Comparing Effectiveness of Machine Learning Methods for Diagnosis of Deep Vein Thrombosis

Ruslan Sorano(✉), Lars V. Magnusson, and Khurshid Abbas

Østfold University College, Halden, Norway
{ruslans,lars.v.magnusson,khurshid.abbas}@hiof.no

Abstract. This paper presents the results of a comparative study of machine learning techniques when predicting deep vein thrombosis. We used the Ri-Schedule dataset with Electronic Health Records of suspected thrombotic patients for training and validation. A total of 1653 samples and 59 predictors were included in this study.

We have compared 20 standard machine learning algorithms and identified the best-performing ones: Random Forest, XGBoost, GradientBoosting and HistGradientBoosting classifiers. After hyper-parameter optimization, the best overall accuracy of 0.91 was shown by GradientBoosting classifier using only 15 of the original variables.

We have also tuned the algorithms for maximum sensitivity. The best specificity was offered by Random Forests. At maximum sensitivity of 1.0 and specificity of 0.41, the Random Forest model was able to identify 23% additional negative cases over the screening practice in use today.

These results suggest that machine learning could offer practical value in real-life implementations if combined with traditional methods for ruling out deep vein thrombosis.

Keywords: Deep vein thrombosis · DVT · VTE · Machine learning

1 Introduction

Deep vein thrombosis (DVT) is a common medical disorder that is caused by a blood clot, also called a thrombus, forming in any section of the venous system. The disease can quickly become lethal if a thrombus dislodges from the location where it coagulated and enters the blood circulation system, obstructing the blood flow at other locations. This often results in pulmonary embolism (PE), when a thrombus blocks a lung artery, affecting the oxygen supply to the cells in the whole body. Timely diagnosis of DVT and PE, commonly known as venous thromboembolism (VTE), is vital to prevent patient's disability or death.

The most common tools for VTE pre-testing are risk assessment models, including Wells [33] and Revised Geneva [22].

These integer-based scoring systems are time and cost-efficient, but their results can be inaccurate. More invasive and expensive clinical testing, such as

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
O. Gervasi et al. (Eds.): ICCSA 2022 Workshops, LNCS 13381, pp. 279–293, 2022.
https://doi.org/10.1007/978-3-031-10548-7_21

compression ultrasound (CUS) or computed tomography (CT) is often required to get a conclusive diagnosis.

In the past two decades, the growing volumes of Electronic Health Records (EHR) and other digital patient data have enabled the extensive application of machine learning (ML) techniques in all areas of medicine, including prediction and diagnosis of DVT and PE. However, the comparative study of the performance of ML algorithms on patient data from the Norwegian healthcare system has not been previously performed.

1.1 Research Objectives and Questions

The data used in this research was collected by Østfold Hospital Trust during the Rivaroxaban for Scheduled Work-up of Deep Vein Thrombosis Study, or Ri-Schedule study [17]. The Ri-Schedule dataset has 1653 suspected DVT patient records registered by a Norwegian regional hospital. Neither this dataset nor its combination of features were previously used in any machine learning based predictions of deep vein thrombosis.

The primary objective of this study was to compare the predictive abilities of a larger number of ML algorithms on the Ri-Schedule dataset.

We would also like to see how ML-based method can compare in performance with the traditional DVT classification model currently used as a standard solution.

To achieve these objectives, we have defined the following research questions:

RQ1: How do state-of-the-art ML algorithms compare in accuracy on the Ri-Schedule dataset?
RQ2: How do state-of-the-art ML algorithms compare in terms of specificity when maximizing sensitivity?
RQ3: To what degree can machine learning models improve existing diagnostic procedures?

In order to answer these questions, we will compare a wide selection of modern machine learning techniques, and we will ensure that we evaluate the techniques fairly.

2 Related Work

Ma et al. [25] trained the XGBoost classification model using a dataset comprising 70 features and 9213 patients, 1165 of which were VTE positive. The model's predictive abilities were found to be comparable to or better than traditional assessment methods.

Wang et al. [32] utilized the ML approach to assess the risk of VTE using the clinical EHR of 376 patients with 22 risk factor variables. The number of samples with positive and negative outcomes was evenly balanced with 188 labels of each class. Nine classification algorithms, including RF, GBDT, LR, KNN, SVM, NB, XGBoost, EBM and EBM-inter, were trained and their performance

was compared with each other and with Padua risk assessment model. The ML model performance was also validated on a holdout set including EHRs for 42 patients with and 1537 without VTE. RF was shown to be the best performing model, but its sensitivity results were inferior to those of Padua model.

