2018 IEEE International Conference on Healthcare Informatics

Machine Learning Methods for Disease Prediction with Claims Data

Tanner Christensen*, Abraham Frandsen†, Seth Glazier*, Jeffrey Humpherys*‡, David Kartchner*
Department of Mathematics
Brigham Young University
Provo, Utah, 84602
Emails: *tkchristensen@byu.edu *[seth.glazier,jeffh,david.kartchner]@math.byu.edu
Department of Computer Science
Duke University
Durham, NC 27708
Email: †abef@cs.duke.edu
Department of Population Health Sciences
University of Utah School of Medicine
Salt Lake City, UT 84158
Email: ‡jeff.humpherys@hsc.utah.edu

***Abstract*—One of the primary challenges of healthcare delivery is aggregating disparate, asynchronous data sources into meaningful indicators of individual health. We combine natural language word embedding and network modeling techniques to learn meaningful representations of medical concepts by using the weighted network adjacency matrix in the GloVe algorithm, which we call Code2Vec. We demonstrate that using our learned embeddings improve neural network performance for disease prediction. However, we also demonstrate that popular deep learning models for disease prediction are not meaningfully better than simpler, more interpretable classifiers such as XGBoost. Additionally, our work adds to the current literature by providing a comprehensive survey of various machine learning algorithms on disease prediction tasks.**

***Keywords*-Machine Learning, Manifold Embeddings, Disease Prediction, Recurrent Neural Networks**

I. Introduction

The increasing volume of healthcare data contained in Electronic Health Records (EHRs) has caused many to consider the possibility of designing automated clinical support and disease detection systems based on patient history and risk factors [1]. A number of past studies have attempted to use patient laboratory tests [2], [3], [4], diagnoses [2], [5], [6], [7], [8], [9], and medications [5] as means of predicting disease onset. Such models have also been used to identify potentially unknown risk factors [9], often while simultaneously improving sensitivity and specificity of detection.

A number of recent studies have been successful in predicting disease via various methods, including support vector machines [6], [10], [11], [12], logistic regression [9], random forests [6], [13], neural networks [5], [2], and time series modeling techniques [3]. Many have noted that deep learning methods have been particularly successful for offering new insight into both data representation and diagnosis in medicine. We note that the following have been particularly salient in recent literature:

- **Embeddings of Medical Concepts:** Many papers have applied word embedding techniques from natural language processing to obtain embedded representations of medications, diagnoses, and procedures using adaptations of word2vec [14] and GloVe [15]. Though techniques are varied, they include supplementing medical corpora with insurance claims [16], jointly learning diagnosis and medical visit representations via modifications of the word2vec loss function [17], and using the vectors obtained from the embedding layer of a sequence-to-sequence RNN designed to forecast upcoming medical encounters [18]. Additional work has shown that clustering these embedded representations produces meaningful groupings of conditions not identified by groupers such as Clinical Classification Software (CCS) [19].
- **Disease Prediction Models:** A number of papers have additionally developed means of predicting future medical conditions using a number of neural network models. The general approach of each paper is to learn an encoded representation of an individual's EHR up to time t and then try to predict events in the time interval $[t+1, t+k]$, where most papers either try to predict the patient's next visit [5], [18] or disease onset in the next 1-2 years [2], [18], [17], [20]. Proposed methods for encoding past medical history include convolutional neural networks (where the direction of the convolution captures a temporal relationship in the data) [2], multi-label RNN [5], graph-based neural attention models combined with RNNs [18], and generative adversarial networks [21].

The vast majority of recent literature on disease prediction has focused almost entirely on deep learning models, to the exclusion of other potentially viable models as baselines for prediction tasks. We have found no other papers that investigate the usefulness of machine learning algorithms other than logistic regression and multi-layer perceptrons in learning from embedded representations of medical concepts.

2575-2634/18/$31.00 ©2018 IEEE
DOI 10.1109/ICHI.2018.00108

Additionally while many papers seek to learn probabilities of future disease onset from past visits, few of them incorporate additional and potentially meaningful personal demographic information (e.g. age, sex, weight) into their analysis, even though such variables have demonstrated correlations with the onset of many chronic diseases [22], [23]. Accordingly, this paper adds to the previous literature in the following ways:

- **General Usefulness of Embeddings:** We show that embeddings can be adapted to work quite well with other machine learning algorithms and propose a new way to obtain such embeddings that more explicitly captures the temporal relationships between observed diagnoses/procedures than previously described methods. Specifically, we demonstrate that embeddings substantially enhance the performance of linear classifiers such as logistic regression by meaningfully condensing information that is otherwise difficult for these algorithms to capture.
- **Non Deep Learning Machine Learning Baselines:** We demonstrate that other machine learning algorithms such as XGBoost [24] can achieve comparable, or in some cases superior, prediction accuracy to RNNs.

