Contrastive Learning for Clinical Outcome Prediction with Partial Data Sources

Meng Xia [1] **Jonathan Wilson** [2] **Benjamin Goldstein** [2] **Ricardo Henao** [1 2 3]

Abstract

The use of machine learning models to predict clinical outcomes from (longitudinal) electronic health record (EHR) data is becoming increasingly popular due to advances in deep architectures, representation learning, and the growing availability of large EHR datasets. Existing models generally assume access to the same data sources during both training and inference stages. However, this assumption is often challenged by the fact that real-world clinical datasets originate from various data sources (with distinct sets of covariates), which though can be available for training (in a research or retrospective setting), are more realistically only partially available (a subset of such sets) for inference when deployed. So motivated, we introduce Contrastive Learning for clinical Outcome Prediction with Partial data Sources (CLOPPS), that trains encoders to capture information across different data sources and then leverages them to build classifiers restricting access to a single data source. This approach can be used with existing cross-sectional or longitudinal outcome classification models. We present experiments on two real-world datasets demonstrating that CLOPPS consistently outperforms strong baselines in several practical scenarios.

1. Introduction

In recent times, a growing number of healthcare institutions have started to routinely collect and leverage electronic health records (EHR) from large collections of patients, resulting in the availability of vast, rich, multimodal and longitudinal EHR datasets. The promise is that these data sources, if harnessed effectively, have the potential to substantially improve the delivery and quality of healthcare. Given the proven success of machine learning algorithms in various fields such as image classification (Simonyan & Zisserman, 2014; He et al., 2016; Dosovitskiy et al., 2020), object detection (Redmon et al., 2016; He et al., 2017; Carion et al., 2020), and natural language processing (Devlin et al., 2018; Brown et al., 2020; Zhang et al., 2022a; Touvron et al., 2023), numerous studies have explored the application of these algorithms to EHR datasets for a variety of medical tasks, including disease detection (Choi et al., 2017; Baumel et al., 2018), diagnosis (Doctor; Lipton et al., 2015) and prognosis (Harutyunyan et al., 2019; Song et al., 2018).

Prediction of clinical outcomes, *e.g.*, mortality, has emerged as a crucial focus due to two main reasons, namely, i) it enables healthcare providers to tailor interventions effectively, improving patient care; and ii) it promises to improve the allocation healthcare resources, for instance, intensive care units (ICUs), staff, procedures and treatments; all of which aid in optimizing the overall healthcare delivery process. Initially, traditional machine learning algorithms were used for mortality prediction tasks (Lemeshow et al., 1994; Crawford et al., 2000; Meyfroidt et al., 2009). Later, researchers leveraged the longitudinal nature of EHRs to develop sequence-based models such as LSTMs (Hochreiter & Schmidhuber, 1997) and Transformers (Vaswani et al., 2017), for more granular (over time) and accurate outcome predictions (Harutyunyan et al., 2019; Song et al., 2018).

Most outcome prediction models share a common assumption requiring the access to identical (clinical) data sources during both training and inference stages. However, this assumption often does not align with the realities of real-world healthcare settings. Common examples are those in which data is collected by different parties (*e.g.*, EHR and claims) or by different systems (*e.g.*, EHR and picture archiving systems). Moreover, there are situations where a richer (*e.g.*, medical narrative) or more complete (*e.g.*, claims) data sources are available for model training but not at inference due to security, privacy, readiness or portability constraints.

To address this important challenge and maximize the utility of complete data sources for the purpose of enhancing prediction of outcomes in clinical settings, we introduce Contrastive Learning for clinical Outcome Prediction with Partial data Sources (CLOPPS). Specifically, we learn encoders

[1]Department of Electrical and Computer Engineering, Duke University, Durham, US [2]Department of Biostatistics and Bioinformatics, Duke University, Durham, US [3]King Abdullah University of Science and Technology, Thuwal, KSA. Correspondence to: Meng Xia <mx41@duke.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

set to preserve information across distinct data sources, but then use them with a single data source during inference, as part of a cross-sectional or longitudinal outcome classification model. In this manner, we can take advantage of the availability of multiple data sources during model development, but without requiring all of them to be available during inference. This is motivated by a real-world scenario (see experiments) where we have data from both a dialysis provider–Dialysis Clinic, Inc (DCI)–and a national data system–United States Renal Data System (USRDS) (U.S. Renal Data System)–but we are interested in a model that can be used by the provider without relying on the less accessible and non-real-time data retrospectively collected by the national system.

Our work offers two main contributions. From a methodological perspective, CLOPPS allows to build representations from multiple data sources that can be leveraged to improve the performance of prediction models using only one data source. To the best of our knowledge, CLOPPS is the first framework developed to tackle this unique challenge that is prevalent in clinical applications. From a practical perspective, we demonstrate CLOPPS on several practical scenarios with two real-world datasets from electronic health records used for the purpose of mortality prediction.

2. Related Work

Sequence Models Architectures specifically designed to model sequences (including time series) have demonstrated exceptional performance across various tasks, including language translation (Devlin et al., 2018; Zhang et al., 2022a; Touvron et al., 2023), weather forecasting (Doreswamy et al., 2018; Hewage et al., 2021) and mortality prediction (Karabacak & Margetis, 2023; Nunez et al., 2023). One of the earliest approaches is the recurrent neural network (RNN) (Rumelhart et al., 1985), which was then enhanced by the long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997). A significant advancement came with the introduction of the Transformer (Vaswani et al., 2017), which utilizes a self-attention mechanism and shows better performance than traditional sequence models when used to encode sequential data. This led to the development of more specialized Transformer models, *e.g.,* BERT (Devlin et al., 2018), OPT (Zhang et al., 2022a), and RoFormer (Su et al., 2024). Recently, to combine the advantages of both RNNs and Transformer models, Peng et al. (2023) proposed RWKV; a RNN approach leveraging self-attention that matches the performance of Transformer models.

Longitudinal prediction of outcomes Initial approaches mostly concerned with mortality prediction exclusively employed traditional machine learning algorithms, such as logistic regression (Lemeshow et al., 1994; Wagner et al., 1994; Vincent & Singer, 2010), decision trees (Crawford et al., 2000; Ramon et al., 2007; Ribas et al., 2011), and support vector machines (Meyfroidt et al., 2009; Citi & Barbieri, 2012; Kim et al., 2011). With the grow in popularity of neural networks, researchers began exploring their application in this domain (Dybowski et al., 1996; Clermont et al., 2001; Nimgaonkar & Sudarshan, 2004). However, some studies indicated that neural networks performed similarly to traditional algorithms in various ICU outcome prediction settings (Doig et al., 1993; Wong & Young, 1999; Clermont et al., 2001). Acknowledging the longitudinal nature of patient data, there was a shift toward employing such data for improved prediction performance. For instance, Harutyunyan et al. (2019) leveraged LSTMs to process medical time-series data for ICU mortality prediction. Similarly, Song et al. (2018) developed the SAnD (Simply Attend and Diagnose) architecture utilizing an attention. Further, Rajkomar et al. (2018) used both LSTM and attention-based models to enhance the accuracy of mortality predictions. Unlike traditional outcome prediction models where full access to data sources is assumed, we address the more challenging scenario where access during inference is partial.

Contrastive learning The core insight of contrastive learning is to learn effective representations by contrasting positive pairs against negative pairs without the need for expertly-acquired labels. Initially perceived as a form of self-supervised learning, contrastive learning has seen diverse applications. For instance, Chen et al. (2020) employed it to learn valuable visual representations, achieving results on par with supervised methods. Similarly, Logeswaran & Lee (2018) utilized it to enhance semantic text understanding. The concept of contrastive learning has expanded into multimodal scenarios, such as Radford et al. (2021) using it to learn joint image-text representations, enabling the model to comprehend a broad array of visual concepts in a zero-shot manner. Recent studies have explored the application of contrastive learning to longitudinal data. For instance, Schneider et al. (2021) leveraged contrastive learning to simulate an infant's learning experiences with a image sequences datasets, where the resulting object representations bear similarities to established neurobiological findings. In another instance, Hong et al. (2022) introduced specific augmentation techniques for longitudinal data, enhancing survival analysis. Leveraging the success of contrastive learning for multi-modal data, we train our encoder to capture information across data sources, generating embeddings for inference when only one of the data sources is available.

Several studies (Franceschi et al., 2019; Tonekaboni et al., 2021; Eldele et al., 2021; Yèche et al., 2021; Kiyasseh et al., 2021; Zhang et al., 2022b; Raghu et al., 2023) have also tried to apply contrastive learning to medical time-series analysis and event prediction. However, our work differen-

tiates itself by addressing a unique challenge where access to data sources varies between the training and inference phases. For instance, while studies such as Tonekaboni et al. (2021); Yèche et al. (2021) utilize time windows to align patient observations for representation learning, they focus on representations within data sources. In contrast, our research focuses on addressing the challenge of information representation across data sources. More specifically, Tonekaboni et al. (2021) uses a discriminator (not contrastive learning) within a single source, and Yèche et al. (2021) does use contrastive learning, but our research assumes that during inference only one data source is available.

3. Approach

Below we describe our approach for longitudinal prediction of outcomes from incomplete data sources. In Section 3.1, we define the problem and modeling framework. Then, in Sections 3.2 and 3.3, respectively, we introduce the optimization objectives used to train the encoding and outcome prediction components of the proposed framework.

3.1. Problem Definition

Predicting from longitudinal data We consider the prediction of (future) outcomes in a longitudinal setting. More formally, a collection of historical data (covariates) for sample p (*e.g.*, a patient) observed at regular intervals is denoted as $x_p = \{x_{p1}, \ldots, x_{pt}, \ldots, x_{p,n_p}\}$, for time points $t_1 < \cdots < t < \cdots < t_{n_p} \leq t_p$, and where t_p represents the last time point at which sample p was observed. More precisely, t_p indicates the outcome or censoring time, if $e_p = 1$ or $e_p = 0$, respectively, hence e_p is an event indicator. Note also that we do not assume covariates are available at t_p, however, that may be the case in practice. It is also worth noting that the assumption of data being sampled at regular intervals (*i.e.*, $|t - t'| \approx \text{const.}$, for $t \neq t'$) is primarily based on the characteristics of the datasets used in our experiments, however, the framework proposed below is general and can be readily extended to longitudinal data with irregular sampling.

Having defined sample p, a dataset of N samples consisting of triplets (x_p, t_p, e_p), and denoted as $\mathcal{D} = \{(x_p, t_p, e_p)\}_{p=1}^{N}$ is used to estimate the risk for an outcome of interest, *e.g.*, mortality, $p(y_{pt} = 1 | \{x_{pj}\}_{j=1}^{t}) \in (0, 1)$ at time t, defined as the probability of a sample experiencing the outcome within a fixed *horizon* of future M time units (*e.g.*, days, moths or years) relative to time t. Note that time points for which y_{pt} cannot be ascertained due to censoring, *i.e.*, $t_p \in [t, t + M]$ for $e_p = 0$, which in general amounts to a small proportion of sample time points, are not used for learning (2.7% for the Private dataset in the experiments).

