# Data Augmentation for EEG Motor Imagery Classification using Diffusion Model

**Nutapol Soingern**[1],**Akraradet Sinsamersuk**[2],**Chaklam Silpasuwanchai**[3]

Asian Institute of Technology, School of Engineering and Technology, Data Science and Artificial Intelligence, Pathum Thani, Thailand

E-mail: nutapol1997@gmail.com, akraradets@gmail.com, chaklam@ait.asia

November 2023

**Abstract.** Motor imagery classification using electroencephalogram (EEG) signals is an important research topic that has been extensively studied in the field of brain-computer interfaces (BCIs). However, due to the limited amount of available data, the possibility of overfitting is a challenge, especially when using a deep-learning classifier. One way to address this is by performing data augmentation. This paper investigates the efficacy of the diffusion model as a data augmentation method for motor imagery classification. We evaluated the diffusion method by comparing it with commonly-used EEG data augmentation techniques namely such as Noise Addition, Fourier Transform Surrogates, Frequency Shift, and SmoothTimeMask. The result shows that the diffusion method outperformed other methods in terms of classification accuracy by 17.49%. The Kullback-Leibler (KL) divergence is used for assessing the similarity between the training set (with and without augmentation) and validation set, thus showing the effectiveness of the diffusion approach compared to other techniques.

*Keywords*: Motor Imagery, EEG, Brain Computer Interface, Deep Learning, Data Augmentation, Diffusion, KL divergence

## 1. Introduction

Brain-computer interfaces (BCI) establish a direct pathway between the human brain and a computer via signal processing and decoding techniques. One classic paradigm of EEG is motor imagery (MI), in which its physiological basis is based on body movements or imagined movements that can produce $\alpha$ (8-13 Hz) and $\beta$ (13-30 Hz) event-related synchronization (ERS) and event-related desynchronization (ERD) rhythms in the motor-sensory areas of the brain [1]. Recently, the deep learning (DL) model has been used for motor imagery classification. For example, EEGNet [2] is a compact convolutional neural network designed for EEG-based brain-computer interfaces that effectively extracts spatial-temporal features from EEG signals. In any case, the paucity of data is a

prevalent issue in the field of EEG classification, as it hinders the development and performance of DL models. Consequently, a common symptom is overfitting, which reduces the model's accuracy and robustness on test set [3].

Data augmentation (DA) has been widely used to improve the robustness and accuracy of DL by artificially increasing the number of training data. Traditional EEG data augmentation methods include `Noise Addition` [4, 5, 6], fourier transform surrogates [7], `Frequency Shifting` [8, 9] and `SmoothTimeMask` [10]. `Noise Addition` [6, 5] adds random white noise to all channels. Fourier transform surrogates [7] randomizes the Fourier-transform (FT) phases of temporal-spatial data and generates surrogates that approximate examples from the data-generating distribution. `Frequency Shifting` [8, 9] randomly shifts the frequency spectrum on all channels. Last, `SmoothTimeMask` [10] randomly masks consecutive time steps of the EEG signal and replaces them with zeros, in which the motivation is to force the model to disregard minor irrelevant events.

Recently, diffusion model [11] was proposed which generates synthetic data based on Langevin dynamics. These models naturally admitted a progressive lossy decompression scheme. Diffusion mode has been used as a DA method to generate synthetic training data for skin disease classification [12], prostate cancer detection [13], chest X-ray imaging [14], etc.

In this work, we demonstrated the use of the diffusion model for data augmentation. Particularly, we developed our diffusion model based on `WaveGrad` [15] as a DA method for motor imagery classification. `WaveGrad` has been initially for audio waveform generation. The proposed approach involves utilizing score matching [16, 17] and diffusion probabilistic [18, 11] models to estimate gradients of the data density within a conditional model. Because `WaveGrad` is shown to be successful in waveform generation, we apply the technique in the EEG signal with sequence length adjusted. The process involves initializing the model with a Gaussian white noise signal and subsequently improving the signal quality through an iterative process that utilizes a gradient-based sampler that is conditioned on the mel-spectrogram. We evaluated the effect of the proposed method by performing DA on BCI Competition IV 2a [19] with various sizes of synthetic data with five standards EEG MI models (EEGNet [2], ATCNet [20], EEG-ITNet [21], Deep ConvNet [22] and ShallowFBCSPNet [22]). The proposed method improves the performance of classification models and outperforms other traditional EEG data augmentation methods.

