Predicting Pulmonary Embolism Risk in Deep Vein Thrombosis Patients Using Supervised Machine Learning Models

Alex McCorriston[1,†]

[1] LinkedIn

https://www.linkedin.com/in/alex-mccorriston-42b437268/

† Address to which correspondence should be addressed:

alexmc808@gmail.com

Abstract

Pulmonary embolism (PE) is a serious and potentially fatal complication of deep vein thrombosis (DVT), making it vital to understand how both a DVT diagnosis and its initial treatment influence PE risk. While anticoagulation (AC) is the treatment standard, alternative treatments like systemic thrombolytics, mechanical thrombectomy (MT), and catheter-directed thrombolysis (CDT) exist, yet their impact on PE risk remains unclear. Supervised machine learning can identify patterns and risk factors that may not be evident through traditional statistical methods, providing a more comprehensive assessment of PE risk by integrating demographic, clinical, and treatment variables. This study experimented with Random Forest (RF), XGBoost (XGB), and Logistic Regression (LR) models to predict PE risk in 4,410 DVT patients using the Medical Information Mart for Intensive Care (MIMIC)-IV database. The models identified history of PE, number of DVT-related admissions, and treatment timing (one to three days post-diagnosis) as key predictors of a future PE event, while treatment type was not a significant predictor, suggesting timely intervention may be more critical than the specific treatment used. However, model performance was limited by class imbalance, as even the best models struggled with recall-precision balance, highlighting challenges in predicting rare events like PE. Additionally, the small dataset size and absence of Current Procedural Terminology (CPT) data may have restricted the ability to fully assess the impact of different treatment types, particularly MT and CDT. These findings emphasize early DVT treatment and show how machine learning enhances PE risk assessment and informs clinical decisions, contributing to a data-driven approach to PE prevention and optimized treatment strategies.

Table of Contents

1 Introduction

Venous thromboembolism (VTE) is a medical condition characterized by the presence of blood clotting in veins, which can lead to serious complications if untreated. The two primary manifestations of VTE are deep vein thrombosis (DVT), where a clot forms in a deep vein, typically in the lower extremities, and pulmonary embolism (PE), where a clot dislodges and travels to the lungs (National Heart, Lung, and Blood Institute 2022). Because DVT increases the risk of PE, which can be fatal, prompt and effective management of DVT is critical (Dexter et al. 2022).

The standard of treatment for DVT has been anticoagulation (AC) since the 1930s. Another historical treatment approach involves systemic (non-targeted) thrombolytics, where clot-dissolving drugs are administered intravenously. However, thrombolytics are no longer recommended because of increased incidence of bleeding complications (Lin 2024). Due to the limitations in traditional AC and systemic thrombolytics, interventional approaches such as mechanical thrombectomy (MT) and catheter-directed thrombolysis (CDT) have been developed as alternative treatment strategies. MT employs a mechanical device to physically break down and remove clot from the vein while CDT involves infusing thrombolytics directly into the clot via a catheter to facilitate dissolution. MT is a faster method than CDT, requiring shorter hospital stays and lower thrombolytic drug doses, reducing bleeding risks (Zhang et al. 2024).

Despite these treatment advancements, there is a lack of comprehensive data evaluating whether the risk of subsequent PE development is influenced by the initial DVT treatment type. Additionally, an estimated 600,000 VTE events occur annually in the United States, but many cases go undiagnosed because they can occur without obvious signs or have symptoms that are compatible with many other conditions (National Heart, Lung, and Blood Institute 2022). This clinical ambiguity underscores a need for a deeper understanding of the risk factors of PE following a DVT to improve diagnostic certainty. A promising approach to addressing these challenges is predictive modeling using supervised machine learning techniques. This study aims to address these questions by developing machine learning models that predict the risk of PE following a DVT diagnosis, incorporating demographic, clinical, and treatment data to enhance risk assessment and support clinical decision-making.

2 Literature Review

The risk of PE following DVT has been well studied in the context of AC therapy, but gaps remain in understanding the impact of other treatments such as systemic thrombolytics, MT, and CDT on PE incidence. For example, Douketis et al. (1998) sought to provide better estimates of the risk of fatal PE in patients with prior DVT or PE who were treated with AC therapy. Their study was motivated by the need to inform clinical decision-making regarding the safety of discontinuing AC therapy. Through a systematic review of 25 prospective studies conducted between 1966 and 1997, they found that the risk of fatal PE following AC therapy was 0.3 per 100 patient-years, suggesting that PE is rare both during and after AC therapy. However, this study only evaluated fatal PE and did not consider non-fatal PE events, limiting its applicability to understanding the broader risk of PE development. Additionally, the analysis was restricted to three-month follow-up periods and did not account for the impact of other treatment types.

Similarly, Patel et al. (2020) conducted a meta-analysis of 12 prospective, cross-sectional, and cohort studies spanning 1974 to 2019 to assess PE incidence in patients with suspected symptomatic DVT. Their analysis revealed that patients diagnosed with lower extremity DVT and treated with AC therapy had a 0% PE incidence at three months, whereas those with upper extremity DVT had a 1.98% incidence of PE at three months. While this study provides valuable insights into the short-term risk of PE, it shares similar limitations with Douketis et al. (1998) in that it only evaluated outcomes over a three-month period and exclusively considered AC therapy as DVT treatment type.

Moving beyond short-term PE incidence, Monreal et al. (1992) conducted a prospective study from 1985 to 1991 involving 434 patients with acute lower extremity DVT, with or without PE, confirmed via venography and lung scan. Their objective was to identify risk factors associated with PE incidence following DVT. Patients received intravenous AC therapy for eight days, followed by oral AC if no contraindications existed. The study found that PE was more frequently observed in patients with proximal DVT compared to those with distal DVT. Additionally, a logistic regression analysis identified a history of VTE as the only statistically significant predictor of PE, with a more than twofold increased risk. Although the study provided critical insight into PE risk factors, it was limited to patients with lower extremity DVT, excluding upper extremity cases. Furthermore, again, only AC was studied, leaving unanswered questions regarding the effectiveness of other treatments in preventing PE.

Dexter et al. (2022) contributed to understanding the effectiveness of other DVT treatment types by investigating the use of the MT “ClotTriever” device created by Inari Medical for DVT treatment in a multicenter, prospective registry. Their study compared MT with AC and

CDT, and found that MT resulted in higher thrombus removal rates and lower post-thrombotic syndrome (PTS) rates at six months compared to AC alone. Additionally, MT offered advantages over CDT by achieving similar thrombus clearance without the use of thrombolytics, thereby reducing bleeding risks and intensive care unit (ICU) admissions. However, while the registry provided valuable data on DVT treatment effectiveness, it did not specifically assess PE development as an outcome. Instead, PE risk was inferred from incidental findings of DVT during follow-up rather than through direct measurement of PE incidence.

The aforementioned studies used traditional statistical analysis methods to garner a greater understanding of PE risk factors. However, machine learning approaches have been introduced to improve PE prediction. Machine learning has the capability to reaffirm previously determined risk factors and may even help identify novel risk factors that may not be immediately obvious through traditional statistical analysis. Ryan et al. (2022) developed a machine learning algorithm to predict the development or clinical detection of PE in hospitalized patients at any point during their stay. Using electronic health record data from 63,798 medical and surgical patients aged 40 and older from a large U.S. medical center (2011–2017), they trained logistic regression, gradient-boosted tree, and neural network models to predict PE occurrence. The gradient-boosted tree model demonstrated the best performance, with feature importance analysis identifying recent fracture, history of surgery, and prior DVT as the most influential predictors. Notably, the inclusion of AC prescription as a predictor did not improve model performance. Despite the study's strengths in utilizing machine learning for PE prediction, it did not specifically assess PE risk after an initial DVT diagnosis, instead considering PE risk in a broad hospitalized patient population. Additionally, it excluded patients under 40, limiting its generalizability to younger populations at risk for DVT-related complications.

Despite extensive research on AC therapy and its impact on PE risk, and newer research into the impact of interventional treatments for DVT, there is a lack of studies evaluating if and how different initial DVT treatments, both drug-based and interventional, affect subsequent PE incidence. Additionally, machine learning approaches to PE prediction have not specifically targeted patients with an initial DVT diagnosis, limiting their applicability to understanding the progression from DVT to PE and the specific risk factors contributing to this transition. This study aims to address these gaps by developing predictive machine learning models to assess PE risk specifically in patients diagnosed with DVT and comparing risk factors across AC, systemic thrombolytics, MT, and CDT treatments. This study will contribute to a more comprehensive understanding of treatment-related PE risk and inform optimal DVT management strategies to prevent life-threatening PE events.

3 Data

3.1 Data Overview

The data for this study came from the Medical Information Mart for Intensive Care (MIMIC)-IV, a publicly available database with de-identified electronic health records from 223,452 unique patients aged 18 years and older admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts between 2008 and 2019 (Johnson et al. 2024). MIMIC-IV has a relational database structure and data are grouped into three modules: hosp, icu, and note. The hosp module contains admission, discharge, and transfer data, lab values, microbiology cultures, medication orders, and administrative data. The icu module stores data documented in the ICU. The note module comprises discharge summaries and radiology reports. The latter module was not used in this study. Each patient was assigned a unique identifier (subject ID) and each hospitalization received a hospital admission ID (hospital admission ID) to ensure deidentification. Additionally, dates were offset but the intervals between any two time points were preserved (Johnson et al. 2023).

3.2 Data Extraction

To extract the required data from MIMIC-IV, PhysioNet, a secure platform for accessing medical datasets, was used (Goldberger et al. 2000). Access required PhysioNet account registration, completion of a human subjects research training, and signing a data use agreement. Data extraction was conducted directly via Google BigQuery.

To identify relevant patient records, the first step was to compile a list of International Classification of Diseases (ICD) codes for DVT and PE from both the ICD-9 and ICD-10 coding systems, as the transition from ICD-9 to ICD-10 occurred in 2015 (U.S. Department of Labor n.d.). This process was challenging due to the lack of a direct conversion chart, requiring the development of a manually curated mapping, presented in Appendix A, based on extensive research (ICD9Data.com n.d.; ICD10Data.com n.d.; Rajagopal 2024). Appendix B provides the full list of relevant ICD diagnosis codes.

