

Introduction

Venous Thromboembolism (VTE), encompassing deep vein thrombosis (DVT) and pulmonary embolism (PE) is a very severe condition that affects 1-2 people per 1000 in the United States alone.¹ VTE is often underdiagnosed especially in African American populations even though they have higher incidence of VTE.^{2,3} Machine Learning has been widely used for VTE risk prediction.⁴ Studies taken from a systematic review by Gu et al. (2025) have proven to be successful with some being more accurate than traditional VTE risk tools.⁴ However, a disparity exists in that many AI risk tools have not been developed using Black populations.^{2,5}

Objectives

The main objective of this study was to develop a Machine Learning Model used for early risk prediction of Venous Thromboembolism in African American patients. The gap of knowledge of this study is that very few VTE predictive models have included primarily Black patients. The hypothesis of this study is that this ML model will accurately predict VTE in Black patients. The rationale is that machine learning can model complex relationships to enable timely clinical decision-making, optimize resource management, and reduce racial health disparities.

Methods

This study utilized electronic health records data from the MIMIC-4 emergency department data subset and used Python for all statistical analysis. The electronic health records data contained emergency department encounters from Beth-Israel Deaconess Medical Center. A total of 92,998 encounters were analyzed for this study. This dataset included only African Americans reporting to the emergency department presenting with VTE or without VTE. A descriptive analysis was performed on the data. Feature selection was supported by logistic GEE, FDR correction, and Spearman correlation. Patients were separated into training, validation, and testing sets. Logistic regression was used for the supervised classification model to predict each patient's chance of VTE. The main metric used for the final model evaluation was AUROC.

Results

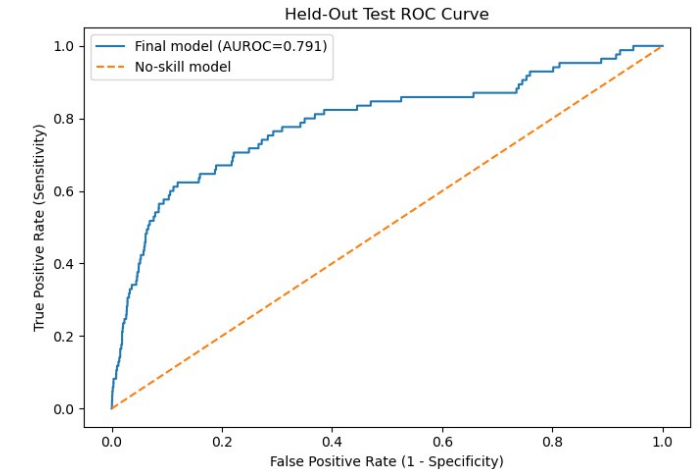
1.

