MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI230970

Time-Series Aware Metrics for the Evaluation of Intraoperative Electroencephalography-Based Ischemia Detection

Amir I. Mina [a], Jeremy U. Espino [a], Allison M. Bradley [a], Parthasarathy Thirumala [b], Kayhan Batmanghelich [a], Shyam Visweswaran [a,1]
[a] *Department of Biomedical Informatics, University of Pittsburgh, Pennsylvania, USA*
[b] *Department of Neurological Surgery, University of Pittsburgh, Pennsylvania, USA*
ORCiD ID: Amir I. Mina https://orcid.org/0000-0002-1231-1696

Abstract. Continuous intraoperative monitoring with electroencephalography (EEG) is commonly used to detect cerebral ischemia in high-risk surgical procedures such as carotid endarterectomy. Machine learning (ML) models that detect ischemia in real time can form the basis of automated intraoperative EEG monitoring. In this study, we describe and compare two time-series aware precision and recall metrics to the classical precision and recall metrics for evaluating the performance of ML models that detect ischemia. We trained six ML models to detect ischemia in intraoperative EEG and evaluated them with the area under the precision-recall curve (AUPRC) using time-series aware and classical approaches to compute precision and recall. The Support Vector Classification (SVC) model performed the best on the time-series aware metrics, while the Light Gradient Boosting Machine (LGBM) model performed the best on the classical metrics. Visual inspection of the probability outputs of the models alongside the actual ischemic periods revealed that the time-series aware AUPRC selected a model more likely to predict ischemia onset in a timely fashion than the model selected by classical AUPRC.

Keywords. Cerebral ischemia, Carotid endarterectomy, Electroencephalography, Time-series aware precision, Time-series aware recall, Area under the precision-recall curve

1. Introduction

Continuous electroencephalography (EEG) is frequently used to detect cerebral ischemia during high-risk surgical procedures such as carotid endarterectomy (CEA)[1]. Typically, a neurophysiologist visually monitors the EEG for changes indicative of ischemia; however, such monitoring is mentally taxing and prone to error. Machine learning (ML) models that detect ischemia in real time can form the basis of automated intraoperative EEG monitoring. We have developed supervised ML models that output a prediction for

[1]Corresponding Author: Shyam Visweswaran, shv3@pitt.edu.

ischemia for each 20-second interval of EEG signals. Assuming that each 20-second interval is an independent data point, these predictions can be evaluated using classical metrics like precision and recall. However, ischemia is a time-dependent phenomenon that occurs over a period, and in our application, a predicted ischemic period consists of a series of contiguous 20-second intervals. Therefore, classical precision and recall metrics cannot capture the crucial time-series characteristics of ischemic periods.

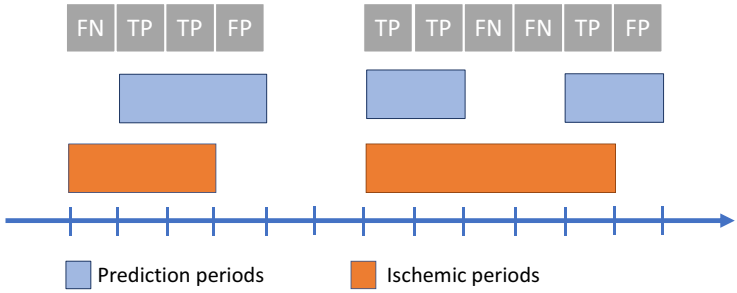

Figure 1. Illustrative examples of two ischemic periods (in orange) and three prediction periods (in blue). The tick marks on the arrow at the bottom represent 20-second intervals. The grey boxes at the top indicate whether a 20-second interval is a TP, FP, or FN. For this example, classical precision is 0.71 (TPs = 5 and FPs = 2) and classical recall is 0.62 (TPs = 5 and FNs = 3). Note that these precision and recall values ignore the overlap length and the overlap's position between the ischemic and prediction periods.

Figure 1 presents illustrative examples of ischemia predictions to shed light on the limitations of classical precision and recall. Under the classical consideration, a predicted ischemic 20-second interval is either a member of the set of real ischemic intervals (referred to as a true positive, or TP) or not (referred to as a false positive, or FP), and a real ischemic 20-second interval is either a member of the set of predicted ischemic intervals (TP) or not (referred to as a false negative, or FN). Precision is the proportion of all predicted 20-second ischemic intervals that are, in fact, ischemic $[TP/(TP+FP)]$, while recall is the proportion of all actual 20-second ischemic intervals that are predicted as ischemic $[TP/(TP+FN)]$. However, under time-series aware considerations, a predicted ischemic period (a series of contiguous intervals) may partially overlap with a real ischemic period, resulting in a prediction that is partly a TP and partly an FP. Consequently, the length of the overlap must be quantified, and the overlap's position must be considered. For example, a real-time ischemia-detection application must accurately predict the onset of an ischemia period (i.e., its "front-end") for the surgical team to respond in a timely fashion.