Nafee et al. [26] developed two ensemble learning algorithms by combining estimates across multiple ML models to predict VTE in acutely ill patients. The ensemble models were built with 39 candidate ML methods in five families of models, such as generalized additive models, elastic net, extreme gradient boosting, RF and Bayesian logistic regression. The dataset comprised 6459 samples, including 6052 with negative and 407 positive DVT outcome. The first ensemble model was trained using 68 predictors, the second model used a subset of 16 variables, identified to have association with VTE. The experiments have shown that ML-based solution outperformed the traditional risk score model in predicting VTE from randomized clinical trial data.

In their recent research, Luo et al. [24] trained XGBoost, SVM, GLM and RF machine learning models to predict DVT based on the records for 518 patients, including 189 with confirmed DVT diagnoses. Among the four classification models, GLM and SVM showed the largest ROC AUC, which was used as the primary performance metric.

These studies were conducted used limited amount of techniques, and dataset sizes were either in the range of several thousand or a few hundred samples. To the best of our knowledge, there isn't a comparative study of ML-based DVT prediction on a dataset from the Norwegian healthcare system.

3 The Setup of the Comparative Study

This section describes the framework for Ri-Schedule dataset preparation and comparison of the performance of classification methods in the prediction of DVT. Figure 1 illustrates the general design of the research methodology framework.

3.1 Data Source

The electronic medical records were collected by Østfold Hospital Trust during the Rivaroxaban for Scheduled Work-up of Deep Vein Thrombosis Study (Ri-Schedule study). The study included patients with suspected DVT referred to the Emergency Department of Østfold Hospital, Norway, between February 2015 and November 2018. The other inclusion criteria were age not younger than 18 years old, agreement to sign written consent and no recent prior enrollment in the study.

The original Ri-Schedule dataset contains 1800 patient records, with 195 attributes per record. The dataset contains both numerical and categorical data. The data includes personal characteristics (age, gender, weight, height), clinical symptoms, risk factors, vital signs, laboratory test results, limb measurements, information from the follow-up visits, prescribed medication and diagnostic outcomes.

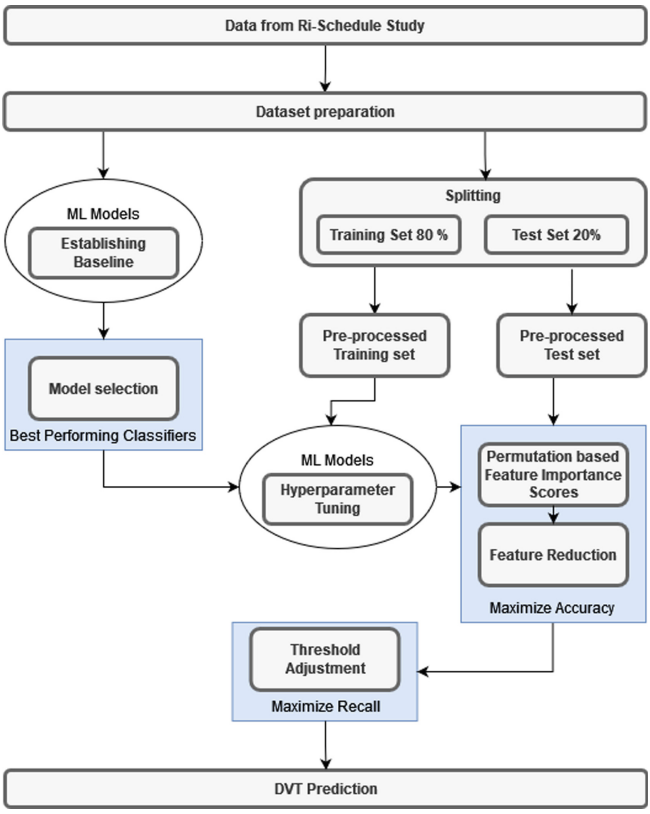


Fig. 1. Framework design for DVT prediction on Ri-Schedule dataset.

Target Variable. The labels are in binary format, representing positive or negative DVT diagnosis. The ground truth values were determined during the Ri-Schedule study based on D-dimer [20] attribute values and confirmed with the compression ultrasonography (CUS) test.

3.2 Dataset Preparation

Here we describe how we have prepared the data for machine learning analysis.

Duplicates. Each data sample in Ri-Schedule dataset contains an identification number that is unique to a patient. Some of the records appear to be duplicate entries, referring to the same patient ID. These duplicate entries were compared with each other to find out if they had identical medical data. We kept the latest entry for each patient ID, which contained the most complete or up-to-date information. After removing duplicate records, the dataset contained 1653 samples. The target attribute had 1342 negative and 311 positive values.

