# DA-NN: A ResNet-Based Architecture with Domain Adaptation for Automatic EEG Artifact Identification

Emmanuel de Jesús Velásquez-Martínez[1][a] Miguel Ángel Porta-García[1][b]

[1]*Centro de Investigación e Innovación en Tecnologías de la Información y Comunicación (INFOTEC), Ciudad de México 14050, México*

*iemmanuelvm@gmail.com, miguel.porta@infotec.mx*

Keywords: Artifact Identification, Domain Adaptation, Semi-Supervised Deep Learning.

Abstract: Electroencephalography (EEG) is a fundamental tool for studying and understanding brain activity. However, EEG signals are often contaminated by physiological and non-physiological artifacts that distort the original signal, which can affect interpretation and diagnostic accuracy. Therefore, accurate classification of these artifacts during preprocessing is essential to ensure more reliable EEG data analysis. In recent years, this problem has been addressed using both traditional methods and deep learning approaches. In the latter case, many models have been trained primarily with synthetic data to estimate clean signals, while other state-of-the-art studies use real, expert-labeled datasets commonly applied for classification tasks. However, when these models are evaluated using data that differs even slightly from the training set, their performance often declines due to a lack of domain adaptation, revealing limited generalization ability. To overcome this limitation, this study proposes an architecture called the Domain Adaptation Neural Network (DA-NN), designed to generalize the identification of EEG artifacts in both synthetic and real signals. The model is based on semi-supervised learning with domain adaptation to reduce the gap between the two types of data. Experiments demonstrated robust performance, achieving 91.59% accuracy in artifact classification.

## 1 INTRODUCTION

Electroencephalography (EEG) is a widely used tool in clinical settings and neuroscience research, as it allows for the direct, non-invasive measurement of brain activity with high temporal resolution (Huang et al., 2024). This technique plays a fundamental role in the diagnosis, monitoring, and prognosis of various neurological disorders (Saba-Sadiya et al., 2021), as well as being a key element in the development of brain-computer interfaces (BCI) (Chuang et al., 2022). EEG is recorded using electrodes placed on the scalp, which capture the electrical activity of the brain in different cortical regions. The international 10–20 system is used for measurement, ensuring coverage of the frontal, occipital, temporal, and parietal areas (Jackson and Bolger, 2014). EEG corresponds to a sequence of electrical voltages generated by the synchronous activity of neurons (Ahmed, 2022); however, due to its low amplitude, it is highly susceptible to external interference. These interferences, known as artifacts, are unwanted signals of non-cerebral origin that can significantly distort the waveform. Depending on their origin, artifacts are classified as physiological (originating from the subject's own body) or non-physiological (coming from external sources) (Quintero-Rincón et al., 2021). Their presence can compromise data analysis and affect diagnostic accuracy(Cai et al., 2025).

Figure 1 illustrates different types of artifacts obtained from the Temple University Hospital EEG dataset (TUH-EEG) (Hamid et al., 2020). For reference, Figure 1 (a) shows a clean EEG signal, while Figures (b)–(f) present common interferences such as artifacts from eye movement, muscle activity, shivering, chewing, and electrode pop. These distortions are mainly reflected as abrupt changes in amplitude and frequency, which makes it difficult to analyze brain signals.

In recent literature, the methods proposed for treating artifacts in EEG are mainly grouped into two approaches: classification and elimination. The first group, corresponding to classification methods, aims to identify the type of interference present in the signal in order to subsequently

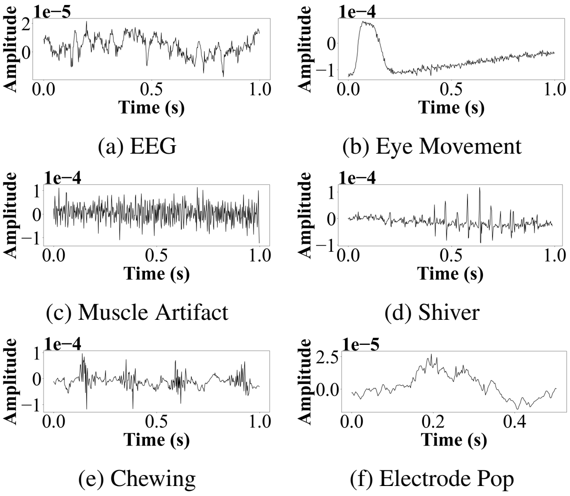
[a] https://orcid.org/0009-0008-7500-3579

[b] https://orcid.org/0000-0002-7524-5833

apply specific correction strategies. Among the most representative works are (Kim and Keene, 2019), (Bahador et al., 2020b), (Brophy et al., 2022), and (Maiwald et al., 2023), which employ architectures based on convolutional neural networks (CNNs), long short-term memory (LSTM) networks, and Transformers, as well as hybrid approaches combining these techniques. In contrast, artifact removal methods seek to directly suppress interference while preserving, as far as possible, the underlying neural activity. In this area, recent proposals include (Mashhadi et al., 2020), (Chuang et al., 2022), (Huang et al., 2024), (Bao et al., 2022), (Xiong et al., 2024), (Choi, 2025), (Gabardi et al., 2023), and (Azhar et al., 2024). These studies explore the use of CNNs, autoencoders, U-Net architectures, Transformers, diffusion models, and hybrid approaches, achieving significant improvements in metrics such as relative root mean square error (RRMSE), correlation coefficient (CC), and signal-to-noise ratio (SNR).

However, deep-learning–based models applied to this preprocessing stage typically adopt a one-size-fits-all approach, assuming that the incidence and nature of artifacts are uniform across subjects (Saba-Sadiya et al., 2021). This assumption constrains their generalization capability, given that EEG signals exhibit high inter- and intra-individual variability (Maiwald et al., 2023). As a result, the effectiveness of many methods decreases when applied to domains whose distributions differ from those present in the training dataset. Furthermore, several studies (Choi, 2025), (Huang et al., 2024) emphasize the importance of evaluating models beyond their original configurations in order to achieve an appropriate balance between artifact removal and the preservation of relevant neural information. In this context, deep learning techniques combined with domain-adaptation strategies have shown significant potential in addressing this challenge.