Predicting from incomplete data sources A key defining characteristic of the scenario we seek to address is that in which we have access to multiple sources of covariates, but only one of such is available during inference. Without loss of generality, we assume the complete set of features consists of two non-overlapping *data sources*, each containing a set of covariates, *i.e.*, $x_{pt} = (x_{pt}^{(1)}, x_{pt}^{(2)})$. Consequently, we are interested in building two models to estimate the outcome of interest $p(y_{pt} = 1 | \{x_{pj}^{(k)}\}_{j=1}^{t}, \{z_{pj}^{(k)}\}_{j=1}^{t})$, for $k \in \{1, 2\}$, and where $\{z_{pj}^{(k)}\}_{j=1}^{t}$ is a longitudinal (latent) representation for data source k aimed at capturing the information in *both* data sources. This implies that i) only one data source $\{x_{pj}^{(k)}\}_{j=1}^{t}$ is available during inference; and ii) both data sources are available only when building the representation model for $\{z_{pj}^{(k)}\}_{j=1}^{t}$. The hypothesis driving our formulation is that if there is common information to be leveraged from both sources to predict the outcome of interest, it will be captured by representation $\{z_{pj}^{(k)}\}_{j=1}^{t}$, thus we expect the model using it to outperform the alternative that only uses $\{x_{pj}^{(k)}\}_{j=1}^{t}$. It is important to note that the formulation above suggests that data sources need to be longitudinally paired, *i.e.*, both having the same time points, however, this can be relaxed as we will discuss below. Further, the assumption that data sources are non-overlapping can be also relaxed as we will show in the experiments.

A two-stage training framework In order to leverage the knowledge from the complete set of data sources, we train the outcome prediction model in two separate stages, namely, *longitudinal feature learning* and *supervised outcome learning*. In the longitudinal feature learning stage, two separate feature encoders $z_{pt}^{(k)} = f^{(k)}(x_{pt}^{(k)})$ for $k \in \{1, 2\}$ are learned to produce *encoded features* $z_{pt}^{(k)}$ from input data source $x_{pt}^{(k)}$ as a means to incorporate the knowledge from both data sources in a longitudinal fashion. Subsequently, in the supervised outcome learning stage, classifiers for each data source $c^{(1)}(\cdot)$ and $c^{(2)}(\cdot)$ are learned to estimate the risk of the outcome of interest based on raw covariates $x_{pt}^{(1)}$ and $x_{pt}^{(2)}$, as well as the *fixed* encoded features $z_{pt}^{(1)}$ and $z_{pt}^{(2)}$, *i.e.*, without further refinement. The outcome prediction models can thus be formulated as

$$p(y_{pt} = 1 | \{x_{pj}^{(k)}\}_{j=1}^{t}) = c^{(k)}(\{z_{pj}^{(k)}\}_{j=1}^{t}, \{x_{pj}^{(k)}\}_{j=1}^{t}) \quad (1)$$

$$z_{pt}^{(k)} = f^{(k)}(x_{pt}^{(k)}), \quad (2)$$

where $k \in \{1, 2\}$. Note that the classifier in (1) takes both the raw covariates $\{x_{pj}^{(k)}\}_{j=1}^{t}$ and encoded features over time $\{z_{pj}^{(k)}\}_{j=1}^{t}$. Conceptually, the former can be seen as a *skip connection* into the classifier while the latter implicitly captures information from both data sources (recall only source k is available at inference). In practice, longitudinal covariates may not be available, thus we can use (1) in a

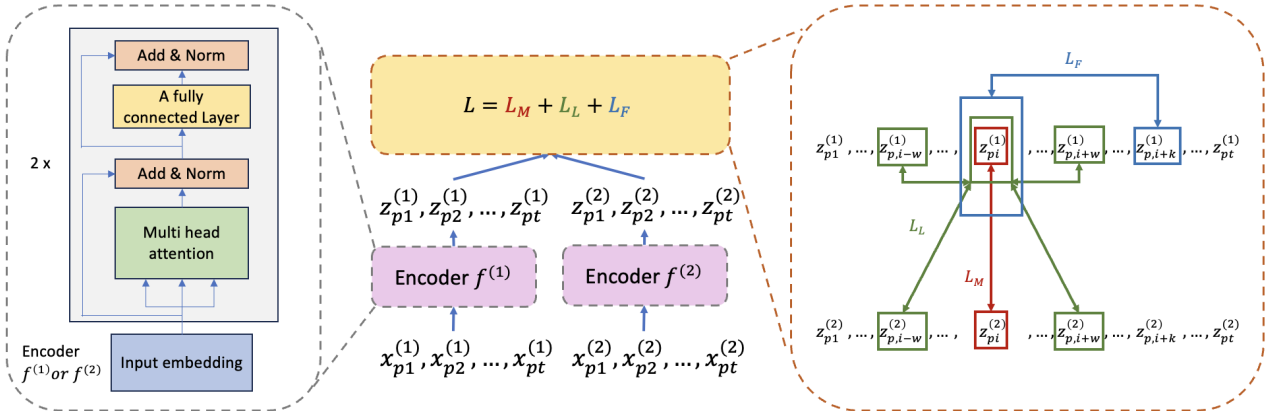

Figure 1. Pretraining of the encoders for CLOPPS. Given two longitudinal observations of sample $\{x_{pj}^{(1)}\}_{j=1}^{t}$ and $\{x_{pj}^{(2)}\}_{j=1}^{t}$, the attention-based encoder (left) produces representations $\{z_{pj}^{(1)}\}_{j=1}^{t}$ and $\{z_{pj}^{(2)}\}_{j=1}^{t}$. The learning objective consist of three components, $\mathcal{L}_M$, $\mathcal{L}_L$, and $\mathcal{L}_F$, that leverage positive pairs defined in terms of time matching, local similarity and future information, respectively.

cross-sectional manner using only $(z_{pt}^{(k)}, x_{pt}^{(k)})$ as inputs, as we will describe below and demonstrate in the experiments.

3.2. Longitudinal Feature Learning

The strategy used to learn the encoders is concisely illustrated in Figure 1 and described in detail below.

Information extraction via matching To produce encoders that remain effective when only one of the data sources is available during inference, we train them to align observations from the same sample and nearby time points across data sources in latent (representation) space. This approach is motivated by the intuition that, at any given time point t, covariates from different sources for sample p collectively and complementary represent their condition (*e.g.* health status) at that moment in time. Consequently, these covariates also likely contain shared (latent) information about the condition of p at time t. Training encoders to extract this shared information encourages that, when only one of the sources is available, such encoders can produce representations infused with valuable information from the data source that will not available at inference.

So motivated, we leverage contrastive learning (Chen et al., 2020; He et al., 2020; Khosla et al., 2020). Specifically, we randomly select a batch of B samples from $\mathcal{D}$, yielding a set of records $\mathcal{S}_B = \{\{x_{pt}^1, x_{pt}^2\}_{t=1}^{n_p}\}_{p=1}^{B}$. Within this batch, only pairs $x_{pt}^{(1)}$ and $x_{pt}^{(2)}$ are considered as a positive match, whereas all other (sample and time point) combinations deemed as negative pairs. Note that for situations in which data sources are not longitudinally concordant, the time matching for $x_{pt}^{(1)}$ and $x_{pt'}^{(2)}$ can be done for a predefined time window ϵ, *i.e.*, they are considered a positive match if $|t - t'| < \epsilon$. The training of encoders $f^{(1)}(\cdot)$ and $f^{(2)}(\cdot)$ is then conducted through a contrastive loss function in the resulting latent space, which encourages the encoders to distinguish between positive and negative pairings. Let $\text{sim}(z, z')$ be the cosine similarity between z and z', then the loss function for a positive pair $x_{pt}^{(1)}$ and $x_{pt}^{(2)}$ is

$$\ell(x_{pt}^{(k)}, x_{pt}^{(\backslash k)}) = - \log \frac{\exp(\text{sim}(z_{pt}^{(k)}, z_{pt}^{(\backslash k)})/\tau)}{\sum_{x_{ij}^{(o)} \sim \mathcal{S}_B \backslash x_{pt}^{(k)}} \exp(\text{sim}(z_{pt}^{(k)}, z_{ij}^{(o)})/\tau)}, \quad (3)$$

where $x_{pt}^{(\backslash k)}$ indicates covariates from the data source that is not $x_{pt}^{(k)}$, $x_{ij}^{(o)} \sim \mathcal{S}_B \backslash x_{pt}^{(k)}$ denotes an item from $\mathcal{S}_B$ excluding $x_{pt}^{(k)}$, $z_{pt}^{(k)}$ is obtained from the corresponding encoder via (2), and τ is the standard temperature parameter used in contrastive learning algorithms (Chen et al., 2020).

We can obtain the *time matching loss* for all samples in batch $\mathcal{S}_B$ via (3) as

$$\mathcal{L}_M = \sum_{x_{pt}^{(1)}, x_{pt}^{(2)} \sim \mathcal{S}_B} \ell(x_{pt}^{(1)}, x_{pt}^{(2)}) + \ell(x_{pt}^{(2)}, x_{pt}^{(1)}), \quad (4)$$

where we make explicit the need to calculate separately the loss for each data source $x_{pt}^{(k)}$ relative to the alternative $x_{pt}^{(\backslash k)}$, provided that the normalization term (denominator) in (3) keeps one data source fixed.

Local information extraction Recall that covariates for a given sample are structured as longitudinal series ordered in time. In practical scenarios, the state of a sample does not exhibit dramatic changes over short periods of time (relative to the time scale of the outcome horizon, M). Consequently, we hypothesize that covariates for a sample within a short time window are likely to share transient (local) information.

Consider a sample p and a predefined time window of w time points, we define sub-sequences at time t as

$\{x_{p,t-w}^{(k)}, \ldots, x_{p,t}^{(k)}, \ldots, x_{p,t+w}^{(k)}\}$, denoted for conciseness as $x_{p,t\pm w}^{(k)}$, for $k \in \{1,2\}$. Then, based on the hypothesis that their covariates share information within and across data sources, we can also consider the sets $(x_{pt}^{(k)}, x_{p,t\pm w}^{(k)})$ (within) and $(x_{pt}^{(k)}, x_{p,t\pm w}^{(\backslash k)})$ (across) as positive pairs in a contrastive learning framework.

However, unlike in standard contrastive learning (Chen et al., 2020), we wish to account for sharing of information over a short period of time, but in relation to the outcome of interest. In this manner, we can encourage the model to capture information that is relevant for outcome prediction. More specifically, we propose to weight the contrastive loss for positive pairs in (3) with the (unknown) outcome probability for a sample, for which we leverage a leave-one-out standard Kaplan-Meier (KM) estimator[1] (Rich et al., 2010), as a means to obtain the outcome distribution for a sample denoted as $S_p(t)$. Consequently, the weight for the loss for positive pairs is obtained as $w_p(t,t') = 1 - |S_p(t) - S_p(t')|$, where $\{t, t'\}$ are any two time points within a set of positive pairs as defined above. The local loss for all positive pairs corresponding to sample p can be written as

$$\ell_L(x_{pt}^{(k)}) = \sum_{j=-w|j\neq 0}^{w} w_p(t,t+j)\left[\ell(x_{pt}^{(k)}, x_{p,t+j}^{(k)}) + \ell(x_{pt}^{(k)}, x_{p,t+j}^{(\backslash k)})\right], \tag{5}$$

where $\ell(\cdot,\cdot)$ is the same as in (3), we have excluded $j=0$ from the calculation because $i)$ $\ell(x_{pt}^{(k)}, x_{pt}^{(k)})$ is trivial; and $ii)$ $\ell(x_{pt}^{(k)}, x_{pt}^{(\backslash k)})$ is already included in (4) since $w_p(t,t) = 1$. The *local loss* for all samples in batch $\mathcal{S}_B$ is

$$\mathcal{L}_L = \sum_{x_{pt}^{(1)}, x_{pt}^{(2)} \sim \mathcal{S}_B} \ell_L(x_{pt}^{(1)}) + \ell_L(x_{pt}^{(2)}). \tag{6}$$

Based on definitions, the $\mathcal{L}_{\mathcal{L}}$ can be viewed as a generalization to the $\mathcal{L}_{\mathcal{M}}$. The separation into two loss terms here is intentional for two reasons: $i)$ $\mathcal{L}_{\mathcal{M}}$ (matching) requires time stamp values (for positive pairs) across data sources to be identical so they can be "matched", which may not be feasible in some scenarios, $ii)$ $\mathcal{L}_{\mathcal{L}}$ (local) uses a time window but weights the loss using an empirical survival function (5), for which additional (outcomes) information is required. Note also that in principle, $\mathcal{L}_{\mathcal{M}}$ is unsupervised while $\mathcal{L}_{\mathcal{L}}$ is (weakly) supervised since the outcomes are used to weight the loss function.

