

Missingness-Aware Mamba for Early Sepsis Prediction from Irregularly Sampled ICU Time Series

Yuze Li¹, Tohgoroh Matsui²

^{1,2}Computer Engineering / Information Technology, Graduate School of Engineering, Chubu University, Japan

Abstract — Sepsis is a leading cause of death in intensive care units (ICUs), and every hour of treatment delay increases mortality, so hourly early-warning models built from routine bedside data have clear practical value. Bedside records are irregularly sampled, with most laboratory values observed in fewer than 10% of hours. Recurrent neural networks remain the dominant sequence model for this task, and the recently proposed Mamba selective state-space model has not been evaluated on hourly sepsis prediction. This paper presents a missingness-aware Mamba model for the PhysioNet/Computing in Cardiology Challenge 2019 dataset (40,336 ICU stays). Each hour is represented by forward-filled, standardised measurements concatenated with a binary observation mask and processed causally by residual Mamba blocks. Under identical preprocessing and an equal hyper-parameter search per model, Mamba is compared with GRU and LSTM baselines over five random seeds using the area under the ROC curve (AUROC), the area under the precision-recall curve (AUPRC) and the official utility score. Mamba reached an AUROC of 0.827 ± 0.005 , an AUPRC of 0.112 ± 0.010 and a utility score of 0.386 ± 0.018 , matching the tuned recurrent baselines within seed-to-seed noise and with the lowest variance in AUROC. Removing the observation mask reduced AUROC by 0.030 and utility by 0.054, several times the seed-to-seed standard deviation. Missingness encoding therefore matters more than the choice of backbone on this task.

Keywords — selective state-space model; informative missingness; recurrent neural networks; PhysioNet Challenge 2019; clinical early warning

I. INTRODUCTION

Sepsis, a dysregulated host response to infection that leads to organ dysfunction [1], accounted for an estimated 48.9 million cases and 11 million deaths worldwide in 2017 [2]. Outcome depends strongly on how quickly treatment begins: in a mandated emergency-care programme covering 49,331 patients, each additional hour before completion of the initial care bundle was associated with a 4% relative increase in in-hospital mortality [3]. Automated early-warning systems that score a patient every hour from data already present in the electronic health record therefore have a clear clinical rationale, and the PhysioNet/Computing in Cardiology Challenge 2019 was organised to benchmark exactly this task: predicting the onset of sepsis at least six hours in advance from hourly ICU records [4].

The data generated at the bedside pose two difficulties for sequence models. First, sampling is irregular and sparse. In the Challenge 2019 cohort, heart rate is recorded in about 90% of hours, but most laboratory values are observed in fewer than 10% of hours and some in fewer than 1%. Second, the missingness is informative rather than random: a lactate or a blood culture is ordered when a clinician already suspects deterioration, so the fact that a measurement exists can be as predictive as its value [5], [6]. A model must therefore represent both the measured values and the measurement pattern.

Recurrent neural networks, in particular the gated recurrent unit (GRU) [7] and long short-term memory (LSTM) [8], have been the standard deep sequence models for ICU prediction, including neural entries to Challenge 2019 [4], [9]. Transformers [10] offer parallel training but their self-attention scales quadratically with sequence length. Structured state-space models [11] and especially the selective state-space model Mamba [12] combine the parallel training of Transformers with linear-time, recurrent inference and have recently been adopted for medical images [13], [14] and for large-scale electronic health record pre-training [15]. To our knowledge, however, Mamba has not been evaluated on hourly bedside sepsis prediction, and it is not known whether its advantages carry over to short, sparse, heavily missing ICU sequences.

This paper addresses that gap with a deliberately simple and fully reproducible study. The contributions are as follows.

1. A missingness-aware Mamba model for hourly sepsis prediction, in which each hour is encoded by forward-filled, standardised measurements concatenated with a binary observation mask and processed causally by residual Mamba blocks.
2. A controlled comparison with GRU and LSTM backbones on the complete public Challenge 2019 training cohort (40,336 stays), using identical preprocessing, an equal per-model hyper-parameter search, five random seeds, and the official challenge utility score in addition to AUROC and AUPRC.
3. An ablation isolating the effect of the observation mask, which shows that missingness encoding, not the sequence backbone, is the dominant design factor on this task.
4. An honest efficiency analysis showing that, at the model sizes and sequence lengths relevant to bedside monitoring, Mamba offers no inference-latency advantage over cuDNN-accelerated recurrent networks.

The code is a single Python script and all experiments run in about one minute each on a laptop GPU, so the study can be reproduced in under an hour.

