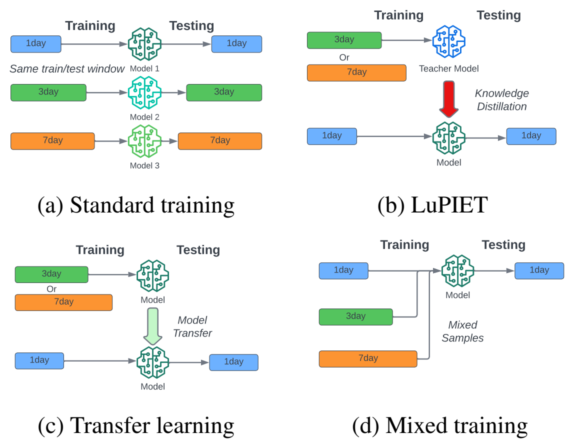
Improving Text-based Early Prediction by Distillation from Privileged Time-Series Text

Jinghui Liu[a,b] **Daniel Capurro**[a] **Anthony Nguyen**[b] **Karin Verspoor**[c,a]
[a] School of Computing and Information Systems, The University of Melbourne
[b] Australian e-Health Research Centre, CSIRO
[c] School of Computing Technologies, RMIT
jinghui.liu@student.unimelb.edu.au
dcapurro@unimelb.edu.au, anthony.nguyen@csiro.au
karin.verspoor@rmit.edu.au

Abstract

Modeling text-based time-series to make prediction about a future event or outcome is an important task with a wide range of applications. The standard approach is to train and test the model using the same input window, but this approach neglects the data collected in longer input windows between the prediction time and the final outcome, which are often available during training. In this study, we propose to treat this neglected text as privileged information available during training to enhance early prediction modeling through knowledge distillation, presented as **L**earning **u**sing **P**rivileged t**I**me-s**E**ries **T**ext (LuPIET). We evaluate the method on clinical and social media text, with four clinical prediction tasks based on clinical notes and two mental health prediction tasks based on social media posts. Our results show LuPIET is effective in enhancing text-based early predictions, though one may need to consider choosing the appropriate text representation and windows for privileged text to achieve optimal performance. Compared to two other methods using transfer learning and mixed training, LuPIET offers more stable improvements over the baseline, standard training. As far as we are concerned, this is the first study to examine learning using privileged information for time-series in the NLP context.



(a) Standard training (b) LuPIET

(c) Transfer learning (d) Mixed training

Figure 1: Different methods to leverage later data from the time-series to assist early prediction. LuPIET refers to learning using privileged time-series text in training via knowledge distillation. Here *1-day* is the baseline prediction window at test time. Models may leverage data from the prolonged training windows, e.g., *3-day*, to enhance the performance for the shorter test window.

1 Introduction

Time-series forecasting, or early prediction, is an important machine learning task with a wide range of applications, such as weather prediction (Krasnopolsky and Fox-Rabinovitz, 2006; Espeholt et al., 2022) and stock forecasting (Xu and Cohen, 2018; Sharma et al., 2017). Predicting future events or outcomes would enable timely responses that can bring significant social and economic benefits. Meanwhile, most existing works on forecasting or early prediction use structured measurements or features as input (Steyerberg, 2009), and studies to leverage unstructured text to explore temporal patterns are still scarce (Assale et al., 2019). Moreover, user-generated, domain-specific textual data, such as clinical and social media texts, can be noisy and complex to model (Baldwin et al., 2013; Huang et al., 2020). This creates challenges in utilizing text for early prediction.

The standard framework for early prediction trains and tests machine learning models using the same input window, depicted in Figure 1a. Though this is the widely adopted approach, it discards data that are outside the prediction window but are collected in practice as part of the training set. Ideally, these data can be utilized to enhance early prediction, such as learning the future trajectory of the time-series to assist modeling. Leveraging data available at training time but not at test time – referred to as *privileged information* – for training has been proposed as Learning using Privileged Information (LuPI) (Vapnik and Vashist, 2009; Vapnik and Izmailov, 2015). Recent studies have shown LuPI can be successfully applied to utilize time-series privileged information for early prediction (Hayashi et al., 2019; K.A. Karlsson et al., 2022). However, these experiments focus on structured features from synthetic data or distributions under certain assumptions. It remains unknown whether the approach applies to text-based early prediction applications, where natural language presents distinct characteristics and variation.

In this study, we adapt the time-series LuPI to textual data, presented as **L**earning **u**sing **P**rivileged t**I**me-s**E**ries **T**ext (LuPIET). We evaluate LuPIET on a range of tasks to evaluate its efficacy. LuPIET trains a more performant model using a longer predictive window that includes data created after the target prediction point as a teacher model. This is applied to guide the training of the early model through knowledge distillation. Figure 1b gives an example where the prediction window is *1-day*, and we aim to guide the training of the early model with the teacher model trained from a *3-day* window. To compare with LuPIET, we also apply two other methods, common in other domains but not well-examined for text-based time-series, to leverage data collected after the prediction time but available for training. These are transfer learning and mixed training, depicted in Figure 1c and Figure 1d, respectively.