Future information extraction Complementary to integrating local information into embeddings, we also aim to encourage that the encoded features capture information spanning multiple time steps into the future. This *future information*, represents the shared information between current and upcoming time steps, which in principle is likely to aid the predictive ability of current embeddings, especially when the goal is to predict future outcomes. Inspired by Contrastive Predictive Coding (CPC) (Oord et al., 2018), we also consider a future information loss.

Unlike the local information extraction loss, the future information loss is computed within data sources. Consider a batch $\mathcal{S}_C^{(k)} = \{x_{p,t+d}^{(k)}\} \cup \{x_{pt}^{(k)} | t \sim U(t_1, \ldots, t_{n_p})\}_{p=1}^{C}$ of C samples and time points selected uniformly at random, and such that each record contains a positive pair $(x_{p,t+d}^{(k)}, x_{pt}^{(k)})$, for a predetermined temporal distance d. The global information loss is defined as

$$\ell_F(x_{pt}^{(k)}) = -\log \frac{g^{(k)}(z_{p,t+d}^{(k)}, z_{pt}^{(k)})}{\sum_{x_{ij}^{(k)} \sim \mathcal{S}_C^{(k)}} g^{(k)}(z_{ij}^{(k)}, z_{pt}^{(k)})}, \tag{7}$$

where $g^{(k)}(z, z') = \exp(z^T W_k z')$ is a log-bilinear encoder with parameters W_k and $z_{pt}^{(k)}$ is obtained from the corresponding encoder via (2). Similarly, the *future loss* for all samples in batch $\mathcal{S}_C^{(k)}$ is

$$\mathcal{L}_F = \sum_{k\in\{1,2\}} \sum_{x_{pt}^{(k)} \sim \mathcal{S}_C} \ell_F(x_{pt}^{(k)}). \tag{8}$$

Finally, the complete loss to optimize the parameters of t $f^{(k)}(\cdot)$ and the auxiliary $g^k(\cdot)$, for $k \in \{1,2\}$ is

$$\mathcal{L} = \mathcal{L}_M + \mathcal{L}_L + \mathcal{L}_F. \tag{9}$$

Feature encoder architecture. Inspired by the ubiquitous success of attention-based models (Vaswani et al., 2017), we adopt the well-known multi-head attention with causal masking specification to account for the historical nature of the data for sample p. A block diagram of the encoder composition is shown in Figure 1, while the detailed network architecture is described in Appendix A.1.

3.3. Supervised Outcome Learning

In the supervised training stage, we concatenate the raw covariates $x_{pt}^{(k)}$, and encoded features $z_{pt}^{(k)}$, obtained from the pre-trained encoder $f^{(k)}(\cdot)$ in (2). These are fed into the classifier $c^{(k)}(\cdot)$ in (1), whose parameters are optimized with the standard cross-entropy loss and ground-truth labels y_{pt} using data from a single data source while keeping the parameters of the encoder $f^{(k)}(\cdot)$ fixed.

We consider feeding both covariates and encoder features into the classifier to account for outcome-relevant information loss during the encoding procedure, while the encoders focus on capturing longitudinal and cross-source information. In the experiments we consider two distinct practical

[1] The KM estimator for p is obtained from $\{t_j, e_j\}_{j=1|j\neq p}^{N}$.

scenarios for the inputs the classifier receives: i) a cross-section, *i.e.*, a single time point $(x_{pt}^{(k)}, z_{pt}^{(k)})$; or ii) longitudinal inputs $\{(x_{pj}^{(k)}, z_{pj}^{(k)})\}_{j=1}^{t}$ as in (1). For the former, we use multilayer perceptrons (MLPs) as a basic nonlinear classifier (Rosenblatt, 1958). For the latter we consider the open pre-trained transformer (OPT) (Zhang et al., 2022a), and the receptance weighted key value model (RWKV) (Peng et al., 2023), namely an attention-based and RNN-style longitudinal modeling architectures, respectively. Further details of these architectures can be found in Appendix A.1.

4. Experiments

Below we first introduce the baselines against which we compare CLOPPS, along with training details and metrics used to assess performance. Subsequently, we describe the datasets used and present results supporting the effectiveness and utility of the proposed framework. The source code used in the experiments is available at `https://github.com/mx41-m/Contrastive-Learning.git`.

Baselines We consider two types of baselines in our experiments, namely cross-sectional and longitudinal models. For the cross-sectional setting we use Elastic Net (Zou & Hastie, 2005) and MLP, as representatives for simple, yet effective, linear and nonlinear models, respectively. In both cases, the inputs to the model are either $x_{tp}^{(1)}$, $x_{tp}^{(2)}$ or $(x_{tp}^{(1)}, x_{tp}^{(2)})$, the latter as a naive alternative in which during inference $x_{tp}^{(k)}$ is used as the input to the model whereas $x_{tp}^{(\setminus k)}$ is imputed with population estimates. For the longitudinal setting we consider three recently proposed models, namely, RoFormer (Su et al., 2024), OPT (Zhang et al., 2022a) and RWKV (Peng et al., 2023). Notably, RoFormer shares the same architecture as CLOPPS (for both encoder and prediction head). For all of these, longitudinal covariates are either $\{x_{pj}^{(1)}\}_{j=1}^{t}$, $\{x_{pj}^{(2)}\}_{j=1}^{t}$ or $\{(x_{pj}^{(k)}, z_{pj}^{(k)})\}_{j=1}^{t}$. Importantly, all models using both data sources are used to quantify the performance of the *idealized* setting where they are available during inference.

Training details To ensure a fair comparison across models, hyperparameter tuning is done separately. We consider three datasets (described below): moving MNIST (MMNIST), a private real-world dataset (Private), and a public real-world dataset (MIMIC). The encoders for CLOPPS are trained for 50, 100 and 100 epochs on MMNIST, Private and MIMIC, respectively. The classifiers for CLOPPS are trained for 10, 5 and 5 epochs on MMNIST, Private and MIMIC, respectively. In CLOPPS, the values for τ, w and d are set to 0.1, 2 and 12, respectively, based on experimental results. Details of the architecture and hyperparameter tuning can be found in Appendix A.1 and A.2, respectively. Regarding the baseline models: for the Elastic Net, we set (via grid search) the L1 regularization ratio to 0.7 and the inverse of regularization strength to 0.5. The MLP is trained for 10, 25 and 25 epochs on MMNIST, Private and MIMIC, respectively. The RoFormer model is trained for 40, 30 and 30 on MMNIST, Private and MIMIC, respectively. The OPT and RWKV models are trained for 10, 20 and 20 epochs for MMNIST, Private and MIMIC, respectively. Details of all baselines can be found in Appendix A.1. For all models (excluding Elastic Net), AdamW (Loshchilov & Hutter, 2017) is employed as the optimizer. The values for the learning rate, beta, weight decay and batch size are set for all models to 10^{-4}, $(0.9, 0.999)$, 0.01, and 64, respectively

Metrics The performance of all models is evaluated in terms of outcome classification accuracy using well known metrics, namely, the area under the receiver operating characteristic (AUC) (Fawcett, 2006), and the area under the precision-recall curve (AP) (Boyd et al., 2013). These metrics are widely used in settings involving longitudinal prediction of outcomes (Tokodi et al., 2020; Choi et al., 2022).

4.1. Moving MNIST

This experiment serves as an idealized and conceptually straightforward illustration of the partial data source scenario considered in this work.

Dataset description Inspired by the methodology presented in Srivastava et al. (2015), we created $20,000$ sequences simulating longitudinal observations of moving digit images. These sequences were then divided into training, validation, and test datasets following an $8:1:1$ ratio. For each sequence, we define the event as the frame where two digits intersect and subsequent frames are discarded, thus yielding sequences of varying length. Labels for prediction y_{pt} for different values of the horizon M are created using whether two digits will intersect in the subsequent M frames relative to frame t. From this dataset, 15% of the sequences are randomly selected and censored uniformly at random, thus discarding frames after the censoring frame. Each frame is then split into two sub-frames along the horizontal midpoint, with the top half designated as one data source, $\{x_{pt}^{(1)}\}$ and the bottom half as the other, $\{x_{pt}^{(2)}\}$. Examples of the generated sequences are provided for reference in the Appendix A.3.

Results The primary objective of this experiment is to assess whether the encoded features effectively capture information from both data sources. To this end, we conduct a comparative analysis involving CLOPPS (with an MLP classifier) and three major baseline models: Elastic Net, MLP, and RoFormer, each evaluated at different horizons. Among these baselines, the MLP model achieved the best

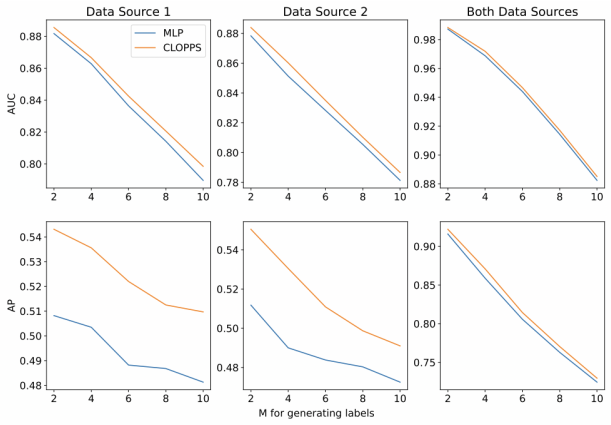

Figure 2. Performance comparisons (AUC and AP) of MLP and CLOPPS as a function of the prediction horizon M (x axis) on the Moving MNIST dataset. From left to right, only $\{x_{pt}^{(1)}\}$, only $\{x_{pt}^{(2)}\}$, and (for reference) both $(\{x_{pt}^{(1)}\}, \{x_{pt}^{(2)}\})$ are available during inference.

performance by a significant margin. Consequently, we only present results comparing the proposed framework and MLP in Figure 2. Recall that CLOPPS uses both data sources only when building the encoder. Detailed performance comparison for all models is included in Table 5 in Appendix A.3.

Figure 2 shows that CLOPPS consistently outperforms MLP across all prediction horizons and with larger performance differences observed in terms of AP. Interestingly, this is the case even in the hypothetical case where both data sources are available at inference. Also, as M increases, we observe a consistent decline in performance of both CLOPPS and MLP. This behavior is expected considering that a larger M requires the model to predict events more distant in the future, which is inherently harder especially as M increases.

4.2. Private Real-World Dataset

We now evaluate the proposed framework using a real-world dataset of end-stage kidney disease patients consisting of two data sources, namely, a *dialysis provider* (DCI) and a *national data system* (USRDS).