In this work, we propose a deep-learning model, Domain Adaptation Neural Network (DA-NN), that incorporates semi-supervised domain-adaptation mechanisms to enhance generalization in EEG artifact-classification tasks. Unlike traditional approaches, our method leverages a combination of real and synthetic data to capture a broader diversity of artifact-related patterns and conditions. In the domain-adaptation classification tasks, only the labels from the source domain are used, while the unlabeled data from the target domain are projected into a shared latent space. Subsequently, a Sinkhorn loss function is applied in this latent space to align the distributions of both domains, reducing their discrepancy and improving model robustness without requiring extensive fine-tuning for each new dataset (Geuter et al., 2025), (Pandya et al., 2025).



(a) EEG (b) Eye Movement

(c) Muscle Artifact (d) Shiver

(e) Chewing (f) Electrode Pop

Figure 1: Class examples for EEG artifact identification.

Finally, the structure of this paper is as follows: Section 2 presents the related works; Section 3 describes the materials and methods; Section 4 discusses Sinkhorn-based domain alignment process applied to deep learning models; Section 5 details the experimental results; and Section 6 provides the conclusions.

## 2 RELATED WORKS

This section summarizes relevant studies that have proposed classification strategies for EEG artifact detection. In particular, prior works have explored a variety of machine learning techniques for identifying artifacts in EEG signals. Among these approaches, several investigations have relied on the TUH-EEG Corpus (Hamid et al., 2020), which has served as a benchmark dataset for training and evaluating artifact-detection models, in addition to other studies that employ alternative datasets for artifact-classification tasks. In (Peh et al., 2022), the TUH-EEG Corpus was utilized, which contains EEG recordings annotated with chewing artifacts, eye movements, muscle activity, shivering, and electrode pop. The authors proposed a convolutional neural network architecture combined with Transformer layers and implemented five individual detectors for each channel and signal segment. Using statistical features as input, their approach achieved a balanced accuracy (BAC) of 0.804. Subsequently, (Kim and Keene, 2019) developed a hybrid model integrating LSTM and CNN networks to classify five types of artifacts from one-second EEG segments, achieving an accuracy of 67.59% and a true-positive rate of 80% with a false-positive rate of 25.8% in binary classification. The model reported in (Kim and Keene, 2019) was lightweight and sufficiently fast to be deployed on portable devices such as a Raspberry Pi, requiring approximately 2 ms to predict a one-second EEG segment. Furthermore, (Maiwald et al., 2023) explored an alternative representation of EEG signals by converting them into temporal images. Several transformation techniques were employed, including Gramian angular fields, recurrence plots, continuous wavelet transform, spectrograms, correlograms, and Markov chains. The best-performing models were Xception and EfficientNetB0 using Markov-based features, achieving F1-scores of 90.5% and 89.4% and accuracies of 89% and 89.9%, respectively. Models based on spectrograms exhibited moderate bias. Beyond studies employing the TUH-EEG Corpus, other approaches have explored alternative datasets to improve artifact classification through different signal representations and neural-network architectures. In this context, (Bahador et al., 2020a) proposed a CNN model that maps multichannel EEG signals into a two-dimensional RGB space, thereby capturing inter-channel correlations. The dataset used was provided by Oulu University Hospital. Using correlation coefficients as input features, the model achieved an accuracy of 92.30% and an area under the ROC curve (AUC) of 0.96. Additionally, (Saba-Sadiya et al., 2021) introduced an unsupervised model for EEG artifact detection and correction based on an encoder–decoder architecture. For training, 58 features related to brain complexity, continuity, and connectivity were extracted. The model substantially outperformed a random baseline classifier (Cohen's kappa = 0.029), achieving a kappa coefficient of approximately 0.49.

# 3 MATERIALS AND METHODS

This section presents the proposed methodology to address the adaptation problem in the DA-NN model. In particular, a domain adaptation scheme is developed between real EEG artifacts and synthetically generated artifacts, with the aim of improving performance in EEG artifact-classification tasks. The experimental procedure adopted in this study is structured into methodological modules that constitute a sequential pipeline for data processing and analysis, as illustrated in Figure 2.

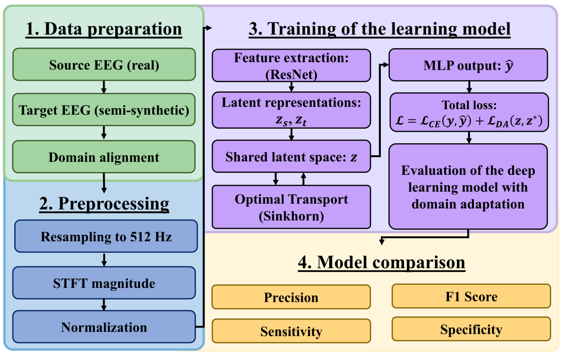

1. Data preparation
Source EEG (real)
Target EEG (semi-synthetic)
Domain alignment
2. Preprocessing
Resampling to 512 Hz
STFT magnitude
Normalization
3. Training of the learning model
Feature extraction: (ResNet)
Latent representations: $z_s, z_t$
Shared latent space: $z$
Optimal Transport (Sinkhorn)
MLP output: $\hat{y}$
Total loss: $\mathcal{L} = \mathcal{L}_{CE}(y, \hat{y}) + \mathcal{L}_{DA}(z, z^*)$
Evaluation of the deep learning model with domain adaptation
4. Model comparison
Precision
F1 Score
Sensitivity
Specificity

Figure 2: Proposed methodology for EEG artifact identification and removal.

