Robust EEG-based Emotion Recognition Using an Inception and Two-sided Perturbation Model

Shadi Sartipi*
Department of Electrical and Computer Engineering
University of Rochester
Rochester, USA
ssartipi@ur.rochester.edu

Mujdat Cetin
Department of Electrical and Computer Engineering
Goergen Institute for Data Science
University of Rochester
Rochester, USA
mujdat.cetin@rochester.edu

Abstract—Automated emotion recognition using electroencephalogram (EEG) signals has gained substantial attention. Although deep learning approaches exhibit strong performance, they often suffer from vulnerabilities to various perturbations, like environmental noise and adversarial attacks. In this paper, we propose an Inception feature generator and two-sided perturbation (INC-TSP) approach to enhance emotion recognition in brain-computer interfaces. INC-TSP integrates the Inception module for EEG data analysis and employs two-sided perturbation (TSP) as a defensive mechanism against input perturbations. TSP introduces worst-case perturbations to the model's weights and inputs, reinforcing the model's elasticity against adversarial attacks. The proposed approach addresses the challenge of maintaining accurate emotion recognition in the presence of input uncertainties. We validate INC-TSP in a subject-independent three-class emotion recognition scenario, demonstrating robust performance.

Index Terms—Adversarial training, Brain-computer interfaces, Emotion recognition, Inception module, Weight perturbation.

I. INTRODUCTION

Brain-computer interfaces (BCI) develop a direct pathway between the electrical activities of the human brain and the external environment, bypassing the need for typical nerves and muscles [1]. This provides a multidisciplinary framework incorporating psychology, electronics, computers, and neuroscience to interpret the electrical activity of the brain and address health-related challenges to improve life quality [2]. Emotions profoundly influence human daily life by impacting physiological activities and decision-making. Automated recognition of emotions can improve human-machine communication [3]. Among various modalities for emotion recognition, electroencephalography (EEG) stands out as a preferred physiological method for collecting brain signals in BCI research due to its affordability, non-invasiveness, high temporal resolution, and portability [3]. Various studies have investigated automated emotion recognition by focusing on feature extraction and classification model construction [4]. Some of the widely used features in this context are differential entropy (DE), power spectral density (PSD), and functional connectivity [5].

Deep learning (DL) approaches have shown significant performance over traditional machine learning techniques for decoding EEG signals [6]. Convolutional neural networks (CNNs) [7] and recurrent neural networks (RNNs) [6] are widely used in this area. In [8], they used the combination of CNN and long short-term memory (LSTM) along with graph-based smoothed EEG data to get the temporal and spatial features for emotion recognition. Popular examples of using CNN-based networks include the DeepCNN and EEGNet proposed in [7] and [9], respectively.

For various BCI applications, most studies focus on improving the performance of DL approaches. However, one drawback of DL models is that they are often susceptible to precisely crafted minor alterations in input data, which could be due to environmental noise, individual variations, and adversarial attacks [10], [11]. These types of perturbations can lead to significant performance degradation [11]. Few studies explored this limitation. For instance, in [12] the susceptibility of machine learning algorithms in EEG-based BCIs is explored. A novel loss function is proposed in [13] to generate universal adversarial perturbations. Since emotion recognition is utilized in mental health applications and various real-time applications [14], the robustness of the model alongside the model's performance should be of concern. To the best of our knowledge, there is currently no study on the robustness of EEG-based emotion recognition.

In this paper, we propose a novel DL approach called Inception feature generator and two-sided perturbation (INC-TSP) to extract effective features from EEG data and learn the model in a way that is robust against adversarial attacks for subject-independent emotion recognition. To achieve this, we integrate the Inception module [15] with a CNN backbone into the deep architecture. As EEG signals contain oscillatory patterns of varying temporal lengths, the Inception module enables a multiscale analysis of the input data. Two-sided perturbation (TSP), which forms an outer maximization problem [16], serves as our defensive mechanism. TSP applies worst-case perturbations to the weights and inputs. We evaluate the proposed approach for three-class emotion recognition. Our main contributions can be summarized as follows:

- We develop an INC-based deep feature generator that performs multiscale analysis of spatial, temporal, and

Block	Details
C1	Conv (64; 5), Maxpool (4; 4), Dropout (0.3)
Inception	Conv (32, 1)
	Conv (96;1), Conv (128; 3)
	Conv (128;3), Conv (32; 5)
	Maxpool (3;1), Conv (32; 5)
C2	Conv (256; 5), Maxpool (4; 4), Dropout (0.3)

Algorithm 1: TSP

Input: Input data (x_i, y_i), mini-batch size $\hat{m}$, network f_θ, PGD iteration T, learning rate α, TSP step size η_2, constraint ϵ.

