Machine Learning and Deep Learning on EEG-Based Emotion Classification: A Comparative Study

Shu Ting Wong
230221855
Prof. Marcus Pearce
Artificial Intelligence (MSc)

***Abstract*— EEG-based emotion classification holds significant potential for applications across various fields, including the medical industry and autonomous vehicles. However, as a relatively new area of research, it still requires further development. In this study, we offer a comprehensive overview and comparison of different preprocessing techniques, feature extraction methods, and both machine learning and deep learning classification models. Our findings indicate that ICA is particularly effective for preprocessing. While traditional machine learning methods, such as random forest (RF), are effective for emotion classification, the potential of deep learning should not be underestimated. We propose that utilizing deeper models or more advanced hybrid approaches could further enhance classification accuracy.**

Keywords—EEG-based emotion classification, machine learning, deep learning, signal preprocessing, feature extraction

I. INTRODUCTION

Emotions and their mechanisms are among the most complex and intriguing topics in neuroscience. Emotion recognition has wide applications in diverse domains, including driving assistance, education, and healthcare (Soroush, 2018). Significant efforts have been made to automate the emotion recognition process through machines, using facial expression, speech, and body gesture analysis (Soroush, 2018). While these metrics can be accurate, they are often susceptible to social masking. With advancements in brain-computer interface (BCI) technology, brain signal analysis has gained popularity in affective computing, as electrical signals cannot easily be consciously controlled by individuals. The rapid development of machine learning and deep learning has further facilitated the creation of brain signal-based emotion classification models, providing more robust methods for analyzing non-linear and complex brain signals. Electroencephalogram (EEG) is a prevalent method for collecting brain signals due to its convenience, non-invasiveness, and cost-effectiveness (Teplan, 2002).

Despite the progress made in EEG-based emotion classification, significant effort is still needed to handle the noisy and non-linear nature of EEG signals and to understand how different classification methods interact with various preprocessing and feature extraction techniques. This paper offers a comprehensive overview of preprocessing, feature extraction, and classification techniques and their implementation. We analyzed, evaluated, and compare various machine learning and deep learning models for emotion classification tasks, alongside different preprocessing and feature extraction methods. The analysis was conducted on multiple EEG datasets, offering a deeper and more unbiased understanding of the techniques and models.

II. RELATED WORK

A. Overview

The brain-computer interface (BCI) focuses on establishing a connection between the human brain and electronic devices, with EEG signal extraction serving as a prime example of its application. This BCI technique is useful for emotion recognition, enhancing the understanding of the human mind and human-machine interaction (Soroush, 2018; Alhalaseh, 2020). It enables machines to better interpret human intentions beyond mere words and commands, fostering seamless cooperation between humans and machines. This forward-looking field warrants significant attention. Extensive research has been conducted on emotion models, the correlation between EEG signals and emotions, and methods for effectively extracting and classifying useful signals.

B. Models of Emotion

There are two main types of emotion models: dimensional and discrete. An example of discrete categorization is Ekman et al.'s six basic emotions: happiness, sadness, anger, fear, surprise, and disgust. However, a disadvantage of discrete models is their limited number of categories, which may not capture the wide range of human emotions (Bazgir 2018). Dimensional models are more widely used in research settings. Among them, the bi-dimensional valence-arousal model is the most popular one. Four emotional states, including high arousal high valence (HAHV), high arousal low valence (HALV), low arousal high valence (LAHV) and low arousal low valence (LALV) are introduced in the model.

C. EEG and Emotion

EEG measures the brain's electrical activity through several electrodes placed on the scalp. Brain signals can be decomposed into sine waves at different frequencies by Fourier decomposition. Different frequency bands are often observed with different strength in specific brain regions and correspond to different mental states (Teplan, 2002). Delta waves (1-3 Hz) are high amplitude brain waves and are associated with deep sleep. Theta band (4-8 Hz) is associated with normal sleep and deep meditation. Alpha band (8-13 Hz) is linked to eye-closing and relaxation, Beta activity (13-30 Hz) represents alertness and tension. Gamma band (30-64 Hz) is related to high cognitive processes. The power of these bands is connected to the metrics in the 3-dimensional emotion model: valence, arousal, and dominance. Higher alpha signals in the frontal area and higher beta signals in the right parietal region suggest high valence. Higher beta power in the parietal lobe and lower alpha activity suggest higher arousal. Increased beta activity indicates higher dominance (Bazgir, 2018).