## 3.1 Data Preparation

Block 1 in Figure 2 corresponds to the data preparation stage. In this phase, real EEG recordings containing artifacts from the TUH-EEG dataset are used as the source domain. This dataset primarily includes muscle artifacts (51,052 events), ocular movements (38,569 events), chewing artifacts (6,482 events), baseline electrical potentials (172 events), and shivering (613 events). These events exhibit different temporal durations, enabling the capture of variability associated with each type of artifact in contaminated EEG signals.

Additionally, synthetic data from the EEGdenoiseNet dataset (Zhang et al., 2021) are used as the target domain in order to increase model variability. EEGdenoiseNet consists of 4,514 clean EEG segments, along with 3,400 segments containing ocular artifacts and 5,598 segments containing muscle artifacts. This structure enables the creation of controlled SNR conditions and provides a standardized environment for model comparison.

Since EEGdenoiseNet only includes muscle and ocular artifacts, longer-duration events from the source domain were selected to synthesize the missing artifact types. These events were superimposed to avoid mixing information across domains and to preserve the integrity of the process. Likewise, for muscle and ocular artifacts from the source domain, a number of samples equivalent to those available in EEGdenoiseNet was selected in order to balance the classes across domains.

Synthetic data generation was performed using artifact-free EEG signals from EEGdenoiseNet as baseline signals. Subsequently, segments containing artifacts from both EEGdenoiseNet and TUH-EEG were randomly selected and incorporated into the clean signals through a controlled noise-addition algorithm (see Algorithm 1). First, the algorithm estimates the power of the clean signal based on the average of the squared samples. Using this value and the desired SNR, the target power that the artifact must exhibit to maintain the specified degradation level is computed. The artifact is then normalized to zero mean and unit variance. Once normalized, the artifact is amplified according to the calculated target power, producing a noise signal with the appropriate magnitude. The contaminated signal is generated via linear summation of the clean signal and the scaled artifact. This procedure allows control over the interference level of each incorporated artifact, with an SNR range between -5 dB and 5 dB, facilitating the creation of a variable dataset for model training.

Finally, both datasets were split into 80% for training and 20% for testing, ensuring an appropriate validation process.

```
Data: $x(n)$: clean EEG signal, $a(n)$: artifact
    signal, SNR: desired SNR (dB)
Result: $y(n)$: artifact-contaminated EEG
    signal
Estimate clean EEG power: $\mathcal{P}_x = \frac{1}{N}\sum_{n=1}^{N} x(n)^2$
Compute target artifact power from SNR:
$\mathcal{P}_a = \frac{\mathcal{P}_x}{10^{\mathrm{SNR}/10}}$
Normalize artifact to zero-mean and unit
variance: $\tilde{a}(n) = \frac{a(n) - \mu_a}{\sigma_a}$
Scale artifact to desired power:
$a_s(n) = \sqrt{\mathcal{P}_a} \cdot \tilde{a}(n)$
Generate synthetic contaminated EEG:
$y(n) = x(n) + a_s(n)$
```

Algorithm 1: Synthetic addition of artifacts into EEG signals

## 3.2 Preprocessing

Block 2 in Figure 2 consists of standardizing the sampling frequency to 512 Hz for the TUH-EEG and EEGdenoiseNet datasets. Subsequently, input features are extracted for deep learning models with domain adaptation. For the classification task, the Short-Time Fourier Transform (STFT) is applied as shown in equation 1 in order to extract frequency information from the EEG signals (Goyal and Pabla, 2015), where $x(\cdot)$ is the input signal, $w(k)$ is the analysis window, $H$ is the hop size, $K$ is the FFT size, $m$ is the time frame index, $k$ is the sample index, and $\omega$ is the frequency bin index. The STFT provides a time-frequency representation that facilitates the identification of patterns associated with different types of artifacts. From this transform, the magnitude spectrum is calculated, which is used as the basis for training the classification model. Additionally, spectrum-based features were used since (Maiwald et al., 2023) mentions that this type of feature has less performance bias. The configuration includes a window of 512 samples, a shift of 60 samples between consecutive windows, an analysis length of 128 samples, and the use of a Hamming window to smooth the signal. Finally, a normalization process is applied to both types of features: those used for classification and those used for artifact removal, ensuring a consistent and stable scale during model training.

$$X(\omega, m) = \sum_{k=0}^{K-1} x(mH + k)\, w(k)\, e^{-j2\pi\omega k/K} \qquad (1)$$

## 3.3 Training of the learning model

Block 3 in Figure 2 corresponds to the training stage of the proposed DA-NN model, which includes neural network feature extraction, domain alignment, and supervised optimization of predictions. First, a ResNet architecture is employed to perform feature extraction, obtaining latent representations from both the source and target domains, denoted as $z_s$ and $z_t$, respectively. Subsequently, these representations are projected into a shared latent space. To ensure alignment between both latent distributions, an Optimal Transport–based scheme is applied, specifically the Sinkhorn algorithm, whose objective is to minimize the divergence between distributions and obtain a joint representation. Once the shared representation $z$ is obtained, a multilayer neural network is used to generate the model output, denoted as $\hat{y}$. The training process optimizes a loss function composed of two terms: a classification loss $\mathcal{L}_{CE}(y, \hat{y})$, based on cross-entropy, and an adaptation loss $\mathcal{L}_{DA}(z, z^*)$, which penalizes discrepancies between the latent spaces of the source and target domains. To monitor the performance of the model throughout training, only the accuracy metric was employed, evaluating the network's ability to correctly classify samples during different training epochs. Additionally, to assess the effectiveness of the domain adaptation process, three metrics are considered: the Kullback–Leibler (KL) divergence (Shlens, 2014), the Jensen–Shannon (JS) divergence (Fuglede and Topsoe, 2004), and the Jensen–Shannon distance (Zunino et al., 2022). These metrics quantify the degree of alignment between the distributions of the source and target domains, verifying that the model successfully learns to reduce the domain gap during training.