We examine LuPIET using two challenging and domain-specific datasets containing clinical and social media text. Specifically, we explore four risk and diagnosis prediction tasks with clinical text and two mental health status prediction tasks with social media. The results show LuPIET can be an effective and stable approach for improving early prediction based on textual input.

In summary, this work examines the usefulness of privileged time-series information in the NLP context to support early prediction. Our main contributions include:

1. Proposing LuPIET to improve time-series modeling for text-based early prediction.

2. Evaluating the performance of LuPIET on two domains using clinical and social media corpora, presenting results on six prediction tasks. We show that when the privileged text is appropriately chosen and represented, LuPIET can improve over the baseline for early prediction by being more sample efficient.

3. Benchmarking the performances of two other competitive methods to support early prediction. We find although they can sometimes outperform LuPIET, they are in certain cases detrimental to modeling. LuPIET offers more consistent and stable improvements over the baseline.

2 Related Work

Early prediction with text Forecasting or early prediction has been widely studied in various domains and applications. For example, Steyerberg (2009) demonstrates the different facets and modeling strategies for clinical prediction modeling. Most initial works in the field focus on structured measurements as input features, with some attempts to extract and include shallow textual features or topics (Suresh et al., 2017; Ghassemi et al., 2015, 2014). More recent studies aim to put more stress on text by applying more powerful models to handle the complexity of language (Matero and Schwartz, 2020; Seinen et al., 2022). They have shown promise in modeling various types of text to support the prediction of mental health issues (Halder et al., 2017), stock market trends (Xu and Cohen, 2018), and clinical outcomes (Hsu et al., 2020).

Learning using privileged information LuPI presents a framework to leverage features only available at the train time but not at test time (Vapnik and Vashist, 2009). It has shown improved results in a range of applications, including recommendation (Xu et al., 2020) and image processing (Lee et al., 2020). Recently, the approach has been applied to improve early prediction using time-series data, which leverages data observed between the prediction time and the future outcome as privileged information. Hayashi et al. (2019) examines this approach on a synthetic dataset and a real-world dataset on air conditions with eight variables. K.A. Karlsson et al. (2022) further formalizes the framework for time-series as Learning using Privileged Time-Series (LuPTS) and proves it is guaranteed to result in more efficient learning when the time-series are drawn from a non-stationary Gaussian-linear dynamic system. However, none of the prior works examines text as input.

Knowledge distillation In knowledge distillation (KD), a more performant teacher model guides a smaller student model to achieve better results by matching the distributions of their predictions or output logits (Hinton et al., 2015). By training with the teacher output, the student model is provided with soft targets that contain more nuanced information about the label distribution compared to the true, hard labels. KD has been widely used for model compression (Sanh et al., 2019; Tung and Mori, 2019; Jiao et al., 2020) and other machine learning applications (Furlanello et al., 2018; Clark et al., 2019) to transfer knowledge across models with different strengths, sizes, or even architectures. In contrast, the classic transfer learning focuses on a single model and transfers knowledge across datasets, often from larger datasets to smaller ones (Devlin et al., 2019). Early works have shown the connection between LuPI and KD, unified them under *generalized distillation* (Lopez-Paz et al., 2016). Distillation has become a standard implementating technique to leverage privileged information (Hayashi et al., 2019; Xu et al., 2020).

3 Methods

Here we present the problem setting for early prediction and then describe how learning using privileged time-series text (LuPIET) works.

3.1 Problem setting

The goal for early prediction is to learn a mapping function $f(\theta)$ between input $X_{t,n} \in \mathcal{R}^{n\times d}$ and future events or outcomes $Y \in \mathcal{R}$, where $t = 1 \dots T$ is a time point of the time-series defined by the prediction window, and $n = 1 \dots N$ is a textual note or post available in the window X_t. Since N can vary across prediction windows, we neglect the notation of n from now on for simplicity. Note that $[X_1, X_2, ..., X_T]$ share the same label Y as they come from the same sample, where X_t is always a subset of X_{t+1}. We assume the *baseline* prediction window by setting $t = 1$, and the baseline model trained in Figure 1a is obtained as

$$\theta_{base} = \arg\min_{\theta\in\Theta} \mathcal{H}(f(X_1), Y) \tag{1}$$

where $\mathcal{H}$ is the cross entropy loss. We then aim to improve θ_{base} by leveraging texts created chronologically after X_1, namely $[X_2, X_3, ...X_T]$.

