Mining for Health: A Comparison of Word Embedding Methods for EHRs

Emily Getzen[1] , Yucheng Ruan[2], Qi Long[1]

[1] Department of Biostatistics ,Epidemiology, and Informatics. [2] Department of Engineering, University of Pennsylvania, USA

Background

Electronic health records (EHRs) offer great promises for advancing precision health, but present significant analytical challenges– EHRs contain data from multiple domains that can be structured or unstructured, and collected at irregular time intervals and frequencies. To harness the power of EHRs, one powerful tool is word embedding algorithms, which take a corpus of text and generate vector representations (embeddings) of individual words that capture word relationships as well as semantic and syntactic similarities. By considering a single medical event as a "word" and a sequence of medical events as a"corpus", the same method can be used for structured events of EHRs. This can be viewed as automated feature extraction. While currently there exists a wide variety of embedding tools, there has been little to no work on comparing their performance for analysis of EHRs data. We extend these methods to embed a patient's entire medical history, and use the resultant embeddings to build prediction models for medical events. We assess performance of multiple state-of-the-art word embedding methods in terms of predictive accuracy and computation time using the Medical Information Mart for Intensive Care (MIMIC) database.

Methods

Embedding Algorithms

Table 1: Comparison of Word Embedding Algorithms

Word Embedding Algorithm	Description
Word2Vec	Shallow neural network with two training mechanisms: continuous bag of words (CBOW) and skip-gram.
FastText	An extension of Word2Vec that represents each word as an n-gram of characters
GloVe	Builds a co-occurrence matrix which counts how frequently two words appear together. The cosine distance between two embeddings = the log probability of their co-occurrence.
ELMo	Uses a deep bidirectional LSTM model to create word representations.
BERT	Deep neural network based on a transformer architecture that trains by masking and predicting a percentage of words.

Word2Vec, FastText, and GloVe can only generate static embeddings (a single representation for a unique word) while ELMo and BERT can also incorporate contextual information to generate different vector representations for a word depending on its meaning (this could be useful since some ICD-9 diagnosis codes are ambiguous). BERT can also embed whole sentences rather than just words

