

Attention-Guided Autoencoder for Retrospective Physiological Anomaly Detection in ICU Vital Signs Using MIMIC-IV Demo Data

Suprayogi¹, Mochamad Agung Wibowo², De Rosal Ignatius Moses Setiadi³

¹²Postgraduate School, Diponegoro University, Semarang, Indonesia

³Informatics Department, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia

Received:

May 1, 2026

Revised:

July 19, 2026

Accepted:

September 2, 2026

Published:

October 24, 2026

Corresponding Author:

Author Name*:

Suprayogi

Email*:

suprayogismg@gmail.com

DOI:

10.63158/journalisi.v8i5.1808

© 2026 Journal of Information Systems and Informatics. This open access article is distributed under a (CC-BY License)



Abstract. This study evaluates threshold-defined physiological anomaly screening in retrospective vital-sign data from the MIMIC-IV Clinical Database Demo v2.2. Six vital signs were temporally aligned; Fahrenheit temperature values were converted to Celsius, physiologically implausible measurements were converted to missing values before within-stay forward/backward filling, and remaining incomplete rows were removed. After cleaning, the ICU stays were separated at stay level into 82 training, 28 validation, and 28 test stays. These partitions produced 9,210, 2,784, and 5,183 windows, respectively. Only normal windows from the training partition were used to fit the neural models, while labeled validation windows were used to choose the operating thresholds. Across five seeds, AGAE achieved accuracy 0.7838, precision 0.3417, recall 0.4954, specificity 0.8335, F1 0.4025, ROC-AUC 0.7709, and PR-AUC 0.3519. The threshold-independent results for AGAE and LSTM-AE were almost the same (ROC-AUC 0.7709 versus 0.7722; PR-AUC 0.3519 versus 0.3530). At the selected operating points, AGAE produced higher accuracy, precision, and specificity and fewer false positives; LSTM-AE produced higher recall and fewer false negatives. Taken together, these results do not show a consistent advantage for AGAE over LSTM-AE, and AGAE recall remained moderate. Prospective streaming and clinical validation are required before any real-time use.

Keywords: Attention-Guided Autoencoder, ICU Vital Signs, MIMIC-IV Demo, Physiological Anomaly Screening, Reconstruction Error

1. INTRODUCTION

Continuous physiological monitoring has become more common with the growth of the Internet of Things, and electronic health-record datasets now make it possible to evaluate monitoring algorithms retrospectively and reproducibly [1], [2], [3], [4]. In this work, MIMIC-IV Clinical Database Demo v2.2 is used only as retrospective data; we do not test an online stream, bedside integration, or a deployed Healthcare IoT system. This distinction matters because ICU monitoring can generate large numbers of alarms, and excessive non-actionable alarms are associated with alarm fatigue and reduced responsiveness [5]. We therefore focus on an evaluation in which the operating trade-off between missed events and false alarms is made explicit.

We distinguish three concepts that can otherwise be conflated in monitoring studies. A sensor anomaly refers to a measurement artifact arising from device malfunction, signal loss, or measurement error. A physiological anomaly, as used here, is a value or temporal pattern that falls outside an operationally specified physiological range. Clinical deterioration is broader and requires patient-level evidence from multiple sources together with clinical judgment. The endpoint rules in this study therefore define physiological anomaly labels for evaluation; they do not adjudicate sensor faults, diagnoses, or deterioration events. The task evaluated here is threshold-defined physiological anomaly screening, not clinical deterioration prediction.

Rule-based early-warning systems such as NEWS2 and MEWS offer interpretable bedside summaries but may not fully represent multivariate temporal dependencies [6], [7]. Fixed physiological thresholds serve here as transparent evaluation labels. In contrast, reconstruction models are trained on normal multivariate windows and learn temporal covariation among six vital signs; their anomaly score can therefore reflect departures from learned temporal normality rather than merely reproducing the endpoint threshold rule. This distinction also makes the direct clinical rule circular when used as a comparator against labels generated by that same rule.

Prior work has examined ICU monitoring and unexpected ICU admission [8], alarm fatigue [5], early-warning scores [6], [7], and reconstruction-based time-series anomaly detection [9], [10], [11], [12], [13]. Attention mechanisms have also been studied in clinical and

multivariate temporal modeling [2], [14], [15], [16], [17], [18]. However, the operating characteristics of temporal attention relative to a matched LSTM autoencoder remain insufficiently characterized for cleaned MIMIC-IV Demo vital signs under stay-level separation, training-only scaling, validation-only threshold selection, multi-seed evaluation, and explicit error analysis.

Accordingly, the aim of this study is to determine whether temporal attention changes anomaly-screening performance relative to a matched LSTM autoencoder under a leakage-aware retrospective evaluation. We compare a Dense Autoencoder (DAE), LSTM-AE, and Attention-Guided Autoencoder (AGAE), together with additional conventional baselines. The study contributes a leakage-aware experimental pipeline with physiological plausibility filtering before imputation; comprehensive reporting of data splits, model architectures, threshold sensitivity, error distributions, attention patterns, statistical robustness, and runtime; and a restrained analysis of the recall-specificity trade-off. Attention weights are treated as descriptive rather than causal explanations.

