Diffusion-Based Data Augmentation for Improving the Classification of EEG Signals

Mischa Tomasz Rauch
Department of Advanced Computing Sciences
Faculty of Science and Engineering
Maastricht University
Maastricht, The Netherlands

***Abstract*—The subfield of psychology, emotion recognition, focuses on identifying emotions and addressing mental conditions like autism. Due to advancements in machine learning and deep learning the automation of emotion recognition was drastically improved [1]. These techniques allow for the development of models that analyze facial expressions, speech, and texts to evaluate human emotions. The emotion evaluation based on Electroencephalogram (EEG) signals offers a promising approach for emotion recognition since its not susceptible to facial suppression. EEG is also cost-effective, non-invasive, and advantageous over traditional methods like Magnetic Resonance Imaging. However, the limited availability of EEG data hampers progress due to restricted access and privacy concerns. By augmenting the existing dataset with artificially generated signals mimicking the actual data distribution, emotion classification can be improved.**

In this thesis, a Probabilistic Diffusion Model (PDM) is proposed as generative artificial intelligence. Three major feature extraction approaches are tested. The first is with a direct comparison to a Generative Adversarial Network [2], the second in which features from the frequency domain are extracted, and a third in which a reduced subset is tested. Additionally, experiments with different augmentation factors [x1,x2,x3,x4] are tested to evaluate the impact of the data augmentation. The experimental results demonstrated that the GAN model achieved higher improvements than the PDM, but the PDM achieves classification improvements of up to 20% if the signal is transformed into the frequency domain before. Lastly, the reduced subset approach yielded similar accuracies proposing future work to use a subset of the initial signals.

I. Introduction

Biosignals, such as Electroencephalogram (EEG) and Electrocardiogram (ECG), are useful tools for recording brain activity and other vital organs. Compared to traditional methods like Magnetic Resonance Imaging (MRI) or invasive techniques, such as extracellular Action Potentials (APs) or local field potentials (LFPs), they are non-invasive and less expensive. EEG signals are recorded by placing electrodes on various locations of the scalp to register brain activity. The signals are useful in detecting brain abnormalities by inspecting the signal and measuring the deviation from normal signal patterns. Biosignals are also used in Human Machine Interfaces (HMI) to interface the brain with machines, this enables complex tasks such as controlling a motorized hand or other external systems [3].

This thesis was prepared in partial fulfilment of the requirements for the Degree of Bachelor of Science in Data Science and Artificial Intelligence, Maastricht University. Supervisor: Enrique Hortal

Another area where EEG signals are extensively used is emotion recognition [4], [5]. Emotions play a significant role in human communication, and EEG signals offer a more accurate approach to detecting them than facial expressions since they are not susceptible to facial suppression. Researchers have been using EEG signals in the last decade to classify emotions, and recent advancements in deep learning have enabled the development of models with higher accuracy in deciphering human emotions through EEG signals.

The progress of deep learning in medical data, such as EEG data, has been hindered by a lack of data availability. Medical data is highly sensitive and task-specific, making it difficult to access or generate datasets for research purposes. Additionally, EEG data is subject-dependent, with differences in variables such as probe placement, head size, location of the experiment, and brain anatomy [6]. Consequently, it is challenging to compare model performance with similar datasets, this is leading to research bottlenecks. Furthermore, training data is subject-dependent, which makes it difficult to achieve similar model performances on other test subjects. Creating data through lab-driven experiments is a difficult task that requires specialized equipment, as well as expert knowledge to extract and annotate the data. Although preliminary data augmentation is generally done to increase the spread of data, the results obtained from the modifications fed into an AI are not as impressive.

The aforementioned drawbacks motivate this project, which is to test a state-of-the-art generative artificial intelligence (AI) to perform data augmentation on EEG signals.

A. Data Augementation

Data augmentation is a technique used to expand an existing dataset by creating new data samples through modification or transformation of existing data. This method can address the problem of overfitting and enhance the generalization ability of models, thereby improving stability and accuracy [7]. Data augmentation is widely used in computer vision, particularly in object/image recognition, where it involves applying geometric transformations such as rotation, translation, scaling, cropping, padding, and flipping. However, since EEG signals are time-series data, applying such transformations can destroy their temporal features and make it challenging to create labels for the transformed data [2]. In recent years, data augmentation has been extensively studied in the context of EEG for various tasks, including emotion recognition, seizure detection, motor task, sleep stage detection, visual task, and mental workload. Various augmentation methods, such as noise addition, Generative Adversarial Networks (GAN), sliding window, sampling, and Fourier transform, have been employed, and their effectiveness varies depending on the task. Lashgari et al. have conducted a detailed study of the available methods [8].

B. Problem Statement

The main problem statement in this research is targeting the question: *Is it possible to achieve better classification performance by using an augmented dataset generated by a probabilistic diffusion model?* Therefore, a probabilistic diffusion model will be used to create artificial EEG signals. The work of a diffusion model can be split into two parts: a diffusion process or forward process which is fixed to a Markov Chain that gradually adds Gaussian noise to the data until the whole image contains only gaussian noise. The second part is the reverse process, which removes gaussian noise from an image [9]. By letting an Artificial Neural Network (ANN) observe the diffusion process, it learns to reverse the process of removing the noise as well which enables the model to sample new images from pure gaussian noise. This thesis will investigate the applicability of this technique in the classification of the dataset DEAP: A dataset for emotion analysis using EEG, physiological and video signals [10]. The dataset is private and thus a suitable EULA was signed to obtain it. The core steps of the project can be summarized as follows:

1) Literature and code review of similar techniques for data augmentation in an EEG data set. Study of diffusion models in a similar context.
2) Understand, pre-process, and frame the data so that it can be used in the classification task.
3) Design a baseline classification model using the data.
4) Develop a diffusion model that would take chunks of data from the above-mentioned data sets and generate synthetic data that would best capture the characteristics of the data.
5) Use of the synthetic data to augment the original dataset and create a newer/larger dataset.
6) Pass new dataset to the original model to generate classification results and make necessary comparisons with the baseline model and previously conducted work on GAN.
7) Test other ways of preprocessing the data and generating synthetic data, such as transformations in the frequency domain, in comparison with previous methods.
8) Explore feature importance to test the classification effect by removing less important/relevant features.

C. Research Question

In this thesis, the following research questions are addressed:

1) Are Diffusion-based approaches better than other augmentation models, for example, GANs?
2) Which features are more suitable for a diffusion model-based data augmentation approach?
3) What are good/significant features to classify EEG signals, with respect to feature importance?

