured at baseline and incorporating other data modalities, such as imaging data, clinical symptoms, signs, etc., into the baseline risk prediction models. Secondly, the bleeding events may occur because of changes that take place after the baseline visit, and therefore, the baseline clinical information might not be informative enough to predict bleeding over time. Finally, the bleeding events for some patients may occur randomly without any clinical predetermination. Although it is hard to prove this point, it should be considered as a reason for the lack of improvement in bleeding risk prediction.

ACKNOWLEDGMENTS

We would like to thank Dr Maria de Winter for her invaluable insights in data preparation and preprocessing. This research was enabled in part by support provided by the Digital Research Alliance of Canada (alliancecan.ca).

FUNDING

This research received no external funding.

ETHICS STATEMENT

All patients provided informed consent. The study was approved by the Institutional Review Boards at each site that contributed patients to the study.

AUTHOR CONTRIBUTIONS

S.S.F. designed and ran the experiments, analyzed the data, interpreted the results, and wrote the manuscript. T.J.P. designed the study, interpreted the results, and revised the manuscript. P.S.W. initiated, designed, and supervised the study; interpreted the results; and revised the manuscript.



RELATIONSHIP DISCLOSURE

There are no competing interests to disclose.

DATA AVAILABILITY

The datasets used in this paper can be shared upon reasonable request to the corresponding author. The code to reproduce the results and the saved models are publicly available in the GitHub link <https://github.com/sorshf/Bleeding-Risk-Prediction-from-Baseline>.

ORCID

Soroush Shahryari Fard  <https://orcid.org/0000-0001-6445-7622>
Philip S. Wells  <https://orcid.org/0000-0002-8657-8326>

X, FORMERLY KNOWN AS TWITTER

Theodore J. Perkins  @PerkinslabC
Philip S. Wells  @PhilWellsMD1

REFERENCES

- [1] Wendelboe AM, Raskob GE. Global burden of thrombosis: epidemiologic aspects. *Circ Res*. 2016;118:1340–7.