Feature Selection. We removed some of the attributes from the original dataset due to the following factors: (a) irrelevance to this research, (b) highly sparse data, (c) follow-up information for positively diagnosed patients, (d) conflicting data values, such as Wells risk scores from three different examiners. The two variables containing circumference measurements of left and right knees were replaced with a new variable containing absolute values of the difference in these measurements. This transformation makes sense, as an individual measurement of a limb does not present predictive value for diagnosing DVT. The non-zero difference between the two values, however, may indicate a swollen limb, which is one of the possible signs of a blood clot. The same procedure was applied to the pairs of measurements for left and right ankles. After removing empty, sparse or irrelevant features, the dataset contained 1653 samples and 59 independent variables.

Categorical Data. We converted categorical features to dummy variables. This transformation creates a binary representation of predictor categories and replaces the variable with $k - 1$ binary features, where k is number of original categories. The number of predictors increased to 197 as we converted categorical features to dummy variables.

Missing Values and Outliers. The Ri-Schedule dataset contains a large number of missing values. Discarding the samples with incomplete data is not a feasible option due to the relatively small dataset size. Extreme values confirmed as valid by medical experts were kept unchanged. Those of erroneous or accidental nature were removed from the dataset. We replace missing, invalid and extreme outlier values using univariate imputation of means, medians or modes, depending on each attribute's nature and valid range. As a general rule, the missing or invalid values were imputed with means when the predictor values were real numbers. The modes were imputed into columns containing binary values. Some of the features contained limited sets of integer values; these columns were imputed with medians.

Processed Dataset. The final class distribution is unbalanced, with a negative to positive label ratio of 0.81:0.19. We adjust class weights using the balanced mode of the relevant hyper-parameter provided in algorithm implementation.

The transformed dataset is randomly split into training (80%) and test (20%) sets. We maintained the same distribution of classes using labels for stratified sampling. After splitting, the training and test sets contained 1322 and 331 samples, respectively. The negative to positive ratios in training and test subsets remained approximately the same 0.81:0.19.

The value magnitudes in different independent variables within the dataset vary greatly. We use standardization to transform features to a common value range [0, 1].

3.3 Data Analysis

Performance Metrics. We use overall classification accuracy for training and selecting the best models. We also used balanced accuracy, sensitivity, precision, specificity, f1-score, f2-score, ROC AUC (Receiver Operator Curve – Area Under Curve), PRC AUC (Precision-Recall Curve – Area Under Curve) for finding the final results.

Machine Learning Methods. Twenty classification methods from different families of ML algorithms were selected to compare their performance in classifying DVT using the Ri-Schedule dataset. The following classification models were used: AdaBoost (AB) [29], Bagging Classifier (BC) [4], Decision Tree (DT) [7], Extra Tree Classifier (ET) [7], Extra Trees Classifier (ETs) [5], GaussianNB (GNB) [8], GaussianProcess (GP) [35], Gradient Boosting (GB) [15], HistGradient Boosting (HGB) [21], K-Neighbors Classifier (KN) [27], Linear Discriminant Analysis (LDA) [14], Linear SVC (LSVC) [2], Logistic Regression (LR) [19], MLP Classifier (MLP) [1], Passive Aggressive Classifier (PA) [12], Quadratic Discriminant Analysis (QDA) [30], Random Forest (RF) [6], Ridge Classifier (RC) [18], SGD Classifier (SGD) [28], SVM Classifier (SVC) [11] and XGBoost Classifier (XGB) [9]. Nineteen models were used from the Scikit-learn package, one from the XGBoost package. Each model was initialized with default parameters, trained and validated on the entire dataset using 10-fold stratified split and cross-validation. This procedure was repeated 5 times, and the average accuracy across all training-validation cycles was calculated. The models that showed top 4 accuracy results were selected for further parameter optimization.

Hyper-parameter Tuning. The hyper-parameter search was implemented on the training dataset using grid search with stratified 5-fold cross-validation. The classification accuracy metric was chosen as the primary metric for predictive performance evaluation of models at this stage of our experiments.

The best performing model was then fit and validated on the training set using five rounds of 10-fold stratified split and cross-validation. The mean and standard deviation of the cross-validation results were reported for accuracy, balanced accuracy, recall, precision, specificity, f1- and f2-scores, ROC AUC and PRC AUC.