D. Related Work

Since the release of the DEAP dataset, numerous researchers have experimented with it. Previous studies have demonstrated the non-linear and multi-fractal nature of EEG signals, which has led to the use of fractal decomposition to extract features that can be used for training machine learning algorithms such as Support Vector Machines (SVM). For instance, Liu and Sourina [11] have applied this approach and reported an accuracy of 53.7% for classifying the four emotion types in the DEAP dataset. This thesis is based on similar research by Bhat and Hortal [2] which applied a Wasserstein GAN to create synthetic data to improve classification. Their research resulted in an increment between 3.75% to 17.5% throughout the four classification emotions: valence, arousal, dominance, and liking. The use of Stable Diffusion has so far only been tested on recreating EEG signals by Torma and Szegletes [12], the approach of extracting features from an EEG signal and testing different approaches to enhance classification however, has not been researched yet and will be the main focus in this work.

E. Diffusion Models

Diffusion Models (DM) are part of generative models. Given observed samples x from a distribution of interest, the goal of a generative model is to learn to model its true data distribution $p(x)$. Once learned, we can generate new samples from the approximate model at will. Furthermore, under some formulations, we are able to use the learned model to evaluate the likelihood of observed or sampled data as well. The generative AI probabilistic diffusion model (PDM) is a deep learning technique that has recently gained popularity in the field of AI [13]. The aim of the model is to generate new samples that follow the same statistical properties as the original data set.

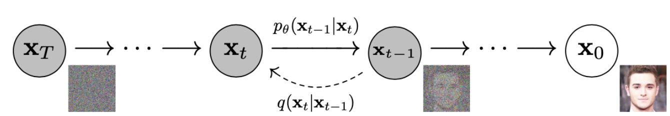

Fig. 1: The diffusion process. Extracted from [9]

At its core, the PDM uses a series of diffusion processes to model the data distribution of the form:

$$p_\theta(x_0) := \int p_\theta(x_{0:T})\, dx_{1:T} \tag{1}$$

where $x_1, ..., x_T$ are latents of the same dimensionality as the data $x_0 \sim q(x_0)$ [9]. In each diffusion process, Gaussian noise is added to the current input, and the procedure is observed by an artificial neural network. The process is repeated iteratively, with the output of one diffusion process serving as the input to the next one. By using multiple diffusion processes, the model is able to capture complex data distributions. The approximate posterior $q(x_{1:T}|x_0)$, called the *forward process* or *diffusion process*, is fixed to a Markov chain that gradually adds Gaussian noise to the data according to a variance schedule $\beta_1, ..., \beta_T$:

$$q(x_{1:T}|x_0) := \prod_{t=1}^{T} q(x_t|x_{t-1}), \tag{2}$$

$$q(x_t|x_{t-1}) := \mathbb{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{3}$$

The joint distribution $p_\theta(x_{0:T})$ is called *reverse process*, in which noise is gradually removed from the images. This process is defined as a Markov chain, starting at $p(x_T) = \mathbb{N}(x_T; 0; I)$:

$$p_\theta(x_0 : T) := p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \tag{4}$$

$$p_\theta(x_{t-1}|x_t) := \mathbb{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{5}$$

A more in depth explanation of equations 2, 3, 4, and 5 can be found by Ho, Jaain, and Abbeel [9]. Because the PDM is a generative AI it can be used in a variety of tasks, like the synthesis, denoising, and inpainting of images. In image synthesis, a model is trained on a dataset of images and can generate new images that are similar to the training set. In denoising, the model can remove noise from an image by iteratively diffusing the noise away. In inpainting, the model can fill in missing parts of an image by generating plausible pixel values based on the surrounding pixels. One of the strengths of the PDM is that it can model complex data distributions with relatively few parameters. This is achieved by using the diffusion process to slowly transform the input data into the target distribution. Another strength of the PDM is that it can handle missing or incomplete data since it can generate plausible values for the missing data. Despite the listed strengths diffusion models do not come without limitations; it can be computationally expensive to train, especially for large datasets till the latent variable is learned. Additionally, the model assumes that the data distribution is continuous and smooth, which might not always be given. In conclusion, the generative AI probabilistic diffusion model is a powerful deep learning technique that can model complex data distributions and be used in a variety of tasks.

II. Methodology

A. Data

The DEAP dataset is a collection of physiological signals used for emotion analysis [10]. It contains EEG data from 32 individuals, who watched 40 one-minute videos and provided valence, arousal, dominance, and liking ratings on a scale of 0-9 during a self-assessment phase. The EEG data was recorded from 32 channels on the scalp sampled at 512Hz, as well as 12 peripheral channels, 3 unused channels, and 1 status channel. In this study, the preprocessed version of the dataset was used, which was provided by the DEAP team and consisted of data files in .dat format for each participant downsampled to 128Hz. Each participant file contains two parts, data and labels, and was considered for experimentation.

1) The 8.064 datapoint EEG data array for each of the 40 experiments was obtained from 40 different channels. This in total generates 322.560 data points per experiment
2) A label array containing a set of 4 values for each experiment representing the emotions: valence, arousal, dominance, and liking.

Table I provides an insight into the data dictionary component for each participant.

TABLE I: Insight of DEAP dataset

Array File	Shape	Contents
Data	40x40x8064	Number of Videos/ experiments x Number of Channels x Number of Datapoints
Labels	40x4	Number of Videos/ experiments x Number of labels

In order to enhance the interpretability of integer labels for a classification task, they are transformed into binary classes. The rating values are partitioned into two classes based on a threshold of 5, where values less than the threshold are assigned to class 0 and those greater than or equal to it are assigned to class 1. This conversion simplifies the application of binary classification algorithms for subsequent analyses. This process was followed by previous work to establish a fairer comparison [2].

B. Data Preprocessing

The data file for each recorded experiment contains 8.064 features per experiment. To accelerate computation it was necessary to reduce the dimensionality. To draw a fair comparison between the work from Bhat and Hortals [2] the data was divided into ten parts of approximately 807 points each. For each batch, the following features are extracted: mean, median, maximum, minimum, standard deviation, variance, range, skewness, and kurtosis. These features were extracted for each of the ten batches and additionally calculated overall the sum of each batch. This resulted in 99 features. The described approach was followed to later have a fair comparison between GAN and the used diffusion model. The described approach was performed for all 40 available channels.

For the second and third research questions, an alternative methodology was employed, involving the transformation of the signals into the frequency domain. Two distinct approaches were considered for this purpose. In the first approach, the signals underwent a Short-time Fourier transform (STFT) procedure. In this method, the signals were segmented into the same ten parts as in research question one. For each segment and EEG band (Theta: 4-7Hz, Alpha: 8-12Hz, Beta: 13-30Hz, Gamma: 31-50Hz), the mean was extracted. Consequently, this resulted in 40 features, alongside an additional 4 data points representing the overall signal.