3.2 Learning with privileged time-series text

LuPIET optimizes a knowledge distillation loss that maps the predictions between the baseline model and the new model trained with privileged text, which can be viewed as a teacher model. We train this teacher model using input from a prolonged prediction window compared to baseline, namely X_t where $t \geq 2$, to obtain

$$\theta_t = \arg\min_{\theta\in\Theta} \mathcal{H}(f(X_t), Y) \tag{2}$$

Then let $p_{base}(x)$ and $p_t(x)$ be the output logits from the base model and teacher model, and we scale them with a temperature τ before taking the softmax, as defined in the original setting of knowledge distillation (Hinton et al., 2015):

$$p^{\tau}_{base}(x_i) = \frac{e^{p_{base}(x_i)/\tau}}{\sum_{j=1}^{K} e^{p_{base}(x_j)/\tau}} \tag{3}$$

$$p^{\tau}_{t}(x_i) = \frac{e^{p_{t}(x_i)/\tau}}{\sum_{j=1}^{K} e^{p_{t}(x_j)/\tau}} \tag{4}$$

where K is the number of labels. We calculate the distillation loss as the KL-divergence between the two scaled logit distributions as

$$\mathcal{L}_{KD} = \mathcal{D}_{KL}(p^{\tau}_{base}(x)||p^{\tau}_{t}(x)) \tag{5}$$

This distillation loss is then added to the cross entropy loss to train the final model that consumes the baseline input X_1, which is obtained by optimizing the following training objective:

$$\mathcal{L} = (1-\alpha)\mathcal{H}(f(X_1), Y) + \alpha\mathcal{L}_{KD} \tag{6}$$

where α is a hyperparameter and $0 \leq \alpha \leq 1$.

3.3 Other options to improve early baseline

Besides LuPIET, we also examine two other simple methods to leverage privileged text to assist early training — transfer learning and mixed training. Transfer learning (Zhuang et al., 2021) refers to further fine-tuning θ_t using X_1 by minimizing the standard cross entropy loss. In other words, we initialize the training in Eq 1 using the model parameters from Eq 2, as shown in Figure 1c.

In the mixed training method, the approach is to mix X_t with X_1 and train the model from scratch. This can be considered as a data augmentation approach (Wen et al., 2021) which enriches the training set and encourages the model to learn from all variations of the same sample. This is depicted in Figure 1d.

4 Experiments

We examine LuPIET using datasets from two challenging textual domains: clinical notes and social media posts. All datasets contain notes or posts that are created chronologically, naturally forming the text-based time series. Here we introduce them in more detail.

4.1 Clinical datasets and tasks

For clinical text, we use the MIMIC-III database (Johnson et al., 2016) to construct the datasets to predict four clinical outcomes and targets, which are in-hospital mortality, in-ICU mortality, length-of-stay (LOS) over 3 days, and diagnostic related groups (DRG). These are popular tasks in the literature for clinical early prediction and we follow previous works to define and extract cohorts for them (Wang et al., 2020; Liu et al., 2021, 2022a). We note that the early DRG prediction is different from the typical medical coding performed post-discharge (Dong et al., 2022; Liu et al., 2022b) as it aims to predict diagnosis and estimate care costs while patients are still in the hospital (Gartner et al., 2015; Islam et al., 2021).

For each patient in the cohort, we extract all clinical notes during the hospital course that are created before the prediction time, sort them chronologically, and remove empty or duplicated notes. The prediction window is defined by the number of days after the patient ICU admission. For example, the *3-day* window would include all clinical notes charted by the end of the third day of ICU admission.

We define the baseline window for the clinical prediction tasks as *1-day*, which would allow timely interventions and resource arrangements to be made for the hospitalized patients. This is also a common choice made in the literature (Wang et al., 2020; Hsu et al., 2020). We then examine two extended prediction windows to train LuPIET, which are *3-day* and *7-day*. Notice there are cases whose time-series lengths are less than *3-day* or *7-day*. We did not distinguish these cases from others in the data extraction process to ensure we can directly compare with baseline results. For example, if a case has a length of 2 days, then the input texts for this case under *3-day* and *7-day* are the same. The numbers of train/validation/test cases are presented in Table 1. Notice the two mortality predictions and LOS prediction share the same cohort.

4.2 Social media datasets and tasks

For social media texts, we focus on the eRisk 2018 datasets (Losada and Crestani, 2016; Losada et al., 2018) to use Reddit posts to predict potential mental health issues, which are depression and anorexia. Predicting mental health status using social media data is an important yet challenging task (Guntuku et al., 2017) that draws much attention in recent years from the NLP community (Benton et al., 2017; Cohan et al., 2018). To parse the datasets, we use both the title and the content as input text for the post. The datasets present the posts in ten chunks, which are evenly split by time and sorted in the chronological order. We follow this format to define the prediction window by the number of chunks used as input, e.g., a *3-chunk* window includes all posts in the first three chunks.

	# train	# validation	# test	# labels
Mortality & LOS	26729	3407	3392	2
DRG	16296	972	1866	570
eRisk Depression	656	82	82	2
eRisk Anorexia	376	47	48	2

Table 1: Number of cases for the train, validation, and test sets of the examined prediction tasks.