II. Code Embedding Algorithms

A. Temporal Diagnosis and Procedure Embeddings: Code2Vec

In [16], Choi et al. utilized word2vec to create vector embeddings for medical diagnoses and procedures. Their model was trained by treating diagnoses and procedures as "words" and patients as "documents." We propose a new framework to obtain diagnosis and procedure code embeddings as an extension of the GloVe word embedding algorithm. The word2vec algorithm and the GloVe algorithm both operate under the Distributional Hypothesis, which states that words that appear in similar contexts tend to have similar semantic meaning [25]. Code2Vec, similarly, assumes that diagnoses and procedures that appear in similar contexts have related clinical meaning.

1) Redefining Disease Co-occurrence: A word2vec model is traditionally trained using a *context window*, meaning if two words appear within 10 words (for example) of each other, they are considered to be in the same context. GloVe handles context somewhat differently by constructing a word co-occurrence matrix that contains counts of how many times a word appears in the same context (also utilizing a context window).

Using a context window works wonderfully for natural language processing problems. However, for the application of clinical diagnosis and procedure codes, some unwanted behavior can arise. For example, imagine a patient goes to her doctor and is diagnosed with bronchitis. Now say she does not go the doctor again for a full year, at which time she happens to be diagnosed with cancer. Using a typical sequence-based context window to define co-occurrences, this means that the data suggests there is a strong link between bronchitis and cancer, which is not necessarily the case.

We circumvent this problem by utilizing a co-occurrence window constructed using the temporal spacing between medical codes. While previous models have only counted co-occurrence between codes from the same visit [18], [17], we create a more flexible approach that allows diseases to interact across multiple days. We believe that this is more realistic since the disease diagnosis and treatment is not likely to occur on the same day, especially for chronic diseases. It is reasonable to see a few days or even weeks between related medical treatments, even for patients hospitalized for acute conditions. Accordingly, we define co-occurrence between codes in as demonstrated in Table I.

Temporal Difference	Co-occurrence Weight
0 - 5 Days	1
6 - 14 Days	0.5
15 - 30 Days	0.25
31 - 60 Days	0.125

TABLE I
Weights of medical code co-occurrence based on temporal difference between codes

This method of redefining co-occurrence is reminiscent of that employed by Choi et al. in [18], where co-occurrence is restricted to codes that occur on the same day. Thus, we investigate whether factoring in long-term disease relationships can enable improved learning of features to predict chronic disease.

It is important to note that this co-occurrence matrix is equivalent to the weighted adjacency matrix of a network of diseases where vertices correspond to diagnosis and procedure codes and edge weights correspond to sum of the co-occurrence scores. We believe that the properties of this network could yield important insights into disease comorbidity, though investigating this possibility is beyond the scope of this paper.

2) Learning Embeddings: Once we have obtained this co-occurrence matrix, we minimize the following cost functional from Pennington et al. [15] in order to obtain our embeddings:

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - log(X_{ij}))^2 \qquad (1)$$

which is essentially a weighted least squares problem. The pieces of J are as follows:

- V is the size of the vocabulary (number of distinct codes). All of the distinct codes are enumerated ranging from 1 to V.
- X_{ij} is the co-occurrence score between the i-th and j-th medical codes.
- $f(x)$ is a weight function on the co-occurrence score. Pennington et al. empirically found that the function,

$$f(x) = \begin{cases} (x/x_{max})^{\alpha} & \text{if } x < x_{max} \\ 1 & \text{otherwise} \end{cases}$$

Disease	ICD-10 Codes
HF	I09.81, I10, I13.0, I13.2, I50.*
AMI	I21.*, I22.*
DM	E11.*
CKD	D63.1, E08.22, E09.22, E10.22, E11.22
COPD	J41.1, J41.8, J42, J43.*, J44.*, J47.*, J68.3, J68.8, J68.9, Q33.4

TABLE II
DIAGNOSIS AND PROCEDURE CODES USED TO IDENTIFY EACH DISEASE. * INDICATES THAT CODES WITH ANY DIGIT(S) IN THIS PLACE CORRESPOND TO THE DISEASE IN QUESTION.

with $\alpha = .75$ and $x_{max} = 100$ works well. We borrow this same weight function for our experiments. This function ensures that no code will be overweighted.

