Predicting First-Episode Venous Thromboembolism Risk Using a Supervised Regression Random Forest Model

Samantha Flottman

Abstract

Venous thromboembolism (VTE), the formation of blood clots in deep veins, kills over 100,000 people in the United States. Many of these deaths occur due to the fact that VTE is not diagnosed until the patient is in critical condition. While AI models have been explored for predicting the risk of recurrent venous thromboembolism (rVTE), there is a paucity of research using these models to predict the risk of first-episode venous thromboembolism. This study compared several AI approaches to identify the superior method for VTE risk stratification. It was determined that a supervised regression random forest machine learning model would be the optimal choice for this task, given its numerous complex factors. While this model can be potentially used in clinical settings, further research must be done in order to determine its accuracy and applicability.

Introduction

In an analysis by Min-Jeoung Kang et al. spanning 3525 patients, the delayed diagnosis of VTE (venous thromboembolism) caused an increased mortality rate from 17 deaths (2.52%) to 217 deaths (8.33%) at Mass General Brigham (MGB), and an increase mortality rate from 2 deaths (4.65%) to 12 deaths (5.97%) at Penn State Health (PSH), exemplifying the importance of diagnosing within a short time frame (Min Jeoung, 2025). VTE's symptoms include shortness of breath, pain while walking, and swollen legs (DVT). However, symptoms of VTE are not uniform and are often symptomless until serious complications occur ("Symptoms"). Thus, difficulty in diagnosing VTE until a critical condition, together with delayed diagnosis in hospitals, poses a significant risk. It is reported to be associated with a significant risk of recurrence and substantial mortality, with reported death rates of up to 40% at 10 years" (Winter, 2017).

In 2020, Martin et al. created three distinctive neural network models with the aim of predicting the risk of recurrent VTE. The models were constructed using a database with 39 clinical factors from 245 patients. Model one utilized all 39 factors as inputs; model two utilized 18 factors as inputs; and model three utilized 15 inputs. Because each model used different training mechanisms, this resulted in variations of predictive outputs and different accuracy for each model. After assessment of accuracies, the most effective model was developed, leading researchers to conclude that it could be used as a valuable tool by physicians (Martins, 2020).

Other examples of the use of AI in healthcare have been in detecting Cancer. Clinical Histopathology Imaging Evaluation Foundation (CHIEF), an AI model created with the purpose of predicting cancer by evaluating imaging, was trained using 15 million unlabeled data into sections of interest. It was further trained on 60,000 images on various tissues. Using databases collected from 24 different hospitals, it was determined that CHIEF had an accuracy reading of 94%, outperforming other AI models (Pesheva, 2024).

In determining the appropriate model to predict the first episode of VTE, a supervised regression machine learning model with labeled relevant clinical inputs and outputs will be employed. The database used must list all factors associated with VTE and indicate whether the patient had VTE. Additionally, a wide demographic should be employed in order to avoid bias in the model. A regression model will be used instead of a classification model in order to best assess the level of risk a patient is at for VTE.

Data acquisition will pose a challenge for this regression model as it will need to include specific factors. In order to evaluate the accuracy of this model, the R-squared value will be calculated. Furthermore, the benefits of this model will be stated clearly along with the challenges this model imposes.

Perspective Results

In order to take the first step toward building the model, a comprehensive, labeled data set with captures all the factors that contribute to the risk of Venous Thromboembolism (VTE) is needed. The search for the acquisition of this data will be geared toward academic and clinical data repositories. Certain datasets that can be used include, but are not limited to, PhysioNet, UCI Machine Learning Repository, and Clinicaltrials.gov. Other publicly available datasets from major clinical trials focused on cardiovascular health may be used. Search terms will be concise, including, but not limited to, “VTE reoccurrence risk,” “VTE patient data,” and “deep vein thrombosis (DVT) cohort.” All datasets will undergo screening to ensure that they adhere to the factors mentioned in Table 1. Diverse demographics should be included in order to avoid bias in the model.

To ensure compatibility with a supervised regression machine learning model means, data must undergo major engineering prior to use. Factors such as surgery must be categorized as shown in Table 1. For example, a patient who underwent a complex spinal deformity correction would be categorized as a major transient surgical risk factor. Rigorous screening on all final data will be conducted before training the model.

In order to predict the probability of VTE, a supervised model will be used. The input of this model will include the data in Table 1. All the data for the input will be labeled.

Table 1

Major Transient Risk Factors	
Surgical Factors	**Nonsurgical Factors**
Orthopedic, general, urologic, or gynecologic surgery (duration > 45 minutes)	Immobilization
Trauma	Bed riddance due to acute disease
	Critical Illness
Minor Transient Risk Factors	
Surgical Factors	**Nonsurgical Factors**
Orthopedic, general, urologic, or gynecologic surgery (duration ≤ 45 minutes)	Pregnancy/Postpartum
Limb Trauma with minor surgery	Acute Infections
	Estrogen Use
	Limb Trauma with or without plaster cast
Chronic Risk Factors	
Major Risk Factors	**Minor Risk Factors**
Cancer	Inflammatory Bowel Disease
Neurologic disease with paresis	Autoimmune disease
Predisposing Conditions	
Increasing Age	
Obesity	
Heart Failure	
Diabetes	
Heart Disease	

Note: *The data in Table 1 is from research by Becattini et al.*

A regression model with labeled inputs will be utilized, with the output ranging from 0-100% risk. Unlike a classification model, which outputs a discrete value (e.g., *risk* or *no risk*), a regression model outputs a continuous value, thereby providing the patient with an accurate assessment of their risk for VTE. For example, a regression model might output patient one to have a VTE risk of 10% and another patient with a VTE risk of 90% while a classification model would output both patients as being at risk for VTE. While both are accurate assessments, the regression model provides a specific risk value, allowing for clinical settings to make nuanced decisions.

