Continually-Adaptive Representation Learning Framework for Time-Sensitive Healthcare Applications

Akash Choudhuri
akash-choudhuri@uiowa.edu
Department of Computer Science
Iowa City, Iowa, USA

Hankyu Jang
hankyu-jang@uiowa.edu
Department of Computer Science
Iowa City, Iowa, USA

Alberto M. Segre
alberto-segre@uiowa.edu
Department of Computer Science
Iowa City, Iowa, USA

Philip M. Polgreen
philip-polgreen@uiowa.edu
Department of Internal Medicine
Iowa City, Iowa, USA

Kishlay Jha
kishlay-jha@uiowa.edu
Department of Electrical and
Computer Engineering
Iowa City, Iowa, USA

Bijaya Adhikari
bijaya-adhikari@uiowa.edu
Department of Computer Science
Iowa City, Iowa, USA

ABSTRACT

Continual learning has emerged as a powerful approach to address the challenges of non-stationary environments, allowing machine learning models to adapt to new data while retaining the previously acquired knowledge. In time-sensitive healthcare applications, where entities such as physicians, hospital rooms, and medications exhibit continuous changes over time, continual learning holds great promise, yet its application remains relatively unexplored. This paper aims to bridge this gap by proposing a novel framework, i.e., Continually-Adaptive Representation Learning, designed to adapt representations in response to changing data distributions in evolving healthcare applications. Specifically, the proposed approach develops a continual learning strategy wherein the context information (e.g., interactions) of healthcare entities is exploited to continually identify and retrain the representations of those entities whose context evolved over time. Moreover, different from existing approaches, the proposed approach leverages the valuable patient information present in clinical notes to generate accurate and robust healthcare embeddings. Notably, the proposed continually-adaptive representations have practical benefits in low-resource clinical settings where it is difficult to training machine learning models from scratch to accommodate the newly available data streams. Experimental evaluations on real-world healthcare datasets demonstrate the effectiveness of our approach in time-sensitive healthcare applications such as Clostridioides difficile (C.diff) Infection (CDI) incidence prediction task and medical intensive care unit transfer prediction task.

CCS CONCEPTS

• **Information systems**; • **Computing methodologies** → **Lifelong machine learning**;

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom

ACM ISBN 979-8-4007-0124-5/23/10.
https://doi.org/10.1145/3583780.3615464

KEYWORDS

dynamic embeddings; clinical notes; continual learning; electronic healthcare records

ACM Reference Format:
Akash Choudhuri, Hankyu Jang, Alberto M. Segre, Philip M. Polgreen, Kishlay Jha, and Bijaya Adhikari. 2023. Continually-Adaptive Representation Learning Framework for Time-Sensitive Healthcare Applications. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3583780.3615464

1 INTRODUCTION

Healthcare has emerged as a prominent domain for applied machine learning research, primarily driven by the widespread availability of fine-grained hospital operations data and advancements in computing capabilities [30]. Researchers have approached various healthcare challenges by formulating them as machine learning tasks. Some common areas of focus in machine learning for healthcare research include assisting in predicting outcomes and risks [4], disease diagnosis and monitoring [23], optimizing decision-making processes [35], and enhancing workflow efficiency [20]. A precursor to many of these healthcare applications is the availability of pre-trained representations of healthcare entities. Representation of healthcare entities, such as patients, doctors, rooms, and medications, can be learned from the diverse data streams by various representation learning models such as Word2Vec [24], GloVE [29], ELMo [28], or BERT [7]. These models can embed the discrete entities into a continuous vector space as distributed, dense embeddings based on the distributional hypothesis that argues the entities that occur in the same contexts tend to have similar semantics [3]. While a majority of these representation learning models approaches have been developed for a general domain, some recent studies such as [5, 12, 37] have attempted to model the special properties of healthcare data and learn high-quality representations.

Despite significant advances made, the existing approaches have two major drawbacks. First, the existing approaches fail to leverage the granular patient information such as patient complaints, disease progression, treatment history, and other crucial information present in clinical notes. Second, the existing approaches are unable to continually (or incrementally) accommodate information from newly available data streams. This becomes limiting in time-sensitive and resource-critical domains such as healthcare, where the efficient adaptation of healthcare entities is of utmost importance.

Prior research has attempted to learn adaptive embeddings through a range of solutions such as knowledge distillation [9], weights pruning [26], and continual learning [6]. Amongst them, the continual learning-based approaches have attracted increasing interest from the community due to their natural ability to adapt the representations to the continuous streams of data. However, directly applying these approaches to the current problem setting would yield unsatisfactory performance. This is because the existing approaches are not designed to model the interaction among heterogeneous entities. To address this, we propose a new continual representation learning scheme that models the co-evolving dynamics of entities and efficiently adapts the representations to the newly available data streams. To effectively learn dynamic embeddings of healthcare entities based on heterogeneous interactions, we design a dedicated objective function for each component and then propose a joint inference mechanism. Specifically, the proposed approach considers the successive data snapshots as a sequence of related tasks and updates the representations that are affected by the new snapshot while preserving those that were well-trained previously. The main challenge in this strategy is to automatically identify the entities whose context (i.e., interactions) evolved over time and thus would require retraining of representations. To address this, we propose a scheme wherein at every new snapshot, we identify and retrain the representations of those entities whose context evolved over time. Following this strategy, the proposed technique is continually (iteratively) applied to the consecutive snapshots, and the entity representations are adapted. Moreover, as the proposed CL formulation facilitates incremental updates of entity representations, it effectively mitigates the expensive retraining of the proposed model whilst acquiring information from data streams. One critical issue in CL based approach is to prevent catastrophic forgetting, i.e., the model abruptly forgets knowledge learned from previous data snapshots when learning on the new data snapshot. To overcome this, we propose a regularization mechanism that constrains the learned entity representations in the embedding space.

