Title

Predicting venous thromboembolism (VTE) risk in cancer patients using machine learning

Permalink

https://escholarship.org/uc/item/74f229fz

Journal

Health Care Science, 2(4)

ISSN

2771-1757

Authors

Townsley, Samir Khan

Basu, Debraj

Vora, Jayneel

et al.

Publication Date

2023-08-01

DOI

10.1002/hcs2.55

Title: Predicting Venous Thromboembolism (VTE) Risk in Cancer Patients Using Machine Learning

Author details:

1. Samir Khan Townsley *

Department of Electrical and Computer Engineering, University of California, Davis

2. Debraj Basu *

Department of Electrical and Computer Engineering, University of California, Davis

3. Jayneel Vora

Department of Computer Science, University of California, Davis

4. Ted Wun

School of Medicine, University of California, Davis Health

5. Chen-Nee Chuah

Department of Electrical and Computer Engineering, University of California, Davis

6. Corresponding author: Dr. Prabhu RV Shankar, M.D., M.S., MRCP (UK)

School of Medicine, University of California, Davis Health

Email: rvpshankar@ucdavis.edu

Mailing address: 4610 X St, Sacramento, CA 95817, USA

Phone number : 01 408-458-6119

Fax number :   01 916-734-7055

* The research work was conducted during his association with University of California, Davis

1. Summary

Objectives

The association between cancer and venous thromboembolism (VTE) is well-established with cancer patients accounting for approximately twenty percent of all VTE incidents. In this paper, we have:

- Performed a comparison of machine learning (ML) methods to traditional clinical scoring models for predicting the occurrence of VTE in a cancer patient population, and
- Identified important features (clinical biomarkers) for ML model predictions and examined how different approaches to reducing the number of features used in the model impact model performance.

Methods

We have developed an ML pipeline including three separate feature selection processes and applied it to routine patient care data from the electronic health records (EHR) of 1910 cancer patients at the University of California Davis Medical Center (UCDMC).

Results

Our ML-based prediction model achieved an area under the receiver operating characteristic (ROC) curve of 0.778 ± 0.006 when trained on a set of 15 features. This result is comparable to the model performance when trained on all features in our feature pool (0.779 ± 0.006 with 29 features). Our result surpasses the most validated clinical scoring system for VTE risk assessment in cancer patients by 16.1%. We additionally found cancer stage information to be a useful predictor after all performed feature selection processes despite not being used in existing score-based approaches.

Conclusion

From these findings, we observe that machine learning can offer new insights and a significant improvement over the most validated clinical VTE risk scoring systems in cancer patients. The results of this study also allowed us to draw insight into our feature pool and identify the features that could have the most utility in the context of developing an efficient machine learning classifier. While a model trained on our entire feature pool of 29 features significantly outperformed the traditionally used clinical scoring system, we were able to achieve an equivalent performance using a subset of only 15 features through strategic feature selection methods. These results are encouraging for potential applications of ML to predicting cancer associated VTE in clinical settings such as in bedside decision support systems where feature availability may be limited.

Keywords

VTE; Cancer; Binary classification; Machine learning pipeline

2. Introduction

Venous thromboembolism (VTE) comprises both deep-vein thrombosis (DVT) and pulmonary embolism (PE) [1]. The association between VTE and cancer is well-established with cancer patients accounting for approximately twenty percent of all VTE incidents [2]. While the estimated prevalence of VTE in the general population is around 1 in 1000 [3, 4], some estimates suggest this number increases 5-fold within the cancer patient population [1, 5, 6]. The risk increases further among patients who receive chemotherapy as shown in a 15-year population-based study [7].

VTE is a multifaceted risk in cancer patients that exacerbates clinical consequences, significantly impacting morbidity, mortality, and cost of patient care [1, 5, 8, 9, 10, 11]. Specifically, VTE associated mortality is 2.2 times more likely in VTE patients with cancer than in those without [10]. VTE is the leading cause of mortality in cancer patients, aside from mortality due to cancer itself [1, 8]. In addition to increasing risk of mortality, VTE burdens the cancer treatment process. When managing VTE in cancer patients, use of anticoagulants, which thin the blood, requires rigorous patient monitoring in order to achieve adequate anticoagulation and to identify complications such as bleeding. Compared to cancer patients without VTE, patients with VTE have over two times the risk of experiencing major bleeding [12]. Bleeding can worsen anemia while reduced blood counts can delay cancer interventions such as chemotherapy and radiotherapy and increase the need for blood transfusions.

The recurrence rates of VTE are also high in patients with cancer. Patients with an active malignancy have a 3-4 fold higher risk of recurrence compared to patients without cancer, and the risk is further increased in those with metastatic cancers. According to one study, the one-year cumulative risk for recurrent VTEs after the first episode was 21% in cancer patients compared to 7% in patients without cancer [12]. All the VTE related factors discussed above can affect cancer management, increase treatment costs, and escalate average price per hospitalization for cancer patients [2, 3, 12, 13].

Treatments such as anticoagulant therapy are available, both for prophylaxis against occurrence, as well as for treatment of VTE in cancer patients. Appropriate and timely use of the prophylactic measures are vital for reducing the risk of both fatal and non-fatal pulmonary embolism as well as the post-thrombotic complications [14]. Anticoagulants are drugs that interfere with blood coagulation cascade to reduce or inhibit blood clotting. The low-molecular-weight heparin (LMWH) has been found in multiple studies to reduce the likelihood of a VTE event occurring in a cancer patient [2, 15, 16, 17]. With these issues in mind, it is evident that effective VTE prophylaxis in cancer patients has the potential to drastically improve cancer survival rates and decrease treatment costs for hospitals and patients alike. However, while anticoagulant prophylaxis and treatment is effective in primary and secondary prevention of VTE, as mentioned above, there are certain implications with their regular use in all cancer patients. In particular, anticoagulants are associated with increased bleeding, require parenteral administration, training, and additional monitoring, all of which can increase both cost and complexity of cancer patient management [2, 12 ,18]. Therefore, it is important to stratify and define high risk cohorts of cancer patients who are prone for VTE. There is thus a need for effective VTE risk stratification systems to ensure that prophylaxis is administered only to high-risk patients. An accurate, reliable, and robust VTE stratification system would help clinicians in decision making about anticoagulant therapy at the point-of-care (POC). Prophylactic measures against VTE are often implemented for hospitalized patients, so high risk stratification is particularly important in ambulatory patients (outpatients) as they cannot be monitored as closely as hospitalized patients.

The importance of delineating which cancer patients are at increased risk of VTE for instituting anticoagulation prophylaxis, particularly ambulatory patients, is critical as anticoagulation is associated with significant risks and costs in already debilitated cancer patients. Decision to provide prophylactic anticoagulation in ambulatory patients clinically alone is often difficult and providers need a decision support tool that pinpoints the most vulnerable groups for VTE. Several Cancer Associated Thromboembolism (CAT) prediction scores have been developed, such as Khorana [19], Vienna CATS [20,] PROTECHT [21] and CONKO [22] based on routinely collected patient care data. These risk-assessment methods all use a simple scoring system where points are added based on each of five to eight different predictors with higher scores indicating a higher risk of developing VTE. Some of the predictors that these scores use include cancer site, platelet count, white blood cell count, hemoglobin, use of red blood cell stimulating factors, and BMI. Of these scores, the Khorana score is the most validated and used [23]. However, despite its acceptance in the research community, the Khorana score still only achieves a positive predictive value of 6.7 %, which is not meaningful enough to make a quantified decision by the clinicians and thus leaves plenty of room for improvement [19]. In another study of 218 patients with cancer initiating chemotherapy, it is shown that the Khorana score was able to stratify ambulatory cancer patients according to the risk of VTE, but not for all cancer types [24]. The Khorana score can be used to select ambulatory cancer patients at high risk of venous thromboembolism for thromboprophylaxis, but most events occur outside this high-risk group [25].