## 2. Related Work

We reviewed commonly-used data augmentation for EEG MI such as `Noise Addition`, `Fourier Transform Surrogates`, `Frequency Shift` and `SmoothTimeMask`.

### *2.1. Noise Addition*

`Noise Addition` has two main categories for adding noise to the EEG signals for DA [23]. The first category regards adding various types of noise, such as Gaussian noise, Poisson noise, salt-and-pepper noise, etc., each of which has its own set of parameters (such as mean and standard deviation), to the original signal. The second category converts EEG signals to image sequences and then adds noise to the resulting image sequences. In any case, the introduction of noise to the training data is assumed to enhance the robustness of the model by compelling it to learn features that were less susceptible to minor fluctuations in the data. Indeed, such simulation of EEG data variability was commonly utilized to replicate the effects of electrode noise or subject movement during experimental procedures. Previous research has demonstrated that the inclusion of Gaussian noise in EEG signals enhances the efficacy of the MI classification model when applied to BCI competition IV dataset 2b [19], resulting in a 10% improvement in performance.

### *2.2. Fourier transform surrogates*

The Fourier transform surrogates (`FTSurrogate`) method utilizes the phase data of frequency elements, which were subsequently rearranged randomly while maintaining their original magnitude spectrum [7]. The generation of synthetic data samples has been utilized as a means to address the underrepresentation of certain classes. This approach has been shown to improve the balance of class distribution and enhance the accuracy of classification. The method proposed in this study has the potential to enhance classification performance either as a standalone technique or in conjunction with other data augmentation methods [7]. The extent of enhancement varies based on the particular dataset and classification issue. The mean F1-score of a convolutional neural network was improved by 7% in a sleep stage classification with the implementation of surrogate-based augmentation on the CAPSLPDB sleep database [24].

### *2.3. Frequency Shift*

In the `Frequency Shift` method, the frequency spectrum of an EEG signal was randomly shifted to a different frequency range while maintaining the amplitude spectrum [9]. The proposed technique involved generating novel EEG signals that exhibit identical spectral characteristics as the initial signal, albeit with altered frequencies. The `Frequency Shift` method was successful in enhancing the classification accuracy of certain EEG datasets. In the `Frequency Shift` method on motor imagery datasets, the implementation of the `Frequency Shift` method resulted in a 2.5% increase in classification accuracy when compared to the baseline method. Moreover, this technique has been compared with various other transformation techniques to generate augmented EEG signals [8]. The study assessed the efficacy of a novel method on various EEG classification tasks and demonstrated its superiority over conventional data augmentation techniques, including

random cropping and flipping, as well as other learned data augmentation methods. These findings indicated that the suggested approach yields optimal performance and requires less time for training compared to gradient-based methods in the class-agnostic context. Additionally, it surpassed gradient-free methods in the class-wise context. The research paper lacked a specific numerical value for the quantity of effects or enhancements. The effectiveness of this method in enhancing classification performance was also observed in the BCI Competition IV 2a dataset [19].

### *2.4. SmoothTimeMask*

`SmoothTimeMask` is a research methodology that utilizes time-domain augmentation to introduce smoothness into a signal. This is achieved by masking contiguous time intervals [10]. This method involved generating a mask by randomly selecting a starting point and masking a fixed length of contiguous samples. A common technique used to create a smooth transition between masked and unmasked regions is the application of a convolution with a Gaussian kernel to the mask. The introduction of smoothness in the augmented signal has the potential to prevent overfitting and enhance the generalization of the model [10]. This technique achieved an accuracy is 85.77% on the emotion recognition task.