Dataset description The dialysis provider data source consists mainly of vital signs and laboratory tests recorded at monthly visits. Complementary, data from the national data system provides diagnoses and procedures codes also recorded on a monthly basis. For the experiments, these are aggregated into clinical classifications software (CCS) codes[2]. Both data sources also contain a set of shared covariates consisting of 4 demographic attributes, *i.e.*, age, sex, race, and ethnicity, which are initially excluded from the

[2]Details about CCS codes can be found at https://hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

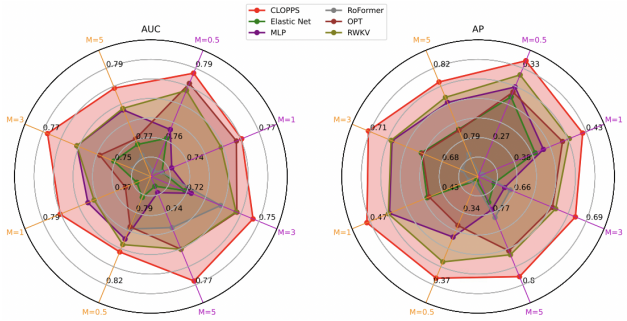

Figure 3. Performance comparisons (AUC and AP) between CLOPPS and five baselines for different prediction horizons $M = \{0.5, 1, 3, 5\}$ on the Private Real-World Dataset. Axes colors indicate the data source available during inference: orange for the dialysis provider and purple for the national data system.

model (see below). In this scenario we regard mortality as the outcome of interest. The resulting dataset consists of $40,752$ samples with 26.94 ± 28.65 (monthly) time points, and 89 and 620 covariates for the dialysis provider and national data system sources, respectively. Moreover, this dataset has an overall survival rate of 28.09% and 2.7% of sample time points were excluded from modeling due to censoring. Among the covariates, 64 are continuous, 288 are discrete (one-hot encoded), 361 are binary.

Results We performed a detailed comparison of CLOPPS with the five baseline models described above for various horizons $M = \{0.5, 1, 3, 5\}$, and the assumption that only one data source is available during inference. Results of this comparison are presented in Figure 3, for which we focus on the version of our model with the RWKV classifier as it outperformed MLP and OPT. However, results with all the alternative classifiers are presented in Appendix A.5. Figure 3 indicates that, regardless of data source, horizon and performance metric, CLOPPS consistently outperforms all baseline models. This consistency underscores the robustness and effectiveness of the proposed framework in capturing information from both data sources as a means to enhancing predictive ability.

Though the aim of CLOPPS is to capture information across data sources to then perform inference with a single data source, for completeness and transparency, we also compared CLOPPS with baseline models assuming access to both data sources, with results in Appendix A.5. These show that, with the availability of both data sources, CLOPPS still performs on par with, or better than the baseline models.

Besides the training mechanism discussed in Section 3.3, we also explore an alternative training strategy on the Private Real-World Dataset, namely, fine-tuning. This approach involves initially training our encoders, followed by a joint fine-tuning phase of both the encoders and the prediction

head tailored to each specific dataset. A detailed performance comparison between these two training strategies is presented in Appendix A.8, where the results reveal no additional gains from the fine-tuning mechanism.

Covariates common to both data sources were excluded from the previous comparison to conform to the more general case in which data sources do not overlap. However, incorporating such (demographic) information during the training of (mortality) classifiers can enhance the performance of the models. To verify that the encoded features (which do not use such information) can still augment mortality prediction performance when these other covariates are introduced, we incorporated them into the classifier for each data source. This augmented dataset was then utilized to also train all the baseline models. Results are presented in Figure 4, while detailed results for other classifier options within CLOPPS are provided in Appendix A.5.

Figure 4 shows results that are consistent with those in Figure 3, implying that even when combined with additional (static) covariates, the learned encoded features generated by CLOPPS enhance the performance of the single-source mortality prediction models. Moreover, Figure 3 and Figure 4 collectively underscore the robustness of the proposed approach, as CLOPPS always matches or improves the performance relative to directly training models on the single data source being available during inference, thus demonstrating the information transfer from the source that is available during training but not during inference.

Further, our experiments primarily consider various data sources from the same dataset. To better accommodate for general real-world scenarios, where data sources may from different datasets, we utilize our Private Real-World Dataset to emulate a prospective scenario, which is more akin to transfer learning. Details about this scenario and its results can be found in Appendix A.7. The results affirm the viability of CLOPPS in real-world prospective applications.

Ablation study The complete loss in (9) used for pretraining the encoders comprises three distinct components: L_M, L_L, and L_F. To explore the specific contribution of each of these, we conducted an ablation study using the Private dataset. Specifically, we set the outcome classifier to the simplest option, *i.e.*, the MLP, and focused the on 1-year mortality prediction, $M = 1$, assuming only the national data system data source is available during inference.

The results in Table 1 indicate that each loss component contributes to the classifier's performance in the mortality prediction task. This underscores the value of each component in generating embeddings that are more effective at improving prediction of outcomes. Among the three components, both $\mathcal{L}_M$ and $\mathcal{L}_L$ demonstrate a comparably substantial influence on improving mortality prediction performance. This finding is in line with our assumptions, considering that as previously mentioned, $\mathcal{L}_M$ can be seen as a particular case of $\mathcal{L}_L$.

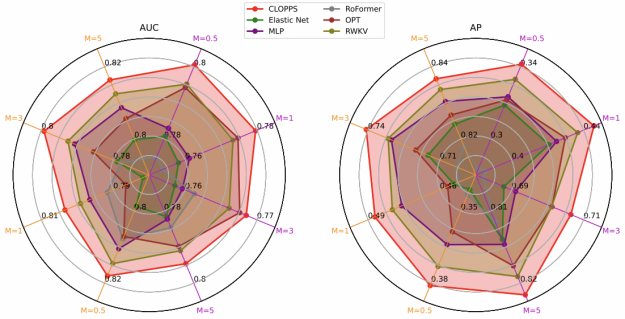

Figure 4. Performance comparisons (AUC and AP) between CLOPPS and five baselines for different prediction horizons $M = \{0.5, 1, 3, 5\}$, and *shared demographic covariates* on the Private Real-World Dataset. Axes colors indicate the data sources available during inference: orange for the dialysis provider and purple for the national data system.

Table 1. Ablation study results for 1-year mortality using the only the data from the national data system. The first row is for the encoder and MLP classifier trained jointly. The second row is for an MLP trained directly on raw covariates. Figures in parentheses represent improvement relative to the first row.

$\mathcal{L}_M$	$\mathcal{L}_L$	$\mathcal{L}_F$	AUC	AP
ENCODER + MLP			0.7336 (—)	0.3632 (—)
MLP			0.7371 (—)	0.4041 (—)
√			0.7496 (1.6%)	0.4127 (4.9%)
	√		0.7509 (1.7%)	0.4135 (5.0%)
		√	0.7414 (0.8%)	0.4038 (4.1%)
√	√		0.7511 (1.7%)	0.4144 (5.1%)
√		√	0.7521 (1.8%)	0.4146 (5.1%)
√	√	√	**0.7527 (1.9%)**	**0.4202 (5.7%)**

To further corroborate these insights, we expanded the ablation study to encompass other time horizons, namely, $M = \{0.5, 3, 5\}$, and data sources available at inference. Results of these additional studies, which support the findings in Table 1, can be found in Appendix A.4.

4.3. MIMIC (Public Real-World Dataset)

Given that the Private dataset is not readily publicly accessible, we also validate CLOPPS using the MIMIC-III clinical database (Johnson et al., 2016).

Dataset description MIMIC-III encompasses de-identified data from $+40,000$ patients who received care in critical care units. This comprehensive dataset consists of EHR data spanning 26 distinct data tables. An in-depth description of the data can be found in Johnson et al. (2016).

In our experiments, we opted for laboratory and proce-

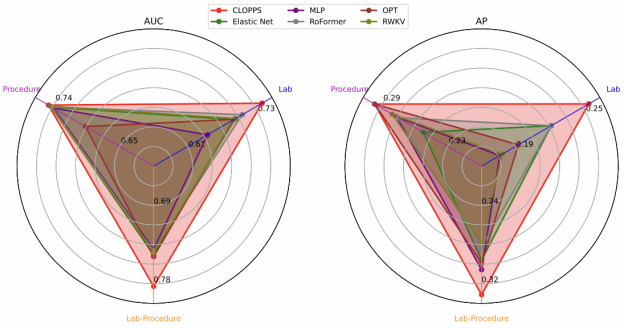

Figure 5. Performance comparison (AUC and AP) between CLOPPS and five baselines, including *shared covariates* on the MIMIC data. Axes colors indicate the data sources available during inference: procedures, labs and both labs and procedures.

dure events to emulate the type of data available in the two sources of our Private dataset. These two data sources will be referred to as Lab and Procedure hereafter. For each patient, we aggregated Lab values and Procedure codes hourly throughout their ICU stay. For Labs, we used the hourly mean if multiple records were available or the population mean (over the training set) in the case of no records being available. For Procedure codes, we employed binary indicators to denote whether a specific procedure occurred within each hour. Further, we also extracted other features such as sex, age, and admission codes, to serve as additional "shared" covariates for the mortality risk prediction model.

Patients with ICU stays exceeding one week were excluded, resulting in a final cohort of $20,784$ patients, with 53.32 ± 35.12 (hourly) time points, and 80, 12, 145 covariates for Lab, Procedure and shared covariates, respectively. Among these, 81 are continuous and 156 are binary. Moreover, this dataset has a survival rate of 88.80%. This cohort was then divided into training, validation, and testing subsets following an 8:1:1 ratio. The task is to predict mortality within the next 168 hours (1 week) relative to time point t.

Results Mirroring the approach used for the Private dataset, we compare the proposed framework with five baselines. This comparison is conducted in the scenario where the additional shared covariates are included in the classifier. Results in Figure 5 are shown for CLOPPS with the RWKV, consistent to those for the Private dataset. Detailed results for other classifier options can be found in Appendix A.6. Figure 5 demonstrates that regardless of the data source accessed during inference, CLOPPS consistently outperforms other baselines, which effectively validates it.

5. Conclusion

We introduced CLOPPS, a framework tailored for longitudinal outcome prediction with incomplete data sources, which can be of wide applicability in real-world scenarios, especially in healthcare settings. Utilizing contrastive learning, CLOPPS makes full use of the complete set of data sources during training, while generating highly informative embeddings for different data sources to be used during the inference stage, thus facilitating accurate predictions and potentially more portable models. Our experiments demonstrated that CLOPPS consistently outperforms strong competing baselines, under conditions of partial access to data sources during inference, thereby validating its efficacy in real-world mortality prediction scenarios. This study has several limitations: i) we only consider tabular data; ii) though readily extensible, we only consider two data sources; and iii) we only consider a single outcome, thus not exploring competing risk. In future work, we plan to expand the types (*e.g.*, images and text) and number of data sources and outcomes considered. Further, we are also interested in exploring the potential applicability of our framework to a variety of medical predictive tasks (Ghassemi et al., 2020; Wang et al., 2020) beyond mortality prediction.

Impact Statement

The proposed framework is intended as a means to improve the portability and reduce the data requirements of clinical outcome prediction models. This is important because it is well recognized that these are two barriers preventing predictive models from being more widely deployed as part of clinical decision support systems. Despite the potential positive impact of CLOPPS on the clinical predictive model ecosystem, it is worth noting that this work does not explore ethical implications of bias caused or amplified by models produced by CLOPPS. Fortunately, there is a growing body of work in machine learning devoted to address problems associated with quantifying and correcting bias, equity or fairness issues in predictive models.

Acknowledgements

We thank the anonymous reviewers for their constructive feedback. This work was supported by NIDDK R01DK123062. The data reported here have been supplied by the United States Renal Data System (USRDS). The interpretation and reporting of these data are the responsibility of the author(s) and in no way should be seen as an official policy or interpretation of the US government.