During informal discussions, clinicians opined that, even a positive predictive value of 20 to 30% will help them with decision making, tipping the decision one way or other with some scientific qualitative basis, and those discussions motivated the team to explore various features (clinical biomarkers) and develop more robust and clinically meaningful predictive models.

In this study we use machine learning to take a data-driven approach to VTE prediction in cancer patients. Our aim in this study is to not only improve upon the performance of known risk assessment scores such as the Khorana score but also to perform an in depth, data-driven exploration of both new and known VTE risk factors.

Traditional approaches to prediction in medicine often focus on capturing medical expertise through a set of carefully designated rules [26]. However, data driven approaches, such as machine learning algorithms instead can learn effective prediction decisions by observing numerical patterns in the input data [26, 27]. One subset of machine learning, known as supervised learning, deals with training a model to accomplish this task of classifying data based on a set of input data with labeled ground truth values [27]. Supervised learning has the advantage over traditional rule-based methods of being able to leverage computational power to identify highly convoluted patterns in massive datasets with large numbers of potential predictors relatively quickly and efficiently [26, 28]. Such an approach has promise in the context of cancer patient VTE prediction, where the currently accepted scoring systems are simple rule-based methods that do not necessarily capture a wide range of the potentially complex interactions between variables [19, 20, 22]. Patrizia Ferroni et al. have designed a precision medicine approach to exploit significant patterns in data to produce VTE risk predictors for cancer outpatients [29]. They have used Multiple kernel learning (MKL) [30] based on support vector machines (SVM) models to predict VTE risk. In our research, we have examined VTE classification performances of several standard ML algorithms including SVM, logistic regression (LR), and Random Forest (RF) and compared these to the baseline performance of the Khorana score.

Methods and results are described in the following sections.

3. Methods

In this retrospective study of a population of cancer patients at the University of California Davis Medical Center, we used machine learning to explore both new and known VTE risk factors. Our goal was to not only develop a machine-learning based VTE risk assessment system for cancer patients but also to examine which risk factors may be useful when taking such an approach. From our efforts, we hope to establish a foundation for using machine learning to eventually answer more complex questions about VTE prediction in cancer patients, such as how changes in a patient's condition, as the patient continues with his/her cancer management, affect the risk of developing VTE over time.

In this study, we examined 29 features in total, including a selection of available features from the Khorana score and biomolecular markers from a previous study of CAT [19, 29]. Since relevant VTE events can occur before or after cancer diagnosis and clinical interventions (i.e., surgery, chemotherapy, radiotherapy), we used a set of time-agnostic features to gain a view of how a patient's general profile over a large period of time may or may not be indicative of VTE risk. Each of the features we used covered information about a patient's background, cancer, lab values, or medications.

We then explored the utility of our feature set in a machine learning context in a two-phased approach. In the first phase we trained several different models with a spectrum of hyperparameter choices on four different feature subsets that were derived both from performed feature selection experiments and from pre-determined feature pools. We then identified the best performing model and feature set combination and, in a second phase of experiments, attempted to reduce the number of used features without sacrificing performance through an iterative feature accumulation process. Finally, we validated the performance of our chosen model on a held-out dataset extracted from our original data.

3.1. Dataset and Data Prepossessing

The dataset used in these experiments was extracted from the UCDMC affiliated hospital's EHR system and combined with curtained and manually curated data elements from the California state cancer network CNExT registry, from 2015-2017 (C/NET Solutions, Berkeley, CA). The organ system-based cancers which are considered high risk for VTE episodes in previous studies were included in the study and are listed in Table 1 [32].

Table 1. Cancer sites contained in the dataset

Site group
Pancreas
Bladder
Non-Hodgkin's Lymphoma
Hodgkin's Disease
Corpus Uteri/Uterus
Prostate
Ovary
Breast
Lung/Bronchus (Small Cell and Non-Small Cell)
Brain

Stomach

In order to study how a given cancer and its attributes may be predictive of VTE events, each cancer instance was treated as a separate entry in our dataset. Thus, a few patients have more than one cancer entry in the dataset. Associated with each cancer instance is a list of features describing the cancer and patient background.

All medications were grouped according to the pharmacologic class of the medication. Medication data was incorporated in the primary cancer entry cohort by assigning a binary variable to each patient for every medication, indicating whether or not that medication was ever administered to the patient.

Lab test values were represented by the mean of all pre-chemotherapy measurements associated with that test in order to eliminate noise and understand how a patient's general condition correlates with VTE risk. We accumulated such values for 45 different lab tests. This set of 45 was then reduced to only the lab tests which were performed on at least 75 % of patients. Of the 45 lab tests, only 12 of the tests satisfied this criterion and were included in our final feature pool. Any missing values among these 12 lab tests were imputed using the mean across all patients for the given test.

Exclusion criteria for our dataset included patients with missing information in any of the listed categories outside of lab tests, patients with benign tumors, patients with mesotheliomas, and patients with extreme outliers (i.e., BMI > 100). These exclusion criteria were applied to the general dataset. After cleaning, the dataset consisted of 1973 cancer entries across 1910 unique patients.

The presence or absence of a VTE diagnosis date served as our binary target variable for prediction in our machine learning models. The full list of features in our curated dataset is detailed in Table 2.

Table 2: Feature pool

Feature Type	Features (29)
Cancer	site, grade, stage, behavior, histopathological type
Patient	gender, body mass index (BMI), age, race list, race count
Binary Medications	antineoplastic - aromatase inhibitors, immunosuppressives, antineoplastic - antiandrogenic agents, steroid antineoplastics, antineoplastic - alkylating agents, antineoplastic systemic enzyme inhibitors, antineoplastic - antimetabolites
Lab Tests	albumin, hematocrit, hemoglobin, creatinine serum, red blood cell count, calcium, white blood cell count, platelet count, mean corpuscular hemoglobin concentration (MCHC), mean corpuscular hemoglobin (MCH), protein, mean corpuscular volume (MCV)

3.2. Model Training

We performed an 80:20 split on the dataset, allocating 80% of the data for cross-validation of different model and feature set combinations. We used the remaining 20% as a hold-out dataset for testing the generalizability of our best performing model. Our approach to performing model training and feature selection was two-fold:

1. First, we trained seven different model configurations, each on four different feature sets. The model configurations and feature set choices are described in the remainder of this subsection and in subsections 3.3.1 and 3.3.2.

2. Second, we took the highest performing model configuration and used a stepwise feature selection approach to attempt to find a reduced subset of features that would provide comparable performance. The implementation of this feature selection approach is described in subsection 3.3.3.

To prevent overfitting, all models were trained and validated on our training dataset using 10-fold cross-validation. We evaluated our trained models using the area under the ROC curve (AUROC) and the DeLong test for statistical significance [33]. We also evaluated the AUROC generated by the Khorana score on our dataset and used this for baseline performance comparisons with our models.

