Simple adversarial model to preserve EEG privacy in brain-machine interfaces

Paras Amitkumar Lad
Lakehead University
Thunder Bay, ON, Canada
plad2@lakeheadu.ca

Vinaydeep Kaur
Lakehead University
Thunder Bay, ON, Canada
vkaur3@lakeheadu.ca

Garima Bajwa
Lakehead University
Thunder Bay, ON, Canada
garima.bajwa@lakeheadu.ca

***Abstract*—*Background*. With the boom in AI and ML, EEG signals are now been used in various applications. It is important to secure EEG data while it is used in EEG-based applications. *Aim*. We have developed a simple architectural design using CNN and adversarial learning that can provide user privacy during task classification. *Method*. We built a CNN model with two sets of dense layers, one for the subjects and the other for task classification. Using alternate weight freezing determined optimal weights that preserved the user's privacy while enhancing task classification. *Results*. We conducted our study over two datasets, Our pilot study over the Dreamer dataset is based on visually evoked videos by 23 subjects. We found an overlap in its data because a single EEG had components of all the output classes. The results of CNN adversarial architecture for privacy gave good results for the alcoholic and controlled dataset of 20 users. It classified alcoholic and controlled with an accuracy of 70%, and user classification was as low as 18%, indicating that the model preserves user identity. *Conclusion*. In EEG task classification, it is possible to ensure the user identity is not exposed by performing adversarial learning, which maximizes information gain and user identity loss using simple CNN adversarial architecture.**

I. Introduction

The field of neural signal and ambulatory sensors has gained significant momentum in recent years, with Electroencephalogram (EEG) emerging as a particularly promising area of investigation. EEG has been found to provide ecologically valid data on human emotions, mental states, and psychological activity, making it an invaluable tool for various applications [1].

Advancements in Machine Learning and Artificial Intelligence have enabled us to leverage EEG signals to derive meaningful results in emotional state identification, mental state identification, and the development of assistive technologies such as EEG-based wheelchairs[19] and spell checkers. Additionally, EEG can also be employed in military applications to control drones [18]. It is relatively feasible for attackers to gain sensitive characteristics and traits about the users using machine learning and their EEG recordings[20]. Hence, it is crucial that these applications safeguard against identity theft or attacks on user identity, to prevent sensitive personal data from being permanently stored on third-party servers and potentially susceptible to hacker attacks.

To improve accountability in the services offered through the use of EEG signals, it is vital to conduct research focused on enhancing the privacy of EEG data. One of the main concerns in this regard is user re-identification, as EEG-based systems record patterns unique to the individual's behavior and thoughts. Any compromise in this information can lead to compromised user identity, which may cause severe repercussions.

With the sophisticated techniques available in the field of Machine Learning and Artificial Intelligence, it may be possible to reverse engineer and determine the user to whom the EEG signal belongs, making it essential to explore measures to prevent user re-identification while preserving the information gain.

The purpose of this study is to investigate methods to prevent user re-identification while maintaining the information gain from EEG signals. In the following sections, we will discuss previous work done in the field of EEG, research conducted in preventing user re-identification, and identify gaps in knowledge to formulate research questions. We will then outline our methodology and present the results we obtained, followed by a comprehensive discussion on the implications of our findings.

II. Literature Review

Several critical applications use EEG. It is used in medical applications, emotion detection, cryptography, and military applications[22]. A study done by Hemdan [11] for seizure detection created spectrogram images of EEG signals and performed Arnold and Chaotic encryption over it, which was passed over a pre-trained CNN model to perform seizure detection. The architecture's performance was around 85% which needs to be improved to provide better results. Haojun Xu [4], on the other hand, performed mental load classification with CNN using a single model which includes multi-channel containing temporal and spatial information and single channel emphasis on one channel spectral map as inputs giving comparable outputs as traditional and state-of-art single CNN models. Two fusion structures from the single model were analyzed, and the fusion of the Pointwise-Gated Boltzmann Machine layer (PGBM) component outperformed all previous models. Efficient ways to preprocess EEG data, and analyze aspects other than working memory needs to be looked upon in the future.

It is also gaining popularity in Cryptographic keys as it is difficult to hack. A study done by Garima[17] exploited the brain as a biometric unclonable function using which a user can generate a unique and repeatable key that is resistant to cryptoanalysis, eavesdropping, and even against an attacker with information about the system using Subject Authentication obtained from task using energy obtained from Discrete Fourier Transform and Discrete Wavelet Transform followed by Neurokey generation by feature selection using normalized threshold and segmentation window protocol. Another study in the field of encryption done by Marcin [9] presented cryptographic authentication over the covariance matrix of EEG data (EEG channels in 10-20 international systems) of 42 subjects from fuzzy commitment scheme and error correcting BCH (Bose-Chaudhuri-Hocquenghem) code. 400 bits of the cryptographic key while tolerating 87 bits of error was generated by the system resulting in EER, TPR, and AUC of 0.024, 0.9974, and 0.927, respectively, for each EEG signal. Wanli [10] demonstrated using associated correction data to generate different keys robustly over EEG data. Their schemes produce different long keys for different applications, thus making sure an attack on one does not give an attack on all. Its security case is founded in extensive research in the application area and in statistical lower bound argument.

There exist studies that attempt to increase the accuracy of EEG-based classification. Jose [3] did research on different deep learning models for the classification of EEG signals, which proved CNN, RNN, and DBN outperformed other models, while CNN performed best in spectrogram images and DBN performed best with calculated features as inputs. Their study narrows the selection of models based on the business problem. Analysis of deep models on raw and denoised EEG signals needs to be assessed. On the other hand, Hong [5] investigated functional near-infrared spectroscopy (fNIRS) and hybrid fNIRS-EEG for BCI over LIS (locked-in syndrome) patients. Mental arithmetic was most suitable for hemodynamic response stating signal mean and signal peak as the most efficient features for classification of it. In hybrid, mean fNIRS is combined with PSD of EEG to improve the model'. LDA was used as the classifier for both the fNIRS and hybrid fNIRS-EEG. Vector phase density also has potential over hemodynamic response, which is yet to be explored.

After understanding the criticality of EEG privacy preservation, we explored the research of user reidentification using the same EEG data. Wenyao Xu [6] performed EEG-based identification using NN over four different scenarios considering ensemble averaging and low pass filter for noise reduction wavelet packet decomposition for feature extraction. Identifying individuals from the other 32 subjects got the best accuracy of 90, whereas identifying all the subjects had the worst. The side-by-side method improved the accuracy of identifying all the subjects by five times to reach 47%. The architecture of NN can be explored and worked to predict more accurately. We studied LSTM for identifying subjects and found that there is a study done in the field. Benny Lo [7] proposed a 1D-convolution LSTM model for 16 channels of EEG-based identification and outperformed the previous state-of-art approach receiving Rank-1 accuracy and EER of 99.58% and 0.41%, respectively. The results showed that the proposed model exploits spatial information residing in EEG channels providing additional features to distinguish subjects. The proposed model can be tested on other public datasets for its scalability; instead of selecting manual channels of EEG, an automatic selection algorithm can be looked upon, and important EEG channels providing identification information can be analyzed too. Miguel [14] presented deep network architecture consisting of CNN and Inception layer, which increased the performance to identify 23 different individuals over EEG data over the DREAMER dataset. 0.9401 accuracy was achieved over their model, with SVM-L securing the second-best accuracy of 0.8879. Identification performance achieved can be improved to be applied over applications. The model can be analyzed over EEG raw data for new findings.