In the second approach, the Fast Fourier Transformation (FFT) technique was applied to the entire signal, from which eight data points were derived per signal. Specifically, the mean and standard deviation were calculated for each of the four, previously mentioned, EEG bands. The data preprocessing in the frequency domain was only applied for the real EEG signals of the preprocessed data which were 32 out of 40 channels. These distinctive approaches were adopted to explore different perspectives and extract relevant insights for the respective research questions, ensuring a comprehensive analysis of the data.

C. *Framework for EEG data augmentation*

To ensure accurate training and evaluation of data generated from the diffusion model, it was necessary to partition the data into separate train and test sets. To establish a benchmark for evaluation, the 32-subject DEAP dataset was randomly shuffled and divided in a manner that allocated 22 subjects to the training dataset. The remaining 10 subjects' data was reserved for testing the model both before and after augmentation. This separation is crucial to prevent potential bias in the classification of the 10 participant test data. If the same 32-participant dataset was used for both training and testing, the model's classification accuracy could be compromised. This is due to the test data already being present in the training dataset, leading to an inflated performance evaluation. By keeping the test data separate from the training data, we ensure a fair and unbiased assessment of the model's classification performance. The diffusion model generates additional participant data comprising 22 participants. This generated data exhibits a probability distribution that closely resembles the data used for training the model. Subsequently, this newly generated data is appended to the original training set, creating an augmented training dataset. This augmented dataset facilitates the training and evaluation of the model's performance after augmentation.

D. *Probabilistic Diffusion Model*

To optimize time utilization and leverage cutting-edge methodologies, an openly available implementation of the diffusion model was employed. This approach allowed to benefit from the latest advancements in the field, as the diffusion model represents a state-of-the-art method. By utilizing an existing implementation, valuable time and effort were conserved, enabling a more efficient exploration and analysis of the diffusion model within the context of this research. This strategic decision aligned to incorporate contemporary methodologies while optimizing productivity and ensuring adherence to the most up-to-date practices in the field. The used model is based on the work from Ho, Jain, and Abbeel [9], which consists mainly of two algorithms a training algorithm and a sampling algorithm.

The implementation used in this study was derived from the work of Phil Wang [14], as documented in their publication.

Algorithm 1 Training

```
repeat
    x_0 ~ q(x_0)
    t ~ Uniform({1, ..., T})
    ε ~ N(0, I)
    Take gradient decent step on
        ∇_θ ||ε − ε_θ(√(ᾱ_t) x_0 + √(1 − ᾱ_t) ε, t)||^2
until converged
```

$$\nabla_\theta ||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)||^2$$

Algorithm 2 Sampling

```
x_T ~ N(0, I)
for t = T, ...1 do
    z ~ N(0, I) if t > 1, else z = 0
    x_{t−1} = 1/√(α_t) (x_t − (1−α_t)/√(1−ᾱ_t) ε_θ(x_t, t)) + σ_t z
end for
return x_0
```

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right) + \sigma_t z$$

This implementation encompassed the use of two distinct algorithms, employed for both training the model using extracted images, as seen in algorithm 1 and generating synthetic images through sampling techniques as seen in algorithm 2. To ensure compatibility with non-squared images, minor modifications were made to the implementation. Additionally, a method was implemented to track the loss over time, facilitating the comparison of different loss functions. These adjustments were deemed necessary to enable the successful execution of the model and enhance its applicability within the scope of this research study. The adaptions can be found in the associated GitHub repository.[1]

III. Experiments

A. *Baseline*

To establish a performance benchmark for the augmented models, a baseline model was constructed, leveraging the original training set comprising 22 subject data. Traditional machine learning algorithms, namely K-nearest neighbours (KNN) and SVMs, were employed for this purpose. The machine learning models were trained using the label attributes of valence, arousal, dominance, and liking. To determine the optimal parameters, a grid search approach was employed. The KNN method was evaluated with varying k values ranging from 5 to 50 for each label type, while SVM was tested with different values of the parameter c, spanning from $1e^{-4}$ to $1e^{2}$. Furthermore, a convolutional neural network (CNN) architecture, comprising a sequence of Conv2d layers, was used to train the actual data. The network optimization was performed using autokeras-based grid search [15], aiming to attain the best possible outcome. An Adam optimizer with a default learning rate of $\alpha = 10^{-3}$ was employed in this process. A detailed discussion of the obtained results can be found in Section IV.

[1] https://github.com/MischaRauch/Stable_Diffusion_on_EEG_Signals

B. Verification: Quality of the generated data

To ensure the similarity of synthetic data properties to real data and enhance the classification results, two tests were conducted. These two tests, K-fold cross-evaluation and the feature distribution comparison provided a comprehensive evaluation of the synthetic data's properties in relation to the real data, ensuring the suitability and effectiveness of the synthetic data for classification tasks.

1) K-fold cross evaluation: Firstly, K-fold cross-evaluation was employed to assess the performance and generalization capability of the classification model using the synthetic data. This technique involved partitioning the dataset into K subsets, utilizing $K-1$ subsets for training and the remaining subset for validation, iteratively repeating the process K times. The process was split into four sub-parts in which the accuracies were collected for:

1) training on the real data and test on the real data
2) training on the real data and test on the synthetic data
3) training on the synthetic data and test on the real data
4) training on the synthetic data and test on the synthetic data

If the accuracies with the same training data are similar to one another, the hypothesis that the synthetic data is similar to the real data would be supported. Additionally, to check for statistical significance a statistical test on the classification results was conducted. This test showed if there exists a statistical difference between training on the real dataset and training on the synthetic dataset.

2) Distribution plots: Additionally, feature distribution plots were generated and compared between the synthetic and real data. These plots visualized the distribution of key features and allowed for a qualitative assessment of their similarity. The overlapping patterns and statistical characteristics reveal insights into the consistency of feature distributions between the two datasets.

C. Verification: Performance of model with data augmentation

The data generated from the DM is employed to augment the existing dataset. Classification algorithms, including KNN, SVM, and neural networks, were evaluated using different combinations of K values, kernels, and network layers, as specified in section III-A. The augmentation process contained four stages: $x1$ (no augmentation), $x2$, $x3$, and $x4$, representing different extends of data augmentation. The analysis and performance of these models will be discussed in Section IV.

1) Comparison with previous GAN model: The initial augmentation experiment involved conducting a comparative analysis between the proposed model and a GANs model that incorporated identical features. This comparative approach aimed to establish a fair and unbiased evaluation by considering both models within the same experimental framework. By employing the GANs model as a reference point, the augmentation performance and effectiveness of the proposed model could be assessed, enabling a comprehensive and reliable comparison between the two approaches. To investigate the capabilities of a DM model two DM were trained. One was trained with 12000 training steps, whereas the other was trained for 100000 training steps. Both models used a learning rate of $8e^{-5}$ throughout the whole training process. Before the training was conducted, an experiment between the $l1$ and $l2$ loss was established to fasten the convergence of the model to find out which loss performs better.