In this research, our contributions can be summarized as follows:

- We propose a new end-to-end continual learning framework that updates the representations of healthcare entities in an online fashion. This strategy greatly improves both the accuracy and computational efficiency of the proposed approach whilst accounting for the time-sensitive nature of healthcare applications.
- The proposed approach leverages the granular information in clinical notes to learn semantically enriched, accurate, and robust representations. This has immediate practical benefits to a variety of downstream predictive health applications.
- Extensive experiments on real-world healthcare datasets through the tasks of Clostridioides difficile (C.diff) Infection (CDI) incidence prediction task and medical intensive care unit (MICU) transfer prediction validates the effectiveness of the proposed approach.

2 METHOD

Overview: Our neural network architecture consists of two primary components. The first component consists of dynamic co-evolving neural networks [19], which are designed to learn meaningful embeddings of entities (including patients, rooms, nurses, doctors, etc.) encountered in healthcare facilities based on the observed heterogeneous interactions (e.g., patient-is cared by-nurse, doctor-prescribes-medication, nurse-visits-room interactions e.t.c.).

The second component further enhances the learned embeddings by infusing the information extracted from clinical notes. Clinical notes are written by healthcare providers, including doctors and nurses, as they provide care to the patient, administer medication, and/or perform procedures. These notes contain fine-grained information about the patient's medical progress. This additional wealth of information can be extremely useful in foreseeing patient outcomes and forecasting probable risk factors. In order to make use of this data, our model uses a natural language processing model to extract pertinent elements from clinical notes and merge them with the embeddings learned from the interactions. By combining both interactions and clinical notes, our model captures a more comprehensive representation of the patient's healthcare journey.

In healthcare applications, the data is continually generated as patients receive care in the healthcare facility. Moreover, this sensitive healthcare data is protected by the Health Insurance Portability and Accountability Act (HIPPA) of 1996 in the United States [1]. The act mandates that healthcare providers do not disclose health-related information to anyone other than the patient and authorizes representatives. Due to government regulation and other data leakage risks, healthcare data is usually stored in low-access machines to minimize the risk of unauthorized access.

The combination of a continual stream of data being generated (which ought to be included for predictive healthcare applications) and difficulty of access poses a conundrum; on the one hand, we would like to train our model with the latest batch of data being generated in near real-time, but on the other hand, the data cannot be easily accessed to train large language-based models over and over again. To address this, we adopt continually adaptive training in which the model can easily integrate streams of newer data that appear over time.

Overall, our proposed model leverages a set of co-evolving neural networks to process heterogeneous interactions between healthcare entities and the natural language processing model to process and extract fine-grained information from clinical notes to provide a holistic view of a patient's healthcare experience to enable more accurate predictions in different downstream tasks and to provide a better understanding of patient dynamics. Finally, we leverage a continual learning framework to train the model on large batches of sensitive clinical data in a temporally dynamic fashion while minimizing access. In the next few sections, we describe the components of our proposed model in detail.

2.1 Language Model Integration

We first describe the language model component. Here the goal is to learn a low-dimensional representation of each clinical note in the corpus while being sensitive to the semantic changes over

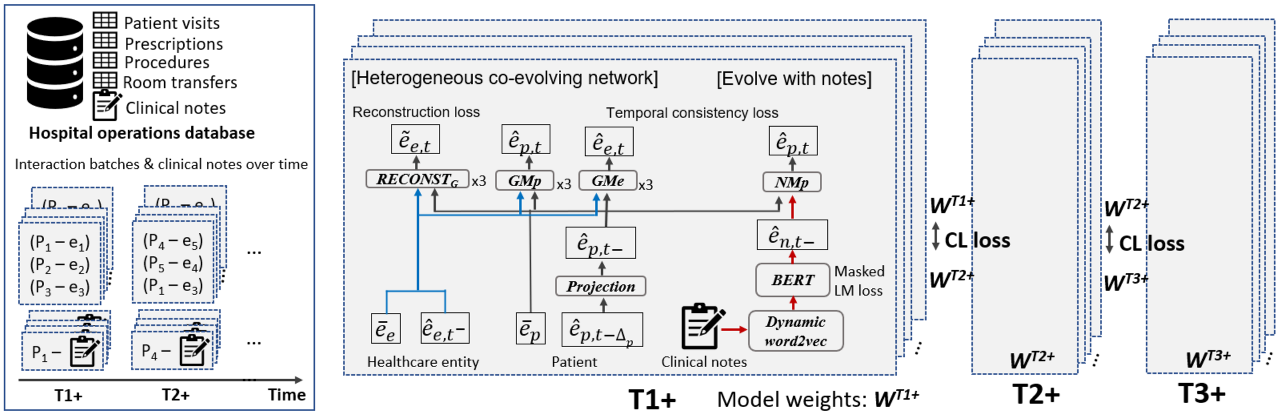

Figure 1: Model figure. Our method is trained in a continual setting. Interactions that belong to a time window (e.g., T1+) are processed in batches. Then, for interactions in the next time window (e.g., T1+), the model is trained to minimize continual learning (CL) loss. Within each time window, the model utilizes (i) clinical notes to update patient embeddings and (ii) heterogenous co-evolving networks that are reconstruction modules and update modules per interaction type.

time. We describe the process of combining node embeddings with embeddings in a later section.

