2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871916

Transformer Convolutional Neural Networks for Automated Artifact Detection in Scalp EEG

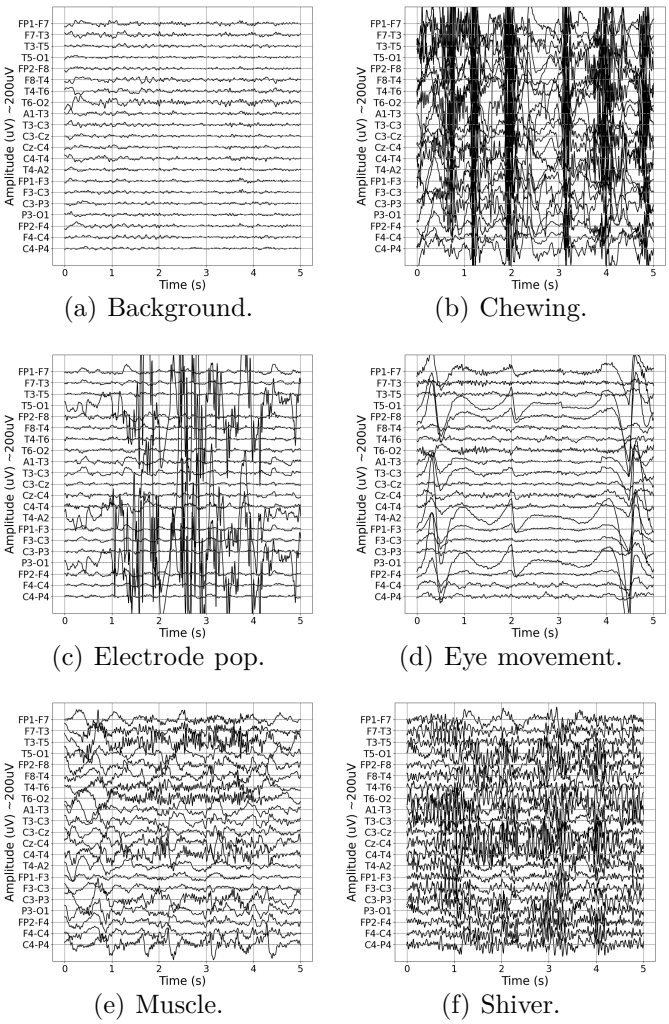
Wei Yan Peh[1*], Yuanyuan Yao[2], and Justin Dauwels[2*]

Abstract— It is well known that electroencephalograms (EEGs) often contain artifacts due to muscle activity, eye blinks, and various other causes. Detecting such artifacts is an essential first step toward a correct interpretation of EEGs. Although much effort has been devoted to semi-automated and automated artifact detection in EEG, the problem of artifact detection remains challenging. In this paper, we propose a convolutional neural network (CNN) enhanced by transformers using belief matching (BM) loss for automated detection of five types of artifacts: chewing, electrode pop, eye movement, muscle, and shiver. Specifically, we apply these five detectors at individual EEG channels to distinguish artifacts from background EEG. Next, for each of these five types of artifacts, we combine the output of these channel-wise detectors to detect artifacts in multi-channel EEG segments. These segment-level classifiers can detect specific artifacts with a balanced accuracy (BAC) of 0.947, 0.735, 0.826, 0.857, and 0.655 for chewing, electrode pop, eye movement, muscle, and shiver artifacts, respectively. Finally, we combine the outputs of the five segment-level detectors to perform a combined binary classification (any artifact vs. background). The resulting detector achieves a sensitivity (SEN) of 42.0%, 32.0%, and 13.3%, at a specificity (SPE) of 95%, 97%, and 99%, respectively. This artifact detection module can reject artifact segments while only removing a small fraction of the background EEG, leading to a cleaner EEG for further analysis.

I. INTRODUCTION

Electroencephalography (EEG) is a widely used technology in neurology, e.g., helpful for the diagnosis of epilepsy [1]. However, EEG recordings often contain artifacts, which can be due to eyeblinks, head movements, chewing, interference from electronic equipment, and other causes [2]. These artifacts may resemble epileptiform abnormalities or other transient waveforms, resulting in mistakes during annotation [3]. Knowledge of the plausible scalp distribution of EEG abnormalities is essential to distinguish artifacts from brain waves [4]. For instance, muscle artifacts usually appear in multiple channels, whereas artifacts such as electrode pop may only be visible in a single channel.