Characteristic	Overall (n=92,998)	VTE (n=543)	No VTE (n=92,455)
Unweighted ED encounters, n	92,998	543	92,455
Unique patients, n	33,016	469	32,900
DEMOGRAPHICS			
Female sex, n (%)	56,103 (60.3%)	330 (60.8%)	55,773 (60.3%)
INITIAL TRIAGE VITAL SIGNS			
Temperature, °F, mean (SD)	98.1 (1.0)	98.1 (0.8)	98.1 (1.0)
Heart rate, beats/min, mean (SD)	84.7 (16.8)	88.2 (18.3)	84.6 (16.7)
Respiratory rate, breaths/min, mean (SD)	17.6 (2.3)	17.9 (2.6)	17.6 (2.3)
Oxygen saturation, %, mean (SD)	98.8 (1.9)	98.4 (2.4)	98.8 (1.9)
Systolic blood pressure, mm Hg, mean (SD)	137.8 (22.9)	138.4 (21.2)	137.8 (23.0)
Diastolic blood pressure, mm Hg, mean (SD)	79.6 (15.0)	80.3 (13.9)	79.6 (15.0)
Pain score, mean (SD)	4.8 (3.9)	4.3 (3.7)	4.8 (3.9)
DERIVED PHYSIOLOGIC MEASURES			
Pulse pressure, mm Hg, mean (SD)	58.2 (19.8)	58.1 (18.7)	58.2 (19.8)
Mean arterial pressure, mm Hg, mean (SD)	99.0 (15.4)	99.7 (14.2)	99.0 (15.4)
Shock index, mean (SD)	0.63 (0.16)	0.65 (0.17)	0.63 (0.16)
Modified shock index, mean (SD)	0.87 (0.22)	0.90 (0.22)	0.87 (0.22)
Tachycardia (heart rate ≥100/min), n (%)	16,783 (18.5%)	135 (25.5%)	16,648 (18.5%)
Tachypnea (respiratory rate ≥20/min), n (%)	13,577 (15.1%)	114 (21.6%)	13,463 (15.1%)
Low oxygen saturation (<95%), n (%)	2,432 (2.7%)	30 (5.7%)	2,402 (2.7%)
Hypotension (SBP <90 mm Hg), n (%)	510 (0.6%)	2 (0.4%)	508 (0.6%)
CHIEF COMPLAINT FEATURES			
Chest pain, n (%)	7,761 (8.3%)	77 (14.2%)	7,684 (8.3%)
Dyspnea / shortness of breath, n (%)	5,747 (6.2%)	107 (19.7%)	5,640 (6.1%)
Leg pain or swelling, n (%)	2,826 (3.0%)	137 (25.2%)	2,689 (2.9%)
Syncope, n (%)	1,406 (1.5%)	6 (1.1%)	1,400 (1.5%)
VTE OUTCOME COMPONENTS			
Deep vein thrombosis, n (%)	290 (0.3%)	290 (53.4%)	0 (0.0%)
Pulmonary embolism, n (%)	266 (0.3%)	266 (49.0%)	0 (0.0%)

2.

Predictor	Adjusted OR	95% CI lower	95% CI upper	P value
Ambulance arrival (vs walk-in)	0.971	0.8	1.179	0.7668
Helicopter arrival (vs walk-in)	0	0	0	
Other arrival mode (vs walk-in)	0			
Unknown arrival mode (vs walk-in)	0.462	0.202	1.053	0.0663
Triage acuity level 1 (vs level 3)	0.896	0.44	1.825	0.7616
Triage acuity level 2 (vs level 3)	1.874	1.542	2.277	0
Triage acuity level 4 (vs level 3)	0.46	0.283	0.748	0.0017
Triage acuity level 5 (vs level 3)	1.036	0.168	6.37	0.9695
Female sex	1.035	0.837	1.28	0.7506
Temperature, per 1°F increase	1.017	0.942	1.098	0.6606
Heart rate, per 10 beats/min increase	1.105	1.043	1.17	0.0007
Respiratory rate, per 5 breaths/min inc	0.898	0.738	1.092	0.2808
Oxygen saturation, per 5 percentage-pt	0.8	0.649	0.987	0.037
Systolic blood pressure, per 10 mm H	0.984	0.943	1.027	0.4514
Diastolic blood pressure, per 10 mm H	1.03	0.964	1.1	0.3826
Pain score, per 1-point increase	0.966	0.943	0.99	0.006
Chief complaint: chest pain	1.683	1.299	2.182	0.0001
Chief complaint: dyspnea	2.408	1.816	3.193	0
Chief complaint: leg pain or swelling	12.505	9.97	15.684	0
Chief complaint: syncope	0.77	0.328	1.806	0.5473

3.



Discussion

This study used the MIMIC-4 dataset to model risk prediction for VTE. The final test metrics produced an AUROC of 0.79102. This shows that the model demonstrated moderate to good at discriminating VTE. Limitations of this study include: low prevalence of VTE in the data and not having age as an available predictor. Because this study intentionally focused on data from African Americans to develop the model, there should be caution when applying these results to other racial or ethnic groups. Before implementation of this model it would be important to train the model on more diverse data as the data used came from one hospital. Future work should validate this model on geographically and demographically diverse data.

References



Figure 1. Results of the descriptive analysis conducted on the MIMIC-IV dataset

Figure 2. Results of the multivariable GEE logistic regression including: adjusted odds ratio, 95% confidence interval upper and lower, and p values

Figure 3. ROC curve & AUC value