II. RELATED WORK

A. Machine learning for early sepsis prediction

Early sepsis prediction from EHR data has been studied extensively; a systematic review found wide variation in sepsis definitions, prediction horizons and evaluation protocols, which limits comparability [9]. Challenge 2019 standardised the task by releasing hourly records of 40,336 ICU stays from two hospitals with Sepsis-3-derived

labels and by introducing a clinically motivated utility score that rewards predictions made 12 to 3 hours before onset and penalises false alarms [4]. The strongest entries relied on gradient-boosted trees over hand-engineered features, such as ensembles of XGBoost models [16], or on path-signature features of the time series [17]; neural entries typically used GRU or LSTM networks. These entries combined the model with substantial feature engineering, whereas the present study isolates the sequence backbone and the missingness representation under a fixed, minimal feature set.

B. Missing data in clinical time series

Clinical time series are multivariate, irregularly sampled and sparsely observed. Lipton et al. [5] showed that simply appending binary missingness indicators to the input of an RNN improves paediatric ICU diagnosis, demonstrating that missingness is informative. GRU-D [6] extended this idea with learned exponential decay of both the hidden state and the last observed value as a function of the time since observation. Related work has found that measurement frequency and indicator features are among the most important predictors of sepsis in the Challenge 2019 data [16], [17], that missingness indicators raise the AUROC of ICU outcome models by about 0.05 on average [18], and that the measurement pattern alone predicts ICU mortality with an AUROC of 0.76 [19]. The present work adopts the simplest of these representations, forward filling plus a binary observation mask, so that its contribution can be measured in isolation with a single ablation.

C. State-space models in healthcare

Structured state-space sequence models (S4) [11] model a sequence through a discretised linear dynamical system that can be computed either as a convolution (parallel training) or as a recurrence (constant-memory inference). Mamba [12] makes the state-transition parameters input-dependent, giving the model a content-based selection mechanism while keeping linear time complexity. In healthcare, Mamba has been applied to medical image classification [13] and segmentation, surveyed in [14], and to large-scale EHR foundation models trained on tokenised clinical events [15]. Closer to the present task, a Mamba-based model has been used for real-time prediction of ICU patient acuity and therapy requirements across three cohorts [20], and a systematic evaluation of long-context architectures on EHR event sequences found that a Mamba-based model outperformed the prior state of the art on 9 of 14 clinical prediction tasks [21]. These studies focus on images or on event-level EHR sequences with large training corpora. The present study examines the opposite regime: short (median 38 h), dense-in-time but sparse-in-variable bedside sequences from a single public cohort, trained from scratch.

III. MATERIALS AND METHODS

A. Dataset

We use the public training data of the PhysioNet/Computing in Cardiology Challenge 2019 [4], which comprises 40,336 ICU stays from two hospital systems (set A: 20,336 stays; set B: 20,000 stays). Each stay is a table with one row per hour and 40 variables: 8 vital signs, 26 laboratory values and 6 demographic or administrative variables (age, sex, two ICU-unit indicators, time from hospital admission to ICU admission, and ICU length of stay). The label SepsisLabel is 1 from six hours before the Sepsis-3 onset time (t_{sepsis}) until the end of the stay and 0 otherwise, so that a correct positive prediction at the first labelled hour is six hours early. Table 1 summarises the cohort. Sepsis develops in 2,932 stays (7.3%); because the label is positive only in the final hours of septic stays, positive hours make up 1.8% of all hours, which makes the hourly task strongly imbalanced. Fig. 1 shows the fraction of hours in which each variable is observed: seven of the eight vital signs are observed in 34–90% of hours (end-tidal CO2 in 3.7%), whereas 24 of the 26 laboratory values are observed in fewer than 10% of hours.

Table 1. Summary of the PhysioNet/CinC Challenge 2019 training cohort used in this study

Characteristic	Value
ICU stays (hospital A / hospital B)	40,336 (20,336 / 20,000)
Hourly rows	1,552,210
Length of stay, median (max), h	38 (336)
Stays with sepsis	2,932 (7.3%)
Hours with SepsisLabel = 1	27,916 (1.8%)
Variables per hour (vital / laboratory / static)	40 (8 / 26 / 6)
Train / validation / test stays (per seed)	28,238 / 6,049 / 6,049