D. Preprocessing and Feature Extraction

There are a few challenges posed by EEG-based emotion classification. The primary challenge being the subjectivity of interpretation towards human emotions. The lack of ground truth makes it difficult to label EEG signals completely accurately (Soroush, 2018). Moreover, EEG signals are non-linear, non-stationary and can contain noise from various sources such as eye movements and environmental interference. Therefore, effective preprocessing methods that can remove artifacts and improve signal quality, and robust feature extraction methods that can extract meaningful features from the signal are of crucial importance.

One of the most common strategies in EEG signal preprocessing is Blind Source Separation (BSS), which separates mixed signals into independent sources. Typically, EEG signals from each electrode do not directly represent the activity of a specific brain area but instead are a linear mixture of source signals generated from different sub-areas of the brain. Therefore, the BSS approach is particularly effective for clustered EEG signals, and it facilitates further analysis to reject biological artifacts such as electrocardiogram (ECG), electromyogram (EMG), and electrooculogram (EOG) activity (Stergiadis, 2022). Independent Component Analysis (ICA) is the most established BSS technique, extracting a set of source signals from a set of mixed signals. This powerful preprocessing technique plays an important role in biological signal analysis (Albera, 2012).

Feature extraction methods for EEG signals can be categorized into four main categories: time domain, frequency domain, time-frequency domain, and non-linear characteristics. Methods in different domains can be applied to different scenarios. The most commonly used methods are Power Spectral Density (PSD), Short-Time Fourier Transform (STFT), Wavelet Transform (DWT), and Empirical Mode Decomposition (EMD). Among these, PSD is the least effective despite its popularity. Methods suitable for non-stationary signals and providing good spectral estimation are preferred (Al-Fahoum, 2014).

E. Machine learning and Deep Learning Classification

Before the advancement of deep learning, various shallow machine learning models were widely used in EEG emotion classification studies, with accuracies ranging from 57.5% to 97.5%. Differences in model architecture, hyperparameters, preprocessing, and feature extraction methods contributed to the variation in accuracy. Some of the most popular shallow models included Support Vector Machine (SVM), k-Nearest Neighbor (KNN), Random Forest (RF), Naive Bayes (NB), and Extreme Learning Machine (ELM), with SVM being particularly favored for its ability to perform classification in higher dimensional spaces (Wang 2015).

Despite the effectiveness of traditional machine learning models, there are drawbacks that make recent deep learning models more favorable. A notable drawback is that shallow models directly use EEG characteristics, requiring substantial prior knowledge for feature extraction, and identifying representative and effective features can be challenging. In contrast, deep learning methods can automatically extract features without manual effort. Examples of deep learning models include Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Generative Adversarial Networks (GAN), and Long Short-Term Memory (LSTM) networks, with accuracies ranging from 63.38% to 97.56% (Wang 2015)., demonstrating significant improvement over traditional machine learning models.

The work doesn't stop at implementing deep learning models, as we still face issues such as computational complexity. Extensive efforts have led to methods like the deep forest-based multi-channel EEG emotion classification method and hybrid convolution recurrent neural networks. The former model achieved impressive average accuracies of 97.69% and 97.53% for valence and arousal, respectively, while the latter reached final classification accuracies of 93.64% and 93.26% (Wang 2015).

III. METHOD

A. Overview

The study performed an emotion classification task employing three machine learning models, namely k-NN, RF, SVM and four deep learning models, including RNNs, LSTM, CNNs and transformers. The analysis utilizes two datasets: the DEAP dataset and the DREAMER dataset. 80% of the data was allocated to the training set, while the remaining 20% was reserved for testing. Prior to training the Machine Learning models, the data underwent preprocessing and feature extraction. Scaling was also performed prior to training both Machine Learning and Deep Learning Models to standardize the features by removing the mean and scaling to unit variance. This approach removes the negative influence brought by the difference in the scale of the features in the data. Although the data was preprocessed, feature extraction was not performed before training the Deep Learning models, as these models can automatically extract features. Various preprocessing methods and feature extraction techniques were evaluated and compared alongside the machine learning and deep learning models. The specifics of these methods and the analysis workflow are detailed in this section.