For the first phase of our study, we trained and evaluated models using the machine learning algorithms and parameter configurations listed in Table 3.

Table 3: Machine Learning Model Configurations

Model	Parameter Choices
Logistic Regression (LR) [34]	-
Support Vector Machine (SVM) [35]	Radial basis function kernel, linear kernel
Random Forest (RF) [36]	50 trees, 100 trees, 200 trees, 500 trees

All LR, SVM, and RF models were implemented using the Scikit-learn library in Python [37]. Each of these models was cross-validated on four different feature sets/subsets:

1. All 29 available features in our feature pool.

2. Features used for calculating the Khorana score: cancer site, platelet count, hemoglobin level, white blood cell count, and BMI.
3. Features selected by our clinical team. We will refer to this feature selection method as the "clinical expert" method.
4. Features selected based on statistical correlation with VTE incidence. We will refer to this feature selection method as the "filtering" method.

For the second phase of the experiment, we identified the model with the highest performance based on AUROC values and DeLong test results for statistical significance. We then used this model to perform a stepwise forward feature selection method to identify a minimum subset of features required to attain equivalent performance. We will refer to this feature selection method as the "wrapper" method. The implementations of this and the clinical expert and filtering methods are described in detail in the following section.

Finally, we tested our best performing model on the held-out dataset to better examine the generalizability of the model and ensure that we did not overfit the training dataset.

3.3. Feature Selection Methods

In training different machine learning models for predicting VTE, we experimented with three different feature selection methods. The first was an expert-driven feature selection process in which we used domain expertise from clinicians and researchers at UCDMC to derive a subset of known clinically relevant features as a feature set for training our machine learning models. The second was a filtering approach which identified the highest statistically correlated features with our target. The third was a wrapper approach that bootstrapped the model training process to iteratively accumulate an optimal set of features for a chosen ML classifier [38].

The clinical expert and filtering approaches were used in the first phase of our study for comparing performances of different machine learning approaches across several feature sets. The goals of performing these feature selection approaches were to:

- Examine the utility of commonly accepted VTE risk factors in a machine learning approach.
- Identify new risk factors or combinations of risk factors which may add value to predicting VTE incidence in cancer patients using machine learning.

The wrapper approach was used in the second phase of our study on the best performing model and feature set from the first phase. The goal of this approach was primarily to:

- Minimize the number of features required for the best performing model configuration to achieve optimal performance.

The implementation details for these feature selection methods are described in the following subsections.

3.3.1 Clinical Expert Method

Our first feature selection method involved consulting with our team of physicians to determine a subset of features that are known risk factors in the development of VTE. The decisions made in this process were based both on clinical expertise and review of literature in the area [19, 29, 20, 21, 22, 32].

3.3.2 Filtering Method

Since our data consists of both categorical and continuous data, we divided our feature filtering approach into two tasks. For the categorical features, we determined the likelihood of each feature being linearly independent of our target variable using a chi-squared test [39].

Meanwhile, for each continuous feature in our dataset, we observed the distribution of the feature across VTE-diagnosed patients as well as the distribution of the feature across patients without a VTE diagnosis. We then compared these distributions to determine the likelihood that they came from one common distribution using a Kolmogorov-Smirnov (KS) test for goodness of fit [40].

We acquired our final statistically filtered feature set by selecting only the features from both of the above tests which resulted in $p < 0.05$.

3.3.3 Wrapper Method

The final feature selection process we used was an empirical forward feature selection method that served the purpose of maximizing the performance of our model while minimizing the dimensionality. While a high-dimensional model is appealing from a performance standpoint, it may not always be practical in a clinical setting due to limitations in available lab test results or other information. Performing a forward feature selection process allows us to directly identify only the *n* best performing features on our dataset and thus reduce the amount of required information without significant sacrifices in performance.

While the filtering method that is discussed in the last subsection is valuable for identifying variables directly correlated with the target, it fails to examine how different combinations of these variables may affect the predictive power of our chosen ML classifier [33]. In order to cover the full space of variable interactions, we would ideally train a model on every possible combination of features from our feature pool, but doing so would take several years of model training and would be computationally infeasible. We used the wrapper method to shortcut this process and only test a small subset of all possible unique feature combinations.

In our approach, we accumulated features one at a time under the assumption that the best performing feature at each iteration is part of the optimal set [41]. This process started by training 29 separate models: one trained on each feature in our set. Each training cycle included 10 iterations of 10-fold cross-validation. The best performing feature was then selected and the process repeated with the remaining 28 features, this time also including the best selected feature(s) from the previous iteration(s) and so on. We continued to accumulate features in this fashion until we no longer saw improvements in performance for a predetermined number of iterations. To provide a small buffer for temporary drops in performance, we set this number to 2 iterations.

It should be noted that, while the clinical expert and filtering feature selection methods are determined independently of any model choices, the wrapper selected features are specific to one model as they are accumulated by iterative model training. Since we used this method in the second phase of our study to optimize the feature set for a selected model, we found it sufficient to only perform the wrapper feature selection process for our best performing model.

4. Results

4.1. Model Selection

The first phase of our study involved training several model configurations on different selected feature sets. Each model was evaluated by generating an AUROC value and confidence interval from 10 iterations of 10-fold cross-validation. The results of this model training and feature selecting are presented in this section and in section 4.2. Table 4 shows the performance of each model configuration on the training dataset (80% of the original dataset) across the four different feature sets listed in section 3.2. Each row represents a unique model algorithm or scoring system and each column represents a unique feature set. To make a fair comparison between different models that are using different feature sets, we have included a model trained on the features that the Khorana score uses as shown in column 3 **Khorana (n=5)** of Table 4. The performance generated by using the standard Khorana scoring system itself is also included as a baseline in the last row of Table 4. All model ROC curves were compared to that of the baseline Khorana score in the last row of Table 4 via the DeLong test. The differences that were statistically significant based on a p-value < 0.05 are marked with an asterisk in the table. The full list of model-to-model DeLong comparisons is also provided in Appendix B.

Table 4: AUROC of predictive models by feature set

	All (n=29)	**Khorana (n=5)**	**Clinical (n=5)**	**Filtered (n=20)**
Logistic Regression	0.684 ± 0.054*	0.668 ± 0.077	0.662 ± 0.074	0.672 ± 0.047*
SVM (RBF Kernel)	0.652 ± 0.061	0.562 ± 0.061*	0.576 ± 0.056*	0.617 ± 0.072
SVM (Linear Kernel)	0.644 ± 0.042	0.577 ± 0.040*	0.589 ± 0.048*	0.669 ± 0.036*
Random Forest (50 trees)	0.751 ± 0.068*	0.672 ± 0.062*	0.681 ± 0.072*	0.748 ± 0.071*
Random Forest (100 trees)	0.752 ± 0.062*	0.676 ± 0.066*	0.683 ± 0.072*	0.743 ± 0.073*
Random Forest (200 trees)	0.762 ± 0.065*	0.684 ± 0.070*	0.692 ± 0.074*	0.746 ± 0.075*
Random Forest (500 trees)	0.761 ± 0.065*	0.684 ± 0.073*	0.696 ± 0.071*	0.755 ± 0.067*
Baseline: Khorana Score	-	0.632 ± 0.019	-	-

* p<0.05 from DeLong test when compared to Khorana score (bottom row)

In general, every model outperformed the Khorana score baseline when trained on our entire feature space (though this difference for the SVM models was not statistically significant). The RF models trained on the same features used in the Khorana score all achieved a small but significant improvement over the Khorana score, suggesting that using ML alone instead of a simple point system may offer an improvement over currently used clinical risk assessment scores. However, the results of the models trained on the other feature sets indicate that this is not the maximum attainable performance and that adding additional risk factors to the model could result in even larger performance improvements.