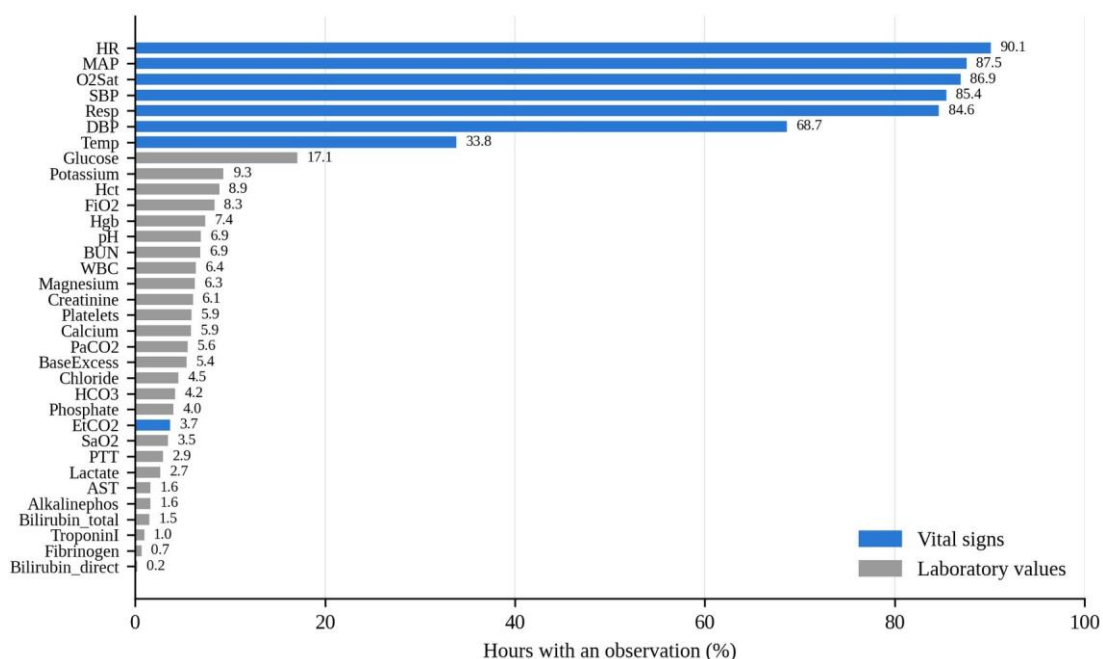


Figure 1. Fraction of hours in which each time-varying variable is observed in the cohort (vital signs in blue, laboratory values in grey).

B. Preprocessing

The released data contain at most 336 h (14 days) per stay, so all 1,552,210 hourly rows are used without truncation. Within each stay, missing values are forward filled from the last observation of the same variable; hours before the first observation remain missing. Every variable is then standardised by its mean and standard deviation computed over all observed values in the cohort, hours that are still missing are set to 0 (the cohort mean), and standardised values are clipped to $[-10, 10]$ to limit the influence of a small number of implausible entries (0.02% of observed values, for example an FiO2 recorded on a percentage rather than a fractional scale). No feature engineering, resampling or external data is used.

C. Missingness-aware input

For each hour t , let $x_t \in \mathbb{R}^{40}$ be the preprocessed measurement vector and $m_t \in \{0,1\}^{40}$ the observation mask, with $m_{t,j} = 1$ if variable j was actually recorded at hour t . The model input is the concatenation $u_t = [x_t; m_t] \in \mathbb{R}^{80}$. The mask lets the network distinguish a fresh observation from a forward-filled or imputed value and exposes the measurement pattern itself, which is known to be informative in ICU data [5], [6]. In the ablation we set $u_t = x_t$, so that the only difference between the two configurations is the presence of the 40 mask channels; all models in the main comparison use the same 80-dimensional input.

D. Sequence backbones

All models share the structure: a linear input projection $u_t \rightarrow h_t \in \mathbb{R}^d$, a causal sequence backbone, and an output head consisting of layer normalisation, dropout and a linear layer producing one logit per hour. Only the backbone differs.

Mamba. The backbone is a stack of two residual blocks, each computing $h \leftarrow h + \text{Dropout}(\text{Mamba}(\text{LayerNorm}(h)))$. The Mamba layer [12] expands the input to 2d channels, applies a causal depth-wise convolution of width 4, and runs a selective state-space scan with a state size of 16 in which the discretisation step and the input and output projections of the state are functions of the current input. The scan is causal by construction, so the output at hour t depends only on hours 1 to t ; this is required for an early-warning system and is the setting in which Mamba's linear-time recurrence can be exploited at deployment. We use the reference CUDA implementation (mamba-ssm 2.3.2).

GRU and LSTM. The baselines replace the Mamba blocks by a two-layer unidirectional GRU [7] or LSTM [8] with hidden size d and inter-layer dropout, using the cuDNN implementation in PyTorch. These are the standard neural backbones for this task [4], [9].

E. Training

Stays are split at the patient level into training, validation and test sets of 70%, 15% and 15% (28,238 / 6,049 / 6,049 stays), stratified on whether sepsis ever occurs; the split is re-drawn for each random seed. The loss is the binary cross-entropy averaged over all valid (non-padded) hours of the batch, with the positive class weighted by $\min(20, N_{\text{neg}}/N_{\text{pos}}) = 20$ to counter the 1.8% positive rate. Models are trained with AdamW [22] (weight decay 0.01, gradient-norm clipping at 1.0, batch size 128 stays) for at most 30 epochs, and training stops when the validation AUPRC has not improved for five

epochs; the checkpoint with the best validation AUPRC is evaluated on the test set. Each run takes about one minute on a single NVIDIA RTX 4060 laptop GPU (8 GB) with PyTorch 2.12.