We started with a well know dataset for our research which is the dataset DREAMER presented by Ramzan [8], which consists of EEG and ECG signals over 23 participants showing them 18 film clips capturing stimuli and baseline and rating their emotional response in terms of valence, arousal, and dominance. The recordings were recorded with a low-cost portable, wireless device, and the results of classification accuracy and F1 score were significantly higher than random voting and voting according to the class ratio and are comparable to other similar works that used non-portable medical equipment. Different algorithms and models can be applied to analyze the data being created. Although the study by Anwar [13] used this dataset for privacy preservation in EEG, it has overlapping data i.e. a single EEG record cannot be classified into a specific output class, it has components of all the output classes. This led us to move to a different dataset called predisposition to alcoholism [16]. It is EEG data to examine the EEG correlation of genetic predisposition to alcoholism. It is measured from 64 electrodes placed on the subject's scalp sampled at 256 Hz for 1 second. There were two groups of subjects alcoholic and controlled. It consisted of a large dataset of 20 users with ten alcoholics and ten controlled. Each subject was exposed to either a single stimulus (S!) or two stimuli S1 and S2. When two stimuli were shown, they were presented in either a matched condition called an S2 match or an unmatched S2 no-match.

A study by Miguel [14] found that user identity is revealed in the EEG task classification. This study presented deep network architecture consisting of CNN and Inception layer, which increased the performance to identify 23 different individuals over EEG data over the DREAMER dataset with 0.9401 accuracy. This suggests that issue of privacy in EEG data is not new. There have been researches that provide privacy in EEG data. Research by Anwar [13] aimed to create a FedEmo model, which would protect user privacy while categorizing emotions in the DREAMER dataset. By utilizing a base model of ANN and employing federated learning techniques, the research team attained accuracy rates of 63.3%, 56.7%, and 52.2% for Valence, Arousal, and Dominance,

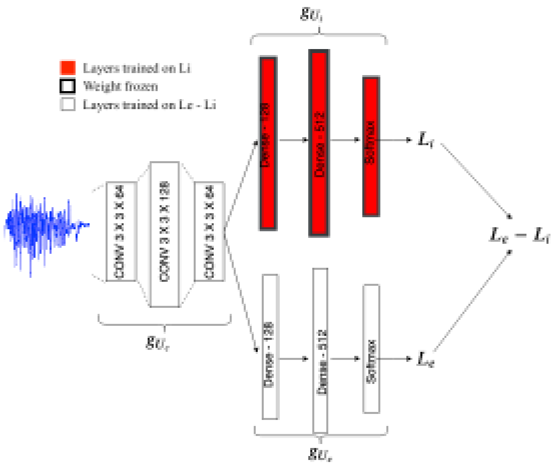


Fig. 1. [1] Hybrid model framework consisting a couple of convolution layers followed by two separate dense layers to perform the task and user identification classification

respectively. However, the research findings indicate that the accuracy of the framework is not satisfactory and needs to be improved. Another study by Popescu [15] used a privacy-preserving technique called homomorphic encryption (HE), which allowed for computations on encrypted data. It was evaluated on seizure detection and prediction of predisposition to alcoholism, but it has practical limitations of high computational complexities and noise accumulation. A study presented by Narula [12] developed an effective architectural approach to utilize anonymization methods to retain emotion-specific information while removing user-specific information, which is crucial for preserving privacy and reducing the possibility of user re-identification for image-based data. The approach performs well in anonymizing image-based data, but it is not explored for EEG data.

III. Research Gap and Motivation

After careful analysis, it has been determined that no significant research has been conducted on the prevention of re-identification of users through the use of EEG data. Numerous applications utilize EEG datasets, yet fail to establish sufficient safeguards against potential attacks and instances of identity misuse or theft. Consequently, these applications are left vulnerable to such threats.

Narula [12] conducted a study focused on utilizing anonymization techniques to extract emotion-specific information while eliminating user-specific data. This approach is essential for preserving privacy and reducing the likelihood of user re-identification over image data through hybrid model performing adversial learning by weight freezing approach over two task of of maximizing emotion detection and minimizing user identity detection as shown in Figure 1 [1].

This gave us the motivation to analyze EEG data through this approach to reach our target. Considering fewer convolution layers also reduces space and time complexity and can achieve great results.

IV. Research Questions

We want to perform task classification which minimizes the identity of the user being exposed which leads us to our Research Question as

How to maximize information gain and identity loss in an EEG application?

We divided our goal into the following sub-research questions:

1a) Is it possible to disclose the user's information while performing task classification over EEG data?

1b) What is the complexity to distinguish EEG signals of one user over another?

2) How to achieve optimal user anonymity while performing task classification over EEG data?

V. Methodology

Due to the ability of Convolution Neural Networks to automatically learn features from hierarchical patterns, CNN excels in extracting relevant features from time-series datasets like EEG. EEG data involves electrical activities of the brain represented in multichannel time series signals. CNN can identify temporal and spatial relationships between them and disclose important information for task classification. Thus we focus our analysis on the CNN model which can unfold meaningful data used across wide applications like BCI, seizure detection, cognitive state recognition, and other neuroscientific investigations.

Chaspari[1] created a hybrid model that performs a similar task over image data but considering EEG data, we need to create a model and set hyperparameters such that we can retrieve meaningful information with less complexity or layers of the model. Lee[14] showcased the extraction of EEG data over CNN layers and helped us to create CNN architecture that correlates with our requirements.

We created a model with an entry point for the EEG data. Input consists of 2D convolutional data with timestamps and channels on both axis respectively. This data is forwarded to 1st convolution layer which consists of 18 filters, kernel size as (32,1), and kernel constraint with max norm as 2 which is used to efficiently capture temporal features from EEG signal.

The temporal features captured are provided to another convolution layer having the same 18 filters with kernel size as (61,1), kernel constraint with max norm as 2, and L2 regularization with a coefficient of 0.001 which captures spatial features over the signal.

Data obtained containing temporal and spatial features is forwarded to the square activation layer which amplifies the features making smaller ones smaller and larger ones larger. This is forwarded to Average pooling with pool size (1,75) and stride (1,22) which reduces complexity and extract significant features. Log activation is performed followed by Batch Normalization which helps in faster convergence during training.