2. METHODS

Figure 1 summarizes the complete retrospective experimental workflow, from MIMIC-IV Demo extraction and data cleaning through stay-level partitioning, model fitting, validation-based threshold selection, and final test evaluation. The steps are detailed in Sections 2.1–2.6.

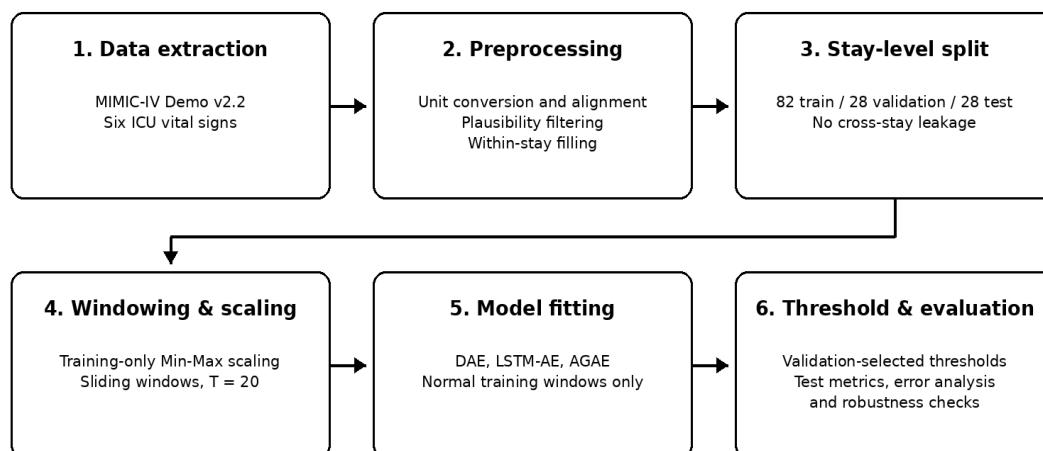


Figure 1. Retrospective experimental workflow from data extraction to final evaluation.

2.1. Dataset and Variables

Experiments used the MIMIC-IV Clinical Database Demo v2.2, a small, deidentified retrospective subset of MIMIC-IV available through PhysioNet [3], [19]. Six routinely recorded vital signs were extracted: heart rate (HR), systolic blood pressure (SBP), diastolic blood pressure (DBP), respiratory rate (RR), oxygen saturation (SpO₂), and temperature. The raw extraction contained 72,881 event rows. After alignment by stay_id and charttime, 19,775 multivariate time points were available before final complete-case handling.

2.2. Preprocessing and Missing-Data Handling

The preprocessing pipeline applied the following order: (1) raw vital-sign extraction; (2) conversion of Fahrenheit item 223761 to Celsius; (3) temporal alignment by stay_id and charttime; (4) broad physiological plausibility filtering before imputation; (5) conversion of invalid measurements to missing values without clipping; (6) forward fill within each ICU stay; (7) backward fill within each stay; (8) complete-case handling; (9) deterministic ICU-stay-level 60/20/20 partitioning with split seed 42; (10) MinMaxScaler fitting exclusively on normal training observations; (11) sliding-window generation with T=20; and (12) assignment of the last-time-step anomaly label to each window. Forward and backward filling were performed separately for each ICU stay, so imputation never carried a value from one stay into another.

The missing-data audit in Table 1 also records what happened to values removed by the plausibility filter. We found 53 implausible source measurements: three zero HR values, 47 zero RR values, and three temperature measurements outside 25–45 °C after unit conversion. Rather than clipping these observations, we set them to missing and processed them with the same within-stay filling procedure. Five rows were still incomplete after both filling steps, leaving 19,770 aligned records. Before imputation, missingness was particularly high for temperature (80.97%). The filling procedure improves continuity of the multivariate series, but it can also carry an older observation forward or smooth a short physiological change. We therefore treat this as a limitation of the preprocessing procedure and do not assume that an imputed trajectory is clinically equivalent to directly measured data.

Table 1. Missingness and within-stay imputation audit before final complete-case handling.

Vital sign	Missing n (%)	Filtered to missing	Forward Filled	Backward Filled	Still missing
Diastolic BP	6,174 (31.22)	0	5995	177	2
Heart Rate	5,865 (29.66)	3	5803	62	0
Respiratory Rate	5,909 (29.88)	47	5841	68	0
SpO2	6,235 (31.53)	0	6088	147	0
Systolic BP	6,174 (31.22)	0	5995	177	2
Temperature	16,011 (80.97)	3	15494	514	3

As Table 2 shows, the cleaned cohort contained 138 ICU stays: 82 training, 28 validation, and 28 test stays. Windows were generated independently within partitions. Neural optimization used only the 7,735 normal training windows; anomalous training windows were retained only for the split audit and were not used to fit the neural models.

Table 2. Cleaned ICU-stay-level partitions and generated windows.

Partition	ICU stays	Records	Total sequences	Normal	Anomalous	Anomaly (%)
Train	82	10,743	9,210	7,735	1,475	16.02
Validation	28	3,316	2,784	2,371	413	14.83
Test	28	5,711	5,183	4,420	763	14.72

The sensitivity of validation F1 to the candidate operating thresholds is visualized in Figure 2.