2) STFT approach: Another experimental approach involved transforming the signal into the frequency domain using the STFT technique. Hereby the window size was the same as in the previous experiment which resulted in 44 features. These features are represented by ten sliding windows multiplied by four extracted means throughout the four EEG bands (Theta, Alpha, Beta, Gamma), and an additional four means which got extracted over the whole signal in the frequency domain. These features were extracted for all pure 32 EEG signals. This approach allowed for a comprehensive representation of the signal characteristics in the frequency domain, facilitating a detailed analysis and exploration of the underlying patterns and trends present within the data.

3) FFT approach: To explore an alternative approach within the frequency domain, the FFT was employed. In this experiment, the entire signal was transformed into the frequency space, allowing for an analysis of frequency components. For each of the previous four mentioned EEG bands, statistical measures such as the mean and standard deviation were extracted from the Fourier-transformed signal. This method yielded a total of eight features for each individual signal, encompassing key statistical characteristics within the frequency domain. This procedure was applied for all real 32 EEG signals. By using this approach, a comprehensive understanding of the frequency characteristics of the signals was obtained, enabling a thorough investigation and comparison of different features within the frequency space. Similar to the previous approach, firstly a comparison between $l1$ and $l2$ loss was conducted to find out which loss performs better with the given data. Secondly, two models were trained one for 12000 training steps and a second one for 100000 training steps, with the same learning rate of $8e^{-5}$ throughout the whole training process.

4) Feature selection within the FFT approach: Reducing the number of EEG channels in an experiment from 32 to a lower subset, for example by utilizing the 10-20 standard [16], offers several advantages in terms of resource utilization and experimental efficiency. Firstly, by reducing the number of channels, the data acquisition and storage requirements are significantly minimized, resulting in optimized resource allocation. This reduction in data volume leads to reduced computational demands during preprocessing, analysis, and storage processes, saving valuable computational resources and memory. Moreover, it allows for more streamlined and focused data collection, reducing the overall experimental setup complexity and associated costs. Additionally, a smaller number of channels simplifies electrode placement and setup, minimizing the potential for electrode misplacements or artefacts that could affect data quality. This targeted approach in channel selection also facilitates the identification of specific brain regions of interest, enabling researchers to concentrate their analysis on the most relevant neural activity patterns. Thus, by strategically reducing the number of EEG channels, researchers can achieve resource-efficient experimentation without compromising the quality and integrity of the acquired data. Therefore, a reduced subset of 16 electrodes is suggested based on the overlapping 32 system with the 10-20 standard. The selected subset therefore is: Fp1, Fp2, F7, F3, F4, F8, T7/T3, C3, C4, T4/T8, T5/P7, P3, P4, T6/P8, O1, O2. Lastly, to check for statistical significance a statistical test on the classification results was conducted. This test showed if there exists a statistical difference between the results on all 32 channels with the results of the reduced subset of 16 channels.

IV. Results and Discussion

A. Baseline

The Baseline model comprises both established baseline classification results from prior studies and the baseline results obtained within this research. By incorporating these two components, a comprehensive foundation for performance assessment and comparative analysis is established. These values were used respectively to compare the data augmentation. The Baseline values can be seen in the tables throughout this paper denoted as $x1$.

B. Verification: Quality of the generated data

1) K-fold cross evaluation: As previously described the four sub-experiments performed with the 32 k-fold cross-evaluation can be evaluated in two parts in which the accuracies for the real and synthetic test are compared while training on the same data set. The objective was to ensure that the accuracies on the respective test sets, whether real or synthetic, achieve a similar level of performance. The accuracies can be found in table VIII and IX in Appendix VII-B. From the table, one can observe that the accuracies are all in a margin of around 5% which is comparable and therefore supports that the synthetic dataset is showing a similar structure to the real dataset. The valence and dominance accuracies for the synthetic-real and the real-real split, however, are below 50% which could be due to a bad splitting of the training and test set. Another interesting observation found was that for both experiments the score on the class dominance was off by 5%, further investigation into the data and insights of how the dominance class is being represented could reveal more information. To test for statistical significance between the obtained classification results on the dataset trained on the real data and synthetic data a test was performed. Since the two samples did not have approximately the same variance, a Welch's t-test was conducted. The test had the following H_0 hypothesis: *The mean performance score on the real dataset is equal to the mean performance score on the synthetic dataset*. Therefore, the alternative hypothesis stated the opposite, that both performance scores are not equal. The p-value for this test was 0.96, since the p-value is higher than the chosen significance value of $\alpha = 0.05$ there is no reason to reject the H_0 hypothesis. This means that the obtained classification results are similar supporting that augmentation should lead to classification improvement. The full test with its assumptions checks can be found in Appendix VII-C.

TABLE II: Maximal improvements for GANs model

	Valence	Arousal	Dominance	Liking
Max improvement	5	3.75	6.25	17.5

2) Distribution plots: To evaluate the distributions and compare the trend of real and synthetic data an external library was used. Synthetic Data Metrics (SDMetrics) [17] is an open-source Python library for evaluating tabular synthetic data. The metric resulted in an overall quality score of 87.58% by applying the previous GANs approach. By transforming the preprocessed data with the FFT into the frequency domain a quality score of 77.1% was reached. The detailed scores can be seen in the table X in Appendix VII-F as well as an example distribution plot for both approaches in figure 4 and figure 5 in appendix VII-D. Based on these results we see that the generated data with the previous GANs approach results in a higher quality score so we would expect a higher increase throughout the data augmentation process for this approach.

C. Verification: Performance of model with data augmentation

1) Comparison with previous GAN model: Before the results for the DM approach are represented the baseline has to be established with the results of a comparative approach using GANs [2]. In this work, a maximal improvement was reached between 3.75% to 17.5%. All improvements according to their classes can be found in table II. This table is summarising the experiments with KNN, SVM and ANN as classifiers. Any classifier performed significantly better than the others, depending on the predicted emotion it would suit to choose the classifier. The detailed results can be found in table XI, XII, and XIII in the appendix VII-F. The same approach with the discussed DM as underlying generative AI resulted in the maximal improvements seen in table III. To evaluate which loss, $l1$ or $l2$, should be used a small experiment was set up in which both losses were tested against each other. The l2 loss was used since it showed a faster and more stable convergence, see figure 2 in the appendix VII-A. In summary, the DM model reached astonishing results in the dominance class but lacks to leverage the accuracies throughout the other classes, especially Liking. It can be seen that the GANs model was able to leverage the accuracies to a higher extent throughout all classes. This was expected by the distribution plots' score values discussed previously. The detailed accuracies, for both models with 1200 and 100000 training steps, can be found in the tables XIV, XV, XVI, XVII, XVIII, and XIX in the appendix VII-F.