Every RF model also outperformed the logistic regression and SVM models on each feature set suggesting that a random forest is likely the best suited algorithm choice for this task among our tested classifiers. For the ease of viewing, the p-values of all pair-wise model comparisons by feature set are not listed here but can be viewed in Appendix B.

The RF models also showed similar trends across feature sets with performance being highest when trained on all features followed by the filtered feature set, clinical expert feature set, and then the Khorana score feature set. The highest performing models were the four RF models trained on all features and on the filtered feature set. Since the difference between these models was generally not statistically significant, we chose the most complex model – the RF model with 500 trees – as our best performing model for the second phase of the study. The reasoning for this choice was that a more complex model, while more prone to overfitting, is also capable of learning more complex variable relationships leading to potential performance improvements. As mentioned in the methodology, we combat and assess overfitting by performing 10-fold cross validation on all experiments and further validating our best performing model on a held-out dataset.

Based on these results, we will focus on the performance of the 500-tree RF model for the remainder of our analysis where we will explore optimizing the set of required features using the wrapper feature selection method and will validate our model performance on our held-out dataset.

But first, details on the results of the clinical expert and filtering feature selection processes are provided in the following section.

4.2. Feature Selection Results

4.2.1 Clinically Important Features

Our first feature selection method involved reducing our feature set to a list of only five features deemed clinically important to the prediction of VTE by a team of UCDMC physicians and researchers. These features are:

platelet count, white blood cell count, hemoglobin, cancer site, cancer stage

The first four of these are the same four features that are common across the Khorana, Vienna CATS, PROTECHT, and CONKO scoring systems while cancer stage is an additional feature deemed relevant by our team [19, 20, 21, 22]. The RF model with 500 trees trained on these features outperforms the AUROC of the Khorana score on our dataset by 10%. This improvement can be attributed to the fact that the RF model is capable of making decisions that are much more nuanced than the decisions made in any of the listed scoring systems, which involve only simple point additions based on binary categorizations of the data [39]. Despite this improvement in performance, the model still falls short of the model trained on the full feature set by 8.5%, indicating that there are other potentially useful features in predicting VTE that were not initially deemed clinically relevant.

4.2.2 Filtered Features

In order to further examine the known clinically relevant features and identify new features, we used statistical methods to filter our feature pool and identify features highly correlated with our target variable. The feature filtering method described previously yielded a set of 20 features that were significantly correlated with the binary presence of VTE. The full list of this filtered feature set includes the following features:

site, grade, stage, histopathological type, gender, age, race list, antineoplastic - aromatase inhibitors, albumin, hematocrit, hemoglobin, creatinine serum, red blood cell count, calcium, white blood cell count, platelet count, MCHC, MCH, protein, MCV

Notably, all of the clinically essential features identified above were also found to be significantly correlated with our target. All of the features used in the Khorana score were also selected with the exception of BMI. All of the lab tests in our feature pool were selected as well while all but one pharmacologic class, i.e., antineoplastic aromatase inhibitors, were left out. The RF model with 500 trees achieved a 19.5% improvement over the Khorana score and did not result in a significant decline in performance based on the DeLong test compared to the model trained on all features.

4.3. Model Optimization

For the second phase of our study, we looked at optimizing the feature set for our best performing model configuration and validating the performance on our held-out test set. Based on the results presented in Table 4, we used the 500-tree RF model trained on our entire feature pool as a baseline for our best performing model. In this section, we present the results of using this model with the previously described wrapper feature selection method to reduce the dimensionality of the feature set while attempting to maintain the same level of model performance.

4.3.1. Wrapper Selected Features

Table 5 compares the cross-validation performance of the 500-tree RF model using the wrapper selected feature set to the results from the first phase of the study. When compared to the model trained on all features, the wrapper and filtered feature sets are the only feature sets that did not result in a statistically significant decline in performance. This confirms that the wrapper method was effective in identifying a reduced subset of features (52% of the whole feature pool and 75% of the filtered feature pool), without sacrificing performance.

Table 5. Cross-Validation of 500-Tree Random Forest on All Feature Sets

	All (n=29)	**Khorana (n=5)**	**Clinical (n=5)**	**Filtered (n=20)**	**Wrapper (n=15)**
Random Forest (500 trees)	0.761 ± 0.065	0.684 ± 0.073*	0.696 ± 0.071*	0.755 ± 0.067	0.769 ± 0.072

* p<0.05 from DeLong test when compared to model trained on all features (first column)

Table 6 contains the ordered list of features accumulated when performing the wrapper feature selection method with the RF model of 500 trees. The curve illustrated in Figure 1 shows the relationship between these features and the AUROC of our model during feature accumulation. Each model evaluation came from the average result of 10 iterations of 10-fold cross-validation. The x-axis represents each iteration of the recursive accumulation of features, while the y-axis represents the AUROC associated with the model trained after each added feature. The model trained on this set of recursively selected features not only matched the performance of the model trained on all features with no statistical difference between ROC outputs, but also did so with only 15 features, reducing the size of our feature set by 14. The ROC and PRC curves resulting from training a model on these 15 features are contained in Figure 3 and Figure 4 respectively.

Table 6. Order of Accumulated Features During Wrapper Selection

1	creatinine serum

2	antineoplastic - aromatase inhibitors
3	MCHC
4	red blood cell count
5	stage
6	Immunosuppressives
7	antineoplastic - antiandrogenic agents
8	protein
9	site
10	MCV
11	antineoplastic - alkylating agents
12	albumin
13	antineoplastic – antimetabolites
14	MCH
15	histopathological type

Figure 1. Mean AUROC of 500-tree RF Model During Wrapper Feature Accumulation

Unlike in the clinical expert and filter-selected feature sets, seven different medications were included in the wrapper-selected feature set, although only two appeared in the first twelve selected features. Furthermore, the white blood cell count and platelet count lab tests were excluded despite being included in both of our other examined feature sets as well as the Khorana score. This exclusion is not to undermine the usefulness of the features to the task of VTE prediction, but rather to show that they were not necessary for achieving optimal performance with reduced dimensionality on our dataset.

4.3.2. Feature Set Comparisons

Table 7 lists the overlap between the feature sets of the three presented feature selection methods. The full list of features selected by each method is provided in Appendix A.

All features deemed clinically relevant were also found to be statistically correlated with the presence of VTE in our filtered feature set. Furthermore, all three feature selection methods selected the cancer site and stage as important features for VTE prediction. While cancer site is a widely used risk factor for VTE, cancer stage is not typically included in currently used scoring systems [19, 20, 21, 22]. The clinical team further concurred with the data driven finding of the importance of clinical staging information.

The overlap of the clinical expert and wrapper feature sets matches the overlap of the clinical expert, filter, and wrapper feature sets and is thus omitted from the table.