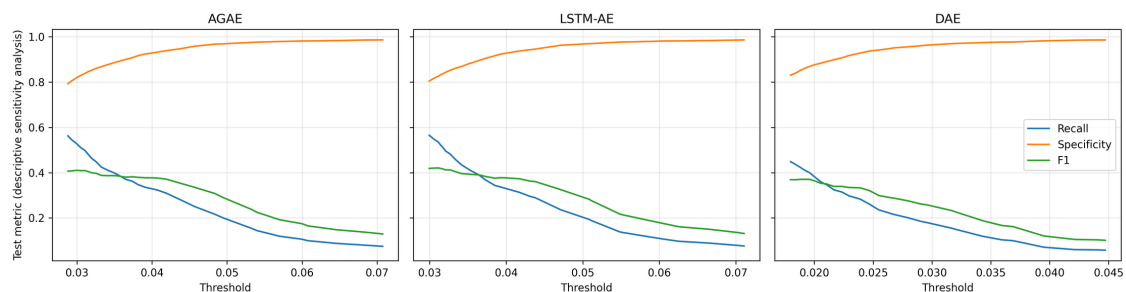


Figure 2. Threshold sensitivity across percentile and mean-plus-k-standard-deviation candidates selected or evaluated using validation data

2.3. Physiological Anomaly Labels

A time point was labeled anomalous when at least one endpoint value met the predefined rule: HR >130 or <45 bpm; SpO₂ <88%; SBP >170 or <85 mmHg; DBP >110 or <40 mmHg; RR >30 or <8 breaths/min; or temperature >39 or <34.5 °C. Each sequence inherited the label of its last time step. These threshold-derived labels were used for validation threshold selection and final evaluation, never for neural model optimization. They operationalize physiological screening and are not clinician-adjudicated deterioration outcomes.

2.4. Model Architectures

As summarized in Table 3, AGAE uses a single-layer LSTM encoder, query/key/value scaled dot-product temporal attention, a 32-dimensional latent projection, and a single-layer LSTM decoder with hidden dimension 64. DAE and LSTM-AE provide nonrecurrent and recurrent reconstruction baselines, respectively. All models calculate the anomaly score as mean squared reconstruction error over the 20×6 window.

Table 3. Architecture of the evaluated reconstruction models.

Model	Encoder	Hidden / latent	Decoder	Activations	Dropout	LSTM layers	Parameters	Error
DAE	Flatten; Linear(120,32); ReLU; Linear(32,32); ReLU	32 / 32	Linear(32,32); ReLU; Linear(32,120); Sigmoid	ReLU, Sigmoid	0	0	9,944	MSE over 20×6
LSTM-AE	LSTM(6,64); Linear(64,32)	64 / 32	Linear(32,64); repeat T=20; LSTM(64,6); Sigmoid	Sigmoid output	0	1 encoder + 1 decoder	24,352	MSE over 20×6
AGAE	LSTM(6,64); query/key/value attention; Linear(64,32)	64 / 32	Linear(32,64); repeat T=20; LSTM(64,6); Sigmoid	Softmax x attention, Sigmoid output	0	1 encoder + 1 decoder	36,832	MSE over 20×6

2.5. Training and Threshold Selection

Each neural model was trained for 50 epochs with batch size 128, Adam, learning rate 0.001, hidden dimension 64 where applicable, latent dimension 32, and sequence length 20. Five model seeds were used: 42, 52, 62, 72, and 82. The primary threshold was the candidate maximizing validation F1, with validation balanced accuracy used as a tie-breaker. Candidates comprised training-score percentiles 80–99 and $\mu + k\sigma$ for $k=0.5-3.0$. The conventional $\mu + 2\sigma$ point was retained for sensitivity analysis. A recall-prioritized operating point was also selected from validation candidates. Test labels were never consulted during threshold selection.

2.6. Evaluation and Statistical Analysis

Threshold-dependent evaluation used accuracy, precision, recall/sensitivity, specificity, false-positive and false-negative rates, balanced accuracy, and F1. ROC-AUC and PR-AUC summarized ranking independently of a single threshold. Primary results are five-seed means from validation-selected thresholds. For seed 42, 1,000 bootstrap resamples estimated 95% confidence intervals, and exact paired McNemar tests compared binary decisions. A significant McNemar result indicates different paired decisions; direction and error counts must be examined before any claim of superiority.

3. RESULTS AND DISCUSSION

3.1. Primary Five-Seed Performance

Table 4 shows a close result for AGAE and LSTM-AE when performance is considered independently of a single threshold. Mean ROC-AUC was 0.7709 for AGAE and 0.7722 for LSTM-AE; mean PR-AUC was 0.3519 and 0.3530. These differences are small and do not support a substantial advantage from attention. At the validation-selected thresholds, AGAE reached higher accuracy (0.7838), precision (0.3417), and specificity (0.8335). LSTM-AE, however, reached higher recall (0.5657) and a slightly higher F1 (0.4071 versus 0.4025). DAE ranked the classes less effectively, with ROC-AUC 0.7068 and PR-AUC 0.2984.

Table 4. Primary five-seed results using validation-selected thresholds (mean $\pm$ SD).