F. Evaluation

Three metrics are reported on the held-out test stays. AUROC and AUPRC are computed over all valid hours pooled across stays; AUPRC is the more informative of the two at a 1.8% positive rate. The third metric is the normalised utility score defined by the challenge organisers [4], re-implemented from the official scoring code with the same constants: a correct positive prediction earns a utility that rises linearly from 0 at 12 h before t_{sepsis} to 1 at 6 h before and falls to 0 at 3 h after; a missed prediction between 6 h before and 3 h after t_{sepsis} incurs a penalty falling linearly to -2 ; and every false alarm in a non-septic patient costs 0.05. The total is normalised so that perfect prediction scores 1 and never predicting sepsis scores 0. The utility score requires a binary decision; the threshold is chosen on the validation set from 50 quantiles of the predicted probabilities to maximise validation utility and is then applied unchanged to the test set.

G. Hyper-parameter selection

To give every backbone the same tuning budget, each model is trained with seed 0 under all eight combinations of hidden size $d \in \{64, 128\}$, dropout $\in \{0.1, 0.3\}$ and learning rate $\in \{10^{-3}, 3 \times 10^{-4}\}$, and the combination with the highest validation AUPRC is selected (Table 2). The selected configuration of each model is then retrained with five random seeds (0–4), each with its own patient split, and all reported results are the mean and standard deviation over these five runs. The ablation without the observation mask uses the selected Mamba configuration.

Table 2. Hyper-parameter search (seed 0): validation AUPRC for each configuration and the selected setting per model. d = hidden size, p = dropout, lr = learning rate; ties are resolved on the unrounded value.

d	p	lr	Mamba	GRU	LSTM
64	0.1	1e-3	0.098	0.110	0.105
64	0.1	3e-4	0.093	0.099	0.097
64	0.3	1e-3	0.096	0.106	0.108
64	0.3	3e-4	0.092	0.107	0.097
128	0.1	1e-3	0.103	0.113	0.101
128	0.1	3e-4	0.095	0.101	0.097
128	0.3	1e-3	0.107	0.113	0.104
128	0.3	3e-4	0.095	0.102	0.095
Selected			$d=128, p=0.3, lr=1e-3$	$d=128, p=0.3, lr=1e-3$	$d=64, p=0.3, lr=1e-3$

IV. RESULTS

A. Main comparison

Table 3 reports the test-set performance of the three backbones with the missingness-aware input, and Fig. 2 shows the ROC and precision–recall curves of one representative seed. All three models reach a similar operating range: AUROC 0.826–0.831, AUPRC 0.111–0.117 and utility 0.384–0.398. The GRU has the highest mean on every metric and Mamba the second highest by a narrow margin, and the differences between means (no more than 0.005 in AUROC, 0.006 in AUPRC and 0.014 in utility) are no larger than the corresponding standard deviations across seeds.

Because every seed uses its own patient split, the fairest comparison is the paired difference on the same split. Across the five seeds, Mamba minus GRU is -0.003 ± 0.009 in AUROC (Mamba higher in three of five seeds), -0.005 ± 0.009 in AUPRC and -0.013 ± 0.019 in utility; Mamba minus LSTM is $+0.002 \pm 0.006$, $+0.002 \pm 0.009$ and $+0.002 \pm 0.011$. None of these paired differences is distinguishable from zero with five seeds, so the three backbones should be regarded as equivalent on this task. Mamba does show the smallest spread across seeds in AUROC (SD 0.005 versus 0.010 for the GRU and 0.009 for the LSTM; Fig. 3), although its spread in utility is comparable to that of the other models, so this stability observation applies to the ranking metric rather than to the thresholded decision.

Table 3. Test-set performance, mean $\pm$ SD over five seeds (each seed uses its own patient-level split). All models use the selected configuration from Table 2. Best mean per column in bold.