## 3.4 Model comparison

Block 4 of Figure 2 evaluates the performance of the DA-NN model in EEG artifact classification, which predicts categorical outputs. Commonly used metrics described by (Grandini et al., 2020), (Tohka and van Gils, 2021), and (Géron, 2019) include precision, sensitivity, specificity, and F1-score, which together provide a comprehensive assessment of classification performance.

# 4 Sinkhorn-Based Domain Alignment Process Applied to Deep Learning Models

Domain adaptation aims to improve the performance of a target model in scenarios where annotated data are scarce or unavailable, by leveraging knowledge from a related source domain that contains sufficient labeled samples (Farahani et al., 2021), (Courty et al., 2017a). The algorithm presented in Algorithm 2 builds upon prior research in related areas, including (Courty et al., 2017b), (Lin et al., 2021), and (Feydy et al., 2019). It is designed to train deep neural networks capable of extracting meaningful latent representations, with the goal of incorporating adaptation mechanisms that enhance the robustness of classification task under distributional shifts. Domain adaptation encompasses a set of techniques aimed at aligning the latent distributions learned by neural networks when covariate shift exists between training and deployment data (Courty et al., 2017a). To achieve this alignment, feature representations from both the source and target domains are utilized, while labeled samples are only required from the source domain. The training paradigm follows a semi-supervised scheme in which source-domain labels guide supervised learning—typically via loss functions such as cross-entropy—while adaptation is performed by extracting latent vectors from the network and minimizing the Sinkhorn divergence (Feydy et al., 2019). The Sinkhorn algorithm is an iterative method used to solve regularized optimal transport problems. Its objective is to obtain a transport matrix that maps one distribution into another while minimizing a given cost and satisfying normalization constraints (Geuter et al., 2025). The latent distributions refer to probability distributions over these latent feature vectors, and the adaptation procedure seeks to minimize a statistical distance between them, thereby reducing domain discrepancy during training.

```
Data: $x_s$: Source domain inputs, $y_s$: Source
      domain labels, $x_t$: Target domain
      inputs, $n$: Batch size per domain, $\lambda_{DA}$:
      Weight of domain loss
Result: Model adapted to the target domain
while not converged do
    $X \leftarrow [x_s, x_t]$;
    $Y, Z \leftarrow \text{model}(X)$;
    $y_s \leftarrow Y$;
    $z_s, z_t \leftarrow Z$;
    $\mathcal{L}_{DA} \leftarrow \text{Sinkhorn}(z_s, z_t, \|z_{si} - z^*_{tj}\|_2, \varepsilon)$;
    $\mathcal{L}_{task} \leftarrow \begin{cases} \mathcal{L}_{CE}(y_s, \hat{y}_s), & \text{classification} \end{cases}$;
    $\mathcal{L} \leftarrow \mathcal{L}_{task} + \lambda_{DA} \cdot \mathcal{L}_{DA}$;
    loss.backward();
    optimizer;
end
```

Algorithm 2: Domain adaptation for classification tasks using Sinkhorn loss in the DA-NN model (Feydy et al., 2019)

The process of Algorithm 2 (see Figure 3) begins with the alignment of the feature matrix of the source domain $x_s$ and the target domain $x_t$, jointly denoted as $X = x_s, x_t$. Next, the DA-NN model generates predictions for each domain: $y_s, y_t \leftarrow Y$. In the section corresponding to the fully connected layers of the model, an intermediate layer is selected from which the latent representations are extracted: $z_s, z_t \leftarrow Z$, corresponding to both domains. The goal of this step is to align the representations of the source and target domains. Using these feature vectors, the Euclidean distance between them is computed, which is then used in the Sinkhorn function to calculate the minimal cost between the latent distributions of both domains. The purpose of this function is to align the source and target domain distributions during training, so that the model is able to generalize correctly in both domains (Feydy et al., 2019). Once these calculations are performed, the appropriate loss function is selected according to the nature of the task. In the next step, both the domain adaptation loss and the neural network prediction loss are added together. Finally, this total loss is used to compute the gradients and update the model parameters during the training process.

