Journal of the American Medical Informatics Association, 2024, **31(6)**, 1322–1330
https://doi.org/10.1093/jamia/ocae088
Advance access publication 29 April 2024
Research and Applications

Research and Applications

Development and external validation of deep learning clinical prediction models using variable-length time series data

Fereshteh S. Bashiri, PhD[1], Kyle A. Carey, MPH[2], Jennie Martin, MS[1], Jay L. Koyner, MD[2], Dana P. Edelson, MD, MS[2], Emily R. Gilbert, MD[3], Anoop Mayampurath, PhD[1,4], Majid Afshar, MD, MSCR[1,4], Matthew M. Churpek, MD, MPH, PhD[1,4,]*

[1]Department of Medicine, University of Wisconsin-Madison, Madison, WI 53792, United States, [2]Department of Medicine, University of Chicago, Chicago, IL 60637, United States, [3]Department of Medicine, Loyola University, Chicago, IL 60153, United States, [4]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53726, United States

*Corresponding author: Matthew M. Churpek, MD, MPH, PhD, Departments of Medicine and Biostatistics and Medical Informatics, University of Wisconsin-Madison, 600 Highland Ave, Madison, WI 53792, United States (mchurpek@medicine.wisc.edu)

Abstract

Objectives: To compare and externally validate popular deep learning model architectures and data transformation methods for variable-length time series data in 3 clinical tasks (clinical deterioration, severe acute kidney injury [AKI], and suspected infection).

Materials and Methods: This multicenter retrospective study included admissions at 2 medical centers that spanned 2007-2022. Distinct datasets were created for each clinical task, with 1 site used for training and the other for testing. Three feature engineering methods (normalization, standardization, and piece-wise linear encoding with decision trees [PLE-DTs]) and 3 architectures (long short-term memory/gated recurrent unit [LSTM/GRU], temporal convolutional network, and time-distributed wrapper with convolutional neural network [TDW-CNN]) were compared in each clinical task. Model discrimination was evaluated using the area under the precision-recall curve (AUPRC) and the area under the receiver operating characteristic curve (AUROC).

Results: The study comprised 373 825 admissions for training and 256 128 admissions for testing. LSTM/GRU models tied with TDW-CNN models with both obtaining the highest mean AUPRC in 2 tasks, and LSTM/GRU had the highest mean AUROC across all tasks (deterioration: 0.81, AKI: 0.92, infection: 0.87). PLE-DT with LSTM/GRU achieved the highest AUPRC in all tasks.

Discussion: When externally validated in 3 clinical tasks, the LSTM/GRU model architecture with PLE-DT transformed data demonstrated the highest AUPRC in all tasks. Multiple models achieved similar performance when evaluated using AUROC.

Conclusion: The LSTM architecture performs as well or better than some newer architectures, and PLE-DT may enhance the AUPRC in variable-length time series data for predicting clinical outcomes during external validation.

Key words: variable-length time series; deep learning; AI in medicine.

Background and significance

In healthcare, structured data (eg, laboratory results, vital signs, etc.) from electronic health records (EHRs) have proven helpful in predicting patient outcomes, diagnosing disease, and augmenting bedside decision-making.[1] For example, structured time-series data have been used to develop machine learning models to predict events in hospitalized patients.[2–4] Conditions such as clinical deterioration,[5,6] sepsis,[7] and acute kidney injury (AKI)[8] are common clinical events that contribute to mortality and morbidity of hospitalized patients and have been popular targets for machine learning algorithms. Recent evidence suggests that implementing machine learning models into clinical practice may improve patient outcomes, increasing the importance of developing accurate models that can be used in real-world applications.[9,10]

Although non-deep learning approaches have historically been used on structured data, recent advancements in deep learning have shown promise in this area.[11] However, several important questions remain. First, time series data from the EHR are recorded irregularly and have variable lengths across patients. The conventional fixed-length input approach of truncating variable-length time series into a limited and predetermined number of time steps used by most benchmarking studies is not pragmatic for healthcare delivery, where decision-making is dynamic over the course of patient care.[12,13] Additionally, many clinical applications, such as identifying clinical deterioration, require continuous output of predictions whenever new data are collected. In contrast, most benchmarking studies predict outcomes at fixed time points (eg, predicting in-hospital mortality using data within 48 hours after admission).[14,15] Therefore, there is a need to test these approaches in more pragmatic settings, where predictions are dynamic. Finally, numerous studies have presented novel deep learning architectures for time series?data as well as advancements in structured data pre-

Received: December 8, 2023. **Revised:** February 27, 2024. **Editorial Decision:** April 4, 2024. **Accepted:** April 5, 2024

processing.[13,16]?However, these studies often perform limited?hyperparameter tuning, which can affect model performance, and most develop and test their models in the same population (eg, using the Medical Information Mart for Intensive Care [MIMIC] dataset[17]) as opposed to testing external validity in other health systems.[13,15] To avoid overfitting and evaluate model generalizability, models must be tested in new cohorts.

Objective

To address these limitations in the field, we develop and externally validate a suite of pre-processing methods and deep learning architectures for structured time series data with variable lengths. Our pipeline provides extensive tuning of hyperparameters through Bayesian Optimization (BO) and is utilized to test the generalizability of models for several common medical applications. Our findings have important implications for researchers developing deep learning models for real-world applications with structured time series data.

Materials and methods

Study population and data collection

All adult (age ≥18 years) inpatients hospitalized at the University of Chicago Medicine (UCM) between 2008 and 2022 (training site) and Loyola University Medical Center (LUMC) between 2007 and 2017 (external test site) were eligible for inclusion. The study protocol was granted a waiver of informed consent by the institutional review board (IRB) (IRB # 2019-1258, 2019-1124, and 2019-1425).