References

Agarap, A. F. Deep learning using rectified linear units. *arXiv preprint arXiv:1803.08375*, 2018.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Baumel, T., Nassour-Kassis, J., Cohen, R., Elhadad, M., and Elhadad, N. Multi-label classification of patient notes: case study on icd code assignment. In *Workshops at the thirty-second AAAI conference on artificial intelligence*, 2018.

Boyd, K., Eng, K. H., and Page, C. D. Area under the precision-recall curve: point estimates and confidence intervals. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, pp. 451–466. Springer, 2013.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Choi, E., Schuetz, A., Stewart, W. F., and Sun, J. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, 24(2):361–370, 2017.

Choi, M. H., Kim, D., Choi, E. J., Jung, Y. J., Choi, Y. J., Cho, J. H., and Jeong, S. H. Mortality prediction of patients in intensive care units using machine learning algorithms based on electronic health records. *Scientific reports*, 12(1):7180, 2022.

Citi, L. and Barbieri, R. Physionet 2012 challenge: Predicting mortality of icu patients using a cascaded svm-glm paradigm. In *2012 Computing in Cardiology*, pp. 257–260. IEEE, 2012.

Clermont, G., Angus, D. C., DiRusso, S. M., Griffin, M., and Linde-Zwirble, W. T. Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models. *Critical care medicine*, 29(2):291–296, 2001.

Crawford, E. D., Batuello, J. T., Snow, P., Gamito, E. J., McLeod, D. G., Partin, A. W., Stone, N., Montie, J., Stock, R., Lynch, J., et al. The use of artificial intelligence technology to predict lymph node spread in men with clinically localized prostate carcinoma. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 88(9):2105–2109, 2000.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Doctor, A. Predicting clinical events via recurrent neural networks edward choi, mohammad taha bahadori, andy schuetz, walter f. *Stewart, Jimeng Sun arXiv (2015-11-18) https://arxiv. org/abs/1511.05942 v11*.

Doig, G., Inman, K., Sibbald, W., Martin, C., and Robertson, J. Modeling mortality in the intensive care unit: comparing the performance of a back-propagation, associative-learning neural network with multivariate logistic regression. In *Proceedings of the annual symposium on computer application in medical care*, pp. 361. American Medical Informatics Association, 1993.

Doreswamy, Gad, I., and Manjunatha, B. Multi-label classification of big ncdc weather data using deep learning model. In *Soft Computing Systems: Second International Conference, ICSCS 2018, Kollam, India, April 19–20, 2018, Revised Selected Papers 2*, pp. 232–241. Springer, 2018.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Dybowski, R., Gant, V., Weller, P., and Chang, R. Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. *The Lancet*, 347(9009):1146–1150, 1996.

Eldele, E., Ragab, M., Chen, Z., Wu, M., Kwoh, C. K., Li, X., and Guan, C. Time-series representation learning via temporal and contextual contrasting. *arXiv preprint arXiv:2106.14112*, 2021.

Fawcett, T. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

Franceschi, J.-Y., Dieuleveut, A., and Jaggi, M. Unsupervised scalable representation learning for multivariate time series. *Advances in neural information processing systems*, 32, 2019.

Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., and Ranganath, R. A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings*, 2020:191, 2020.

Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., and Galstyan, A. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96, 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Hewage, P., Trovati, M., Pereira, E., and Behera, A. Deep learning-based effective fine-grained weather forecasting model. *Pattern Analysis and Applications*, 24(1):343–366, 2021.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Hong, C., Yi, F., and Huang, Z. Deep-csa: Deep contrastive learning for dynamic survival analysis with competing risks. *IEEE Journal of Biomedical and Health Informatics*, 26(8):4248–4257, 2022.

Johnson, A., Pollard, T., and Mark, R. Mimic-iii clinical database (version 1.4). *PhysioNet*, 10(C2XW26):2, 2016.

Karabacak, M. and Margetis, K. Embracing large language models for medical applications: Opportunities and challenges. *Cureus*, 15(5), 2023.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Kim, S., Kim, W., and Park, R. W. A comparison of intensive care unit mortality prediction models through the use of data mining techniques. *Healthcare informatics research*, 17(4):232–243, 2011.

Kiyasseh, D., Zhu, T., and Clifton, D. A. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning*, pp. 5606–5615. PMLR, 2021.

Lemeshow, S., Klar, J., Teres, D., Avrunin, J. S., Gehlbach, S. H., Rapoport, J., and Rué, M. Mortality probability models for patients in the intensive care unit for 48 or 72 hours: a prospective, multicenter study. *Critical care medicine*, 22(9):1351–1358, 1994.

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

Logeswaran, L. and Lee, H. An efficient framework for learning sentence representations. *arXiv preprint arXiv:1803.02893*, 2018.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Meyfroidt, G., Güiza, F., Ramon, J., and Bruynooghe, M. Machine learning techniques to examine large patient databases. *Best Practice & Research Clinical Anaesthesiology*, 23(1):127–143, 2009.

Narayan, S. The generalized sigmoid activation function: Competitive supervised learning. *Information sciences*, 99(1-2):69–82, 1997.

Nimgaonkar, A. and Sudarshan, S. Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models. *Intensive Care Med*, 30:248–253, 2004.

Nunez, J.-J., Leung, B., Ho, C., Bates, A. T., and Ng, R. T. Predicting the survival of patients with cancer from their initial oncology consultation document using natural language processing. *JAMA Network Open*, 6(2):e230813–e230813, 2023.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Peng, B., Alcaide, E., Anthony, Q., Albalak, A., Arcadinho, S., Cao, H., Cheng, X., Chung, M., Grella, M., GV, K. K., et al. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Raghu, A., Chandak, P., Alam, R., Guttag, J., and Stultz, C. Sequential multi-dimensional self-supervised learning for clinical time series. In *International Conference on Machine Learning*, pp. 28531–28548. PMLR, 2023.

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., et al. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):18, 2018.

Ramon, J., Fierens, D., Güiza, F., Meyfroidt, G., Blockeel, H., Bruynooghe, M., and Van Den Berghe, G. Mining data from intensive care patients. *Advanced Engineering Informatics*, 21(3):243–256, 2007.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

Ribas, V. J., López, J. C., Ruiz-Sanmartín, A., Ruiz-Rodríguez, J. C., Rello, J., Wojdel, A., and Vellido, A. Severe sepsis mortality prediction with relevance vector machines. In *2011 annual international conference of the IEEE engineering in medicine and biology society*, pp. 100–103. IEEE, 2011.

Rich, J. T., Neely, J. G., Paniello, R. C., Voelker, C. C., Nussenbaum, B., and Wang, E. W. A practical guide to understanding kaplan-meier curves. *Otolaryngology—Head and Neck Surgery*, 143(3):331–336, 2010.

Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

Rumelhart, D. E., Hinton, G. E., Williams, R. J., et al. Learning internal representations by error propagation, 1985.

Schneider, F., Xu, X., Ernst, M. R., Yu, Z., and Triesch, J. Contrastive learning through time. In *SVRHM 2021 Workshop@ NeurIPS*, 2021.

Shaw, P., Uszkoreit, J., and Vaswani, A. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Song, H., Rajan, D., Thiagarajan, J., and Spanias, A. Attend and diagnose: Clinical time series analysis using attention models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Srivastava, N., Mansimov, E., and Salakhudinov, R. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pp. 843–852. PMLR, 2015.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Tokodi, M., Schwertner, W. R., Kovács, A., Tősér, Z., Staub, L., Sárkány, A., Lakatos, B. K., Behon, A., Boros, A. M., Perge, P., et al. Machine learning-based mortality prediction of patients undergoing cardiac resynchronization therapy: the semmelweis-crt score. *European heart journal*, 41(18):1747–1756, 2020.

Tonekaboni, S., Eytan, D., and Goldenberg, A. Unsupervised representation learning for time series with temporal neighborhood coding. *arXiv preprint arXiv:2106.00750*, 2021.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

U.S. Renal Data System. *2023 USRDS Annual Data Report: Epidemiology of Kidney Disease in the United States.* National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Vincent, J.-L. and Singer, M. Critical care: advances and future perspectives. *The Lancet*, 376(9749):1354–1361, 2010.

Wagner, D. P., Knaus, W. A., Harrell, F. E., Zimmerman, J. E., and WATIS, C. Daily prognostic estimates for critically ill adults in intensive care units: results from a prospective, multicenter, inception cohort analysis. *Critical care medicine*, 22(9):1359–1372, 1994.

Wang, S., McDermott, M. B., Chauhan, G., Ghassemi, M., Hughes, M. C., and Naumann, T. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of the ACM conference on health, inference, and learning*, pp. 222–235, 2020.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Wong, L. and Young, J. A comparison of icu mortality prediction using the apache ii scoring system and artificial neural networks. *Anaesthesia*, 54(11):1048–1054, 1999.

Yèche, H., Dresdner, G., Locatello, F., Hüser, M., and Rätsch, G. Neighborhood contrastive learning applied to online patient monitoring. In *International Conference on Machine Learning*, pp. 11964–11974. PMLR, 2021.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022a.

Zhang, X., Zhao, Z., Tsiligkaridis, T., and Zitnik, M. Self-supervised contrastive pre-training for time series via

time-frequency consistency. *Advances in Neural Information Processing Systems*, 35:3988–4003, 2022b.

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

Table 2. Tuning Results for τ on the Private validation dataset. Throughout the tuning phase we only include $\mathcal{L}_M$. The best performances are in highlighted in bold text.

M (YEARS)	INFERENCE DATASET	METRIC	τ VALUE 0.01	0.1	0.5
0.5	NATIONAL DATA SYSTEM	AUC	0.7729	**0.7749**	0.7740
		AP	0.3231	**0.3246**	0.3174
	DIALYSIS PROVIDER	AUC	0.8050	**0.8108**	0.8099
		AP	0.3726	**0.3768**	0.3781
1	NATIONAL DATA SYSTEM	AUC	0.7451	**0.7488**	0.7475
		AP	0.4188	**0.4225**	0.4173
	DIALYSIS PROVIDER	AUC	0.7771	**0.7816**	0.7790
		AP	0.4718	**0.4763**	0.4736
3	NATIONAL DATA SYSTEM	AUC	0.7367	**0.7418**	0.7405
		AP	0.6803	**0.6855**	0.6820
	DIALYSIS PROVIDER	AUC	0.7585	**0.7632**	0.7595
		AP	0.7077	**0.7133**	0.7086
5	NATIONAL DATA SYSTEM	AUC	0.7606	**0.7664**	0.7644
		AP	0.8045	**0.8086**	0.8065
	DIALYSIS PROVIDER	AUC	0.7801	**0.7816**	0.7800
		AP	0.8182	**0.8197**	0.8184
BEST PERFORMANCE COUNT			0/16(0.00%)	16/16(100.00%)	0/16(0.00%)

A. Appendix

A.1. Model Architecture Details

Encoders The two encoders $f^{(1)}$ and $f^{(2)}$ in CLOPPS are identical in architecture. Each encoder starts with an input embedding layer, followed by two attention blocks. Each block consists of a multi-head self-attention layer, featuring 4 heads with each head having a dimension of 128, and a fully connected feed-forward layer. A residual connection (He et al., 2016) and a layer normalization (Ba et al., 2016) are employed after each of sub-layer. The final output dimension of the encoders is set at 512. Diverging from traditional position embeddings like absolute position embedding (Vaswani et al., 2017), or relative position embedding (Shaw et al., 2018), our encoders utilize a rotary position embedding (Su et al., 2024).

