Graph Representation Learning for Medical Event Recurrence Prediction*

Nicholas Glaze
Department of Electrical and Computer Engineering
Rice University
Houston, Texas, USA
nglaze00@gmail.com

Artun Bayer
Department of Electrical and Computer Engineering
Rice University
Houston, Texas, USA
email address or ORCID

Xiaoqian Jiang
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

other UTH mentors
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

Santiago Segarra
Department of Electrical and Computer Engineering
Rice University
Houston, Texas, USA
email address or ORCID

6th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—We introduce a graph representation learning architecture to predict recurrence of medical events. We demonstrate its effectiveness for predicting recurrent strokes using electronic health records (EHRs), which are common, have high mortality rates, and can be mitigated with preventative care**

***Index Terms*—TODO**

I. Introduction

1 pg

A. Recurrent strokes

- Strokes are very common, and those who suffer them can permanently lose many of the skills necessary for daily life. Also, strokes account for 10% of deaths worldwide.
- Stroke recurrence increases mortality rate 17x, so understanding how to predict and mitigate recurrent strokes is therefore incredibly important to reducing patients' rates of mortality, disability, and financial hardship.

B. Related work

- Machine learning has been applied to predict recurrent strokes using electronic health records (describe here), but hasn't given great results (do we need to describe other recent research, if we'll only compare with basic benchmarks?)
- Graph representation learning (in particular, the HORDE framework) has been successfully applied to other downstream medical tasks using EHRs, so it's a good candidate to use here.

Identify applicable funding agency here. (do UTHealth or D2K go here?)

C. Overview of the paper

Other papers usually have a summary at the end of the intro, so these are the items we might include in that summary (would probably pick a few)

- The problem we're solving
- The HORDE graph framework
- The architecture we introduce
- The stroke dataset we evaluate it on, plus experiment results + our analysis
- Conclusion, which includes practical stroke-related benefits for this particular model

II. Problem Definition

0.5 pg We leverage data observed about a patient during medical appointments to predict whether a medical event they previously experienced will recur. Given a matrix $\mathbf{X} \in \mathbb{R}^{T \times N}$ describing N features of a patient's condition at each of T timesteps, and a vector $\mathbf{d} \in \mathbb{R}^{M}$ describing M time-invariant characteristics of the patient, we propose an architecture that learns a parametrized function $f_\theta : (\mathbf{X}, \mathbf{d}) \rightarrow [0, 1]$, with parameters θ, that computes the probability that the event will recur for the patient within an m-month window.

Additional things:

- Timesteps can be irregular, both inter- and intra-patient (do we need to justify that we ignore this irregularity? or can we just say that it works anyway)
- Not really sure how people generally format mathematical descriptions of ML problems, would like some advice there – I feel like I'm probably missing stuff

III. Proposed framework

1-1.5 pg, leaning towards 1?

- We propose a graph representation learning framework for recurrent event prediction based on time-varying medical data.
- The framework has two parts: first, we construct the HORDE graph out of the available data. Second, we apply our new architecture, (Name here)

A. HORDE framework

- We use the HORDE framework, introduced by (reference), to represent the time-varying matrix X as a heterogeneous network.
- the HORDE graph is constructed with both patients and features as nodes, drawing edges, weighted by feature value, between patient and feature nodes at timesteps where that patient has a non-zero value for that feature
- Why is this good? Efficient (EHRs are sparse), allows application of GCNs, supports heterogeneous feature types + including relationships (edges) between features

B. Our architecture

- We propose a new architecture, inspired by the one used in the HORDE framework, to predict event recurrence.
- Transform each timestep graph using standard GCN layer (mathematically describe, but not that much since people know them now). This gives us a vector representing each patient at each timestep (we take only the patient nodes)
- Should we describe architecture variations (Laplacian, HetGCN) here or in experiments section?
- Stack each patient's vectors into a time series, then apply an LSTM layer (mathematically describe, say why it's good here) This gives us a single vector per patient
- Concat the vector with the demographics vector, then put in dense layer to get final probability
- do we need to discuss why we chose this architecture, beyond "gcns are good for graphs and lstms are good for time series"?

IV. Experiments

2 pg

2 experiments: 1. overall comparison, ours vs. baselines vs. variations of ours, 2. stroke type breakdown

A. Recurrent stroke dataset

In working with the University of Texas Health Science Center at Houston (UTHealth) for my senior capstone, I have been granted access to a labeled database of EHRs, over up to 69 encounters, for 4,776 UTHealth hospital patients—one of the most comprehensive in the world of its kind. Each patient in the database suffered at least one stroke, and approximately 25% of these patients suffered recurrent strokes within one year as well. The EHRs are represented in tabular form, and consist of approximately 5,000 features describing patients' demographics, along with their diagnoses, lab tests, measurements taken, and other information throughout their medical appointments. The raw text of approximately 2 million clinical notes is also included in the dataset, along around 8,000 CT and MRI scans; however, this project leverages only the EHRs, and the other modalities will be incorporated in future work.

- UTH stroke patient dataset 2019-2020
- Size of dataset, distribution of stroke recurrence window lengths
- EHR description; breakdown by category?
- Could mention image / text data, but it's not related to this project

B. Experiment parameters

- Describe hyperparameters common for all experiments
- Describe baselines + variations we compare to
- Describe train / test splits (CV) and evaluation metrics

C. Overall model performance comparison

TABLE I
PERFORMANCE ON 1-YEAR RECURRENCE WINDOW.

Architecture	Metric			
	Specificity	***Sensitivity***	***AUROC***	***AUPRC***
Logistic Regression	0.	0.	0.	0.
SVM	0.	0.	0.	0.
Gradient-boosted trees	0.	0.	0.	0.
More	0.	0.	0.	0.
Benchmarks	0.	0.	0.	0.
Here	0.	0.	0.	0.
Ours	0.	0.	0.	0.
Ours$_L$	0.	0.	0.	0.
Ours$_H$	0.	0.	0.	0.
Ours$_{L+H}$	0.	0.	0.	0.

Analysis of results, which will hopefully be that ours is the best!

Is it interesting to include architectures like (first few timesteps only), (more GCN layers), etc.?

D. Performance by stroke type

Remove spec / sens? they're interesting though

TABLE II
PERFORMANCE ON 1-YEAR RECURRENCE WINDOW.

Initial stroke type	Test dataset		Metric			
	Samples	***Pos. %***	***Specificity***	***Sensitivity***	***AUROC***	***AUP***
All	0.	0.	0.	0.	0.	0.
AIS	0.	0.	0.	0.	0.	0.
ICH	0.	0.	0.	0.	0.	0.
SAH	0.	0.	0.	0.	0.	0.
TIA	0.	0.	0.	0.	0.	0.
Multi-type	0.	0.	0.	0.	0.	0.
A few unions	0.	0.	0.	0.	0.	0.

- Detailed breakdown for best model
- Comparison of the top few models on the most interesting stroke type subset

- Practical applications; why is this experiment important?
- how much to emphasize that we are working with UTHealth and our results will practically benefit them? I imagine that could be cool, if EUSIPCO people care about that kind of thing

V. Conclusion

- We introduced a medical event recurrence prediction architecture, and demonstrated its effectiveness for the specific task of recurrent stroke.
- Our model can enable more efficient application of preventative care, allowing hospitals to give necessary care to more patients, saving more lives.
- Future work includes incorporating additional data modalities, which our model supports.
- Anything else?

Acknowledgment

Capstone team members + anyone else at UTH who helped

References

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.