Creating Dynamic Word Vectors: Previous works by Yao et al. [38], Gulordava et al. [10] and Sagi et al. [33] ascertain that words which appear in documents undergo a semantic change as time progresses. Intuitively, this observation seems to apply in the medical setting as well. As newer diseases, medications, procedures, and symptoms emerge, words associated with these concepts gain new meanings/use cases and lose old ones. For example, Hydroxychloroquine is a drug primarily used to treat malaria. However, it gained traction as a drug that could cure COVID-19. Note that the presence of the word 'Hydroxychloroquine' in a clinical note in a pre-COVID era was a strong indication of malaria-related cases. However, this may no longer be true in COVID/post-COVID era. As evidenced by this example, we need to account for semantic changes in the words themselves before we leverage them to learn clinical note embeddings. Here, we address the concern by modifying the original architecture of the BERT [7] by initializing the model with learned word embeddings instead of random word embeddings. We use the popular **DynamicWord2Vec** model proposed by Yao et al. [38] to learn the representations of words found in clinical notes in evolving contexts. The **DynamicWord2Vec** model takes pre-processed (Stemming, tokenization, stop-word removal) clinical notes as input and outputs dynamic embeddings of works. The learned embeddings then initialize the BERT model.

BERT Model and Pre-training: After we obtain word embeddings for clinical notes in different periods, we learn clinical note embeddings (a single embedding for each clinical note) by passing the sequence of learned word embeddings through the BERT architecture. However, contrary to the original architecture, we only minimize the Masked Language Model Loss. This is because we have removed all punctuation marks from the clinical notes during our pre-processing step to learn word embeddings. The Masked Language Loss is given by:

$$\mathcal{L}_{LM} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \tag{1}$$

Here, y denotes the true binary label of the word, and p denotes the predicted likelihood of the word given by the BERT model. N denotes the total number of samples.

2.2 Construction of Dynamic Patient Embeddings

We design a pair of co-evolving neural networks for each type of entity (doctor D, medication M, or room R) with whom a patient interacts in the healthcare facility. Let $\mathcal{PR}$ denote a set of patient p - doctor d interactions (p, d, t), $\mathcal{MD}$ denote a set of patient p - medication m interactions (p, m, t), and $\mathcal{TR}$ denote a set of patient p - room r interactions (p, r, t). Let GM_p and GM_e denote co-evolving neural networks that update patient p's embedding and entity e's embedding, respectively. Here, $\mathrm{GM} \in \{\mathrm{PM}, \mathrm{MM}, \mathrm{TM}\}$, where PM, MM, and TM denote modules for $\mathcal{PR}$, $\mathcal{MD}$, and $\mathcal{TR}$, respectively.

When a patient p interacts with a hospital entity e at time t, we simultaneously update the patient's embedding $\hat{\mathbf{e}}_{p,t}$ and the entity's embedding $\hat{\mathbf{e}}_{e,t}$ with GM_p and GM_e. Specifically, we use their dynamic embeddings at t^-, that is $\hat{\mathbf{e}}_{p,t^-}$ and $\hat{\mathbf{e}}_{e,t^-}$, which is just before time t. We also use patient p's static and dynamic features $\mathbf{p}_p$ and $\hat{\mathbf{p}}_{p,t}$, respectively, to update both $\hat{\mathbf{e}}_{p,t}$ and $\hat{\mathbf{e}}_{e,t}$. Finally, for GM_p, we use the time elapsed from the patient p's previous interaction $\Delta_{p,t}$, and for GM_e, we compute time elapsed from the entity e's previous interaction $\Delta_{e,t}$. Here are the update equations for GM_p and GM_e:

$$\begin{aligned}\hat{\mathbf{e}}_{p,t} &= \sigma\left[\mathbf{W}_p^{\mathrm{GM}}[\hat{\mathbf{e}}_{p,t^-} \mid \hat{\mathbf{e}}_{e,t^-} \mid \Delta_{p,t} \mid \mathbf{p}_p \mid \hat{\mathbf{p}}_{p,t}] + \mathbf{B}_p^{\mathrm{GM}}\right] \\ \hat{\mathbf{e}}_{e,t} &= \sigma\left[\mathbf{W}_e^{\mathrm{GM}}[\hat{\mathbf{e}}_{e,t^-} \mid \hat{\mathbf{e}}_{p,t^-} \mid \Delta_{e,t} \mid \mathbf{p}_p \mid \hat{\mathbf{p}}_{p,t}] + \mathbf{B}_d^{\mathrm{GM}}\right]\end{aligned} \tag{2}$$

We use the symbol | to denote vector concatenation. σ is a non-linear activation function (e.g., tanh activation). $\mathbf{W}_p^{\mathrm{GM}}$ and $\mathbf{W}_e^{\mathrm{GM}}$ denote weight matrices that parameterize GM_p and GM_e, respectively, for $\text{GM} \in \{\text{PM}, \text{MM}, \text{TM}\}$. $\mathbf{B}_p^{\text{GM}}$ and $\mathbf{B}_e^{\text{GM}}$ are bias for GM_p and GM_e, respectively.