Flatten layer is added to convert 2D features to 1-D features followed by a Dropout layer that randomly drops 50% of the neurons while training which reduces overfitting. Two different tasks are split from this layer. The first task uses the Dense

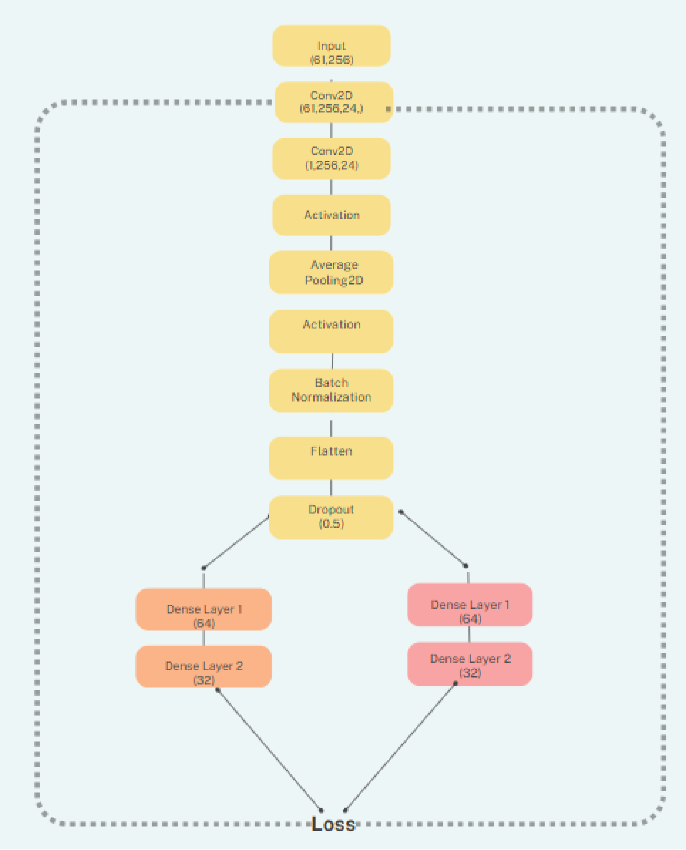
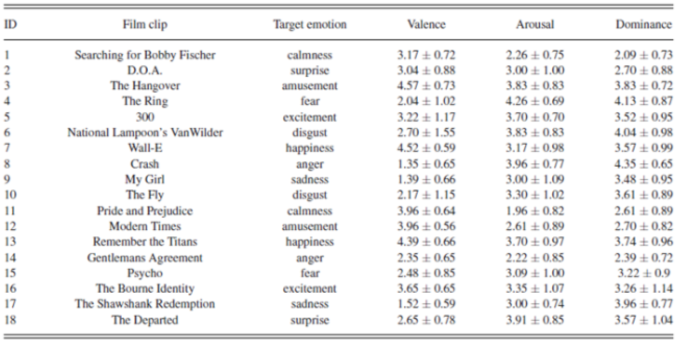


Fig. 2. Hybrid model performing adversarial learning



ID	Film clip	Target emotion	Valence	Arousal	Dominance
1	Searching for Bobby Fischer	calmness	3.17 ± 0.72	2.26 ± 0.75	2.09 ± 0.73
2	D.O.A.	surprise	3.04 ± 0.88	3.00 ± 1.00	2.70 ± 0.88
3	The Hangover	amusement	4.57 ± 0.73	3.83 ± 0.83	3.83 ± 0.72
4	The Ring	fear	2.04 ± 1.02	4.26 ± 0.69	4.13 ± 0.87
5	300	excitement	3.22 ± 1.17	3.70 ± 0.70	3.52 ± 0.95
6	National Lampoon's VanWilder	disgust	2.70 ± 1.55	3.83 ± 0.83	4.04 ± 0.98
7	Wall-E	happiness	4.52 ± 0.59	3.17 ± 0.98	3.57 ± 0.99
8	Crash	anger	1.35 ± 0.65	3.96 ± 0.77	4.35 ± 0.65
9	My Girl	sadness	1.39 ± 0.66	3.00 ± 1.09	3.48 ± 0.95
10	The Fly	disgust	2.17 ± 1.15	3.30 ± 1.02	3.61 ± 0.89
11	Pride and Prejudice	calmness	3.96 ± 0.64	1.96 ± 0.82	2.61 ± 0.89
12	Modern Times	amusement	3.96 ± 0.56	2.61 ± 0.89	2.70 ± 0.82
13	Remember the Titans	happiness	4.39 ± 0.66	3.70 ± 0.97	3.74 ± 0.96
14	Gentlemans Agreement	anger	2.35 ± 0.65	2.22 ± 0.85	2.39 ± 0.72
15	Psycho	fear	2.48 ± 0.85	3.09 ± 1.00	3.22 ± 0.9
16	The Bourne Identity	excitement	3.65 ± 0.65	3.35 ± 1.07	3.26 ± 1.14
17	The Shawshank Redemption	sadness	1.52 ± 0.59	3.00 ± 0.74	3.96 ± 0.77
18	The Departed	surprise	2.65 ± 0.78	3.91 ± 0.85	3.57 ± 1.04

Fig. 3. [2] Data captured over 18 Film clips representing different emotions - DREAMER Dataset

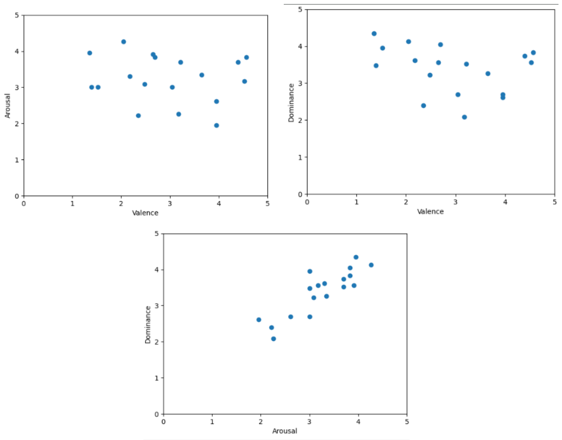


Fig. 4. (a) Arousal vs Valence Plot (b) Valence vs Dominance Plot (c) Arousal vs Dominance Plot

layer to perform task classification and the other task uses the Dense layer to perform user classification. Architecture parameters are further described during experiments. The architecture of our proposed model is described in figure 2.

VI. Experiments

A. Dataset

The experiments were performed on the DREAMER dataset. It consists of Electroencephalographic (EEG) and Electrocardiogram (ECG) signals over affect elicitation through audio-video signals. 18 film clips were represented from which 2 of them were used to represent one of the nine emotions - amusement, excitement. happiness, calmness, anger, disgust, fear, sadness, and surprise as described in figure 3. These 9 emotions were been divided into 3 divisions as Valence, Arousal, and Dominance. 23 healthy participants were been analyzed after each stimulus (signal) to determine emotions in terms of Valence, Arousal, and Dominance.

Valence depicts humans having positive or negative feelings (ranging from unpleasant/stressed to happy/elated). Arousal shows whether humans feel bored or excited (ranging from uninterested/bored to excited/alert). Dominance showcases how humans feel without control or empowered (ranging from helpless to empowered). Each film clip ranges between 65 to 393 seconds. As the emotional state of a person changes over time, to avoid contaminating data recordings with multiple emotions, only the recordings captured during the last 60 s of each film clip were used for further analysis.

EEG signals were been recorded with a sampling rate of 128 Hz using the Emotiv EPOC system which uses 16-gold plated contact sensors whose locations are placed according to international 10-20 system - AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4. Experiments were started by showing a neutral film clip, a video clip considered to have no valence to establish the baseline signals. figure 2 shows information regarding the film clips.

B. Data Analysis

We started with analyzing the three emotions- valence, arousal, and dominance being classified by the participants. We took the mean assessment by all participants for each film clip in terms of valence, arousal, and dominance in the form of a scatter plot which shows the relation between them. Plots are depicted in figure 4.