Model	Accuracy	Precision	Recall	Specificity	F1	Balanced acc.	ROC-AUC	PR-AUC
AGAE	0.7838 $\pm$	0.3417 $\pm$	0.4954	0.8335 $\pm$	0.4025	0.6645 $\pm$	0.7709	0.3519
	0.0198	0.0182	$\pm$	0.0323	$\pm$	0.0112	$\pm$	$\pm$
			0.0536		0.0070		0.0056	0.0068
LSTM-AE	0.7574 $\pm$	0.3181 $\pm$	0.5657	0.7905 $\pm$	0.4071	0.6781 $\pm$	0.7722	0.3530
	0.0082	0.0086	$\pm$	0.0115	$\pm$	0.0052	$\pm$	$\pm$
			0.0155		0.0069		0.0045	0.0065
DAE	0.7729 $\pm$	0.3083 $\pm$	0.4359	0.8310 $\pm$	0.3611	0.6335 $\pm$	0.7068	0.2984
	0.0049	0.0067	$\pm$	0.0066	$\pm$	0.0036	$\pm$	$\pm$
			0.0091		0.0052		0.0035	0.0034

3.2. Threshold Sensitivity Analysis

Figure 2 shows how the validation F1 changed across the percentile and $\mu + k\sigma$ threshold candidates. For the primary analysis, we used the candidate with the highest validation F1; $\mu + 2\sigma$ was kept as a conventional, more specificity-oriented reference. We also examined a validation-selected recall-prioritized point. Across five seeds, mean test recall at that point was 0.5523 for AGAE, 0.5759 for LSTM-AE, and 0.4372 for DAE; mean specificity was 0.7998, 0.7830, and 0.8290, respectively. The pattern illustrates the operating trade-off in these data: moving toward higher recall introduces more false positives, while moving toward higher specificity leaves more threshold-defined events undetected.

3.3. Additional Baselines

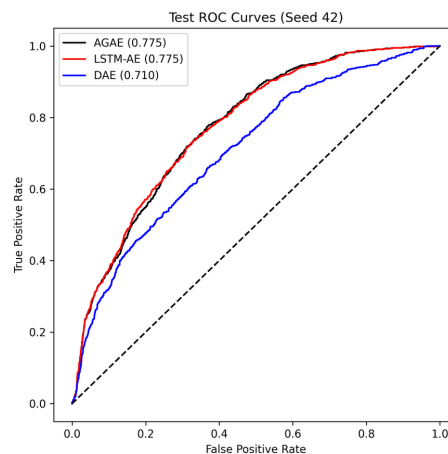
Table 5 shows that Isolation Forest achieved the highest accuracy among the non-neural unsupervised baselines, while Z-score provided relatively strong ROC-AUC; PCA reconstruction had greater recall but lower specificity. The Direct Clinical Rule is a circular upper-bound/reference check, not a fair predictive comparator, because the same endpoint rule defines the evaluation labels. It's perfect values verify label implementation only and do not demonstrate generalization.

Table 5. Additional baselines and seed-42 neural results.

Method	Accuracy	Precision	Recall	Specificity	F1	ROC-AUC	PR-AUC
Z-score	0.7957	0.3263	0.3644	0.8701	0.3443	0.7552	0.3021
Isolation Forest	0.8163	0.3682	0.3460	0.8975	0.3568	0.7444	0.3279
PCA reconstruction	0.7301	0.2623	0.4600	0.7767	0.3341	0.6581	0.2524
Direct clinical rule*	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
DAE	0.7743	0.3135	0.4482	0.8305	0.3689	0.7098	0.3004
LSTM-AE	0.7691	0.3326	0.5649	0.8043	0.4186	0.7749	0.3584
AGAE	0.7997	0.3566	0.4482	0.8604	0.3972	0.7748	0.3561

3.4. ROC and Precision–Recall Analysis

Threshold-independent discrimination on the test partition is further examined with the ROC curves in Figure 3. Figures 3 and 4 reinforce that AGAE and LSTM-AE have comparable ranking performance. Their five-seed mean ROC-AUC differs by approximately 0.0013 and mean PR-AUC by approximately 0.0011, both slightly favoring LSTM-AE. DAE separates the classes less effectively. These curves do not establish that attention substantially outperforms recurrent reconstruction without attention.

**Figure 3.** ROC curves for DAE, LSTM-AE, and AGAE on the test partition (seed 42).

The corresponding precision-recall behavior is shown in Figure 4, which is particularly informative under the observed class imbalance.

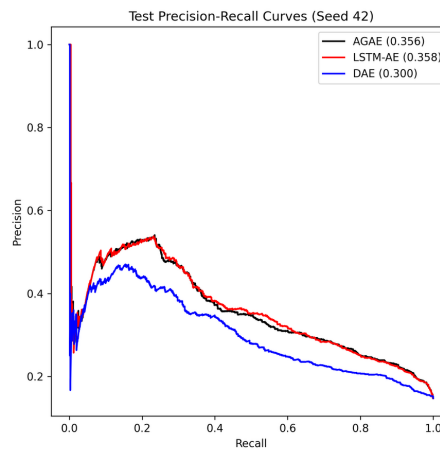


Figure 4. Precision–recall curves for DAE, LSTM-AE, and AGAE on the test partition (seed 42).

3.5. Confusion Matrices and Error Analysis

To examine the direction of classification errors at a concrete operating point, Table 6 reports the seed-42 confusion-matrix counts. At seed 42, the error counts in Table 6 and Figure 5 differed in a clinically relevant direction. AGAE produced 617 false positives and 421 false negatives, compared with 865 and 332 for LSTM-AE and 749 and 421 for DAE. Relative to LSTM-AE, AGAE therefore generated fewer false alarms but left more threshold-defined anomalies undetected. In a monitoring setting, false-positive decisions may add alarm burden, whereas false-negative decisions represent missed rule-defined events. These labels are physiological screening rules, however, so neither a detected nor a missed event should be interpreted as a confirmed deterioration outcome.