Since social media text could present noisier temporal patterns, we examine two baseline windows for the prediction tasks, which are *3-chunk* and *7-chunk*, and only use the full *10-chunk* as prolonged prediction window. We make this choice also because the eRisk datasets are relatively small and the performance can be unstable. We use all samples from the datasets for each task and split them in a 0.8/0.1/0.1 ratio, and the numbers are again presented in Table 1.

4.3 Text representation and modeling

We examine two text representations for the text-based time-series modeling. The first one is to model all input text as a single text string by concatenating all notes or posts together. This allows us to model the input as a sequence of words, which considers the word-level details. This is a standard practice in NLP. The other representation is to encode all notes or posts into document embeddings and model them at the document level. This method, on the other hand, may lose details during the document encoding process, but it helps to maintain the temporal patterns of the texts.

For word-level representation, we use domain-specific pretrained word embeddings for the two datasets. Specifically, we use BioWordVec (Zhang et al., 2019) for clinical text and GloVe-840B (Pennington et al., 2014) for social media. Since the concatenated text string is rather long, we adapt MultiResCNN (Li and Yu, 2020) for modeling, which is a CNN-based model that has shown strong results in long-document classification. The model enhances the vanilla text-CNN model by adding more filters with residual connections. Given space restrictions here, we do not introduce the architecture in detail and refer readers to the original paper for more information.

For document-level modeling, we first encode notes or posts using BERT (Devlin et al., 2019) by extracting the [CLS] token as the final representation. We again try to align BERT with the domains of the text by using ClinicalBERT (Alsentzer et al., 2019) for clinical notes and RoBERTa (Liu et al., 2019) for Reddit posts. We then apply LSTM (Hochreiter and Schmidhuber, 1997) to model the sequence of notes and use the last hidden state for final prediction. We use LSTM due to its power to extract and model temporal patterns.

4.4 Training and evaluation

We tune the hyperparameters for all examined methods based on the validation set and then retrain the model with the best configuration with multiple random seeds. Specifically, we tune the filter number and sizes for the CNN model, hidden size for the LSTM model, and the number of layer, dropout, learning rate, weight decay for both models. Adam (Kingma and Ba, 2014) is used as the optimizer throughout the experiments. After obtaining the best architectural configuration for the baseline model, we fix the setup and tune τ and α for LuPIET. We adopt random search for the clinical datasets and grid search for social media datasets given the former takes much longer to run. We facilitate the hyperparameter searching with asynchronous successive halving algorithm (Li et al., 2020), based on the implementation from Ray Tune (Liaw et al., 2018). We perform all our experiments using pytorch (Paszke et al., 2019) and pytorch-lightning[1] on V100 GPUs.

For evaluation, we use area under the receiver operating characteristics curve (AUROC) and precision-recall curve (AUPR) for the binary classification tasks, and accuracy and macro F1 for multiclass classification. We run the final, tuned model for each task with five random seeds for the clinical datasets and average the results, and run with ten seeds for the social media datasets given the small data size.

[1] https://www.pytorchlightning.ai/

5 Results

5.1 Comparing LuPIET with baseline

We present our main results with the word-level modeling for clinical prediction tasks in Table 2. We observe LuPIET using the extended *3-day* and *7-day* windows improves over the standard baseline window on *1-day* for in-hospital mortality, in-ICU mortality, and DRG predictions. Meanwhile, LuPIET does not provide much benefit for predicting LOS>3days. We believe this is related to the nature of the task as the extended windows are already longer than the target in consideration, so the teacher models can take advantage of this shortcut to make accurate and confident predictions that are close to the true labels and can no longer serve as soft targets (Hinton et al., 2015; Cho and Hariharan, 2019).

The results for social media datasets are presented in Table 3. Here we examine *3-chunk* and *7-chunk* as the baseline windows and apply LuPIET trained using full length, i.e., *10-chunk*. We again observe the benefit of LuPIET over the baseline results, especially under AUROC. However, we note that due to the much smaller dataset size, the variances in the results can be large. In a couple of cases, LuPIET achieves slightly lower AUPR scores than the baseline, but the results are still comparable given the variance.

5.2 Comparing LuPIET with other methods

We focus on the clinical datasets to compare LuPIET with transfer learning and mixed training approaches given they have relatively sufficient data to observe their behavior. In Table 2, we find LuPIET still performs strongly when compared to these two approaches, but other methods may outperform LuPIET, such as mixed training on DRG prediction. In this particular case, we believe this is caused by the DRG dataset being a multiclassification task with 570 labels, thus not having enough samples for training. By mixing different windows of the same sample together serves as a data augmentation strategy, which alleviates the low-resource situation for DRG to achieve better results. Similarly, when modeling at the note level (Table 4), transfer learning may be able to better transfer the temporal relations across windows thus achieving slightly better results.

