doi:10.3233/SHTI210800

Encoding Health Records into Pathway Representations for Deep Learning

Marco Luca SBODIO[a,1], Natasha MULLIGAN[a], Stefanie SPEICHERT[a], Vanessa LOPEZ[a] and Joao BETTENCOURT-SILVA[a]
[a] *IBM Research Europe*

Abstract. There is a growing trend in building deep learning patient representations from health records to obtain a comprehensive view of a patient's data for machine learning tasks. This paper proposes a reproducible approach to generate patient pathways from health records and to transform them into a machine-processable image-like structure useful for deep learning tasks. Based on this approach, we generated over a million pathways from FAIR synthetic health records and used them to train a convolutional neural network. Our initial experiments show the accuracy of the CNN on a prediction task is comparable or better than other autoencoders trained on the same data, while requiring significantly less computational resources for training. We also assess the impact of the size of the training dataset on autoencoders performances. The source code for generating pathways from health records is provided as open source.

Keywords. Patient Representation, Convolutional Neural Networks, EHR

1. Introduction

Despite the promising results of deep learning techniques for performing analytics tasks, several open challenges remain in dealing with the heterogeneous data from Electronic Health Records (EHRs) and the lack of model intelligibility and interpretability required for real-world applications [1],[2],[3]. Data representation and encoding plays a key role in training successful models for prediction tasks and explainability; additionally, compact representations have been used to address challenges such as sparseness of EHR data [4]. Multi-source EHR data has been modelled as patient trajectories through time [5] or by representing EHR information as a 2D matrix [6] of appointments and diagnoses codes where convolutional neural networks (CNNs) were used for risk prediction [7]. The most common type of representation is a sequential ordering of a patient's data used as input to a Recurrent Neural Network (RNN) for application areas such as, for example, prediction or phenotyping [8].

This paper proposes a pathway representation that maps patient's health records into different classes over time, to form a machine-processable image-like structure for further analyses and deep learning tasks. A Patient Pathway Extractor application was developed and used to transform EHR data into the new representations (described in section 2.2). The application is shared in an open-source git repository. The pathways generated using the proposed representations were then validated using three mainstream

[1] Corresponding Author, IBM Research Europe, Dublin, Ireland; E-mail: marco.sbodio@ie.ibm.com

deep learning algorithms, described in section 2.3, and the preliminary results using synthetic data are included in section 3.

2. Methods

2.1. Pathway Representation

We propose an encoding of pathway data with an accompanying open-source application called Patient Pathway Extractor[2]. We use a set of predefined classes to classify data and pathway events; we build a structured representation that shows the discretized values of the data along a time dimension. We use Synthea [7], a FAIR dataset generator, to produce the CSV input for the Patient Pathway Extractor. More precisely, we classify data generated by Synthea along the following classes: *demographics* (patient details), *observations* (results of clinical exams and vitals), *conditions* (diagnoses and care plans), *medications*, *procedures* and *outcomes* (readmission, death, survival at a point in time).

Data may consist of isolated events happening at a specific point in time, of events having a duration. Events can be visualized along a timeline: isolated events can be shown as dots, while events having a duration can be shown as horizontal bars. When multiple events happen at the same time, or when an event include a set of data values, the timeline visualization can display these using an overlay information box.

Humans can easily understand the timeline visualization of a pathway, but such a representation is not helpful when trying to analyse data using machine learning or deep learning algorithms. For this reason we propose a novel image-like representation of the pathway data. We build such image-like representation using a three steps process: (1) representing the discretized data points in a 3-dimensional grid, (2) projecting into a bi-dimensional grid, and (3) numerical encoding.

Firstly, we map the discretized values of the data in a 3-dimensional grid. The dimensions of the grid represent respectively the order of the events (time), the different classes (demographics, observations, conditions, medications, procedures and outcomes), and co-occurrence of events (values of a given class having the same timestamp). Figure 1 shows (on the left) the 3-dimensional grid representation of the pathway timeline. Note that we do not encode timestamps along the time dimension, but only retain the order of events (recording timestamps is possible with a simple extension of the proposed representation). We use a configurable set of rules that discretize values into custom bins. We use spreadsheet (easily interpretable by practitioners) to define the rules, and parse them into executable formats using the Drools[3] rule engine. Our current sets of 246 rules cover demographics, medications, observations (based on patient age and gender, the LOINC code and its units), as well as outcomes. For example, a rule that takes as input a body mass index (BMI) observation (LOINC code 39156-5) and when its value is in the range [18.5, 25 kg/m^2], it maps it to the bin value “normal BMI”.