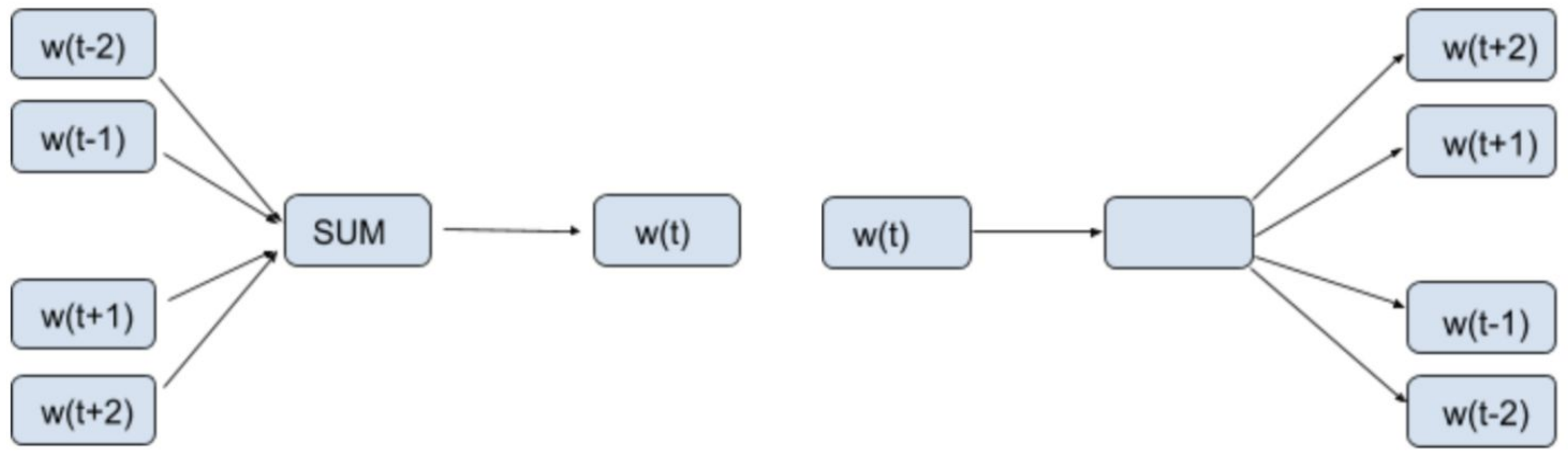

Figure 1. a) CBOW training mechanism b) Skip-Gram training mechanism

Embedding of EHRs

We used the approach from Farhan et al. (2016) to generate vector representations of patient medical histories. "History" is defined as events prior to the most recent visit to the ICU. The approach involves multiplying each event's embedding in the history by a temporal factor (to give more importance to more recent events) and then summing across the adjusted embeddings. For BERT whole sentence embedding, we treat the entire patient history as a "sentence".

Methods

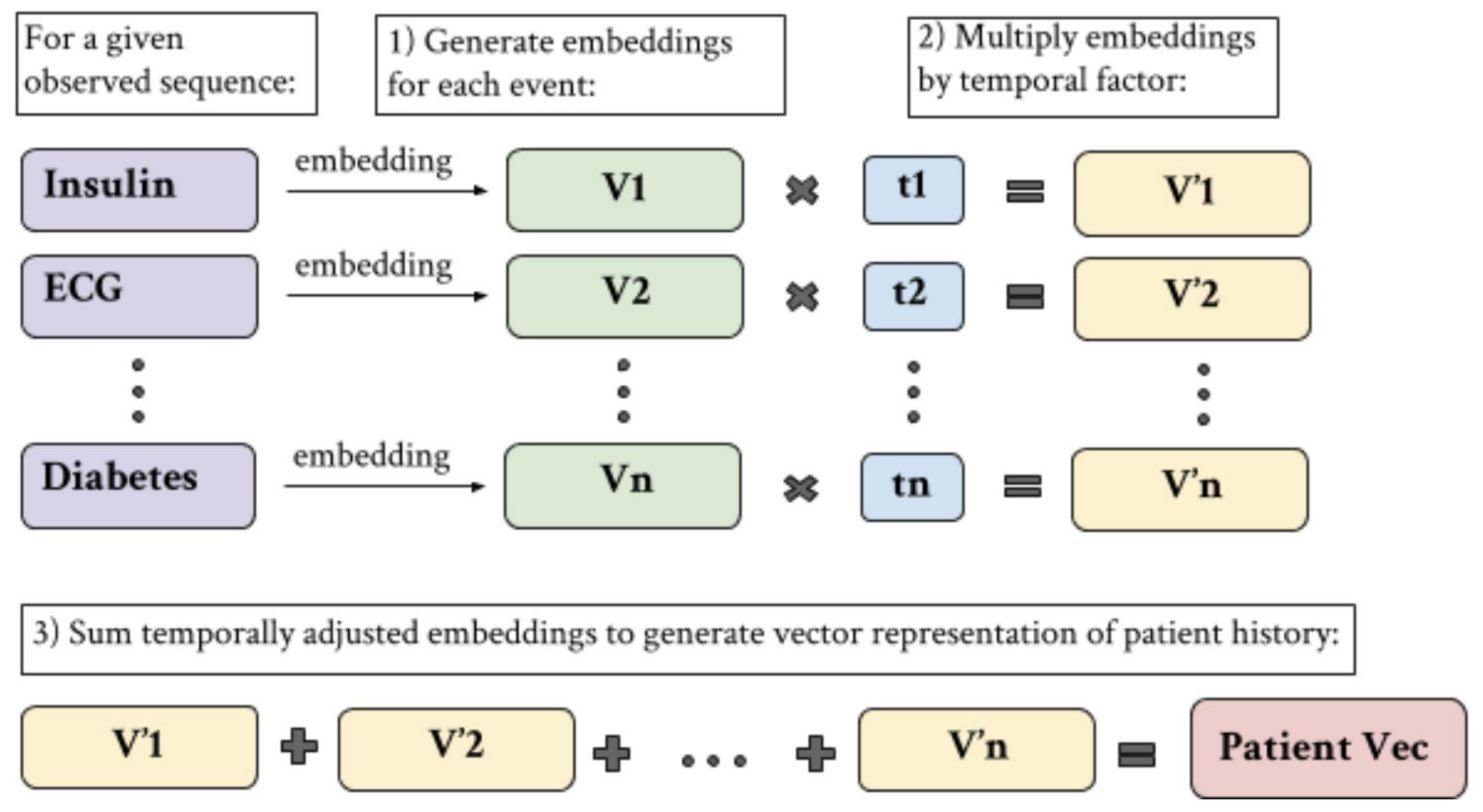
For a given observed sequence:
1) Generate embeddings for each event:
2) Multiply embeddings by temporal factor:
Insulin
ECG
Diabetes
embedding
embedding
embedding
V1
V2
Vn
t1
t2
tn
V'1
V'2
V'n
3) Sum temporally adjusted embeddings to generate vector representation of patient history:
V'1
V'2
V'n
Patient Vec

