GAN-Based Data Augmentation For Improving The Classification Of EEG Signals

GAN-Based Data Augmentation For Improving The Classification Of EEG Signals

Sudhanva Bhat*
sudhanva.nd@gmail.com
Maastricht University, Department of Data Science and Knowledge Engineering
The Netherlands

Enrique Hortal*
enrique.hortal@maastrichtuniversity.nl
Maastricht University, Department of Data Science and Knowledge Engineering
The Netherlands

ABSTRACT

Emotion recognition is a field of psychology that involves the process of identifying emotions and treating mental conditions like autism. The advancements in the field of machine learning and deep learning have paved the way for scientists to develop models for evaluating emotions by analyzing facial expressions, speech and text. However, the task of evaluating emotions could be best done by processing the bio-signals and neural imaging of the brain. In that sense, bio-signals such as Electroencephalogram (EEG) are less expensive to use and non-invasive, giving them an edge over traditional methods like Magnetic Resonant Imaging (MRI). However, not many datasets are publicly available due to privacy issues and their availability is highly limited by the classification task. These constraints, along with the problem of data scarcity, motivates this work as an attempt to enhance the accuracy scores by generating synthetic features that are close to actual data distribution. In this research, we propose a Wasserstein Generative Adversarial Network with gradient penalty (WGAN-GP) based model that can help tackle this problem. The dataset that is investigated is DEAP, one of the benchmark datasets for evaluating emotion recognition algorithms. In the method proposed, nine descriptive features are extracted from the original data and baseline models are evaluated. Subsequently, a WGAN-GP is trained on these extracted features and it is used to generate a new set of synthetic data features. The synthetic features are then analysed for quality and appended to the original data to expand this dataset. Experiments with different augmentation factors (x2, x3, x4) are investigated to evaluate the impact of the data augmentation procedure. The experimental results demonstrate that the proposed method gives a considerable enhancement of the classification task's performance.

KEYWORDS

emotion recognition, eeg, data augmentation, generative adversarial networks

*Both authors contributed equally to this research.

PETRA 2021, June 29-July 2, 2021, Corfu, Greece

ACM ISBN 978-1-4503-8792-7/21/06…$15.00
https://doi.org/10.1145/3453892.3461338

ACM Reference Format:
Sudhanva Bhat and Enrique Hortal. 2021. GAN-Based Data Augmentation For Improving The Classification Of EEG Signals. In *The 14th PErvasive Technologies Related to Assistive Environments Conference (PETRA 2021), June 29-July 2, 2021, Corfu, Greece.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3453892.3461338

1 INTRODUCTION

Bio-signals such as Electroencephalogram (EEG) or Electrocardiogram (ECG) help to map the activity of the brain and other vital organs. They are less expensive and non-invasive, giving them an edge over traditional methods like Magnetic Resonance Imaging (MRI) and other invasive techniques like extracellular Action Potentials (APs) or the local field potentials (LFPs). EEG is used to register the activity of the brain by placing electrodes on the scalp at various locations. Among other applications, EEG signals are used to detect brain abnormalities by inspection of the signal and measuring the deviation from normal signal patterns.

Emotions play a vital role in human communication. In earlier days, facial expressions were used to detect emotions. However, humans have the ability to suppress facial emotions while EEG signals are not susceptible to such suppression, as the data is collected directly from the brain. Therefore, the use of EEG signals as a means of classifying emotions has been extensively researched [6][2].

In the past decade, researchers have placed tremendous importance on the process of automating emotion recognition [11]. The advancements in machine learning, particularly deep learning, have sparked new ways to classify signals and develop models which offer better performance. However, the progress of deep learning is being hindered by the availability of data. This is especially notorious in tasks involving private or confidential information as is the case of medical data. Regarding the EEG recordings, data is highly sensitive as they belong to the individuals' bio-signal recordings undergoing experiments in controlled environments. Additionally, medical data is often very task-specific. In the case of EEG, for example, the characteristics of the data are subject dependent with differences due to variables like probe placement, size of the head, location of the experiment and brain anatomy [10]. Additionally, access to such data is highly restricted and not many datasets are publicly available due to privacy issues. Additionally, it is highly difficult to compare model performance with similar available datasets, leading to research bottlenecks. Moreover, as above-mentioned, the training data is subject dependent in most cases, meaning that it is difficult to get similar model performance on other test subjects. Finally, creating such data using lab driven

experiments is a demanding task as it needs expert knowledge to extract and annotate the data along with specialized equipment. Preliminary data augmentation is generally performed to increase the spread of data. Nonetheless, the results obtained from the model fed with such data is not so remarkable. The drawbacks mentioned above are the main motivation for this research.