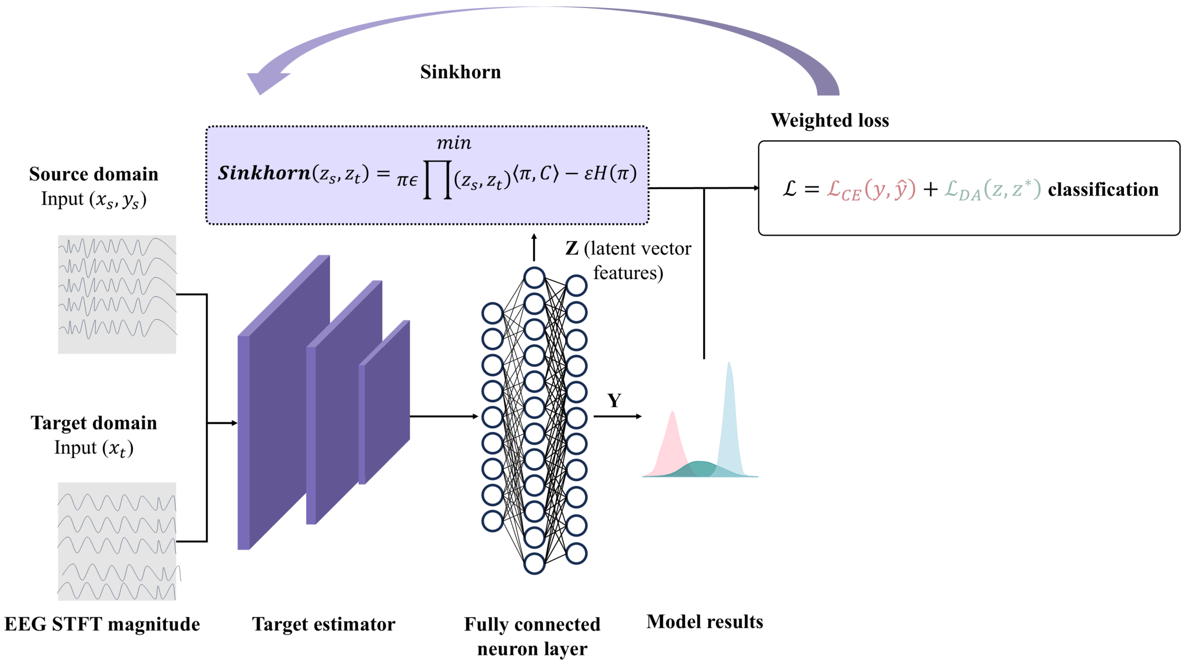

Sinkhorn
Weighted loss
$\textbf{Sinkhorn}(z_s, z_t) = \underset{\pi\epsilon \prod(z_s,z_t)}{min} \langle \pi, C\rangle - \varepsilon H(\pi)$
$\mathcal{L} = \mathcal{L}_{CE}(y, \hat{y}) + \mathcal{L}_{DA}(z, z^*)$ classification
Source domain
Input $(x_s, y_s)$
Target domain
Input $(x_t)$
Z (latent vector features)
Y
EEG STFT magnitude
Target estimator
Fully connected neuron layer
Model results

Figure 3: Neural network architecture with domain adaptation (DA-NN).

# 5 EXPERIMENTATION

This section presents the experimental results obtained for the DA-NN model applied to the artifact classification problem under the domain adaptation scenario.

## 5.1 Neural network architecture

For the EEG artifact classification task, the proposed DA-NN model is based on a two-dimensional residual network (2D-ResNet) with a 2-2-2-2 configuration, designed for classification tasks on time-frequency representations (see Table 1). The choice of this ResNet architecture was motivated by the study conducted by (Maiwald et al., 2023). The network receives single-channel inputs of size $1 \times 65 \times 9$. The DA-NN architecture begins with a $7 \times 7$ convolutional layer with 64 filters, a stride of 2, batch normalization, and ReLU activation, followed by a $3 \times 3$ max-pooling layer. Subsequently, four residual stages are implemented, each consisting of two basic residual blocks. The first stage maintains the channel dimensionality ($64 \rightarrow 64$), whereas the following stages increase the number of filters to 128, 256, and 512, respectively. Spatial resolution is reduced at the beginning of each stage by applying a stride of 2. After the final residual block, global average pooling is applied before a fully connected layer that produces the logits for the six target classes. The architecture contains a total of 11.18 M trainable parameters.

Table 1: Network architecture of the proposed DA-NN (2D-ResNet configuration 2-2-2-2).

| Stage | Description | Output Shape | Params |
|---|---|---|---|
| Input | – | $(B,1,65,9)$ | – |
| Conv1 | $7 \times 7$ conv, 64 ch, stride 2, pad 3; BN, ReLU | $(B,64,112,112)$ | 3,328 |
| MaxPool | $3 \times 3$ max pool, stride 2, pad 1 | $(B,64,56,56)$ | 0 |
| Layer 0 | 2× ResidualBlock ($64 \rightarrow 64$) | $(B,64,56,56)$ | 148,224 |
| Layer 1 | ($64 \rightarrow 128$) stride 2 + downsample; ($128 \rightarrow 128$) | $(B,128,28,28)$ | 526,208 |
| Layer 2 | ($128 \rightarrow 256$) stride 2 + downsample; ($256 \rightarrow 256$) | $(B,256,14,14)$ | 2,100,992 |
| Layer 3 | ($256 \rightarrow 512$) stride 2 + downsample; ($512 \rightarrow 512$) | $(B,512,7,7)$ | 8,396,288 |
| Global AvgPool | Adaptive average pooling | $(B,512,1,1)$ | 0 |
| Flatten | – | $(B,512)$ | 0 |
| Fully Connected | Linear ($512 \rightarrow 6$) | $(B,6)$ | 3,078 |
| **Total params** | | | **11,178,118** |
| **Trainable params** | | | **11,178,118** |

The training procedure was formulated under a domain adaptation scheme (as described in Algorithm 2 and Figure 3), using labeled data from the source domain and both labeled and unlabeled samples from the target domain. The DA-NN model was trained for 100 epochs using the Adam optimizer with a learning rate of $1 \times 10^{-3}$ and a weight decay factor of $1 \times 10^{-10}$. The main optimization objective was to minimize the cross-entropy loss on the source-domain data. Additionally, a regularization term based on Sinkhorn optimal transport was incorporated. To improve numerical stability, gradient clipping with a threshold of 1.0 was applied.

## 5.2 Learning behavior of the neural network

Figure 4 presents the accuracy metric behavior of the DA-NN model, applied to the artifact classification task under the domain adaptation approach. The accuracy curve corresponding to the source domain during training, represented with blue markers, reaches an average value close to 97.96%, indicating that the model effectively learns from these data. In contrast, the curve associated with the target domain during training, shown in orange, achieves an average accuracy of approximately 92.17%. For the source domain test data, represented by the green line, the curve stabilizes at an average of 89.34%, while for the target domain test set, represented by the red line, an average accuracy of 91.04% is obtained.

On the other hand, the markers below 70% correspond to the results in the target domain without applying the domain adaptation algorithm (see Algorithm 2) to the learning model. This behavior demonstrates that when the model performs inference on a dataset different from the one used during training and no domain adaptation is applied, its performance tends to decrease. This effect is associated with the lack of variability in the training data, resulting in an average accuracy of 67%.