B. Dataset

1. DEAP

The DEAP dataset was designed for studying human affective states and included multimodal data. It involved 32 participants who viewed 40 one-minute excerpts of music videos, during which their EEG and peripheral physiological signals were recorded. The EEG data was collected using 32 channels arranged according to the international 10-20 system. Participants also provided self-reported ratings on valence, arousal, dominance, like-dislike, and familiarity. The selection of videos was based on affective tags from the last.fm website, video highlight detection, and an online assessment tool.

To ensure data quality, the dataset underwent preprocessing by the development team. This included down sampling to 128 Hz, bandpass filtering from 4.0 to 45 Hz, segmentation into 60-second trials, and removal of the 3-second pre-trial baseline. These preprocessing steps aimed to mitigate EOG artifacts effectively (Koelstra, 2012).

2. DREAMER

The DREAMER dataset is a multimodal dataset where 23 participants were exposed to audio-visual stimuli lasting between 65 to 393 seconds. During these sessions, their EEG and electrocardiogram (ECG) signals were recorded at a sample frequency of 128 Hz. Participants also provided self-assessments of valence, arousal, and dominance levels. The signals were captured with 14 channels using the Emotiv EPOC wireless EEG headset and the Shimmer2 ECG sensor, the electrodes were placed according to the international 10-20 system (Katsigiannis, 2018).

C. Model of emotion

In this paper, emotions are represented using the 3-dimensional model featuring valence, arousal and dominance. Valence describes positive, happy emotions, while arousal describes the level of excitement. Dominance measures whether a person feels in control or empowered. The additional axis expands the traditional bi-dimensional plane, offering deeper insights into the analysis (Alhalaseh, 2020). Binary classifications were performed separately on arousal, valence, and dominance, categorizing each into high and low groups, with a threshold set at 5 for the DEAP dataset and 3 for the DREAMER dataset.

D. Workflow

Figure 1 illustrates the proposed workflow. The initial step involves preprocessing the data through bandpass filtering and Fast Independent Analysis (Fast ICA), performed separately. The analysis and comparison of the extraction methods focuses on the time-frequency domain methods. The preprocessed data is fed into Short-time Fourier Transform (STFT) and wavelet transform in the feature extraction step. The final step involves classification using machine learning classifiers (SVM, k-NN, RF) and deep learning classifiers (CNNs, RNNs, LSTMs, transformers). The results are then subjected to performance analysis.

E. Preprocessing

1.Band pass filtering

A band pass filter of 4-45Hz was applied to the datasets. Prior to the filtering, the signals were converted from the time domain to the frequency domain using fast Fourier transform (FFT). This allows for the removal of noise outside the required frequency range. For instance, electrooculogram (EOG) artifacts are a major source of noise in EEG signals and tend to concentrate at lower frequencies (Bhandari, 2007). By eliminating frequencies below 4Hz, a significant proportion of EOG artifacts can be effectively removed.

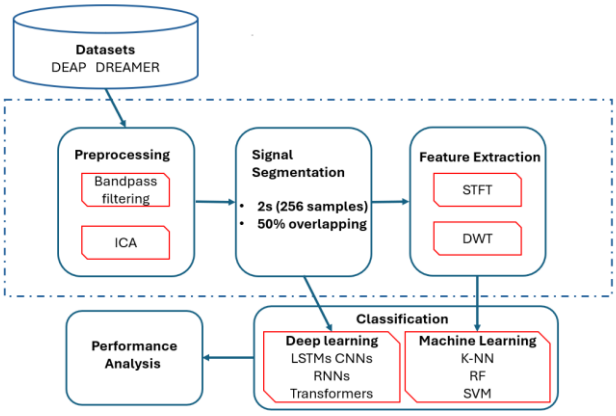


Fig.1. Flow chart of the EEG signal analysis and classification process

2.Fast ICA

In this paper, we utilize Fast ICA for preprocessing. ICA assumes that the observed signals are mixtures of source signals. An estimated matrix A transforms the source signals s(t) into the mixed observed signals x(t). A matrix W, which is the inverse of A, is constructed to decompose the observed signal into a linear combination of independent components (source signals), as shown in the equation below (Yang, 2009; Ullsperger & Debener, 2010):

$$s(t)=Wx(t) \qquad (1)$$

In this use case, we obtain source signals that correspond to the number of channels, which can be seen as virtual channels. This transformation allows data analysis in the source space, potentially improving classification performance (Pontifex. 2017).