Output: Robust network f_θ

repeat

for $i = 1, \cdots, \hat{m}$ **do**

   $x'_i \leftarrow x_i + \delta$, where $\delta \sim \text{Uniform}(-\epsilon, +\epsilon)$

   if *attack==PGD* **then**

      for $t = 1, \cdots, T$ **do**

         | Update x'_i based on (2)

      end

   else

      if *attack==FGSM* **then**

         | Update x'_i based on (1)

      end

   end

   $v \leftarrow v + \eta_2 \nabla_v (\mathcal{L}(f_{\theta+v}(x'_i), y_i))$

   $v \leftarrow \gamma \frac{v}{\|v\|} \|\theta\|$

end

$\theta \leftarrow (\theta + v) - \alpha \nabla_{\theta+v} \frac{1}{m} \mathcal{L}(f_{\theta+v}(x'_i, y_i)) - v$

until training converge

---

spectral characteristics of EEG data.

- A novel learning approach is used for robust EEG-based emotion recognition against input perturbations.

II. METHOD

Inception-based Feature Generator: The process of feature extraction plays a pivotal role in EEG emotion recognition studies. To achieve this, we draw inspiration from image classification techniques and employ the Inception module [15] to extract features. The designed inception-based (INC) feature generator architecture comprises a combination of multiple convolutional layers, a pooling layer, and an activation function.

Let the input data be denoted as $X \in \mathbb{R}^{n \times c \times t}$, where n, c, and t are the number of frequency subbands, EEG channels, and temporal length of the data, respectively. Table I shows specific configurations and architectural choices for three different CNN blocks. The feature extraction process involves the initial application of a convolutional layer, C1, to acquire shallow features. Subsequently, the inception module,

with different convolutional kernels, is employed to capture deeper features. The various kernel sizes allow the model to capture diverse information in both the temporal and spatial domains. The output features from the parallel convolutional layers are then merged using a concatenation layer. The concatenated features pass through another convolutional layer, C2, followed by pooling and dropout layers in the final step. ReLU activation function and batch normalization are applied in CNN layers. Finally, the features are fed to three fully connected layers with dimensions of 512, 256, and 64.

Robust Generalization: The performance of deep neural networks for EEG-based emotion recognition is often vulnerable to perturbations applied to input data, which could significantly diminish the model's performance across different subjects [11]. These perturbations are also known as adversarial attacks [10].

To overcome such attacks, we employ adversarial training (AT) with adversarial weight perturbation as a defensive scheme inspired by [16]. This approach involves the use of two-sided perturbation (TSP), which applies worst-case perturbations to both the input data and the model's weights.

1) Adversarial Attack: To begin, we briefly explain two widely used adversarial attack techniques targeted at computer vision models. Fast gradient sign method (FGSM) [17] is a single-step gradient-based method to find the perturbed example in one step by the amount of ϵ in a direction specified by the sign of the gradient of the loss function:

$$x'_i = x_i + \epsilon.\text{sign}(\nabla_{x_i} \mathcal{L}(f_\theta(x'_i), y_i)) \tag{1}$$

Another attack is projected gradient descent (PGD) [18] which perturbs the original sample for T iterations with stepsize η. At each iteration, PGD projects the perturbed sample back onto the ϵ-ball at the tth iteration as:

$$x_i^{'(t+1)} = \Pi_\epsilon(x_i^{'(t)} + \eta.\text{sign}(\nabla_{x'_i} \mathcal{L}(f_\theta(x_i^{'(k)}), y_i))) \tag{2}$$

where Π is the projection operator.