Study variables were collected from the clinical research data warehouse at each site and contained timestamped records rounded to the minute, which formed the longitudinal structured dataset. These data included patient demographics, hospital unit, vital signs, laboratory results, and nursing assessments from the hospital encounter. Orders, medication administrations, and prior billing codes from previous encounters within the hospital system were also accessible for use in our study. The study cohort was used to create datasets for 3 tasks: (1) clinical deterioration; (2) severe AKI; and (3) suspected infection, as described below. Figure 1 depicts an overview of the study design.

Clinical deterioration task

All patients admitted to the medical-surgical wards were included in this task. The outcome of clinical deterioration was defined as either a direct ward to ICU transfer or an in-hospital death within 24 hours of a ward record. The event of death was considered to occur at the time of the last vitals and was determined using the encounter's discharge disposition.

Severe AKI task

The severe AKI cohort was created by excluding hospitalized encounters with any of the following conditions to remove patients with pre-existing severe kidney disease: (1) prior end-stage renal disease (ESRD) using International Classification of Diseases (ICD) billing codes; (2) undergoing dialysis within 48 hours of their initial creatinine measurement; (3) first creatinine ≥3mg/dL; or (4) no creatinine records during the encounter, as previously described.[3,18] As per the creatinine-based Kidney Disease Improving Global Outcomes (KDIGOs) definition, the outcome of severe AKI (stage 2/3) was characterized by 1 of 3 conditions: (1) the current creatinine level was ≥2 times higher than the baseline creatinine recorded 1 week prior; (2) the current creatinine level was ≥4mg/dL; or (3) the patient underwent dialysis or renal replacement therapy (RRT).[19] The baseline creatinine was defined as the first creatinine measurement of the encounter, and this baseline was continuously updated on a rolling basis, following the KDIGO guidelines.

Suspected infection task

The task of identifying suspected infection—a key criterion for sepsis—included all hospitalized patients. The outcome for this task is based on the infection criteria published by Seymour et al.[20] The outcome occurred if an encounter had received antibiotics (excluding those administered in the operating room [OR] and single doses of non-OR antibiotics) and had any culture order, and either (1) a culture was ordered within 24 hours after antibiotics were given, or (2) antibiotics were given within 72 hours after a culture was ordered. The accuracy of this definition for identifying infected patients has been confirmed previously.[21]

Predictor variables and preprocessing

Predictor variables included patient demographics, location, vital signs, laboratory values, and nursing documentation (eg, Braden Scale) selected through a review of existing literature[2,3,12,18] and insights provided by subject matter experts (M.M.C., D.P.E., J.L.K., M.A.). Table S1 provides a list of all predictor variables, demonstrating 60 variables in the AKI and infection datasets and 57 variables in the deterioration dataset (excluding location because all patients were on the wards). Three different data transformation methods for the predictor variables were compared within each task based on the data distribution in training (UCM) datasets: (1) min-max normalization of features into a fixed [0, 1] range; (2) standardization to achieve zero-mean and unit-variance features; and (3) piece-wise linear encoding with decision tree (PLE-DT)[22] bin edges.

The variable-length time series data allowed the natural flow of data by keeping the timestamp of variables rounded to a minute, unlike prior studies that use a common practice of hour-blocking, enabling the generation of predictions at every time step as new observations become available. When multiple values for the same variable occurred within a minute, 1 value was chosen at random. An "hours since admission" variable was included as a feature so that the models were aware of the elapsed time since the start of the admission and the time between observations. For each task, the time series included data only up to hospital discharge or the initial time point when the actual clinical outcome occurred, whichever came first. Because the aim was to predict clinical events before their occurrence, binary outcome labels (0/1) for each time step were determined based on the difference between the observation time of each time step and the time of the future outcome event. Specifically, a label of "1" indicated an event occurring within the next 24 hours for the deterioration and AKI tasks and within 6 hours for the infection task. These time horizons are consistent with prior studies for clinical deterioration and AKI, and the shorter time horizon for infection was due to many cases occurring soon after admission. All other timesteps had a "0" for the outcome. Non-physiologic outliers were removed, as shown

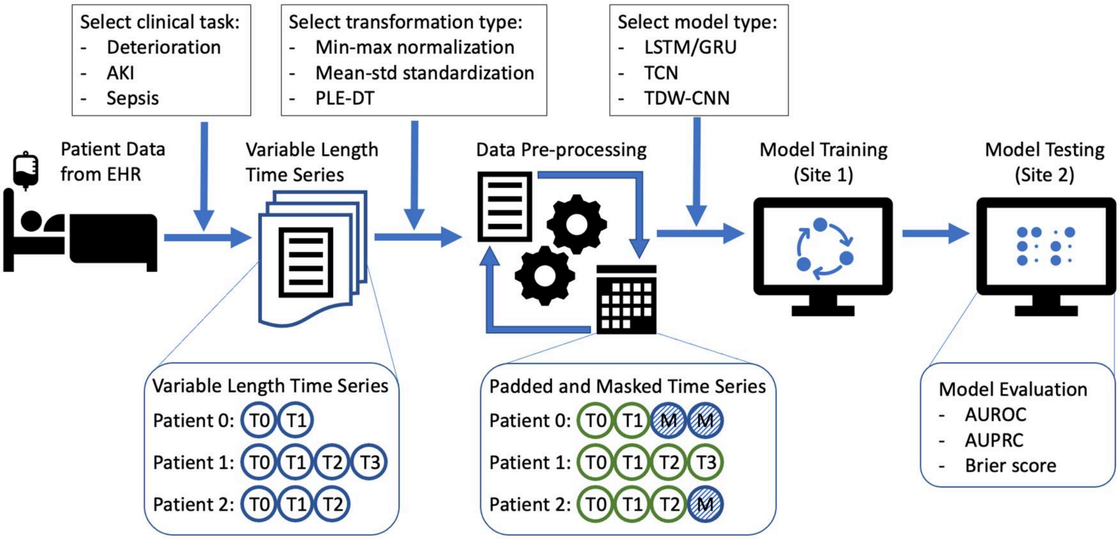

Figure 1. Study design overview. Abbreviations: EHR = electronic health record; AKI = acute kidney injury; PLE-DT = piece-wise linear encoding with a decision tree; LSTM = long short-term memory; GRU = gated recurrent unit; TCN = temporal convolutional network; TDW-CNN = time-distributed wrapper with convolutional neural network; T0-3 = time step 0-3; M = masked value.