### *2.5. WaveGrad*

`WaveGrad` [15] is a generative model for waveform generation that uses score matching [16, 17] and denoising [18, 11] to improve the quality of generated waveforms. The basic idea behind the method is to estimate the probability density function (PDF) of a dataset using a generative model, and then use this estimate to generate new data points that are similar to the original data. To achieve this, `WaveGrad` uses an autoregressive architecture that predicts each sample of the waveform conditioned on the previous samples. Specifically, `WaveGrad` uses a modified version of the WaveNet architecture that replaces the dilated convolutions with a set of learned gates and skip connections, which reduces the computational cost of the model.

As stated earlier, `WaveGrad` model is trained using score matching for comparing the log-density function of both generated data and real data. With score matching, the gradient of the log-density (score) function is easier to estimate than the function itself and matching the gradients is sufficient to match the distributions. In other words, we can estimate the PDF of a dataset by matching the score function of a model to the true score function of the PDF of the target dataset. Thus, the objective of `WaveGrad` is to minimize the difference between the score function of both generated data and the target dataset.

In the implementation of `WaveGrad`, the idea of score matching is extended to a weighted denoising score matching. The denoising autoencoder is trained to remove noise from the input data, and the score function of the denoised data is used to estimate the

true score function of the PDF. Then, we weigh the cleaner samples (less noise) with a higher value while the noisier samples get a lower weight. Then our loss function is:

$$L(\theta) = \sum_i w(x_i) |\nabla_x \log \hat{p}(\tilde{x}_i) - \nabla_x \log \hat{p}(x_i)|^2 \quad (1)$$

where $\theta$ are the parameters of the model, $x_i$ is a data point, $\hat{p}(x_i)$ is the true probability density function of the dataset, $\tilde{x}_i$ is the denoised version of $x_i$, and $w(x_i)$ is a weighting function that assigns a weight to each sample based on the level of noise in its label.

In addition, `WaveGrad` further improves the optimization of the model by using a variant of stochastic gradient descent called Stochastic Gradient Hamiltonian Monte Carlo (SGHMC). SGHMC uses Hamiltonian dynamics to simulate the motion of particles in a potential energy landscape, which improves the exploration of the parameter space during optimization.

Overall, `WaveGrad` is able to generate high-quality waveforms that are comparable to or better than previous state-of-the-art methods. It achieves this by combining denoising and score matching with a modified version of the WaveNet architecture and SGHMC optimization.