Notice that a patient p's embedding at time t^-, that is $\hat{\mathbf{e}}_{p,t^-}$, can be quite different from $\hat{\mathbf{e}}_{p,t-\Delta_{p,t}}$ if $\Delta_{p,t}$ is somewhat large. Our model handles this using projection operation [19]. Specifically, we project the p's embedding from time $t - \Delta_{p,t}$ to t^-.

$$\hat{\mathbf{e}}_{p,t^-} = (1 + \mathbf{W} \times \Delta_{p,t}) + \hat{\mathbf{e}}_{p,t-\Delta_{p,t}} \tag{3}$$

where $\mathbf{W}$ is a linear weight matrix.

Furthermore, we preserve the information on each patient p's interaction with other entity e at time t^- by reconstructing the concatenation of e's dynamic embedding $\hat{\mathbf{e}}_{e,t^-}$ and static embedding $\bar{\mathbf{e}}_e$, that is of size $|\hat{\mathbf{e}}_{e,t^-}| + |\bar{\mathbf{e}}|$. To reconstruct $\tilde{\mathbf{e}}_{e,t^-}$, we use $\bar{\mathbf{e}}_e$, $\hat{\mathbf{e}}_{e,t^-}$ as well as patient p's information, such as p's static features $\mathbf{p}_p$, static embedding $\bar{\mathbf{e}}_p$ and dynamic embedding $\hat{\mathbf{e}}_{p,t^-}$. Note that we design a reconstruction module for each entity $\mathit{RECONST}_{\text{D}}$, $\mathit{RECONST}_{\text{M}}$, and $\mathit{RECONST}_{\text{R}}$ for doctor, medication, and room, respectively. We define $\mathit{RECONST}_{\text{E}}$ for $E \in \{D, M, R\}$:

$$\tilde{\mathbf{e}}_{e,t^-} = \mathbf{W}_e \left[\hat{\mathbf{e}}_{p,t^-} \mid \bar{\mathbf{e}}_p \mid \mathbf{p}_p \mid \hat{\mathbf{e}}_{e,t^-} \mid \bar{\mathbf{e}}_e\right] + \mathbf{B}_e \tag{4}$$

$\mathbf{W}_e$ and $\mathbf{B}_e$ are weight matrix and bias for $\mathit{RECONST}_{\text{E}}$.

2.3 Co- evolution with Clinical Notes

In addition to constructing dynamic patient embeddings based on doctor, medication, and room interactions, we add information from clinical notes by incorporating an evolving neural network architecture to update dynamic patient embeddings. Let $\mathcal{NM}$ denote a set of patient p - clinical note n interactions (p,n,t). Contrary to other types of hospital entity interactions, clinical note interactions only update patient interactions and not vice-versa.

When patient p gets a clinical note n at time t, we obtain note embedding $e_{n,t}$ from the modified BERT as mentioned in Section 2.1 We refine the latent space of BERT by minimizing the Masked Language Model Loss $\mathcal{L}_{LM}$. In addition, we obtain the note embeddings through a fine-tuning Feed-Forward layer. After obtaining the note embedding, we update the dynamic embedding of p via a neural network NM_p:

$$\hat{\mathbf{e}}_{p,t} = \sigma\left[\mathbf{W}_p^{NM}[\hat{e}_{p,t^-} \mid e_{n,t} \mid \Delta_{p,t} \mid p_p \mid \hat{p}_{p,t}]\right] \tag{5}$$

where $\mathbf{W}_p^{NM}$ denotes the weight matrix for NM_p and $e_{n,t}$ denotes the clinical note embedding obtained from BERT. Note that the clinical notes co-evolution is jointly trained with other co-evolving networks.

2.4 Continual- Learning Framework

We elaborate on the continual learning framework used to train the model on data that appears in periods. Let $\{Period_1, \cdots, Period_n\}$ and $\{\theta_1, \cdots, \theta_n\}$ denote a set of periods and the overall model parameters, respectively. We train θ_1 as described in Section 2.2.

Given the model parameters θ_1 generated from $Period_1$ (via constructing dynamic patient features), we propose to incrementally account for the model parameters of the successive periods by initializing the model parameters θ_n of $Period_n$ with θ_{n-1} of $Period_{n-1}$. The initialization scheme is motivated by a similar idea given in [16] which aligns learned embeddings in the unified coordinate space. We do the same to the model parameters of the subsequent periods to enable continual knowledge infusion from previous periods and reduce the time needed for retraining the model parameters for data that appears in a new period.