in Table S2. For missing values, a last value carry-forward approach was performed, followed by imputation using location-specific medians calculated from the training cohort if the variable was still missing. The former method addresses gaps between consecutive measurements, while the latter ensures a value is provided for time points occurring before the initial recording of a variable. Lastly, training datasets were truncated at the 95th percentile of the length of stay (LOS) for computational efficiency, as some outlier patients remained hospitalized for several months.

Model development

For each prediction task, time series deep learning models were trained and optimized in the datasets created from the UCM cohort (see Figure 1). Three distinct deep learning architectures were compared for all tasks: (1) multi-layer recurrent neural network (RNN); (2) multi-layer temporal convolutional network (TCN)[23,24]; and (3) time-distributed wrapper with convolutional neural network (TDW-CNN).[25–27] Each model predicted the probability of an outcome event at every time step. In RNN models, we specifically harnessed long short-term memory (LSTM)[28,29] and gated recurrent unit (GRU),[30] as they make predictions at each time step based solely on features from the current time step and the cell state derived from previous time steps. We refer to this architecture as LSTM/GRU, with the slash denoting the cell type as a hyperparameter. For TCN models, the causal padding feature within a TCN layer ensured predictions depend on past and current inputs but not future time steps. In TDW-CNN models, a multi-layer CNN, with each CNN layer wrapped in a time-distributed layer, applied convolutions to sequential data. Additionally, an average pooling layer and a multi-layer LSTM/GRU network were integrated into the TDW-CNN model. To ensure the stability of the optimization process during model fitting, we employed layer normalization[31] for recurrent layers and batch normalization[32] layers for non-recurrent layers.

The BO algorithm was used to find the optimal hyperparameters that maximize the area under the receiver operating characteristic curve (AUROC) by training models in a randomly selected 80% of the training encounters and internally validating them in the remaining 20% of the training encounters. This also ensured that model performance comparisons were fair across each method, as different settings may require different optimal hyperparameter. Training with BO used 20 epochs and stopped early if the validation AUROC failed to improve by at least 0.05 in 5 consecutive epochs. Variations of mini-batch gradient descent optimization algorithms with a batch size of 64 were utilized. A complete list of hyperparameters and their ranges are presented in Table S3. All encounters within a batch were padded with a mask value to accommodate the variable-length data (Figure 1). Masking was employed to ensure models disregarded the padded time steps.

Model evaluation

All combinations of data preprocessing transformations and deep learning architectures were evaluated for each clinical task in the external LUMC independent test cohort (Figure 1). During model evaluation, predicted probabilities were calculated at each time step for each model for each task. Model discrimination was assessed using the area under the precision-recall curve (AUPRC) as the primary metric, and the AUROC and its 95% confidence intervals (CI) as the secondary metric, calculated via the DeLong method.[33,34] Furthermore, the sensitivity of models at the 25% precision threshold were compared. Model performance comparisons with *P*-values were omitted because small changes in performance could be statistically significant even when numerically similar due to the size of the datasets. Additionally, model calibration was assessed using the Brier score.

Due to the imbalanced outcome ratio, a sensitivity analysis was performed to determine the impact of incorporating class weights during model training. For each task, we trained a new LSTM/GRU model with class weights that rebalanced the outcome ratio to 1:1 using the feature transformation that yielded the highest AUROC in the external test dataset. In instances of matching results, the transformation with the lowest computational burden was chosen. Class weights were exclusively introduced during model training, and discrimination in the external test datasets was calculated without class weights.

Data cleaning and descriptive analysis were conducted using Stata version 16.0 (StataCorp, College Station, Texas). Python version 3.9.16 was employed to develop deep learning models, along with several libraries, including Keras version 2.12.0, Keras Tuner version 1.3.4, and TCN version 3.5.0. Additionally, calculations of AUROC and its 95% CI were carried out using R version 3.6.0 and the pROC package version 1.16.2. The code for our developed pipeline is available at https://git.doit.wisc.edu/smph-public/dom/uw-icu-data-science-lab-public/variable-length-sequence-modeling/.

Results

Cohort characteristics

A total of 373 825 admissions at UCM and 256 128 admissions at LUMC occurred during the study period. The characteristics of the final cohorts across the 3 tasks are presented in Table 1. Patients from the LUMC cohort were older, more likely to be of Black race, had a shorter LOS, and had lower outcome rates. The clinical deterioration dataset comprised 363 037 encounters for training (UCM) and 242 805 encounters for testing (LUMC). Approximately 5% and 4% of encounters at UCM and LUMC had an outcome event, respectively. Similarly, AKI cohorts encompassed a total of 286 947 training and 199 559 test encounters, with 3% labeled with severe AKI at both sites. Lastly, the infection dataset consisted of 371 847 encounters for training and 253 975 encounters for testing. In the training dataset, approximately 30% of encounters experienced the outcome, while in the test dataset, this percentage was 27%. In the training data, after truncation at the 95th percentile, the time steps ranged in maximum count across the 3 clinical tasks between 2264 and 2727. The test data, which was not truncated, had a maximum count between 2328 and 3819.