Figure 2. Process for generating vector representation of patient medical history

Results

A patient gets a positive label if the diagnosis of interest is present in their most recent visit to the ICU. We created disease-specific models for each word embedding algorithm using both penalized logistic regression (lasso) and deep learning (DL) by using the overall patient history embeddings as features.

We assess prediction performance with regard to static embeddings only, compare static and contextual performance for ELMo and BERT, and evaluate training times for each algorithm. Our diagnoses of interest are cardiac dysrhythmia, esophageal disease, and mitral valve disorder which have varying prevalences.

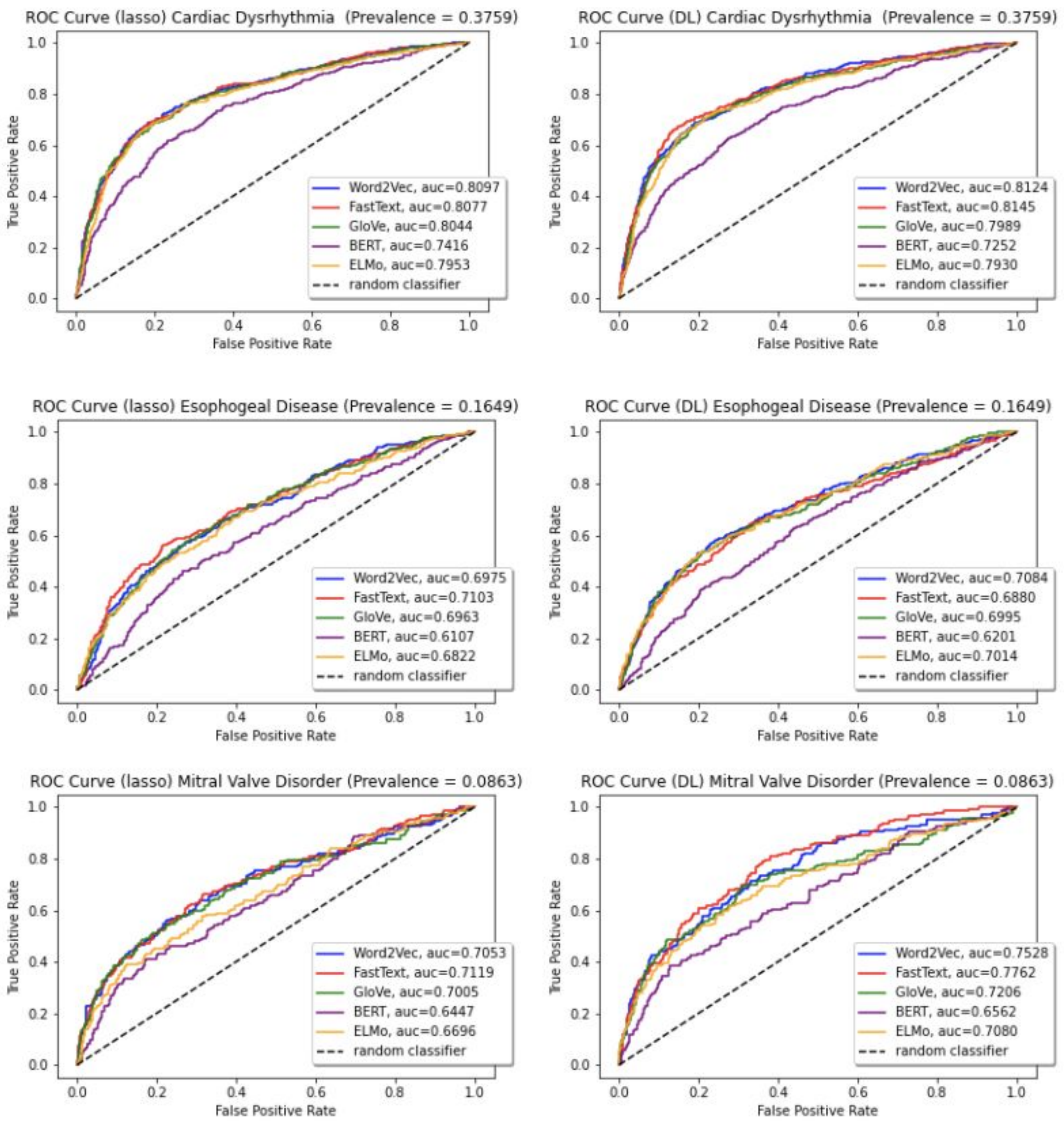

Figure 3. ROC Curves for disease-specific models

Results

Table 2: Comparison of Contextual vs. Static Embedding for ELMo and BERT

Cardiac Dysrhythmia	**ELMo (contextual)**	**ELMo (static)**	**BERT (contextual)**	**BERT (whole sentence)**	**BERT (static)**
AUC (lasso)	0.8037	0.7964	0.7168	0.7187	0.7416
AUC (DL)	0.7970	0.7974	0.7007	0.7321	0.7240
Esophageal					
AUC (lasso)	0.6993	0.7076	0.5991	0.6243	0.6107
AUC (DL)	0.6090	0.6963	0.6047	0.6150	0.6141
Mitral Valve					
AUC (lasso)	0.7054	0.6696	0.6350	0.6090	0.6447
AUC (DL)	0.6878	0.7080	0.6374	0.6044	0.6562

Table 3: Comparison of Train Times for Embedding Algorithms

Embedding Algorithm	**Train Time CPU (s)**	**Train Time GPU (s)**
Word2Vec	89	24
FastText	262	48
GloVe	11	14
ELMo	1830	60
BERT	382	72

**CPU consists of 2 x Intel Xeon E5-2660 v3 @ 2.60 GHz CPUs with 128 GB RAM installed.
**GPU consists of 2x GenuineIntel Intel(R) Xeon(R) Silver 4216

Conclusion

- Models that used FastText and Word2Vec embeddings tend to yield better prediction results in our models. ELMo performs well at a higher disease prevalence, but performs comparatively worse at lower prevalence. BERT performs consistently worse than the other algorithms.