2) STFT approach: Experiments were initially conducted using the STFT approach. However, the obtained results from these experiments did not demonstrate substantial promise or

TABLE III: Maximal improvements for DM model compared with GANs approach

	Valence	Arousal	Dominance	Liking
Max improvement	3.5	3.5	16	0

TABLE IV: Accuracies for STFT approach on actual data

	Valence	Arousal	Dominance	Liking
KNN	56.49	56.75	69.75	62.5
SVM	59.25	56.25	73.00	62.5
ANN	51.75	51.25	59.25	59.25

satisfactory outcomes. The results obtained for each respective class were around an accuracy of 51.25% up to 73% which are around the same range as the previous experiment. For the detailed accuracies see table IV and table V. Therefore, a different approach used the FFT methodology was adopted as an alternative. This shift in approach was motivated by the desire to explore different avenues and potentially uncover more favourable results by leveraging the advantages offered by the FFT technique. By transitioning from the STFT to the FFT approach, the study aimed to enhance the efficacy and effectiveness of the analysis, ultimately contributing to a more comprehensive understanding of the underlying data patterns and trends.

3) FFT approach: At the beginning the two losses were compared, $l1$ and $l2$, the loss graphs can be seen in figure 3 in the appendix VII-A. From the graph, it was clear that the $l2$ loss converged faster and reached lower levels therefore, the $l2$ loss was used. The classification results obtained with the Fast Fourier Transformation were more promising and leveraged the classification on the actual dataset without augmentation by up to 20%, for a more detailed look the values can be found in table V and VI.

Despite the inherent advantages offered by the augmented data in relation to the actual data, the classification scores achieved using the augmented dataset did not exhibit significant levels of promise or discernible improvements. The highest improvement could be made for the class dominance by 4.5%. The remaining improvements can be seen in table VII. During this experiment, the ANN was able to make use of the augmentation the most, as three out of four maximal improvements came from the ANN as classifier. The full experiment can be seen in tables XX, XXI, XXII, XXIII, XXIV, and XXV in Appendix VII-F. In summary, it can be said that the transformation of the signal into the frequency domain results in higher accuracies during the classification but the DM seems to struggle more than the previous approach, to generate synthetic data that can leverage classification.

TABLE V: Classification score on statistical features without argumentation (stratified from Bhat and Hortal [2])

	Valence	Arousal	Dominance	Liking
KNN	57.50	56.75	59.00	64.00
SVM	62.50	53.75	53.50	62.75
ANN	58.00	55.75	57.25	53.50

TABLE VI: Classification scores on FFT approach without augmentation

	Valence	Arousal	Dominance	Liking
KNN	78.25	64.50	66.50	75.00
SVM	77.75	63.00	73.00	75.00
ANN	76.75	63.00	62.50	74.00

TABLE VII: Maximal improvements for DM model in frequency domain

	Valence	Arousal	Dominance	Liking
Max improvement	1	1.5	4.5	1

4) Feature selection within the FFT approach: The application of the reduced subset approach yielded comparable accuracies to the non-reduced approach, indicating that the reduced subset of features effectively captured the essential discriminative information necessary for accurate classification. Despite the valence value for the KNN approach all classification scores on the actual data were as high or even higher on the reduced subset. However, despite employing data augmentation techniques aimed at enhancing the classification performance, the results did not demonstrate a substantial improvement in classification accuracy. This observation suggests that the augmentation methods implemented in this study failed to significantly enhance the discriminatory power of the feature space for the reduced subset, leading to minimal impact on the overall classification performance. The findings highlight the need for further exploration and refinement of augmentation strategies to achieve more substantial improvements in classification outcomes. Despite the lack of improvement during the augmentation process, it is visible that a reduced channel approach can be used without losing information. The obtained accuracies can be found in the tables XXVI, XXVII, XXVIII, XXIX, XXX, and XXXI in Appendix VII-F. The conducted t-test, which should test for statistical significance between the classification results on the full channels and reduced channels, had the following H_0 hypothesis: *The mean performance score on the whole dataset is equal to the mean performance score on the reduced dataset*. Therefore, the alternative hypothesis stated the opposite, that both performance scores are not equal. The p-value for this test was 0.52, since the p-value is higher than the chosen significance value of $\alpha = 0.05$ there is no reason to reject the H_0 hypothesis. This means that the obtained classification results are similar supporting that the reduced subset of features is representing the whole data distribution. The full t-test with its assumptions checks can be found in Appendix VII-E.

V. Research Question

In this work the following research questions are addressed:

1) **Are diffusion based approaches better than other augmentation models, for example GAN's?**
To answer this research question the classification results from the augmentation with the GAN approach have to be compared with the DM approach. The GAN approach yielded improvements for the emotions: valence, arousal, dominance, and liking of 5%, 3.75%, 6.25%, and 17.5% respectively. The DM with the same underlying feature extraction approach on the other hand achieved improvements of 3.5%, 3.5%, 16%, and 0% respectively. Although the DM achieved a very high improvement for the dominance classification it lacks in performance for the other emotion types. Therefore, in this case, the DM model approach is not as good as the GAN alternative.
2) **Which features are more suitable for a diffusion model based data augmentation approach?**
A total of three different approaches were tested in this study. The first is the comparison approach to the previously established GAN methodology, the second moving from the time domain into the frequency domain with a STFT approach, and the last moving into the frequency domain with a FFT. Since the highest accuracies were reached with the last approach, transforming the signal into the frequency domain with a FFT, the features from this technique seem most promising to classify emotions on EEG signals. In summary, extracting features over a whole signal in the frequency domain yielded the best results and turned out to outperform feature extraction with a slicing window approach as done in the STFT approach.
3) **What are good features to classify EEG signals, with respect to feature importance**
The use of the reduced feature approach resulted in comparable outcomes to those obtained with the complete feature space, indicating that the selected subset of features effectively captured the essential discriminative information. This finding suggests that the reduced feature approach presents a favourable tradeoff, as it offers comparable results while reducing the computational burden and complexity associated with the analysis. Similar results achieved using the reduced feature approach imply that the selected features retain the essential characteristics required for accurate classification. In summary, reducing the features from 32 to 16 with a 10-20 approach, achieved comparable classification results. Therefore, employing the reduced feature approach in further research endeavours provides a practical and efficient means of data analysis, enabling researchers to focus on the most informative features.