The main query retrieved patients with DVT diagnoses, categorizing them by chronicity (acute, chronic, or unspecified) and anatomical location (upper extremity, lower extremity, or unspecified). PE diagnoses were also extracted using ICD codes to establish the target outcome field, a binary field indicating whether the patient had a PE event after the initial DVT diagnosis or not. Additional patient-level attributes, including demographics (age, gender, race, marital status, and insurance type), hospital admission details (admission type, location, length of hospital stay, discharge status, and mortality status), and clinical history (prior DVT, prior PE, prior VTE, and history of long-term AC use) were extracted. To obtain a subject-level dataset, only the first recorded DVT diagnosis per patient was included, selecting the earliest DVT admission for each subject. Patients with a concurrent PE during their initial DVT admission were excluded to focus the analysis on patients who had an initial DVT without immediate PE complications. Additional features were created to capture DVT-related data, including a flag to indicate whether DVT was the primary admission diagnosis, the total number of DVT diagnoses across all admissions, and the total number of DVT-related hospital admissions per patient.

To account for underlying patient health conditions, the Charlson Comorbidity Index (CCI) was computed from historical diagnosis records. The CCI is a widely used scoring system that quantifies the burden of 17 major comorbidities, including cancer, chronic pulmonary disease, diabetes, and heart failure. The CCI was originally developed as a prognostic tool for mortality risk stratification in longitudinal studies (Charlson et al. 1987). The original study demonstrated an increase in mortality risk with higher comorbidity scores, validating the index as a strong predictor of long-term patient outcomes. To incorporate comorbidity burden into the analysis, each comorbidity was identified using ICD-9 and ICD-10 codes, mapped to the CCI scoring system, and assigned a weighted score following the established methodologies (Johnson et al. 2017; Johnson et al. 2024; Johnson et al. 2025).

To investigate the potential influence of biomarkers in PE risk prediction, relevant lab test data was extracted. This included D-dimer and oxygenation tests, which are commonly used in VTE diagnosis workups (National Heart, Lung, and Blood Institute 2022). For each patient, lab measurements within a 7-day window following the initial DVT diagnosis were identified, and flags were created to indicate whether the relevant tests were performed. The extracted data was merged with the DVT cohort based on subject ID and hospital admission ID to ensure the lab tests were linked to the correct DVT admission event.

Finally, treatment data was extracted for AC therapy, thrombolytics, MT, and CDT. Appendix C provides a detailed mapping of treatment types. Medication names for AC and thrombolytics were identified through extensive research (Drugs.com n.d.; Moriarty 2015; WebMD 2024). MT and CDT procedures were classified using ICD-9 and ICD-10 Procedural Coding System (PCS) codes, which were identified through research. The ICD-PCS system serves as a standardized framework for categorizing hospital-performed procedures (Mouawad 2019; Penumbra Inc. 2018; Shafi et al. 2023).

Treatment initiation dates were identified by querying prescriptions and procedure records, enabling the calculation of how much time passed from the initial DVT diagnosis to the initiation of treatment. The extracted treatment data was merged with the DVT cohort based on subject ID and hospital admission ID to ensure treatments were linked to the correct DVT admission event. Patients who received multiple treatments had each treatment type flagged, and time to each treatment was recorded to facilitate further analysis of treatment timing and its impact on PE risk.

As illustrated in Figure 1, after applying the DVT inclusion criteria based on the manually curated ICD-9 and ICD-10 code list, selecting each patient’s first recorded DVT diagnosis, and excluding cases with concurrent PE, the final cohort comprised 4,723 unique patients. However, patients who expired during their initial DVT admission were excluded from the primary dataset since they could not have developed PE post-discharge. These patients were retained in a separate dataset for exploratory data analysis to investigate mortality trends. This resulted in a final cohort of 4,410 unique patients. 4,193 of these patients (95.1%) did not have a future PE event and 217 patients (4.9%) did. The data was exported as comma-separated value (CSV) files and brought into Python for further preprocessing and analysis.

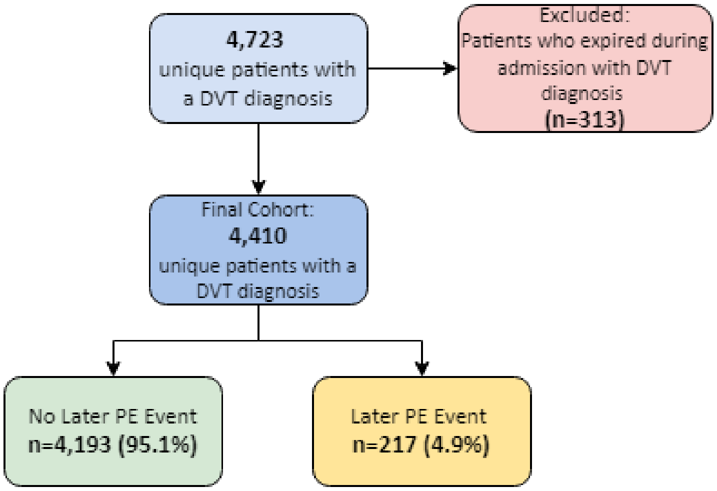

4,723
unique patients with a DVT diagnosis
Excluded:
Patients who expired during admission with DVT diagnosis
(n=313)
Final Cohort:
4,410
unique patients with a DVT diagnosis
No Later PE Event
n=4,193 (95.1%)
Later PE Event
n=217 (4.9%)

Figure 1. Cohort selection for patients with DVT. Of 4,723 identified cases, 313 were excluded due to in-hospital mortality, leaving 4,410 patients. Among them, 4,193 (95.1%) had no later PE event, while 217 (4.9%) developed a PE.

3.3 Data Preprocessing

Data preprocessing was performed in a Python Jupyter Notebook using the Pandas, NumPy, Matplotlib, and Seaborn libraries. First, the four CSV files were merged into a single dataset using "subject ID" as the identifier. Next, correct data types were ensured. Categorical variables were all made to be the object type, numeric variables either integer or float types, and binary variables the integer type. Next, the data was checked for duplicates, which it did not contain. Regarding the handling of missing data, for demographic and admission-related fields (discharge location, insurance, and marital status), missing values were replaced with “Unknown” to retain all available patient records. Missing values in PE-related fields (PE ICD code, ICD version, and diagnosis) were assigned “No PE” to explicitly document the absence of a recorded PE diagnosis. Additionally, patients with a prior history of DVT or PE who lacked a flag for history of VTE were appropriately flagged to ensure accurate classification of historical VTE cases. Likewise, patients with a history of long-term AC use were verified to ensure they had the AC flag and a “days to AC" value of 0, indicating they were already receiving AC therapy at admission.

3.4 Feature Engineering

Feature engineering was performed both during the data collection phase in Google BigQuery as well as in Python for further feature derivation. The engineered features were created through transformations, categorization, and consolidation of variables with the goal of enhancing the predictive power of subsequent machine learning models. A detailed breakdown of the engineered features is provided in Appendix C for features derived using Google BigQuery and in Appendix D for those derived in Python.

3.5 Exploratory Data Analysis

To better understand the dataset, exploratory data analysis (EDA) was conducted by building and using an interactive Power BI dashboard to examine demographics, comorbidities, treatment distribution, and PE outcomes across subgroups.

3.5.1 Demographics & Comorbidities

Across the full dataset of 4,410 patients, 24.2% presented with DVT as their primary diagnosis. Over half (51.4%) had a history of VTE, 49.2% had a history of DVT, and 12.7% had a history of PE, highlighting that many patients were already at risk for recurrent clotting events. Demographically, the majority (68.0%) of patients were White, followed by Black (15.6%), Hispanic/Latino (4.3%), Asian (2.6%), and the rest unknown or other. Insurance coverage was largely Medicare (50.9%), private insurance (29.1%), and Medicaid (16.5%), reflecting an older or lower socioeconomic patient population. The gender distribution was nearly even (47.6% female and 52.4% male), and the mean age was 63.7 years, suggesting that DVT and PE are more common in older adults.

The CCI peaked at 5, indicating a high burden of chronic conditions. As CCI increased, the percentage of primary DVT diagnoses decreased, likely due to other competing health concerns. Among comorbidities, cancer (13.5%) was the most prevalent, followed by renal disease (11.5%), diabetes (10.8%), chronic pulmonary disease (10.3%), and congestive heart failure (9.8%). The high prevalence of cancer aligns with research showing malignancy as a major risk factor for hypercoagulability and thrombosis (Caine et al. 2002). Additionally, chronic conditions such as diabetes, kidney disease, and heart failure are associated with endothelial dysfunction, which can contribute to clot formation and poor vascular health (Baaten et al. 2023).

3.5.2 DVT Diagnosis & Treatment Trends

On average, patients had 1.4 DVT-related admissions and 1.7 DVT diagnoses, indicating that many experienced multiple thrombotic events requiring hospitalization. Most DVTs were acute (85.3%) and lower extremity DVTs were the most common (60.4%), aligning with prior research (National Heart, Lung, and Blood Institute 2022).

Regarding treatment, 81.7% received AC alone, reflecting the standard of care (Lin 2024), while 8.5% received systemic thrombolytics. Interventional treatments were rarely used, with only 1.2% receiving CDT, 0.7% undergoing MT, and 0.9% receiving multiple interventions. Additionally, 7.0% of patients had no identified treatment, suggesting the need for further investigation into alternative therapies or potential gaps in documentation. The timing of treatment initiation showed that 74.3% of patients started treatment on the day of diagnosis, while 16.3% began within one to three days, 1.0% between four and seven days, and 0.5% after more than a week. This suggests that while most patients received timely intervention, delays in treatment initiation may still contribute to PE risk and require further examination.

3.5.3 DVT-Related Mortality

Among patients with a DVT diagnosis, 6.6% expired during their admission. Only 2.9% of these patients had DVT as their primary diagnosis, suggesting that most had other critical conditions contributing to their mortality. Additionally, a high percentage (75.4%) of expired patients received AC therapy alone, which raises concerns about whether AC was sufficient for high-risk cases or if alternative or more aggressive treatments might have improved outcomes.