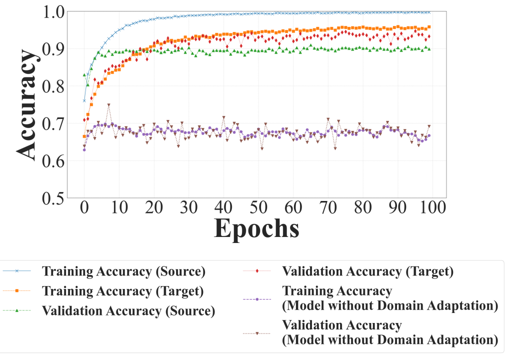


Figure 4: Training and validation accuracy over epochs.

Another metric used to monitor the proper functioning of the DA-NN domain adaptation mechanism corresponds to three divergence measures (see Figure 5). In particular, the Jensen–Shannon divergence (green line), its scaled version (orange line), and the Kullback–Leibler divergence (blue line) exhibit a progressive reduction over the epochs. This behavior indicates that the distribution of latent representations between the source and target domains becomes increasingly similar, demonstrating that the model is capable of classifying artifacts in both real and synthetic EEG data.

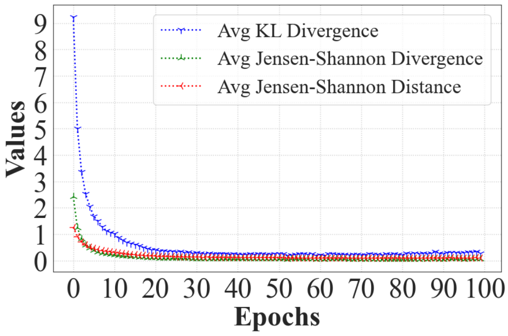


Figure 5: Domain adaptation distance and divergence metrics over epochs.

## 5.3 Results of artifact classification in EEG on the validation set

The performance of the DA-NN model was assessed using a normalized confusion matrix (Figure 6), which summarizes the proportion of correctly and incorrectly classified samples for each signal category. The model was designed to distinguish between five artifact types — Chewing, Electrode Pop, Eye Movement, Muscle, and Shiver — as well as clean EEG activity. The results demonstrate strong classification performance across all classes, with 98% accuracy achieved for EEG segments. Among the artifact classes, Eye Movement and Muscle artifacts were correctly identified with accuracies of 92% and 91%, respectively. Chewing and Shiver artifacts achieved 88% accuracy, while Electrode Pop reached 82%. Although overall performance is high, the confusion matrix reveals misclassification patterns in the off-diagonal entries. The most notable confusions occurred between Chewing and Muscle artifacts and between Electrode Pop and Eye Movement. These errors align with the spectral and temporal similarities shared by these artifact types, which can overlap in frequency and time domains, making them inherently more challenging to discriminate. The dominant diagonal entries confirm that the model effectively distinguishes EEG signals from both physiological and non-physiological artifacts in real and synthetic EEG recordings. These results support the suitability of the model for automated EEG artifact detection and preprocessing pipelines, aiding in the enhancement of signal quality for downstream analyses.

Table 2 presents the experimental results obtained for the DA-NN model on the TUH-EEG and EEGDenoiseNet datasets in the artifact classification task, considering precision, sensitivity, F1-score, and specificity as performance metrics. For the EEG class, TUH-EEG achieved a precision of 0.9896,

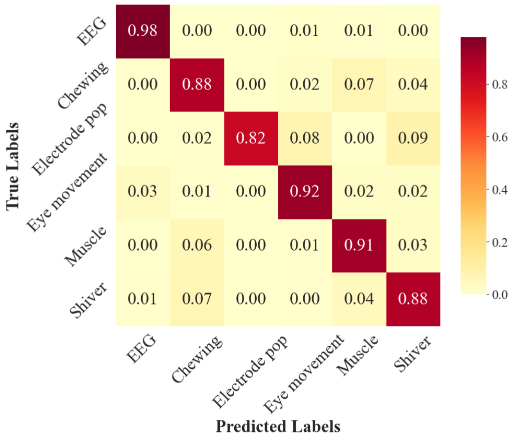


Figure 6: Confusion matrix — Source and Target domains.

sensitivity of 0.9566, F1-score of 0.9728, and specificity of 0.9972, while EEGDenoiseNet obtained 0.9325, 1.0, 0.9651, and 0.9795, respectively. For the Chewing artifact class, TUH-EEG reported a precision of 0.8878, sensitivity of 0.8440, F1-score of 0.8654, and specificity of 0.9709, compared to EEGDenoiseNet values of 0.8999, 0.9071, 0.9035, and 0.9724, showing improvements in most metrics. For the Electrode Pop artifact class, TUH-EEG achieved a precision of 1.0, sensitivity of 0.8485, F1-score of 0.9180, and specificity of 1.0, whereas EEGDenoiseNet recorded 0.7879 for precision, sensitivity, and F1-score, and 0.9983 specificity, indicating lower performance for this class. For the Eye Movement artifact class, TUH-EEG achieved 0.9449 precision, 0.9255 sensitivity, 0.9351 F1-score, and 0.9841 specificity, while EEGDenoiseNet obtained 0.9727, 0.9222, 0.9468, and 0.9923, showing improvements in precision and specificity. The most notable difference was observed in the muscle artifact class, where TUH-EEG achieved 0.8116 precision, 0.8814 sensitivity, 0.8450 F1-score, and 0.9451 specificity, while EEGDenoiseNet improved these values to 0.9769, 0.9349, 0.9554, and 0.9941, respectively. Finally, for the Shiver artifact class, TUH-EEG reported 0.8350 precision, 0.8755 sensitivity, 0.8548 F1-score, and 0.9772 specificity, whereas EEGDenoiseNet reached 0.8563, 0.8797, 0.8678, and 0.9805.