Subsequently, we project the 3-dimensional grid into a bi-dimensional grid: the horizontal axis denote the order of events, while the vertical axis denote the various classes of our representation. The projection places values having the same timestamp (co-occurring) one after the other along the horizontal axis. The order of events is

[2] Pathway Extractor, https://github.com/Alvearie/patient_pathway_extractor/

[3] Drools, https://www.drools.org

preserved along the x-axis. The pathway representation does not impose any restriction on the order of the classes on the y-axis and downstream applications may use different ordering. In practice, we consider every bi-dimensional corresponding to a value along the time dimension, we rotate each slice along the class dimension, and finally concatenate them as illustrated in Figure 1.

As a final step, we use a numeric space to encode the values of the bi-dimensional grid in a numeric space. The encoding space is $\mathbb{R}$ but can also be $\mathbb{N}$ depending on the downstream analysis task; for some applications we may encode values in the RGB space, which translates our representation into an image. This encoding step of the process produces a numeric representation of the pathway, which, while retaining meaningful dimensions, is also easy to use as input for machine learning and deep learning tasks.

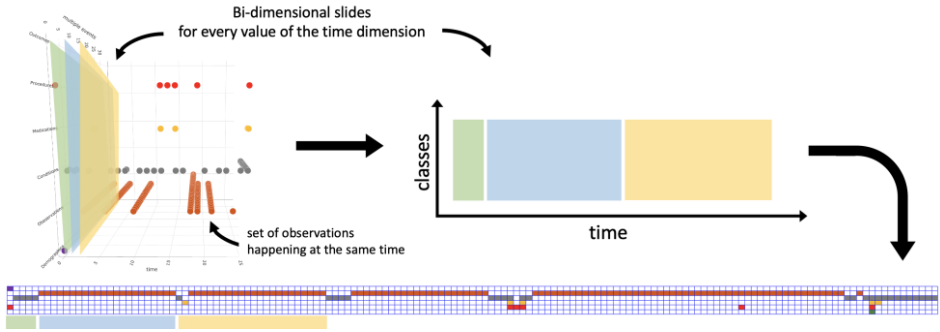


Figure 1. Transformation of patient pathway data from the 3-dimensional grid to the final bi-dimensional grid representation.

2.2. Synthetic EHR Data and Pathways Generation

We tested our approach using synthetic data generated by Synthea [9]. The generator creates realistic patient data with the help of a rule-based backend that determines the course of a condition. As with real-world EHR data, patients may have multiple concurrent conditions, which, over the course of their life, may interact or influence each other. Using Synthea, we generated a population of 500,000 patients, from which we extracted 1,073,105 pathways considering only a set of ten conditions including common chronic (e.g. diabetes) and acute (e.g. appendicitis, fractures) conditions.

2.3. Deep Learning Frameworks used to test the Pathway Representation

We used our pathways representation to create three pathway encodings using three types of autoencoders: Multilayer perceptron (MLP) autoencoder (Denoising Autoencoder), Sequence-to-Sequence (RNN) autoencoder, and CNN autoencoder.

An analysis of the pathways in our dataset has revealed that the maximum pathway length was 5128 data points on the x-axis; however, around 98% of the pathways fitted into a 6*400 grid (6 classes by 400 points on the x-axis). Pathways with a length of less than 400 were padded with zeros and only pathways not exceeding this size were used for training and testing with an 80/20% split. All autoencoders were designed to generate pathway encodings of the same length and their architectures are described in Figure 2.

- The MLP autoencoder consists of 3 layers for both encoder and decoder and a vector (bag) of pathway events was used as an input. Temporal dimension is not supported.

- The RNN autoencoder was built using GRU cells. The pathway grid was sliced vertically (time axis) and all events in that slice were concatenated providing a sequence of input vectors.
- The CNN autoencoder directly supports the pathways as input. It has an encoder with three convolutional layers followed by a fully connected layer to produce the pathway encoding. The decoder has the same architecture but in the reverse order.

We evaluated the performance of the three autoencoders using the following prediction task: given an input pathway, we remove from its representation the data identifying the medical condition that originated the patient pathway, and we use the trained autoencoder to predict such condition.