3.5.4 PE Incidence & Timing

A total of 217 patients (4.9%) experienced a PE event. Compared to the overall cohort, these patients had higher rates of prior VTE (84.3%), DVT (81.6%), and PE (51.2%), reinforcing that history of VTE increases PE risk (Monreal et al. 1992). Among these patients, 35.0% had primary DVTs and 57.6% had acute, lower extremity DVTs. This higher proportion of primary

DVTs compared to the overall cohort suggests that patients who initially present with DVT as their primary diagnosis may have an elevated risk of developing PE.

The timing of PE events relative to DVT diagnosis showed that 26.7% of PE events occurred within one month, 13.8% between one and three months, 9.2% between three and six months, 9.7% between six and twelve months, and 40.1% more than a year later. This highlights that a large number of PEs occur long after DVT diagnosis, underscoring the need for long-term risk assessment and monitoring in these patients, an aspect not emphasized in prior studies (Douketis et al. 1998; Patel et al. 2020).

3.5.5 PE Treatment Trends

Among treatment groups, CDT had the highest PE event rate (9.6%), followed by thrombolytics (5.9%), AC-only (5.0%), and MT (3.2%). These findings align with Dexter et al. (2022), which reported better post-DVT outcomes with MT compared to AC-only and CDT, though their study did not directly assess PE incidence. The findings also suggest that AC alone may not be sufficient for all DVT cases. Patel et al. (2020) reported a 0% PE incidence at three months in patients with lower extremity DVTs and 1.98% in those with upper extremity DVTs receiving AC therapy. In contrast, this study's EDA found an overall PE rate of 5.0% in patients treated with AC therapy alone, though this was measured across different time frames rather than at a standardized three-month interval. Lastly, 89.9% of patients who developed a later PE received initial DVT treatment on the same day of their diagnosis, 6.5% one to three days after their diagnosis, and 3.7% received no treatment. This suggests that while early DVT treatment is common, PE risk may still be influenced by other factors like patient characteristics, disease severity, or treatment effectiveness.

3.5.6 PE Key Takeaways

This EDA highlights several important trends. Older age, history of VTE, and comorbid conditions, especially cancer, appear to be key risk factors for PE in patients diagnosed with DVT. Lower extremity DVTs were more strongly linked to PE than upper extremity DVTs, and a large proportion of PEs occurred more than a year after DVT diagnosis, underscoring the need for long-term monitoring. Among treatments, CDT had the highest PE event rate, while MT had the lowest, suggesting that AC therapy alone may not be sufficient for all cases. These findings lay the groundwork for predictive modeling to enhance risk stratification and guide personalized treatment strategies for DVT.

4 Methods

This section details the key steps taken in data extraction, data preprocessing, EDA, feature engineering, and modeling. A full breakdown of the data pipeline is provided in Figure 2.

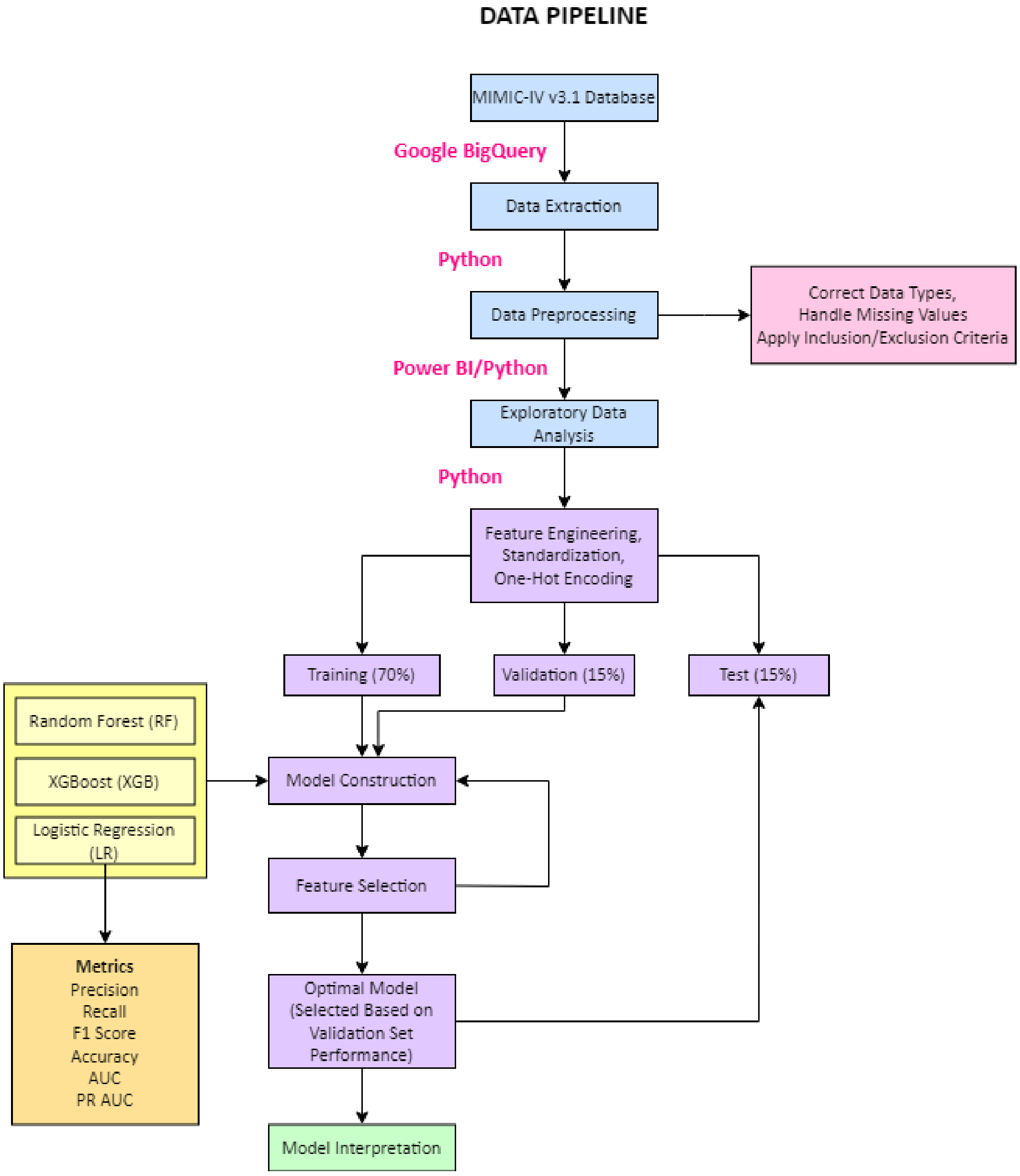


Figure 2. Data pipeline for PE risk prediction using MIMIC-IV v3.1. Data was extracted via BigQuery, processed in Python, and analyzed with Power BI/Python. After feature engineering, data was split into training (70%), validation (15%), and test (15%) sets. Models underwent feature selection, with the best model chosen based on validation performance. Evaluation metrics included precision, recall, F1-score, accuracy, AUC, and PR AUC, followed by model interpretation.

4.1 Model Preparation

After completion of EDA, the data was prepared for baseline modeling in Python. First, unnecessary fields were dropped from the dataset. A field was determined to be unnecessary if it would not have any predictive value or if it was only relevant for exploration. For example, fields like subject ID and hospital admission ID were dropped because they are just unique identifiers and have no inherent relationship with the target variable. Next, the data was split into training (70%), validation (15%), and test (15%) sets using the scikit-learn package. Subsequently, numeric fields in the training set were standardized to ensure all features were on a comparable scale, preventing features with larger ranges from disproportionately influencing the model. This same transformation was applied to the validation and test set numeric fields. Finally, categorical variables were one-hot encoded in all datasets to convert them into a numerical format suitable for machine learning models while preserving their categorical nature.

4.2 Model Performance Metrics

To evaluate the performance of baseline models in predicting PE events, multiple performance metrics were employed. Given the severe class imbalance in the dataset, where 95.1% of cases did not have a PE event, relying on accuracy alone would not suffice as a model that simply predicts an outcome of no PE for every case would achieve 95% accuracy but would fail to identify any actual PE events, making it unusable. Instead, a combination of metrics was selected to evaluate model performance. These included precision, recall, F1 score, area under the receiver operating characteristic (AUC-ROC) curve, and area under the precision-recall curve

(PR-AUC). Recall was prioritized because it measures the proportion of actual PE cases correctly identified by the model. Since missing a PE (false negative) could have life-threatening consequences, a high recall was important. However, an overemphasis on recall can lower precision, and this can lead to a high number of false positives where non-PE cases are incorrectly classified as PE. Low precision could have real-life consequences like unnecessary follow-ups and additional testing. However, since the clinical implications of missing a PE are far more severe than the burden of additional testing, recall was prioritized over precision.

To balance recall and precision, the F1-score was used, which is the harmonic mean of the two. This metric is particularly useful in situations where both false negatives and false positives have consequences and provides a single measure of model effectiveness. Additionally, the AUC-ROC was used to evaluate the model's ability to distinguish between PE and non-PE cases across different probability thresholds. A higher AUC value indicates stronger discrimination between the two classes. Since precision-recall trade-offs are especially important in imbalanced datasets, the PR-AUC was also considered, as it directly evaluates how well the model maintains precision as recall increases.

Model performance was first assessed on the validation set to compare different algorithms, model parameters, and class imbalance strategies. The best-performing models were evaluated on the test set to estimate real-world generalizability.

4.3 Baseline Models

Multiple baseline models were developed to assess their effectiveness in predicting PE. Three commonly used classification algorithms were selected: random forest (RF), XGBoost (XGB), and logistic regression (LR). These models were evaluated under different conditions, incorporating various resampling and weighting strategies to address the dataset’s large class imbalance.