2) Adversarial Defense: AT directly integrates adversarial examples into the training process to flatten the loss changes with respect to the input via the optimization problem presented below [18]:

$$\min_\theta \rho(\theta), \text{ where } \rho(\theta) = \frac{1}{n} \sum_{i=1}^{n} \max_{\delta \le \epsilon} \mathcal{L}(f_\theta(x'_i), y_i) \tag{3}$$

where $\rho(\theta)$ is the adversarial loss, n is the number of samples, $\delta = \|x'_i - x_i\|_p$, $\|.\|_p$ is the norm (threat model), x'_i is the adversarial example within the ϵ-ball centered at original sample x_i, f_θ is the deep learning architecture, i.e., INC, with weights θ, $\mathcal{L}$ is the classification loss, and y_i are the true labels. Following [16], to incorporate the flatness of the loss change with respect to the weight, we propose solving the following optimization problem:

$$\min_\theta \max_v \rho(\theta + v) = \min_\theta \max_v \frac{1}{n} \sum_{i=1}^{n} \max_{\delta \le \epsilon} \mathcal{L}(f_{\theta+v}(x'_i), y_i) \tag{4}$$

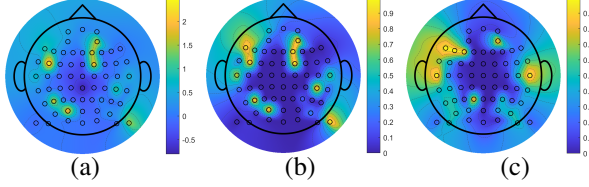|	PGD-10		PGD-20		FGSM	
Threat Model	R-Accuracy	R-F1-score	R-Accuracy	R-F1-score	R-Accuracy	R-F1-score
L_2	0.91 ± 0.04	0.89 ± 0.05	0.86 ± 0.07	0.85 ± 0.08	0.85 ± 0.05	0.84 ± 0.05
L_∞	0.82 ± 0.07	0.81 ± 0.07	0.81 ± 0.08	0.080 ± 0.08	0.84 ± 0.08	0.83 ± 0.07

Fig. 1. Sample EEG features for (a) original data, (b) perturbed with PGD-10, and (c) perturbed with PGD-20.

TABLE III
ABLATION STUDY FOR A PGD-10 ATTACK.

Defense	R-Accuracy	Accuracy
TSP	$\mathbf{0.91 \pm 0.04}$	$\mathbf{0.93 \pm 0.03}$
AT	0.81 ± 0.06	0.88 ± 0.04
Without defense	0.55 ± 0.08	0.86 ± 0.06

TABLE IV
COMPARISON WITH PREVIOUS WORKS.

Study	Accuracy
Proposed method	$\mathbf{0.93 \pm 0.03}$
SDDA [19]	0.91 ± 0.07
MSFR-GCN [20]	0.87 ± 0.05
SECT [21]	0.85 ± 0.06

where v is selected from the feasible region for weight perturbation. Let γ be the constraint on weight perturbation size. The weight perturbation on the lth layer with weight θ_l is $\|v_l\| \leq \gamma \|\theta_l\|$.

3) Learning process: The optimization and learning process for INC-TSP is shown in Algorithm 1. Let $\hat{m}$ be the mini-batch size. The input perturbation is applied by choosing the attack type subsequently the weight perturbation is calculated. Finally, the model parameters, θ, are updated via the Adam optimizer.

III. EXPERIMENTAL STUDY

Dataset: Here, we use the publicly available SEED dataset [5] for emotion recognition. The dataset comprises fifteen movie clips designed to elicit three emotions: happiness, sadness, and neutrality. The experiments involved a total of 15 participants, consisting of 8 females and 7 males, and were instructed to immerse themselves in the movie clips to evoke the corresponding emotions. The EEG signals were recorded using the international 10-20 system with 62 channels. Each trial followed a specific sequence: a 5-second starting hint before the film clip, 4 minutes of the clip as an emotional stimulus, 45 seconds for self-assessment, and a 15-second break. The recorded EEG data were downsampled from 1000 Hz to 200 Hz, and a band-pass filter with a frequency range of 0.5-70 Hz was applied. We calculated the differential entropy (DE) features every 1-second with no overlap in five frequency subbands, namely, delta, theta, alpha, beta, and gamma.