- [2] Ortel TL, Neumann I, Ageno W, Beyth R, Clark NP, Cuker A, et al. American Society of Hematology 2020 guidelines for management of venous thromboembolism: treatment of deep vein thrombosis and pulmonary embolism. *Blood Adv.* 2020;4:4693–738.
- [3] Klok FA, Huisman MV. How I assess and manage the risk of bleeding in patients treated for venous thromboembolism. *Blood.* 2020;135:724–34.
- [4] De Winter MA, Van Es N, Büller HR, Visseren FLJ, Nijkeuter M. Prediction models for recurrence and bleeding in patients with venous thromboembolism: a systematic review and critical appraisal. *Thromb Res.* 2021;199:85–96.
- [5] Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning.* New York: Springer; 2009.
- [6] Wells PS, Tritschler T, Khan F, Anderson DR, Kahn SR, Lazo-Langner A, et al. Predicting major bleeding during extended anticoagulation for unprovoked or weakly provoked venous thromboembolism. *Blood Adv.* 2022;6:4605–16.
- [7] Pisters R, Lane DA, Nieuwlaar R, de Vos CB, Crijns HJGM, Lip GYH. A novel user-friendly score (HAS-BLED) to assess 1-year risk of major bleeding in patients with atrial fibrillation: the Euro Heart Survey. *Chest.* 2010;138:1093–100.
- [8] Klok FA, Hösel V, Clemens A, Yollo WD, Tilke C, Schulman S, et al. Prediction of bleeding events in patients with venous thromboembolism on stable anticoagulation treatment. *Eur Respir J.* 2016;48:1369–76.
- [9] Ruíz-Giménez N, Suárez C, González R, Nieto JA, Todolí JA, Samperiz AL, et al. Predictive variables for major bleeding events in patients presenting with documented acute venous thromboembolism. Findings from the RIETE Registry. *Thromb Haemost.* 2008;100:26–31.
- [10] Kearon C, Akl EA, Ornelas J, Blaivas A, Jimenez D, Bounameaux H, et al. Antithrombotic therapy for VTE disease: CHEST Guideline and Expert Panel Report. *Chest.* 2016;149:315–52.
- [11] Beyth RJ, Quinn LM, Landefeld CS. Prospective evaluation of an index for predicting the risk of major bleeding in outpatients treated with warfarin. *Am J Med.* 1998;105:91–9.
- [12] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *Mach Learn Res.* 2011;12:2825–30.
- [13] The Pandas Development Team. pandas-dev/pandas: Pandas; 2023. <https://doi.org/10.5281/ZENODO.7794821>
- [14] Caswell TA, Lee A, De Andrade ES, Droettboom M, Hoffmann T, Klymak J, et al. matplotlib/matplotlib: REL: v3.7.1; 2023. <https://doi.org/10.5281/ZENODO.7697899>
- [15] Waskom ML. seaborn: statistical data visualization. *J Open Source Softw.* 2021;6:3021. <https://doi.org/10.21105/joss.03021>
- [16] Sarker IH. Machine learning: algorithms, real-world applications and research directions. *SN Comput Sci.* 2021;2:160. <https://doi.org/10.1007/s42979-021-00592-x>
- [17] Schölkopf B, Smola A, Müller KR. Kernel principal component analysis. In: Gerstner W, Germond A, Hasler M, Nicoud J-D, eds. *Artificial Neural Networks – ICANN '97.* 1327. Berlin, Heidelberg: Springer; 1997:583–8.
- [18] van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res.* 2008;9:2579–605.
- [19] Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. *Science.* 2000;290:2319–23.
- [20] Arthur D, Vassilvitskii S. k-means++: the advantages of careful seeding. In: *Proceedings of the 18th annual ACM-SIAM symposium on Discrete algorithms (SODA '07).* Philadelphia, PA: Society for Industrial and Applied Mathematics; 2007:1027–35.
- [21] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc.* 1958;53:457–81.
- [22] Crammer K, Singer Y. On the algorithmic implementation of multi-class Kernel-based vector machines. *J Mach Learn Res.* 2002;2:265–92.
- [23] Breiman L. Random forests. *Mach Learn.* 2001;45:5–32.
- [24] Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci.* 1997;55:119–39.
- [25] Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat.* 2001;29:1189–232.
- [26] Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: *Proceedings of the 22nd international conference on Machine learning (ICML '05).* New York, NY: Association for Computing Machinery; 2005:625–32.
- [27] Bengio Y, Delalleau O, Roux N. The curse of highly variable functions for local Kernel machines. In: Weiss Y, Schölkopf B, Platt J, eds. *Advances in Neural Information Processing Systems.* 18. Cambridge, MA: MIT Press; 2005:107–14.
- [28] Rosenberg A, Hirschberg J. V-measure: a conditional entropy-based external cluster evaluation measure. In: *Conference on Empirical Methods in Natural Language Processing.* Prague, Czech Republic: Association for Computational Linguistics; 2007:410–20.
- [29] Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev.* 1950;78:1–3.
- [30] Murphy AH. A new vector partition of the probability score. *J Appl Meteorol.* 1973;12:595–600.
- [31] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol.* 1995;57:289–300.
- [32] Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J Am Stat Assoc.* 1937;32:675–701.
- [33] Wilcoxon F. Individual comparisons by ranking methods. *Biom Bull.* 1945;1:80–3.
- [34] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods.* 2020;17:261–72.
- [35] Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with python. Scipy.org. In: *9th Python in Science Conference.* 2010:57–61.
- [36] Pölsterl S. scikit-survival: a library for time-to-event analysis built on top of scikit-learn. *J Mach Learn Res.* 2020;21:1–6.
- [37] Xiong H, Wu J, Chen J. K-means clustering versus validation measures: a data-distribution perspective. *IEEE Trans Syst Man Cybern B Cybern.* 2009;39:318–31.
- [38] Kawai VK, Cunningham A, Vear SI, Van Driest SL, Oginni A, Xu H, et al. Genotype and risk of major bleeding during warfarin treatment. *Pharmacogenomics.* 2014;15:1973–83.
- [39] Bova C, Rossi V, Ricchio R, Greco A, Bloise A, Daniele F, et al. Incidence of post-thrombotic syndrome in patients with previous pulmonary embolism. A retrospective cohort study. *Thromb Haemost.* 2004;92:993–6.
- [40] Kahn SR. The post-thrombotic syndrome. *Hematology Am Soc Hematol Educ Program.* 2016;2016:413–8.
- [41] Ten Cate V, Prochaska JH, Schulz A, Nagler M, Robles AP, Jurk K, et al. Clinical profile and outcome of isolated pulmonary embolism: a systematic review and meta-analysis. *EClinicalMedicine.* 2023;59: 101973. <https://doi.org/10.1016/j.eclinm.2023.101973>
- [42] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One.* 2015;10:e0118432. <https://doi.org/10.1371/journal.pone.0118432>
- [43] Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics.* 2006;7:91. <https://doi.org/10.1186/1471-2105-7-91>

- [44] Wallace BC, Dahabreh IJ. Improving class probability estimates for imbalanced data. *Knowl Inf Syst.* 2014;41:33–52.
- [45] Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J.* 2014;35:1925–31.
- [46] Aithal GP, Day CP, Kesteven PJ, Daly AK. Association of polymorphisms in the cytochrome P450 CYP2C9 with warfarin dose requirement and risk of bleeding complications. *Lancet.* 1999;353:717–9.
- [47] Mohammed R, Rawashdeh J, Abdullah M. Machine learning with oversampling and undersampling techniques: overview study and experimental results. In: 2020 11th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan. IEEE; 2020:243–8.
- [48] Mora D, Mateo J, Nieto JA, Bikdeli B, Yamashita Y, Barco S, et al. Machine learning to predict major bleeding during anticoagulation for venous thromboembolism: possibilities and limitations. *Br J Haematol.* 2023;201:971–81.

SUPPLEMENTARY MATERIAL

The online version contains supplementary material available at <https://doi.org/10.1016/j.rpth.2024.102403>.