However, even though this scheme works in reducing training time and resources, a critical issue can be the phenomenon called "catastrophic forgetting", a term that was first coined in [18]. As the model is trained on the data from subsequent periods, the embedding space of the model parameters may become distorted and the model might forget information that it had learned earlier. So, similar to the method proposed in [15], we minimize the variations of the model parameters by introducing an additional loss called continual loss ($\mathcal{L}_{CL}$) which minimizes the L2- norm between the model parameters of the subsequent periods. The formula is given below:

$$\mathcal{L}_{CL} = \lambda||\theta_i - \theta_{i-1}||_2 \tag{6}$$

where λ is a regularization hyperparameter.

2.5 Losses and Overall Training Scheme

In addition to the loss functions mentioned before, we also use the below-mentioned losses in our overall heterogenous co-evolving network architecture. They are as follows:

Reconstruction Loss: This loss computes the difference between the predicted and the ground truth embeddings of the entity a patient interacts with. It is written as:

$$\begin{aligned}\mathcal{L}_{reconstruct} = &\sum_{(p,d,t)\in PR} \left\|\tilde{\mathbf{e}}_{d,t^-} - \left[\hat{\mathbf{e}}_{d,t^-} \mid \bar{e}_d\right]\right\|_2 \\ &+ \sum_{(p,m,t)\in MD} \left\|\tilde{\mathbf{e}}_{m,t^-} - \left[\hat{\mathbf{e}}_{m,t^-} \mid \bar{e}_m\right]\right\|_2 \\ &+ \sum_{(p,r,t)\in TR} \left\|\tilde{\mathbf{e}}_{r,t^-} - \left[\hat{\mathbf{e}}_{r,t^-} \mid \bar{e}_r\right]\right\|_2\end{aligned} \tag{7}$$

Temporal Consistency Loss: It is the L_2 norm of the difference between the embeddings of each entity between each consecutive interaction. The equation is:

$$\mathcal{L}_{temp} = \sum_{(p,e,t)\in S} ||\hat{e}_{p,t} - \hat{e}_{p,t^-}||_2 + ||\hat{e}_{e,t} - \hat{e}_{e,t^-}||_2 \tag{8}$$

Overall Loss: The overall loss is represented as the sum of the previous losses. After pre-training them individually for the first period, we jointly train the heterogenous co-evolving networks and the BERT. In the subsequent periods, we only do the joint training. Our overall loss formulation is as follows:

$$\mathcal{L} = \mathcal{L}_{reconstruct} + \mathcal{L}_{temp} + \mathcal{L}_{LM} + \mathcal{L}_{CL} \tag{9}$$

We optimize the overall loss using the Adam optimization algorithm [17]. We use the Adam optimizer with the learning rate of 1e-3 and the weight decay of 1e-5. The size of the dynamic embeddings is set to 128. The overall training schema is given in Algorithm 1.

Algorithm 1 Training scheme for our proposed model

```
Require: $\mathcal{PR}$, $\mathcal{TR}$, $\mathcal{MD}$, $\mathcal{NM}$, $p_p$, $\hat{p}_{p,t}$, period number ($period$)
  out of a total of $k$ periods, and number of epochs $E$.
  if $period$==1 then
      Initialize word embeddings with Dynamic Word2Vec vector
      Pre-train BERT on Eqn (1)
      $\mathcal{L}_{\mathcal{LM}} = 0$ and $\mathcal{L}_{CL} = 0$
      Pre-train co-evolving Neural Network architecture on $\mathcal{PR}$,
  $\mathcal{TR}$, $\mathcal{MD}$ by using Eqns (2-4)
      Minimize $\mathcal{L}$ in Eqn (9)
  end if
  if $period > 1$ then
      $\theta_{period} = \theta_{period-1}$
  end if
  $e = 1$
  while $e \leq E$ do
      Perform Joint training on $\mathcal{PR}$, $\mathcal{TR}$, $\mathcal{MD}$ and $\mathcal{NM}$ with Eqns
  (1-8).
      if $period > 1$ then
          Compute $\mathcal{L}_{CL}$ by Eqn (6)
      else if $period$==1 then
          $\mathcal{L}_{CL} = 0$
      end if
      Minimize $\mathcal{L}$ in Eqn (9)
      Save model parameters
      $e = e + 1$
  end while
```

3 EXPERIMENT

We provide code for academic purposes [1]. We conducted experiments on AMD EPYC 7763 64-Core Processor with 2 TB memory and on 8 NVIDIA A30 GPUs.

Data: The dataset is collected from the University of Iowa Hospitals and Clinics (UIHC), which is a large (800-bed) tertiary care teaching hospital in Iowa City, Iowa. The dataset consists of de-identified Electronic Healthcare Records (EHR) and admission-discharge-transfer (ADT) records on patients. Each patient visit has a set of diagnoses, a timestamped record of room transfers and procedures performed by physicians, prescribed medications, and clinical notes. We extract patient interactions with medications, doctors, and rooms, along with the clinical notes written for the patient in 2008, between May 04 to August 31. We split the data into three overlapping periods of time chunks:

Period 1: Between May 04 and June 25. There are 245,043 doctor, medication, and room interactions and 149,685 clinical notes.

Period 2: Between June 13 and August 07. There are 252,089 doctor, medication, and room interactions and 152,037 clinical notes.

Period 3: Between July 10 and August 31. There are 257,994 doctor, medication, and room interactions and 163,158 clinical notes.

We compare the performance of our method with state-of-the-art methods in all applications.

