Beyond the ROC Curve: Activity Monitoring to Evaluate Deep Learning Models in Clinical Settings

Hyunwoo CHOO[1,*], Kyung Hyun LEE[1], Sungsoo HONG[1], Sungjun HONG[2], Ki-Byung LEE[1,3], and Chang Youl LEE[3]

[1] AITRICS Inc., 218 Teheran-ro, Gangnam-gu, 06221Seoul, Republic of Korea
[2] Medical AI Research Center, Research Institute for Future Medicine, Samsung Medical Center, 81 Irwon-ro, Gangnam-gu, 06351 Seoul, Republic of Korea
[3] Division of Pulmonary, Allergy and Critical Care Medicine, Hallym University Chuncheon Sacred Heart Hospital, 77 Sakju-ro, 24253 Chuncheon, Republic of Korea
E-mails: sshong@aitrics.com; hwchoo@aitrics.com; lkh256@aitrics.com; hsj2864@skku.edu; hasej@aitrics.com; doclcy@hallym.or.kr

* Author to whom correspondence should be addressed

Received: 12 July 2024/Accepted: 11 September 2024 / Published online: 21 November 2024

Abstract
We evaluated 'VITALCARE-SEPS', a deep learning model for sepsis prediction, using the activity monitoring operator characteristics curve with two different scoring algorithms. This evaluation is crucial as the AMOC curve addresses the time-dependent nature of predictions, providing a more nuanced performance assessment than traditional ROC metrics. Our findings demonstrate that the AMOC curve significantly enhances the evaluation of time-series predictions, enabling more accurate and continuous performance monitoring of machine learning models in clinical settings. This approach can improve model deployment and ultimately lead to better patient outcomes in healthcare.

Keywords: Deep learning; Sepsis; Activity monitoring; ROC (receiver operating characteristic) curve

Introduction

The area under the receiver operating characteristic (ROC) curve (AUC-ROC) is a well-established metric for evaluating machine learning models in binary classification tasks. It provides a single scalar value summarizing the trade-off between true positive and false positive rates across various thresholds. However, AUC-ROC may not be ideal for time-series contexts or dynamic environments, such as clinical settings, as it treats each inference point equally without considering temporal aspects or the sequence of predictions.

To address this limitation, time-dependent ROC curves were introduced [1,2] to evaluate model performance at different time points, showing how prediction accuracy evolves over time. Yet, this approach still falls short in capturing the timing of predictions relative to clinical outcomes, failing to reward early predictions or penalize late ones, which is crucial in time-sensitive scenarios like sepsis detection.

Tom Fawcett proposed the Activity Monitoring Operator Characteristics [3] (AMOC) curve to offer a more nuanced evaluation of time-series predictions. The AMOC curve extends traditional and time-dependent ROC analyses by incorporating the timing of predictions, rewarding early correct predictions and penalizing late or incorrect ones. This approach provides a more comprehensive assessment, especially where timely interventions are critical.

Despite its advantages, the AMOC curve remains underutilized for evaluating time-series predictions. This study introduces the AMOC curve to assess the deep learning model 'AITRICS-VC SEPS,' used for early sepsis detection in clinical settings. By comparing the AMOC curve with traditional and time-dependent ROC curves, we show that the AMOC curve captures both accuracy and timing of predictions, highlighting its effectiveness and advocating for its broader use in scenarios where timely predictions are crucial.

Materials and Methods

Study Design

We evaluated the deep learning model 'VITALCARE-SEPS', deployed in clinical settings. The model uses five vital signs, twelve lab results, and demographic information to generate a pseudo-probability of sepsis (0 to 1) upon new events (e.g., new lab results). Model inference outputs were retrospectively collected from EMR data of patients with sepsis infected from Hallym University Chuncheon Sacred Heart Hospital (2018-2022). Each model output was binarily labeled based on whether sepsis occurred within four hours. AMOC curves were drawn from these labeled inference results.

Activity Monitor Operator Characteristics

To illustrate a typical curve, we plot the false alarm rate against the average score. We evaluate each model inference output against threshold values from 0 to 1, at 0.01 intervals. This systematic quantization allows assessment across all thresholds. Scoring each inference across all patients and aggregating the results yields counts of true and false alarms and the average score.

Scoring Functions Used to Draw AMOC Curve

To evaluate model performance and construct the AMOC curve, we employed two scoring functions that account for both prediction accuracy and the timing of predictions:

1) Simple Score Calculation:

This function assigns a binary score where a true positive (score = 1) is recorded if the predicted probability (PredictionScore) exceeds a defined threshold and the event occurs (ObservedOutcome = 1). All other cases receive a score of 0.

2) Time-Sensitive Scoring:

This advanced function rewards early and moderate correct predictions differently while penalizing missed events and false positives. Specifically:

- Early Correct Predictions: Receive a higher reward if the event is predicted before a pre-defined time threshold.
- Moderate Correct Predictions: Receive a lower reward if predicted after this threshold.
- Penalties: Applied for missed events and false positives, scaled based on the time to event occurrence.