When reading EEGs, one must distinguish artifacts from brain waves. An automatic artifact detection system can improve the readability of an EEG [2]. Common artifact rejection and detection methods includes high amplitude rejection [5], common average referencing (CAR) [6], independent component analysis (ICA) [4], [3], wavelet transforms (WT) [4], machine learning [7], convolutional neural networks (CNN) [8], [3], and generative adversarial networks (GAN) [3]. The majority of the studies are not validated on large datasets (more than 100 patients), but instead on small datasets (less than 100 patients) or semi-simulated datasets [4], [8] by injecting noise into regular EEGs. Additionally, most studies detect artifacts directly from multi-channel segments [7]; as a result, many of those methods are only applicable to a fixed number of channels, whereas the proposed method can be applied to EEG with any number of channels. Ultimately, most studies failed



(a) Background. (b) Chewing.

(c) Electrode pop. (d) Eye movement.

(e) Muscle. (f) Shiver.

Fig. 1. Examples of clean EEG and EEG with various types of artifacts.

[1]School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798, Singapore (*corresponding author email: pehw0012@e.ntu.edu.sg).

[2]Department of Microelectronics, Delft University of Technology, 2628 CD Delft, Netherlands (*corresponding author email: j.h.g.dauwels@tudelft.nl).

Fig. 2. Channel-level and segment-level analysis of EEG.

TABLE I
Summary of the TUH ART EEG dataset.

Background/Artifacts	Duration (hr)	Number of Events
Background	62.533	-
Chewing	0.995	804
Electrode Pop	5.554	6090
Eye Movement	8.04	9480
Muscle	16.303	11267
Shiver	0.047	31

to deploy proper evaluation metrics to measure the effectiveness of their methods [5], [6]. Consequently, it is challenging to compare the existing artifact detectors.

In this paper, we propose a CNN equipped with a transformer (CNN-Transformer) trained through a belief matching loss (BM) to detect five different types of artifacts (see Figure 1) from the TUH Artifact (TUH-ART) dataset. The proposed system detects artifacts in individual EEG channels and also in multi-channel EEG segments (see Figure 2) [9]. The artifact detector can detect specific artifacts at segment-level with a balanced accuracy (BAC) of 0.947, 0.735, 0.826, 0.857, and 0.655 for chewing, electrode pop, eye movement, muscle, and shiver artifacts, respectively. When combined to perform binary artifact classification (any artifact type vs. background EEG), the binary artifact detector achieves a sensitivity (SEN) of 42.0%, 32.0%, and 13.3% at 95%, 97%, and 99% specificity (SPE), respectively. This artifact detector can detect specific artifacts and reject them from EEGs, resulting in a cleaner EEG for a better reviewing experience.

II. Methods

A. Scalp EEG recordings and preprocessing

In this study, we analyzed the public TUH Artifact Corpus (TUH-ART), containing EEGs with artifact annotations [10]. The dataset consists of five artifacts types: chewing, electrode pop, eye movement, muscle, and shiver (see Table I). On each EEG, we applied a Butterworth notch filter (4^{th} order) at 60Hz (USA) to remove interference and a 1Hz high-pass filter (4^{th} order) to remove noise [9]. We downsampled all EEGs to 128Hz. We trained the artifact detectors via 5-fold cross-validation (CV), where each fold contains different patients and similar distribution across all five artifact types.

B. Channel-level Artifact Detection

First, we develop a system to detect artifacts at individual EEG channels (channel-level analysis). We train a separate channel-wise detector on the TUH-ART dataset for each of the five artifact types. The channel-level artifact detector is a CNN cascaded with a transformer, while the learning objective function is a BM loss [11] (see Figure 3).

A CNN is not adequate for modeling correlations between distant data points. This inherent limitation makes

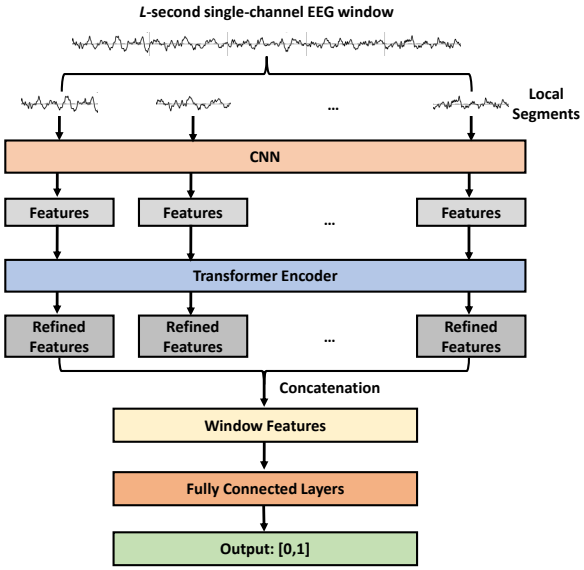


Fig. 3. CNN with transformer encoder.

