STEVENS
INSTITUTE OF TECHNOLOGY
1870

Ph.D. DISSERTATION DEFENSE

Candidate:	Chang Lu
Degree:	Doctor of Philosophy
School/Department:	School of Engineering and Science
Date:	Tuesday, February 27th, 2024
Time/Location:	1:00 pm / https://stevens.zoom.us/j/92514619334
Title:	Knowledge-Guided Deep Learning for Health Data Mining and Augmentation
Chairperson:	Dr. Yue Ning, Department of Computer Science, Stevens Institute of Technology
Committee Members:	Dr. Samantha Kleinberg, Department of Computer Science, Stevens Institute of Technology
	Dr. Tian Han, Department of Computer Science, Stevens Institute of Technology
	Dr. Yu Gan, Department of Biomedical Engineering, Stevens Institute of Technology
	Dr. Chandan Reddy, Department of Computer Science, Virginia Tech

ABSTRACT

The widespread application of Electronic Health Records (EHR) in healthcare facilities has sparked a growing interest in employing deep learning for health event prediction. Health event prediction tasks include two primary categories: 1) temporal event prediction, such as forecasting future diagnoses based on historical records; and 2) static event prediction, such as disease coding from clinical notes. Despite notable accomplishments in these domains, several challenges still exist. Existing deep learning models often lack a comprehensive exploration of medical domain knowledge, disease correlations, and disease development schemes, leading to unsatisfying prediction accuracy and the absence of explainability. Moreover, constrained accessibility to EHR data and imbalanced disease distributions in real-world records limit the application of large-scale deep learning in health event prediction.

In this dissertation, we introduce novel methods to address these challenges. For temporal prediction, we propose a collaborative graph learning method to enhance the integration of medical domain knowledge in graph neural networks based on the International Classification of Diseases (ICD). Next, we introduce a context-aware health event prediction framework via transition functions on dynamic disease graphs, to analyze disease correlations and development.

To tackle the issue of insufficient EHR data, we propose several pre-training and data generation methods. We design a self-supervised learning task to fully leverage EHR data. It can incorporate single-visit data, which are often neglected in existing work. Additionally, we design a multi-label time-series generative adversarial network to generate high-quality synthetic EHR data, specifically addressing imbalanced data for uncommon diseases. Finally, for static prediction, we focus on ICD coding that extracts disease from clinical notes and propose to automatically segment clinical notes into sections. We also employ a contrastive learning approach to pre-train existing models and boost their prediction accuracy.

In terms of evaluation, we conduct extensive experiments on real-world EHR datasets. The results verify the effectiveness of our proposed methods in achieving state-of-the-art performance for temporal event prediction. Finally, our innovative pre-training and data generation methods demonstrate their capability to alleviate issues related to insufficient and imbalanced data in both temporal and static prediction tasks.