From the graph plotted and the correlation matrix of the variables we can see that there is high correlation between arousal and dominance (0.69) which depicts that if a participant felt excited about the stimuli (high arousal), they also

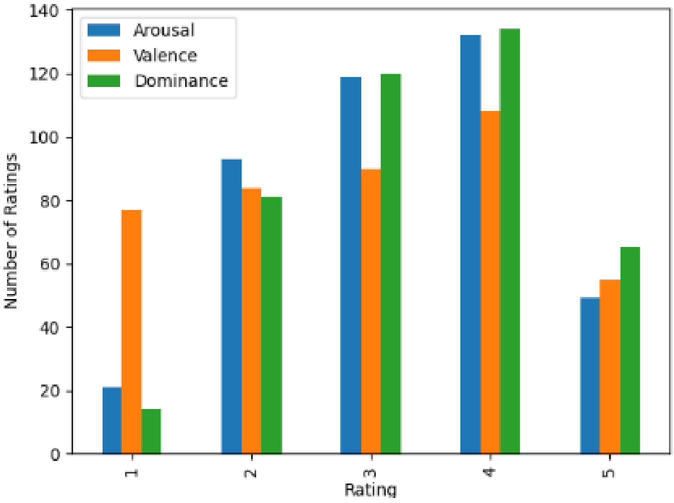


Fig. 5. [2] Valence, Arousal and Dominance Ratings

felt empowered (high dominance) and when they felt bored (low arousal), they also felt helpless and without control (low dominance). A weak negative correlation was found between valence and dominance.

These results are consistent with the emotions that the film clips used to elicit which is also analyzed in [1]. We also analyzed the emotions over all the ratings leading to a balanced dataset except the arousal and dominance being less for rating 1 as depicted in Figure 5.

As emotions range from 1 to 5, we have to convert them into 1 for high and 0 for low and thus we took 2.5 as the threshold. This has led to imbalance classes for some participants and scales - dominance(52%,48%), arousal(56%,44%), and valence(61%,39%) for class 0 and 1 respectively

We then started working with EEG signals. The Dreamer dataset contains EEG and ECG signals. We fetched the last 60 seconds of the stimuli signal which is captured at 128 Hz sampling frequency leading to 60x128 - 7680 time points and the last 4 seconds of the baseline signal leading to 4 x 128 - 512 time points. We plotted the stimuli and baseline EEG data from the ‘AF4’ electrode to view the signal depicted in figure 6.

EEG signals contain information relative to affect recognition in 4-30 Hz split into theta(4-8), alpha(8-13), and beta(13-30). So we applied Hamming linear phase FIR filters over the signal over 4-30 Hz frequencies. Figure 7 contains EEG signals of stimuli and the baseline of the ‘AF4’ electrode after applying the filter.

The above signal was further processed under 3 different scenarios. First scenario: Overall Signal with Log PSD - we applied the Welch’s overlapped segment averaging estimator to estimate the Power Spectrum Density (PSD) of the EEG band over theta, alpha, and beta using 256 sample windows with an overlap of 128 samples. We then plotted the PSD being generated of the overall signal and then found the average bandpower of all three bands (theta, alpha, and beta) and stored

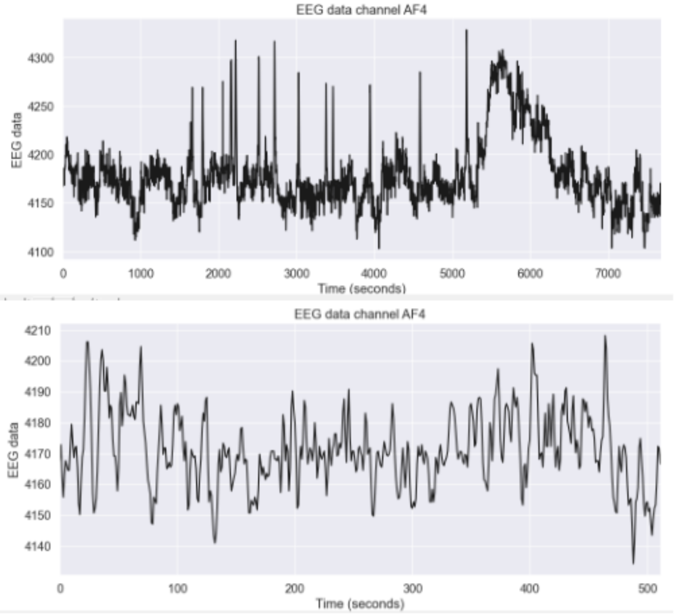


Fig. 6. RAW EEG signal representing stimuli (60s) and baseline (4s)

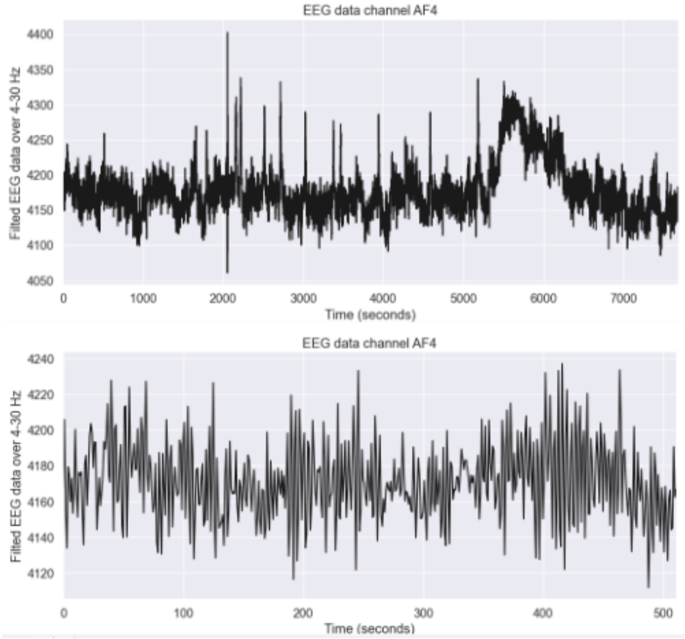


Fig. 7. EEG signal representing stimuli (60s) and baseline (4s) after performing Hamming filter over 4-30 Hz frequencies

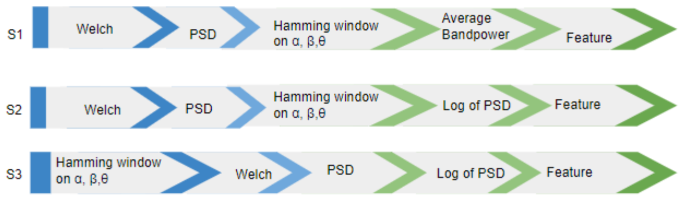


Fig. 8. Three different preprocessing scenarios S1, S2, and S3 experimented to analyze how data behaves over different Machine Learning algorithms

TABLE I
Classification task was performed through Support Vector Machine (SVM) with Radial Basis Function Kernal (RBF)

Metric	Measure	Valence	Arousal	Dominance
Overall Signal Average Bandpower	Accuracy	0.5847	0.6934	0.763
	F1 - score	0.5101	0.7246	0.784
Overall Signal PSD	Accuracy	0.6108	0.6934	0.7782
	F1 - score	0.5231	0.7275	0.7999
Divided Signal PSD	Accuracy	0.6043	0.6934	0.763
	F1 - score	0.5478	0.7188	0.784

them as features in a CSV file. Figures are depicted below for stimuli of the 'AF4' electrode of the EEG signal.