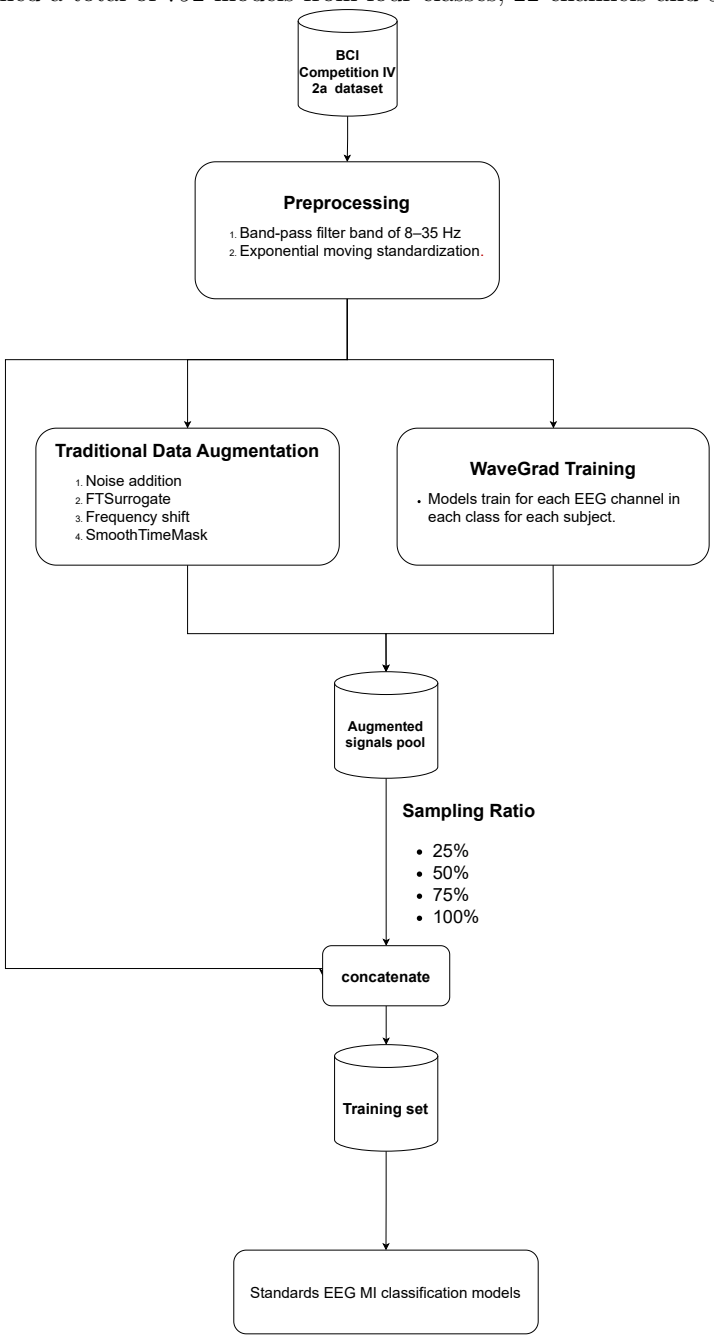
## 3. Methodology

We compared the diffusion method against four DA methods and the baseline method without augmentation. Figure 1 shows how the training sets were obtained using different combinations of the DA method and sampling size. The sampling size was chosen at the ratio of 25, 50, 75, 100%. Five commonly used models for MI classification were used. The models were trained using a subject-dependent scheme and evaluated on their respective testing sets.

### *3.1. Datasets*

BCI Competition IV 2a [19] was a collection of EEG data from 9 subjects who participated in a cue-based BCI paradigm involving four distinct motor imagery tasks: imagining movement of the left hand (class 1), the right hand (class 2), both feet (class 3) and the tongue (class 4). Each subject completed the tasks in two distinct sessions on different days, with each session consisting of six runs separated by brief pauses, resulting in a total of 48 trials (12 for each of the four classes). The data were captured while the participants sat in a comfortable armchair in front of a computer screen, and a fixation cross appeared at the start of each trial. A cue consisting of an arrow pointing to the left, right, down, or up was used to prompt the subjects to perform the desired motor imagery task. The subjects were instructed to perform the motor imagery task until the fixation cross disappeared from the screen without receiving any feedback. Signals were sampled at 250 Hz and filtered between 0.5 Hz and 100 Hz using a 50 Hz notch filter to reduce line noise. We used one section for training sets and another for test sets.

**Figure 1.** Showed how to obtain training sets when using traditional DA and `WaveGrad`. `WaveGrad` was trained for each EEG channel in each class for each subject. Thus, we trained a total of 792 models from four classes, 22 channels and 9 subjects.




### *3.2. Data Preprocessing*

EEG signal was filtered with a bandpass filter of 8 - 35 Hz followed by exponential moving standardization from `Braincode` library.

For exponential moving standardization, we computed the exponential moving mean $m_t$ at time $t$ as shown in equations (2). Then, we computed the exponential moving variance $v_t$ at time $t$ as shown in equation (3). Finally, we standardized the data point

$x_t$ at time $t$ with $\alpha$ as 0.001 and $\epsilon$ as 0.0001 as shown in equation (4).

$$m_t = \alpha \cdot \bar{x}_t + (1 - \alpha) \cdot m_{t-1}, \tag{2}$$

$$v_t = \alpha \cdot (m_t - x_t)^2 + (1 - \alpha) \cdot v_{t-1}, \tag{3}$$

$$x'_t = \frac{(x_t - m_t)}{\max(\sqrt{v_t}, \epsilon),} \tag{4}$$

### *3.3. Data Augmentation*

Our objective was to assess the performance of each DA method with different ratios of augmented/synthetic data. Five DA methods (`Noise Addition`, `FTSurrogate`, `Frequency Shift`, `SmoothTimeMask`, `WaveGrad`) and four ratios (25%, 50%, 75%, 100%) = 20 combinations (5 methods x 4 ratios) of the training set are to be created. Here, ratios refer to the amount of augmented/synthetic data used. For example, given a 25% ratio, 25% augmented data is randomly sampled and added to the original training set. By comparing different ratios, we can better understand the impact of the size of augmentations on accuracy improvement.