Model	Input	Params	AUROC	AUPRC	Utility
GRU	values + mask	208,897	0.831 $\pm$ 0.010	0.117 $\pm$ 0.010	0.398 $\pm$ 0.026
LSTM	values + mask	71,937	0.826 $\pm$ 0.009	0.111 $\pm$ 0.015	0.384 $\pm$ 0.016
Mamba (ours)	values + mask	244,225	0.827 $\pm$ 0.005	0.112 $\pm$ 0.010	0.386 $\pm$ 0.018
Mamba, no mask (ablation)	values only	239,105	0.797 $\pm$ 0.008	0.093 $\pm$ 0.008	0.331 $\pm$ 0.020

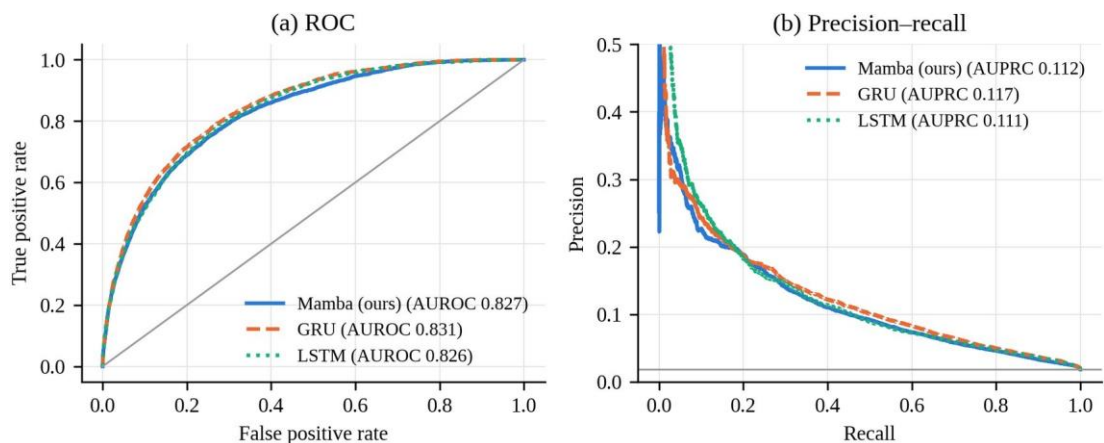


Figure 2. Test-set (a) ROC and (b) precision–recall curves of the three backbones for seed 3; the legend gives the five-seed mean AUROC/AUPRC. The grey line in (b) marks the 1.8% positive rate.

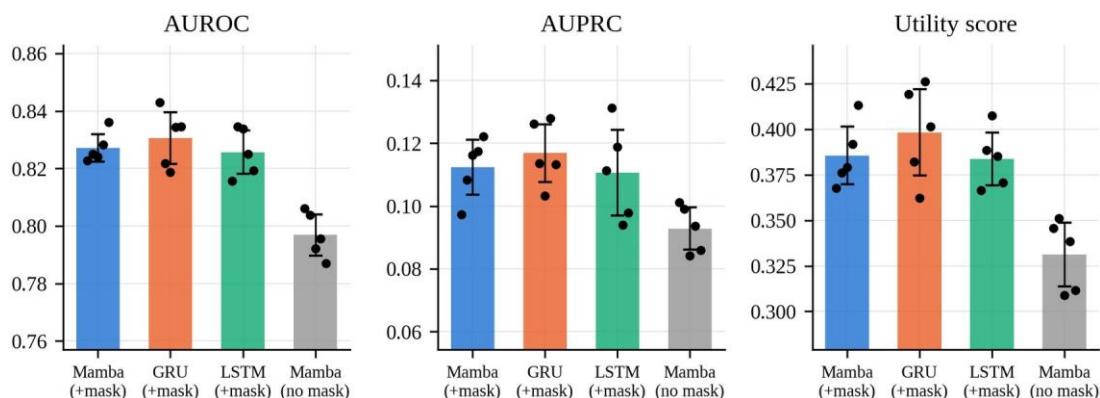


Figure 3. Test-set AUROC, AUPRC and utility score of the three backbones with the observation mask and of Mamba without it. Bars show the five-seed mean, error bars one standard deviation, and dots the individual seeds.

B. Effect of the observation mask

The last row of Table 3 and the grey bars in Fig. 3 show the Mamba model trained on the measurement values alone. Removing the 40 mask channels lowers every metric in every seed: the paired loss is 0.030 ± 0.006 in AUROC (smallest loss over seeds 0.022), 0.020 ± 0.005 in AUPRC (smallest 0.013) and 0.054 ± 0.010 in utility (smallest 0.041). These effects are several times larger than the between-backbone differences of Section IV-A and larger than any seed-to-seed standard deviation, and the model without the mask is worse than every masked model in every seed. The parameter count changes by only 2% between the two configurations, so the gain is attributable to the information in the mask rather than to model capacity. The result confirms, for a selective state-space backbone, the observation made for recurrent networks that the measurement pattern in ICU data is itself a strong predictor [5], [6].

C. Training dynamics

Fig. 4 plots the validation AUPRC per epoch for seed 0. Mamba reaches its best validation score earliest, after 2.8 ± 0.8 epochs on average over the five seeds, against 8.0 ± 1.6 for the GRU and 6.8 ± 2.2 for the LSTM, and its validation score then declines steadily, so that early stopping ends its training after 7.8 epochs on average compared with 13.0 and 11.8 for the recurrent models. The steeper decline indicates that the Mamba model fits the training set faster and overfits sooner at the same dropout rate, which is consistent with the hyper-parameter search (Table 2), where Mamba gained slightly more than the other backbones from raising dropout from 0.1 to 0.3 at $d = 128$. Early stopping on validation AUPRC was therefore necessary for all models and most important for Mamba.