Feature Importance. We inspected feature importance for our fitted estimators using the model agnostic permutation technique. We fitted models on the training set and then estimated how important each feature is for a particular model.

Individual features were permuted one at a time, and the model performance was estimated on the holdout test set. The model performance metrics reflect the loss of relationship between the permuted feature and the output variable so that we can arrange and output the predictors in the order of their importance

for each trained classifier. The procedure is repeated ten times, and the mean and standard deviation of the importance scores are reported.

In the next step we created N subsets of training and test datasets, containing top n most importance features for an estimator, n varied from 1 to N, where N is number of independent variables with non-zero importance scores. The mode was then refitted on each training subset and evaluated on a test subset. The final model and corresponding data subset were selected based on the results of classification metrics. The feature names and their importance scores were then reported for each classifier.

Threshold Adjustment. In the medical domain the tolerance for false negatives is often much lower than for false positives. In the case of DVT prediction, it is more acceptable to send a patient to an extra CUS testing to confirm a negative result, than to discharge a patient with thrombosis who was falsely identified as negative. We explored different values for classification decision thresholds to find the value that gives predictions with zero number of false negatives. In the case of an imbalanced dataset, it is expected that the default cut-off value for the binary classification threshold of 0.5 does not deliver satisfactory results. We iterated over decision thresholds, trying all values from 1.0 to 0.0 with the step of 0.01, and we recorded the cut-off point when the recall reaches 100%. When we did not find an optimal classification decision threshold value that gives a perfect recall score, we registered threshold as equal to zero. In the next step, we counted the number of patients correctly identified as negative and whose D-dimer value is above or equal to the default cut-off point of 0.5 lg mL^{-1} [16]. The ratio of this number to the total number of patients sent to CUS with a D-dimer ≥ 0.5 lg mL^{-1} gives us the estimated savings in CUS testings that our models can propose.

3.4 Hardware and Software

Our project was developed in Python 3.8.10. The machine learning component was implemented in Scikit-learn v.1.0, XGBoost v.1.4.2. The experiments were run on HPC with $\times 2$ AMD EPYC 7551 32-Core CPUs with 2 threads per core. Operating system Ubuntu 20.04.1.

4 Results

This section presents the results of our comparative study.

4.1 Baseline Results

In the baseline performance test, the top result was shown by HistGradientBoosting Classifier, followed closely by Random Forest, and XGBoost and GradientBoosting classifier sharing the third place. The summary is presented in Table 1.

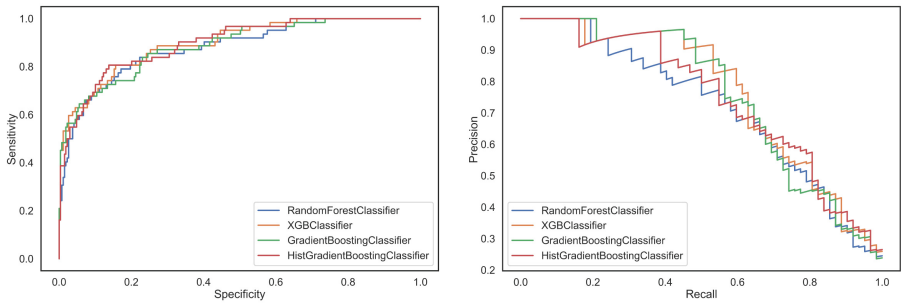
4.2 Results After Hyper-parameter Tuning

The four selected algorithms that showed the best baseline results were optimized using hyper-parameter tuning.

Table 1. Baseline performance metrics on complete dataset. Averaged results with standard deviation shown after 5 rounds of 10-fold stratified split and cross-validation.