Table 6. Seed-42 confusion-matrix counts at validation-selected thresholds.

Model	TN	FP	FN	TP
AGAE	3803	617	421	342
LSTM-AE	3555	865	332	431
DAE	3671	749	421	342

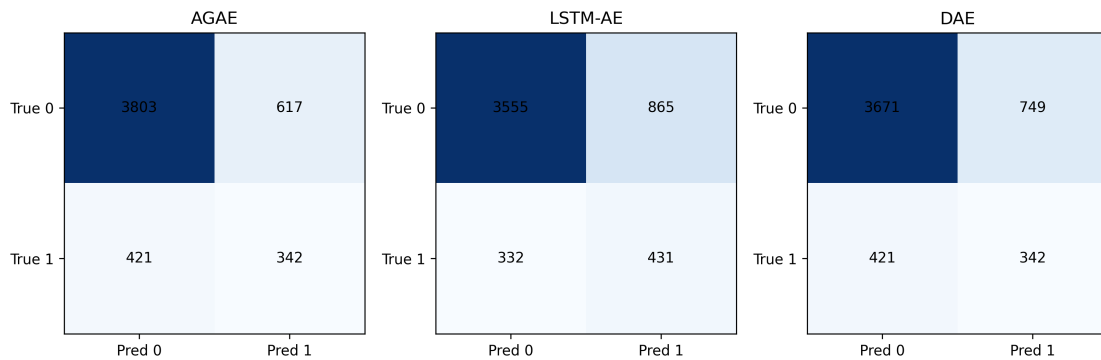


Figure 5. Seed-42 confusion matrices at validation-selected thresholds.

3.6. Reconstruction-Error Distributions

Figure 6 shows that, for AGAE, mean reconstruction error was 0.02377 for normal windows and 0.04038 for anomalous windows; corresponding LSTM-AE means were 0.02416 and 0.04106. DAE produced lower absolute errors (0.01368 and 0.02251) yet substantially lower ROC-AUC and PR-AUC. Lower reconstruction error therefore does not guarantee better anomaly discrimination; separation and overlap between class-conditional score distributions, rather than absolute loss magnitude alone, determine ranking quality.

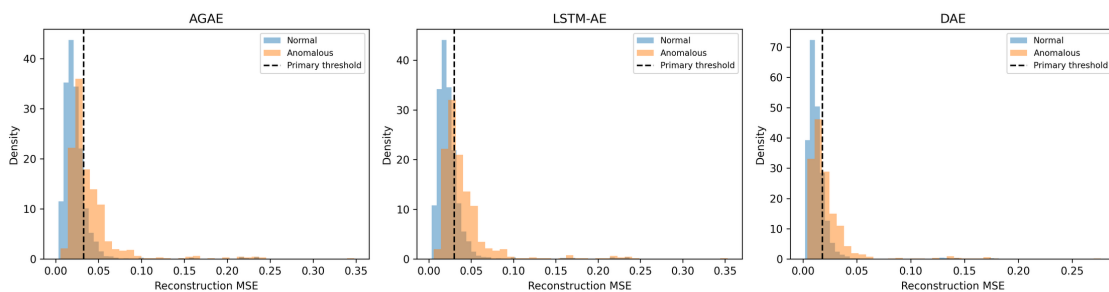


Figure 6. Reconstruction-error distributions for normal and anomalous test windows

3.7. Attention-Weight Visualization

Figure 7 presents representative true-positive, false-negative, false-positive, and true-negative examples from seed 42. The plotted weights show where the trained model allocated temporal weight within each selected window, but attention weights are descriptive and should not be interpreted as causal or definitive clinical explanations [20]. They neither identify a clinical mechanism nor validate the physiological importance of a particular time point.