Model discrimination

The number of time steps and the prevalence of outcome events at the timestamp level for each external validation dataset are presented in Table S4. The AUPRC and AUROC values for all models across each clinical task are presented in Table 2, and the 95% CI of the AUROC is presented in Table S5. Additionally, the precision-recall and ROC curves are shown in Figures 2 and 3.

The TDW-CNN had the best average AUPRC (0.16) across all architectures for the deterioration task, followed by LSTM/GRU (0.15). The LSTM/GRU had the highest average AUPRC (0.25) in the AKI task, while in the infection task LSTM/GRU and TDW-CNN tied (0.30). Similarly, comparing data transformation methods within each clinical task, PLE-DT had the best average AUPRC (0.16) in the deterioration task, min-max outperformed other methods with an AUPRC of 0.24 in the AKI task, while mean-std and PLE-DT tied in the infection task with average AUPRC of 0.31. Altogether, the LSTM/GRU architecture with PLE-DT transformation had the highest AUPRC in all 3 tasks.

With regards to the AUROC, all model architectures tied in predicting deterioration (average AUROC 0.81). However, LSTM/GRU had the highest average AUROC in both AKI (0.92) and infection (0.87) tasks. Overall, when averaged across all tasks, the LSTM/GRU architecture had the highest mean AUROC (0.87), while both TCN and TDW-CNN had a mean AUROC of 0.86.

The sensitivity values at the 25% precision cut-off are presented in Table S6 and do not conclude the superiority of a model architecture or data transformation. The TCN with PLE-DT did best in deterioration, LSTM/GRU with min-max transformation did best in severe AKI, and LSTM/GRU with PLE-DT along with TDW-CNN with mean-std tied in the infection task. However, when averaged across all transformations for the same model architecture within each clinical task, the TDW-CNN had the highest mean sensitivity (22%) for the clinical deterioration task, while the LSTM/GRU had the highest mean sensitivity in both AKI (46%) and infection (50%) tasks.

In the sensitivity analysis using class weights, the LSTM/GRU model, along with the min-max normalization technique, was selected for clinical deterioration and AKI tasks, while the LSTM/GRU model with the standardization method was utilized for the infection task. These models presented the highest AUROC at each task and trained faster than models with PLE-DT transformation. Incorporating class weights led to lower performance, with AUROCs of 0.82, 0.92, and 0.87 for the clinical deterioration, AKI, and infection tasks. The AUPRCs were also reduced across all clinical tasks compared to models trained without class weights.

Finally, Figures S1-S4 show the distribution of the hyperparameters with notable patterns optimized via BO.

Model calibration

Table 3 provides the results for model calibration, highlighting the best score (ie, the lowest Brier score) within each clinical task. Among the highlighted models, no singular architecture or data transformation method consistently showed superior calibration. On average across transformation methods, TDW-CNN had the best average Brier score (0.0138) in the deterioration task, while LSTM/GRU achieved the best average score (0.0089) in the AKI task. These 2 architectures had similar calibration for predicting infection. Among transformation methods, min-max had the best average score in the AKI task, while PLE-DT had the best average score in both deterioration and infection tasks.

Discussion

In this multicenter study that included over half a million admissions, we compared various deep learning architectures and feature transformation methods for variable-length time series data across 3 commonly studied clinical tasks. We found that the LSTM/GRU model had better discrimination than TCN and TDW-CNN models in mean AUROC and sensitivity metrics. Our investigation did not discern a superiority among the different transformation methods regarding the AUROC metric; however, models that utilized the PLE-DT data transformation achieved higher AUPRCs than normalization and standardization methods. Additionally, models predicting severe AKI demonstrated the highest AUROC and models trained for predicting deterioration had the lowest AUROC. The high performance of models for predicting AKI is consistent with prior literature related to models targeting this task.[3,18] Because AKI is defined by creatinine changes, and models that predict AKI typically include creatinine and other variables highly correlated with renal function, this task is likely easier to predict than other

Table 1. Population characteristics of datasets derived from UCM and LUMC cohorts.

Variable	Clinical Det. UCM ($n = 363\,037$)	Clinical Det. LUMC ($n = 242\,805$)	Severe AKI UCM ($n = 286\,947$)	Severe AKI LUMC ($n = 199\,559$)	Suspected infection UCM ($n = 371\,847$)	Suspected infectionLUMC ($n = 253\,975$)
Age, year, median (IQR)	57 (39, 69)	59 (45, 71)	58 (43, 69)	60 (47, 72)	57 (40, 69)	59 (45, 71)
Female, *n* (%)	207 992 (57%)	128 024 (53%)	156 321 (54%)	100 340 (50%)	211 269 (57%)	131 576 (52%)
Black race, *n* (%)	208 880 (58%)	56 627 (23%)	160 522 (56%)	45 269 (23%)	213 163 (57%)	58 700 (23%)
Length of stay, hours, median (IQR)	75 (43, 144)	62 (28, 123)	86 (47, 151)	68 (31, 138)	75 (42, 144)	62 (28, 124)
Encounter with outcome, *n* (%)	17 424 (4.8%)	10 371 (4.3%)	8984 (3.1%)	5628 (2.8%)	109 840 (29.5%)	67 303 (26.5%)
First outcome, hours, median (IQR)	79 (31, 178)	69 (25, 159)	130 (54, 261)	105 (48, 237)	11 (5, 34)	9 (4, 34)

Abbreviations: IQR = interquartile range; AKI = acute kidney injury; UCM = University of Chicago Medicine; LUMC = Loyola University Medical Center.

Table 2. Model performance AUPRC/AUROC in the external validation datasets across different tasks, data transformation methods, and model architectures.