Model	Accuracy	Recall	Precision	Specificity
AB	0.853 (0.025)	0.498 (0.109)	0.652 (0.094)	0.936 (0.026)
BC	0.858 (0.020)	0.431 (0.089)	0.707 (0.082)	0.959 (0.019)
DT	0.813 (0.030)	0.520 (0.103)	0.505 (0.093)	0.879 (0.030)
ET	0.772 (0.036)	0.356 (0.091)	0.383 (0.061)	0.860 (0.036)
ETs	0.852 (0.015)	0.260 (0.083)	0.847 (0.118)	0.989 (0.009)
GNB	0.244 (0.016)	0.943 (0.038)	0.192 (0.006)	0.081 (0.021)
GP	0.833 (0.011)	0.127 (0.055)	0.899 (0.217)	0.997 (0.005)
GB	0.863 (0.021)	0.460 (0.104)	0.715 (0.086)	0.956 (0.018)
HGB	0.865 (0.020)	0.468 (0.099)	0.718 (0.083)	0.956 (0.017)
KN	0.815 (0.018)	0.159 (0.062)	0.530 (0.144)	0.967 (0.014)
LDA	0.856 (0.019)	0.371 (0.081)	0.740 (0.100)	0.969 (0.016)
LSVC	0.860 (0.017)	0.392 (0.077)	0.752 (0.095)	0.968 (0.017)
LR	0.861 (0.018)	0.344 (0.082)	0.814 (0.115)	0.980 (0.013)
MLP	0.850 (0.023)	0.441 (0.097)	0.659 (0.096)	0.941 (0.027)
PA	0.813 (0.111)	0.571 (0.256)	0.658 (0.226)	0.844 (0.250)
QDA	0.261 (0.024)	0.896 (0.046)	0.190 (0.008)	0.114 (0.033)
RF	0.864 (0.015)	0.329 (0.079)	0.866 (0.101)	0.988 (0.010)
RC	0.858 (0.016)	0.317 (0.081)	0.822 (0.105)	0.983 (0.012)
SGD	0.851 (0.030)	0.476 (0.173)	0.702 (0.133)	0.944 (0.087)
SVC	0.788 (0.031)	0.595 (0.094)	0.454 (0.062)	0.833 (0.031)
XGB	0.863 (0.021)	0.481 (0.099)	0.704 (0.091)	0.951 (0.021)

HistGradient Boosting Classifier and Gradient Boosting Classifier slightly improved overall accuracy, while the other two estimators showed small decline compared to baseline results. XGB, GB and RF showed higher recall values and f-scores after hyper-parameter optimization step. Table 2 presents the average results for classification metrics for each model's performance after this stage. HGB had largest ROC AUC and PRC AUC, see Fig. 2.



(a) ROC area under curve (b) PRC area under curve

Fig. 2. Averaged ROC and PRC curves generated with four top performing models during training with 5 rounds of 10-fold stratified split and cross-validation.

Table 2. Model performance metrics after hyper-parameter tuning on training data. Averaged results with standard deviation shown after 5 rounds of 10-fold stratified split and cross-validation.

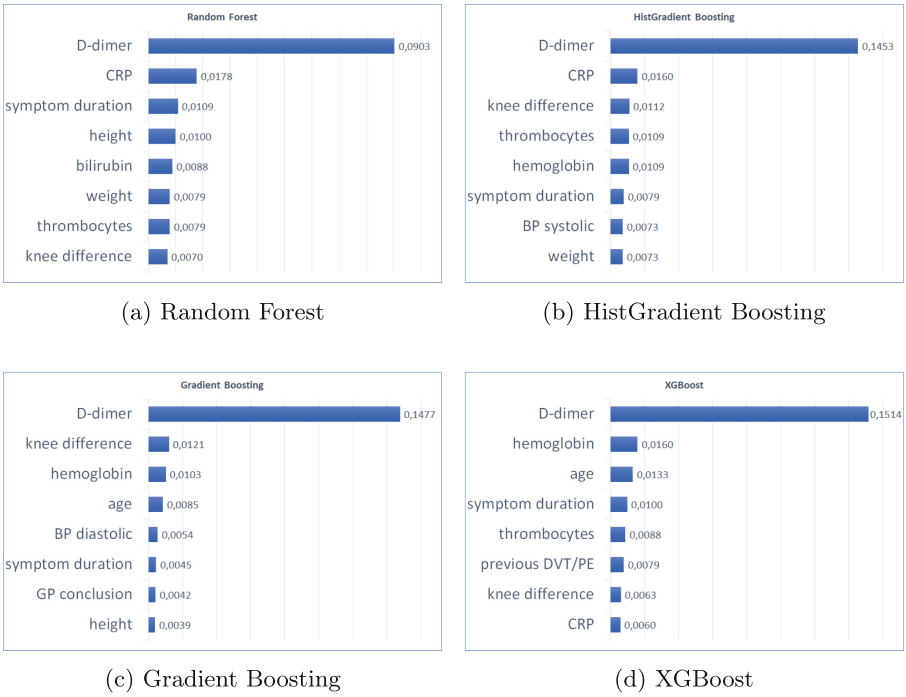
	RF	HGB	GB	XGB
Accuracy	0.862 (0.022)	0.867 (0.019)	0.865 (0.02)	0.859 (0.023)
Balanced accuracy	0.701 (0.051)	0.711 (0.039)	0.715 (0.042)	0.735 (0.045)
Recall	0.444 (0.103)	0.462 (0.08)	0.474 (0.087)	0.537 (0.088)
Precision	0.720 (0.095)	0.742 (0.088)	0.720 (0.085)	0.657 (0.081)
Specificity	0.959 (0.018)	0.961 (0.019)	0.955 (0.019)	0.934 (0.021)
f1-score	0.542 (0.094)	0.564 (0.071)	0.566 (0.074)	0.587 (0.075)
f2-score	0.478 (0.100)	0.497 (0.077)	0.506 (0.082)	0.555 (0.082)
ROC AUC	0.848 (0.038)	0.859 (0.034)	0.846 (0.038)	0.852 (0.038)
PRC AUC	0.647 (0.078)	0.653 (0.07)	0.639 (0.072)	0.651 (0.074)