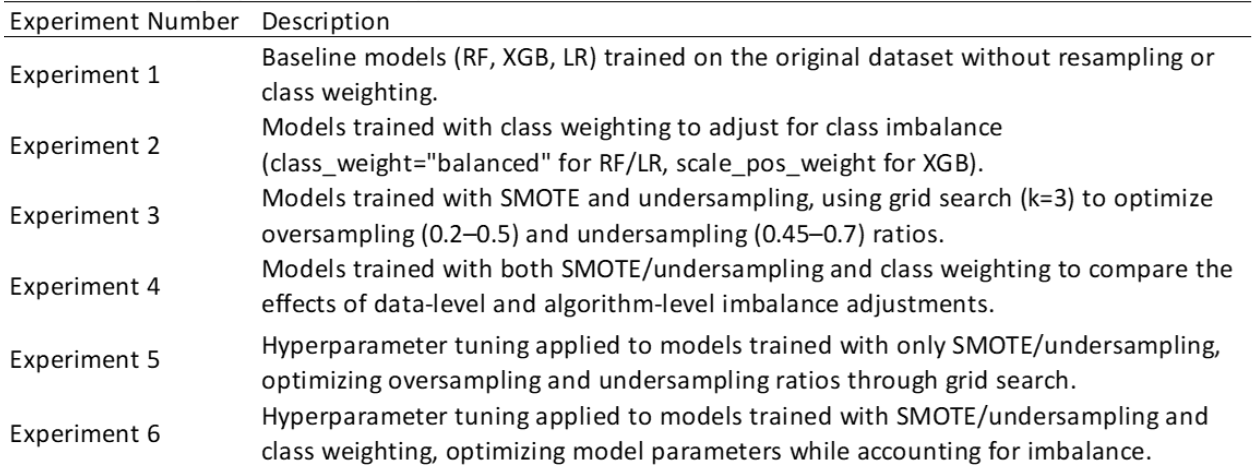
The first set of baseline models was trained on the original dataset without any resampling or weighting. These models served as the most basic benchmark, reflecting how standard classifiers perform on an imbalanced dataset. To mitigate class imbalance, a second set of models incorporated class weighting, adjusting the importance of the minority class. These adjustments ensured that the models placed greater emphasis on correctly identifying PE cases.

To further address class imbalance, a third set of models applied synthetic minority oversampling technique (SMOTE) and random undersampling, using grid search with cross-validation (k=3) to determine optimal oversampling and undersampling ratios. The SMOTE ratios tested were 0.2, 0.25, 0.3, 0.4, 0.45, and 0.5, while the undersampling ratios tested were 0.45, 0.5, 0.55, 0.6, 0.65, and 0.7. These values were chosen to explore varying minority-to-majority class balances, ensuring that models were evaluated under different class distributions while balancing the retention of majority class information with the need to mitigate class imbalance. These resampled models were tested both with and without class weighting to compare the impact of resampling alone versus resampling combined with weighting.

Building on these models, the final set of experiments incorporated hyperparameter tuning using grid search with cross-validation (k=3) in addition to SMOTE and undersampling. These models were also tested in both weighted and unweighted versions to systematically evaluate the combined impact of resampling, class weighting, and model optimization. Details of the hyperparameter tuning process, including specific parameter grids for each model, can be provided upon request. Table 1 provides an overview of the six experimental conditions tested, highlighting the different combinations of resampling, weighting, and tuning strategies. The results from these baseline models served as the foundation for further refinements and performance improvements in subsequent modeling stages.

Table 1

Machine Learning Experiment Descriptions



Experiment Number	Description
Experiment 1	Baseline models (RF, XGB, LR) trained on the original dataset without resampling or class weighting.
Experiment 2	Models trained with class weighting to adjust for class imbalance (class_weight="balanced" for RF/LR, scale_pos_weight for XGB).
Experiment 3	Models trained with SMOTE and undersampling, using grid search (k=3) to optimize oversampling (0.2–0.5) and undersampling (0.45–0.7) ratios.
Experiment 4	Models trained with both SMOTE/undersampling and class weighting to compare the effects of data-level and algorithm-level imbalance adjustments.
Experiment 5	Hyperparameter tuning applied to models trained with only SMOTE/undersampling, optimizing oversampling and undersampling ratios through grid search.
Experiment 6	Hyperparameter tuning applied to models trained with SMOTE/undersampling and class weighting, optimizing model parameters while accounting for imbalance.

4.4 Feature Selection and Model Refinement

Following baseline model evaluation, feature selection was conducted to remove redundant or low-impact features, improving model interpretability and efficiency. Three selection methods were applied before rerunning all models. First, low variance filtering was performed using a variance threshold of 0.01 to eliminate features with minimal variability across samples, as these features provided little discriminatory power for classification. Next, correlation-based feature elimination was conducted using correlation heatmaps and variance inflation factor (VIF) analysis to identify and remove highly correlated features, reducing redundancy and mitigating multicollinearity. Finally, top feature selection was performed by extracting the ten most important features from the best-performing RF, XGB, and LR models.

Feature importance scores from tree-based models and coefficient values from LR were used to determine the most predictive variables.

After feature selection, all six baseline modeling pipelines were rerun using the reduced feature set to assess whether feature reduction improved performance. This approach ensured that models were trained on the most relevant predictors, enhancing both interpretability and computational efficiency.

5 Results

The following section presents the performance of RF, XGB, and LR models under the various experimental conditions.

5.1 Baseline Models

Tables 2, 3, and 4 present the validation results of the RF, XGB, and LR models trained on the full training dataset, evaluating the impact of different resampling, class weighting, and tuning strategies.

Table 2

Performance Metrics on Validation Data for Baseline Random Forest Models

Experiment	Model	Precision	Recall	F1-Score	Accuracy	ROC AUC	PR AUC
Experiment 1	Random Forest	0.750	0.188	0.300	0.958	0.824	0.313
Experiment 2	Random Forest	0.625	0.156	0.250	0.955	0.839	0.290
Experiment 3	Random Forest	0.579	0.344	0.431	0.956	0.846	0.293
Experiment 4	Random Forest	0.450	0.281	0.346	0.949	0.845	0.312
Experiment 5	Random Forest	0.293	0.375	0.329	0.926	0.860	0.307
Experiment 6	Random Forest	0.262	0.500	0.344	0.908	0.861	0.342

Table 3

Performance Metrics on Validation Data for Baseline XGBoost Models

Experiment	Model	Precision	Recall	F1-Score	Accuracy	ROC AUC	PR AUC
Experiment 1	XGBoost	0.385	0.156	0.222	0.947	0.777	0.230
Experiment 2	XGBoost	0.308	0.125	0.178	0.944	0.802	0.218
Experiment 3	XGBoost	0.171	0.375	0.235	0.882	0.791	0.174
Experiment 4	XGBoost	0.227	0.469	0.306	0.897	0.781	0.226
Experiment 5	XGBoost	0.241	0.594	0.342	0.890	0.872	0.352
Experiment 6	XGBoost	0.191	0.688	0.299	0.844	0.869	0.315

Table 4

Performance Metrics on Validation Data for Baseline Logistic Regression Models

Experiment	Model	Precision	Recall	F1-Score	Accuracy	ROC AUC	PR AUC
Experiment 1	Logistic Regression	0.133	0.688	0.223	0.769	0.800	0.170
Experiment 2	Logistic Regression	0.129	0.656	0.215	0.769	0.805	0.174
Experiment 3	Logistic Regression	0.141	0.688	0.234	0.782	0.820	0.181
Experiment 4	Logistic Regression	0.150	0.750	0.250	0.782	0.808	0.176
Experiment 5	Logistic Regression	0.132	0.781	0.226	0.741	0.851	0.225
Experiment 6	Logistic Regression	0.133	0.812	0.228	0.734	0.851	0.226

Table 5 presents the performance metrics of the best-performing model for each model type. Experiment 6—incorporating SMOTE, random undersampling, class weighting, and hyperparameter tuning—achieved the highest performance across all model types. Table 6 summarizes the performance of these top models when evaluated on the test set, providing an estimate of their generalizability. Figure 3 visualizes the test set AUC-ROC curves for the best-performing baseline models, with RF achieving an AUC of 0.789, XGB 0.797, and LR 0.774. Additionally, Figure 3 includes confusion matrices illustrating model classification performance on the test data.

Table 5

Performance Metrics on Validation Data for Best Baseline Models

Experiment	Model	Precision	Recall	F1-Score	Accuracy	ROC AUC	PR AUC
Experiment 6	Random Forest	0.133	0.813	0.228	0.734	0.851	0.226
Experiment 6	XGBoost	0.262	0.500	0.344	0.908	0.861	0.342
Experiment 6	Logistic Regression	0.191	0.688	0.299	0.844	0.869	0.315

Table 6

Performance Metrics on Test Data Using Best Baseline Models

Model	Precision	Recall	F1-Score	Accuracy	ROC AUC	PR AUC
Random Forest	0.208	0.333	0.256	0.903	0.789	0.219
XGBoost	0.071	0.939	0.132	0.385	0.797	0.197
Logistic Regression	0.126	0.667	0.212	0.752	0.774	0.261