However, we observe that transfer learning and mixed training do not always improve over the baseline, which is further shown when we examine

	In-hospital Mortality		In-ICU Mortality		LOS>3days		DRG	
	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	Acc	MacroF1
Baseline								
1-day	0.867 (0.0032)	0.474 (0.0151)	0.862 (0.0110)	0.373 (0.0252)	0.690 (0.0034)	0.625 (0.0047)	0.282 (0.0058)	0.105 (0.0106)
LuPIET								
3-day → 1-day	0.876 (0.0027)	0.491 (0.0039)	**0.880 (0.0020)**	**0.429 (0.0092)**	0.693 (0.0042)	0.626 (0.0058)	0.287 (0.0068)	0.116 (0.0067)
7-day → 1-day	**0.879 (0.0029)**	**0.501 (0.0093)**	0.880 (0.0024)	0.413 (0.0076)	0.692 (0.0035)	**0.627 (0.0041)**	0.290 (0.0097)	0.115 (0.0130)
Transfer								
3-day → 1-day	0.863 (0.0026)	0.466 (0.0063)	0.866 (0.0060)	0.381 (0.0231)	**0.695 (0.0036)**	0.612 (0.0049)	0.293 (0.0033)	0.134 (0.0059)
7-day → 1-day	0.865 (0.0040)	0.482 (0.0073)	0.860 (0.0051)	0.379 (0.0089)	0.691 (0.0046)	0.616 (0.0068)	0.284 (0.0027)	0.113 (0.0072)
7-day → 3-day → 1-day	0.864 (0.0040)	0.482 (0.0110)	0.855 (0.0073)	0.374 (0.0098)	0.683 (0.0031)	0.606 (0.0054)	0.285 (0.0093)	0.112 (0.0081)
Mix-train								
1-day + 3-day + 7-day	0.866 (0.0031)	0.490 (0.0031)	0.860 (0.0079)	0.398 (0.0077)	0.642 (0.0037)	0.561 (0.0057)	**0.298 (0.0025)**	**0.140 (0.0081)**

Table 2: Results of word-level modeling for the four clinical prediction tasks.

	Depression		Anorexia	
	AUROC	AUPR	AUROC	AUPR
Baseline - *3-chunk*	0.819 (0.0636)	**0.458 (0.1606)**	0.796 (0.0742)	0.334 (0.1071)
+ ***LuPIET***	**0.868 (0.0203)**	0.431 (0.1150)	**0.830 (0.0742)**	**0.339 (0.0963)**
Baseline - *7-chunk*	0.836 (0.0303)	0.470 (0.1374)	0.798 (0.0401)	**0.429 (0.1217)**
+ ***LuPIET***	**0.869 (0.0332)**	**0.495 (0.0792)**	**0.807 (0.0239)**	0.423 (0.0705)

Table 3: Results for the two mental health prediction tasks using social media posts.

the note-level modeling results in Table 4. For example, though it could bring benefits to tasks like DRG, mixed training is rather detrimental to the LOS>3days prediction. On the other hand, LuPIET either improves over the baseline or at least maintains the performance under both text representation and modeling strategies.

We also see LOS>3days task does not benefit much from transfer learning and mixed training either, similar to the results with LuPIET. This shows when adopting these methods to improve the time-series modeling, it can be important to consider the nature of the task and to choose proper windows accordingly.

5.3 Comparing text representations

The results presented in Table 2 and Table 4 for word- and note-level modeling are comparable for all the four tasks. Overall, we see modeling at word level performs much better than at note level, demonstrating the need to attend to the fine-grained textual details for these tasks. The LOS>3days is again an exception where the two sets of results are similar, showing the task can be inherently more difficult. Furthermore, modeling at word level tends to benefit more from LuPIET. For example, we see much better mortality prediction scores with LuPIET in Table 2. This indicates that the proper training of LuPIET exploits the nuances in the data and it would benefit more when modeling at a finer granularity.

We do not present the results with document embeddings for the social media tasks as their AUROC results are sub-optimal and barely over 0.5. We suspect this is because the Reddit posts are much noisier and the model needs to sift away much unrelated information, so encoding all posts into embeddings is unhelpful.

6 Discussion

6.1 Sampling efficiency

K.A. Karlsson et al. (2022) shows under certain conditions, such as when the time-series is from linear dynamical systems with Gaussian noise and is in Markov structure, privileged information is guaranteed to improve the learning efficiency of time-series models. Given these conditions do not necessarily hold for the natural language that has distinctive data distribution, here we empirically ex-