- $w_i, \tilde{w_j}$ are the embedding vectors associated with the i-th and j-th medical codes, respectively. More specifically, $\tilde{w_j}$ is referred to as a context embedding vector. Because w and $\tilde{w}$ iterate over all of the distinct medical codes, at the end of training we will have two different sets of word vectors, each of which can be considered equally valid. We combine these two sets by addition for improved results.
- $b_i, \tilde{b_j}$ are the biases associated with the i-th and j-th medical codes, respectively. These biases are thrown away after training.

For the sake of clarity, we refer to codes embeddings learned with a naive co-occurrence as `GloVe embeddings` and those learned with our redefined co-occurrence as `Code2Vec embeddings`.

Total Patients	466,715
HF	758
AMI	306
DM	1,787
CKD	1,367
COPD	661
Proportion White	0.869
Proportion Ethnically Hispanic	0.081
Mean Age	30.1

TABLE III
BASIC STATISTICAL AND DEMOGRAPHIC CHARACTERISTICS OF DATASET

III. Disease Prediction Experiments

A. Disease Identification and Data Preprocessing

Our disease prediction experiments focused on predicting five high-impact, preventable, chronic conditions: heart failure (HF), acute myocardial infarction (AMI), type II diabetes mellitus (DM), chronic kidney disease (CKD), and chronic obstructive pulmonary disease (COPD). We identify each with the appearance of at least two medical codes, as done in [18], [2]. Requiring two codes instead of one avoids false positive identification due to miscoding. Disease identification relies on diagnosis codes from the International Classification of Diseases, 10th Edition (ICD-10) [26]. The codes used to identify each disease are described in Table II.

Our data consists of a deidentified subset of $466,715$ patients from SelectHealth®, a health insurance provider in the Intermountain West of the United States. Patients in the dataset were monitored beginning in November, 2015 through January, 2018. Basic summary data on our patient population is given in Table III. In order to maximize the number of cases of disease onset in our time window, we aligned patients by time of disease onset as done in [27]. For each patient identified as positive according to above criteria, we randomly choose a `last_observed_date` in year prior to diagnosis and throw away all observations after this date. For each disease, we throw away all individuals who met one of the following conditions:

1) **Ambiguous Patients:** Patients had exactly one diagnosis code corresponding to disease in the given time period
2) **Preexisting Conditions:** Patients who already had disease or were diagnosed during the first year of our data. These patients were identified as all individuals who received two diagnosis codes corresponding to the disease in the first year.
3) **Insufficient History:** Patients who had less than one year of medical history after the above cleaning procedure.

Due to the vast class imbalance in our data, we stratified our population to maintain a ratio 50:1 healthy:diseased individuals.

B. Models

For each of the above diseases, we evaluated the performance of the following models. The parameters of each were chosen via a grid search:

- **Logistic Regression (LR):** An L_1-regularized logistic regression with $\lambda = 1$. We use a 1:50 class weighting to account for the stratification of our dataset.
- **Random Forest (RF):** A random forest [28] from Scikit-Learn [29] with 200 trees, a minimum of six samples per leaf, and no max depth.
- **XGBoost (XGB):** XGBoost [24] is a gradient boosting algorithm that has demonstrated exceptional performance

Model	HF	AMI	DM	CKD	COPD
LR	.7347	.7205	.6261	.6852	.6195
LR+GloVe	.9265	.8914	.8602	.9216	.8870
LR+TC2V	.9391	.8742	.8623	.9203	.9126
RF	.9267	.8851	.8781	.9282	.9013
RF+GloVe	.9206	.8738	.8345	.9199	.8844
RF+TC2V	.9208	.8770	.8345	.9233	.8880
XGB	**.9430**	.9126	**.8867**	**.9300**	**.9171**
XGB+GloVe	.9347	**.9136**	.8611	.9201	.9036
XGB+TC2V	.9347	.9097	.8626	.9203	.9126
RNN	.9217	.8204	.8498	.9163	.8433
RNN+GloVe	.9305	.8544	.8710	.9205	.8901
RNN+TC2V	.9317	.8834	.8619	.9234	.8747

TABLE IV
RESULTS: ROC AUC SCORES FROM DISEASE PREDICTION EXPERIMENTS.

on a variety of tasks. Accordingly, we train an XGBoost classifier with 200 trees and parameters *learning_rate* = 0.1, *max_depth* = 6, *colsample_bytree* = 0.2.