		Model architecture		
Dataset	Data transformation	LSTM/GRU	TCN	TDW-CNN
Clinical deterioration	Min-max	0.13/**0.82**	0.11/0.81	**0.16/**0.81
	Mean-std	0.15/0.81	0.13/0.80	**0.16/**0.81
	PLE-DT (10 bins)	**0.16/**0.81	**0.16/0.82**	**0.16/**0.81
Severe AKI	Min-max	0.25/**0.92**	0.22/0.91	0.24/**0.92**
	Mean-std	0.25/**0.92**	0.18/0.89	0.21/0.91
	PLE-DT (10 bins)	**0.26/0.92**	0.22/**0.92**	0.17/0.88
Suspected infection	Min-max	0.28/0.86	0.27/0.85	0.29/0.86
	Mean-std	**0.31/0.87**	0.30/0.86	**0.31/**0.86
	PLE-DT (10 bins)	**0.31/0.87**	**0.31/**0.86	**0.31/0.87**

The best performance for each metric in each clinical task is in bold font.
Abbreviations: AUROC = area under the receiver operating characteristic curve; AUPRC = area under the precision-recall curve; LSTM = long short-term memory; GRU = gated recurrent unit; TCN = temporal convolutional network; TDW-CNN = time-distributed wrapper with convolutional neural network; PLE-DT = piece-wise linear encoding with decision tree; AKI = acute kidney injury.

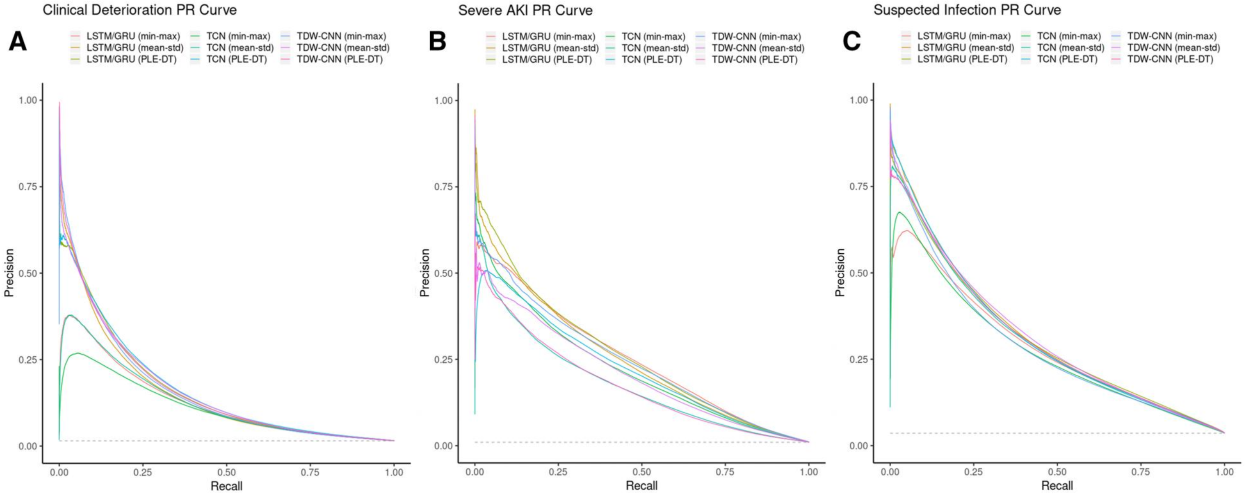

Figure 2. The precision-recall curve of developed models for identifying: (A) clinical deterioration, (B) severe AKI, (C) suspected infections. Abbreviations: PR = precision-recall; AKI = acute kidney injury; PLE-DT = piece-wise linear encoding with a decision tree; LSTM = long short-term memory; GRU = gated recurrent unit; TCN = temporal convolutional network; TDW-CNN = time-distributed wrapper with convolutional neural network.

tasks. In contrast, deterioration can be due to myriad conditions, such as respiratory failure and hemorrhage, each of which have their own specific constellation of physiologic abnormalities. This makes deterioration a more difficult task to predict than AKI. Furthermore, we found that including class weights to balance class distribution during model training decreased discrimination. This work provides important insights related to comparing and externally validating various architectures and data transformation techniques for modeling variable-length time series healthcare data.

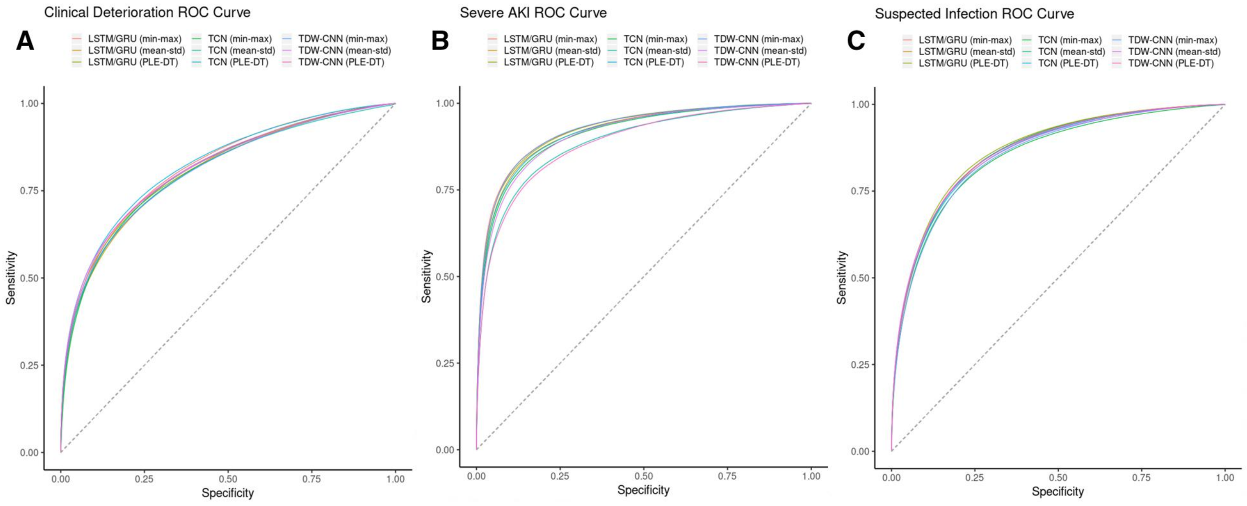

Figure 3. The ROC curve of developed models for identifying: (A) Clinical Deterioration, (B) Severe AKI, (C) Suspected Infections. Abbreviations: ROC = receiver operating characteristics; AKI = acute kidney injury; PLE-DT = piece-wise linear encoding with a decision tree; LSTM = long short-term memory; GRU = gated recurrent unit; TCN = temporal convolutional network; TDW-CNN = time-distributed wrapper with convolutional neural network.