Table 7. Overlapping features between feature sets

Feature Selection Methods	**Features**
*Expert + Filter + Wrapper	site, stage

Filter + Wrapper	site, stage, antineoplastic - aromatase inhibitors, albumin, creatinine serum, red blood cell count, mean corpuscular hemoglobin concentration (MCHC), mean corpuscular hemoglobin (MCH), protein, mean corpuscular volume (MCV), histopathological type
Filter + Expert	site, stage, hemoglobin, platelet count, white blood cell count

*The overlap of only the expert and wrapper feature sets produces the same list of features

4.4 Performance Validation on Held-Out Data

The remainder of the results section shows the performance when validating our RF model trained with 500 trees on our held-out data (20% of the original dataset).

4.2.1 All features

Figure 2. Performance comparison on held out test set between Khorana score and RF model with all features

The ROC curve in Figure 2 illustrates the test performance of the RF model with 500 trees being trained on our entire feature pool in comparison to the ROC curve generated from the Khorana score on our held-out test dataset. The model achieves a statistically significant improvement in AUROC of 16.1**%** compared to the Khorana score. This increase in performance confirms the potential for improving VTE prediction through the inclusion of new risk factors in a machine learning approach. Next, we validated the 500-tree RF model with each of the previously examined feature subsets.

Figure 3. ROC Performance by Feature Set on Held-Out Data

The ROC curves in Figure 3 show this performance by feature set when run on our held-out data. As in the results in section 4.3.1., the model trained on the wrapper selected features did not result in a statistically significant decline in performance compared to the model trained on the entire feature pool. This validates our takeaway that the wrapper feature selection process provided an effective way to reduce the feature space without impacting performance. A full list of DeLong test comparisons for the 500-tree RF models on the held-out dataset are provided in Appendix B.

For additional validation, we evaluated the precision-recall curve (PRC) for the 500-tree RF model on each feature set. These results are displayed in Figure 4.

Figure 4. PRC Performance by Feature Set on Held-Out Data

Similar to the ROC results, the PRC curves in Figure 4 show that the models trained on all features and on the wrapper-selected features are the best performing models and achieve comparable performance.

5. Discussion

In this study, we examined the utility of using machine learning to predict VTE in cancer patients. We accomplished this through a carefully designed set of steps adhering to a typical machine learning pipeline. First, we selected a feature pool based on the data availability within our patient population. We also set aside 20% of the data in a held-out dataset for final model validation. We then performed a number of feature selection methods and trained multiple machine learning classifiers with different hyperparameter configurations to identify a best performing model for our use case. Finally, we iteratively trained the best performing model in order to accumulate a minimum set of required features and thus reduce the complexity of the model without impacting model performance.

The results of this process allow us to draw insight into how a machine learning classifier might offer an improvement in performance over traditionally used clinical VTE risk assessment systems in cancer patients. With these results, we are able to examine our feature pool and identify those features that are most useful in the context of developing an efficient machine learning classifier by comparing the selected features and resulting model performance across multiple unique feature selection methods.

This project was an effort to showcase the improved predictive performance of various ML models over the Khorana score in predicting VTE in cancer patients. We compared the performance of models trained on different feature sets selected by domain experts, statistical methods, and ML techniques. We identified features that were common across these selected feature sets to better understand which features are meaningful in this context.

Our trained classifiers achieved encouraging results on numerous feature subsets. We found that a 500-tree RF model trained using only the features used in the Khorana score achieved a statistically significant 14.6% improvement in AUROC over the standard point-based Khorana score on our held-out test set with an AUROC of 0.769 ± 0.007. Meanwhile, we achieved a peak AUROC of 0.779 ± 0.006 on a held-out dataset when training the 500-tree RF model on our entire feature pool. This surpassed the performance of the Khorana score on the same dataset by 16.1%. We were additionally able to reduce the number of required features to 15 total (a 48% reduction) without a statistically significant impact on model performance by using a wrapper method to iteratively accumulate features. We also used two model-agnostic feature selection methods – a statistical filtering method and a clinical expert method – which both achieved AUROCs of 0.771 ± 0.007 and 0.757 ± 0.004 respectively on our held-out dataset. All of these results showed statistically significant improvements in performance over that of the Khorana score.

The results in Table 7 depict the overlap between the features selected by our three described feature selection methods. Only cancer site and cancer stage were common across all three feature sets. Cancer site is already a common risk factor considered in current VTE risk stratification systems [19, 20, 21, 22]. Based on our experimental results, cancer stage merits inclusion in future VTE prediction systems using an ML approach. Meanwhile, all of the features deemed clinically relevant were also found to be statistically significant in the filtered feature set. Unlike the other two feature sets, the wrapper-selected feature set did not include hemoglobin. However, it did identify three related metrics - corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), and mean corpuscular volume (MCV) - as essential metrics for VTE prediction. While these metrics are not identical to hemoglobin, they are likely inter-related. Furthermore, since the wrapper method optimizes the feature space based on empirical performance of different feature combinations, an excluded feature is not by necessity unimportant. Instead, an excluded feature may be redundant when compared to the optimal set of features, making its inclusion unnecessary for improving prediction performance.

In comparison to the features used in the Khorana score, all but BMI are included in the filtered and clinically relevant feature sets. Furthermore, the cancer site, which is the most heavily weighted risk factor in the Khorana score, was selected in all three feature sets. Interestingly, BMI, which is included in the Khorana score, Vienna CATS, and PROTECHT, was not identified as useful in any of our acquired feature sets [19, 20, 21]. Aside from BMI, however, the results of this study suggest that the predictors used in the Khorana score have a relatively high predictive power when used in a machine learning context. The results also suggest that the stage of the cancer is useful in predicting VTE and should be considered in future machine learning applications. Because staging information is not always readily available in medical notes, future studies could look to reliably extract this information from free medical text using NLP methods. Since cancer staging can vary over time as new information comes in and is incorporated in the staging determination, this problem is particularly challenging with past efforts achieving only limited success [42, 43]. One approach that may improve this performance without sacrificing too much predictive power in VTE risk assessment could involve reducing the cancer stage to a binary variable that simply indicates a presence or absence of metastasis [44].

While the results of this study are promising, it is important to note that the dataset uses a small sample size, especially for certain subgroups, (i.e., only a few pharmacological groups were used in the patient population). Also, the study did not include cancer patients who had radiation therapy. There is increasing evidence implicating radiotherapy in cancer associated thrombosis (CAT) in cancer patients, however accessing data from the radiation therapy information system (RTIS) was not possible for this study. This study dealt with the patient population at only one location, so before we generalize these results across the general population, the findings in this study should be validated in other patient populations. Furthermore, this study takes a time-agnostic approach to identify useful predictors for VTE in cancer patients. Therefore, this approach highlights VTE predictors that may be useful in a machine learning context but does not yet reflect an implementable clinical scenario. With this being the case, the aim of this study was to effectively identify these useful predictors in order to provide the groundwork for exploration of this problem in specific clinical scenarios (i.e., at different stages of pre-diagnosis presentation, establishing diagnosis, and post-diagnosis treatment phases of a patient's cancer management).

The methods used in this study could be generalizable to other clinical conditions, particularly ambulatory settings, where there is moderate to strong increased risk for developing VTE, such as, congestive heart or respiratory failure, hormone replacement and oral contraceptive therapy, antiphospholipid antibody and other thrombophilia syndromes [45]. Even though multiple studies have demonstrated that thromboprophylaxis using anticoagulant treatments such as low-molecular-weight heparin (LMWH) can reduce the likelihood of VTE events, due to the need for training the patients and care-givers to administer (parenteral) the LMWH, regular lab monitoring and dose adjustment, as well as the potential for bleeding complications, all of which add to the cost and quality of care, such prophylaxis may not always be feasible and risk-free. There is thus a need for effective VTE risk stratification and decision support systems to ensure that prophylaxis is administered only to high-risk patients.