1.1 Data Augmentation

Data Augmentation is a technique for increasing the existing dataset by adding new samples, modifying/transforming existing ones. Data augmentation helps to curb the problem of over-fitting by giving it an effect of regularization. This makes the models generalize better to the patterns of the data [9] and thereby helps to improve the stability and accuracy. Data augmentation is a very common technique that is used, for instance, in the field of computer vision, particularly in object/image recognition. For images, the dataset is augmented by applying transformations like rotation, translation, scaling, cropping, padding and flipping. However, EEG is a time-series signal. Applying such geometric transformations would undermine the time-dependent features. Additionally, creating labels for the transformed data is not a straightforward task, due to the fact that EEG data is not visually informative as it is the case of images.

In the last decade, we have seen tremendous effort in the field of data augmentation. Data augmentation in the context of EEG is used in various tasks like emotion recognition, seizure detection, motor task, sleep stage detection, visual task, mental workload, etc. In these aforementioned tasks, augmentation methods like noise addition, generative models, sliding window, sampling, Fourier transform, etc. have been used. Each method is task-dependent meaning that certain methods work well only for a particular task. A detailed study of the available methods is done by the researchers Lashgari et al [6].

1.2 Problem Statement

The classification of EEG signals is challenging and often, does not yield great results. This paper focuses on the usability of Generative Adversarial Networks (GANs) in generating augmented dataset (synthetic data) to improve classification performance for emotion recognition tasks. In this regard, this research aims at generating synthetic medical data by modelling a Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP).

2 DATA

DEAP is a dataset for emotion analysis using EEG, physiological and video signals [5]. From the time the dataset was made available, many researchers have experimented on it. Early researchers like Liu and Sourina [7] exploited the property that EEG is a nonlinear and multi-fractal signal, hence its fractal decomposition would give features that can be used to train machine learning algorithms like Support Vector Machines (SVM). Their results reported an accuracy of 53.7% on the classification of emotions. A very recent study published by Luo and Lu [8] used a Conditional Wasserstein GAN (CWGAN) to generate Data Entropy features instead of raw EEG data and used it to augment the data. Their research produced astonishing results with an accuracy of 78.17% and 73.89%, an increase of up to 9.15% and 20.13% for two emotion types.

Table 1: Insight of DEAP dataset

Array file	Shape
Data	40 x 40 x 8064 (#videos/trial x #channels x #data)
Labels	40 x 4 (#videos x #labels)

In this work, we investigate the applicability of WGAN-GP for data augmentation for the classification task in the DEAP dataset. The data is private and could only be accessed after signing a EULA. DEAP consists of EEG data from 32 participants, each of them watched 40 videos of 1-minute duration. The participants rated videos with an integer between 0-9 during the self-assessment phase. A set of four ratings were given for valence, arousal, dominance and liking respectively. In this study, 40 data channels were used to record the data from different locations on the scalp and monitor other parameters. In this research, we use the preprocessed data provided by the DEAP team which consists of a data file for each of the 32 participants down-sampled to 128Hz including both data and labels distributed as follows:

(1) An 8064 EEG datapoints array for each of the 40 experiments obtained from 40 different channels representing 322560 data points per experiment.
(2) A label array containing a set of four values for each experiment representing the emotions, namely valence, arousal, dominance, and liking.