Results: In this section, we present the performance of the proposed approach. We employed the L_2 and L_∞ threat models. The value of ϵ is set to $8/225$, and the PGD step size is set to $15/255$. We evaluated the approach against various

adversarial attacks. Samples of the perturbed input with 10 and 20 iterations, PGD-10 and PGD-20, are depicted in Fig. 1. The weight perturbation size and learning rate are both set to 0.03 and 9×10^{-5}, respectively.

We employed leave-one-subject-out (LOSO) cross validation to assess performance. The input data are normalized by subtracting the mean and dividing by the standard deviation. To evaluate the performance and show the effectiveness of the defense mechanism, we computed the robust accuracy and F1-score (R-Accuracy and R-F1-score). R-Accuracy and R-F1-score present the accuracy and F1-score on the perturbed test data. The average performance across subjects is presented in Table II. As shown in Table II, the best robust performance is 0.91 ± 0.04 under the L_2 threat model and PGD-10 adversarial attacks. Furthermore, these results suggest that INC-TSP consistently maintains R-Accuracy and R-F1-scores, indicating its effectiveness in countering adversarial perturbations and ensuring accurate classification.

Furthermore, we performed the ablation study shown in Table III. R-Accuracy and Accuracy columns are the performance of each approach on test data with PGD-10 attack and non-perturbed data, respectively. Comparing the results without defense to those with AT (3) and TSP (4), the proposed learning process achieves the highest R-Accuracy and accuracy. For detailed performance analysis, we present the confusion matrices for classification using the INC architecture without a defense and with the proposed INC-TSP approach with PGD-20 attack in Fig. 3. This demonstrates TSP's ability to accurately detect all three emotional states even when the model encounters perturbed inputs. Table IV compares INC-TSP's performance to previous works, demonstrating INC-TSP's superiority. All methods considered here use a LOSO cross-validation process.

Effect of TSP on robustness and generalization: To investigate the generalization of the proposed learning approach, we present the learning curves in the training and test phases with and without TSP (Fig. 2 (a) and (b)). The major drawback of DL approaches in BCI is the overfitting of the model on the training data and poor performance on new test subjects. As presented in Fig. 2 (b), the learning curves of INC-TSP for
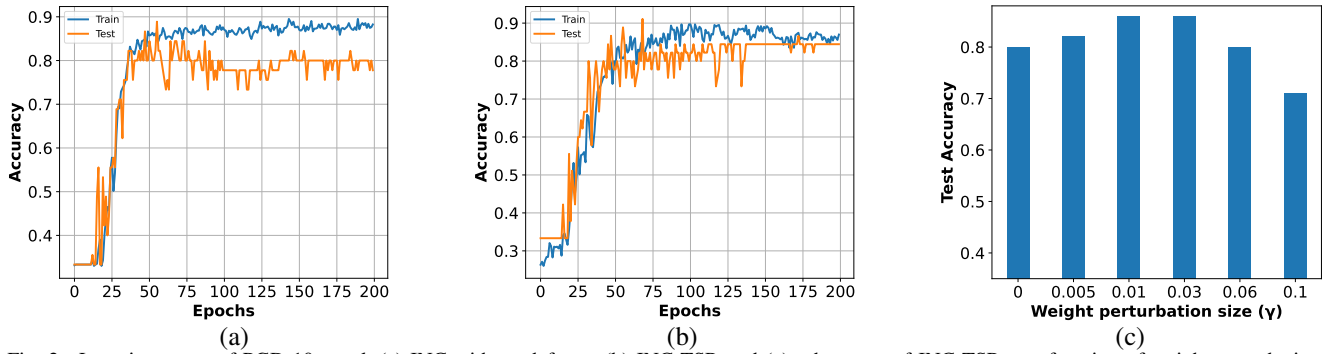
Fig. 2. Learning curve of PGD-10 attack (a) INC without defense, (b) INC-TSP, and (c) robustness of INC-TSP as a function of weight perturbation size.