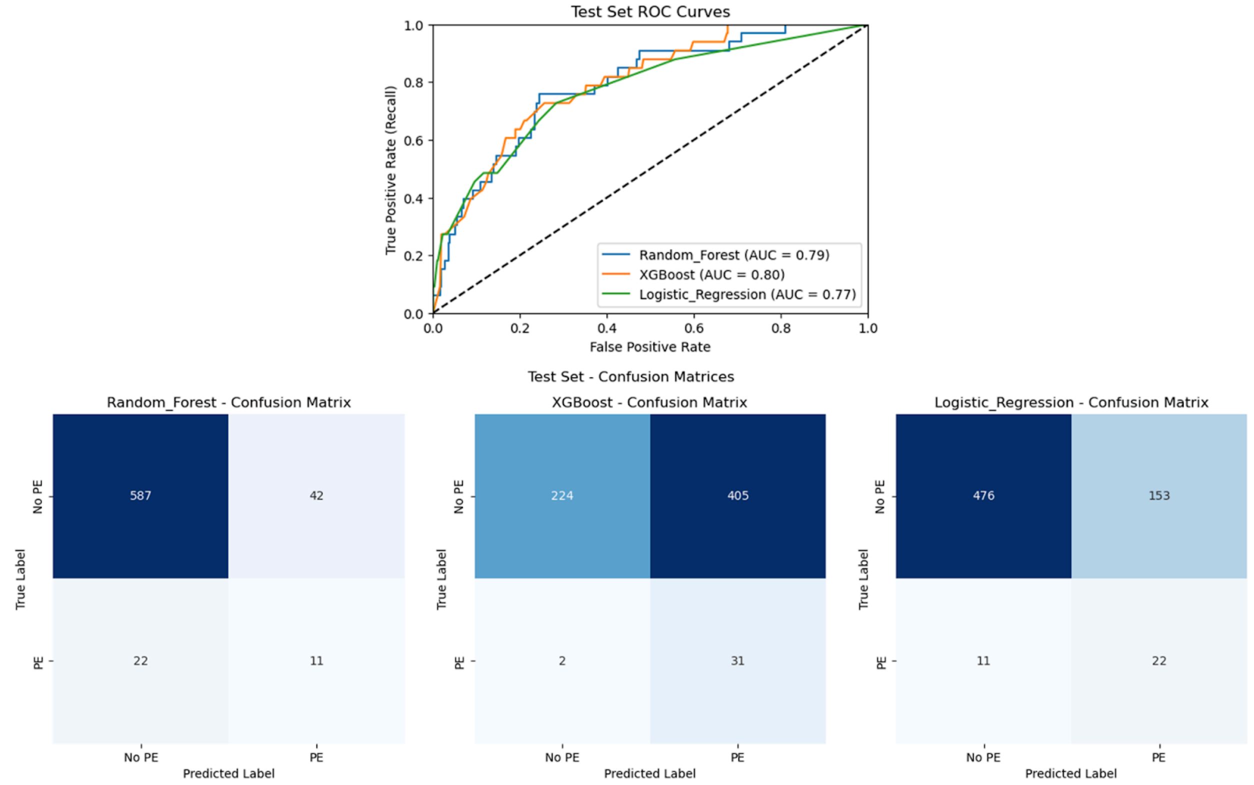


Figure 3. Test set AUC-ROC curves and confusion matrices for the best baseline models. The top plot shows ROC curves with AUC scores: RF (0.79), XGB (0.80), and LR (0.77). The bottom

section presents test set confusion matrices, comparing model performance in classifying PE and no PE cases.

5.2 Models After Feature Selection

After conducting experiments on the full training data, feature selection was applied as outlined in the Methods section, resulting in a refined dataset with reduced features. The six experiments were then repeated using this dataset. Tables 7, 8, and 9 present the results for the RF, XGB, and LR models, respectively.

Table 7

Performance Metrics on Validation Data for Random Forest Models After Feature Selecton

Experiment	Model	Precision	Recall	F1-Score	Accuracy	ROC AUC	PR AUC
Experiment 1	Random Forest	0.212	0.219	0.215	0.923	0.836	0.274
Experiment 2	Random Forest	0.219	0.219	0.219	0.924	0.839	0.263
Experiment 3	Random Forest	0.146	0.375	0.211	0.864	0.838	0.237
Experiment 4	Random Forest	0.160	0.406	0.230	0.868	0.836	0.242
Experiment 5	Random Forest	0.197	0.469	0.278	0.882	0.878	0.274
Experiment 6	Random Forest	0.157	0.688	0.256	0.806	0.864	0.260

Table 8

Performance Metrics on Validation Data for XGBoost Models After Feature Selecton

Experiment	Model	Precision	Recall	F1-Score	Accuracy	ROC AUC	PR AUC
Experiment 1	XGBoost	0.320	0.250	0.281	0.938	0.809	0.261
Experiment 2	XGBoost	0.241	0.219	0.230	0.929	0.801	0.255
Experiment 3	XGBoost	0.207	0.375	0.267	0.900	0.803	0.200
Experiment 4	XGBoost	0.194	0.406	0.263	0.890	0.815	0.201
Experiment 5	XGBoost	0.176	0.688	0.280	0.829	0.861	0.335
Experiment 6	XGBoost	0.153	0.750	0.254	0.787	0.864	0.288

Table 9

Performance Metrics on Validation Data for Logistic Regression Models After Feature Selecton

Experiment	Model	Precision	Recall	F1-Score	Accuracy	ROC AUC	PR AUC
Experiment 1	Logistic Regression	0.136	0.688	0.227	0.773	0.852	0.189
Experiment 2	Logistic Regression	0.144	0.719	0.240	0.779	0.849	0.192
Experiment 3	Logistic Regression	0.159	0.719	0.260	0.802	0.850	0.189
Experiment 4	Logistic Regression	0.135	0.719	0.227	0.762	0.867	0.192
Experiment 5	Logistic Regression	0.132	0.781	0.226	0.741	0.843	0.204
Experiment 6	Logistic Regression	0.128	0.875	0.224	0.707	0.849	0.220

Table 10 presents the performance metrics of the best-performing model for each model type. As with the full feature set, experiment 6—which incorporated SMOTE, random undersampling, class weighting, and hyperparameter tuning—yielded the highest performance across all model types. Table 11 summarizes the test set performance of the best models after feature selection. Figure 4 visualizes the AUC-ROC curves for these models, showing AUC values of 0.776 for RF, 0.789 for XGB, and 0.774 for LR. Additionally, Figure 4 includes confusion matrices for the best-performing models to further illustrate their classification performance on the test data.

Table 10

Performance Metrics on Validation Data for Best Models After Feature Selection

Experiment	Model	Precision	Recall	F1-Score	Accuracy	ROC AUC	PR AUC
Experiment 6	Random Forest	0.128	0.875	0.224	0.707	0.849	0.220
Experiment 6	XGBoost	0.157	0.688	0.256	0.806	0.864	0.260
Experiment 6	Logistic Regression	0.153	0.750	0.254	0.787	0.864	0.288

Table 11

Performance Metrics on Test Data Using Best Models After Feature Selection

Model	Precision	Recall	F1-Score	Accuracy	ROC AUC	PR AUC
Random Forest	0.152	0.606	0.242	0.811	0.776	0.171
XGBoost	0.068	0.939	0.126	0.352	0.789	0.194
Logistic Regression	0.119	0.727	0.204	0.718	0.774	0.261

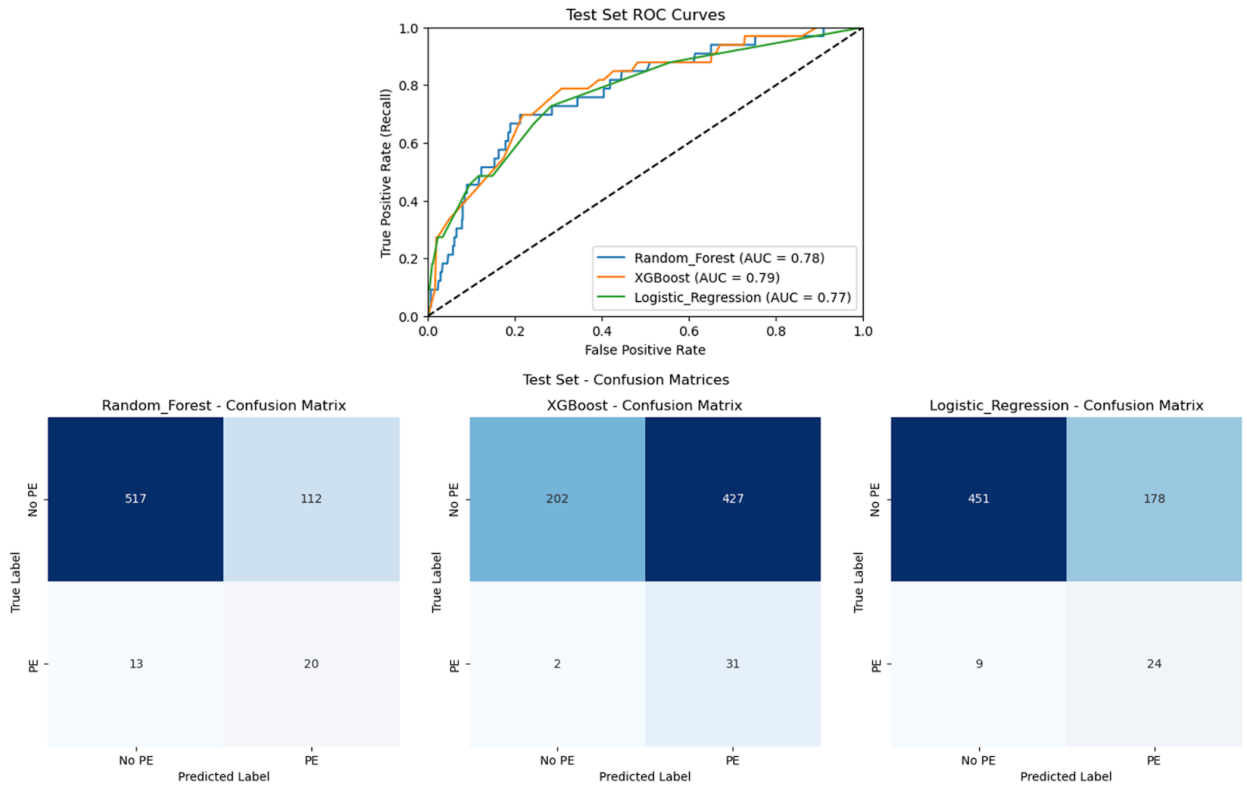


Figure 4. Test set AUC-ROC curves and confusion matrices for the best models after feature selection. The top plot shows ROC curves with AUC scores: RF (0.78), XGB (0.79), and LR (0.77). The bottom section presents test set confusion matrices, comparing model performance in classifying PE and no PE cases.

5.3 Feature Importance

After testing the models with the reduced features dataset, the most influential predictors for PE classification were identified. Table 12 presents the key features selected by the RF, XGB, and LR models.