	In-hospital Mortality		In-ICU Mortality		LOS>3days		DRG	
	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	Acc	MacroF1
Baseline								
1-day	0.844 (0.0058)	0.417 (0.0122)	0.860 (0.0049)	0.369 (0.0172)	0.695 (0.0059)	0.629 (0.0061)	0.228 (0.0057)	0.056 (0.0043)
LuPIET								
3-day → 1-day	0.850 (0.0040)	0.428 (0.0108)	0.861 (0.0060)	0.362 (0.0178)	0.696 (0.0034)	0.631 (0.0054)	**0.238 (0.0063)**	0.056 (0.0055)
7-day → 1-day	0.851 (0.0033)	0.430 (0.0125)	0.860 (0.0046)	0.370 (0.0168)	**0.698 (0.0027)**	**0.635 (0.0039)**	0.237 (0.0042)	0.057 (0.0029)
Transfer								
3-day → 1-day	**0.853 (0.0038)**	**0.439 (0.0080)**	0.862 (0.0069)	**0.384 (0.0093)**	0.688 (0.0022)	0.615 (0.0044)	0.232 (0.0067)	0.059 (0.0043)
7-day → 1-day	0.833 (0.0037)	0.415 (0.0086)	0.861 (0.0065)	0.371 (0.0123)	0.686 (0.0056)	0.623 (0.0057)	0.230 (0.0070)	0.064 (0.0039)
7-day → 3-day → 1-day	0.829 (0.0073)	0.405 (0.0108)	0.861 (0.0080)	0.362 (0.0147)	0.688 (0.0029)	0.621 (0.0022)	0.232 (0.0052)	**0.066 (0.0049)**
Mix-train								
1-day + 3-day + 7-day	0.838 (0.0019)	0.416 (0.0113)	**0.866 (0.0055)**	0.382 (0.0125)	0.671 (0.0035)	0.605 (0.0066)	0.236 (0.0073)	0.061 (0.0043)

Table 4: Results of document-level modeling for the four clinical prediction tasks.

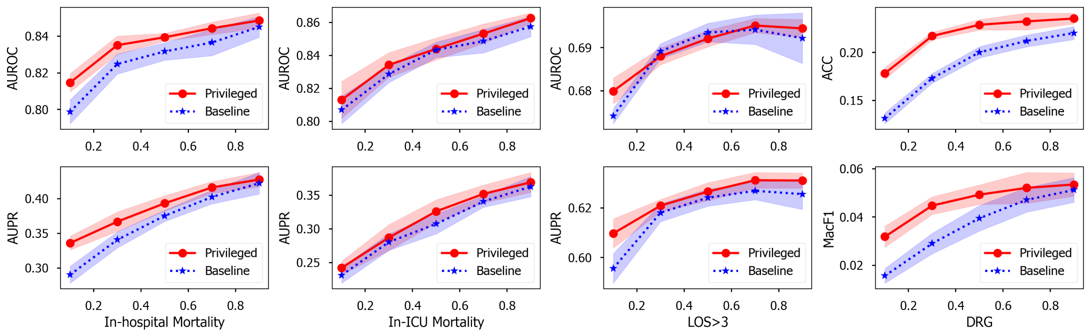


Figure 2: Learning curves for the four clinical prediction tasks under two evaluation metrics. The blue curves depict the results of standard training and the red curves depict those of LuPIET.

amine if privileged text can still benefit NLP modeling efficiency without making other assumptions. Figure 2 shows the learning curves for the four clinical prediction tasks, with x-axis is the ratio of the full dataset used to train the model. We find LuPIET can lead to more sample efficient learning in multiple tasks. For example, when training with only 10% of the whole dataset, privileged learning achieves significant improvements over the baseline on in-hospital mortality, LOS, and DRG. When the model is fed with more samples, we see the difference between LuPIET and the baseline gradually converge. This also happens to the much smaller social media datasets, where we see the larger extent of improvements with LuPIET under AUROC in Table 3. Furthermore, LuPIET could reduce the modeling variance compared to baseline, consistent with findings on structured time-series (K.A. Karlsson et al., 2022).

Meanwhile, in certain scenarios, such as for the in-ICU mortality prediction and for the LOS prediction with sufficient data, we find the benefits of LuPIET are much weaker. This may reflect our observations on LOS in Sec 5.1, where the choice of input windows can be important. Extending input windows in certain scenarios may not necessarily include more data. For instance, in the ICU admissions stays are much shorter compared to hospital stays. This may explain why little benefit is observed for in-ICU mortality prediction.

6.2 Limitations

There are a few limitations with our study. Firstly, we find it is important to apply LuPIET to appropriate task settings but we did not formalize the definition of appropriateness, though we offer some possible intuitions (e.g., on the case of LOS in Sec 5.1). Domain knowledge about the nature of the task may be needed to realize optimal results with LuPIET, which could in some ways correspond to the assumptions about data distribution made for successful LuPI in previous works (Lopez-Paz et al., 2016; K.A. Karlsson et al., 2022). Future works are needed to provide the theoretical explanation for the empirical results and to guide more effective application of LuPIET.

Secondly, we find sometimes the teacher models do not benefit from extending the input window and consuming more time-series data, for example, the shorter *3-chunk* can outperform *7-chunk* for anorexia prediction (Table 3). We did not further investigate this phenomenon and leave it to future work to explore the potential causes. We also focused on the specific time steps for baseline and extended input windows and did not evaluate in more time steps, but in the future we would like to consider various time steps in the time-series for both training and evaluation. Similar evaluation setup has been examined in prior work (Harutyunyan et al., 2019), such as framing patient deterioration assessment as hourly patient mortality predictions. We regard this setup a promising extension of our current experiments to examine LuPIET. Lastly, we do not explore how factors in successful KD applications (Gou et al., 2021), such as creating consistent teacher models (Beyer et al., 2022), affect LuPIET, which we consider as a future direction to better utilize privileged information and to enhance LuPI in general.