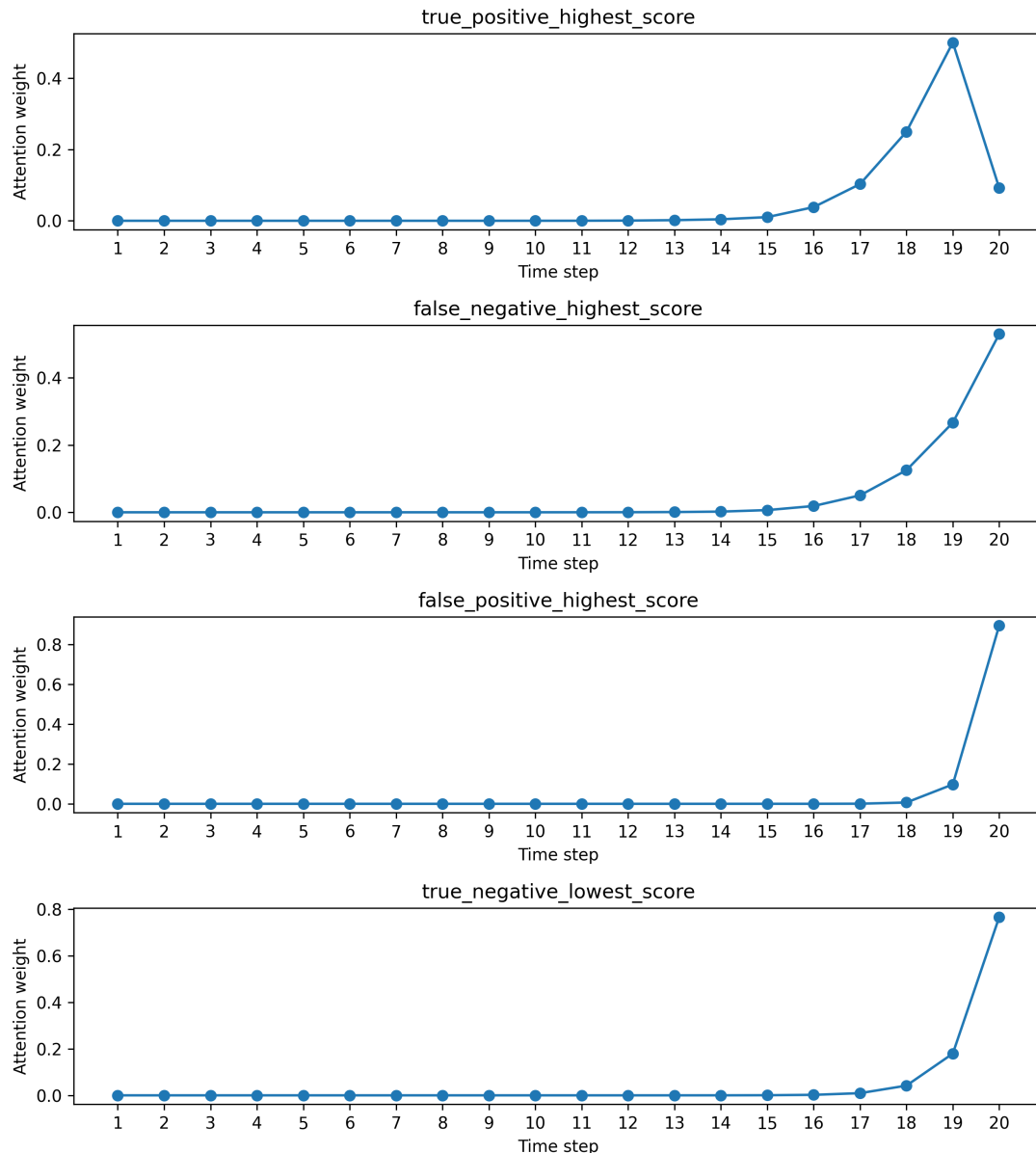


Figure 7. Descriptive AGAE attention weights for representative true-positive, false-negative, false-positive, and true-negative windows.

3.8. Statistical Robustness

Statistical uncertainty and paired decision differences were examined in two complementary analyses. Table 7 reports bootstrap confidence intervals. Tables 7 and 8 add two checks on the seed-42 results. The bootstrap intervals for AGAE and LSTM-AE overlap substantially for both ROC-AUC and PR-AUC, which is consistent with their similar discrimination. The exact McNemar tests nevertheless show different paired binary decisions (AGAE versus LSTM-AE, $p=1.69 \times 10^{-18}$; AGAE versus DAE, $p=1.47 \times 10^{-8}$). We do not

interpret those p-values as evidence of overall AGAE superiority. The discordant counts and confusion matrices instead show the direction of the operating trade-off: compared with LSTM-AE, AGAE made fewer false-positive but more false-negative decisions.

Table 7. Seed-42 bootstrap 95% confidence intervals (1,000 valid resamples)

Model	Accuracy	F1	ROC-AUC	PR-AUC
AGAE	0.7885–0.8107	0.3682–0.4262	0.7573–0.7917	0.3258–0.3902
LSTM-AE	0.7565–0.7806	0.3914–0.4440	0.7585–0.7911	0.3256–0.3942
DAE	0.7627–0.7862	0.3419–0.3955	0.6892–0.7299	0.2717–0.3296

Paired binary decisions for the same seed are then compared with exact McNemar tests in Table 8.

Table 8. Seed-42 exact paired McNemar tests

Comparison	AGAE only correct	Comparator only correct	Exact p-value
AGAE vs LSTM-AE	248	89	1.690e-18
AGAE vs DAE	336	204	1.473e-08

3.9. Computational Characterization

Table 9 reports CPU inference timing after three warm-up runs. DAE was computationally lightest, followed by LSTM-AE and AGAE. These offline inference-time measurements characterize computation only; they do not demonstrate prospective clinical real-time readiness, end-to-end streaming latency, device integration, reliability, or safety.

Table 9. CPU inference-time characterization for 5,183 test windows.

Model	Total (s)	ms/window	Windows/s	Device
AGAE	0.6545	0.1263	7,919.5	CPU
LSTM-AE	0.5799	0.1119	8,937.2	CPU
DAE	0.0388	0.0075	133,584.0	CPU

3.10. Discussion

The main finding is not a uniform advantage for either recurrent model, but a difference in their operating behavior. AGAE and LSTM-AE produced almost the same threshold-

independent discrimination. At their validation-selected thresholds, AGAE had higher mean accuracy, precision, and specificity and fewer false positives; LSTM-AE had higher mean recall, a slightly higher F1, and fewer false negatives. Thus, the attention mechanism did not consistently improve performance over the matched LSTM-AE. Which operating point is preferable depends on whether the application places greater weight on missed events or on false alarms.