Table 3. Brier score of deep learning models for each clinical task.

		Model architecture		
Dataset	Data transformation	LSTM/GRU	TCN	TDW-CNN
Clinical deterioration	Min-max	0.0145	0.0149	0.0138
	Mean-std	0.0139	0.0142	0.0138
	PLE-DT (10 bins)	0.0138	**0.0137**	0.0139
Severe AKI	Min-max	**0.0086**	0.0090	0.0093
	Mean-std	0.0088	0.0094	0.0091
	PLE-DT (10 bins)	0.0092	0.0089	0.0096
Suspected infection	Min-max	0.0298	0.0299	0.0299
	Mean-std	0.0292	0.0296	**0.0291**
	PLE-DT (10 bins)	0.0292	0.0292	0.0292

The best score in each clinical task is in bold font.
Abbreviations: LSTM = long short-term memory; GRU = gated recurrent unit; TCN = temporal convolutional network; TDW-CNN = time-distributed wrapper with convolutional neural network; PLE-DT = piece-wise linear encoding with decision tree; AKI = acute kidney injury.

Our investigation did not identify a single architecture as a clear winner for all tasks. Across various task-metric combinations, multiple models exhibited competitive performance, especially when considering AUROC. This may be due to the fact that we utilized extensive hyperparameter tuning using BO to optimize AUROC, which allowed all the architectures to flexibly learn from the data and avoid overfitting. However, the LSTM/GRU architecture did obtain the highest AUPRC for all tasks when paired with PLE-DT and achieved the highest AUROC and sensitivity values most often when considering all variable engineering methods across tasks. LSTM/GRU also had the highest average AUROC and average sensitivity, although as noted above the differences in AUROC values were small across models. Prior work comparing different deep learning architectures are mixed, likely due to differences in methodology and tasks. For example, Ayad et al compared various deep learning models, including LSTM, TCN, and CNN-based models, for predicting abnormalities in laboratory values.[35] Although a CNN-based model demonstrated superior performance, the study utilized fixed-length sequences and fewer hyperparameters with a smaller search range. In another investigation, Gopali et al compared LSTM and TCN models in detecting anomalies in time series data, revealing similar performance, with TCN having a slight edge.[36] Discrepancies with our results might be attributed to differences in model hyperparameter tuning. The authors prioritized model complexity and the number of layers in model creation rather than fine-tuning hyperparameters for the optimal performance of each model. In contrast, a comparative study by Almqvist focusing on time series forecasting and employing sequential grid search for hyperparameter selection indicated comparable performance between TCN and LSTM models.[37] Finally, in a study by Tomašev from DeepMind and colleagues, which compared more than 10 deep learning architectures to predict AKI and performed an extensive hyperparameter search, LSTM tied for the highest AUROC.[38] These findings establish the LSTM/GRU model as a strong choice for time series predictions while demonstrating comparable performance to the TCN and the TDW-CNN models.

In this study, we found that models trained with data transformed using the PLE-DT method exhibited better performance, achieving the highest AUROC 5 times and the highest AUPRC 7 times for combinations of tasks and architectures. Comparatively, the min-max normalization and standardization methods were easy to implement, executed swiftly, and incurred no additional memory costs. Conversely, using the PLE-DT method, particularly with 10 decision tree bins, resulted in an expanded feature size. Our findings emphasize the context-dependent nature of data preprocessing algorithms. For example, Lima et al conducted a recent study exploring the impact of 11 different data preprocessing methods on time series data across 38 tasks.[39] Their results indicated no single transformation method performs optimally across all tasks. Notably, the study did not include the PLE technique in its comparison. Furthermore, Gorishniy et al introduced a retrieval-augmented model for classification and regression problems and showed their model performs better when incorporating embeddings of numerical features into the input module.[40] In conclusion, the PLE-DT method had high discrimination, but its memory requirements should be considered when selecting an appropriate data transformation technique.

This investigation has several important strengths. First, it focused on predicting critical outcomes that are commonly tested in the literature. For example, a review by Islam et al identified 42 machine learning papers focused on the early prediction of sepsis, while Vagliano et al identified 46 articles that used machine learning models for predicting AKI.[41,42] Moreover, the comprehensive model comparison across various clinical tasks not only serves as a guide for future research but also offers valuable insights into model selection. In addition, evaluating these models against an external cohort underscores their generalizability and provides a robust representation of their real-world performance. Incorporating BO across a large search space provides a fairer comparison across model architectures. Finally, employing models that provide predictions based only on information available prior to the clinical events highlights the requirement for reliable computer-aided diagnosis in clinical settings.