7 Conclusion

In this study, we present LuPIET, a framework to incorporate longer-range time-series data available during training to improve text-based early predictions. Though similar ideas have been examined recently for structured time-series (Hayashi et al., 2019; K.A. Karlsson et al., 2022), we are not aware of any previous studies on the use of this privileged information in the context of text-based time-series. We find LuPIET is an effective strategy for enhancing early prediction and for efficient time-series modeling when applied to appropriate task settings.

LuPIET is implemented by simply optimizing a distillation loss. Therefore, future works may extend LuPIET by training with more advanced distillation techniques, e.g., matching hidden state instead of logits (Zhang et al., 2018), or combining with other inputs, e.g., using multi-modal privileged information.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful reviews and comments. This research was undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne. This Facility was established with the assistance of LIEF Grant LE170100200.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Michela Assale, Linda Greta Dui, Andrea Cina, Andrea Seveso, and Federico Cabitza. 2019. The revival of the notes field: Leveraging the unstructured content in electronic health records. *Frontiers of medicine*, 6:66.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364, Nagoya, Japan. Asian Federation of Natural Language Processing.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162, Valencia, Spain. Association for Computational Linguistics.

Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. 2022. Knowledge distillation: A good teacher is patient and consistent. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10915–10924.

Jang Hyun Cho and Bharath Hariharan. 2019. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4794–4802.

Kevin Clark, Minh-Thang Luong, Urvashi Khandelwal, Christopher D Manning, and Quoc V Le. 2019. BAM! Born-Again Multi-Task networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5931–5937, Florence, Italy. Association for Computational Linguistics.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018.

SMHD: a Large-Scale resource for exploring online language usage for multiple mental health conditions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1485–1497, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hang Dong, Matúš Falis, William Whiteley, Beatrice Alex, Joshua Matterson, Shaoxiong Ji, Jiaoyan Chen, and Honghan Wu. 2022. Automated clinical coding: what, why, and where we are? *NPJ digital medicine*, 5(1):159.

Lasse Espeholt, Shreya Agrawal, Casper Sønderby, Manoj Kumar, Jonathan Heek, Carla Bromberg, Cenk Gazen, Rob Carver, Marcin Andrychowicz, Jason Hickey, Aaron Bell, and Nal Kalchbrenner. 2022. Deep learning for twelve hour precipitation forecasts. *Nature communications*, 13(1):5145.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born again neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1607–1616. PMLR.

Daniel Gartner, Rainer Kolisch, Daniel B Neill, and Rema Padman. 2015. Machine learning approaches for early DRG classification and resource allocation. *INFORMS journal on computing*, 27(4):718–734.

Marzyeh Ghassemi, Tristan Naumann, Finale Doshi-Velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits. 2014. Unfolding physiological state: mortality modelling in intensive care units. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '14, pages 75–84, New York, NY, USA. Association for Computing Machinery.

Marzyeh Ghassemi, Marco A F Pimentel, Tristan Naumann, Thomas Brennan, David A Clifton, Peter Szolovits, and Mengling Feng. 2015. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015:446–453.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International journal of computer vision*, 129(6):1789–1819.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Kishaloy Halder, Lahari Poddar, and Min-Yen Kan. 2017. Modeling temporal progression of emotional status in mental health forum: A recurrent neural net approach. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 127–135, Copenhagen, Denmark. Association for Computational Linguistics.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96.

Shogo Hayashi, Akira Tanimoto, and Hisashi Kashima. 2019. Long-Term prediction of small Time-Series data using generalized distillation. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

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

Chao-Chun Hsu, Shantanu Karnwal, Sendhil Mullainathan, Ziad Obermeyer, and Chenhao Tan. 2020. Characterizing the value of information in medical notes. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2062–2072, Online. Association for Computational Linguistics.

Rongtao Huang, Bowei Zou, Yu Hong, Wei Zhang, Aiti Aw, and Guodong Zhou. 2020. NUT-RC: Noisy user-generated text-oriented reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2687–2698, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Md Mohaimenul Islam, Guo-Hung Li, Tahmina Nasrin Poly, and Yu-Chuan Jack Li. 2021. DeepDRG: Performance of artificial intelligence model for Real-Time prediction of Diagnosis-Related groups. *Healthcare (Basel, Switzerland)*, 9(12).

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035.

Rickard K.A. Karlsson, Martin Willbo, Zeshan M Hussain, Rahul G Krishnan, David Sontag, and Fredrik Johansson. 2022. Using time-series privileged information for provably efficient learning of prediction models. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 5459–5484. PMLR.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Vladimir M Krasnopolsky and Michael S Fox-Rabinovitz. 2006. Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction. *Neural networks: the official journal of the International Neural Network Society*, 19(2):122–134.