VI. Conclusion and Future Research

A. Limitations

Several limitations should be highlighted in this study. Firstly, there remains potential for further fine-tuning of the hyperparameters within the DM, which could lead to improved model performance. Despite efforts to optimize hyperparameters, a broader exploration of parameter configurations may reveal additional opportunities for enhancing the models efficacy. A more in-depth hyperparameter search could have been in the number of batch sizes, the gradient accumulation steps, and the exponential moving average decay. Secondly, due to limited data availability, it was not feasible to use distinct datasets for training the DM and providing input data to calculate the classification. Although this practice is recommended to assess the generalization capabilities of the DM, the use of similar data for both purposes introduces the potential for biased results. The used DM implementation required the input data to be stored as an image. To accomplish this the dataloader was saved as an image separately. This additional saving step had the effect of floating-point precision which lead to information loss. A more elegant method would have been if the input of the model would have been adapted to take the already at-hand dataloader. To establish a more direct and meaningful comparison with research question two, it is recommended that future investigations rerun the GAN model with the same number of channels (32). Lastly, the k-fold approach to verify the similarity between the real and synthetic data should have been included for all different feature extraction methods. By aligning the number of channels, a more accurate and comprehensive assessment of the comparative performance between the DM in the frequency domain as well as the GAN in the frequency domain can be achieved, leading to more robust conclusions.

B. Future Research

Based on the insights gained from the findings and limitations encountered in this study, several directions for future research are proposed to propel the field of analysis forward and enhance classification performance. Firstly, exploring the use of Variational Diffusion Models (VDM), as outlined by Lou [18] could provide a valuable alternative for comparison with GANs. VDMs offer a different approach to generative modelling and may yield more insightful comparisons and insights into the generation of synthetic data. Additionally, further investigation into hyperparameter fine-tuning for the DM is recommended, as stated in the limitations section. While efforts were made to optimize the hyperparameters in this study, the exploration of a broader range of configurations and more comprehensive tuning strategies could lead to enhanced performance and more accurate synthetic data generation. Furthermore, considering the reduced subset approach, it would be beneficial to explore alternative feature extraction methods. The investigation of different techniques, such as wavelet transforms or principal component analysis, could provide valuable insights into the effectiveness of feature selection and its impact on classification accuracy. Another possible venue would be the use of spectrograms as input for the diffusion model. This exploration would contribute to a more comprehensive understanding of feature extraction approaches and their applicability in the context of synthetic data generation. Overall, future research endeavours should strive to improve the understanding and performance of generative models, considering alternative methodologies, fine-tuning approaches, and exploring diverse feature extraction techniques. Addressing these aspects, researchers can advance the field and enhance the effectiveness of synthetic data generation for various applications.

References

[1] Björn W Schuller. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61(5):90–99, 2018.

[2] Sudhanva Bhat and Enrique Hortal. Gan-based data augmentation for improving the classification of eeg signals. In *The 14th PErvasive Technologies Related to Assistive Environments Conference*, pages 453–458, 2021.

[3] Yoky Matsuoka, Pedram Afshar, and Michael Oh. On the design of robotic hands for brain–machine interface. *Neurosurgical focus*, 20(5):1–9, 2006.

[4] Robert Jenke, Angelika Peer, and Martin Buss. Feature extraction and selection for emotion recognition from eeg. *IEEE Transactions on Affective computing*, 5(3):327–339, 2014.

[5] Danny Oude Bos et al. Eeg-based emotion recognition. *The influence of visual and auditory stimuli*, 56(3):1–17, 2006.

[6] Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering*, 16(5):051001, 2019.

[7] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

[8] Elnaz Lashgari, Dehua Liang, and Uri Maoz. Data augmentation for deep-learning-based electroencephalography. *Journal of Neuroscience Methods*, 346:108885, 2020.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[10] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011.

[11] Yisi Liu, Olga Sourina, and Minh Khoa Nguyen. Real-time eeg-based emotion recognition and its applications. *Lecture Notes in Computer Science*, page 256–277, 2011.

[12] Szabolcs Torma and Luca Szegletes. Brain signal generation and data augmentation with a single-step diffusion probabilistic model. 2023.

[13] Anwaar Ulhaq, Naveed Akhtar, and Ganna Pogrebna. Efficient diffusion models for vision: A survey. *arXiv preprint arXiv:2210.09292*, 2022.

[14] Lucidrains. Lucidrains/denoising-diffusion-pytorch: Implementation of denoising diffusion probabilistic model in pytorch.

[15] Haifeng Jin, Qingquan Song, and Xia Hu. Auto-keras: An efficient neural architecture search system. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1946–1956, 2019.

[16] Herbert H Jasper. Ten-twenty electrode system of the international federation. *Electroencephalogr. Clin. Neurophysiol.*, 10:371–375, 1958.

[17] DataCebo. Sdmetrics.

[18] Calvin Luo. Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*, 2022.

VII. Appendix

A. Loss comparison

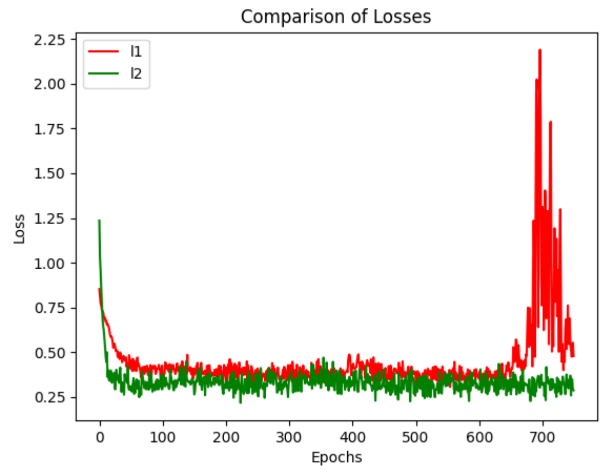

Fig. 2: Loss comparison with GAN approach features

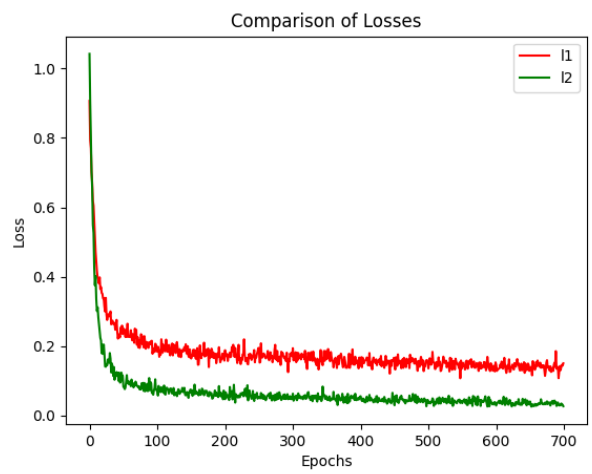

Fig. 3: Loss comparison with FFTs approach features

B. Tables for the k-fold cross evaluation

TABLE VIII: (32) K-fold cross evaluation

	Valence	Arousal	Dominance	Liking
Train Real Test Real	47.19	54.69	51.41	60.55
Train Real Test Synthetic	52.66	53.91	56.02	60.39

TABLE IX: (32) K-fold cross evaluation

	Valence	Arousal	Dominance	Liking
Train Synthetic Test Real	47.89	51.25	49.84	64.38
Train Synthetic Test Synthetic	50.86	53.98	55.08	62.50