Second Scenario: Overall Signal with Average Bandpower - we applied the Welch's overlapped segment averaging estimator to estimate the PSD of the EEG band over theta, alpha, and beta using a 256 sample window with an overlap of 128 samples. The logarithm of PSD was captured and stored as features in a CSV file. Third Scenario: Divided Signal - We applied three separate Hamming windows linear phase FIR filters over the signal over theta(4-8), alpha(8-13), and beta(13-30) frequencies. Then this filtered data was applied to Welch's overlapped segment averaging estimator to estimate the PSD of the EEG band over theta, alpha, and beta using a 256-sample window with an overlap of 128 samples. The logarithm of PSD was captured and stored as features in a CSV file. All 3 experiments are described in figure 8. After performing preprocessing, we have our data to be fed to the model to perform our classification tasks.

Classification of emotions is to be performed, as valence (high/low), arousal (high/low), and dominance (high/low). As emotions range from 1 to 5, we have to convert them into 1 for high and 0 for low and thus we took 2.5 as the threshold. This has led to imbalance classes for some participants and scales - dominance(52%,48%), arousal(56%,44%), and valence(61%,39%) for classes 0 and 1 respectively. To cope up with an unbalanced dataset, F1-score has also been taken along with classification accuracy for the analysis. The classification was performed considering the record of each participant individually over a Support Vector Machine(SVM) as a classifier with Radial Basis Function(RBF) as the kernel. 10-fold cross-validation was used, with 8 folds of 2 samples and 2 folds of 1 sample. At each step, one fold was set as testing and other as training data. This whole process was performed 10 times such that each fold will be a part of testing data.

Overall analysis was performed over all the scenarios. In the Average Bandpower of PSD, it calculates the average over the overall frequency range of the particular band whereas the maximum value is fetched and the log of that value is stored as a feature. Applying the hamming filter to divide the signal into different bands performed better than performing PSD first and applying filters later so we moved ahead with the third scenario and the results are described in Table-I. Support Vector Machine with Radial Basis Kernel, Decision Tree, and Random Forest was experimented over our data and SCM with RBF outperformed others as SVM is better in classifying binary class data which is 0 or 1 for all three emotions in our output features.

We can see from the above table that model is not able to predict results quite accurately as expected as we received maximum accuracy as 61%, 69%, and 77% respectively over the signal. User identification was performed with EEG data over 23 participants. 80% of the record of each participant was considered as training and the rest 20% of each participant is part of testing. SVM with RBF kernel, Decision Tree, and Random Forest were used over Divided Signal PSD data and we received results of 10%, 24%, and 47% respectively. We found SVM performed better in task classification as it is good in separating binary class here 0 and 1 for all 3 emotions. Random Forest performed better over the multi-class classification of 23 users. These results were not acceptable to prove our theory of user identity being exposed over EEG data and thus we started analysis over the results.

We found that EEG signals can elect more than one emotion. Due to this, we are not able to classify each emotion as EEG signal contains features of more than one emotion. To overcome this challenge, we filtered out EEG data that elect unique emotions which are recorded in the below table.

TABLE II
Count of unique emotions from EEG signals of each subject

Subjects	Valence	Dominance	Arousal
0	2	4	0
1	3	0	0
2	4	2	0
3	4	0	0
4	3	0	1
6	2	0	2
7	4	1	0
8	1	0	0
9	4	3	0
10	0	0	1
11	2	0	0
12	2	0	0
13	5	0	0
14	0	3	0
15	2	1	0
16	1	1	0
17	2	1	0
18	1	0	0
19	2	1	1
21	2	0	0
22	5	1	1

From Table-II we see that data has been reduced. We dropped the Arousal emotion as the data is negligible. Considering Valence and Dominance emotions there are 3 subjects (Subject 0, 2 and 9) which have unique records of both emotions which can be taken for analysis. So further analysis was performed on 3 subjects and their unique records of valence and dominance. CNN model used above for emotion classification of Valence and Dominance and accuracy achieved was 55%. Same CNN model used above for subjects classification and accuracy achieved was 66%. From this we can see that emotion classification has still remained same but subject classification has increased. This can be justified as the subjects has been decreased as well as data which has affected the classification of emotions.

To overcome this barrier we created sub-trails for 3 subjects using 1s, 2s 5s, and 10s data intervals for valence and dominance uniquely identified trials (out of the 18 total), using the same architecture CNN model as used in the previous experiment. We subtracted the mean of the baseline (respective channel) —and performed the emotion and user identification. Results are recorded in Table III. We got the best result at 10s time interval sub-trials with emotion classification prediction to be around 71% and user classification accuracy around 36%. Through these results, it is difficult to prove that user identity can be exposed at a larger rate through EEG signals.

C. Observations from Dreamer Dataset

While analyzing the Dreamer dataset we found that there is more than one emotion elected from an EEG signal which means an EEG signal can induce Valence, Arousal as well as Dominance. Due to this, it is difficult to identify features that will help us to distinguish output classes. This has not been addressed in any of the previous studies that work towards this dataset to the best of our knowledge. To overcome this overlapping emotion issues, we filtered out the EEG signals that elect unique emotions which have reduced the size of the dataset. Due to less data, we are not achieving satisfactory results. So we moved on to another dataset ”Alcoholism”.

D. Alcoholism Dataset

We started working with the Alcoholism dataset which consist of EEG signals captured at 256 Hz sampling frequency from 20 users (10 alcoholic - 10 controlled). There are 3 conditions, either a person is shown one stimuli S1, two stimuli where second stimuli matches the first (S2 match) or the second stimuli don’t match with the first (S2 no-match). Our goal is to perform task classification whether the user is alcoholic or controlled and also want to secure user credentials that means while performing this task classification, the user’s identity should not be exposed.

E. Data Analysis

We started analyzing how our EEG data behaves over CNN architecture. We found insights from Kim[6] which helped us to define the architecture and hyper-parameters of our model. We want to keep our model simple through less complexity we can achieve better results comparable to different complex models that include many Convolution and Pooling Layers. So we started with creating 2 convolution layers - one for fetching temporal features and second one for fetching spatial features from the EEG signal.

We started exploring the size of the kernel for the convolution layers. We experimented with kernel size based on sampling frequency. Let’s assume most of our informative signal lies above 4 Hz, so our kernel size will be decided by dividing the sampling frequency by the frequency above which we want our recordings [21]. Here it will be 256/4 that is 64. We experimented with frequencies above 4, 8, 10, and 12 Hz, and the results are updated in Table IV. We found the best results for the kernel size of 20 which means most of our important data lies above 12 Hz.

TABLE IV
EXPERIMENTING WITH KERNEL SIZE

Classification Frequency	Kernels	Alcoholic/Controlled Accuracy	Subject Accuracy
4Hz	64	0.55	0.15
8Hz	32	0.60	0.12
12Hz	**20**	**0.67**	**0.21**
10Hz	24	0.63	0.14

TABLE V
EXPERIMENTING WITH A NUMBER OF FILTERS

Number of Kernels	Alcoholic/Controlled Accuracy	Subject Accuracy
4	0.59	0.07
12	0.61	0.12
18	**0.67**	0.23
24	0.65	0.19
32	0.67	**0.25**
40	0.65	0.19

We analyzed a number of filters required in the convolution layers. We don’t want fewer filters so that model is under-trained and we don’t need many filters such that some kernels are not been utilized. So we experimented with filters in the range of 4 to 40 as described in Table V. We found 18 filters gave us better results and propagated to our model.