Table 1 gives an insight into the data dictionary for each participant. The label information is a single integer value for each of the valence, arousal, dominance and liking classes. Figure 1 represents the distribution of the ratings for different emotion types. In the first three graphs, the mean is centered around 5, indicating that, in general, they have mixed emotions for each of the videos and are not completely sure of their ratings. Using this kind of label information to classify the tasks might result in weak performances. On the contrary, there is a mild skew centered around a score of 7 in the last (liking rate). The hypothesis is that the binary classification would work better on the liking class than on the other ratings, namely valence, arousal and dominance.

For a classification task, the integer classes should be converted into something more interpretable for the model. Each of the label data is therefore divided into two classes (being rates of 5 and below considered as class 0 and class 1 otherwise). This allows binary classification to be evaluated.

2.1 Data pre-processing

The feature dimensionality of the data at hand is high, consisting of 8064 features per experiment. The dimension of this data must be reduced to train the models in a quick and efficient way. To that end, various feature representation schemes were used in previous research. In [12], research gave state-of-the-art results in DEAP dataset classification. Therefore, we follow a similar approach, dividing the data (8064 data points, representing 1 minute of EEG recordings approximately) into 10 batches of approximately

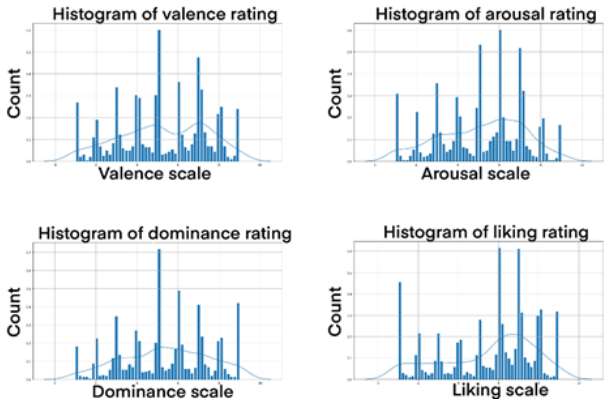

Figure 1: Gaussian distribution of rating types- valence, arousal, dominance and liking- for 1280 experiments (32 participants x 40 experiments). The y-axis represents the counts of the ratings and the X-axis represents the annotations and each graph is an emotion: valence, arousal, dominance and liking ratings respectively.

807 points (around 6-seconds). Then, for each batch, the following nine features are extracted: mean, median, maximum, minimum, standard deviation, variance, range, skewness and kurtosis. Consequently, 90 features are extracted per file and channel. Additionally, the above-mentioned features are calculated on the whole set of 8064 data points and appended to them, increasing the feature space from 90 to 99. The steps above are repeated for each of the 40 channels and the 32-participant data. This pre-processing task is specially important in this work due to the complex system implemented (GAN-based architecture) and the limited number of samples available.

3 METHODOLOGY

Conventional GANs have been proved to give more "realistic" outputs than other solutions such as autoencoders. However conventional GANs have two main problems. On the one hand, the generator and discriminator losses tend to oscillate leading to perturbations that indirectly affect the performance of the generated data. On the other hand, this architecture is prone to problems like mode collapse. Mode collapse is a problem wherein the generator finds a few samples that are able to fool the discriminator into recognising them as original data. Over time the generator would only produce samples from this limited space that can mislead the discriminator and would not learn further. Eventually, the gradient of the loss function will collapse to almost 0, meaning that the training is not able to learn (and consequently generate) more apart from this small sample space. Additionally, the conventional GANs have problems with hyperparameter tuning and convergence. They mostly fail to converge and thus, Wasserstein GAN (WGAN) or its counterpart with Gradient Penalty (WGAN-GP) has been the baseline/starting point of most of the current GAN-based architectures. In these models, a new metric called Earth-Mover distance (also called Wasserstein distance) replaces the Jensen-Shannon divergence of traditional GANs (whose discontinuity divergence makes it difficult to obtain gradients to train the GANs which makes the training process unstable). The Wasserstein distance metric provides a useful gradient during WGAN training times and also improves stability [1]. In this research we use WGAN-GP to overcome the above-mentioned limitations of conventional GANs.