- DOMAIN: In healthcare analytics, features are often handcrafted based on domain knowledge. We handcraft features from the domain-based methods for each application [22, 25].
- JODIE: This is an exemplary co-evolutionary neural network, which learns embeddings over time from the stream of interactions, and the learned embeddings are shown to outperform in predictive modeling tasks. We train patient embeddings using JODIE [19] with the stream of patient interactions.
- DECEnt: This is another co-evolutionary neural network that considers the heterogeneity in the interactions, designed to learn dynamic patient embeddings. DECEnt has shown to perform well in healthcare predictive modeling tasks [12].

3.1 Evaluation of the Continual Learning Framework

Our motivation for incorporating a continually- adaptive representation learning framework into our architecture was to reduce both time and resources to train the entire framework from scratch as new data is available. To experimentally validate our motivation, we compute the total number of Multiply–Accumulate Operations (MACs) which were required to train our proposed model architecture continually. The results are shown in Table 1.

In UIHC, we notice that the number of MACs drops by 68.40 % from Period 1 to Period 2 and by 65.83 % from Period 2 to Period 3 for loss convergence. This validates the utility of the continual adaptation present in the model architecture which leads to the formation of a scaleable lifelong-learning model.

3.2 Application: CDI Incidence Prediction

Clostridioides Difficile Infection (CDI) is an HAI, that can lead to severe health outcomes once an immunocompromised patient gets infected with it. Due to this reason, healthcare facilities are keen to prevent the spread of CDI.

We design the CDI prediction as a binary classification problem. The embedding of a CDI patient is taken three days before the patient's positive test date [8, 13]. This was to ensure no data leakage due to the potential treatment given to patients for treating severe diarrhea [27]. The embedding of a non-CDI patient is selected randomly from their stay at the hospital. Note that getting CDI is a rare event, for which we have about 150:1 class imbalance during the period when the data was collected.

Table 2 shows the prediction results on each method, tested on three periods over time, on three classifiers logistic regression (LR), support vector machines (SVM), and random forest (RF). Notice that our method performs consistently better than the baselines in all the periods, regardless of the classifier that we use. We observe

Table 1: Flop counts for datasets using our method in a continual setting. Note that as the best model parameters from the previous period are used, convergence is much faster, thus reducing the number of MACs in the subsequent periods.

Dataset	Period	Number of MACs (In G)
UIHC	Period 1	492,772.60
	Period 2	155,706.32
	Period 3	53,189.76

[1] https://github.com/Soothysay/CL-EHR

Table 2: ROC-AUC Scores for CDI Incidence Prediction for UIHC. We perform 3- fold cross-validation with 30 repetitions. Note that our proposed method outperforms all the baseline methods for all classifiers.

Period	Method	LR	SVM	RF
Period 1	DOMAIN	0.49 ± 0.20	0.52 ± 0.07	0.34 ± 0.07
	JODIE	0.44 ± 0.12	0.36 ± 0.09	0.52 ± 0.03
	DECEnt	0.62 ± 0.07	0.57 ± 0.01	0.61 ± 0.06
	Ours	**0.65 ± 0.05**	**0.60 ± 0.04**	**0.73 ± 0.07**
Period 2	DOMAIN	0.60 ± 0.11	0.54 ± 0.13	0.76 ± 0.19
	JODIE	0.50 ± 0.05	0.47 ± 0.06	0.52 ± 0.18
	DECEnt	0.71 ± 0.02	0.59 ± 0.16	0.77 ± 0.04
	Ours	**0.74 ± 0.08**	**0.62 ± 0.06**	**0.78 ± 0.19**
Period 3	DOMAIN	0.67 ± 0.19	0.56 ± 0.09	0.71 ± 0.18
	JODIE	0.61 ± 0.08	0.55 ± 0.14	0.59 ± 0.03
	DECEnt	0.68 ± 0.12	0.63 ± 0.04	0.71 ± 0.19
	Ours	**0.69 ± 0.14**	**0.66 ± 0.07**	**0.72 ± 0.23**

Table 3: ROC-AUC Scores for MICU Transfer Prediction for UIHC. We perform 3- fold cross-validation with 30 repetitions. Note that our proposed method outperforms most of the baseline methods.

Period	Method	LR	SVM	RF
Period 1	DOMAIN	0.63 ± 0.20	0.52 ± 0.03	0.86 ± 0.13
	JODIE	0.54 ± 0.15	0.51 ± 0.02	0.66 ± 0.04
	DECEnt	0.85 ± 0.07	0.71 ± 0.05	0.83 ± 0.05
	Ours	**0.89 ± 0.05**	**0.77 ± 0.08**	**0.87 ± 0.03**
Period 2	DOMAIN	0.68 ± 0.12	0.57 ± 0.13	0.71 ± 0.07
	JODIE	0.59 ± 0.05	0.52 ± 0.10	0.55 ± 0.01
	DECEnt	0.72 ± 0.07	0.65 ± 0.10	0.86 ± 0.03
	Ours	**0.76 ± 0.02**	**0.72 ± 0.03**	**0.89 ± 0.09**
Period 3	DOMAIN	0.67 ± 0.13	0.56 ± 0.02	0.81 ± 0.03
	JODIE	0.61 ± 0.08	0.52 ± 0.18	0.62 ± 0.12
	DECEnt	**0.85 ± 0.07**	0.67 ± 0.01	**0.87 ± 0.18**
	Ours	0.84 ± 0.12	**0.71 ± 0.01**	**0.87 ± 0.08**

a gain of up to 19.7 % in Period 1 when compared to the next best-performing baseline method, DECEnt. Notice that DECEnt learns embeddings from heterogeneous interactions but not using the clinical notes, which highlights the importance of clinical notes for learning patient embeddings.