We now describe two time-series aware precision and recall metrics that have been described in the literature. Both employ a weighted score comprised of two terms, called detection and coverage, which are averaged across periods to determine either precision or recall (see Equation 1).

$$Precision/Recall = \alpha \times Detection + (1-\alpha) \times Coverage \quad (1)$$

The **range-based recall and precision** (RPR)[2] approach considers an ischemic period to be detected if there exists any overlap with any prediction periods. Two multipliers are applied to the length of overlap to determine the coverage. The first multiplier is determined by a positional bias and can be selected to favor an overlap occurring at the "front," "middle," or "end" of the ischemic period. If no positional bias

is desired, the "flat" value is employed. The second multiplier is cardinality, which emphasizes and rewards having fewer prediction periods for a given ischemic period.

The **time-series aware precision and recall** (TaPR)[3] approach employs flexible hyperparameters for the proportion of overlap required for precision (θ_p) and recall (θ_r) for the detection term. The computation of the overlap is affected by the ambiguity of the time ranges of the ischemic periods. The ambiguous period is assigned an inverse sigmoid curve, which defines the area within which any portion of the prediction within the ambiguous region can be computed and added to the overlap. TaPR does not consider cardinality or positional bias, whereas RPR does not take into account tunable detection thresholds or ambiguous periods.

In this study, we compare the area under the precision-recall curve (AUPRC) derived from time-series aware precision and recall computed using the RPR and TaPR approaches to the AUPRC derived from classical precision and recall for evaluating the performance of ML models that predict ischemia.

2. Methods

2.1. Data and modeling

We created a training data set from intraoperative EEG recordings on 766 patients who underwent CEA between 2009 and 2017 at a large academic medical center. All EEG recordings captured eight channels: F3-P3, P3-O1, F3-T3, T3-O1, F4-P4, P4-O2, F4-T4, and T4-O2. Channels with odd numbers (1, 3) correspond to electrode placement on the left hemisphere, whereas channels with even numbers (2, 4) correspond to placement on the right hemisphere. A neurophysiologist annotated the time periods in each recording that were indicative of ischemia. We extracted 10 minutes of post-clamp EEG signal from each recording and partitioned it into 20-second intervals after applying low-pass (70Hz), high-pass (0.166Hz), and notch (60Hz) filters[4]. Since the risk of ischemia is highest during the first 10 minutes after the diseased carotid artery is clamped, we chose this time frame. Each 20-second interval yielded 111 features, the majority of which were computed from a Fourier Transform power spectrogram. Based on the neurophysiologists' annotations, we labeled each 20-second interval as ischemia or no ischemia.

We trained several supervised ML models to predict the ischemia / no ischemia labels. Since current state-of-the-art models for tabular data[5] are tree-based, we trained random forest (RF) and XGBoost random forest models. We also included additional boosting models such as Histogram Gradient Boosting (HGB) and Light Gradient Boosting Machine (LGBM) classifiers. Other models included logistic regression (LR) and Support Vector Machine for classification (SVC). For each model, we obtained a set of 10 trained versions through 10-fold cross-validation.

2.2. Evaluation

Each model was evaluated on an independent test data set from the corresponding iteration of the cross-validation. With the prediction probabilities, we computed the AURPC from time-series aware precision and recall values using the RPR and TaPR approaches, as well as the AUPRC from the classical precision and recall values. To

calculate the time-series aware precision and recall, we concatenated the predictions from all patients into a single contiguous sequence. To ensure that there was no accidental continuity of a prediction period spanning two adjacent patients, we added padding between the predictions from the two patients. The padding consisted of 100 20-second intervals, each of which was assigned a label of no ischemia and a prediction probability of zero.

For each ML algorithm, we trained 1000 models from 1000 bootstrapped data sets and obtained the AUPRCs at each iteration of the bootstrap. We identified the model that was ranked first among the bootstrap iterations most frequently for each AUPRC approach and recorded how often that model was ranked first. To examine the behavior of models visually, we plotted the probability outputs of the models alongside the spectrogram of the signals at each channel and the labeled time ranges for patients who experienced an ischemic period.