D. Model size and inference latency

Table 4 compares the three backbones at equal hidden size ($d = 128$) in parameter count and in batched inference latency on the same GPU for a batch of 128 stays at sequence lengths from 48 h to 4,000 h. Mamba has 17% more parameters than the GRU of equal hidden size, because of its channel expansion and state projections, and 11% fewer than the LSTM; the LSTM actually selected by the search ($d = 64$) has 71,937 parameters, 3.4 times fewer than Mamba. In latency, the cuDNN GRU is the fastest at every length and Mamba the slowest: at the longest stay in the data (336 h) Mamba takes 12.0 ms per batch against 5.0 ms for the GRU and 8.1 ms for the LSTM, and the ratio widens rather than narrows up to 4,000 h. All three latencies are far below any clinical requirement for hourly scoring, but the result shows that the linear-time complexity of Mamba does not translate into a speed advantage over optimised recurrent kernels at the hidden sizes (≤ 128) and sequence lengths (≤ 336 h) relevant to this task.

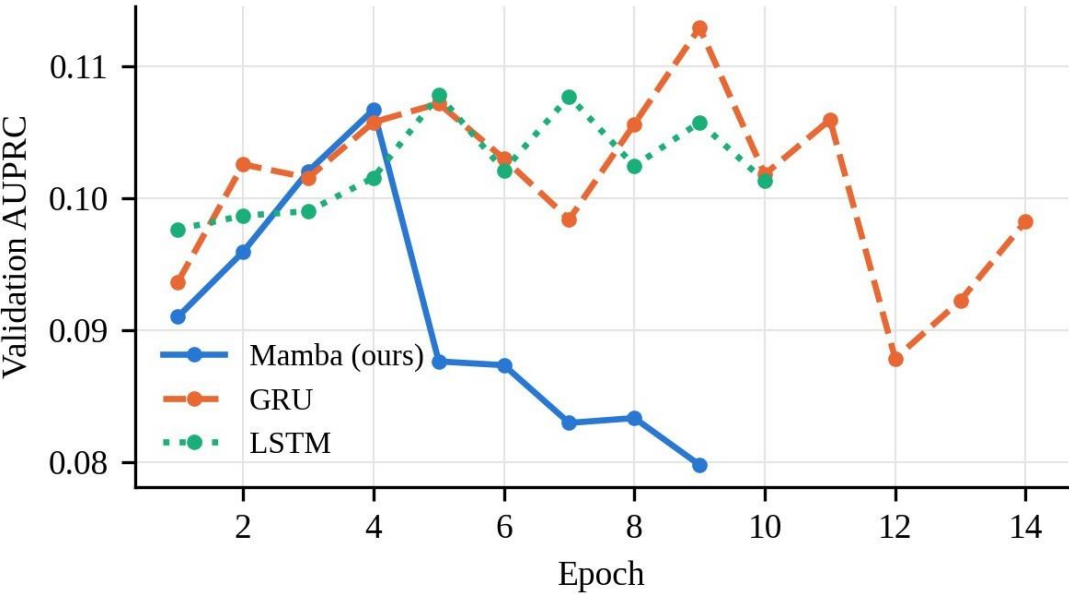


Figure 4. Validation AUPRC per epoch for the three backbones (seed 0, selected configurations). Training stops five epochs after the best validation AUPRC.

V. DISCUSSION

Mamba is a competitive but not superior backbone for hourly sepsis prediction. Under equal preprocessing, an equal tuning budget and five paired seeds, Mamba, GRU and LSTM reach the same operating range on all three metrics, and the paired differences are within noise. This is a different picture from the image and event-sequence settings in which Mamba has been reported to surpass convolutional or attention-based models [13], [15], and two properties of the present task explain it. First, the sequences are short: the median stay is 38 h and no stay exceeds 336 h, so the long-range modelling capacity that distinguishes state-space models from gated recurrences is rarely exercised. Second, the training signal is sparse (1.8% positive hours) and the cohort is modest (28,000 training stays), a regime in which the stronger inductive bias and lower capacity of a gated recurrence can be an advantage; the faster overfitting of Mamba in Fig. 4 points in that direction. The lower seed-to-seed variance of Mamba in AUROC is a practical advantage for model selection, but it should be confirmed on other cohorts before being relied upon.

Table 4. Model size and batched inference latency (ms per batch of 128 stays, mean of 10 runs, NVIDIA RTX 4060 laptop GPU) as a function of sequence length T. All three backbones are instantiated at d = 128 for a like-for-like comparison; the LSTM selected in Table 2 (d = 64) is smaller and faster.