Our study also has limitations. First, we focused on 3 popular architectures for time series data. Although recent studies have introduced numerous alternative architectures, a comprehensive comparison would be impossible given our dataset size and interest in exploring feature transformation approaches. Second, there are myriad ways to handle missing data during variable pre-processing and modeling, and we chose a commonly used approach. Since the purpose of the study was to compare models to each other for each clinical task, each model would be impacted by the chosen method for managing missing data because they all used the same datasets. Third, our examination encompassed only 3 specific clinical tasks, which restricts the generalizability of our findings to a broader range of in-hospital scenarios. Fourth, this study includes encounters until their first outcome without exploring potential reoccurrences within a single encounter, such as patients who are discharged from the ICU and are at risk again for another deterioration event. Lastly, we focused on structured data, and incorporating unstructured data lies outside the scope of this study.

Conclusion

In conclusion, our study demonstrates that the LSTM/GRU model achieves similar and sometimes superior performance to TCN and TDW-CNN architectures when utilizing variable-length time series structured data for clinical outcome predictions. Furthermore, we found models demonstrate better discrimination in terms of AUPRC when utilizing the PLE-DT transformation method. These results show that accurate prediction of patient outcomes while leveraging the full extent of structured data is feasible and offer valuable insights for real-time computer-aided diagnoses.

Author contributions

Matthew M. Churpek, Majid Afshar, and Anoop Mayampurath conceptualized the study. Matthew M. Churpek, Majid Afshar, Dana P. Edelson, and Emily R. Gilbert managed data collection and secured IRB approvals. Kyle A. Carey and Fereshteh S. Bashiri led the data preprocessing. Fereshteh S. Bashiri and Jennie Martin are responsible for data analysis and methodology. Fereshteh S. Bashiri prepared the original draft, and all authors participated in revising and editing the article.

Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

Funding

This work was supported by the National Institutes of Health (NIH), National Institute of General Medical Sciences grant number R35-13362546 (M.M.C.), NIH National Heart, Lung, and Blood Institute grant numbers R01HL157262 (M. M.C.), and NIH National Institute of Diabetes, Digestive and Kidney Diseases grant number R01-DK126933A (M.M.C.).

Conflicts of interest

Drs Churpek and Edelson have a patent issued (#11,410,777) for risk stratification algorithms for hospitalized patients. Dr Edelson is employed by and has an equity interest in AgileMD, San Francisco, CA. Litmus Health, Austin, TX, has contracted Dr Mayampurath for consulting services outside of the work submitted here. The other authors have declared no potential conflicts of interest.

Data availability

The data used in this study were acquired from 2 hospital systems following approval from the IRBs. The data use agreements prohibit sharing data due to regulatory and legal constraints, and therefore, the data cannot be shared publicly.

References

1. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet*. 2012;13(6):395-405.
2. Green M, Lander H, Snyder A, Hudson P, Churpek M, Edelson D. Comparison of the between the flags calling criteria to the MEWS, NEWS and the electronic cardiac arrest risk triage (eCART) score for the identification of deteriorating ward patients. *Resuscitation*. 2018;123:86-91. https://doi.org/10.1016/j.resuscitation.2017.10.028
3. Churpek MM, Carey KA, Edelson DP, et al. Internal and external validation of a machine learning risk score for acute kidney injury. *JAMA Netw Open*. 2020;3(8):e2012892. https://doi.org/10.1001/jamanetworkopen.2020.12892