4.3 Results of Permutation-Based Feature Importance Scoring

We further optimized the tuned models by investigating permutation-based feature importance scores. GB reached the accuracy of 0.91, using only 15 features for both training and predicting.

The other three models reached a classification accuracy of around 0.90. RF needed 31 predictors to reach the accuracy of 0.90 while HGB used only 12, reaching a marginally lower accuracy of 0.897. XGB showed the highest values for recall, f2-score and both ROC AUC and PRC AUC, while GB had heist precision and f1-score results. Table 3 shows the complete results for accuracy, balanced accuracy, recall, precision, specificity, f1- and f2-scores, ROC and PRC areas under curve.

Figure 3 shows the features ranged by their importance scores for each model. D-dimer value was the most significant feature for all models, while the other top 5 most important variables varied greatly between different ML models.



(a) Random Forest (b) HistGradient Boosting

(c) Gradient Boosting (d) XGBoost

Fig. 3. Permutation-based feature importance. The top 8 features with names and scores shown for each algorithm.

4.4 Results of Threshold Adjustment

We attempted to optimize classification decision thresholds for each model to find the value that gives us perfect recall. At this stage, Random Forest showed the highest specificity of 0.405 when reaching a maximum recall of 1.0 at the classification threshold level of 0.14. The ratio of predicted negatives to all negatives for the samples with a D-dimer value over 0.5 was approximately 23%. XGBoost and HGBC had specificities, ratios and thresholds of 0.04/16%/0.04 and 0.03/6%/0.03 accordingly. GradientBoosting model did not have a discrimination threshold giving 100% recall. Table 4 presents recorded threshold values, selected performance metrics, the number of correctly predicted negatives for samples with D-dimer 0.5 or higher, and their proportion in the total amount of cases requiring further diagnostic confirmation.

Table 3. Model performance metrics after permutation-based feature importance scoring and feature reduction on validation data.

	RF	HGB	GB	XGB
Number of features	31	12	15	18
Accuracy	0.900	0.897	0.909	0.903
Balanced accuracy	0.790	0.776	0.814	0.816
Recall	0.613	0.581	0.661	0.677
Precision	0.809	0.818	0.820	0.778
Specificity	0.967	0.970	0.967	0.955
f1-score	0.697	0.679	0.733	0.724
f2-score	0.644	0.616	0.688	0.695
ROC AUC	0.90	0.89	0.89	0.91
PRC AUC	0.77	0.73	0.77	0.80

Table 4. Results of threshold adjustment. The 3 algorithms that has reached perfect recall of 1.0 are RF, HGB and XGB. Gradient Boosting did not have a threshold which could give a perfect recall.

	RF	HGB	XGB
Threshold	0.14	0.03	0.04
Accuracy	0.517	0.411	0.468
Precision	0.279	0.241	0.261
Specificity	0.405	0.275	0.346
True negatives (D-dimer $\geq$ 0.5)	41 (23%)	11 (6%)	28 (16%)

5 Discussion

This study explored the predictive abilities of ML classification algorithms on the Ri-Schedule dataset of patients with suspected deep vein thrombosis.

All four models that performed best on our dataset rely on tree-based ensemble ML algorithms for binary classification. These results confirm that ensemble methods such as Random Forest and gradient boosting algorithms perform well on high-dimensional data with non-linear relationships, such as the Ri-Schedule dataset. Similar trends were reported in the studies of ML method classification performance on VTE related datasets [31,32,36].

Removing redundant features based on their insignificant or zero importance score improved all observed classification metrics for all four models. The results indicate that many features in the Ri-Schedule dataset either do not contribute to the identification of DVT or introduce noise that potentially reduces prediction accuracy. The sparse data content of the input variables and their irrelevance to blood clotting condition could be the two reasons for low or near-zero feature importance for DVT prediction.