Fig. 3. Confusion matrices for a perturbed input with PGD (**left**) without defense and (**right**) with TSP.

test and training follow each other closely which shows the generalization and robustness of the learned model. Additionally, Fig. 2 (c) presents the model performance across different weight perturbation sizes. These results show that while the model gets the best performance with $\gamma = 0.01$ and 0.03, the performance does not drop significantly even with large perturbations.

IV. CONCLUSION

This study proposed the INC-TSP model as a novel approach to enhance the robustness of emotion recognition in BCIs. By combining the Inception module and TSP as a defensive mechanism, the model demonstrates durability against adversarial attacks and input uncertainties. Results indicate that the INC-TSP model consistently achieves robust accuracy and F1-scores across various threat models and adversarial attack scenarios which shows its efficacy in countering perturbations. The robustness and generalization investigations suggest the use of INC-TSP as a possible defense mechanism in future BCI implementations.

REFERENCES

[1] Jerry J Shih, Dean J Krusienski, and Jonathan R Wolpaw, "Brain-computer interfaces in medicine," *Mayo clinic proceedings*, vol. 87, no. 3, pp. 268–279, 2012.

[2] Sarah N Abdulkader, Ayman Atia, and Mostafa-Sami M Mostafa, "Brain computer interfacing: Applications and challenges," *Egyptian Informatics Journal*, vol. 16, no. 2, pp. 213–230, 2015.

[3] Raymond J Dolan, "Emotion, cognition, and behavior," *science*, vol. 298, no. 5596, pp. 1191–1194, 2002.

[4] Robert Jenke, Angelika Peer, and Martin Buss, "Feature extraction and selection for emotion recognition from EEG," *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 327–339, 2014.

[5] Wei-Long Zheng and Bao-Liang Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Transactions on autonomous mental development*, vol. 7, no. 3, pp. 162–175, 2015.

[6] Alexander Craik, Yongtian He, and Jose L Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: a review," *J. Neural Eng.*, vol. 16, no. 3, pp. 031001, 2019.

[7] Robin Tibor Schirrmeister and et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human brain mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[8] Shadi Sartipi, Mastaneh Torkamani-Azar, and Mujdat Cetin, "A hybrid end-to-end spatio-temporal attention neural network with graph-smooth signals for EEG emotion recognition," *IEEE Trans. Cogn. Develop. Syst.*, 2023.

[9] Vernon J Lawhern and et al., "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, pp. 056013, 2018.

[10] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[11] Xiao Zhang and Dongrui Wu, "On the vulnerability of CNN classifiers in EEG-based BCIs," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 5, pp. 814–825, 2019.

[12] Lubin Meng and et al., "White-box target attack for EEG-based BCI regression problems," in *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part I 26*. Springer, 2019, pp. 476–488.

[13] Zihan Liu, Lubin Meng, Xiao Zhang, Weili Fang, and Dongrui Wu, "Universal adversarial perturbations for CNN classifiers in EEG-based BCIs," *J. Neural Eng.*, vol. 18, no. 4, pp. 0460a4, 2021.

[14] Adnan Qayyum, Junaid Qadir, Muhammad Bilal, and Ala Al-Fuqaha, "Secure and robust machine learning for healthcare: A survey," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 156–180, 2020.

[15] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017.

[16] Dongxian Wu, Shu-Tao Xia, and Yisen Wang, "Adversarial weight perturbation helps robust generalization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2958–2969, 2020.

[17] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[19] Zhunan Li and et al., "Dynamic domain adaptation for class-aware cross-subject and cross-session EEG emotion recognition," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 12, pp. 5964–5973, 2022.

[20] Deng Pan, Haohao Zheng, Feifan Xu, Yu Ouyang, Zhe Jia, Chu Wang, and Hong Zeng, "Msfr-gcn: A multi-scale feature reconstruction graph convolutional network for EEG emotion and cognition recognition," *IEEE Trans. Neural Syst. Rehabil. Eng.*, 2023.

[21] Zhongli Bai and et al., "Sect: A method of shifted EEG channel transformer for emotion recognition," *IEEE J. Biomed. Health Inform.*, 2023.