The project goal was to select the necessary and sufficient features from our available feature pool that would maximize the predictive power of various statistical ML models. It can be a hard decision to initiate prophylaxis against VTE, especially in ambulatory cancer patients where anti-thrombosis prophylaxis can be expensive and cumbersome. Evidence based decision support is crucial for minimizing risk in this decision process and improving patient outcomes.

At the point of care where the decisions are made, ideally, prediction tools and scoring systems should automatically retrieve the required features and inform the clinicians to help make decisions. For ease of use and interpretability, the list of features should be small, but should provide meaningful enough information to supplement the current evidence and clinicians' evaluations. We found cancer staging information to be particularly meaningful as a predictor of VTE as it was selected in all of our feature selection processes. The Khorana score does not include the cancer staging information as often it can be hard to retrieve accurate staging information from clinical notes. Accurate staging information is often established by cancer registrars retrospectively, which may take up to six months. Our study emphasizes the importance of cancer staging information as a predictor of VTE in cancer patients and highlights the need for its timely evaluation. Simplifying the cancer stage variable into a binary value indicating whether the cancer is metastatic (stage 4) or non-metastatic could improve the accessibility and real-time accuracy of staging but would require further studies and additional validation.

6. Conclusion

Machine learning offers a promising avenue for improving the performance of current VTE prediction scores in cancer patients. A combination of a time-agnostic approach and three unique feature selection methods demonstrates that at least four of the features that are used to calculate the Khorana score can also provide high predictive power to a machine learning classifier. We also observe that cancer stage information is generally more useful than BMI as a predictor in our ML classifiers. Consultation with clinicians reveal a potential reason - BMI can vary as patients lose significant weight due to cancer itself, chemotherapy, and associated anorexia or other adverse effects. Furthermore, with significant improvements in the generated ROC curve, it is clear that a machine learning classifier can make complex deductions that may allow it to outperform currently used VTE risk scores. The results in this study offer a foundation from which future machine learning approaches to VTE prediction in cancer patients can be built. Future studies should consider the identified relevant variables in the context of a temporal analysis in which machine learning may be used to dynamically assess at all levels how cancer management progress, including medical intervention, over time can alter a patient's risk of developing VTE.

Declaration

Ethics approval and consent to participate

Appropriate institutional review board (IRB) review and approval was obtained from the UCDMC IRB, bearing number: UCDMC.

Conflict of interests

The authors declare no conflict of interests.

References

1. Khorana AA. The NCCN Clinical Practice Guidelines on Venous Thromboembolic Disease: strategies for improving VTE prophylaxis in hospitalized cancer patients. Oncologist. 2007;12(11):1361-1370. doi:10.1634/theoncologist.12-11-1361
2. Lyman GH, Khorana AA, Falanga A, et al. American Society of Clinical Oncology guideline: recommendations for venous thromboembolism prophylaxis and treatment in patients with cancer. J Clin Oncol. 2007;25(34):5490-5505. doi:10.1200/JCO.2007.14.1283
3. Silverstein MD, Heit JA, Mohr DN, Petterson TM, O'Fallon WM, Melton LJ 3rd. Trends in the incidence of deep vein thrombosis and pulmonary embolism: a 25-year population-based study. Arch Intern Med. 1998;158(6):585-593. doi:10.1001/archinte.158.6.585
4. Spencer FA, Emery C, Lessard D, et al. The Worcester Venous Thromboembolism study: a population-based study of the clinical epidemiology of venous thromboembolism. J Gen Intern Med. 2006;21(7):722-727. doi:10.1111/j.1525-1497.2006.00458.x
5. Kessler CM. The link between cancer and venous thromboembolism: a review. Am J Clin Oncol. 2009;32(4 Suppl):S3-S7. doi:10.1097/COC.0b013e3181b01b17
6. Lee AY, Levine MN. Venous thromboembolism and cancer: risks and outcomes. Circulation. 2003;107(23 Suppl 1):I17-I21. doi:10.1161/01.CIR.0000078466.72504.AC

7. Heit JA, Silverstein MD, Mohr DN, Petterson TM, O'Fallon WM, Melton LJ 3rd. Risk factors for deep vein thrombosis and pulmonary embolism: a population-based case-control study. Arch Intern Med. 2000;160(6):809-815. doi:10.1001/archinte.160.6.809
8. Khorana AA, Francis CW, Culakova E, Kuderer NM, Lyman GH. Thromboembolism is a leading cause of death in cancer patients receiving outpatient chemotherapy. J Thromb Haemost. 2007;5(3):632-634. doi:10.1111/j.1538-7836.2007.02374.x
9. Khorana AA, Francis CW, Culakova E, Kuderer NM, Lyman GH. Frequency, risk factors, and trends for venous thromboembolism among hospitalized cancer patients. Cancer. 2007;110(10):2339-2346. doi:10.1002/cncr.23062
10. Sørensen HT, Mellemkjaer L, Olsen JH, Baron JA. Prognosis of cancers associated with venous thromboembolism. N Engl J Med. 2000;343(25):1846-1850. doi:10.1056/NEJM200012213432504
11. Elting LS, Escalante CP, Cooksley C, et al. Outcomes and Cost of Deep Venous Thrombosis Among Patients With Cancer. Arch Intern Med. 2004;164(15):1653–1661. doi:10.1001/archinte.164.15.1653
12. Prandoni P, Lensing AW, Piccioli A, et al. Recurrent venous thromboembolism and bleeding complications during anticoagulant treatment in patients with cancer and venous thrombosis. Blood. 2002;100(10):3484-3488. doi:10.1182/blood-2002-01-0108
13. Mandalà M, Falanga A, Roila F; ESMO Guidelines Working Group. Management of venous thromboembolism (VTE) in cancer patients: ESMO Clinical Practice Guidelines. Ann Oncol. 2011;22 Suppl 6:vi85-vi92. doi:10.1093/annonc/mdr392
14. Cayley WE Jr. Preventing deep vein thrombosis in hospital inpatients. BMJ. 2007;335(7611):147-151. doi:10.1136/bmj.39247.542477.AE
15. Samama MM, Cohen AT, Darmon JY, et al. A comparison of enoxaparin with placebo for the prevention of venous thromboembolism in acutely ill medical patients. Prophylaxis in Medical Patients with Enoxaparin Study Group. N Engl J Med. 1999;341(11):793-800. doi:10.1056/NEJM199909093411103
16. Leizorovicz A, Cohen AT, Turpie AG, et al. Randomized, placebo-controlled trial of dalteparin for the prevention of venous thromboembolism in acutely ill medical patients. Circulation. 2004;110(7):874-879. doi:10.1161/01.CIR.0000138928.83266.24
17. Cohen AT, Davidson BL, Gallus AS, et al. Efficacy and safety of fondaparinux for the prevention of venous thromboembolism in older acute medical patients: randomised placebo controlled trial. BMJ. 2006;332(7537):325-329. doi:10.1136/bmj.38733.466748.7C
18. Kuderer NM, Khorana AA, Lyman GH, Francis CW. A meta-analysis and systematic review of the efficacy and safety of anticoagulants as cancer treatment: impact on survival and bleeding complications. Cancer. 2007;110(5):1149-1161. doi:10.1002/cncr.22892
19. Khorana AA, Kuderer NM, Culakova E, Lyman GH, Francis CW. Development and validation of a predictive model for chemotherapy-associated thrombosis. Blood. 2008;111(10):4902-4907. doi:10.1182/blood-2007-10-116327
20. Ay C, Dunkler D, Marosi C, et al. Prediction of venous thromboembolism in cancer patients. Blood. 2010;116(24):5377-5382. doi:10.1182/blood-2010-02-270116