A Random Forest Regression model will be used to determine the risk for VTE. The features listed in Table 1 will serve as the input for the ensemble of the decision trees. Each individual decision tree will output a continuous risk value between clamped function 0 and 1. The values of the individual decision trees will be averaged unweighted. Additionally, the factor importance, which determines which factor has the most weight, will be extracted in order to optimize the data.

In order to assess the accuracy of the model, the R-squared coefficient will be determined for the model. An R-squared coefficient of 1 indicates perfect predictive power of the model, while an R-squared score of 0 indicates no predictive power of the model. The R-squared coefficient is determined by the following formula:

$$R^2=1 - SS_{tot}/SS_{res}$$

An accurate risk evaluation can be determined through this supervised regression machine learning model. It could be used for clinical use and evaluation. Additionally, the inputs of the model allow patients to realize what the factors are that contribute to VTE. This model can accurately evaluate risk while providing pertinent information to patients.

Additionally, this model will help spot VTE by alerting patients and care providers before a critical condition ensues. According to the CDC, around 60,000 to 100,000 American patients die due to Venous thromboembolism every year. Having an accurate model that evaluates the risk can help mitigate this number (Data, 2025).

Previous integration of AI in healthcare suggests that it can provide accurate predictions in certain fields. For example, CHIEF, a cancer detection AI model, outperformed other AI models and performed with great accuracy. Specifically, the model performed at 96 percent accuracy with multiple different cancer types. With unseen data, it performed at an accuracy rate of 90% (Pesheva, 2024).

Similarly, the advancement of ECG accuracy can be partially attributed to the integration of AI. Specifically, it was found that AI-based ECG significantly improved the detection of STEMI heart attack and reduced the number of false positives (Herman, 2025).

To the same effect, a supervised regression machine learning model can accurately calculate the risk of VTE or rVTE of a patient.

Discussion

The challenge posed is the sheer amount of data and the revision of the data necessary for the model. Extensive screening of the data must be completed before using any type of data. Additionally, clinicians often miss factors, causing inaccurate outputs.

Furthermore, the artificial intelligence is limited to algorithmic bias. Although this can be mitigated by employing a wide range of demographics and age ranges, an artificial intelligence model is at risk of overfitting and bias.

Future studies could possibly specify this model further by splitting Venous Thromboembolism into its different types of VTE. Additionally, it could be further broken down into the male and female categories, further specifying the details and noting how certain factors determine the VTE based on sex.

Conclusion

Tens of thousands of deaths occur due to venous thromboembolism. Late detection of blood clots contributes to this number of deaths. With the help of artificial intelligence, venous thromboembolism can be mitigated by predicting the risk before a critical condition.

A supervised regression model will be employed with inputs being of the following categories: major transient, minor transient, major chronic, and minor chronic. To evaluate the model's accuracy, the R-squared value will be calculated. This model will potentially help mitigate the number of deaths caused by VTE.

However, an artificial intelligence model will always have its limitations, which include both data acquisition limitations and bias. Future study that delves into the specific types of VTE and the sex differences is encouraged, as they will reveal additional insight into how different factors affect different scenarios of VTE.

References

Becattini, Cecilia, and Cimini Anna Ludovica. “Provoked vs Minimally Provoked vs Unprovoked VTE: Does It Matter? | Hematology, Ash Education Program | American Society of Hematology.” *Hematology, ASH Education Program*, 8 Dec. 2023, ashpublications.org/hematology/article-abstract/2023/1/600/506484/Provoked-vs-minimally-provoked-vs-unprovoked-VTE.

“Data and Statistics on Venous Thromboembolism.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 27 Jan. 2025, www.cdc.gov/blood-clots/data-research/facts-stats/index.html.

Herman, Robert, et al. “AI-Enabled ECG Analysis Improves Diagnostic Accuracy and Reduces False STEMI Activations: A Multicenter U.S. Registry.” *JACC Journals*, 28 Oct. 2025, www.jacc.org/journal/jacc.

Martins, T D, et al. “Artificial Neural Networks for Prediction of Recurrent Venous Thromboembolism.” *International Journal of Medical Informatics*, U.S. National Library of Medicine, 18 June 2020, pubmed.ncbi.nlm.nih.gov/32593848/.

Min-Jeoung, Kang. “Delayed Venous Thromboembolism Diagnosis and Mortality Risk | Hematology | Jama Network Open | Jama Network.” *Delayed Venous Thromboembolism Diagnosis and Mortality Risk*, 26 Sept. 2025, jamanetwork.com/journals/jamanetworkopen/fullarticle/2839378.

Pesheva, Ekatrina. “A New Artificial Intelligence Tool for Cancer.” *Home*, 4 Sept. 2024, hms.harvard.edu/news/new-artificial-intelligence-tool-cancer.

“Symptoms.” *National Heart, Lung, and Blood Institute*, U.S. Department of Health and Human Services, www.nhlbi.nih.gov/health/venous-thromboembolism/symptoms. Accessed 3 Nov. 2025.

Winter, M-P, et al. “Chronic Complications of Venous Thromboembolism.” *Journal of Thrombosis and Haemostasis : JTH*, U.S. National Library of Medicine, 15 Aug. 2017, pubmed.ncbi.nlm.nih.gov/28762624/.