Model	Params	T = 48	T = 336	T = 1,000	T = 2,000	T = 4,000
Mamba (d = 128)	244,225	2.8	12.0	30.5	89.6	223.3
GRU (d = 128)	208,897	0.8	5.0	15.3	31.1	62.8
LSTM (d = 128)	274,945	1.5	8.1	23.7	47.9	94.6

Missingness encoding is the dominant design factor. The single ablation in this study produced an effect (0.030 AUROC, 0.054 utility) several times larger than any difference between backbones. The mask tells the model when a laboratory value was actually drawn, and in ICU practice tests such as lactate, blood gases and coagulation panels are ordered when a clinician suspects deterioration; the measurement pattern therefore encodes clinical judgement that precedes the formal onset of sepsis. The same conclusion has been reached for recurrent models [5], [6] and for the feature-engineered entries to Challenge 2019 [16], [17], and the present study shows that it holds equally for a selective state-space model. A practical implication is that any comparison of sequence backbones on bedside data that does not control the missingness representation is likely to be dominated by that representation rather than by the backbone.

Relation to the challenge leaderboard. The utility scores reported here (0.38–0.40) are obtained on a random held-out split of the public training data and are not comparable with the leaderboard values of the hidden challenge test set, which included a third hospital and on which the best entry scored 0.360 [4]. The purpose of this study is a controlled comparison of backbones and input representations rather than a leaderboard submission, and no feature engineering, ensembling or threshold tuning beyond the validation set was used.

Limitations

This study has several limitations that bound its conclusions. First, all results come from random splits of a single public cohort of two hospitals; generalisation to other institutions, in particular under a distribution shift in measurement practice, which directly affects the mask channels, was not tested. Second, five seeds are sufficient to show that the between-backbone differences are smaller than the seed variance but not to detect differences of the order of 0.005 AUROC with confidence; the conclusion is one of practical equivalence rather than of proven equality. Third, the hyper-parameter search covered eight configurations per model with a fixed depth of two layers; deeper or wider Mamba models, or Mamba-specific settings such as the state size, were not explored, and the ranking could change with a larger budget. Fourth, the missingness representation was deliberately the simplest one (forward filling plus a binary mask); time-since-observation decay as in GRU-D [6] and learned imputation were not combined with the Mamba backbone and remain a natural extension. The latency comparison is specific to one consumer GPU and to batched inference; the recurrent inference mode of Mamba, which processes one hour

at a time with constant memory, was not benchmarked.

VI. CONCLUSION

This paper evaluated a missingness-aware Mamba model for hourly early sepsis prediction on the complete public PhysioNet/CinC Challenge 2019 cohort, under a controlled protocol with identical preprocessing, an equal hyperparameter budget per backbone, five paired random seeds and the official challenge utility score. Mamba achieved an AUROC of 0.827, an AUPRC of 0.112 and a utility score of 0.386, matching tuned GRU and LSTM baselines within seed-to-seed noise and with the lowest variance in AUROC, which establishes the selective state-space model as a valid backbone for bedside early-warning systems built on short, sparse ICU sequences. The study's clearest finding, however, concerns the input rather than the backbone: encoding which variables were measured at each hour improved AUROC by 0.030 and utility by 0.054, an effect several times larger than any difference between sequence models. Future work should combine the Mamba backbone with richer missingness representations such as time-aware decay, validate the approach on external cohorts with different measurement practices, and benchmark Mamba's recurrent inference mode for streaming bedside deployment.

DECLARATIONS

CRedit author contribution statement

Author 1: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Author 2:** Conceptualization, Supervision, Writing – review & editing, Project administration.

Conflicts of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability statement

No new data were created in this study. The PhysioNet/Computing in Cardiology Challenge 2019 training data analysed here are publicly available from PhysioNet (<https://physionet.org/content/challenge-2019/1.0.0/>) under the Open Data Commons Attribution License [4]. The data are fully de-identified and were released by the organisers for public research use, so no additional ethics approval was required. The complete source code (a single Python script covering preprocessing, models, training and evaluation), the per-run result files and the figure-generation script will be released under the MIT license in a public GitHub repository upon publication of this article.

Use of AI tools

Claude (Anthropic; Claude Code with the Claude Opus and Fable model families, 2026) was used to assist in writing and debugging the experiment code, in searching the literature, and in drafting and editing the text of all sections of this manuscript. All experiments were executed, and all results, tables and figures were generated, by the authors' own code on the authors' hardware; no results, figures or references were generated by the AI tool, and every cited work was independently verified by the authors against its primary source. The authors critically reviewed and revised all AI-assisted text and take full responsibility for the content of the publication.