21. Verso M, Agnelli G, Barni S, Gasparini G, LaBianca R. A modified Khorana risk assessment score for venous thromboembolism in cancer patients receiving chemotherapy: the Protecht score. Intern Emerg Med. 2012;7(3):291-292. doi:10.1007/s11739-012-0784-y
22. Pelzer U, Sinn M, Stieler J, Riess H. Primary pharmacological prevention of thromboembolic events in ambulatory patients with advanced pancreatic cancer treated with chemotherapy?. Dtsch Med Wochenschr. 2013;138(41):2084-2088. doi:10.1055/s-0033-1349608
23. van Es N, Di Nisio M, Cesarman G, et al. Comparison of risk prediction scores for venous thromboembolism in cancer patients: a prospective cohort study. Haematologica. 2017;102(9):1494-1501. doi:10.3324/haematol.2017.169060
24. Overvad TF, Ording AG, Nielsen PB, Skjøth F, Albertsen IE, Noble S, Vistisen AK, Gade IL, Severinsen MT, Piazza G, Larsen TB. Validation of the Khorana score for predicting venous thromboembolism in 40 218 patients with cancer initiating chemotherapy. Blood Adv. 2022 May 24;6(10):2967-2976. doi: 10.1182/bloodadvances.2021006484. PMID: 35045569; PMCID: PMC9131922.
25. Mulder FI, Candeloro M, Kamphuisen PW, Di Nisio M, Bossuyt PM, Guman N, Smit K, Büller HR, van Es N; CAT-prediction collaborators. The Khorana score for prediction of venous thromboembolism in cancer patients: a systematic review and meta-analysis. Haematologica. 2019 Jun;104(6):1277-1287. doi: 10.3324/haematol.2018.209114. Epub 2019 Jan 3. PMID: 30606788; PMCID: PMC6545838.
26. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. N Engl J Med. 2016;375(13):1216-1219. doi:10.1056/NEJMp1606181
27. Deo RC. Machine Learning in Medicine. Circulation. 2015;132(20):1920-1930. doi:10.1161/CIRCULATIONAHA.115.001593
28. Kotsiantis, S., Zaharakis, I, Pintelas, P. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 2007, 160, 3–24.
29. Ferroni P, Zanzotto FM, Scarpato N, et al. Risk Assessment for Venous Thromboembolism in Chemotherapy-Treated Ambulatory Cancer Patients. Med Decis Making. 2017;37(2):234-242. doi:10.1177/0272989X16662654
30. Mehmet Gönen, Ethem Alpaydin. Multiple Kernel Learning Algorithms. Journal of Machine Learning Research 12 (2011) 2211-2268.
31. Hanna DL, White RH, Wun T. Biomolecular markers of cancer-associated thromboembolism. Crit Rev Oncol Hematol. 2013;88(1):19-29. doi:10.1016/j.critrevonc.2013.02.008
32. Wun T, White RH. Epidemiology of cancer-related venous thromboembolism. Best Pract Res Clin Haematol. 2009;22(1):9-23. doi:10.1016/j.beha.2008.12.001
33. 31. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988 Sep 1:837-45.

34. Bagley SC, White H, Golomb BA. Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. J Clin Epidemiol. 2001;54(10):979-985. doi:10.1016/s0895-4356(01)00372-9
35. Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge: Cambridge University Press; 2000. doi:10.1017/CBO9780511801389
36. Breiman, L. Random forests, Machine learning. 2001. 45(1), 5–32.
37. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V and others. Scikit-learn: Machine learning in Python. Journal of machine learning research. (2011). 12(Oct), 2825–2830.
38. Jiliang Tang, Salem Alelyani, Huan Liu. Feature selection for classification: A review. Data classification: Algorithms and applications, (2014). 37.
39. McHugh ML. The chi-square test of independence. Biochem Med (Zagreb). 2013; 23(2):143-149. doi:10.11613/bm.2013.018
40. Massey Jr, F. J. The Kolmogorov-Smirnov test for goodness of fit. Journal of the American Statistical Association (1951). 46(253), 68-78. DOI: 10.2307/2280095
41. Guyon, Isabelle, and André Elisseeff. An Introduction to Variable and Feature Selection, Journal of Machine Learning Research. (2003) 3, 1157–1182.
42. Warner JL, Levy MA, Neuss MN, Warner JL, Levy MA, Neuss MN. ReCAP: Feasibility and Accuracy of Extracting Cancer Stage Information From Narrative Electronic Health Record Data. J Oncol Pract. 2016;12(2):. doi:10.1200/JOP.2015.004622
43. AAlAbdulsalam AK, Garvin JH, Redd A, Carter ME, Sweeny C, Meystre SM. Automated Extraction and Classification of Cancer Stage Mentions from Unstructured Text Fields in a Central Cancer Registry. AMIA Jt Summits Transl Sci Proc. 2018;2017:16-25. Published 2018 May 18.
44. Soysal E, Warner JL, Denny JC, Xu H. Identifying Metastases-related Information from Pathology Reports of Lung Cancer Patients. AMIA Jt Summits Transl Sci Proc. 2017;2017:268-277. Published 2017 Jul 26.
45. Anderson Jr, F. A., & Spencer, F. A. (2003). Risk factors for venous thromboembolism. Circulation, 107(23_suppl_1), I-9.

Appendix A

Table A.1. Full list of selected features by feature selection method

Feature Selection Method	Features
Clinical Expert Method	site, stage, hemoglobin, platelet count, white blood cell count
Filter Method	site, grade, stage, histopathological type, gender, age, race list, antineoplastic - aromatase inhibitors, albumin, hematocrit,

	hemoglobin, creatinine serum, red blood cell count, calcium, white blood cell count, platelet count, MCHC, MCH, protein, MCV
Wrapper Method	site, stage, histopathological type, albumin, creatinine serum, red blood cell count, MCHC, MCH, protein, MCV, antineoplastic - aromatase inhibitors, immunosuppressives, antineoplastic - antiandrogenic agents, antineoplastic - alkylating agents, antineoplastic - antimetabolites

Appendix B

The following tables show the comprehensive results of performing the DeLong test for statistical significance between ROC curves of the various models we trained during the study. Each table is a grid of DeLong p-values. For this study, we used p<0.05 as our cutoff for statistical significance. The first four tables are most pertinent to the results discussed in the main text while the following tables contain a more comprehensive coverage of pairwise prediction comparisons.