## 5.4 Discussion of Results

Table 3 shows that the proposed DA-NN, trained on representations based on the STFT magnitude and Sinkhorn, attains the highest average accuracy at 91.59 %, surpassing DenseNet201 at 88.30 %, CNN+Transformer at 79.22 %, and LSTM Ensemble + 2 CNN at 67.59 %. This ranking aligns with the learning dynamics in Figure 4: the DA-NN model reaches source and target training accuracies near 97.96 % and 92.17 %, with validation near 89.34 % (source) and 91.04 % (target). In contrast, the non-adapted counterpart drops to roughly 67 % on the target domain, a clear indication of covariate shift that is mitigated by the alignment term in Algorithm 2.

The confusion matrix in Figure 6 corroborates the class-wise behavior: clean EEG around 98 %, ocular near 92 %, muscular near 91 %, chewing and shiver near 88 %, and electrode pop near 82 %. Most errors cluster in classes with spectrotemporal similarity, notably chewing versus muscle and electrode pop versus ocular, suggesting that residual overlap in time–frequency signatures remains the principal source of ambiguity. Complementarily, Figure 5 shows a decrease in the KL/JS divergences between the latent spaces $z_s$ and $z_t$, which is consistent with improved cross-domain alignment and explains the transfer from TUH-EEG to EEGdenoiseNet.

From a clinical standpoint (Table 2), high target-domain values were obtained for both F1-score and specificity for ocular events (F1-score = 0.9468, specificity = 0.9923) and for muscle artifacts (F1-score = 0.9554, specificity = 0.9941). The most challenging class remains electrode pop, with an F1-score of approximately 0.9180 on TUH and 0.7879 on EEGdenoiseNet, consistent with its intra-class variability and partial overlap with transient ocular components.

The DA-NN methodology is central to these results. First, STFT-magnitude–based representations enhance class separability by preserving time- and frequency-localized patterns, which explains the consistent margin over baselines trained on raw EEG or on purely statistical features. Second, Sinkhorn-based regularization reduces cross-domain mismatch by aligning source and target distributions in the latent space, thereby maintaining performance under domain shift without sacrificing source-domain accuracy. Taken together, these elements yield both higher mean accuracy—91.59% versus 88.30%, 79.22%, and 67.59%—and greater robustness to false positives, as reflected in the specificity values.

Performance on minority classes (e.g., electrode pop) motivates targeted data-augmentation strategies that emulate rapid spectral onsets and mitigate class imbalance. The proposed evaluations under controlled SNR regimes would strengthen claims of generalization. Finally, combining STFT-magnitude features with the Sinkhorn objective in DA-NN yields superior artifact-classification performance, maintains reliability under domain shift, and is

Table 2: Comparison of classification performance metrics per class between TUH-EEG and EEGdenoiseNet Custom.

| **Class** | **TUH-EEG** | | | | **EEGdenoiseNet Custom** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Precision** | **Sensitivity** | **F1-score** | **Specificity** | **Precision** | **Sensitivity** | **F1-score** | **Specificity** |
| EEG | 0.9896 | 0.9566 | 0.9728 | 0.9972 | 0.9325 | 1.0000 | 0.9651 | 0.9795 |
| Chewing | 0.8878 | 0.8440 | 0.8654 | 0.9709 | 0.8999 | 0.9071 | 0.9035 | 0.9724 |
| Electrode pop | 1.0000 | 0.8485 | 0.9180 | 1.0000 | 0.7879 | 0.7879 | 0.7879 | 0.9983 |
| Eye movement | 0.9449 | 0.9255 | 0.9351 | 0.9841 | 0.9727 | 0.9222 | 0.9468 | 0.9923 |
| Muscle | 0.8116 | 0.8814 | 0.8450 | 0.9451 | 0.9769 | 0.9349 | 0.9554 | 0.9941 |
| Shiver | 0.8350 | 0.8755 | 0.8548 | 0.9772 | 0.8563 | 0.8797 | 0.8678 | 0.9805 |

Table 3: Comparison of the proposed method versus state-of-the-art approaches

| **Method** | **Feature Type** | **Accuracy** |
|---|---|---|
| (Kim and Keene, 2019) | Baseline EEG | 67.59% |
| (Peh et al., 2022) | Baseline EEG statistics | 79.22% |
| (Maiwald et al., 2023) | Spectrogram | 88.30% |
| **DA-NN model** | **STFT magnitude** | **91.59%** |

suitable for deployment as a robust preprocessing block within EEG analysis pipelines.

# 6 CONCLUSIONS

In conclusion, this study proposed the DA-NN model, a method designed for semi-supervised domain adaptation in a classification setting. The Sinkhorn-based domain adaptation strategy proved effective for artifact classification on both real and synthetic EEG data, where the source domain lacks both features and labels, while the target domain lacks only labels. For this classification task, using a source domain based on real data enabled training the DA-NN model capable of accurately identifying different artifact types, whereas incorporating a target domain modified through controlled adjustments to the SNR enhanced the model's generalization capability under adverse conditions, achieving an overall accuracy of 91.59%. Importantly, the proposed DA-NN reduced the performance gap between real and synthetic domains by 24%, demonstrating the effectiveness of domain adaptation in improving cross-domain robustness. As future research, we intend to investigate fully unsupervised learning approaches leveraging real unlabeled EEG recordings, with the goal of removing the dependency on labeled datasets while preserving competitive artifact classification performance.

# ACKNOWLEDGEMENTS

We gratefully acknowledge the College of Engineering, Temple University, for providing the real EEG data used in this work. We also acknowledge the support of the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) and Centro de Investigación e Innovación en Tecnologías de la Información y Comunicación (INFOTEC). The authors declare no conflicts of interest.

# REFERENCES

Ahmed, A. (2022). A quick survey of eeg signal noise removal methods. *Global Journal of Engineering and Technology Advances*, 11:098–104.

Azhar, M., Shafique, T., and Amjad, A. (2024). A convolutional neural network for the removal of simultaneous ocular and myogenic artifacts from eeg signals. *Electronics*, 13(22).