AUPRC	*Most frequently first-ranked model*	*Percent iterations*
Classical	LGBM	64.1
RPR (front)	SVC	95.5
TaPR	SVC	83.5

Table 1. Most frequently first-ranked model for three AUPRC metrics and the percent bootstrapped iterations that the first-ranked model was the top model.

3. Results

Based on the time-series aware AUPRC, the most frequently first-ranked model was SVC for both RPR and TaPR approaches; it was the top model in 95.5% and 83.5% of the bootstrap iterations, respectively (see Table 1). Based on the classical AUPRC, the most frequently top-ranked model was LGBM; it was the top model in 64.1% of the bootstrap iterations (see Table 1). Visual examination of the predictions of SVC and LGBM showed that SVC was more likely to accurately predict the onset of an ischemia period than LGBM. Figure 3 provides an example of a patient.

4. Discussion

We found that the top-performing model differed between classical metrics and time-series aware metrics, indicating that it is crucial to consider the adequacy of the evaluation metric used to compare the performances of EEG-based models to detect ischemia. Visual examination revealed that the behavior of the model with the highest time-series aware AUPRC was more suitable for the EEG monitoring application, as early detection of cerebral ischemia during surgery is crucial.

5. Conclusions

The advent of time-series-aware metrics provides researchers interested in predictions in time-series data with an improved means to evaluate models. Models that detect changes

in EEG are typically evaluated using classical metrics; however, the utilization of time-series-aware metrics developed for other applications is better suited for our application.

These metrics are relatively new, so their adoption, particularly for healthcare-related applications, will take time. We note their benefits for the purpose of selecting an EEG-based model to detect ischemia during CEA. The next steps include using time-series-aware metrics during training and calibration.

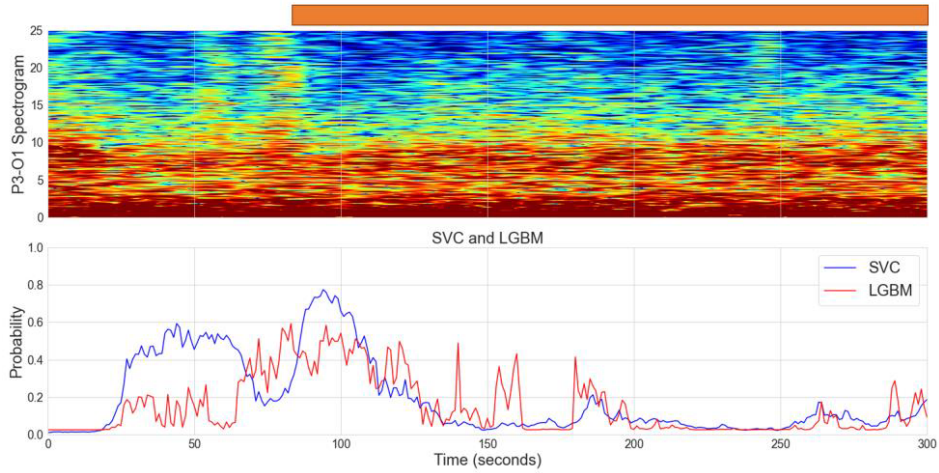

Figure 2. An example patient's spectrogram (top panel) and the predicted probabilities for the SVC and LGBM models (bottom panel). The orange bar at the top indicates an ischemic period. The SVC model (in blue) detects the beginning of the ischemic period earlier than the LGBM model (in red).

References

[1] Plestis KA, Loubser P, Mizrahi EM, Kantis G, Jiang ZD, Howell JF. Continuous electroencephalographic monitoring and selective shunting reduces neurologic morbidity rates in carotid endarterectomy. *Journal of Vascular Surgery*. 1997;25(4):620-628.

[2] Tatbul N, Lee TJ, Zdonik S, Alam M, Gottschlich J. Precision and recall for time series. *Advances in neural information processing systems*. 2018;31.

[3] Hwang WS, Yun JH, Kim J, Kim HC. Time-series aware precision and recall for anomaly detection: considering variety of detection result and addressing ambiguous labeling. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2019:2241-2244.

[4] Libenson MH. Practical Approach to Electroencephalography. Elsevier Health Sciences. 1st Edition. 2009 Dec 4.

[5] Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*. 2022 Dec 6;35:507-20.