Results

From 2018 to 2022 in Hallym University Chuncheon Sacred Heart Hospital, we retrospectively collected EMR data from 3,460 sepsis positive patients. From those patients, we generated 93,922 model inference outputs. We illustrated two AMOC curves applying different scoring algorithms on Figure 1.

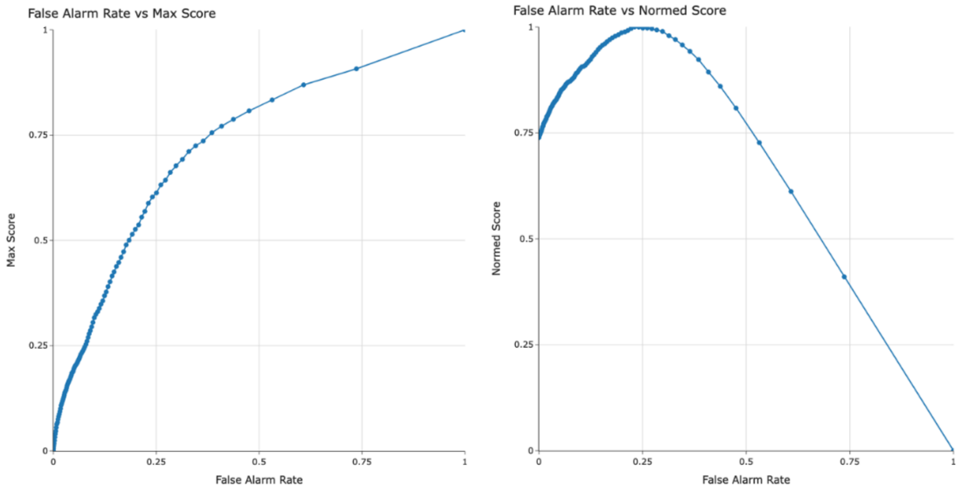

Figure 2 (Left) AOMC curve drawn using simple scoring function **(Right)** AOMC curve drawn using time-sensitive scoring function

Discussion

In this study, we used the Activity Monitoring Operator Characteristics (AMOC) curve to evaluate the 'VITALCARE-SEPS' deep learning model for sepsis prediction, applying two different scoring functions. This analysis highlights each approach's strengths and limitations in capturing the timing of predictions, crucial in clinical settings.

The AMOC curve using the simple scoring function (Figure 2, left) treats all correct predictions equally, regardless of timing. This method results in a curve where increases in the false alarm rate often correspond to proportional increases in the score. While straightforward, it does not account for the timing of predictions, potentially overestimating model performance in time-sensitive contexts.

Conversely, the AMOC curve with the time-sensitive scoring function (Figure 2, right) addresses this by incorporating prediction timing into the evaluation. It rewards early correct predictions and penalizes missed events and false positives based on their timing. This results in a more nuanced curve that better reflects the practical needs of clinical decision-making, distinguishing between models that perform well in a timely manner and those that do not.

The time-sensitive scoring function enhances the AMOC curve's intuitiveness for real-time monitoring and evaluation. By accounting for the timing of predictions, it provides a more accurate assessment of model effectiveness in critical settings like sepsis detection. This refined approach could improve the deployment and performance of machine learning models in healthcare, potentially leading to better patient outcomes.

Further research is needed to explore optimal reward and penalty parameters for the time-sensitive scoring function and their impact on model evaluation.

Conclusions

The AMOC curve enhances the evaluation of time-series predictions, leading to more accurate and continuous monitoring of clinical machine learning models.

List of Abbreviations: activity monitoring operator characteristics (AMOC)

Author Contributions: HC defined the research aim, designed the experiment, created the figures, and wrote the manuscript. HSS HSJ, and KL participated in the research design, conducted the experiments, and critically reviewed the manuscript.

Funding: This research received no funding.

Ethics Statement: This study is approved by the institute review board of Hallym University Chuncheon Sacred Heart Hospital (IRB No: CHUNCHEON 2023-03-007-002). The informed consent is waived as this is a retrospective study that has minimal risk to patients.

Data Availability Statement: The data utilized in this study were obtained under the supervision and with the grant support of Hallym Chuncheon Sacred Heart Hospital. For access to the data, individual contact with the hospital is required.

Conflict of Interest: HC, SSH, KHL and KBL are employee of AITRICS.

References

1. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics. 2000;56(2):337-344. doi: 10.1111/j.0006-341x.2000.00337.x.
2. Kamarudin AN, Cox T, Kolamunnage-Dona R. Time-dependent ROC curve analysis in medical research: current methods and applications. BMC Med Res Methodol. 2017;17(1):53. doi: 10.1186/s12874-017-0332-6.
3. Fawcett T, Provost F. Activity monitoring: Noticing interesting changes in behavior. In: KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. 1999:53-62. doi: 10.1145/312129.312195.