CNN less suitable for time series, especially when correlations over relatively long periods are expected, such as long artifacts patterns (e.g., eye movement artifacts). Therefore, we augment the CNN with a transformer to compensate for this limitation since transformers can extract long-range patterns in the features extracted by the CNN.

In addition, it is essential to have a reliable measure of the uncertainty associated with a detection (output of the neural network) such that we can be confident in the detections with low uncertainty. To this end, we deploy a BM loss instead of the traditional softmax (SM) loss, as it yields more reliable uncertainty estimates [11]. The BM framework is a Bayesian approach that views the binary classification from a distribution matching perspective, making it a more reliable detector. Moreover, Joo et al. observed improvements in generalization, a desirable property for the application at hand [11]. The BM loss is defined as:

$$\mathcal{L}(\mathrm{W}) \approx -\frac{1}{m}\sum_{i=1}^{m} l_{\mathrm{EB}}\left(y^{(i)}, \alpha^{\mathrm{W}}(\mathrm{x}^{(i)})\right), \qquad (1)$$

where $\mathrm{x}^{(i)}$ and $y^{(i)}$ are the i-th training data and its label, respectively, m is the total number of training data, l_{EB} is the evidence lower bound (ELBO) [11], and $\alpha^{\mathrm{W}} = \exp(\mathrm{W})$, where W are the weights of the neural network classifier.

The input of the CNN-Transformer is the L-second single-channel EEG window that is split into 0.5s local segments with 25% overlap (see Figure 3). We trained the model with different window lengths L, i.e., 1, 3, and 5s. We varied the window lengths to determine the best window length to detect artifacts. For instance, a short window length of 1s is adequate to remove short-duration eye blinks, while a long window length of 5s is suitable to remove long-duration muscle artifacts.

The CNN architecture consists of 5 convolution layers,

Fig. 4. Artifact detection pipeline.

all with a filter size of 3 and a ReLU activation function. The number of filters is 8, 16, 32, 64, and 128 in layers 1, 2, 3, 4, and 5, respectively. After each convolution layer, we apply max-pooling with stride 2. Next, we deploy a transformer encoder to identify patterns in the features extracted by the CNN [12]. The encoder relies on an activation function that maps the query and a set of key-value pairs to an output. Here, the local features extracted by the CNN are the query, key, and value simultaneously. We set the number of heads in the transformer to the commonly chosen value of 8 [12] and the number of neurons in the hidden layer of the feed-forward network (FNN) module to 1024. Two fully connected (FC) layers containing 100 and 2 neurons follow the CNN-Transformer module. Before the final FC layer, we include a dropout layer with a probability of 0.5. The output of the second FC is the prediction for that particular window of EEG. Finally, we applied the Adam optimizer [13] with an initial learning rate equal to 10^{-4} to minimize the BM loss. The batch size for training is 1000. We applied balanced training to avoid overfitting during training by applying weights to each class. We optimized the hyperparameters of the CNN-Transformer via nested CV on the training data with an 80%:20% split for training and validation.

C. Segment-level Artifact Detection

Next, we wish to detect artifacts in multi-channel segments (see Figure 4). To perform binary classification of a specific artifact type, we performed the following:

1) Perform channel-level predictions on all channels in a multi-channel segment.
2) With the set of probabilities outputs and knowledge of their location, we distribute probability outputs accordingly to seven regions: frontal, frontal-temporal, non-frontal (all non-frontal channels), central, parietal, occipital, and the entire scalp.
3) From each region, we extract statistical features: mean, median, standard deviation, maximum values, minimum values, and the histogram features (5 bins, range: [0,1]). This corresponds to 10 features per region.

This results in 70 features for each artifact type. Additionally, we include the cross-correlation and auto-correlation of the signals between channel FP1/FP2 and channel F7/F8 to account for eye blink features in each feature set. Eventually, we obtain 74 features from each multi-channel segment for the training and testing of each artifact class. We performed segment-level binary classification (specific artifact vs. background) for each artifact type. We classify the features with CatBoost [14], and optimize the hyperparameters by grid search.

Lastly, we concatenate the probability outputs from the five segment-level classifiers. With all the features, we trained a CatBoost classifier to detect any of the five artifacts types; in other words, this system is designed to determine whether a multi-channel EEG segment is clean or contains artifact(s). Finally, we evaluate the systems with the following metrics: area under the receiver operator characteristic (AUC), area under the precision-recall curve (AUPRC), accuracy (ACC), balanced accuracy (BAC), sensitivity (SEN), and specificity (SPE) [9]. NVIDIA GeForce GTX1080 GPU machines, Keras 2.2.0 and TensorFlow 2.6.0 were adopted in this study.