3.3 Application: MICU Transfer Prediction

Some hospitalized patients get transferred to MICU when there is a need for intensive care and continuous patient monitoring. Such an event may indicate a deterioration of care; hence detecting patients' risk of being transferred to MICU beforehand may help HCPs to better care for high-risk patients. Furthermore, predicting such patients would help hospital officials to better allocate hospital resources over time.

Similar to Section 3.2, We design the MICU transfer prediction as a binary classification problem. Inpatients that get transferred to MICU are positive instances. From them, we take the embedding one day before the MICU transfer event. For the remaining inpatients (aka, negative instances), we randomly sample the embedding from their hospital stay. The MICU transfer prediction task is also a rare event of a class imbalance of about 100:1.

Table 3 shows the results of MICU transfer prediction. Here, our method outperforms all the other methods in Period 1 and Period 2, with a gain of up to 3.5 %. Notice that 3.5 % gain is impressive since the resources are scarce in MICU, and hence this could have help HCPs to better utilize the limited resources and hence lead to saving patients' lives. We observe a comparable performance in Period 3 with DECEnt.

4 CONCLUSION

This work proposes a novel framework to learn patient embeddings over time for time-sensitive healthcare applications. The learned embeddings incorporate both the interactions and the clinical notes. We use continual learning to reduce the time for training incoming batches of interactions and notes. For each batch of interactions, we jointly train the heterogeneous co-evolving networks with clinical notes and refine the latent space of BERT. We show that our model outperforms all state-of-the-art baselines in predictive modeling tasks, such as MICU transfer and CDI incidence prediction.

5 RELATED WORK

LLM for clinical notes Recently, biomedical communities have adapted large language models (LLMs), such as BERT [7], to learn to embed clinical notes. BioBERT initializes with general BERT weights then use PMC full-text articles and PubMed abstracts to pre-train their model [21].

Continual learning In continual learning, various methods were developed to combat catastrophic forgetting [18]. iCARL stores a subset of samples per class that best approximates class means and re-uses them in training new batches [31]. Elastic Weight Consolidation (EWC) estimates the importance of neural network parameters, then penalizes if there are changes made to important parameters [18]. Some other methods, such as progressive networks, instantiate new branches for new tasks but enable knowledge transfer via lateral connections [32].

Healthcare analytics Various predictive modeling tasks are considered in Healthcare Analytics, such as mortality prediction [34] or CDI prediction [22], that leverage electronic health records. Some other works utilize patient mobility logs to solve inference problems, such as outbreak detection [2], missing infection [14, 36]. The role of the architectural layout of the hospital is also explored [11]. Other methods learn patient embeddings. DECEnt uses heterogeneous co-evolving networks [12], whereas MiME utilizes multilevel structure of EHR data [5].

6 ACKNOWLEDGEMENTS

This project is partially funded by the CDC MInD Healthcare Network grant U01CK000594 and the associated COVID-19 supplemental funding. K. Jha's research is partially supported by the University of Iowa Office of Vice President for Research (OVPR) Early Career Scholars Award. The authors acknowledge feedback from other University of Iowa CompEpi group members.

REFERENCES

[1] Accountability Act. 1996. Health insurance portability and accountability act of 1996. *Public law* 104 (1996), 191.

[2] Bijaya Adhikari, Bryan Lewis, Anil Vullikanti, José Mauricio Jiménez, and B Aditya Prakash. 2019. Fast and near-optimal monitoring for healthcare acquired infection outbreaks. *PLoS CompBio* (2019).

[3] Felipe Almeida and Geraldo Xexéo. 2019. Word embeddings: A survey. *arXiv preprint arXiv:1901.09069* (2019).

[4] Fatemeh Amrollahi, Supreeth P Shashikumar, Andre L Holder, and Shamim Nemati. 2022. Leveraging clinical data across healthcare institutions for continual learning of predictive risk models. *Scientific Reports* 12, 1 (2022), 8380.

[5] Edward Choi, Cao Xiao, Walter F. Stewart, and Jimeng Sun. 2018. MiME: Multilevel Medical Embedding of Electronic Health Records for Predictive Healthcare. arXiv:1810.09593

[6] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. 2021. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence* 44, 7 (2021), 3366–3385.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[8] Erik R Dubberke, Kimberly A Reske, Margaret A Olsen, Kathleen M McMullen, Jennie L Mayfield, L Clifford McDonald, and Victoria J Fraser. 2007. Evaluation of Clostridium difficile–associated disease pressure as a risk factor for C difficile–associated disease. *Archives of internal medicine* 167, 10 (2007), 1092–1097.

[9] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision* 129 (2021), 1789–1819.