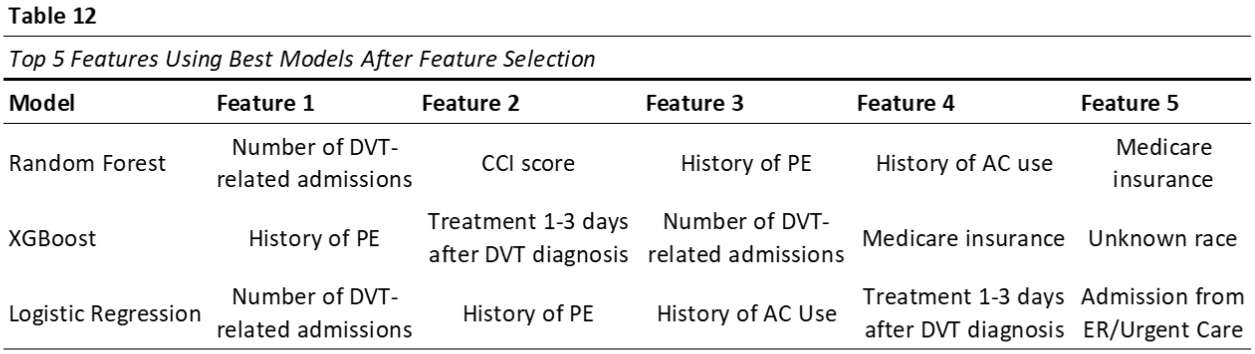


Table 12

Top 5 Features Using Best Models After Feature Selection

Model	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
Random Forest	Number of DVT-related admissions	CCI score	History of PE	History of AC use	Medicare insurance
XGBoost	History of PE	Treatment 1-3 days after DVT diagnosis	Number of DVT-related admissions	Medicare insurance	Unknown race
Logistic Regression	Number of DVT-related admissions	History of PE	History of AC Use	Treatment 1-3 days after DVT diagnosis	Admission from ER/Urgent Care

The results indicate that the number of DVT-related admissions, history of PE, and treatment timing (one to three days post-DVT diagnosis) were the most influential predictors across all models. These results will be examined further in the Discussion.

6 Discussion

This study aimed to understand the key variables involved in predicting PE risk in DVT patients, with a particular focus on whether specific DVT treatments influenced PE occurrence. To achieve this, the study evaluated the performance of RF, XGB, and LR models for predicting PE using various resampling techniques, class weighting, and hyperparameter tuning both before feature selection and after feature selection.

6.1 Interpretation of Results

Across both full-feature and reduced-feature datasets, experiment 6, which incorporated SMOTE, random undersampling, class weighting, and hyperparameter tuning using grid search and k-fold cross validation, consistently yielded the highest model performance. These findings reinforce the importance of addressing class imbalance and optimizing hyperparameters to improve PE prediction.

Feature selection had a notable impact on recall, particularly for RF (0.333 to 0.606) and LR (0.667 to 0.727), indicating that these models became more sensitive to identifying PE cases, which was a primary goal of modeling due to the consideration in a clinical setting where missing a PE case can have serious consequences. XGB maintained a consistently high recall (0.939) both before and after feature selection, suggesting that the reduced feature set preserved its ability to detect PE cases. However, the increase in recall resulted in a decline in precision across all models, indicating a higher rate of false positives. For instance, precision in RF decreased from 0.208 to 0.152, while precision in XGB dropped from 0.071 to 0.068, and precision in LR dropped from 0.126 to 0.119.

The F1-score slightly declined across all models after feature selection, with RF decreasing from 0.256 to 0.242, XGB from 0.132 to 0.126, and LR from 0.212 to 0.204. This indicates that while feature selection improved recall, it did not enhance the balance between precision and recall, which remained low. The ROC-AUC scores remained largely consistent across models, with only minimal variation (RF: 0.789 to 0.776, XGB: 0.797 to 0.789, and LR: 0.774 to 0.774), indicating that feature selection did not meaningfully impact the models' overall ability to distinguish between PE and no PE cases. This suggests that while feature selection reduced the number of predictors, it had minimal impact on overall model performance, indicating that the retained features preserved most of the predictive power. As a result, feature selection effectively reduced model complexity and improved interpretability without compromising classification performance. However, the PR-AUC scores showed slight declines, particularly for RF (0.219 to 0.171), again highlighting that while recall improved, precision suffered in the process.

To identify the most important predictors in the best-performing models trained on the reduced dataset, the top five features from each model were examined. Across all models, the number of DVT-related admissions and a history of PE consistently emerged as the strongest predictors of PE risk. This aligns with existing research, as DVT is a well-established risk factor for PE, and patients with a prior PE are at higher risk for recurrence (Monreal et al. 2002; Ryan et al. 2022). Additionally, the XGB and LR models identified treatment timing (one to three days post-DVT diagnosis) as a significant factor. While EDA showed that 89.1% of patients who later developed a PE had received initial treatment on the same day as diagnosis, a delay of even a few days was still observed in some cases. This suggests that treatment delays, though uncommon, may increase PE risk, underscoring the importance of early intervention.

Additionally, RF and XGB models identified Medicare insurance status as a predictor, likely reflecting older age, which was a defining characteristic of the cohort based on the mean age (63.7 years) observed in EDA. RF models also highlighted CCI, reinforcing that patients with multiple comorbidities are at greater risk for PE. XGB uniquely identified unknown race as a factor, which may indicate disparities in healthcare access or potential data quality limitations. LR models emphasized admission from an ER or Urgent Care setting, suggesting that patients who initially sought emergency care may have been at higher risk for PE due to more severe clinical presentations at the time of diagnosis.

Interestingly, while EDA indicated potential differences in PE occurrence across treatment types, the predictive models did not identify specific treatments—AC-only, thrombolytics, MT, or CDT—as key predictors of PE risk. This suggests that although treatment type may be associated with PE development, it does not provide significant predictive value when accounting for other clinical and demographic factors. One possible explanation is that treatment selection is closely tied to disease severity and clinician discretion—variables that were not captured in the dataset. For instance, patients undergoing more invasive treatments like MT or CDT could have had more severe DVT cases, which inherently increased their risk of PE, possibly making treatment type more reflective of underlying severity rather than an independent predictor.

Additionally, the distribution of treatment types was highly imbalanced, with 81.7% of patients receiving AC-only, while thrombolytics (8.5%), MT (0.7%), CDT (1.2%), and multiple interventions (0.9%) were far less common. This skewed distribution may have limited the models' ability to detect meaningful patterns related to treatment type, as machine learning algorithms tend to prioritize features with more variation. Moreover, the timing of treatment, rather than the specific type, was identified as a key predictor in XGB and LR models, reinforcing the idea that delayed intervention (one to three days after DVT diagnosis) may have a greater impact on PE risk than the treatment modality itself. This finding suggests that early initiation of treatment, regardless of the specific intervention, may be more critical in preventing PE.

Overall, these results highlight an important distinction between associative trends observed in EDA and the predictive significance of features in machine learning models. While EDA findings suggested potential differences in PE risk among treatment groups, predictive modeling emphasized factors such as history of PE, number of DVT-related admissions, and treatment timing as more impactful.

6.2 Strengths & Contributions

This study makes several key contributions to PE risk prediction in DVT patients. First, it is among the few studies to compare the impact of multiple DVT treatment types—AC-only, thrombolytics, MT, and CDT—on PE risk using machine learning models rather than traditional statistical methods. Prior research has primarily focused on AC therapy alone, leaving gaps in understanding the long-term effects of interventional treatments on PE development (Douketis et al. 1998; Monreal et al. 2002; Patel et al. 2020). Additionally, while Dexter et al. (2022) analyzed MT's effectiveness for DVT treatment, their study focused on thrombus removal and post-thrombotic syndrome rather than PE incidence, highlighting the need for further research in this area.

Second, this study applied a comprehensive set of data balancing and model optimization techniques, including SMOTE, random undersampling, class weighting, hyperparameter tuning, and feature selection. The consistent performance improvements observed with these methods highlight their importance in addressing class imbalance, which is a common challenge in medical datasets with rare events like PE.

Lastly, this study identifies treatment timing (one to three days post-DVT diagnosis) as a key predictor of PE risk, underscoring the importance of early treatment timing. Feature importance analysis further highlights the number of DVT-related admissions and history of PE as strong risk factors. These findings highlight the critical role of early treatment and risk assessment in optimizing patient outcomes, supporting more personalized strategies to mitigate PE risk.

6.3 Limitations

Despite its contributions, this study has several limitations. The relatively small dataset size may affect the generalizability of the findings, particularly in detecting nuanced relationships between treatment type and PE risk. With only 4,410 patients and 217 PE events, the sample may not fully capture the complexity of PE risk factors across diverse populations. Expanding the dataset to include multi-institutional data could enhance model reliability and improve generalizability.

Class imbalance also posed challenges in model development, as PE cases accounted for only 4.9% of the dataset. This imbalance increased the risk of models favoring the majority class, potentially reducing sensitivity in detecting true PE cases. While techniques such as SMOTE, random undersampling, and class weighting were employed to address this issue, imbalanced data remains a limitation in rare event prediction and may have influenced model performance.

Another key limitation was the absence of Current Procedural Terminology (CPT) codes in the MIMIC-IV database, restricting access to detailed procedural data on mechanical thrombectomy (MT) and catheter-directed thrombolysis (CDT). Since these interventions are primarily documented through procedural coding rather than standard diagnosis or medication records, their effects on PE risk may have been underrepresented. The inclusion of CPT data would allow for a more comprehensive analysis of treatment pathways and provide better insight into how interventional procedures influence PE outcomes.

7 Conclusion

This study applied supervised machine learning to predict PE risk in DVT patients, identifying history of PE, number of DVT-related admissions, and treatment timing (one to three days post-diagnosis) as key predictors. Treatment type (AC, thrombolytics, MT, or CDT) was not a significant predictor, suggesting that delayed treatment initiation may pose a greater risk than treatment modality itself. The absence of procedural coding (CPT) data and treatment distribution imbalances may have influenced these findings, highlighting the need for larger, more representative datasets and enhanced feature availability in future research. Additionally, this study underscores the importance of addressing class imbalance through data resampling and model tuning to improve predictive performance.