A square activation layer which amplifies the features making a smaller one smaller and a larger one larger was applied after the convolution layer. After convolution and square activation, we analyzed the pooling layer. We can loose information if incorrect parameters are set in pooling layers. The appropriate kernel size of the pooling layer was determined based on the kernel size of convolution layers. We tried kernel size by multiplying sampling frequency by 0.10, 0.15, 0.30, 0.45, and 0.50. We found better results with 0.30 which is 20. Then we look over the overlapping ratio in the pooling layer. We started with 0.1 and went up to 0.8 and we found the overlapping ratio of 0.7 worked best for our data and thus stride with (1,22) was assigned based on it. We also investigated max pooling and average pooling with the above parameters and found Average pooling gave better results as described in Table VI and we move ahead with that. This helps to reduce complexity and extract significant features.

Log activation is performed followed by Batch Normalization which helps in faster convergence during training. Flatten layer is added to convert 2D features to 1-D features followed by a Dropout layer that randomly drops 50% of the neurons while training which reduces overfitting.

TABLE VI
EXPERIMENTS FOR POOLING

Classification	Alcoholic/Controlled Accuracy	Subject Accuracy
Max Pooling	0.60	0.21
Average Pooling	**0.67**	**0.23**

We performed four experiments and all the above layers and parameters remain the same. The first experiment is used to perform task classification, here we are classifying whether a user is alcoholic or controlled. So Dense layer was added with one neuron which either says 0 for controlled and 1 for alcoholic.

The second experiment is used to perform user classification, here we are classifying 20 different users. So dense layer was added with 20 neurons which are used to identify each user. Through this experiment, we are trying to prove that EEG signals can be trained to perform user classification.

The third experiment performs multi-head classification. All the above layers are used and then the model is divided into two different tasks one to perform task classification with one output class of dense layer and the other to perform user classification with 20 output classes of dense layer. Both tasks are trained and weights are updated. We will show that while performing task classification, if the model is trained to classify users, it will be able to classify them. We can prove that users' data can be exposed while performing task classification.

The fourth experiment is our proposed model that performs adversarial learning over a multi-head model. Model architecture remains the same as the multi-head model. In adversarial learning, model weights of the Dense layer to perform user identification are frozen and the model is trained to maximize task classification. After that, all layers of the model are kept frozen except the user classification dense layer, and it is trained to maximize user identity loss. This adversarial learning helps to maximize task classification gain and minimize user classification gain. This process helps to preserve user identity that can be exposed while performing task classification. Model architecture is described in figure 2.

VII. RESULTS

A. *Quantitative Results*

We performed qualitative analysis on our study. The first experiment deals with task classification, and we found our model was able to predict alcoholic or controlled users with 98% accuracy. Moving ahead with the second experiment, we were able to classify users with 95% accuracy. Both experiments prove that EEG signals can be used to perform the task as well as user classification in this scenario.

In the third experiment, we performed task and user classification on a multi-headed model. The convolution layer was followed by two sets of dense layers, one for task classification and the other for user classification. The same features extracted in the convolution layer were able to classify the binary classification of Alcoholic/ Controlled with an accuracy of 96% (Fig 10) and multi-class classification of subjects with an accuracy of 80% (Fig 11), indicating that the same features extracted for task classification can also classify the user with good prediction rate. This shows the user's identity information is revealed during the task classification.

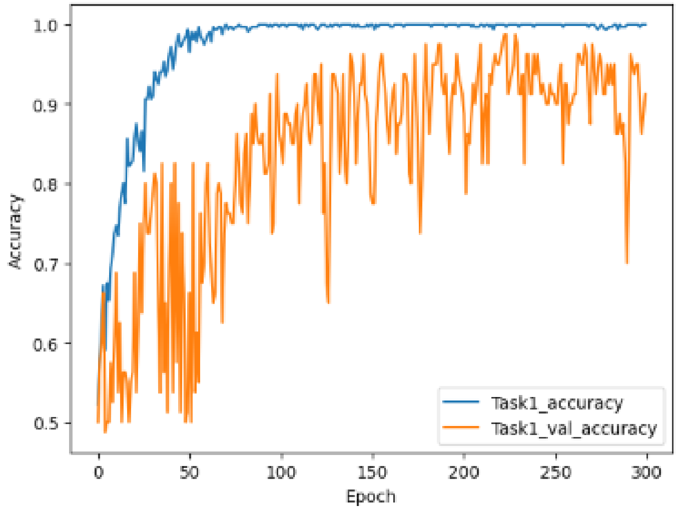


Fig. 9. Accuracy plot for Alcoholic Controlled classification in Multihead model

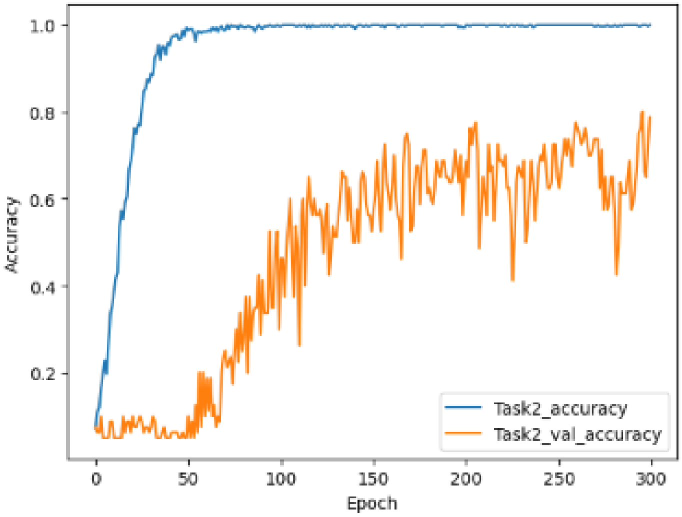


Fig. 10. Accuracy plot for Subject classification in Multihead model

In the model explained in Fig (1,2), the convolution layers of the CNN leaned and preserved the information of task classification and leaned to withhold the subject information. The Hybrid model could classify the task into Alcoholic and Controlled by 70% and user classification by 18%, which is a 77% decrease in user classification. This is a significant drop in the ability of the CNN model to identify the user, hence providing privacy to the user. The task classification, on the other hand, was 70%. However, there is a decrease in the accuracy of task classification by 26%, which is not significant

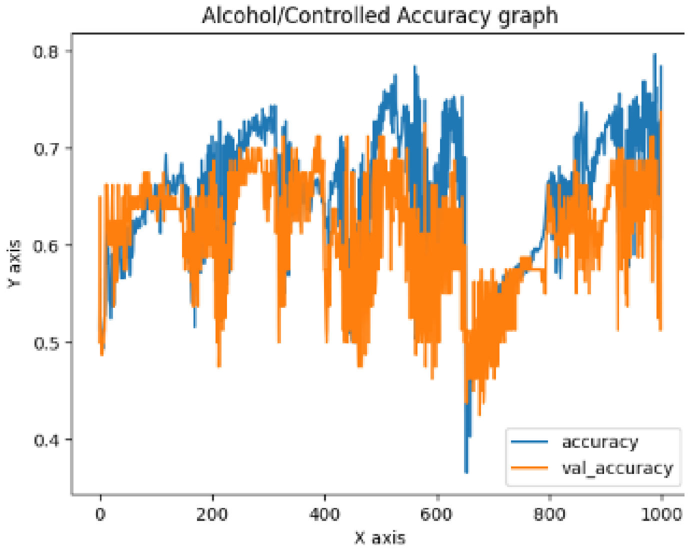


Fig. 11. Accuracy plot for Alcoholic Controlled classification in the proposed model through adversarial learning