4. van Doorn WPTM, Stassen PM, Borggreve HF, et al. A comparison of machine learning models versus clinical evaluation for mortality prediction in patients with sepsis. *PLoS One*. 2021;16(1):e0245157. https://doi.org/10.1371/journal.pone.0245157
5. Padilla RM, Mayo AM. Clinical deterioration: a concept analysis. *J Clin Nurs*. 2018;27(7-8):1360-1368. https://doi.org/10.1111/jocn.14238
6. McFarlan SJ, Hensley S. Implementation and outcomes of a rapid response team. *J Nurs Care Qual*. 2007;22(4):307-313.
7. Rhee C, Jones TM, Hamad Y, et al.; Centers for Disease Control and Prevention (CDC) Prevention Epicenters Program. Prevalence, underlying causes, and preventability of sepsis-associated mortality in US acute care hospitals. *JAMA Netw Open*. 2019;2(2):e187571. https://doi.org/10.1001/jamanetworkopen.2018.7571
8. Ronco C, Bellomo R, Kellum JA. Acute kidney injury. *Lancet*. 2019;394(10212):1949-1964.
9. Arnolds DE, Carey KA, Braginsky L, et al. Comparison of early warning scores for predicting clinical deterioration and infection in obstetric patients. *BMC Pregnancy Childb*. 2022;22(1):295. https://doi.org/10.1186/s12884-022-04631-0
10. Bartkowiak B, Snyder AM, Benjamin A, et al. Validating the electronic cardiac arrest risk triage (eCART) score for risk stratification of surgical inpatients in the postoperative setting: retrospective cohort study. *Ann Surg*. 2019;269(6):1059-1063.
11. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J Biomed Health Inform*. 2017;22(5):1589-1604.
12. Bashiri FS, Caskey JR, Mayampurath A, et al. Identifying infected patients using semi-supervised and transfer learning. *J Am Med Inform Assoc*. 2022;29(10):1696-1704. https://doi.org/10.1093/jamia/ocac109
13. Purushotham S, Meng C, Che Z, Liu Y. Benchmarking deep learning models on large healthcare datasets. *J Biomed Inform*. 2018;83:112-134. https://doi.org/10.1016/j.jbi.2018.04.007
14. Hou N, Li M, He L, et al. Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost. *J Transl Med*. 2020;18(1):462. https://doi.org/10.1186/s12967-020-02620-5
15. Caicedo-Torres W, Gutierrez J. ISeeU: visually interpretable deep learning for mortality prediction inside the ICU. *J Biomed Inform*. 2019;98:103269. https://doi.org/10.1016/j.jbi.2019.103269
16. Tipirneni S, Reddy CK. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Trans Knowl Disc Data (TKDD)*. 2022;16(6):1-17.
17. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3(1):160035. https://doi.org/10.1038/sdata.2016.35
18. Koyner JL, Carey KA, Edelson DP, Churpek MM. The development of a machine learning inpatient acute kidney injury prediction model. *Crit Care Med*. 2018;46(7):1070-1077. Accessed January 19, 2023. https://journals.lww.com/ccmjournal/Fulltext/2018/07000/The_Development_of_a_Machine_Learning_Inpatient.5.aspx
19. Kellum JA, Lameire N, Aspelin P, et al. Kidney disease: improving global outcomes (KDIGO) acute kidney injury work group. KDIGO clinical practice guideline for acute kidney injury. *Kidney Int Suppl (2011)*. 2012;2(1):1-138.
20. Seymour CW, Liu VX, Iwashyna TJ, et al. Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA*. 2016;315(8):762-774. https://doi.org/10.1001/jama.2016.0288
21. Churpek MM, Dumanian J, Dussault N, et al. Determining the electronic signature of infection in electronic health record data. *Crit Care Med*. 2021;49(7):e673-e682. Accessed May 18, 2021. https://journals.lww.com/ccmjournal/Fulltext/2021/07000/Determining_the_Electronic_Signature_of_Infection.32.aspx
22. Gorishniy Y, Rubachev I, Babenko A. On embeddings for numerical features in tabular deep learning. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, eds. *Advances in Neural Information Processing Systems*. Vol 35. Curran Associates, Inc.; 2022:24991-25004. Accessed January 12, 2023. https://proceedings.neurips.cc/paper_files/paper/2022/file/9e9f0ffc3d836836ca96cbf8-fe14b105-Paper-Conference.pdf
23. Bai S, Kolter JZ, Koltun V. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. ArXiv, abs/1803.01271, preprint: not peer reviewed. https://doi.org/10.48550/arXiv.1808.0127
24. Remy P. Temporal convolutional networks for Keras. GitHub repository. Published online 2020. Accessed March 1, 2021. https://github.com/philipperemy/keras-tcn
25. Dallanoce F. Neural network for input of variable length using tensorflow timedistributed wrapper. Towards data science. Published August 24, 2021. Accessed December 6, 2023. https://towardsdatascience.com/neural-network-for-input-of-variable-length-using-tensorflow-timedistributed-wrapper-a45972f4da51
26. Montaha S, Azam S, Rafid A, Hasan MZ, Karim A, Islam A. Time-distributed-CNN-LSTM: a hybrid approach combining CNN and LSTM to classify brain tumor on 3d MRI scans performing ablation study. *IEEE Access*. 2022;10:60039-60059.
27. Siddique LA, Junhai R, Reza T, Khan SS, Rahman T. Analysis of real-time hostile activitiy detection from spatiotemporal features using time distributed deep CNNs, RNNs and attention-based mechanisms. In: *2022 IEEE 5th International Conference on Image Processing Applications and Systems (IPAS)*. IEEE; 2022:1-6.
28. Hochreiter S, Schmidhuber J. LSTM can solve hard long time lag problems. In: *Proceedings of the 9th International Conference on Neural Information Processing Systems*. NIPS'96. MIT Press; 1996:473-479.
29. Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw*. 2005;18(5-6):602-610. https://doi.org/10.1016/j.neunet.2005.06.042
30. Cho K, van Merrienboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. Published online 2014. Accessed March 1, 2021. http://arxiv.org/abs/1406.1078
31. Ba JL, Kiros JR, Hinton GE. 2016. Layer normalization. arXiv, arXiv:160706450, preprint: Published online.
32. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: *International Conference on Machine Learning*. PMLR; 2015:448-456.
33. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837-845. https://doi.org/10.2307/2531595
34. Sun X, Xu W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Process Lett*. 2014;21(11):1389-1393. https://doi.org/10.1109/LSP.2014.2337313
35. Ayad A, Hallawa A, Peine A, et al. Predicting abnormalities in laboratory values of patients in the intensive care unit using different deep learning models: comparative study. *JMIR Med Inform*. 2022;10(8):e37658. https://doi.org/10.2196/37658
36. Gopali S, Abri F, Siami-Namini S, Namin AS. A comparison of TCN and LSTM models in detecting anomalies in time series data. In: *2021 IEEE International Conference on Big Data (Big Data)*. 2021:2415-2420. https://doi.org/10.1109/BigData52589.2021.9671488

37. Almqvist O. A comparative study between algorithms for time series forecasting on customer prediction: an investigation into the performance of ARIMA, RNN, LSTM, TCN and HMM. Published online 2019. Accessed November 14, 2023. https://www.diva-portal.org/smash/get/diva2:1321224/FULLTEXT01.pdf
38. Tomašev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*. 2019;572(7767):116-119. https://doi.org/10.1038/s41586-019-1390-1
39. Lima FT, Souza VMA. A large comparison of normalization methods on time series. *Big Data Res*. 2023;34:100407. https://doi.org/10.1016/j.bdr.2023.100407
40. Gorishniy Y, Rubachev I, Kartashev N, Shlenskii D, Kotelnikov A, Babenko A. 2023. TABR: unlocking the power of retrieval-augmented tabular deep learning. arXiv, arXiv:230714338, preprint: Published online.
41. Islam KR, Prithula J, Kumar J, et al. Machine learning-based early prediction of sepsis using electronic health records: a systematic review. *J Clin Med*. 2023;12(17):5658.
42. Vagliano I, Chesnaye NC, Leopold JH, Jager KJ, Abu-Hanna A, Schut MC. Machine learning models for predicting acute kidney injury: a systematic review and critical appraisal. *Clin Kidney J*. 2022;15(12):2266-2280.

Journal of the American Medical Informatics Association, 2024, 31, 1322–1330
https://doi.org/10.1093/jamia/ocae088
Research and Applications