C. Welch's t-test on the classification results of the real and synthetic dataset

For testing the statistical significance of both classification results, with a significance value $\alpha = 0.05$, the following scores were used:

- Training on real data: 47.1875, 54.6875, 51.40625, 60.546875, 52.65625, 53.90625, 56.015625, 60.390625
- Training on synthetic data: 47.890625, 51.25, 49.84375, 64.375, 50.859375, 53.984375, 55.078125, 62.5

These results can be obtained from tables VIII and IX in appendix VII-B

Therefore the H_0 and H_1 hypothesis are:

H_0: *The mean performance score on the real dataset is equal to the mean performance score on the synthetic dataset*

H_1: *The mean performance score on the real dataset is not equal to the mean performance score on the synthetic dataset*

The following assumptions have to hold for a t-test:

1) The observations in one sample should be independent of the observations in the other sample.
2) The data in both samples was obtained using a random sampling method.
   a) The two statements above are given for free in this case.
3) The data should be approximately normally distributed.
   a) To test normality a Shapiro Wilk test was performed on both data series. They resulted in a $p-value$ of 0.7310628294944763 and 0.18823620676994324 respectively. Since both $p-values$ are larger than 0.05 we can assume normality.
4) The two samples should have approximately the same variance. If this assumption is not met, a Welch's t-test should be performed.
   a) The variances are 17.539119720458984 and 31.411361694335938 respectively, since these are not approximately the same a Welch's t-test was performed.

The obtained result of the Welch's t.-teest was a statistic value of 0.04800803518032314 and a $pvalue$ of 0.9624418239297778. Since the $pvalue$ is > 0.05 we have no reason to reject the H_0 hypothesis.

D. Results with SDMetrics and distribution plots

TABLE X: Quality Scores with SDMetrics

	GANs aproach	FFT approach
Overall	87.58	77.1
Column Shapes	82.47	75.88
Column Pair Trends	92.69	78.33

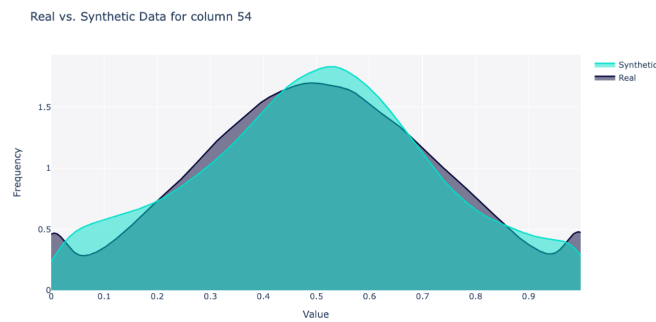

Fig. 4: GAN approach feature 54.

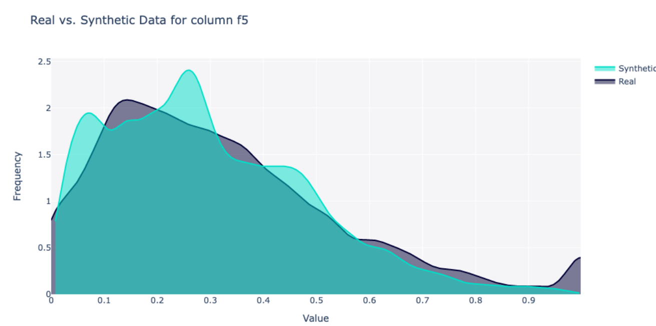

Fig. 5: FFT approach feature 5.

E. T-test on the classification results of the full dataset and the reduced dataset

For testing the statistical significance of both classification results, with a significance value $\alpha = 0.05$, the following scores were used:

- Training on full data: 78.25, 65.50, 67.50, 75.50, 78.75, 63, 73, 75, 76.75, 64.50, 67, 75
- Training on reduced data: 79.75, 65, 69.50, 75, 78.75, 65, 73, 75, 76.75, 63, 66.5, 75

These results can be obtained from tables XX, XXI, XXII, XXIII, XXIV, XXV, XXVI, XXVII, XXVIII, XXIX, XXX, and XXXI in appendix VII-F. For each classifier the highest obtained score was used.

Therefore the H_0 and H_1 hypothesis are:

H_0: *The mean performance score on the whole dataset is equal to the mean performance score on the reduced dataset*

H_1: *The mean performance score on the whole dataset is not equal to the mean performance score on the reduced dataset*

The following assumptions have to hold for a t-test:

1) The observations in one sample should be independent of the observations in the other sample.
2) The data in both samples was obtained using a random sampling method.
   a) The two statements above are given for free in this case.
3) The data should be approximately normally distributed.
   a) To test normality a Shapiro Wilk test was performed on both data series. They resulted in a $p - value$ of 0.11654853075742722 and 0.23475410044193268 respectively. Since both $p-values$ are larger than 0.05 we can assume normality.
4) The two samples should have approximately the same variance. If this assumption is not met, a Welch's t-test should be performed.
   a) The variances are 30.14019097222223 and 30.952690972222225 respectively, since these are approximately the same a t-test was performed.

The obtained result was a statistic value of -0.670881385239266 and a $pvalue$ of 0.5161354818838517. Since the $pvalue$ is > 0.05 we have no reason to reject the H0 hypothesis.

F. Experiment tables for augmentation classification

TABLE XI: GANs approach using KNN

Using KNN				
Data	Valence	Arousal	Dominance	Liking
x1	57.50	56.75	59.00	64.00
x2	61.75	57.00	**61.50**	**71.50**
x3	62.00	59.00	61.25	70.25
x4	**62.50**	**59.25**	61.00	70.50
Max Improvement	5	2.5	2.5	7.5

TABLE XII: GANs approach using SVM

Using SVM				
Data	Valence	Arousal	Dominance	Liking
x1	**62.50**	53.75	53.50	62.75
x2	60.25	55.75	**59.50**	**70.50**
x3	60.50	**57.50**	59.75	70.50
x4	59.75	57.00	58.25	70.50
Max Improvement	0	3.75	6.25	7.75

TABLE XIII: GANs approach using ANN

Using ANN				
Data	Valence	Arousal	Dominance	Liking
x1	58.00	55.75	57.25	53.50
x2	59.25	**59.00**	**59.75**	**71.00**
x3	**61.50**	59.00	59.50	70.50
x4	58.25	57.75	59.25	70.50
Max Improvement	3.50	3.25	2.5	17.5

TABLE XIV: DM approach using KNN with 1200 training steps

Using KNN with 1200 training steps				
Data	Valence	Arousal	Dominance	Liking
x1	53.75	56.75	58.75	**63.00**
x2	**57.25**	**57.75**	68.25	63.00
x3	55.75	55.25	**69.75**	62.75
x4	56.00	55.75	66.50	62.50
Max Improvement	3.5	1	11	0

