AI-Powered Early Detection of Embolism Using Multimodal Clinical Data

ABSTRACT

Early detection of pulmonary embolism (PE) and deep vein thrombosis (DVT) remains a critical challenge in healthcare, with delayed recognition contributing to significant morbidity and mortality. This study explores the development of a multimodal deep learning system that integrates vital signs, laboratory values, clinical notes, and imaging reports to predict embolism risk 8-24 hours before clinical manifestation. Using data from the MIMIC-IV database, we implemented a novel architecture combining LSTM networks for temporal data, BERT embeddings for clinical text, and attention mechanisms for feature importance. Despite promising initial results on retrospective data (AUROC 0.82), our model struggled with prospective validation and clinical integration. This paper details our approach, challenges faced, and lessons learned in developing explainable AI for time-sensitive clinical decision support.

CCS CONCEPTS

- Computing methodologies → Machine learning; • Applied computing → Health informatics

KEYWORDS

pulmonary embolism, deep vein thrombosis, multimodal learning, clinical decision support, explainable AI, electronic health records

1 INTRODUCTION

Thromboembolic diseases, including pulmonary embolism (PE) and deep vein thrombosis (DVT), represent significant causes of preventable hospital mortality, with PE alone accounting for approximately 100,000 deaths annually in the United States [1]. The timely identification of embolic events is crucial, as prompt diagnosis and treatment can significantly reduce mortality rates, yet current diagnostic approaches often rely on overt clinical manifestations or physician suspicion [2].

The abundance of data in electronic health records (EHRs) presents an opportunity to leverage artificial intelligence for earlier detection of thromboembolic risk. Previous approaches have largely focused on individual data streams (e.g., vital signs or laboratory values) without integrating the rich information contained in clinical notes, medication data, and patient context [3, 4]. Moreover, the "black box" nature of many machine learning models has limited their adoption in clinical settings where interpretability is crucial [5].

This project aimed to develop and evaluate a multimodal deep learning system capable of predicting embolism risk 8-24 hours before clinical criteria are met, while providing explainable predictions to support clinical decision-making. We hypothesized that combining structured data (vitals, labs) with unstructured text (clinical notes, radiology reports) would yield superior performance compared to unimodal approaches, and that attention mechanisms could provide clinically relevant explanations for model predictions.

2 METHODOLOGY

2.1 Data Source and Preprocessing

We utilized the MIMIC-IV database, containing de-identified health data from over 40,000 patients admitted to Beth Israel Deaconess Medical Center between 2008 and 2019 [6]. Embolism cases were identified using ICD codes for PE and DVT, validated by manual review of radiology reports confirming these diagnoses.

For each patient, we extracted:

- Vital signs: heart rate, respiratory rate, blood pressure, oxygen saturation, temperature
- Laboratory values: D-dimer, troponin, BNP, complete blood count, basic metabolic panel
- Demographic information: age, gender, BMI, comorbidities
- Clinical notes: nursing notes, physician notes, and radiology reports
- Medication data: anticoagulants, contraceptives, immobility status

Data preprocessing included handling missing values, normalizing numerical features, converting clinical notes to BERT embeddings using ClinicalBERT [7], and creating temporal windows of 48 hours with 4-hour increments.

2.2 Model Architecture

Our multimodal architecture comprised three main components:

1. A bidirectional LSTM network processed temporal vital signs and laboratory values, capturing trends and patterns over time.
2. A BERT-based encoder extracted semantic features from clinical notes and radiology reports, identifying relevant clinical observations and concerns documented by healthcare providers.
3. A multilayer perceptron handled demographic and static features, integrating patient risk factors into the prediction model.

These components were integrated through a cross-attention mechanism that allowed the model to identify relationships between different data modalities (e.g., how abnormal vital signs correlated with

specific mentions in clinical notes). The final classification layer outputted the probability of embolism within the next 8-24 hours.

2.3 Explainability Approach

To address the black box problem, we implemented three complementary approaches:

1. SHAP (SHapley Additive exPlanations) values quantified feature importance for individual predictions, highlighting which data elements contributed most to high-risk assessments.
2. Attention visualization highlighted relevant sections of clinical notes, allowing clinicians to quickly locate the textual evidence supporting the model's predictions.
3. A clinical explanation generator translated model outputs into natural language summaries, providing context-aware interpretations of risk predictions.