Missed events remain an important limitation in both recurrent models. Mean recall across five seeds was 0.4954 for AGAE and 0.5657 for LSTM-AE, so neither detected all events defined by the physiological rules. If reducing missed rule-defined events is the priority, the LSTM-AE result or a recall-oriented threshold is more favorable in this experiment, at the cost of additional false positives. If false alarms are given greater weight, the higher-specificity AGAE operating point becomes more attractive. These are retrospective operating trade-offs; their clinical consequences would have to be measured prospectively before either choice could be justified for practice.

In AGAE, temporal attention forms a learned weighted summary of the encoder states. In this experiment, adding that mechanism changed the operating characteristics but did not produce a general performance advantage over LSTM-AE. A plausible interpretation of this result is that both recurrent models already encode much of the temporal information available in these six cleaned vital-sign sequences, so the additional weighting mechanism mainly redistributes emphasis across time rather than creating a large new separation between normal and rule-defined anomalous windows. This interpretation is consistent with the nearly identical mean ROC-AUC and PR-AUC values and with the persistence of both false-positive and false-negative errors. The data-quality safeguards also came from preprocessing—not from attention itself—through unit conversion, plausibility filtering, within-stay handling of missing values, and independent validation. For the same reason, we treat the plotted attention weights as descriptive model behavior rather than causal clinical explanations.

The DAE results help separate reconstruction magnitude from anomaly discrimination. Although DAE reconstructed both classes with lower absolute error, the normal and anomalous score distributions overlapped in a way that produced lower ROC-AUC and PR-AUC. This shows why a small reconstruction error alone is not sufficient: screening

depends on how well the score distributions separate, not simply on the absolute magnitude of the error. A second interpretive constraint comes from the labels: they were generated from fixed endpoint thresholds rather than clinician-adjudicated outcomes. The reconstruction models can learn multivariate temporal regularities, but the reported performance measures agreement with these threshold-defined physiological anomalies, not prediction of clinical deterioration. The Direct Clinical Rule is therefore useful only as a circular implementation check.

Several limitations define the scope of these findings. First, MIMIC-IV Demo contains a relatively small ICU cohort and is not a substitute for evaluation on the full MIMIC-IV database or independent centers. Second, the labels come from fixed physiological thresholds rather than clinician adjudication, so they cannot reliably separate all measurement artifacts from genuine physiological change. Missingness was also high, especially for temperature, and forward/backward filling can carry stale observations forward or smooth brief changes. We filtered implausible HR, RR, and temperature values before imputation, but did not have a broader artifact-detection procedure. Finally, recall remained moderate and false negatives persisted, while all evaluation was retrospective and offline; no prospective stream, bedside integration, multi-center validation, or clinical test was performed.

4. CONCLUSION

In this retrospective MIMIC-IV Demo evaluation, AGAE provided competitive reconstruction-based screening of threshold-defined physiological anomalies, but it did not clearly outperform LSTM-AE. Their ROC-AUC and PR-AUC were nearly comparable, while their validation-selected operating points emphasized different errors: AGAE had higher specificity and fewer false positives, whereas LSTM-AE had higher recall and fewer false negatives. AGAE recall remained moderate, leaving a meaningful number of rule-defined events undetected. These findings should therefore be read as an offline comparison of operating characteristics, not as evidence of clinical or real-time readiness. Future work should evaluate the full MIMIC-IV database and independent multi-center cohorts; use clinician-reviewed labels; develop calibrated or adaptive thresholds and improved missing-data and artifact handling; evaluate explanation methods beyond

descriptive attention weights; and conduct prospective streaming evaluation followed by appropriately governed prospective clinical testing.

ACKNOWLEDGMENT

The authors thank PhysioNet and the MIMIC-IV research team for providing access to the deidentified clinical data and acknowledge institutional support for this study.

REFERENCES

- [1] F. Ali, S. El-Sappagh, S. M. R. Islam, A. Ali, M. Attique, M. Imran, and K.-S. Kwak, "An Intelligent Healthcare Monitoring Framework Using Wearable Sensors and Social Networking Data," *Future Gener. Comput. Syst.*, vol. 114, pp. 23–43, 2021, doi: 10.1016/j.future.2020.07.047.
- [2] W. Nie, Y. Yu, C. Zhang, D. Song, L. Zhao, and Y. Bai, "Temporal-Spatial Correlation Attention Network for Clinical Data Analysis in Intensive Care Unit," *IEEE Trans. Biomed. Eng.*, vol. 71, no. 2, pp. 583–595, 2024, doi: 10.1109/TBME.2023.3309956.
- [3] A. E. W. Johnson, L. Bulgarelli, L. Shen, et al., "MIMIC-IV, a Freely Accessible Electronic Health Record Dataset," *Sci. Data*, vol. 10, Art. no. 1, 2023, doi: 10.1038/s41597-022-01899-x.
- [4] K. A. Awan, S. Khan, K. Cengiz, H. Song, and I. Alrashdi, "MediNet: Self-Supervised Framework for Multimodal Analysis and Patient Care in IoMT," *IEEE Internet Things J.*, 2025, doi: 10.1109/JIOT.2025.3614207.
- [5] K. Albanowski, K. J. Burdick, C. P. Bonafide, R. Kleinpell, and J. J. Schlesinger, "Ten Years Later, Alarm Fatigue Is Still a Safety Concern," *AACN Adv. Crit. Care*, vol. 34, no. 3, pp. 189–197, 2023, doi: 10.4037/aacnacc2023662.
- [6] Royal College of Physicians, National Early Warning Score (NEWS) 2: Standardising the Assessment of Acute-Illness Severity in the NHS, Updated Report of a Working Party. London, U.K.: Royal College of Physicians, 2017, ISBN: 978-1-86016-682-2.
- [7] C. P. Subbe, M. Kruger, P. Rutherford, and L. Gemmel, "Validation of a Modified Early Warning Score in Medical Admissions," *QJM*, vol. 94, no. 10, pp. 521–526, 2001, doi: 10.1093/qjmed/94.10.521.