  - ELMo and particularly BERT are complex models with more parameters. We likely did not have enough training data to get state-of-the-art results
- Models that used contextual embeddings did not perform better than static models.
  - Again, could be a lack of training data and the fact that the majority of events in the data were uniquely coded.
- For unique medical events in a patient history, we recommend Word2Vec or FastText given the faster training times and higher performance.

References

- Mikolov T, Sutskever I, Chen K, Corrado GS and Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems 26, volume 7. Curran Associates, Inc., pp. 3111-3119.
- Joulin A, Grave E, Bojanowski P and Mikolov T (2017) Bag of tricks for efficient text classification. In: Procedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. Association for Computational Linguistics, pp. 427-431.
- Pennington J, Socher R and Manning CD (2014) Glove: Goobal vectors for word representation. In: Empirical Methods in Natural Language PRocessing (EMNLP). pp 1532-1543.
- PEters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, and Zettlemoyer L (2018) Deep contextualized word representations. In: Proc. of NAACL.
- Devlin J, Ming-Wei C, Lee K and Toutanova K (2018) Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- Johnson AEW, Pollard TK, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA and Mark RG (2016) Mimic-iii, a freely accessible critical care database. Scientific Data 3: 160035.
- Farhan W, Wang Z, Huang Y, Wang S, Wang F and Jiang X (2016) A predictive model for medical events based on contextual embedding of temporal sequences. JMIR medical informatics 4(4):e39.

Acknowledgements

- The Long Group in the Department of Biostatistics, Epidemiology, and Informatics at the University of Pennsylvania.