Similar to previous works [8, 10, 25, 24], we used the subject-dependent scheme which augments data in subject level (i.e., each subject is treated separately) and channel level (i.e., each channel is treated independently).

*3.3.1. Noise Addition:* `Noise Addition` entails the inclusion of diverse forms of noise, such as Gaussian, Poisson, and others, that possess varying parameters to the original EEG signal. The raw EEG signal was subjected to additive noise by incorporating a Gaussian distribution with a standard deviation of 0.1.

*3.3.2. Fourier Transform Surrogates:* `Fourier transform surrogates` are a type of data generated by randomizing the phases of temporal-spatial data.

*3.3.3. Frequency Shift:* The technique of `Frequency Shift` is characterized by the alteration of the frequency of the EEG signal by a specific amount. We random shift the frequency by $\pm 2$ Hz.

*3.3.4. SmoothTimeMask:* `SmoothTimeMask` involves applying a smooth window function to mask a continuous segment of the signal and optimizing it using gradient-based methods. The signal was randomly masked with a range of 100 sample points.

*3.3.5. WaveGrad:* It is first important that `WaveGrad` is originally a generative model, not a formal augmentation technique. Thus, contrary to other DA methods, `WaveGrad` has to be trained before it can be used to generate a synthetic EEG signal. The dataset consists of 9 subjects and four MI classes. The EEG recording has 22 channels. Thus, the total number of `WaveGrad` models was (9 subjects x 22 channels x 4 classes) = 792

models. The training procedure and parameters were the same across all `WaveGrad` models. The learning rate was set to 0.0001 and the diffusion steps to 1000. talk more about what you add/modify

### *3.4. Evaluation*

First, it is important to evaluate the quality of the augmented/synthetic data. A common way is through dimensionality reduction KL divergence. The success of the diffusion method in comparison to other methods is demonstrated by the Kullback-Leibler (KL) divergence, which is used to measure the similarity between the training set (with and without augmentation) and the validation set. We expect that high-quality augmented or synthetic data should exhibit similarity between the training set (with and without augmentation) and the validation set.

Second, once the quality of the augmented/synthetic data is quantified, we are now ready to quantify the usefulness of data augmentation techniques on actual EEG tasks. We first selected five commonly used motor imagery classification models (EEGNet, ATCNet, EEG-ITNet, Deep ConvNet and ShallowFBCSPNet) which would allow us to understand how the complexity of the model relates to data augmentation. Here, note that we simply define the complexity based on the model's number of parameters. Accuracy was then measured across all 21 combinations (5 DA methods x 4 sampling ratios + 1 baseline method without augmentation). The details of each model were as follows:

*3.4.1. EEGNet* EEGNet is a single CNN architecture that can accurately classify EEG signals from different BCI paradigms while being as compact as possible. The authors introduced the use of depthwise and separable convolutions to construct an EEG-specific model that encapsulates well-known EEG feature extraction concepts for BCI. They compared EEGNet to current state-of-the-art approaches across four BCI paradigms and showed that EEGNet generalizes across paradigms and achieves comparably high performance when only limited training data is available across all tested paradigms.

*3.4.2. ATCNet* ATCNet was initially developed for predicting the onset of epileptic seizures using electroencephalogram (EEG) signals. The ATCNet consists of two blocks: an attention-based temporal convolutional (ATC) block and a transformer-based classification (TC) block. The ATC block is used to extract relevant features from the EEG signals, while the TC block is used to classify the extracted features into seizure and non-seizure classes. The proposed model was evaluated using the BCI Competition IV-2a (BCI-2a) dataset. The obtained accuracy ranged from 60.5% to 89.5%.