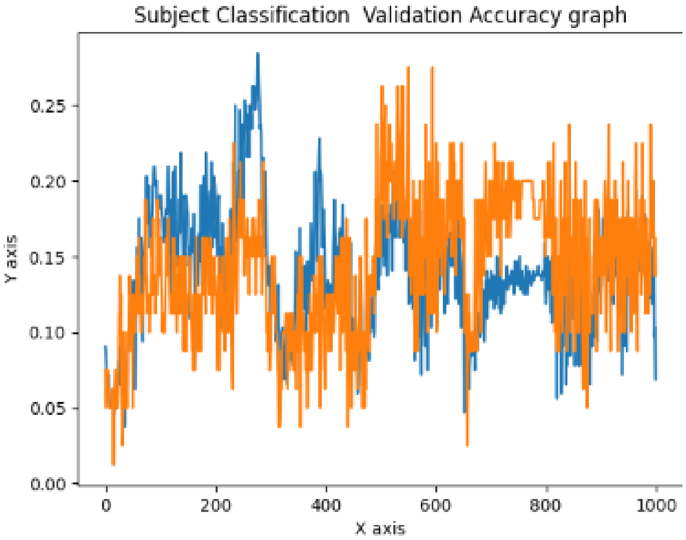


Fig. 12. Accuracy plot for Subject classification in the proposed model through adversarial learning

compared to the user classification drop.

To analyze the behavior of our proposed model, we analyzed the output after the flattening layer depicted in figure 11. The variation of Alcoholic/Controlled classification varies from -1 to -6, and User Classification varies from -3 to -5. If we take a look at the multi-head model, the signal is distributed in a way that all parts over the range are been captured to gain insights which resulted in classifying both tasks as well as user classification. In our proposed hybrid model, the signal varies over a range similar to our task classification, but the signal is not capturing more data over the -3 to -5 range as it did in the multi-head model. This shows that model is trying not to capture data specific to users, which helps us to reduce user classification prediction.

B. *Qualitative Analysis*

1) Is it possible to disclose the user's information while performing task classification over EEG data?

We illustrated with a multi-headed CNN model of alcohol and controlled dataset that the model trained to extract features for alcoholic/controlled classification can also be used to perform subject classification. Alcohol and Controlled classification can be performed on a multi-headed model with 91% accuracy on an average over all three stimuli S1, S2 match, and S2 no match with exposing user-related information above 80%. Any attacker can use the features extracted for task classification and can train them to fetch the information related to users. So it is necessary to secure these features to avoid privacy breaches.

2) What is the complexity of distinguishing the EEG signals of one user over another?

To reduce the complexity associated with the model architecture, we carefully selected the parameters and hyperparameters of the Shallow Convolution Network model. We used minimal convolution layers required to achieve our goal of adversarial weight training. We used two convolution layers, one for temporal and the other for spatial features. Also, the kernel size was optimally selected to avoid overfitting and avoid having unused filters/kernels.

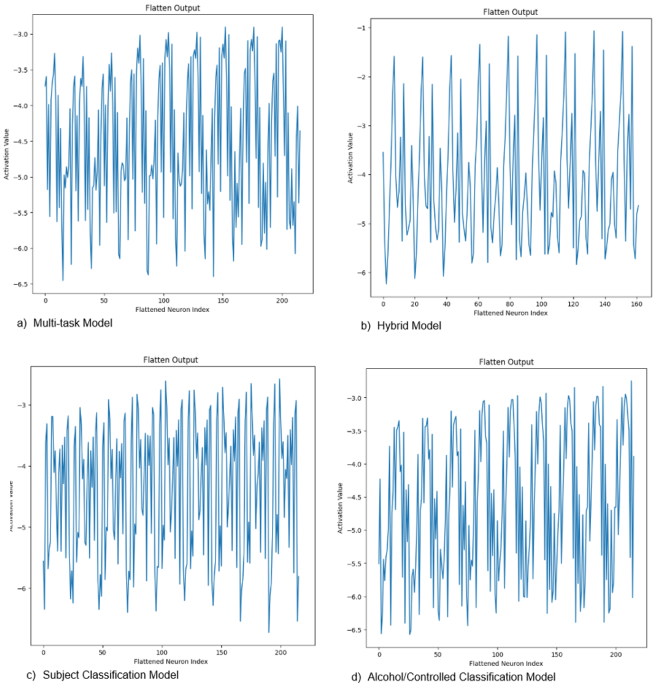
3) How to achieve optimal user anonymity while performing task classification over EEG data?

The Adversarial learning algorithm has proved to be a simple architectural algorithm that can preserve user-specific information that can distinguish one user from the other by maximizing the information gain by keeping the weights frozen over the user identification and ensuring identity loss by keeping weights frozen over task classification. Thus we ensured user anonymity by maximizing information gain and identity loss while performing task classification.

VIII. Discussion

In our research, we explored different preprocessing steps to fetch features that can be transferred to different classic Machine Learning models to perform different activities. We discussed how different ML techniques like Support Vector Machine, Decision Tree, and Random Forest differ from each other and when to use appropriate methods to gain better insights and results. To overcome different preprocessing steps and to reduce complexity, we explored Convolution Neural Networks, which can extract information from raw EEG signals to get better insights.

We propagated our work towards two different datasets, DREAMER and Alcoholism. The DREAMER dataset consists of 23 subjects with more than 60 seconds of EEG recording recorded at 128 Hz gives us a huge dataset to work on and make our model more generalized. The DREAMER dataset was analyzed to perform emotion classification in terms of Valence, Arousal, and Dominance. While performing experiments on this, we found that it contains overlapping emotions



a) Multi-task Model

b) Hybrid Model

c) Subject Classification Model

d) Alcohol/Controlled Classification Model

Fig. 13. Visualization of the performance of a) Multi-Task model b) Hybrid model c) Subject classification Model d) Alcoholic and Controlled classification model

which means an EEG signal can elect more than one emotion. This makes us difficult to classify our output classes as our data contains features of more than one output class. These insights into overlapping emotions are novel and have not been addressed in any previous studies to the best of our knowledge. We filtered EEG records that elect unique emotion but that has reduced our dataset and subjects. Due to that, we were not able to predict task and user classification as accurately as we expected, which directed us to move on to a new dataset.

We started with the Alcoholism dataset. It consists of EEG recording captured at 256 Hz for 1 second over 122 subjects which will help us to perform the task as well as user classification and can make our model generalized over more subjects. We explored over 20 subjects performing task classification to identify between alcoholic and controlled users by taking insights into user classification. Our motive was to create a simple CNN model that can give us better results which is comparable to different complex models. We explained our architecture consisting of 2 convolution layers to fetch temporal and spatial features from EEG data followed by one average pooling layer and we were able to achieve good results. We explained the selection of different hyperparameters we took over our architecture to obtain better results without the model being overtrained or undertrained.