The discriminator in WGAN is called a critic. The critic does not directly distinguish the fake samples from the real, but it is used to learn the weight parameters w of the K-Lipschitz continuous function. This loss function is a direct measure of the Wasserstein distance between real and generated distributions. So as the value of the loss function decreases, it indirectly implies that the Wasserstein distance between real data and generated data is close to zero, meaning that the generator's output is closer to that of real data distribution.

The training process of critic includes the key addition of gradient penalty loss to the original WGAN loss. The gradient penalty term is a squared difference between the norm of the gradient of predictions and 1. The model ensures that the Lipschitz constraint is met by finding the weights that minimize the gradient penalty term. It is complicated to calculate the gradient term at every point of the training process. Instead, the authors of the WGAN-GP proposed a solution to evaluate gradients only at a handful of points (interpolated images) which is the random average of the real data and generated data. These interpolated points lie on the line that connects the batch of real and fake data.

The proposed framework involves using a random noise as input to the WGAN-GP based generator and critic networks. The network is trained for 100 epochs and the data generated at the end of each epoch is stored. The WGAN loss is plotted and evaluated to find the most promising synthetic data that can be used for data augmentation. The hypothesis here is that the data with WGAN-GP losses (generator and discriminator/critic loss) close to zero is of good quality. This high-quality data is appended to the original one to be used to evaluate the classification model. The proposed framework for synthetic data generation is shown in Figure 2.

In order to train and evaluate the data generated from the WGAN-GP model, it is essential that data is split into two parts: training and test sets. In this regard, out of 32 subject data available, different training and test set ratios were evaluated internally to find a proper balance for augmenting datasets. Finally, to set up a baseline for evaluation, the 32 subjects from the DEAP dataset were shuffled and split in such a way that data from 22 subjects formed the training dataset. The remaining 10 subject data was used to test the model before and after data augmentation. This ensures that the model performance is checked on the same test dataset thereby allowing a fair comparison. To train and generate the data using WGAN-GP, the same 22-participant set was used instead of the whole dataset. Using only this 22-participant set to train WGAN-GP would prevent any biases and information leaking from train to test phases. The data generated using the WGAN-GP (with probability distribution close to the training data) is then appended to the original training set and this constitutes the new augmented training data that can be used to evaluate the performance of the augmented model. Figure 3 shows the whole procedure of splitting the data and data augmentation.

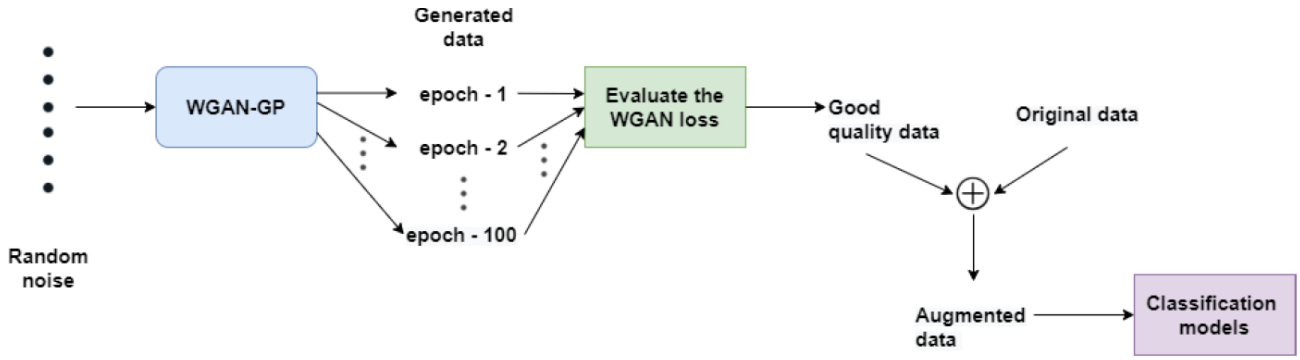

Figure 2: The proposed framework for synthetic data generation.