Bahador, N., Erikson, K., Laurila, J., Koskenkari, J., Ala-Kokko, T., and Kortelainen, J. (2020a). Automatic detection of artifacts in eeg by combining deep learning and histogram contour processing. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 138–141.

Bahador, N., Erikson, K., Laurila, J., Koskenkari, J., Ala-Kokko, T., and Kortelainen, J. (2020b). A correlation-driven mapping for deep learning application in detecting artifacts within the eeg. *Journal of Neural Engineering*, 17(5):056018.

Bao, C., Hao, Z., and Dou, W. (2022). Automatic removal of scalp eeg artifacts using an interpretable hybrid deep learning method. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1451–1456.

Brophy, E., Redmond, P., Fleury, A., De Vos, M., Boylan, G., and Ward, T. (2022). Denoising eeg signals for real-world bci applications using gans. *Frontiers in Neuroergonomics*, Volume 2 - 2021.

Cai, Y., Meng, Z., and Huang, D. (2025). Dhct-gan: Improving eeg signal quality with a dual-branch hybrid cnn–transformer network. *Sensors*, 25(1).

Choi, B. J. (2025). Removing neural signal artifacts with autoencoder-targeted adversarial transformers (at-at).

Chuang, C.-H., Chang, K.-Y., Huang, C.-S., and Jung, T.-P. (2022). Ic-u-net: A u-net-based denoising autoencoder using mixtures of independent components for automatic eeg artifact removal. *NeuroImage*, 263:119586. Epub ahead of print, August 27, 2022.

Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. (2017a). Joint distribution optimal transportation for domain adaptation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3733–3742, Red Hook, NY, USA. Curran Associates Inc.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. (2017b). Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865.

Farahani, A., Voghoei, S., Rasheed, K., and Arabnia, H. R. (2021). A brief review of domain adaptation. In Stahlbock, R., Weiss, G. M., Abou-Nasr, M., Yang, C.-Y., Arabnia, H. R., and Deligiannidis, L., editors, *Advances in Data Science and Information Engineering*, pages 877–894, Cham. Springer International Publishing.

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trouve, A., and Peyré, G. (2019). Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690.

Fuglede, B. and Topsoe, F. (2004). Jensen-shannon divergence and hilbert space embedding. In *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, pages 31–.

Gabardi, M., Saibene, A., Gasparini, F., Rizzo, D., and Stella, F. A. (2023). A multi-artifact eeg denoising by frequency-based deep learning.

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2nd edition.

Geuter, J., Kornhardt, G., Tomasson, I., and Laschos, V. (2025). Universal neural optimal transport. In Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., and Zhu, J., editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 19196–19232. PMLR.

Goyal, D. and Pabla, B. (2015). Condition based maintenance of machine tools—a review. *CIRP Journal of Manufacturing Science and Technology*, 10:24–35.

Grandini, M., Bagli, E., and Visani, G. (2020). Metrics for multi-class classification: an overview. *ArXiv*, abs/2008.05756.

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

Huang, X., Li, C., Liu, A., Qian, R., and Chen, X. (2024). Eegdfus: A conditional diffusion model for fine-grained eeg denoising. *IEEE Journal of Biomedical and Health Informatics*, pages 1–13.

Jackson, A. F. and Bolger, D. J. (2014). The neurophysiological bases of eeg and eeg measurement: a review for the rest of us. *Psychophysiology*, 51(11):1061–1071. Epub 2014 Jul 17.

Kim, D. and Keene, S. (2019). Fast automatic artifact annotator for eeg signals using deep learning. In *2019 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pages 1–5.

Lin, H.-Y., Tseng, H.-H., Lu, X., and Tsao, Y. (2021). Unsupervised noise adaptive speech enhancement by discriminator-constrained optimal transport.

Maiwald, A., Ackermann, L., Kalcher, M., and Wu, D. J. (2023). Image-based data representations of time series: A comparative analysis in eeg artifact detection.

Mashhadi, N., Khuzani, A. Z., Heidari, M., and Khaledyan, D. (2020). Deep learning denoising for eog artifacts removal from eeg signals. *2020 IEEE Global Humanitarian Technology Conference (GHTC)*, pages 1–6.

Pandya, S., Patel, P., Nord, B. D., Walmsley, M., and Ćiprijanović, A. (2025). Sidda: Sinkhorn dynamic domain adaptation for image classification with equivariant neural networks.

Peh, W. Y., Yao, Y., and Dauwels, J. (2022). Transformer convolutional neural networks for automated artifact detection in scalp eeg. *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3599–3602.

Quintero-Rincón, A., Giano, C. D., and Batatia, H. (2021). Chapter 11. artefacts detection in eeg signals. In *Artefacts Detection in EEG Signals*.

Saba-Sadiya, S., Chantland, E., Alhanai, T., Liu, T., and Ghassemi, M. M. (2021). Unsupervised eeg artifact detection and correction. *Frontiers in Digital Health*, 2.

Shlens, J. (2014). Notes on kullback-leibler divergence and likelihood. *ArXiv*, abs/1404.2000.

Tohka, J. and van Gils, M. (2021). Evaluation of machine learning algorithms for health and wellness applications: A tutorial. *Computers in Biology and Medicine*, 132:104324.

Xiong, W., Ma, L., and Li, H. (2024). A general dual-pathway network for eeg denoising. *Frontiers in Neuroscience*, 17.

Zhang, H., Zhao, M., Wei, C., Mantini, D., Li, Z., and Liu, Q. (2021). Eegdenoisenet: a benchmark dataset for deep learning solutions of eeg denoising. *Journal of Neural Engineering*, 18(5):056057.

Zunino, L., Olivares, F., Ribeiro, H. V., and Rosso, O. A. (2022). Permutation jensen-shannon distance: A versatile and fast symbolic tool for complex time-series analysis. *Phys. Rev. E*, 105:045310.