We discussed about multi-head model, which describes how user identity can get exposed while performing task classification. We then further proposed a hybrid model that performs adversarial learning by maximizing the task information gain and minimizing the user identity information gain. We analyzed our data after the Flatten layer which depicts how different models disclose different information according to different tasks. While performing task and user classification, the model exposes features in the range where most of the information related to the tasks resides - from -1 to -6 in Alcohol/Controlled and -3 to -5 in user classification. The multi-head model captures more general features containing data with features of both tasks as well as users' identities. The hybrid model tries to fetch more specific features for task classification of Alcohol and Controlled users rather than exposing generic features outlined by the multi-head model. It tries not to capture/disclose information specific to the user's identity. This helps us to secure EEG data by keeping the privacy of the identity of the user while performing any task classification.

We showed our model using adversarial learning secure EEG data. With just two convolution layers, we are able to receive good results, which are comparable to different complex models having multiple convolution and pooling layers. Thus we tried to maintain less time and space complexity of the model to achieve better results. In the future, we would like to expand our model from 20 users to 122 users. This will help us to make our model more generalized over a wide range of users. We would also like to see insights on how different features are been preserved over task classification. We have analyzed over one layer of the model and would like to explore how data interprets over different layers over the model. This will give us insights on how model behaves after each layer and how it expose task specific features and preserve general features to secure our data. This architecture can be expanded over different domains to get better insights on how it will react of different EEG signals.

IX. Conclusion

We explored DREAMER dataset in which an EEG signal was used to elect more than one emotion. This makes it difficult to classify emotions from one EEG signal as it contains features of more than one emotion. This has not been addressed in any previous research to the best of our knowledge. We illustrated on Alcoholic dataset with the use of a multi-headed model that the user identity gets revealed while performing task classification. We highlighted our research in preserving the identity of the user while performing task classification over EEG signals. We performed adversarial learning, which maximizes the information gained from task classification and maximizes the identity loss from user classification, which helped to secure user identity that can be exposed when performing task classification. We also highlighted how simple CNN architecture with just 2 convolution layers fetching temporal and spatial features can fetch good results, which is comparable to different complex architecture. We described how we set different hyper-parameters, which helped us to achieve good outcomes and provide better space and time complexity. Hence, with the use of simple adversarial CNN architecture and optimal parameters, we can avoid high computational complexities in comparison to the state-of-art privacy-preserving methods for EEG data.

X. Acknowledgement

We would like to thank our supervisor Dr. Garima Bajwa for her consistent guidance and feedback on this research.

References

[1] Shoka, A.A.E., Dessouky, M.M., El-Sayed, A. and Hemdan, E.E.D., 2023. An efficient CNN based epileptic seizures detection framework using encrypted EEG signals for secure telemedicine applications. Alexandria Engineering Journal, 65, pp.399-412

[2] S. Katsigiannis and N. Ramzan, ”DREAMER: A Database for Emotion Recognition Through EEG and ECG Signals From Wireless Low-cost Off-the-Shelf Devices,” in IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 1, pp. 98-107, Jan. 2018, doi: 10.1109/JBHI.2017.2688239

[3] Craik, Alexander, Yongtian He, and Jose L. Contreras-Vidal. ”Deep learning for electroencephalogram (EEG) classification tasks: a review.” Journal of neural engineering 16.3 (2019): 031001.

[4] Zhang, Pengbo, et al. ”Spectral and temporal feature learning with two-stream neural networks for mental workload assessment.” IEEE Transactions on Neural Systems and Rehabilitation Engineering 27.6 (2019): 1149-1159.

[5] Li, Rihui, et al. ”Concurrent fNIRS and EEG for brain function investigation: a systematic, methodology-focused review.” Sensors 22.15 (2022): 5865.

[6] Gui, Qiong, Zhanpeng Jin, and Wenyao Xu. ”Exploring EEG-based biometrics for user identification and authentication.” 2014 IEEE Signal Processing in Medicine and Biology Symposium (SPMB). IEEE, 2014.

[7] Sun, Yingnan, Frank P-W. Lo, and Benny Lo. ”EEG-based user identification system using 1D-convolutional long short-term memory neural networks.” Expert Systems with Applications 125 (2019): 259-267.

[8] Katsigiannis, Stamos, and Naeem Ramzan. ”DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices.” IEEE journal of biomedical and health informatics 22.1 (2017): 98-107.

[9] Damaševičius, Robertas, et al. ”Combining cryptography with EEG biometrics.” Computational intelligence and neuroscience 2018 (2018).

[10] Nguyen, Dang, et al. ”On the study of EEG-based cryptographic key generation.” Procedia computer science 112 (2017): 936-945.

[11] Zandi, Ali Shahidi, et al. ”Predicting epileptic seizures in scalp EEG based on a variational Bayesian Gaussian mixture model of zero-crossing intervals.” IEEE Transactions on Biomedical Engineering 60.5 (2013): 1401-1413.

[12] Narula, Vansh, Kexin Feng, and Theodora Chaspari. ”Preserving privacy in image-based emotion recognition through user anonymization.” Proceedings of the 2020 International Conference on Multimodal Interaction. 2020.

[13] Anwar, Mohd Ayaan, et al. ”FedEmo: A Privacy-Preserving Framework for Emotion Recognition using EEG Physiological Data.” 2023 15th International Conference on COMmunication Systems NETworkS (COMSNETS). IEEE, 2023.

[14] Hosseini, Mohammad Saleh Khajeh, et al. ”Personality-Based Emotion Recognition Using EEG Signals with a CNN-LSTM Network.” Brain Sciences 13.6 (2023): 947.

[15] Popescu, Andreea Bianca, et al. ”Privacy preserving classification of eeg data using machine learning and homomorphic encryption.” Applied Sciences 11.16 (2021): 7360.

[16] L. Ingber, EEG database, UCI machine learning repository, University of California, Irvine, School of Information and Computer Sciences, http://archive.ics.uci.edu/ml/datasets/EEG+Database (1997).

[17] Bajwa, Garima, and Ram Dantu. ”Neurokey: Towards a new paradigm of cancelable biometrics-based key generation using electroencephalograms.” computers security 62 (2016): 95-113.

[18] Singandhupe, Ashutosh, Hung Manh La, and David Feil-Seifer. ”Reliable security algorithm for drones using individual characteristics from an EEG signal.” IEEE Access 6 (2018): 22976-22986.

[19] Al-Qaysi, Z. T., et al. ”A review of disability EEG based wheelchair control system: Coherent taxonomy, open challenges and recommendations.” Computer methods and programs in biomedicine 164 (2018): 221-237.

[20] Landau, Ofir, et al. ”Mind your privacy: Privacy leakage through BCI applications using machine learning methods.” Knowledge-Based Systems 198 (2020): 105932.

[21] Kim, Sung-Jin, Dae-Hyeok Lee, and Seong-Whan Lee. ”Rethinking CNN architecture for enhancing decoding performance of motor imagery-based EEG signals.” IEEE Access 10 (2022): 96984-96996.

[22] Craik, A., He, Y. and Contreras-Vidal, J.L., 2019. Deep learning for electroencephalogram (EEG) classification tasks: a review. Journal of neural engineering, 16(3), p.031001.