III. Results

The channel- and segment-level artifact detection results are displayed in Table II and III. We achieved the best BAC for the chewing artifacts (above 90%), while the BAC varies between 65% to 86% for the other artifacts. Moreover, the segment-level results reported improved performance over the channel-level results. This can be because the multi-channel segments contain more information than the single-channel segments due to lower annotation uncertainty.

The best BAC is achieved at different window lengths for each artifact type. Generally, we should deploy small window lengths to detect electrode pops and eye movement artifacts, and larger window lengths to detect chewing, muscle, and shiver artifacts. Finally, we report the SEN for the combined binary segment-level artifact detector in Table IV. At an SPE of 95%, 97%, and 99%, the highest SEN achieved is 42.0%, 32.0%, and 13.3%, at a window length of 3s, 3s, and 5s, respectively. For practical applications, one may choose a window length of 3s and a threshold (Th) where SPE is 95% to avoid rejecting too many clean EEG segments.

TABLE II
Results for channel-level artifact detection.

Artifact	L	AUC	AUPRC	ACC	BAC	SEN	SPE
Chewing	1	0.961	0.95	0.901	0.901	0.894	0.907
	3	0.966	0.904	0.933	0.911	0.856	0.966
	5	0.967	0.864	0.95	0.906	0.835	0.978
Electrode Pop	1	0.792	0.926	0.802	0.663	0.914	0.412
	3	0.802	0.802	0.709	0.716	0.718	0.713
	5	0.783	0.681	0.711	0.684	0.559	0.809
Eye Movement	1	0.856	0.921	0.795	0.734	0.905	0.564
	3	0.866	0.799	0.807	0.788	0.723	0.853
	5	0.895	0.758	0.877	0.792	0.644	0.94
Muscle	1	0.794	0.949	0.936	0.76	0.99	0.529
	3	0.931	0.973	0.907	0.836	0.961	0.711
	5	0.934	0.957	0.888	0.861	0.942	0.78
Shiver	1	0.657	0.138	0.527	0.527	0.091	0.994
	3	0.61	0.056	0.986	0.563	0.182	0.943
	5	0.756	0.066	0.994	0.621	0.364	0.878

IV. Discussion

In the following, we compare our results to the literature. Roy performed multi-class artifact classification and reported a SEN of 72.39% for the background class [7]. Abdi et al. deployed wavelet transform to detect and reject artifacts. They measured the effectiveness of their artifact detector indirectly by performing brain-computer interface (BCI) classification and achieved

TABLE III

Results for segment-level artifact detection.

Artifact	L	AUC	AUPRC	ACC	BAC	SEN	SPE
Chewing	1	0.968	0.719	0.971	0.910	0.842	0.977
	3	0.972	0.791	0.982	0.947	0.909	0.986
	5	0.892	0.701	0.986	0.921	0.851	0.991
Electrode Pop	1	0.792	0.306	0.803	0.731	0.632	0.830
	3	0.818	0.270	0.855	0.734	0.587	0.881
	5	0.818	0.223	0.875	0.735	0.577	0.894
Eye Movement	1	0.895	0.607	0.826	0.826	0.824	0.829
	3	0.887	0.440	0.874	0.810	0.731	0.888
	5	0.869	0.351	0.873	0.820	0.761	0.880
Muscle	1	0.896	0.739	0.808	0.840	0.934	0.746
	3	0.906	0.676	0.817	0.857	0.937	0.778
	5	0.901	0.583	0.810	0.857	0.931	0.783
Shiver	1	0.691	0.027	0.993	0.516	0.034	0.997
	3	0.770	0.068	0.994	0.530	0.062	0.998
	5	0.661	0.308	0.996	0.655	0.311	0.998

TABLE IV

Results for segment-level artifact detection , where all artifacts are grouped together.

L	AUC	AUPRC	SPE @95%		SPE @97%		SPE @99%	
			SEN	Th	SEN	Th	SEN	Th
1	0.876	0.866	0.401	0.821	0.306	0.855	0.127	0.916
3	0.870	0.796	0.420	0.812	0.320	0.871	0.114	0.927
5	0.873	0.732	0.367	0.827	0.269	0.853	0.133	0.899

an ACC improvement from 63% to 72.5% with the artifact rejection module [4]. Mashhadi et al. reject ocular artifacts by means of U-NET, and reported a mean square error (MSE) of 0.00712 [8]. Meanwhile, Dhindsa performed artifact classification on an EEG dataset with four channels and achieved an ACC of 93.3% and AUC of 0.923 [15]. Finally, Pion-Tonachini et al. deployed ICA and CNN/GAN to classify EEG independent components (IC), and achieved BAC of 0.855, 0.623, and 0.597, for 2, 5, and 7-class classification, respectively [3]. For all scenarios, they reported SEN of 73% for the background class.