- **LSTM:** An LSTM cell with a 128-dimensional hidden state, preceded by a learnable dense embedding layer of dimension 100, ending with a dense softmax classification layer. Demographic information is concatenated to the output of our RNN and then passed into the softmax layer.

C. Data Representations

One important question that we try to answer in this paper is the best means of encoding data for each of the above algorithms in question. For each of the above models, we test the following data configurations:

- **Raw Codes:** Vectors corresponding to ICD-10 diagnosis and procedure codes, as well as . RNN models are fed a sequence of one-hot vectors while LR, RF, and XGB models use a count vector of all of an individual's codes over the past year.
- **GloVe:** Vector embeddings generated by the GloVe algorithm using their original definition of a symmetric context window with size 15.
- **Code2Vec:** Vector embeddings using our redefined definition of co-occurrence. This method is described in section II-A.

We test each of the above configurations with the addition of demographic data, which for our dataset is the age and sex of each individual. In order to use GloVe and Code2Vec embeddings with logistic regression, random forests, and XGBoost, we sum up the codes in a patient's history and normalize for the amount of time the patient was observed.

IV. Results

The results of our experiments are shown in Table IV. Our consistently best performer was XGBoost on raw diagnosis, procedure. Moreover, XGBoost performed better than LSTMs in every scenario. This suggests that quality disease predictions do not require incredibly complex models. Moreover, XGBoost trains significantly faster than LSTMs and requires minimal parameter tuning and should, therefore, not be disregarded in disease prediction tasks. Additionally, and perhaps most importantly, XGBoost is a highly interpretable model, with the ability to give the user insight as to how and why it makes its predictions. It is also worth noting that LSTMs trained with pre-defined embeddings consistently outperformed LSTMs that learned their own embedding layer, with no clear winner between temporally-sensitive Code2Vec embeddings and naive embeddings. This could indicate that most meaningful disease co-occurrence happens in the same visit, though more extensive testing of temporal weighting windows is required.

V. Conclusions

It was surprising to see XGBoost – a model that entirely disregards the sequential and temporal relationships in the data – perform better than an LSTM. Its performance demonstrates the importance of not overlooking simpler models in favor of more complicated ones. For future work, we are interested in seeing if there is any way we can combine the strengths of these two algorithms to create an even better classifier.

We also recognize that when we feed a sequence of medical codes to the LSTM, the LSTM treats these codes as though they are dispersed uniformly in time, which does not reflect reality. We believe the LSTMs inability to perceive the varying amounts of time between visits significantly limits is predictive ability, and we are currently exploring ways to better represent the time between medical codes with hopes of improving results.

VI. Acknowledgements

This work was supported in part by the National Science Foundation, Grant Number 1323785 and the Defense Threat Reduction Agency, Grant Number HDRTA1-15-0049.

References

[1] R. Snyderman, "Personalized health care: from theory to practice," *Biotechnology Journal*, vol. 7, no. 8, pp. 973–979, Aug. 2012.

[2] N. Razavian and D. Sontag, "Temporal convolutional neural networks for diagnosis from lab tests," *CoRR*, vol. abs/1511.07938, 2015. [Online]. Available: http://arxiv.org/abs/1511.07938

[3] A. Perotte, R. Ranganath, J. S. Hirsch, D. Blei, and N. Elhadad, "Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis," *Journal of the American Medical Informatics Association: JAMIA*, vol. 22, no. 4, pp. 872–880, Jul. 2015.

[4] N. Tangri, L. A. Stevens, J. Griffith, H. Tighiouart, O. Djurdjev, D. Naimark, A. Levin, and A. S. Levey, "A predictive model for progression of chronic kidney disease to kidney failure," *Jama*, vol. 305, no. 15, pp. 1553–1559, 2011.

[5] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor ai: Predicting clinical events via recurrent neural networks," in *Proceedings of the 1st Machine Learning for Healthcare Conference*, ser. Proceedings of Machine Learning Research, F. Doshi-Velez, J. Fackler, D. Kale, B. Wallace, and J. Wiens, Eds., vol. 56. Children's Hospital LA, Los Angeles, CA, USA: PMLR, 18–19 Aug 2016, pp. 301–318. [Online]. Available: http://proceedings.mlr.press/v56/Choi16.html

[6] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," *BMC medical informatics and decision making*, vol. 11, no. 1, p. 1, 2011.