Classifiers The MLP classifier in CLOPPS consists of three layers. The output dimensions for these layers are 512, 256, and 1, respectively. Each layer integrates a ReLU activation function (Agarap, 2018), except from the final layer which is followed by a sigmoid activation function (Narayan, 1997). For the construction of the OPT and RWKV models, we leverage the Hugging-face's transformers library (Wolf et al., 2019). The OPT model has 2 hidden layers, 4 attention heads, and a hidden size of 512. The RWKV model is similarly designed with an attention and overall hidden size of 512 units, as well as 2 hidden layers.

Baselines For the Elastic Net baseline, we configured the L1 regularization ratio to 0.7 and set the inverse of regularization strength to 0.5. The RoFormer baseline consists of two attention layers, following the architecture used for the encoders, and then followed by a three-layer MLP classifier. The MLP dimensions are 512, 256, and 1, respectively, with ReLU activation functions. The MLP, OPT, and RWKV baselines are architecturally identical to the classifiers in CLOPPS.

A.2. Hyperparameter Tuning

In this section, we detail the tuning of hyperparameters τ, w and d for CLOPPS, utilizing the Private validation dataset without demographic features. Initially, τ was tuned under the assumption that our encoders are optimized solely with the loss $\mathcal{L}_M$. After setting τ, we proceeded to independently tune w and d, implying that while adjusting one hyperparameter, the other was set to 0. Detailed results of this tuning process can be found in Table 2 for τ, Table 3 for w, and in Table 4 for d. Based on these results, we selected $\tau = 0.1$ $w = 2$ and $k = 12$ as the optimal parameters for CLOPPS, as these settings yielded the highest frequency of top performance across multiple scenarios.

Table 3. Tuning Results for w on Private validation dataset. Throughout the tuning phase, setting $d = 0$ signifies the exclusion of $\mathcal{L}_F$. Similarly, $w = 0$ indicates that $\mathcal{L}_L$ is not incorporated. The best performances are highlighted in bold text.

M (YEARS)	INFERENCE DATASET	METRIC	w VALUE 0	1	2	3
0.5	NATIONAL DATA SYSTEM	AUC	0.7749	0.7784	**0.7785**	0.7757
		AP	0.3246	0.3284	**0.3291**	0.3263
	DIALYSIS PROVIDER	AUC	0.8108	**0.8155**	0.8145	0.8149
		AP	0.3768	**0.3834**	0.3821	0.3823
1	NATIONAL DATA SYSTEM	AUC	0.7488	**0.7519**	0.7514	0.7498
		AP	0.4225	0.4250	**0.4263**	0.4233
	DIALYSIS PROVIDER	AUC	0.7816	0.7855	0.7852	**0.7868**
		AP	0.4763	0.4833	0.4795	**0.4834**
3	NATIONAL DATA SYSTEM	AUC	0.7418	0.7425	**0.7434**	0.7424
		AP	0.6855	0.6865	**0.6866**	0.6842
	DIALYSIS PROVIDER	AUC	0.7632	0.7677	0.7667	**0.7679**
		AP	0.7133	0.7183	0.7162	**0.7184**
5	NATIONAL DATA SYSTEM	AUC	**0.7664**	0.7636	0.7660	0.7644
		AP	**0.8086**	0.8056	0.8069	0.8048
	DIALYSIS PROVIDER	AUC	0.7816	0.7854	0.7858	**0.7864**
		AP	0.8197	0.8226	**0.8230**	0.8229
BEST PERFORMANCE COUNT			2/16(12.50%)	3/16(18.75%)	6/16(37.50%)	5/16(31.25%)

A.3. Moving MNIST

In this section, we firstly show specific examples from the generated Moving MNIST (MMNIST) dataset to provide a better understanding of it. The first example, illustrated in Figure 6, demonstrates a sequence where an event occurs, specifically where the two numbers in the sequence intersect. Another example, depicted in Figure 7, represents censored sequences within the dataset. In both instances, the sequence labels are generated based on the criterion that $M = 2$, indicating whether the two numbers will intersect in the subsequent 2 frames.

Additionally, to provide a comprehensive view of the performance across different models, we show the full results from all models on the Moving MNIST dataset in Table 5, offering a broader perspective of the effectiveness of each model in this specific context.

A.4. Full Ablation Study Results on Private Dataset

In Table 6, we present the complete results of the ablation study on the Private dataset. The results for half-year, 3-year and 5-year mortality predictions in Table 6 assume we only have access to one data source (the dialysis provider or the national data system) during inference, support the findings in Table 1.

A.5. Full results of Private Dataset

In this section, we provide a comprehensive comparison of CLOPPS against all baselines using the Private dataset. Figure 8 and 9 provide a detailed performance comparison between CLOPPS, with various classifier options, and all baseline models. These figures show results on Private dataset with and without shared demographic covariates, respectively. Across different mortality prediction scenarios, CLOPPS, especially when employing the RWKV classifier, consistently outperforms all other models.

In Table 7, we include a comparison of CLOPPS performance against the baseline models, specifically in scenarios where both data sources are available during inference. According to the results in Table 7, there are instances where some baseline models achieve performance comparable to, or in rare cases, better than CLOPPS. This is primarily because, when both data sources are available, the raw covariates tend to encompass the insights provided by our encoded features, leading to similar performance levels in other baseline models. Despite this, it is important to note that in the majority of cases, as indicated in Table 7, CLOPPS features contribute to an improvement in mortality prediction performance. These findings highlight the robustness and adaptability of CLOPPS, demonstrating its effectiveness in a variety of data availability scenarios.

Table 4. Tuning Results for d on the Private validation dataset. Throughout the tuning phase, setting $w = 0$ signifies the exclusion of $\mathcal{L}_L$. Similarly, $d = 0$ indicates that $\mathcal{L}_F$ is not incorporated. The best performances are highlighted in bold text.

M (YEARS)	INFERENCE DATASET	METRIC	d VALUE 0	6	12	18
0.5	NATIONAL DATA SYSTEM	AUC	0.7749	0.7754	**0.7771**	0.7754
		AP	0.3246	0.3247	**0.3257**	0.3246
	DIALYSIS PROVIDER	AUC	0.8108	**0.8137**	0.8132	0.8123
		AP	0.3768	**0.3812**	0.3800	0.3777
1	NATIONAL DATA SYSTEM	AUC	0.7488	0.7490	**0.7504**	0.7490
		AP	0.4225	0.4218	0.4225	**0.4231**
	DIALYSIS PROVIDER	AUC	0.7816	**0.7850**	0.7843	0.7835
		AP	0.4763	**0.4796**	0.4794	0.4782
3	NATIONAL DATA SYSTEM	AUC	0.7418	0.7425	**0.7438**	0.7421
		AP	**0.6855**	0.6846	0.6852	0.6849
	DIALYSIS PROVIDER	AUC	0.7632	**0.7654**	0.7647	0.7639
		AP	0.7133	**0.7144**	0.7136	0.7130
5	NATIONAL DATA SYSTEM	AUC	0.7664	0.7672	**0.7681**	0.7656
		AP	0.8086	0.8082	**0.8091**	0.8077
	DIALYSIS PROVIDER	AUC	0.7816	0.7828	**0.7829**	0.7826
		AP	0.8197	0.8201	**0.8203**	0.8202
BEST PERFORMANCE COUNT			1/16(6.25%)	6/16(37.50%)	8/16(50%)	1/16(6.25%)

A straightforward approach to address the challenge of partial data availability during inference is to train models using complete data sources, and then substitute (impute) the missing data with the mean values from the training dataset during inference. We applied this method to Elastic Net and MLP on the Private dataset without shared (demographic) information, and the results are presented in Table 8. For comparison, we also include the performance metrics of CLOPPS using MLP as classifier in the same table. The results presented in Table 8 demonstrate that CLOPPS consistently outperforms the baseline models, even when the baseline models employ this simple approach.

A.6. Full results on the Public Dataset

In this section, we provide a comprehensive comparison of CLOPPS against all baselines using the Public dataset (MIMIC-III). Figure 10 provides a detailed performance comparison between CLOPPS, with various classifier options, and all baseline models. The results in Figure 10 demonstrate that CLOPPS consistently outperforms all other baselines.

A.7. Transfer Learning Scenarios Simulation

In this section, we detail the methods employed to simulate prospective transfer learning scenarios and present the results. Specifically, we allocated patient records that fully occurring before 2013 to the training set, amassing 12,916 samples for training our framework. Further, the testing dataset comprises patient records that started and fully occurred after 2013, with 2,201 samples, to evaluate the efficacy of our trained framework. The detailed comparative results across different models, as shown in Table 9, demonstrate that CLOPPS either outperforms or is comparable to other baseline models in simulated transfer learning scenarios. This underscores the effectiveness of CLOPPS in real-world prospective scenarios.

A.8. Fine-tuning Results

In this section, we provide the performance comparison between the training mechanism introduced in Section 3.3 and fine-tuning mechanism in Table 10. Fine-tuning mechanism refers to the initial training of our encoders, followed by a joint fine-tuning phase of both the encoders and prediction head for each specific dataset. The comparative results indicate a subtle difference between the two training strategies, thus indicating no additional gains from fine tuning.

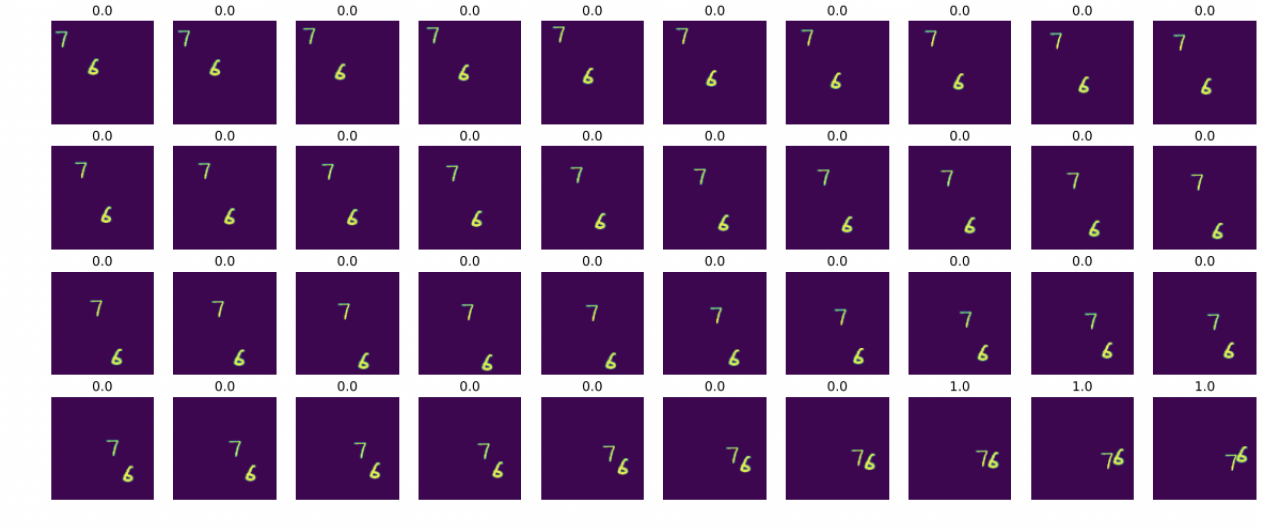

Figure 6. Sequence showcasing movement of digits 6 and 7. Throughout the frames, both numbers move randomly and eventually intersect, signifying an event. Frame-specific labels are presented above each frame.