Acknowledgements

The authors thank the organisers of the PhysioNet/Computing in Cardiology Challenge 2019 for making the dataset and the scoring code publicly available.

Funding

This work received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

- [1] M. Singer et al., "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)," *JAMA*, vol. 315, no. 8, pp. 801–810, 2016, doi: 10.1001/jama.2016.0287.
- [2] K. E. Rudd et al., "Global, regional, and national sepsis incidence and mortality, 1990–2017: Analysis for the Global Burden of Disease Study," *Lancet*, vol. 395, no. 10219, pp. 200–211, 2020, doi: 10.1016/S0140-6736(19)32989-7.
- [3] C. W. Seymour et al., "Time to treatment and mortality during mandated emergency care for sepsis," *N. Engl. J. Med.*, vol. 376, no. 23, pp. 2235–2244, 2017, doi: 10.1056/NEJMoa1703058.
- [4] M. A. Reyna, C. S. Josef, R. Jeter, S. P. Shashikumar, M. B. Westover, S. Nemati, G. D. Clifford, and A. Sharma, "Early prediction of sepsis from clinical data: The PhysioNet/Computing in Cardiology Challenge 2019," *Crit. Care Med.*, vol. 48, no. 2, pp. 210–217, 2020, doi: 10.1097/CCM.0000000000004145.
- [5] Z. C. Lipton, D. C. Kale, and R. Wetzl, "Directly modeling missing data in sequences with RNNs: Improved classification of clinical time series," in *Proc. 1st Mach. Learn. Healthcare Conf. (MLHC)*, PMLR vol. 56, 2016, pp. 253–270.
- [6] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Sci. Rep.*, vol. 8, Art. no. 6085, 2018, doi: 10.1038/s41598-018-24271-9.
- [7] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Language Process. (EMNLP)*, 2014, pp. 1724–1734, doi: 10.3115/v1/D14-1179.
- [8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
- [9] M. Moor, B. Rieck, M. Horn, C. R. Jutzeler, and K. Borgwardt, "Early prediction of sepsis in the ICU using machine learning: A systematic review," *Front. Med.*, vol. 8, Art. no. 607952, 2021, doi: 10.3389/fmed.2021.607952.
- [10] A. Vaswani et al., "Attention is all you need," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017.
- [11] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.
- [12] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *Proc. 1st Conf. Language Modeling (COLM)*, 2024, arXiv:2312.00752.
- [13] Y. Yue and Z. Li, "MedMamba: Vision Mamba for medical image classification," arXiv:2403.03849, 2024.
- [14] S. Bansal et al., "A comprehensive survey of Mamba architectures for medical image analysis: Classification, segmentation, restoration and beyond," arXiv:2410.02362, 2024.
- [15] A. Fallahpour, M. Alinoori, W. Ye, X. Cao, A. Afkanpour, and A. Krishnan, "EHRMamba: Towards generalizable and scalable foundation models for electronic health records," in *Proc. 4th Mach. Learn. Health Symp. (ML4H)*, PMLR vol. 259, 2025, pp. 291–307.
- [16] M. Zabihi, S. Kiranyaz, and M. Gabbouj, "Sepsis prediction in intensive care unit using ensemble of XGboost models," in *Proc. Computing in Cardiology (CinC)*, 2019, doi: 10.22489/CinC.2019.238.
- [17] J. Morrill, A. Kormilitzin, A. Nevado-Holgado, S. Swaminathan, S. Howison, and T. Lyons, "The signature-based model for early detection of sepsis from electronic health records in the intensive care unit," in *Proc. Computing in Cardiology (CinC)*, 2019, doi: 10.22489/CinC.2019.014.
- [18] A. Sharafoddini, J. A. Dubin, D. M. Maslove, and J. Lee, "A new insight into missing data in intensive care unit patient profiles: Observational study," *JMIR Med. Inform.*, vol. 7, no. 1, Art. no. e11605, 2019, doi: 10.2196/11605.
- [19] J. Shi, A. E. Hubbard, N. Fong, and R. Pirracchio, "Implicit bias in ICU electronic health record data: Measurement frequencies and missing data rates of clinical variables," *BMC Med. Inform. Decis. Mak.*, vol. 25, Art. no. 241, 2025, doi: 10.1186/s12911-025-03058-9.
- [20] M. Contreras et al., "Real-time prediction of intensive care unit patient acuity and therapy requirements using state-space modelling," *Nat. Commun.*, vol. 16, Art. no. 7315, 2025, doi: 10.1038/s41467-025-62121-1.
- [21] M. Wornow, S. Bedi, M. A. Fuentes Hernandez, E. Steinberg, J. Fries, C. Ré, S. Koyejo, and N. Shah, "Context Clues: Evaluating long context models for clinical prediction tasks on EHR data," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2025.
- [22] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.