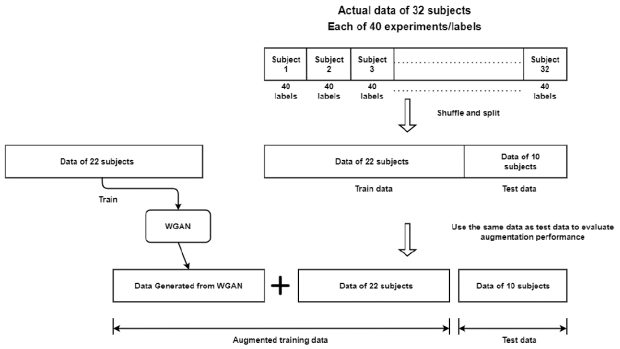

Figure 3: Train/test split strategy for DEAP dataset augmentation.

3.1 WGAN-GP

The critic and generator networks were tested for different numbers of layers and configurations. The generator takes in randomly sampled noise of length 256. It then yields a final output of the shape (40,99) representing each experiment of a particular subject (channels and features respectively).

The critic is crucial to calculate the Wasserstein loss and minimize it. In the critic model proposed, the input is data of shape (40,99) followed by a series of convolutional blocks. The final layer contains a single unit with a linear activation. The linear activation is used instead of a sigmoid activation to predict a realness score where -1 represents a real data point and 1 represents a fake data point.

The training process is executed till the WGAN-GP loss converges. Different epoch sizes (namely 100 and 200) were analyzed, obtaining similar results. Additionally, different batch sizes ranging from 32 to 128 were evaluated. The best results were obtained at batch sizes 32 and 40. In this report, the results executing 100 epochs and with a batch size of 40 are discussed.

3.2 Evaluation of generated data

As mentioned before, the WGAN-GP is trained for a set of 100 epochs, while the data generated after each epoch is saved. The generator and critic loss are plotted per epoch and the data is evaluated. The hypothesis is - the closer the loss to zero, the better is the quality of generated data. Due to the limited amount of data, there is not a third set to generate new samples after the training process. Instead, the data generated at epochs 89, 91, 97, 100 were chosen for augmenting the data by 2x, 3x, 4x factors. Then, the performance of the baseline model (without data augmentation) is compared against the data augmented models for the classification task at hand.

4 EXPERIMENTS

Two different experiments were conducted to 1) evaluate the quality of the data generated and 2) to evaluate our model and its performance in comparison with an established baseline.

4.1 Quality of the generated data

As an initial step to check if the data generated is of high quality, an experiment is set up to compare the performance of the model that is trained only on the real data in comparison with the performance of a model trained only on generated data. The hypothesis is that, if both datasets have a similar distribution, their relevant models should reach similar classification performances. K-fold cross-validation is the state-of-the-art approach when it comes to evaluating the performance of datasets with limited data. Hence, 32-fold cross-validation on the original data is compared against the 32-fold cross-validation score of the generated data. The closer the scores are, the more likely is that the generated and the actual data have a similar probability distribution.

4.2 Performance of model with data augmentation

Finally, data generated from the WGAN-GP is used to augment the existing dataset. To this end, classic machine learning models, namely K-nearest neighbors (KNN) and Support Vector Machines (SVM), were used to establish our baseline. The machine learning models were trained on the four abovementioned emotions -valence, arousal, dominance and liking. A grid search was applied to find the best parameters. KNN method with K value from 5-50 was evaluated for each class. SVM was also tested with different kernels: linear, RBF and polynomial functions. Best results were obtained

Table 2: Comparison of classification accuracy (in %) of original and synthetic data using K-fold cross-validation.

Data	Valence	Arousal	Dominance	Liking
Actual Data	55.16	57.03	59.61	66.80
Generated Data	55.16	56.09	60.86	66.41

by linear kernels. The parameter c was searched from range 1e-4 to 1e2.

Additionally, neural networks were also utilized to train on the actual data to establish an equitable baseline. For that purpose, a simple Convolutional Neural network that contains a series of Conv2D layers in succession was implemented. The network was optimized using autokeras [3] based grid-search to infer the best performance. An Adam optimizer with a default learning rate of 0.001 was applied.

Then, the data augmentation technique is evaluated. Several augmentation levels were considered (namely, x2, x3, x4) for the purpose of analysing the potential improvement in comparison with the original dataset. For example, x2 stands for the original data while adding once synthetic/generated set (same size as the original dataset).