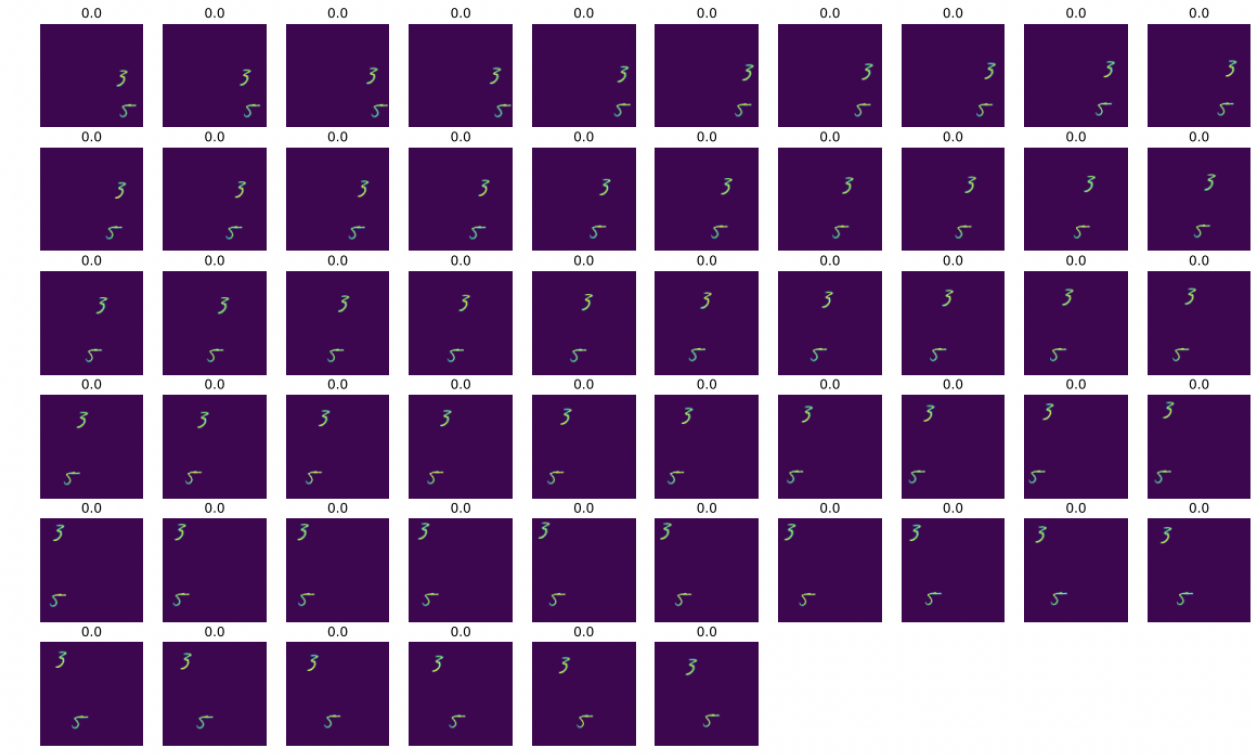

Figure 7. Censored sequence illustrating the movement of digits 3 and 5. The frames depict both numbers moving randomly without intersecting, indicating the absence of an event. Labels specific to each frame are provided above each frame.

Table 5. Performance comparison on Moving MNIST (MMNIST). We show the performance for three baselines, Elastic Net, RoFormer, MLP, and CLOPPS using MLP as classifier, where the M represents that two digits will intersect each other in the next M frames.

M FRAME	MODEL	TOP HALF AUC	TOP HALF AP	BOTTOM HALF AUC	BOTTOM HALF AP	WHOLE FRAME AUC	WHOLE FRAME AP
2	ELASTIC NET	0.5746	0.1647	0.5381	0.1487	0.5886	0.1337
	ROFORMER	0.8001	0.2914	0.7712	0.2622	0.9757	0.8055
	MLP	0.8817	0.5082	0.8783	0.5118	0.9873	0.9160
	CLOPPS(MLP)	**0.8855**	**0.5431**	**0.8839**	**0.5505**	**0.9884**	**0.9221**
4	ELASTIC NET	0.5800	0.1947	0.5467	0.1765	0.5873	0.1619
	ROFORMER	0.8165	0.3668	0.7973	0.3491	0.9675	0.8022
	MLP	0.8628	0.5035	0.8513	0.4900	0.9688	0.8588
	CLOPPS(MLP)	**0.8666**	**0.5356**	**0.8602**	**0.5305**	**0.9719**	**0.8709**
6	ELASTIC NET	0.5826	0.2216	0.5530	0.2021	0.5828	0.1877
	ROFORMER	0.8142	0.4064	0.7897	0.3739	0.9093	0.5909
	MLP	0.8364	0.4882	0.8283	0.4838	0.9440	0.8057
	CLOPPS(MLP)	**0.8425**	**0.5220**	**0.8350**	**0.5109**	**0.9467**	**0.8144**
8	ELASTIC NET	0.5846	0.2468	0.5560	0.2253	0.5768	0.2119
	ROFORMER	0.8004	0.4186	0.7868	0.4006	0.8965	0.5967
	MLP	0.8142	0.4868	0.8055	0.4803	0.9143	0.7629
	CLOPPS(MLP)	**0.8206**	**0.5125**	**0.8103**	**0.4987**	**0.9171**	**0.7704**
10	ELASTIC NET	0.5846	0.2677	0.5556	0.2452	0.5692	0.2332
	ROFORMER	0.7912	0.4377	0.7730	0.4165	0.8812	0.6045
	MLP	0.7897	0.4813	0.7814	0.4725	0.8824	0.7244
	CLOPPS(MLP)	**0.7985**	**0.5097**	**0.7867**	**0.4910**	**0.8851**	**0.7295**

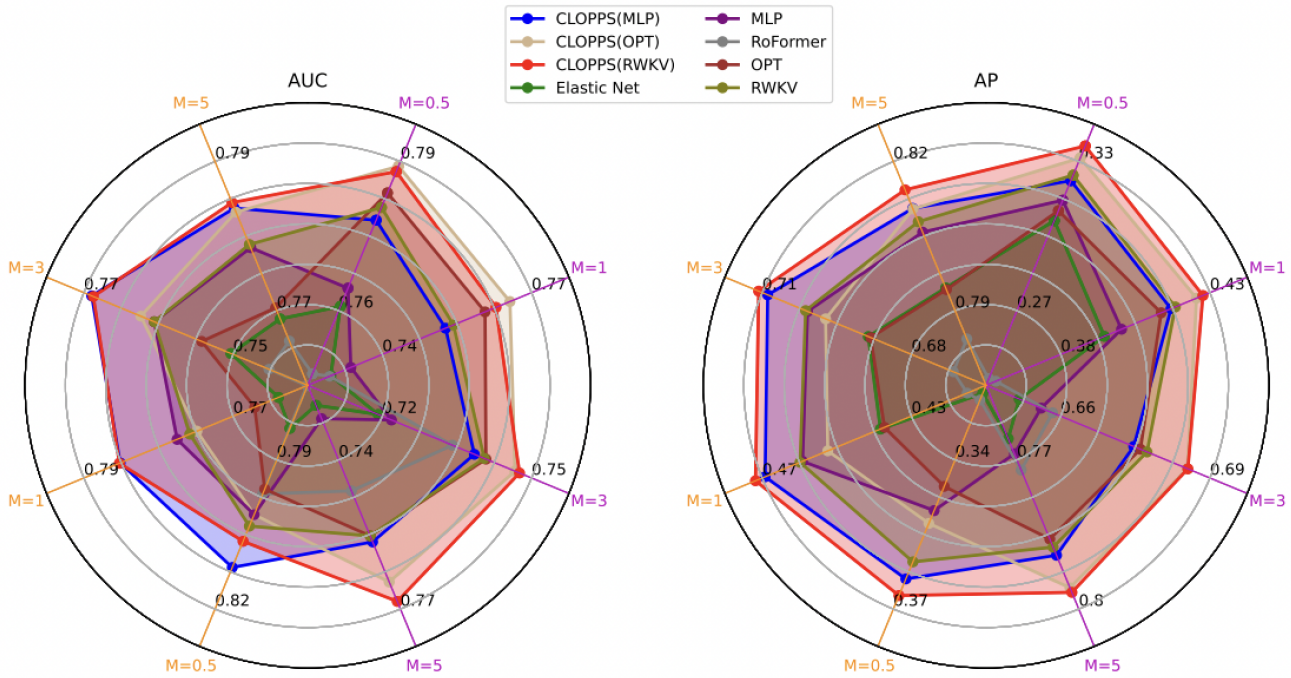

Figure 8. Performance comparisons (AUC and AP) between CLOPPS and five baselines for different prediction horizons $M = \{0.5, 1, 3, 5\}$ on the Private Real-World Dataset. Axes colors indicate the data sources available during inference: orange for the dialysis provider and purple for the national data system. The name in parenthesis following CLOPPS indicates which classifier used.

Table 6. Full results for the ablation study. We pretrained the encoders using various loss component configurations, subsequently we used the generated embeddings for mortality prediction task assuming only one data source (the dialysis provider or the national data system) is available during inference, and where the M represents the prediction horizon. The Encoder + MLP denotes the encoder and MLP classifier trained jointly.

M (YEARS)	LOSS COMPONENT $\mathcal{L}_M$	$\mathcal{L}_L$	$\mathcal{L}_G$	NATIONAL DATA SYSTEM AUC	NATIONAL DATA SYSTEM AP	DIALYSIS PROVIDER AUC	DIALYSIS PROVIDER AP
0.5	ENCODER + MLP			0.7519	0.2425	0.7979	0.3231
	√			0.7744	0.3128	0.8052	0.3542
		√		0.7763	0.3100	0.8014	0.3462
			√	0.7681	0.3060	0.7977	0.3426
	√	√		0.7759	0.3130	0.8102	0.3595
	√		√	0.7766	0.3135	0.8072	0.3576
	√	√	√	**0.7772**	**0.3184**	**0.8100**	**0.3608**
1	ENCODER + MLP			0.7336	0.3632	0.7601	0.4147
	√			0.7496	0.4127	0.7794	0.4617
		√		0.7509	0.4135	0.7769	0.4557
			√	0.7414	0.4038	0.7739	0.4526
	√	√		0.7511	0.4144	0.7848	0.4650
	√		√	0.7521	0.4146	0.7816	0.4642
	√	√	√	**0.7527**	**0.4202**	**0.7837**	**0.4650**
3	ENCODER + MLP			0.7349	0.6540	0.7449	0.6669
	√			0.7383	0.6681	0.7634	0.7022
		√		0.7378	0.6680	0.7623	0.7004
			√	0.7273	0.6544	0.7551	0.6917
	√	√		0.7399	0.6709	0.7683	0.7079
	√		√	0.7410	0.6705	0.7653	0.7039
	√	√	√	**0.7402**	**0.6711**	**0.7673**	**0.7060**
5	ENCODER + MLP			0.7474	0.7746	0.7656	0.7880
	√			0.7549	0.7864	0.7783	0.8080
		√		0.7543	0.7876	0.7803	0.8092
			√	0.7440	0.7762	0.7727	0.8026
	√	√		0.7550	0.7890	0.7840	0.8126
	√		√	0.7574	0.7899	0.7806	0.8100
	√	√	√	**0.7558**	**0.7893**	**0.7824**	**0.8105**

Table 7. Performance comparison. CLOPPS (MLP), CLOPPS (OPT) and CLOPPS (RWKV) represent using different classifier options in CLOPPS, *i.e.,* MLP, OPT, RWKV. This comparison is conducted using the Private dataset, under the assumption of full data sources accessibility during training and inference.