Ranging features by their permutation-based importance allowed us to identify the D-dimer variable as having the highest score for all models across all experiments. This finding confirms its high predictive value in ruling out DVT [3,16]. It is interesting to see that apart from the D-dimer assay value, the ML models identified a limited number of features in the Ri-Schedule dataset that can also contribute to the DVT prediction accuracy. The frequently seen variables among the highest scored features were values for hemoglobin, c-reactive protein, thrombocytes, height, weight, symptom duration and knee measurements. To the best of our knowledge, the combination of these variables is not routinely used as biomarkers for the diagnosis of DVT or PE [10,23].

The standard approach to identifying DVT in patients is to start with prediction rule validation and screen with D-dimer assay [34]. This simple model identifies samples with a D-dimer value below 0.5 as negative and excludes them from further examinations. Those with D-dimer equal to or above 0.5 are subject to more resource- and time-consuming procedures. This model applied to the Ri-Schedule test dataset would yield an overall accuracy of only 0.468 with a recall of 1.0 and specificity of 0.346. We have observed that the RF model, tuned for maximum recall, had reached higher specificity of 0.405 and higher overall accuracy than the D-dimer model. However, the RF model had several false positives for samples with D-dimer values below 0.5.

The D-dimer based model's most important advantage is that it can safely exclude all cases with the D-dimer below the standard cut-off value. At the same time, the D-dimer model produces zero false negatives, which is crucial for correct diagnoses of a potentially lethal health condition. Combining the D-dimer model with the tuned RF model can ensure that no patient with DVT is falsely identified as healthy and reduce the number of patients sent to further examinations. First, the negative cases are ruled out based on the D-dimer cut-off value of 0.5. Then, the RF model with the new classification discrimination threshold is applied to exclude healthy patients from unnecessary examinations.

In the next iteration of this research, we should collect more data for further validation and model optimization. A new set of features could help identify other markers that could contribute to better class discrimination in the future study of DVT prediction on Ri-Schedule data. Another approach that could increase classification accuracy is using an age-adjusted D-dimer for ruling out VTE. This method can especially prove more efficient with older patients [13].

6 Conclusion

This work is the first comparative analysis of the performance of ML algorithms in identifying DVT among patients whose EHRs were collected by Østfold Hospital Trust, Norway.

The tree-based ensemble algorithms classified DVT from previously unseen data with a higher level of accuracy than other ML methods.

Random Forests achieved the highest specificity while reaching a perfect recall score.

The findings in this study confirm the importance of the D-dimer variable for diagnosing DVT. Furthermore, the results of our experiments suggest that the proposed combination of the standard D-dimer model and ML-based prediction can exclude a larger number of negative cases than by using the D-dimer based method alone. This proposed approach can have practical application in medicine and help save time and resources in real-life scenarios.

References

1. Bishop, C.M., et al.: Neural Networks for Pattern Recognition. Oxford University Press, Oxford (1995)
2. Bordes, A., Bottou, L., Gallinari, P.: SGD-QN: careful quasi-newton stochastic gradient descent. J. Mach. Learn. Res. **10**, 1737–1754 (2009)
3. Božič, M., Blinc, A., Stegnar, M.: D-dimer, other markers of haemostasis activation and soluble adhesion molecules in patients with different clinical probabilities of deep vein thrombosis. Thromb. Res. **108**(2), 107–114 (2002). https://doi.org/10.1016/S0049-3848(03)00007-0
4. Breiman, L.: Bagging predictors. Mach. Learn. **24**(2), 123–140 (1996). https://doi.org/10.1007/BF00058655
5. Breiman, L.: Rejoinder: arcing classifiers. Ann. Stat. **26**(3), 841–849 (1998). http://www.jstor.org/stable/120059
6. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324
7. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Routledge, New York (2017)
8. Chan, T., Golub, G., LeVeque, R.: Technical report STAN-CS-79-773, Department of Computer Science (1979)
9. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
10. Coleman, D.M., Wakefield, T.W.: Biomarkers for the diagnosis of deep vein thrombosis. Expert Opin. Med. Diagn. **6**(4), 253–257 (2012)
11. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995). https://doi.org/10.1007/BF00994018
12. Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., Singer, Y.: Online passive aggressive algorithms (2006)
13. Douma, R.A., et al.: Using an age-dependent D-dimer cut-off value increases the number of older patients in whom deep vein thrombosis can be safely excluded. Haematologica **97**(10), 1507 (2012)
14. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann. Eugen. **7**(2), 179–188 (1936)
15. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**(5), 1189–1232 (2001). https://doi.org/10.1214/aos/1013203451
16. Fronas, S.G., et al.: Safety of D-dimer testing as a stand-alone test for the exclusion of deep vein thrombosis as compared with other strategies. J. Thromb. Haemost. **16**(12), 2471–2481 (2018). https://doi.org/10.1111/jth.14314