Compared to these studies, our system performs better in terms of SPE, as our system reports a high SPE of 95% while achieving decent SEN of 42.0% for the artifact class, making it suitable for real-world application. In contrast, the studies by [7], [4], [8], [3] might be less suitable for real-world applications which require high SPE to avoid rejecting too much clean EEG. The majority of existing studies reported low SEN for the clean EEG (less than 75%), which is unacceptable as it can lead to a significant loss of valuable EEG information.

V. Conclusion

We have proposed a neural system for automated detection of five artifact classes: chewing, electrode pop, eye movement, muscle, and shiver artifacts. The channel-wise detector consists of a CNN followed by a transformer optimized via a BM loss. The outputs of the CNN-Transformer at multiple channels are then combined via another classifier for artifact detection in multi-channel EEG segments. The proposed system can reject a substantial fraction of artifacts while only removing a small fraction of clean EEG, thus potentially improving the readability of EEG recordings.

References

[1] C.-Y. Chang, S.-H. Hsu, L. Pion-Tonachini, T.-P. Jung, Evaluation of artifact subspace reconstruction for automatic eeg artifact removal, in: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2018, pp. 1242–1245.

[2] X. Jiang, G.-B. Bian, Z. Tian, Removal of artifacts from eeg signals: a review, Sensors 19 (5) (2019) 987.

[3] L. Pion-Tonachini, K. Kreutz-Delgado, S. Makeig, Iclabel: An automated electroencephalographic independent component classifier, dataset, and website, NeuroImage 198 (2019) 181–197.

[4] B. Abdi-Sargezeh, R. Foodeh, V. Shalchyan, M. R. Daliri, Eeg artifact rejection by extracting spatial and spatio-spectral common components, Journal of Neuroscience Methods 358 (2021) 109182.

[5] J. Thomas, P. Thangavel, W. Y. Peh, J. Jing, R. Yuvaraj, S. S. Cash, R. Chaudhari, S. Karia, R. Rathakrishnan, V. Saini, et al., Automated adult epilepsy diagnostic tool based on interictal scalp electroencephalogram characteristics: A six-center study, International Journal of Neural Systems (2021) 2050074.

[6] P. Thangavel, J. Thomas, W. Y. Peh, J. Jing, R. Yuvaraj, S. S. Cash, R. Chaudhari, S. Karia, R. Rathakrishnan, V. Saini, et al., Time–frequency decomposition of scalp electroencephalograms improves deep learning-based epilepsy diagnosis, International Journal of Neural Systems (2021) 2150032.

[7] S. Roy, Machine learning for removing eeg artifacts: Setting the benchmark, arXiv preprint arXiv:1903.07825.

[8] N. Mashhadi, A. Z. Khuzani, M. Heidari, D. Khaledyan, Deep learning denoising for eog artifacts removal from eeg signals, in: 2020 IEEE Global Humanitarian Technology Conference (GHTC), IEEE, 2020, pp. 1–6.

[9] W. Y. Peh, J. Thomas, E. Bagheri, R. Chaudhari, S. Karia, R. Rathakrishnan, V. Saini, N. Shah, R. Srivastava, Y.-L. Tan, et al., Multi-center validation study of automated classification of pathological slowing in adult scalp electroencephalograms via frequency features, International Journal of Neural Systems (2021) 2150016.

[10] A. Hamid, K. Gagliano, S. Rahman, N. Tulin, V. Tchiong, I. Obeid, J. Picone, The temple university artifact corpus: An annotated corpus of eeg artifacts, in: 2020 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), IEEE, 2020, pp. 1–4.

[11] T. Joo, U. Chung, M.-G. Seo, Being bayesian about categorical probability, in: International Conference on Machine Learning, PMLR, 2020, pp. 4950–4961.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.

[13] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.

[14] A. V. Dorogush, V. Ershov, A. Gulin, Catboost: gradient boosting with categorical features support, arXiv preprint arXiv:1810.11363.

[15] K. Dhindsa, Filter-bank artifact rejection: High performance real-time single-channel artifact detection for eeg, Biomedical Signal Processing and Control 38 (2017) 224–235.