5 RESULTS

In this section, the results of the experiments described in Section 4 are presented.

5.1 Quality of generated data

The results obtained after conducting the 32-fold cross-validation on real and generated data is shown in Table 2. The generated data gives similar classification performance across different emotions-valence, arousal, dominance and liking. The accuracies of the model range from 55% and around 70% approximately. For the valence class, the accuracy scores are identical. Regarding the dominance class, the score of generated data is slightly higher than the actual data. Finally, for arousal and liking, the accuracy scores are slightly lower for the generated data. From the results in Table 2, it is evident that the performances are not significantly different.

5.2 Performance of the model with data augmentation

Tables 3, 4 and 5 shows the values of each baseline and augmented model results of the models using KNN, SVM and simple Convolutional Neural networks respectively. The second row of each table indicates the performance of the baseline models (labelled as X1). Every next row shows the accuracy scores of augmentation levels x2, x3 and x4 respectively. The bold numbers indicate the maximum accuracy score across the three augmentation levels. The last row of the table shows the maximum performance improvement which is the difference between the baseline score and the maximum score that is encountered across the augmentation levels. All the scores mentioned in the table are represented as a percentage. The overall results clearly show an increase in the accuracy scores across different emotions after data augmentation. In terms of percentage of improvement, the neural network-based models performed better due to the excess data available. It is a known fact that deep learning models learn better when more data is available. The neural networks have shown a noteworthy increase in scores of up to 17.5%. On the other hand, all models tend to perform better on the liking class as expected in the hypothesis posted before.

Table 3: Comparison of KNN accuracies of original and augmented data.

Data size	Valence	Arousal	Dominance	Liking
X1(Actual Data)	57.5	56.8	59.0	64.0
X2	61.8	57.0	**61.5**	**71.5**
X3	62.0	59.0	61.3	70.3
X4	**62.5**	**59.3**	61.0	70.5
Max. improv	**5.0**	**2.5**	**2.5**	**7.5**

Table 4: Comparison of SVM accuracies of original and augmented data.

Data size	Valence	Arousal	Dominance	Liking
X1(Actual Data)	62.5	53.8	53.5	62.8
X2	60.3	55.8	59.5	**70.5**
X3	60.5	**57.5**	**59.8**	**70.5**
X4	59.8	57.0	58.3	**70.5**
Max. improv.	**0.0**	**3.75**	**6.25**	**7.75**

Table 5: Comparison of neural network accuracies of original and augmented data.

Data size	Valence	Arousal	Dominance	Liking
X1(Actual Data)	58.0	55.8	57.3	53.5
X2	59.3	**59.0**	**59.8**	**71.0**
X3	**61.5**	59.0	59.5	70.5
X4	58.3	57.8	59.3	70.5
Max. improv	**3.5**	**3.25**	**2.5**	**17.5**

The KNN outperforms every other model by yielding max scores of 62.5%, 59.25%, 61.5% and 71.5% respectively. It is interesting to see that for SVM, there is no improvement across valence class but it increased the performance for dominance and liking by 6.25% and 7.75 respectively. To deduce a general trend in the performance of models, an augmentation of 2x would lead to more promising results. After 2x augmentation, the next best performances were obtained applying x4 and x3 respectively but this improvement is, in general, marginal. The overall results vary from a range of 53.5% to 71.5% in the experiments conducted.

6 DISCUSSION

In this work, a new method for EEG data augmentation data using Wasserstein Generative adversarial networks with gradient penalty (WGAN-GP) was introduced. It was the first attempt of its type in using a set of nine explainable features. The task involved extracting high-level features from the DEAP dataset. The set of