- [8] S. Choi, G. Kim, J. E. Kwon, and H. K. Lee, "Sequential Anomaly Detection for Continuous Prediction of Unexpected ICU Admission from Emergency Department," *IEEE Access*, vol. 12, pp. 96304–96318, 2024, doi: 10.1109/ACCESS.2024.3426675.
- [9] J. Fan, Z. Wang, H. Wu, D. Sun, J. Wu, and X. Lu, "An Adversarial Time–Frequency Reconstruction Network for Unsupervised Anomaly Detection," *Neural Netw.*, vol. 168, pp. 44–56, 2023, doi: 10.1016/j.neunet.2023.09.018.
- [10] Z. Li and X. Zhang, "An Enhanced Autoencoder-Based Anomaly Detection Model for Time Series Data From Wearable Medical Devices," *IEEE J. Biomed. Health Inform.*, vol. 29, no. 10, pp. 7222–7233, 2025, doi: 10.1109/JBHI.2024.3434420.
- [11] E. Karapalidou, N. Alexandris, E. Antoniou, S. Vologiannidis, J. Kalomiros, and D. Varsamis, "Implementation of a Sequence-to-Sequence Stacked Sparse Long Short-Term Memory Autoencoder for Anomaly Detection on Multivariate Timeseries Data of Industrial Blower Ball Bearing Units," *Sensors*, vol. 23, no. 14, Art. no. 6502, 2023, doi: 10.3390/s23146502.
- [12] S. Guan, Z. He, S. Ma, and M. Gao, "Conditional Normalizing Flow for Multivariate Time Series Anomaly Detection," *ISA Trans.*, vol. 143, pp. 231–243, 2023, doi: 10.1016/j.isatra.2023.09.002.
- [13] W. Zhu, W. Li, E. R. Dorsey, and J. Luo, "Unsupervised Anomaly Detection by Densely Contrastive Learning for Time Series Data," *Neural Netw.*, vol. 168, pp. 450–458, 2023, doi: 10.1016/j.neunet.2023.09.038.
- [14] J. Mi, Y. Zu, Z. Wang, and J. He, "ACDNet: Attention-Guided Collaborative Decision Network for Effective Medication Recommendation," *J. Biomed. Inform.*, vol. 149, Art. no. 104570, 2024, doi: 10.1016/j.jbi.2023.104570.
- [15] R. Suzui, J. Natsume, M. Kawaguchi, Y. Ito, H. Yamamoto, T. Nakata, H. Kidokoro, Y. Takahashi, S. Tatsuki, and K. Fujiwara, "Prediction of Long-Term Prognosis in Infantile Epileptic Spasms Syndrome of Unknown Etiology Based on Hypsarrhythmia by Self-Attention Autoencoder," *IEEE J. Biomed. Health Inform.*, vol. 30, no. 7, pp. 5858–5868, 2026, doi: 10.1109/JBHI.2025.3647091.
- [16] C. Wang, Z. Wang, H. Dong, S. Lauria, W. Liu, Y. Wang, F. Fadzil, and X. Liu, "Fusionformer: A Novel Adversarial Transformer Utilizing Fusion Attention for Multivariate Anomaly Detection," *IEEE Transactions on Neural Netw. and Learning Systems*, vol. 36, no. 8, pp. 14479–14492, 2025, doi: 10.1109/TNNLS.2025.3542719.

- [17] W. Hu, L. Cao, Q. Ruan, and Q. Wu, "Research on Anomaly Network Detection Based on Self-Attention Mechanism," *Sensors*, vol. 23, no. 11, Art. no. 5059, 2023, doi: 10.3390/s23115059.
- [18] Y. Chen, H. Zhang, Y. Wang, Y. Yang, X. Zhou, and Q. M. J. Wu, "MAMA Net: Multi-Scale Attention Memory Autoencoder Network for Anomaly Detection," *IEEE Trans. Med. Imaging*, vol. 40, no. 3, pp. 1032–1041, 2021, doi: 10.1109/TMI.2020.3045295.
- [19] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark, "MIMIC-IV Clinical Database Demo," version 2.2, PhysioNet, 2023, doi: 10.13026/dp1f-ex47.
- [20] S. Jain and B. C. Wallace, "Attention Is Not Explanation," in Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), vol. 1, pp. 3543–3556, 2019, doi: 10.18653/v1/N19-1357.