*3.4.3. EEG-ITNet* EEG-ITNet uses inception modules and causal convolutions with dilation to extract rich spectral, spatial, and temporal information from multi-channel EEG signals with less complexity than other existing end-to-end architectures. The

paper also provided a methodology for achieving intuitive visualization structures such as topographic maps. The proposed EEG-ITNet model showed up to a 5.9% improvement in classification accuracy compared to its competitors in different scenarios.

*3.4.4. Deep ConvNet* The deep ConvNet had four convolution-max-pooling blocks, with a special first block designed to handle EEG input, followed by three standard convolution-max-pooling blocks and a dense softmax classification layer. The authors found that recent advances in machine learning, including batch normalization and exponential linear units, together with a cropped training strategy, boosted the Deep ConvNets decoding performance, reaching at least as good performance as the widely used filter bank common spatial patterns (FBCSP) algorithm.

*3.4.5. ShallowFBCSPNet* The shallow ConvNet is similar to the transformations of FBCSP. Concretely, the first two layers of the shallow ConvNet perform a temporal convolution and a spatial filter, as in the deep ConvNet. These steps are analogous to the bandpass and CSP spatial filter steps in FBCSP.

## 4. Results

### *4.1. Kullback-Leibler divergence(KL divergence)*

To understand the quality of the generated data, we measured the similarity of the signal by the KL divergence process. We measure the KL divergence by comparing of train set and test set, test set and train set with 25% of data augmentation, test set and train set with 50% of data augmentation, test set and train set with 75% of data augmentation, test set and train set with 100% of data augmentation in Table 1, and test set and data augmentation and train set and data augmentation in Table 2. We randomized the augmentation data for this process 100 times and averaged the result. Overall, our method increases similarity as the ratios of augmented data are increased. When ratios are increased, the similarity for the SmoothTimeMask, Noise Addition, and FTSurrogate approaches that of the non-augmented data. On the other hand, as the augmented data from frequency shift increases, the similarity declines.

## References

[1] Wolpaw J R 2013 Brain–computer interfaces *Handbook of clinical neurology* vol 110 (Elsevier) pp 67–74

[2] Lawhern V J, Solon A J, Waytowich N R, Gordon S M, Hung C P and Lance B J 2018 *Journal of Neural Engineering* **15** 056013

[3] Bilbao I and Bilbao J 2017 Overfitting problem and the over-training in the era of data: Particularly for artificial neural networks *2017 eighth international conference on intelligent computing and information systems (ICICIS)* (IEEE) pp 173–177

[4] Wang F, Zhong S h, Peng J, Jiang J and Liu Y 2018 Data augmentation for eeg-based emotion recognition with deep convolutional neural networks *MultiMedia Modeling: 24th International*

**Table 1.** The result of the study of KL divergence of raw EEG dataset and data augmentation from `WaveGrad`

.

| Subject ID | test set vs train set with x% of augmentation | | | | |
|---|---|---|---|---|---|
| | 0% | 25% | 50% | 75% | 100% |
| 1 | 2949.63 | 2757.15 | 2623.40 | 2525.57 | 2486.21 |
| 2 | 1641.60 | 1887.98 | 2009.06 | 2118.49 | 2222.43 |
| 3 | 2203.27 | 2298.26 | 2382.90 | 2416.10 | 2410.30 |
| 4 | 2559.50 | 2390.33 | 2291.63 | 2240.40 | 2176.35 |
| 5 | 2238.52 | 2215.49 | 2170.81 | 2146.55 | 2167.71 |
| 6 | 1935.63 | 1999.40 | 2061.10 | 2067.91 | 2103.15 |
| 7 | 2704.38 | 2564.74 | 2482.31 | 2424.12 | 2381.11 |
| 8 | 3245.50 | 2946.14 | 2752.42 | 2628.08 | 2497.71 |
| 9 | 2313.45 | 2290.44 | 2253.48 | 2231.11 | 2201.39 |

**Table 2.** The result of the study of KL divergence of raw EEG dataset and data augmentation from `WaveGrad`

.