nine features was extracted at every 6-second interval and was used as a base for classification algorithms, KNN, SVM and neural networks. A WGAN-GP based augmentation framework was then implemented with different design decisions for critic and generator. The DEAP dataset was split into two parts: one for training the model and WGAN-GP and the other part for performing consistent tests across different augmentation models. The WGAN-GP was trained for 100 epochs and the data generated at the end of the epochs was saved and evaluated. The 32-fold cross-validation accuracy scores revealed that the performance of the generated data was similar to that of the original data, thus emphasizing the fact that the distribution of the generated data must be close to the real data. Furthermore, after the data augmentation procedure, the models using KNN, SVM and neural networks showed slight boosts in their accuracy scores. The highest improvement was, as expected, observed for the liking emotion (17.5%). The remaining emotions (valence, arousal and dominance) produced a maximum enhancement of 5%, 3.75% and 6.25% respectively. This research has set up a path for future research using state-of-the-art WGAN-GPs for EEG data augmentation. The overall research could have been improved but is hindered by the nature of data and the type of features selected in the earlier stages.

7 CONCLUSIONS

This research was done as an early exploration of the use of WGANs in augmenting EEG data for affective computing tasks. This research should therefore be considered as a stepping stone in the development of future research. In this work, we developed a model able to generate synthetic data with the aim of improving the performance of a emotion classification task. To that end, and with a view to reduce the complexity of the features used to facilitate the convergence of a complex model as it is the case of GANs, nine primitive features are used (namely, mean, median, maximum, minimum, standard deviation, variance, range, skewness and kurtosis). Although this work shows a significant improvement in some cases (specially when using data from the *liking* emotion) suggesting a potential application of this technique for the objective pursued, the research gave marginal and inconsistent enhancement of accuracy across different emotion types.

The main limiting factor in this research is the data. The 32 subjects with 40 experiments, while sufficient when using traditional classification approaches, is unsuitable when it comes to training WGAN-GP. Notice that only 22-participants data was available for its training, which is close to 880 experiments. This is far from ideal since a higher amount of data would be expected to properly train a WGAN-GP model. Moreover, verifying the performance of WGAN-GPs on other standard emotion recognition dataset like SEED should be evaluated.

The second factor that limited performance in this research is the simple feature engineering used. As this research was about exploring the use of GANs, simple yet descriptive features were selected and an emphasis on using complex feature engineering was not considered in this early step. Further research and experiments would be needed to verify the performance of WGAN-GP in tandem with more powerful EEG features extracted using methods like data entropy or frequency-based data engineering.

The other important concern with this research is the difficulty to draw a fair comparison with similar works. It is complicated to systematically compare the results of this research to other similar ones. This is mainly because the test data in our research is not similar to other works. Developing methods that allow a fair comparison of results with other researches that involve the same test data is a problem in itself. A way of framing a suitable metric and comparison framework would enable future researchers to contribute towards this growing field of using data augmentation in understanding human emotions and addressing the issue of privacy and curbing high costs of data acquisition.

Future research can make use of a newer PATE-GAN for generating synthetic data that offers differential privacy guarantees [4] or the implementation of a schema considering temporal information such as Long-Short Term Memory (LSTM) approaches. This could open up new spheres in the field of data augmentation for EEG.

REFERENCES

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein gan. arXiv 2017. *arXiv preprint arXiv:1701.07875* 30 (2017).

[2] Vikrant Doma and Matin Pirouz. 2020. A comparative analysis of machine learning methods for emotion recognition using EEG and peripheral physiological signals. *Journal of Big Data* 7, 1 (2020), 1–21.

[3] Haifeng Jin, Qingquan Song, and Xia Hu. 2019. Auto-keras: An efficient neural architecture search system. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1946–1956.

[4] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. 2018. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*.

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

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

[7] Yisi Liu, Olga Sourina, and Minh Khoa Nguyen. 2011. Real-time EEG-based emotion recognition and its applications. In *Transactions on computational science XII*. Springer, 256–277.

[8] Yun Luo and Bao-Liang Lu. 2018. EEG data augmentation for emotion recognition using a conditional Wasserstein GAN. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2535–2538.

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

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

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

[12] Samarth Tripathi, Shrinivas Acharya, Ranti Dev Sharma, Sudhanshi Mittal, and Samit Bhattacharya. 2017. Using deep and convolutional neural networks for accurate emotion classification on DEAP dataset. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. 4746–4752.