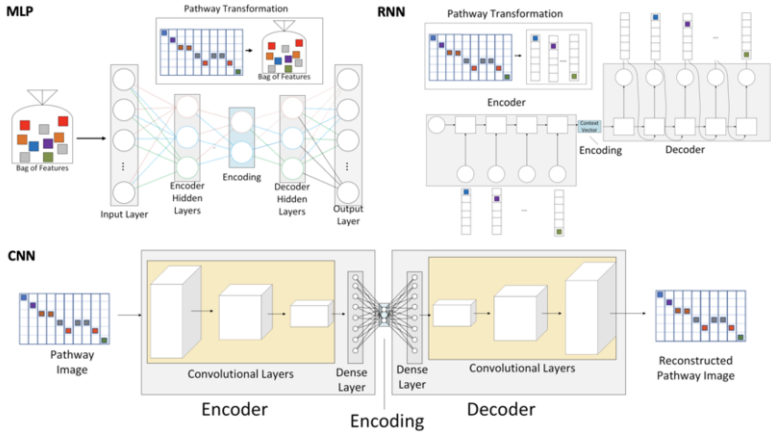
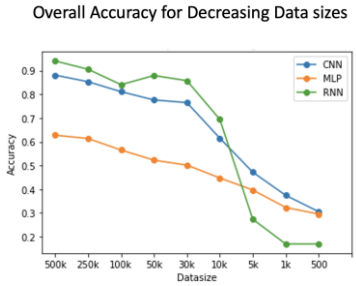


Figure 2. The three architectures used for compact representation of pathways.

3. Results and Discussion

In our experiments, the Pathway Extractor generated 1,073,105 pathways from a large EHR dataset. The flexible way in which pathways and conditions are represented allows for new classes to be added or existing ones to be dropped. Similarly, the codes and values may be adapted by modifying the discretization rules. Our approach is not limited to Synthea and may be applied to other EHR datasets.

In our prediction task, RNN gave the best accuracy (94.0%) followed by CNN (88.1%) and MLP (62.8%). We then tested the models performance using different training datasets to understand the impact on their accuracy. Figure 3 shows the overall accuracies for the three autoencoders when trained on the pathway data extracted from synthetic populations (generated with Synthea) of decreasing size (from 500,000 to 500 people). CNN and RNN, as expected, outperformed MLP, and larger training sets increase prediction accuracies. We note that for RNN and CNN there is a significant increase in accuracy with training sets computed with a population larger than 10,000. Overall CNN achieves good accuracy while requiring considerably less computational resources for training compared to RNN: 37.5% less memory, and over 98% less time (see Figure 3). Additionally, CNN and our pathway image-like representation may help in explaining predictions by using existing techniques such as attention to highlight important events in the input pathway grid [5].



Accuracy and computational resources for the population size of 500,000

	CNN	RNN	MLP
Accuracy (%)	88.1	94.0	62.8
Time (min)	15	1320	65
RAM (GB)	25	40	25

$$\text{Accuracy} = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}$$

Figure 3. Accuracies and computational resources of the three autoencoders.

4. Conclusions

This paper describes an approach to represent patient pathways and provides an accompanying open-source application that transforms health records into a machine-processable representation for deep learning tasks. We have evaluated this approach using data generated by Synthea, and observed that in a prediction task a CNN performed almost as well as an RNN, while being significantly less expensive to train in terms of computational resources and training time, and enabling further work on predictions explainability. Our results also give insight on how much data may be required for model training. Further work includes expanding the pathway representation with additional classes and data beyond EHRs.

References

[1] Si Y, Du J, Li Z, Jiang X, Miller T, Wang F, Jim Zheng W, Roberts K. Deep representation learning of patient data from Electronic Health Records (EHR): A systematic review. J Biomed Inform. 2021 Mar;115:103671..

[2] Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE J Biomed Health Inform. 2018 Sep;22(5):1589-1604. doi: 10.1109/JBHI.2017.2767063. Epub 2017 Oct 27.

[3] Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Sci Rep. 2016 May 17;6:26094.

[4] Yuqi S, Jingcheng D, Zhao L, Xiaoqian J, Timothy M, Fe, W, Zheng WJ, Kirk R. Deep representation learning of patient data from Electronic Health Records (EHR): A systematic review. Journal of Biomedical Informatics 115 (2021), pp.103671

[5] Nguyen-Duc T., et al. Deep EHR Spotlight: a Framework and Mechanism to Highlight Events in Electronic Health Records for Explainable Predictions. AMIA 2021 Virtual Informatics Summit (2021)

[6] Cheng Y, Wang F, Zhang P, Hu J. Risk prediction with electronic health records: A deep learning approach. InProceedings of the 2016 SIAM International Conference on Data Mining 2016 Jun 30 (pp. 432-440). Society for Industrial and Applied Mathematics.

[7] Suo Q, Ma F, Yuan Y, Huai M, Zhong W, Gao J, Zhang A. Deep patient similarity learning for personalized healthcare. IEEE transactions on nanobioscience. 2018 May 16;17(3):219-27.

[8] Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor ai: Predicting clinical events via recurrent neural networks. InMachine learning for healthcare conference 2016 Dec 10 (pp. 301-318). PMLR.

[9] Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, Duffett C, Dube K, Gallagher T, McLachlan S. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. Journal of the American Medical Informatics Association. 2018 Mar 1;25(3):230-8.