By applying machine learning for PE risk prediction, this study underscores the critical role of timely DVT treatment and highlights the potential of predictive modeling to refine risk stratification, inform treatment decisions, and ultimately improve patient outcomes.

8 Directions for Future Work

Future research should expand dataset size and diversity to improve PE risk prediction across diverse patient populations. Incorporating multi-institutional data with a more balanced sample of MT and CDT patients, along with CPT codes, would enable a more comprehensive analysis of treatment effects. Additionally, an expanded dataset should better balance the target variable (PE vs. no PE), enhancing the robustness of predictive modeling and ensuring more representative risk estimates. Alternative methods for addressing class imbalance, such as cost-sensitive learning, should also be explored to enhance model sensitivity in detecting rare PE events.

Another important direction involves refining feature engineering to better capture disease severity and clinical decision-making. Integrating structured severity scores or physician treatment rationales could provide deeper insights into how patient characteristics influence treatment selection and PE risk. Furthermore, distinguishing between fatal and non-fatal PE outcomes could refine risk stratification and guide more targeted interventions. Finally, leveraging explainable artificial intelligence (AI) techniques, such as Shapley Additive Explanations (SHAP) analysis, could improve model interpretability, enabling more personalized risk assessment and informed treatment decisions for high-risk DVT patients.

9 Data Availability

The data used in this study was obtained from the MIMIC-IV v3.1 database, a publicly available electronic health record dataset maintained by the Massachusetts Institute of Technology (MIT) Laboratory for Computational Physiology. Access to MIMIC-IV requires completion of human subjects training and approval through PhysioNet. Due to data use agreement restrictions, the dataset cannot be shared directly but the database can be accessed by eligible researchers through the official MIMIC-IV portal.

10 Code Availability

The code for data extraction, preprocessing, feature engineering, and machine learning model development is available upon reasonable request. Due to compliance with PhysioNet's

Data Use Agreement, direct access to MIMIC-IV data is restricted. However, scripts for data processing and modeling, excluding raw patient data, can be shared upon request.

References

Baaten, Constance C.F.M.J., Sonja Vondenhoff, and Heidi Noels. 2023. "Endothelial Cell Dysfunction and Increased Cardiovascular Risk in Patients With Chronic Kidney Disease." *Circulation Research* 132, no. 8 (April): 970–92. https://doi.org/10.1161/circresaha.123.321752.

Caine, Graham J, Paul S Stonelake, Gregory Y H Lip, and Sean T Kehoe. 2002. "The Hypercoagulable State of Malignancy: Pathogenesis and Current Debate." *Neoplasia* 4, no. 6 (January 1): 465–73. https://doi.org/10.1038/sj.neo.7900263.

Charlson, Mary E., Peter Pompei, Kathy L. Ales, and C.Ronald MacKenzie. 1987. "A New Method of Classifying Prognostic Comorbidity in Longitudinal Studies: Development and Validation." *Journal of Chronic Diseases* 40, no. 5 (January): 373–83. https://doi.org/10.1016/0021-9681(87)90171-8.

Dexter, David J., Herman Kado, Jonathan Schor, Suman Annambhotla, Brandon Olivieri, Hamid Mojibian, Thomas S. Maldonado, et al. 2022. "Interim Outcomes of Mechanical Thrombectomy for Deep Vein Thrombosis from the All-Comer Clout Registry." *Journal of Vascular Surgery: Venous and Lymphatic Disorders* 10, no. 4 (July): 832-40. https://doi.org/10.1016/j.jvsv.2022.02.013.

Douketis, James D., Clever Kearon, Shannon Bates, et al. 1998. "Risk of Fatal Pulmonary Embolism in Patients with Treated Venous Thromboembolism." *JAMA* 279, no. 6 (February): 458-62. https://doi.org/10.1001/jama.279.6.458.

Drugs.com. n.d. "Thrombolytics." Accessed March 15, 2025. https://www.drugs.com/drug-class/thrombolytics.html?condition_id=583#filterSection.

Goldberger, Ary L., Luis A. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. 2000. "Physiobank, PhysioToolkit, and PhysioNet." *Circulation* 101, no. 23 (June): 1-6. https://doi.org/10.1161/01.cir.101.23.e215.

ICD9Data.com. n.d. "The Web's Free ICD-9-CM & ICD-10-CM Medical Coding Reference." Accessed March 15, 2025. http://www.icd9data.com/.

ICD10Data.com. n.d. "The Web's Free 2025 ICD-10-CM/PCS Medical Coding Reference." Accessed March 15, 2025. https://www.icd10data.com/.

Lin, John L. 2024. "Deep Venous Thrombosis (DVT) Treatment & Management." Medscape. Accessed February 12, 2025. https://emedicine.medscape.com/article/1911303-treatment.

Johnson, Alistair E, David J Stone, Leo A. Celi, and Tom J. Pollard. 2017. "The Mimic Code

Repository: Enabling Reproducibility in Critical Care Research.” *Journal of the American Medical Informatics Association* 25, no. 1 (September): 32–39. https://doi.org/10.1093/jamia/ocx084.

Johnson, Alistair E., Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, et al. 2023. “Mimic-IV, a Freely Accessible Electronic Health Record Dataset.” *Scientific Data* 10, no. 1 (January): 1-9. https://doi.org/10.1038/s41597-022-01899-x.

Johnson, Alistair E., Lucas Bulgarelli, Tom J. Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo A. Celi, et al. 2024. “MIMIC-IV.” PhysioNet. Accessed February 12, 2025. https://doi.org/10.13026/kpb9-mt58.

Johnson, Alistair E., Tom J. Pollard, A-Chahin, Jim Blundell, Brian Gow, Erinhong, Michael Schubert, et al. 2025. “MIT-Lcp/Mimic-Code: V2.5.0.” Zenodo. Accessed February 13, 2025. https://doi.org/10.5281/zenodo.821871.

Monreal, Manuel, Joan Ruíz, Angel Olazabal, Antoni Arias, and Josep Roca. 1992. “Deep Venous Thrombosis and the Risk of Pulmonary Embolism. A Systematic Study.” *CHEST Journal* 102, no. 3 (September): 677-81. https://doi.org/10.1378/chest.102.3.677.

Moriarty, John. 2015. "A New ICD-10-PCS Code For Extracorporeal Removal of Thrombi and Emboli from Venous System." Centers for Medicare & Medicaid Services. Accessed March 15, 2025. https://www.cms.gov/files/document/icd-10-pcstopicslides-8-12.pdf.

Mouawad, Nicolas. 2019. "ICD-10 Coding Change Request for Intravascular Ultrasound Assisted Thrombolysis in the treatment of Venous Thromboembolism and Peripheral Arterial Occlusion." Centers for Medicare & Medicaid Services. Accessed March 15, 2025. https://www.cms.gov/files/document/september-2019icd-10-pcs-topic-slides-4-7.pdf.

National Heart, Lung, and Blood Institute. 2022. “ Venous Thromboembolism Diagnosis.” Accessed February 13, 2025. https://www.nhlbi.nih.gov/health/venous-thromboembolism/diagnosis.

National Heart, Lung, and Blood Institute. 2022. “What Is Venous Thromboembolism?” Accessed February 12, 2025. https://www.nhlbi.nih.gov/health/venous-thromboembolism.

Patel, Payal, Parth Patel, Meha Bhatt, Cody Braun, Housne Begum, Robby Nieuwlaat, Rasha Khatib, et al. 2020. “Systematic Review and Meta-Analysis of Outcomes in Patients with Suspected Deep Vein Thrombosis.” *Blood Advances* 4, no. 12 (June): 2779–88. https://doi.org/10.1182/bloodadvances.2020001558.

Penumbra Inc. 2018. "Reimbursement Guide." Accessed March 15, 2025. https://www.penumbrainc.com/wp-content/uploads/2018/12/11629.F.pdf.

Rajagopal, Rajeev. 2024. “Deep Vein Thrombosis (DVT) - Diagnosis, Documentation and Coding.” Outsource Strategies International. Accessed March 15, 2025. https://www.outsourcestrategies.com/blog/deep-vein-thrombosis-diagnosis-documentation-and-coding/#:~:text=ICD%2010%20Codes%20for%20DVT,without%20complication%20(ACP%20Hospitalist).

Ryan, Logan, Jenish Maharjan, Samson Mataraso, Gina Barnes, Jana Hoffman, Qingqing Mao, Jacob Calvert, and Ritankar Das. 2022. “Predicting Pulmonary Embolism among Hospitalized Patients with Machine Learning Algorithms.” *Pulmonary Circulation* 12, no. 1 (January): 1-9. https://doi.org/10.1002/pul2.12013.

Shafi, Irfan, Brooke Zlotshewer, Matthew Zhao, Vladimir Lakhter, Behnood Bikdeli, Anthony Comerota, Huaqing Zhao, and Riyaz Bashir. 2023. “Association of Vena Cava Filters and Catheter-directed Thrombolysis for Deep Vein Thrombosis With Hospital Readmissions.” *Journal of Vascular Surgery Venous and Lymphatic Disorders* 12, no. 1 (September): 1-13. https://doi.org/10.1016/j.jvsv.2023.08.016.

U.S. Department of Labor. “Transition to ICD-10.” n.d. Accessed February 13, 2025. https://www.dol.gov/agencies/owcp/FECA/ICD10transition.

WebMD. 2024. “Thrombolytic Therapy for Deep Vein Thrombosis.” Accessed March 15, 2025. https://www.webmd.com/dvt/thrombolytic-therapy-dvt.

Zhang, Hao, Xiao-ye Li, Jia-si Li, Shi-bo Xia, Chao Song, Qing-sheng Lu, Wei Zhao, and Lei Zhang. 2024. “Which One Is the Best in Treating Deep Venous Thrombosis —— Percutaneous Mechanical Thrombectomy, Catheter-Directed Thrombolysis or Combination of Them?” *Journal of Cardiothoracic Surgery* 19, no. 1 (July): 1-8. https://doi.org/10.1186/s13019-024-02908-3.