[7] X. Wang, D. Sontag, and F. Wang, "Unsupervised learning of disease progression models," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 85–94.

[8] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzell, "Learning to diagnose with lstm recurrent neural networks," *arXiv preprint arXiv:1511.03677*, 2015.

[9] N. Razavian, S. Blecker, A. M. Schmidt, A. Smith-McLallen, S. Nigam, and D. Sontag, "Population-Level Prediction of Type 2 Diabetes From Claims Data and Analysis of Risk Factors," *Big Data*, vol. 3, no. 4, pp. 277–287, Dec. 2015. [Online]. Available: http://online.liebertpub.com/doi/full/10.1089/big.2015.0020

[10] N. Barakat, A. P. Bradley, and M. N. H. Barakat, "Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 4, pp. 1114–1120, Jul. 2010.

[11] "Prediction Modeling Using EHR Data: Challenges, Strategies,... : Medical Care." [Online]. Available: http://journals.lww.com/lww-medicalcare/Fulltext/2010/06001/Prediction_Modeling_Using_EHR_Data__Challenges,.17.aspx

[12] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes," *BMC Medical Informatics and Decision Making*, vol. 10, p. 16, 2010. [Online]. Available: http://dx.doi.org/10.1186/1472-6947-10-16

[13] A. V. Lebedev, E. Westman, G. J. P. Van Westen, M. G. Kramberger, A. Lundervold, D. Aarsland, H. Soininen, I. Kłoszewska, P. Mecocci, M. Tsolaki, B. Vellas, S. Lovestone, and A. Simmons, "Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness," *NeuroImage: Clinical*, vol. 6, pp. 115–125, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2213158214001326

[14] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 3111–3119.

[15] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162

[16] Y. Choi, C. Y.-I. Chiu, and D. Sontag, "Learning Low-Dimensional Representations of Medical Concepts," *AMIA Summits on Translational Science Proceedings*, vol. 2016, pp. 41–50, Jul. 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5001761/

[17] E. Choi, M. T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost, J. Tejedor-Sojo, and J. Sun, "Multi-layer representation learning for medical concepts," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 1495–1504. [Online]. Available: http://doi.acm.org/10.1145/2939672.2939823

[18] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun, "Gram: Graph-based attention model for healthcare representation learning," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '17. New York, NY, USA: ACM, 2017, pp. 787–795. [Online]. Available: http://doi.acm.org/10.1145/3097983.3098126

[19] D. Kartchner, T. Christensen, J. Humpherys, and S. Wade, "Code2vec: Embedding and clustering medical diagnosis data," in *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, Aug 2017, pp. 386–390.

[20] R. Miotto, L. Li, B. A. Kidd, and J. Dudley, "Deep patient: An unsupervised representation to predict the future of patients from the electronic health records," in *Scientific reports*, 2016.

[21] Z. Che, Y. Cheng, S. Zhai, Z. Sun, and Y. Liu, "Boosting deep learning risk prediction with generative adversarial networks for electronic health records," in *2017 IEEE International Conference on Data Mining (ICDM)*, Nov 2017, pp. 787–792.

[22] C. GA, W. WC, R. A, and M. JE, "Weight gain as a risk factor for clinical diabetes mellitus in women," *Annals of Internal Medicine*, vol. 122, no. 7, pp. 481–486, 1995. [Online]. Available: +http://dx.doi.org/10.7326/0003-4819-122-7-199504010-00001

[23] N. K, B. JP, T. TJ, S. SW, and W. DF, "Lifetime risk for diabetes mellitus in the united states," *JAMA*, vol. 290, no. 14, pp. 1884–1890, 2003. [Online]. Available: +http://dx.doi.org/10.1001/jama.290.14.1884

[24] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 785–794. [Online]. Available: http://doi.acm.org/10.1145/2939672.2939785

[25] Z. S. Harris, "Distributional structure," *Word*, vol. 10, pp. 146–162, 1954.

[26] W. H. Organization, *International statistical classification of diseases and related health problems*. World Health Organization, 2004, vol. 1.

[27] R. Ranganath, A. Perotte, N. Elhadad, and D. Blei, "Deep survival analysis," *arXiv preprint arXiv:1608.02158*, 2016.

[28] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.