TABLE XV: DM approach using SVM with 1200 training steps

Using SVM with 1200 training steps				
Data	Valence	Arousal	Dominance	Liking
x1	**59.25**	56.00	**73.00**	**62.50**
x2	59.25	**57.00**	73.00	62.50
x3	59.25	57.25	73.00	62.50
x4	59.25	56.00	73.00	62.50
Max Improvement	0	1	0	0

TABLE XVI: DM approach using ANN with 1200 training steps

Using ANN with 1200 training steps				
Data	Valence	Arousal	Dominance	Liking
x1	**59.25**	52.50	57.00	**62.50**
x2	59.25	**56.00**	**73.00**	62.50
x3	59.00	53.00	62.75	62.50
x4	59.25	56.00	71.50	62.50
Max Improvement	0	3.5	16	0

TABLE XVII: DM approach using KNN with 100000 training steps

Using KNN with 100000 training steps				
Data	Valence	Arousal	Dominance	Liking
x1	53.75	56.75	**58.75**	**63.00**
x2	52.50	56.00	52.25	63.00
x3	**54.50**	**59.50**	52.75	62.75
x4	54.00	58.50	53.25	62.50
Max Improvement	0.75	2.75	0	0

TABLE XVIII: DM approach using SVM with 100000 training steps

Using SVM with 100000 training steps				
Data	Valence	Arousal	Dominance	Liking
x1	59.25	**56.00**	**73.00**	**62.50**
x2	59.25	56.00	73.00	62.50
x3	**59.50**	56.00	73.00	62.50
x4	59.25	56.00	73.00	62.50
Max Improvement	0.25	0	0	0

TABLE XIX: DM approach using ANN with 100000 training steps

Using ANN with 100000 training steps				
Data	Valence	Arousal	Dominance	Liking
x1	**59.25**	52.50	57.00	**62.50**
x2	59.25	**56.00**	**72.50**	62.50
x3	59.00	56.00	72.50	62.50
x4	59.25	56.00	72.00	62.50
Max Improvement	0	3.5	15.5	0

TABLE XX: DM in FFT domain using KNN with 12000 training steps

Using KNN				
Data	Valence	Arousal	Dominance	Liking
x1	**78.25**	64.50	66.50	75.00
x2	77.75	64.50	67.00	75.00
x3	78.25	64.50	**67.50**	**75.50**
x4	77.75	**65.50**	67.00	75.50
Max Improvement	0	1	1	0.5

TABLE XXI: DM in FFT domain using SVM with 12000 training steps

Using SVM				
Data	Valence	Arousal	Dominance	Liking
x1	77.75	63.00	73.00	75.00
x2	**78.75**	63.00	73.00	75.00
x3	78.75	63.00	73.00	75.00
x4	78.75	63.00	73.00	75.00
Max Improvement	1	0	0	0

TABLE XXII: DM in FFT domain using ANN with 12000 training steps

Using ANN				
Data	Valence	Arousal	Dominance	Liking
x1	76.75	63.00	62.50	74.00
x2	76.75	63.00	64.50	**75.00**
x3	76.75	63.00	**66.50**	74.00
x4	76.75	**64.50**	65.50	74.00
Max Improvement	0	1.5	4	1

TABLE XXIII: DM in FFT domain using KNN with 100000 training steps

Using KNN				
Data	Valence	Arousal	Dominance	Liking
x1	78.25	64.50	66.50	75.00
x2	78.25	**65.00**	**67.50**	75.00
x3	78.25	64.50	66.00	75.00
x4	78.25	64.50	67.50	75.00
Max Improvement	0	0.5	1	0

TABLE XXIV: DM in FFT domain using SVM with 100000 training steps

Using SVM				
Data	Valence	Arousal	Dominance	Liking
x1	77.75	63.00	73.00	75.00
x2	**78.75**	63.00	73.00	74.00
x3	78.75	63.00	73.00	74.00
x4	77.75	63.00	73.00	74.00
Max Improvement	1	0	0	0

TABLE XXV: DM in FFT domain using ANN with 100000 training steps

Using ANN				
Data	Valence	Arousal	Dominance	Liking
x1	76.75	63.00	62.50	74.00
x2	76.75	**64.50**	**67.00**	74.00
x3	76.75	64.50	62.50	74.00
x4	76.75	63.00	65.50	**75.00**
Max Improvement	0	1.5	4.5	1

TABLE XXVI: Reduced set with DM in FFT domain using KNN with 12000 training steps

Using KNN				
Data	Valence	Arousal	Dominance	Liking
x1	77.75	65.00	69.50	75.00
x2	77.75	64.50	67.50	75.00
x3	77.75	64.50	68.50	75.00
x4	77.25	64.50	67.50	75.00
Max Improvement	0	0	0	0

TABLE XXVII: Reduced set with DM in FFT domain using SVM with 12000 training steps

Using SVM				
Data	Valence	Arousal	Dominance	Liking
x1	77.75	65.00	73.00	75.00
x2	76.75	64.50	73.00	75.00
x3	76.75	64.50	73.00	75.00
x4	**78.75**	63.50	73.00	75.00
Max Improvement	1	0	0	0

TABLE XXVIII: Reduced set with DM in FFT domain using ANN with 12000 training steps

Using ANN				
Data	Valence	Arousal	Dominance	Liking
x1	76.75	63.00	62.50	74.00
x2	76.75	**64.50**	**66.50**	**75.00**
x3	76.75	63.00	63.00	74.00
x4	76.75	63.00	64.50	75.00
Max Improvement	0	1.5	4	1

TABLE XXIX: Reduced set with DM in FFT domain using KNN with 100000 training steps

Using KNN				
Data	Valence	Arousal	Dominance	Liking
x1	77.75	65.00	69.50	75.00
x2	**79.75**	65.00	67.50	75.00
x3	78.25	64.50	67.50	75.00
x4	77.25	64.50	67.50	75.00
Max Improvement	2	0	0	0

TABLE XXX: Reduced set with DM in FFT domain using SVM with 100000 training steps

Using SVM				
Data	Valence	Arousal	Dominance	Liking
x1	77.75	65.00	73.00	75.00
x2	77.75	64.50	73.00	75.00
x3	77.75	63.50	73.00	74.00
x4	77.75	64.50	73.00	74.00
Max Improvement	0	0	0	0

TABLE XXXI: Reduced set with DM in FFT domain using ANN with 100000 training steps

Using ANN				
Data	Valence	Arousal	Dominance	Liking
x1	76.75	63.00	62.50	74.00
x2	76.75	63.00	**66.50**	**75.00**
x3	76.75	63.00	63.00	74.00
x4	76.75	63.00	63.50	75.00
Max Improvement	0	0	4	1