[10] Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus.. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*. Association for Computational Linguistics, Edinburgh, UK, 67–71. https://aclanthology.org/W11-2508

[11] Hankyu Jang, Samuel Justice, Philip M Polgreen, Alberto M Segre, Daniel K Sewell, and Sriram V Pemmaraju. 2019. Evaluating Architectural Changes to Alter Pathogen Dynamics in a Dialysis Unit. In *IEEE/ACM ASONAM*.

[12] Hankyu Jang, Sulyun Lee, DM Hasibul Hasan, Philip M Polgreen, Sriram V Pemmaraju, and Bijaya Adhikari. 2022. Dynamic Healthcare Embeddings for Improving Patient Care. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 52–59.

[13] Hankyu Jang, Sulyun Lee, DM Hasibul Hasan, Philip M Polgreen, Sriram V Pemmaraju, and Bijaya Adhikari. 2022. Dynamic Healthcare Embeddings for Improving Patient Care. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 52–59.

[14] Hankyu Jang, Shreyas Pai, Bijaya Adhikari, and Sriram V Pemmaraju. 2022. Risk-aware temporal cascade reconstruction to detect asymptomatic cases. *Knowledge and Information Systems* 64, 12 (2022), 3373–3399.

[15] Kishlay Jha and Aidong Zhang. 2021. Continual knowledge infusion into pre-trained biomedical language models. *Bioinformatics* 38, 2 (09 2021), 494–502. https://doi.org/10.1093/bioinformatics/btab671 arXiv:https://academic.oup.com/bioinformatics/article-pdf/38/2/494/49007514/btab671.pdf

[16] Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal Analysis of Language through Neural Language Models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. Association for Computational Linguistics, Baltimore, MD, USA, 61–65. https://doi.org/10.3115/v1/W14-2517

[17] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG]

[18] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences* 114, 13 (2017), 3521–3526. https://doi.org/10.1073/pnas.1611835114 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1611835114

[19] Srijan Kumar, Xikun Zhang, and Jure Leskovec. 2019. Predicting dynamic embedding trajectory in temporal interaction networks. In *ACM SIGKDD*.

[20] Isotta Landi, Benjamin S Glicksberg, Hao-Chih Lee, Sarah Cherng, Giulia Landi, Matteo Danieletto, Joel T Dudley, Cesare Furlanello, and Riccardo Miotto. 2020. Deep representation learning of electronic health records to unlock patient stratification at scale. *NPJ digital medicine* 3, 1 (2020), 96.

[21] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.

[22] Benjamin Y Li, Jeeheh Oh, Vincent B Young, Krishna Rao, and Jenna Wiens. 2019. Using machine learning and the electronic health record to predict complicated Clostridium difficile infection. In *OFID*.

[23] Jingyi Liu and Yixiang Duan. 2012. Saliva: a potential media for disease diagnostics and monitoring. *Oral oncology* 48, 7 (2012), 569–577.

[24] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[25] Xu Min, Bin Yu, and Fei Wang. 2019. Predictive modeling of the hospital readmission risk from patients' claims data using machine learning: a case study on COPD. *Sci. Rep.* (2019).

[26] Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. 2016. Pruning convolutional neural networks for resource efficient inference. *arXiv preprint arXiv:1611.06440* (2016).

[27] Mauricio Monsalve, Sriram Pemmaraju, Sarah Johnson, and Philip M Polgreen. 2015. Improving risk prediction of Clostridium difficile infection using temporal event-pairs. In *IEEE ICHI*.

[28] Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474* (2019).

[29] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[30] Oleg S Pianykh, Steven Guitron, Darren Parke, Chengzhao Zhang, Pari Pandharipande, James Brink, and Daniel Rosenthal. 2020. Improving healthcare operations management with machine learning. *Nature Machine Intelligence* 2, 5 (2020), 266–273.

[31] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2001–2010.

[32] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. 2016. Progressive neural networks. *arXiv preprint arXiv:1606.04671* (2016).

[33] Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2012. *Tracing semantic change with Latent Semantic Analysis*. De Gruyter Mouton, Berlin, Boston, 161–183. https://doi.org/doi:10.1515/9783110252903.161

[34] Eli Sherman, Hitinder Gurm, Ulysses Balis, Scott Owens, and Jenna Wiens. 2017. Leveraging clinical time-series data for prediction: a cautionary tale. In *AMIA*.

[35] Yuqi Si and Kirk Roberts. 2019. Deep patient representation of clinical notes via multi-task learning for mortality prediction. *AMIA Summits on Translational Science Proceedings* 2019 (2019), 779.

[36] Shashidhar Sundareisan, Jilles Vreeken, and B Aditya Prakash. 2015. Hidden hazards: Finding missing nodes in large graph epidemics. In *SDM*.

[37] Li Wang, Qinghua Wang, Heming Bai, Cong Liu, Wei Liu, Yuanpeng Zhang, Lei Jiang, Huji Xu, Kai Wang, and Yunyun Zhou. 2020. EHR2Vec: representation learning of medical concepts from temporal patterns of clinical notes based on self-attention mechanism. *Frontiers in Genetics* 11 (2020), 630.

[38] Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic Word Embeddings for Evolving Semantic Discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (Marina Del Rey, CA, USA) *(WSDM '18)*. Association for Computing Machinery, New York, NY, USA, 673–681. https://doi.org/10.1145/3159652.3159703