M (YEARS)	METRIC	MODEL ELASTIC NET	MLP	ROFORMER	OPT	RWKV	CLOPPS (MLP)	CLOPPS (OPT)	CLOPPS (RWKV)
0.5	AUC	0.8164	0.8203	0.7956	0.8144	0.8197	**0.8244**	0.8151	0.8220
	AP	0.3876	**0.3980**	0.3102	0.3775	0.3894	0.3971	0.3755	0.3960
1	AUC	0.7911	0.7955	0.7836	0.7920	0.7933	**0.7999**	0.7931	0.7979
	AP	0.4873	0.4948	0.4524	0.4791	0.4865	**0.4959**	0.4785	0.4930
3	AUC	0.7779	0.7816	0.7717	0.7801	0.7836	**0.7878**	0.7808	0.7873
	AP	0.7206	0.7254	0.6983	0.7167	0.7247	**0.7318**	0.7188	0.7291
5	AUC	0.7944	0.7971	0.7844	0.7929	0.7982	0.8026	0.7953	**0.8028**
	AP	0.8240	0.8259	0.8097	0.8214	0.8271	0.8312	0.8232	**0.8315**

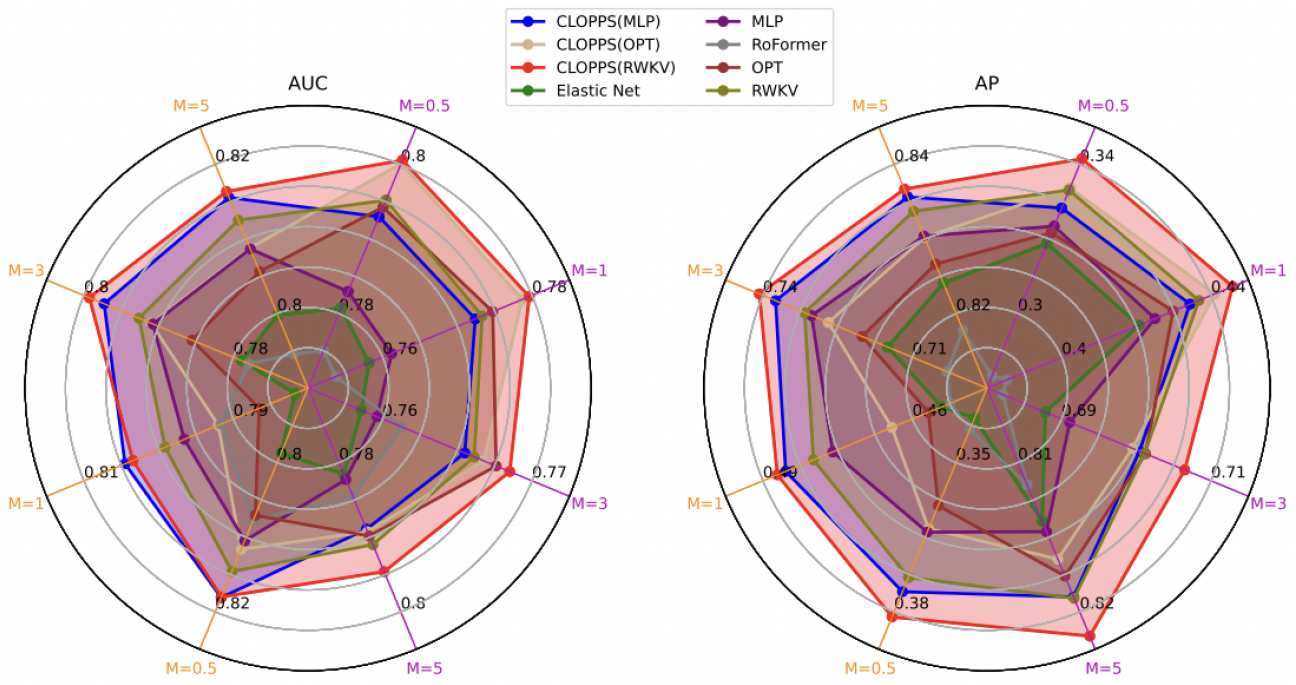

Figure 9. Performance comparisons (AUC and AP) between CLOPPS and five baselines for different prediction horizons $M = \{0.5, 1, 3, 5\}$ and *shared demographic covariates on the Private Real-World Dataset*. Axes colors indicate the data sources available during inference: orange for the dialysis provider and purple for the national data system. The name in parenthesis following CLOPPS indicates which classifier used.

Table 8. Results of substituting missing data sources (imputation) with training dataset statistics during inference where *substitute dialysis provider* and *substitute national data system* indicate the use of mean values from the *dialysis provider* and *national data system* training datasets, respectively, to compensate for (impute) their absence during inference.

M (YEARS)	MODEL	NATIONAL DATA SYSTEM (SUBSTITUTE DIALYSIS PROVIDER) AUC	NATIONAL DATA SYSTEM (SUBSTITUTE DIALYSIS PROVIDER) AP	DIALYSIS PROVIDER (SUBSTITUTE NATIONAL DATA SYSTEM) AUC	DIALYSIS PROVIDER (SUBSTITUTE NATIONAL DATA SYSTEM) AP
0.5	ELASTIC NET	0.7414	0.2852	0.7814	0.3105
	MLP	0.7529	0.2988	0.7785	0.3197
	CLOPPS(MLP)	**0.7772**	**0.3184**	**0.8100**	**0.3608**
1	ELASTIC NET	0.7142	0.3784	0.7592	0.4268
	MLP	0.7228	0.3877	0.7637	0.4388
	CLOPPS(MLP)	**0.7527**	**0.4202**	**0.7837**	**0.4650**
3	ELASTIC NET	0.7002	0.6311	0.7425	0.6755
	MLP	0.7045	0.6329	0.7485	0.6856
	CLOPPS(MLP)	**0.7402**	**0.6711**	**0.7673**	**0.7060**
5	ELASTIC NET	0.7158	0.7575	0.7412	0.7719
	MLP	0.7139	0.7530	0.7482	0.7815
	CLOPPS(MLP)	**0.7558**	**0.7893**	**0.7824**	**0.8105**

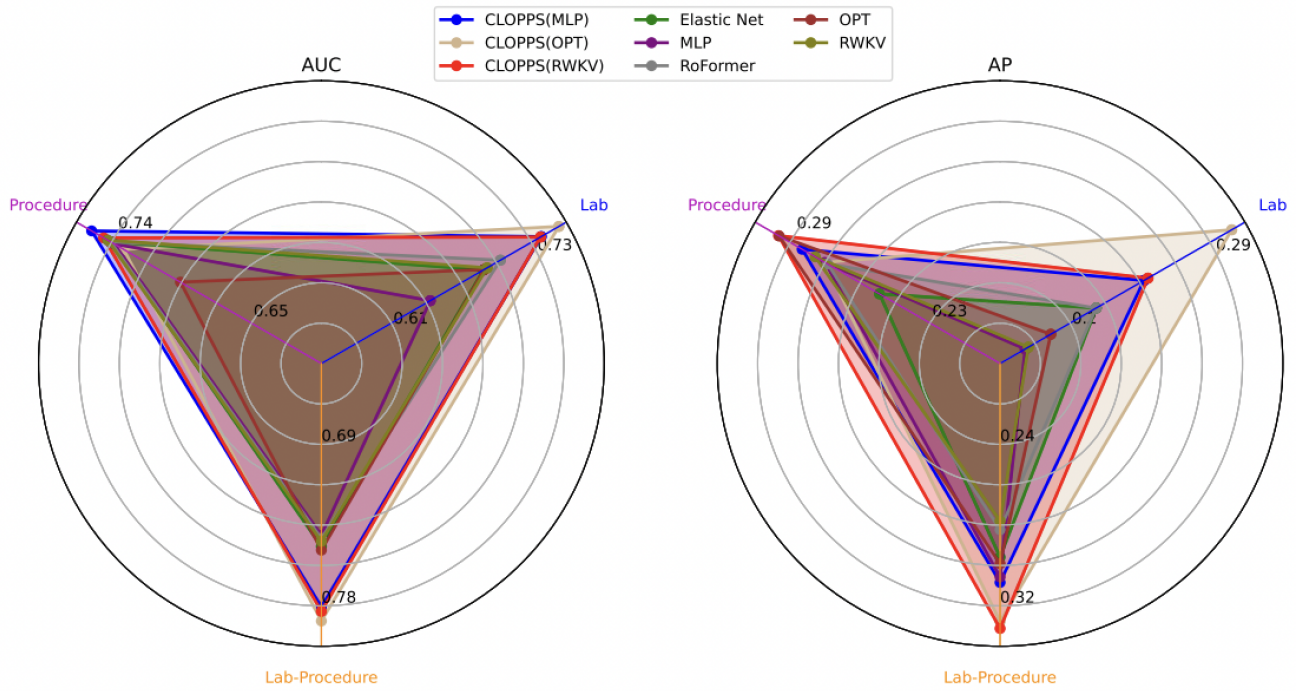

Figure 10. Performance comparison (AUC and AP) between CLOPPS and five baselines, including *shared covariates* on the MIMIC data. Axes colors indicate the data sources available during inference: procedures, labs and both labs and procedures. The name in parenthesis following CLOPPS indicates which classifier used.

Table 9. Performance comparison on simulated transfer learning scenarios on the Private Real-World Dataset. We show the performance for three baselines, Elastic Net, RoFormer, MLP, and CLOPPS using MLP as classifier.

M (YEARS)	MODEL	NATIONAL DATA SYSTEM AUC	NATIONAL DATA SYSTEM AP	DIALYSIS PROVIDER AUC	DIALYSIS PROVIDER AP
0.5	ELASTIC NET	0.7652	**0.3159**	0.7924	0.2957
	ROFORMER	0.7733	0.3023	0.7657	0.2589
	MLP	0.7634	0.3138	0.7846	0.2992
	CLOPPS(MLP)	**0.7798**	0.3145	**0.7987**	**0.3282**
1	ELASTIC NET	0.7353	0.4126	0.7640	0.4280
	ROFORMER	0.7499	0.4161	0.7481	0.4134
	MLP	0.7406	0.4168	0.7625	0.4451
	CLOPPS(MLP)	**0.7520**	**0.4235**	**0.7743**	**0.4584**
3	ELASTIC NET	0.7183	0.6439	0.7454	0.6635
	ROFORMER	0.7310	0.6501	0.7403	0.6592
	MLP	0.7218	0.6473	0.7516	**0.6789**
	CLOPPS(MLP)	**0.7368**	**0.6564**	**0.7568**	0.6773
5	ELASTIC NET	0.7157	0.6721	0.7393	0.6869
	ROFORMER	0.7149	0.6599	0.7137	0.6508
	MLP	0.7142	0.6685	0.7411	**0.6992**
	CLOPPS(MLP)	**0.7341**	**0.6844**	**0.7425**	0.6840

Table 10. Performance comparison between two training strategies on Private Real-World Dataset, where CLOPPS (FT) refers to fine-tuning mechanism and CLOPPS refers to the training mechanism introduced in Section 3.3.

M (YEARS)	MODEL	NATIONAL DATA SYSTEM		DIALYSIS PROVIDER	
		AUC	AP	AUC	AP
0.5	CLOPPS	**0.7772**	**0.3184**	**0.8100**	0.3608
	CLOPPS (FT)	0.7761	0.3162	0.8090	**0.3627**
1	CLOPPS	0.7527	**0.4202**	**0.7837**	0.4650
	CLOPPS (FT)	**0.7535**	0.4156	0.7824	**0.4661**
3	CLOPPS	0.7402	**0.6711**	**0.7673**	**0.7060**
	CLOPPS (FT)	**0.7439**	0.6711	0.7653	0.7042
5	CLOPPS	0.7558	**0.7893**	**0.7824**	0.8105
	CLOPPS (FT)	**0.7607**	0.7892	0.7818	**0.8113**