17. Fronas, S.G., et al.: Safety and feasibility of rivaroxaban in deferred workup of patients with suspected deep vein thrombosis. Blood Adv. **4**(11), 2468–2476 (2020). https://doi.org/10.1182/bloodadvances.2020001556
18. Hoerl, A.E., Kennard, R.W.: Ridge regression: biased estimation for nonorthogonal problems. Technometrics **12**(1), 55–67 (1970). https://doi.org/10.1080/00401706.1970.10488634
19. Hosmer, D.W., Jr., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression, vol. 398. Wiley, Hoboken (2013)
20. Johnson, E.D., Schell, J.C., Rodgers, G.M.: The D-dimer assay. Am. J. Hematol. **94**(7), 833–839 (2019)
21. Ke, G., et al.: LightGBM: a highly efficient gradient boosting decision tree. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017). https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf
22. Le Gal, G., et al.: Prediction of pulmonary embolism in the emergency department: the revised Geneva score. Ann. Intern. Med. **144**(3), 165–171 (2006). https://doi.org/10.7326/0003-4819-144-3-200602070-00004
23. Lippi, G., Cervellin, G., Franchini, M., Favaloro, E.J.: Biochemical markers for the diagnosis of venous thromboembolism: the past, present and future. J. Thromb. Thrombolysis **30**(4), 459–471 (2010). https://doi.org/10.1007/s11239-010-0460-x
24. Luo, L., Kou, R., Feng, Y., Xiang, J., Zhu, W.: Cost-effective machine learning based clinical pre-test probability strategy for DVT diagnosis in neurological intensive care unit. Clin. Appl. Thromb. Hemost. **27** (2021). https://doi.org/10.1177/10760296211008650
25. Ma, H., et al.: A novel hierarchical machine learning model for hospital-acquired venous thromboembolism risk assessment among multiple-departments. J. Biomed. Inform. **122**, 103892 (2021). https://doi.org/10.1016/j.jbi.2021.103892
26. Nafee, T., et al.: Machine learning to predict venous thrombosis in acutely ill medical patients. Res. Pract. Thromb. Haemost. **4**(2), 230–237 (2020). https://doi.org/10.1002/rth2.12292
27. Peterson, L.E.: K-nearest neighbor. Scholarpedia **4**(2), 1883 (2009)
28. Ruder, S.: An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747 (2016)
29. Schapire, R.E.: Explaining AdaBoost. In: Schölkopf, B., Luo, Z., Vovk, V. (eds.) Empirical Inference, pp. 37–52. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41136-6_5
30. Tharwat, A.: Linear vs quadratic discriminant analysis classifier: a tutorial. Int. J. Appl. Pattern Recogn. **3**(2), 145–180 (2016)
31. Wang, K.Y., et al.: Using predictive modeling and supervised machine learning to identify patients at risk for venous thromboembolism following posterior lumbar fusion. Glob. Spine J. (2021). https://doi.org/10.1177/21925682211019361
32. Wang, X., Yang, Y.Q., Liu, S.H., Hong, X.Y., Sun, X.F., Shi, J.H.: Comparing different venous thromboembolism risk assessment machine learning models in Chinese patients. J. Eval. Clin. Pract. **26**(1), 26–34 (2020). https://doi.org/10.1111/jep.13324
33. Wells, P.S., et al.: Value of assessment of pretest probability of deep-vein thrombosis in clinical management. The Lancet **350**(9094), 1795–1798 (1997). https://doi.org/10.1016/S0140-6736(97)08140-3
34. Wilbur, J., Shian, B.: Diagnosis of deep venous thrombosis and pulmonary embolism. Am. Fam. Physician **86**(10), 913–919 (2012)

35. Williams, C.K., Rasmussen, C.E.: Gaussian Processes for Machine Learning, vol. 2. MIT Press, Cambridge (2006)
36. Xue, B., et al.: Use of machine learning to develop and evaluate models using preoperative and intraoperative data to identify risks of postoperative complications. JAMA Netw. Open **4**(3), e212240 (2021). https://doi.org/10.1001/jamanetworkopen.2021.2240