2.4 Evaluation Protocol

We employed a rigorous evaluation approach with a temporal train-test split (2008-2016 for training, 2017-2019 for testing) to assess real-world generalizability. Performance was measured using AUROC, AUPRC, sensitivity, specificity, and timeliness (hours before clinical diagnosis). We compared our model against baseline approaches including Wells criteria, PERC rule, and a random forest model trained on structured data only. Additionally, emergency medicine and critical care physicians provided qualitative evaluation through case studies.

3 RESULTS

3.1 Prediction Performance

Our model achieved an AUROC of 0.82 (95% CI: 0.79-0.85) and AUPRC of 0.46 (95% CI: 0.43-0.49) for predicting embolism 8-24 hours before clinical diagnosis. Table 1 compares performance against baseline models.

Table 1: Performance comparison across models

Model	AUROC	AUPRC	Sensitivity	Specificity
Our approach	0.82	0.46	0.77	0.78
Random Forest	0.76	0.41	0.70	0.74
Wells criteria	0.67	0.32	0.62	0.69
PERC rule	0.63	0.28	0.54	0.74

3.2 Feature Importance

SHAP analysis revealed that the most predictive features were (in descending order):

1. Elevated D-dimer levels
2. Tachycardia (heart rate >100 bpm)
3. Mentions of "leg swelling" or "chest pain" in clinical notes
4. Hypoxemia (O2 saturation <94%)
5. Recent surgery or immobilization

The integration of clinical notes proved particularly valuable, as the model identified subtle linguistic patterns associated with embolism risk that would not be captured in structured data alone. For example, nursing documentation of mild leg swelling or non-specific chest discomfort often preceded formal diagnosis by 12-18 hours.

3.3 Challenges and Limitations

Despite promising initial results, we encountered significant challenges:

1. Dataset shift: Performance degraded when applied to newer data, suggesting temporal drift in diagnostic practices and documentation patterns.
2. Alert fatigue: Initial implementation generated excessive false positives, leading to physician desensitization to alerts.
3. Integration barriers: Workflow integration proved more complex than anticipated, with resistance to adopting AI-based decision support into established clinical pathways.
4. Explanation quality: Clinicians found some explanations counterintuitive or lacking sufficient context, highlighting the gap between statistical associations and clinical reasoning.
5. Rare event prediction: The relatively low incidence of PE/DVT created class imbalance challenges, requiring sophisticated sampling approaches and loss functions.

4 CONCLUSION

This project explored the development of a multimodal approach to early embolism detection using deep learning. While our model demonstrated improved performance compared to clinical scoring systems, the challenges encountered highlight the complexity of translating algorithmic advances into clinical practice.

Key lessons learned include:

1. The importance of continuous model updating to address dataset shift
2. The need for balancing sensitivity and specificity to avoid alert fatigue

3. The value of interdisciplinary collaboration throughout model development
4. The critical role of explainability in fostering clinical trust
5. The difficulty of predicting relatively rare events despite large datasets

Future work should focus on prospective validation, improving explanation quality, and developing more robust approaches to temporal drift. Despite not achieving all our initial objectives, this high-risk project provided valuable insights into the practical challenges of deploying AI for thromboembolic risk prediction in clinical settings.

REFERENCES

[1] Raskob, G.E., et al. (2014). Thrombosis: a major contributor to global disease burden. Arteriosclerosis, thrombosis, and vascular biology, 34(11), 2363-2371.

[2] Kline, J.A. (2018). Diagnosis and exclusion of pulmonary embolism. Thrombosis Research, 163, 207-220.

[3] Fernandes, M., et al. (2020). Risk of mortality and pulmonary embolism in patients with COVID-19 pneumonia: A prediction model based on machine learning. PLOS ONE, 15(12), e0244053.

[4] Mezzatesta, S., et al. (2019). A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis. Computer Methods and Programs in Biomedicine, 177, 9-15.

[5] Ghassemi, M., et al. (2021). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 3(11), e745-e750.

[6] Johnson, A., et al. (2021). MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data, 8(1), 1.

[7] Alsentzer, E., et al. (2019). Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323.