Appendix A

Appendix A

ICD-9 & ICD-10 Codes for Deep Vein Thrombosis by Chronicity and Location

DVT Category	ICD-9 Code	ICD-10 Code	Description
Acute, Lower Extremity	451.2	I80.20%	Phlebitis and thrombophlebitis of unspecified deep vessels of lower extremities
	451.2	I80.29%	Phlebitis and thrombophlebitis of other deep vessels of lower extremity
	453.4	I82.40%	Unspecified deep vein of lower extremity
	453.4	I82.41%	Femoral vein
	453.4	I82.42%	Iliac vein
	453.4	I82.43%	Popliteal vein
	453.4	I82.44%	Tibial vein
	453.5	I82.45%	Peroneal vein
	453.5	I82.46%	Calf muscular vein
	453.4	I82.49%	Other specified deep vein of lower extremity
Chronic, Lower Extremity	453.5	I82.50%	Unspecified deep vein of lower extremity
	453.5	I82.51%	Femoral vein
	453.5	I82.52%	Iliac vein
	453.5	I82.53%	Popliteal vein
	453.5	I82.54%	Tibial vein
	453.5	I82.55%	Peroneal vein
	453.5	I82.56%	Calf muscular vein
	453.5	I82.59%	Other specified deep vein of lower extremity
Acute, Upper Extremity	451.8	I82.60%	Acute embolism and thrombosis of unspecified veins of upper extremity
	453.8	I82.62%	Deep veins of upper extremity
	453.8	I82.A1%	Axillary vein
	453.9	I82.B1%	Subclavian vein
	453.9	I82.C1%	Internal jugular vein
	453.9	I82.290%	Acute embolism and thrombosis of other thoracic veins
Chronic, Upper Extremity	453.7	I82.70%	Chronic embolism and thrombosis of veins of upper extremity
	453.7	I82.72%	Deep veins of upper extremity
	453.7	I82.A2%	Axillary vein
	453.8	I82.B2%	Subclavian vein
	453.8	I82.C2%	Internal jugular vein
	453.8	I82.291%	Chronic embolism and thrombosis of other thoracic veins
Acute, Unknown Location	453.9	I82.890%	Acute embolism and thrombosis of other specified veins
	453.9	I82.90%	Acute embolism and thrombosis of unspecified vein
Chronic, Unknown Location	453.8	I82.891%	Chronic embolism and thrombosis of other specified veins
	453.8	I82.91%	Chronic embolism and thrombosis of unspecified vein
Other	453.2	I82.22%	Embolism and thrombosis of inferior vena cava

Appendix B

Appendix B

ICD-9 & ICD-10 Codes for Deep Vein Thrombosis, Pulmonary Embolism, and Related Conditions

Condition	**ICD-9 Code**	**ICD-10 Code**
DVT	451.19, 451.19, 453.40, 453.41, 453.41, 453.41, 453.42, 453.52, 453.52, 453.42, 453.50, 453.51, 453.51, 453.51, 453.52, 453.52, 453.52, 453.52, 451.83, 453.82, 453.84, 453.85, 453.86, 453.87, 453.73, 453.72, 453.74, 453.75, 453.76, 453.77, 453.89, 453.89, 453.79, 453.79, 453.2	I80.20%, I80.29%, I82.40%, I82.41%, I82.42%, I82.43%, I82.44%, I82.45%, I82.46%, I82.49%, I82.50%, I82.51%, I82.52%, I82.53%, I82.54%, I82.55%, I82.56%, I82.59%, I82.60%, I82.62%, I82.A1%, I82.B1%, I82.C1%, I82.290%, I82.70%, I82.72%, I82.A2%, I82.B2%, I82.C2%, I82.291%, I82.890%, I82.90%, I82.891%, I82.91%, I82.22%
PE	415.11, 415.13, 415.19, 416.2	I26%
History of DVT	V12.51	Z86.718
History of PE	V12.55	Z86.711
Long-Term AC Use	V58.61	Z79.01

Appendix C

Appendix C

Treatment Medications and ICD-9/ICD-10 Procedure Codes for Deep Vein Thrombosis

Treatment Type	Medication Names
Anticoagulation	Heparin, Coumadin, Warfarin, Enoxaparin (Lovenox), Dalteparin, Tinzaparin, Fondaparinux, Argatroban, Bivalirudin, Desirudin, Rivaroxaban (Xarelto), Apixaban (Eliquis), Dabigatran (Pradaxa), Edoxaban
Systemic Thrombolytics	Alteplase, Tenecteplase, Reteplase, Urokinase

Treatment Type	ICD-9 Code	ICD-10 Code
Mechanical Thrombectomy	39.79	05C53ZZ, 05C63ZZ, 05C73ZZ, 05C83ZZ, 05C93ZZ, 05CA3ZZ, 05CB3ZZ, 05CC3ZZ, 05CD3ZZ, 05CF3ZZ, 06C93ZZ, 06CB3ZZ, 06CC3ZZ, 06CD3ZZ, 06CF3ZZ, 06CG3ZZ, 06CH3ZZ, 06CJ3ZZ, 06CM3ZZ, 06CN3ZZ, 06CP3ZZ, 06CQ3ZZ, 06CR3ZZ, 06CS3ZZ, 06CT3ZZ, 06CV3ZZ, 06CY3ZZ
Catheter-Directed Thrombolyss	99.1	6A75, 6A750, 6A750Z, 6A751, 6A751Z, 3E03017, 3E03317, 3E04017, 3E04317

Appendix D

Appendix D

Engineered Features Using Google BigQuery

Derived Feature (BigQuery)	Source Columns	Description
first_dvt_flag	admissions.admittime, diagnoses_icd.icd_code	Indicates if the patient's DVT admission is their first
had_dvt_as_pri_diagnosis	diagnoses_icd.seq_num	Binary flag indicating whether DVT was the primary diagnosis during the hospital admission
pe_outcome	diagnoses_icd.icd_code	Binary indicator for PE occurrence post-DVT
had_icu_stay	icu_stays.subject_id	Binary indicator for whether a patient had an ICU stay
num_dvt_admissions	admissions.hadm_id	Count of DVT-related hospital admissions per patient
num_dvt_diagnoses	diagnoses_icd.icd_code	Number of DVT-related diagnosis codes recorded
ac_flag	prescriptions.drug	Binary flag indicating anticoagulation prescription
lytics_flag	prescriptions.drug	Binary flag indicating thrombolytic prescription
mt_flag	procedures_icd.icd_code	Binary flag indicating mechanical thrombectomy procedure
us_cdt_flag	procedures_icd.icd_code	Binary flag indicating catheter-directed thrombolysis (CDT)
length_of_stay	admissions.admittime, admissions.dischtime	Number of days between hospital admission and discharge
charlson_comorbidity_ index (CCI)	diagnoses_icd.icd_code, patients.age	Comorbidity score calculated using ICD codes
days_to_ac	prescriptions.starttime, admissions.admittime	Days from admission to first anticoagulation prescription
days_to_lytics	prescriptions.starttime, admissions.admittime	Days from admission to first thrombolytics prescription
days_to_mt	procedures_icd.chartdate, admissions.admittime	Days from admission to mechanical thrombectomy procedure
days_to_cdt	procedures_icd.chartdate, admissions.admittime	Days from admission to catheter-directed thrombolysis (CDT)
hx_dvt	diagnoses_icd.icd_code (Z86.718, V12.51)	Binary indicator for history of DVT
hx_pe	diagnoses_icd.icd_code (Z86.711, V12.55)	Binary indicator for history of PE
hx_ac	diagnoses_icd.icd_code (Z79.01, V58.61)	Binary indicator for history of long-term anticoagulation use
days_to_pe	admissions.admittime, diagnoses_icd.icd_code	Days from first DVT diagnosis to PE occurrence
dvt_chronicity	diagnoses_icd.icd_code	Categorization of DVT cases as Acute, Chronic, or Unspecified based on ICD codes
dvt_location	diagnoses_icd.icd_code	Categorization of DVT by anatomical location (e.g., Lower, Upper, or Unspecified based on ICD codes

Appendix E

Appendix E

Engineered Features Using Python

Derived Feature (Python)	Source Columns	Description
num_pe_events	subject_id, pe_outcome	Number of PE occurrences post-DVT per patient
days_to_init_treatment	days_to_ac, days_to_lytics, days_to_mt, days_to_cdt	Number of days from initial DVT admission to first recorded treatment (AC, thrombolytics, MT, or CDT)
cat_days_to_init_treatment	days_to_init_treatment	Categorical version of days_to_init_treatment, grouping patients into bins (e.g., Same Day, 1-3 days, 4-7 days, >7 days, No Treatment, Unknown).
cat_days_to_pe	days_to_pe	Categorical version of days_to_pe, grouping time to PE occurrence into clinically meaningful intervals (e.g., Within 1 month, 1-3 months, 3-6 months, 6-12 months, More than 1 year, No PE)
log_length_of_stay	length_of_stay	Log-transformed hospital length of stay to reduce skewness (log(1 + length_of_stay))
log_num_dvt_admissions	num_dvt_admissions	Log-transformed count of hospital admissions related to DVT for each patient (log(1 + num_dvt_admissions))
log_num_dvt_diagnoses	num_dvt_diagnoses	Log-transformed count of DVT-related diagnosis codes recorded (log(1 + num_dvt_diagnoses))
treatment	ac_flag, lytics_flag, mt_flag, us_cdt_flag	Consolidated treatment category based on intervention type (e.g., AC Only, MT, CDT, Lytics, Multiple Interventions, No Treatment)
race_grouped	race	Consolidated race categories into standardized groups (e.g., Black, White, Asian, Hispanic/Latino, Portuguese, American Indian/Alaska Native, Native Hawaiian/Pacific Islander, Multiracial, Unknown)
admission_location_grouped	admission_location	Consolidated admission locations into standardized groups (e.g., Emergency/Urgent Care, Referral-Based Admissions, Transfer from Another Facility, Scheduled/Procedure-Based Admissions, and Unknown
discharge_location_grouped	discharge_location	Consolidated discharge locations into standardized groups (e.g., Home/Community-Based Care, Facility-Based Care, Against Medical Advice, and Unknown