Table B.1. DeLong p-values for Models Compared to Khorana Score

	All (n=29)	**Khorana (n=5)**	**Clinical (n=5)**	**Filtered (n=20)**
Logistic Regression	0.00142	0.07314	0.101754	0.004921
SVM (RBF Kernel)	0.150591	0.00036	0.001697	0.27491
SVM (Linear Kernel)	0.18518	3.2E-05	0.004174	0.000772
Random Forest (50 trees)	0.0	0.023375	0.017531	0.0
Random Forest (100 trees)	0.0	0.020919	0.015383	2E-06
Random Forest (200 trees)	0.0	0.011794	0.006736	2E-06
Random Forest (500 trees)	0.0	0.014679	0.003016	0.0

Table B.2. DeLong p-values for Models Compared to Same Model Trained on All Features

	Khorana (n=5)	**Clinical (n=5)**	**Filtered (n=20)**
Logistic Regression	0.307395	0.234885	0.300637

SVM (RBF Kernel)	0.00089	0.003027	0.130158
SVM (Linear Kernel)	0.000331	0.005092	0.08326
Random Forest (50 trees)	0.00465	0.016925	0.466185
Random Forest (100 trees)	0.005323	0.014444	0.387342
Random Forest (200 trees)	0.006923	0.016309	0.321548
Random Forest (500 trees)	0.009481	0.020839	0.431354

Table B.3. DeLong p-values for 500-tree RF Models on Held-Out Test Dataset

	All (n=29)	**Khorana (n=5)**	**Clinical (n=5)**	**Filtered (n=20)**	**Wrapper (n=15)**
All (n=29)	0.5	0.000465	0.0	0.00303	0.369048
Khorana (n=5)	0.000465	0.5	1.0E-06	0.301592	0.001222
Clinical (n=5)	0.0	1.0E-06	0.5	0.0	0.0
Filtered (n=20)	0.00303	0.301592	0.0	0.5	0.006966
Wrapper (n=15)	0.369048	0.001222	0.0	0.006966	0.5

Table B.4. DeLong p-values for 500-tree RF Models vs. Khorana Score on Held-Out Test Dataset

	All (n=29)	**Khorana (n=5)**	**Clinical (n=5)**	**Filtered (n=20)**	**Wrapper (n=15)**
Baseline: Khorana Score	0.0	0.0	0.0	0.0	0.0

Below are the results of performing the DeLong test for statistical significance between ROC curves on every pairwise combination of models for each feature set we examined in the study.

Table B.5. DeLong p-values for Models Trained on All Features

	Logistic Regression	**SVM (RBF Kernel)**	**SVM (Linear Kernel)**	**Random Forest (50 trees)**	**Random Forest (100 trees)**	**Random Forest (200 trees)**	**Random Forest (500 trees)**	**Baseline: Khorana Score**
Logistic Regression	0.5	0.116269	0.037197	0.010025	0.006254	0.002805	0.003274	0.001447
SVM (RBF Kernel)	0.116269	0.5	0.367859	0.00054	0.000257	0.000104	0.000127	0.150591
SVM (Linear Kernel)	0.037197	0.367859	0.5	2.7E-05	7E-06	2E-06	3E-06	0.18518

Random Forest (50 trees)	0.010025	0.00054	2.7E-05	0.5	0.48744	0.367113	0.379221	0.0
Random Forest (100 trees)	0.006254	0.000257	7E-06	0.48744	0.5	0.372979	0.385744	0.0
Random Forest (200 trees)	0.002805	0.000104	2E-06	0.367113	0.372979	0.5	0.487627	0.0
Random Forest (500 trees)	0.003274	0.000127	3E-06	0.379221	0.385744	0.487627	0.5	0.0
Baseline: Khorana Score	0.001447	0.150591	0.18518	0.0	0.0	0.0	0.0	0.5

Table B.6. DeLong p-values for Models Trained on Khorana Score Features

	Logistic Regression	**SVM (RBF Kernel)**	**SVM (Linear Kernel)**	**Random Forest (50 trees)**	**Random Forest (100 trees)**	**Random Forest (200 trees)**	**Random Forest (500 trees)**	**Baseline: Khorana Score**
Logistic Regression	0.5	0.000618	0.000882	0.459024	0.416674	0.330223	0.329241	0.073683
SVM (RBF Kernel)	0.000618	0.5	0.266912	7.4E-05	7.4E-05	4.2E-05	6.2E-05	0.00036
SVM (Linear Kernel)	0.000882	0.266912	0.5	5.9E-05	6.6E-05	3.8E-05	6.3E-05	3.2E-05
Random Forest (50 trees)	0.459024	7.4E-05	5.9E-05	0.5	0.450544	0.350922	0.349703	0.023375
Random Forest (100 trees)	0.416674	7.4E-05	6.6E-05	0.450544	0.5	0.399317	0.396751	0.020919
Random Forest (200 trees)	0.330223	4.2E-05	3.8E-05	0.350922	0.399317	0.5	0.494963	0.011794
Random Forest (500 trees)	0.329241	6.2E-05	6.3E-05	0.349703	0.396751	0.494963	0.5	0.014679
Baseline: Khorana Score	0.073683	0.00036	3.2E-05	0.023375	0.020919	0.011794	0.014679	0.5

Table B.7. DeLong p-values for Models Trained on Clinical Expert Features

	Logistic Regression	**SVM (RBF Kernel)**	**SVM (Linear Kernel)**	**Random Forest (50 trees)**	**Random Forest (100 trees)**	**Random Forest (200 trees)**	**Random Forest (500 trees)**	**Baseline: Khorana Score**

Logistic Regression	0.5	0.002724	0.006279	0.288285	0.272988	0.197527	0.163826	0.102482
SVM (RBF Kernel)	0.002724	0.5	0.302347	0.000265	0.00023	9.2E-05	3.5E-05	0.001697
SVM (Linear Kernel)	0.006279	0.302347	0.5	0.000648	0.000563	0.000226	8.7E-05	0.004174
Random Forest (50 trees)	0.288285	0.000265	0.000648	0.5	0.480818	0.380343	0.336385	0.017531
Random Forest (100 trees)	0.272988	0.00023	0.000563	0.480818	0.5	0.398935	0.354845	0.015383
Random Forest (200 trees)	0.197527	9.2E-05	0.000226	0.380343	0.398935	0.5	0.456638	0.006736
Random Forest (500 trees)	0.163826	3.5E-05	8.7E-05	0.336385	0.354845	0.456638	0.5	0.003016
Baseline: Khorana Score	0.102482	0.001697	0.004174	0.017531	0.015383	0.006736	0.003016	0.5

Table B.8. DeLong p-values for Models Trained on Filter Features

	Logistic Regression	**SVM (RBF Kernel)**	**SVM (Linear Kernel)**	**Random Forest (50 trees)**	**Random Forest (100 trees)**	**Random Forest (200 trees)**	**Random Forest (500 trees)**	**Baseline: Khorana Score**
Logistic Regression	0.5	0.027277	0.451289	0.003532	0.007221	0.00583	0.001173	0.005015
SVM (RBF Kernel)	0.027277	0.5	0.024625	4.4E-05	0.00011	8.8E-05	1.2E-05	0.27491
SVM (Linear Kernel)	0.451289	0.024625	0.5	0.001453	0.003458	0.002766	0.000373	0.000772
Random Forest (50 trees)	0.003532	4.4E-05	0.001453	0.5	0.436936	0.477591	0.414938	0.0
Random Forest (100 trees)	0.007221	0.00011	0.003458	0.436936	0.5	0.460575	0.354352	2E-06
Random Forest (200 trees)	0.00583	8.8E-05	0.002766	0.477591	0.460575	0.5	0.395179	2E-06
Random Forest (500 trees)	0.001173	1.2E-05	0.000373	0.414938	0.354352	0.395179	0.5	0.0
Baseline: Khorana Score	0.005015	0.27491	0.000772	0.0	2E-06	2E-06	0.0	0.5