| Subject ID | Train set vs Augmentation | Test set vs Augmentation |
|---|---|---|
| 1 | 1988.35 | 1582.51 |
| 2 | 2749.47 | 2970.01 |
| 3 | 2671.86 | 2800.51 |
| 4 | 1775.89 | 2193.34 |
| 5 | 2045.00 | 1870.57 |
| 6 | 2247.53 | 2069.20 |
| 7 | 2024.06 | 1808.16 |
| 8 | 1779.83 | 1398.65 |
| 9 | 2145.52 | 2123.10 |

*Conference, MMM 2018, Bangkok, Thailand, February 5-7, 2018, Proceedings, Part II 24* (Springer) pp 82–93

[5] Parvan M, Ghiasi A R, Rezaii T Y and Farzamnia A 2019 Transfer learning based motor imagery classification using convolutional neural networks *2019 27th Iranian Conference on Electrical Engineering (ICEE)* (IEEE) pp 1825–1828

[6] Li Y, Zhang X R, Zhang B, Lei M Y, Cui W G and Guo Y Z 2019 *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **27** 1170–1180

[7] Schwabedal J T, Snyder J C, Cakmak A, Nemati S and Clifford G D 2018 *arXiv preprint arXiv:1806.08675*

[8] Rommel C, Moreau T, Paillard J and Gramfort A 2021 *arXiv preprint arXiv:2106.13695*

[9] Rommel C, Paillard J, Moreau T and Gramfort A 2022 *Journal of Neural Engineering* **19** 066020

[10] Mohsenvand M N, Izadi M R and Maes P 2020 Contrastive representation learning for electroencephalogram classification *Machine Learning for Health* (PMLR) pp 238–253

[11] Ho J, Jain A and Abbeel P 2020 *Advances in neural information processing systems* **33** 6840–6851

[12] Akrout M, Gyepesi B, Holló P, Poór A, Kincső B, Solis S, Cirone K, Kawahara J, Slade D, Abid L *et al.* 2023 *arXiv preprint arXiv:2301.04802*

[13] Hao R, Namdar K, Liu L, Haider M A and Khalvati F 2021 *Journal of Digital Imaging* **34** 862–876

[14] Motamed S, Rogalla P and Khalvati F 2021 *Informatics in Medicine Unlocked* **27** 100779

[15] Chen N, Zhang Y, Zen H, Weiss R J, Norouzi M and Chan W 2020 *arXiv preprint arXiv:2009.00713*

[16] Song Y, Garg S, Shi J and Ermon S 2020 Sliced score matching: A scalable approach to density

and score estimation *Uncertainty in Artificial Intelligence* (PMLR) pp 574–584
[17] Song Y and Ermon S 2020 *Advances in neural information processing systems* **33** 12438–12448
[18] Sohl-Dickstein J, Weiss E, Maheswaranathan N and Ganguli S 2015 Deep unsupervised learning using nonequilibrium thermodynamics *International conference on machine learning* (PMLR) pp 2256–2265
[19] Brunner C, Leeb R, Müller-Putz G, Schlögl A and Pfurtscheller G 2008 *Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology* **16** 1–6
[20] Altaheri H, Muhammad G and Alsulaiman M 2022 *IEEE Transactions on Industrial Informatics* **19** 2249–2258
[21] Salami A, Andreu-Perez J and Gillmeister H 2022 *IEEE Access* **10** 36672–36685
[22] Schirrmeister R T, Springenberg J T, Fiederer L D J, Glasstetter M, Eggensperger K, Tangermann M, Hutter F, Burgard W and Ball T 2017 *Human brain mapping* **38** 5391–5420
[23] Lashgari E, Liang D and Maoz U 2020 *Journal of Neuroscience Methods* **346** 108885
[24] Terzano M G, Parrino L, Sherieri A, Chervin R, Chokroverty S, Guilleminault C, Hirshkowitz M, Mahowald M, Moldofsky H, Rosa A *et al.* 2001 *Sleep medicine* **2** 537–554
[25] Leeb R, Brunner C, Müller-Putz G, Schlögl A and Pfurtscheller G 2008 *Graz University of Technology, Austria* 1–6