Wonkyung Lee, Junghyup Lee, Dohyung Kim, and Bumsub Ham. 2020. Learning with privileged information for efficient image Super-Resolution. In *Computer Vision – ECCV 2020*, pages 465–482. Springer International Publishing.

Fei Li and Hong Yu. 2020. ICD coding from clinical text using Multi-Filter residual convolutional neural network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(5):8180–8187.

Liam Li, Kevin Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Jonathan Ben-tzur, Moritz Hardt, Benjamin Recht, and Ameet Talwalkar. 2020. A system for massively parallel hyperparameter tuning. In *Proceedings of Machine Learning and Systems*, volume 2, pages 230–246.

Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez, and Ion Stoica. 2018. Tune: A research platform for distributed model selection and training.

Jinghui Liu, Daniel Capurro, Anthony Nguyen, and Karin Verspoor. 2021. Early prediction of diagnostic-related groups and estimation of hospital cost by processing clinical notes. *NPJ digital medicine*, 4(1):103.

Jinghui Liu, Daniel Capurro, Anthony Nguyen, and Karin Verspoor. 2022a. "Note Bloat" impacts deep learning-based NLP models for clinical prediction tasks. *Journal of biomedical informatics*, 133:104149.

Leibo Liu, Oscar Perez-Concha, Anthony Nguyen, Vicki Bennett, and Louisa Jorm. 2022b. Hierarchical label-wise attention transformer model for explainable ICD coding. *Journal of biomedical informatics*, 133:104161.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. 2016. Unifying distillation and privileged information. In *International Conference on Learning Representations*.

David E Losada and Fabio Crestani. 2016. A test collection for research on depression and language use. In *Proc. of Experimental IR Meets Multilinguality, Multimodality, and Interaction, 7th International Conference of the CLEF Association, CLEF 2016*, pages 28–39. Springer International Publishing.

David E Losada, Fabio Crestani, and Javier Parapar. 2018. Overview of erisk: Early risk prediction on the internet. In *Proc. of Experimental IR Meets Multilinguality, Multimodality, and Interaction, International Conference of the CLEF Association, CLEF 2018*, pages 343–361. Springer International Publishing.

Matthew Matero and H Andrew Schwartz. 2020. Autoregressive affective language forecasting: A Self-Supervised task. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2913–2923, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, High-Performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.

Tom M Seinen, Egill A Fridgeirsson, Solomon Ioannou, Daniel Jeannetot, Luis H John, Jan A Kors, Aniek F Markus, Victor Pera, Alexandros Rekkas, Ross D Williams, Cynthia Yang, Erik M van Mulligen, and Peter R Rijnbeek. 2022. Use of unstructured text in prognostic clinical prediction models: a systematic review. *Journal of the American Medical Informatics Association: JAMIA*, 29(7):1292–1302.

Ashish Sharma, Dinesh Bhuriya, and Upendra Singh. 2017. Survey of stock market prediction using machine learning approach. In *2017 International conference of Electronics, Communication and*

Aerospace Technology (ICECA), volume 2, pages 506–509. ieeexplore.ieee.org.

E W Steyerberg. 2009. Applications of prediction models. In Ewout W Steyerberg, editor, *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*, pages 11–31. Springer New York, New York, NY.

Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. 2017. Clinical intervention prediction and understanding with deep neural networks. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 322–337. PMLR.

Frederick Tung and Greg Mori. 2019. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1365–1374.

Vladimir Vapnik and Rauf Izmailov. 2015. Learning using privileged information: Similarity control and knowledge transfer. *Journal of machine learning research: JMLR*, 16(61):2023–2049.

Vladimir Vapnik and Akshay Vashist. 2009. A new learning paradigm: learning using privileged information. *Neural networks: the official journal of the International Neural Network Society*, 22(5-6):544–557.

Shirly Wang, Matthew B A McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C Hughes, and Tristan Naumann. 2020. MIMIC-Extract: a data extraction, preprocessing, and representation pipeline for MIMIC-III. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, pages 222–235, New York, NY, USA. Association for Computing Machinery.

Qingsong Wen, Liang Sun, Fan Yang, Xiaomin Song, Jingkun Gao, Xue Wang, and Huan Xu. 2021. Time series data augmentation for deep learning: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4653–4660. International Joint Conferences on Artificial Intelligence Organization.

Chen Xu, Quan Li, Junfeng Ge, Jinyang Gao, Xiaoyong Yang, Changhua Pei, Fei Sun, Jian Wu, Hanxiao Sun, and Wenwu Ou. 2020. Privileged features distillation at taobao recommendations. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pages 2590–2598, New York, NY, USA. Association for Computing Machinery.

Yumo Xu and Shay B Cohen. 2018. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1970–1979, Melbourne, Australia. Association for Computational Linguistics.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific data*, 6(1):52.

Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. 2018. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4320–4328.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2021. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76.