F. Feature Extraction

1. Short-Time Fourier Transform

The significance of the band power in five frequency bands—delta (0-4 Hz), theta (4-8 Hz), alpha (8-13 Hz), beta (13-30 Hz), and gamma—is well-documented in affective computing research, where they correspond to various functions of the human brain. These frequency bands provide valuable information about emotional state, cognitive load, and sleep state. In this paper, we focused on the theta, alpha, beta, and gamma bands (Dadebayev, 2022).

To compute the band power of these frequencies, we convert the time-domain EEG data into the time-frequency domain. During the STFT operation, a sliding window segments the data into short-time signals with overlapping (Zabidi, 2012). Window sizes of 2 seconds and a 50% overlapping were implemented. The band powers were derived in each segment from the frequency domain. This results in a feature vector with a length equal to the number of channels multiplied by 4 (representing theta, alpha, beta, and gamma bands) (Katsigiannis 2018).

2. Discrete Wavelet Transform

Another method to extract features from the EEG signal is the Discrete Wavelet Transform (DWT). After segmenting the data according to the protocol mentioned in the SFST section, the signal is decomposed using DWT. This multi-scale method is more suitable for the non-stationary EEG signal. In fixed-window methods like SFST, there is a trade-off between time resolution and frequency resolution depending on the window size. The varying sized window in

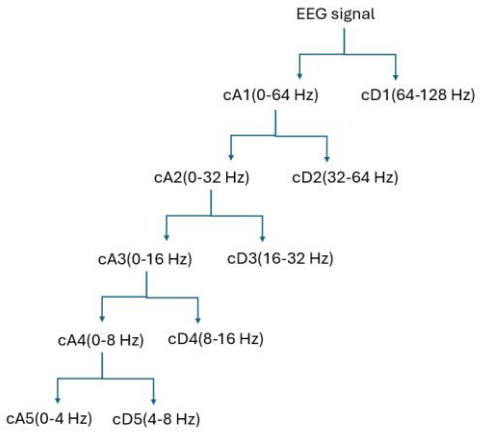


Fig.2. The tree of the multi-level decomposition of the EEG signal

DWT allows for shorter time windows for high-frequency resolution and longer windows for low-frequency resolution (Al-Fahoum 2014). This multi-resolution approach has the potential to improve the interpretation of the signal. In DWT, the signal is decomposed onto an orthonormal basis by translating and dilating a function ψ(x), described as the mother wavelet, where:

$$\psi 2j(x) = \sqrt{2}j\ \psi(2jx). \qquad (2)$$

By correlating the original signal with wavelets of different sizes, the details of the signal can be obtained at several scales in a hierarchical order (Shaker 2007; Hazarika 1997), as illustrated in Figure 2.

At each level, the signal is separated by a high-pass filter and a low-pass filter, resulting in 'details' and 'approximation,' with detailed coefficients cD and approximation coefficients cA. The detailed coefficients roughly correspond to the brainwave bands: cD5 (4-8 Hz) corresponds to the theta band, cD3 (8-16 Hz) corresponds to the alpha band, cD3 (16-32 Hz) corresponds to the beta band, and cD2 (32-64 Hz) corresponds to the gamma band. The power of the bands is derived from the detailed coefficients for further analysis. In this study, Symlets function of order 9 was chosen as the mother wavelet as it has the highest compatibility with EEG datasets, suggest by a study (Al-Qazzaz, 2015).

G. Machine Learning Models

1. k-Nearest Neighbors

k-Nearest Neighbors (k-NN) is a supervised machine learning algorithm used for classification and regression analysis. The classifier is trained using input-output pairs from the training dataset, and it classifies inputs in the test dataset based on the classifications of the k nearest neighbors. These nearest neighbors are selected based on the Euclidean distance or Mahalanobis distance from the input data point to known data points. The k-NN classifier determines the class of a data point based on its similarity to the selected neighbors (Alhalaseh 2020). In this study, the k value was set to the square root of the number of samples, and the Euclidean distance was used.

2. Random Forest

Random Forest (RF) is an ensemble method that employs a forest of decision trees for classification and regression analysis. In RF, a sample set is randomly selected from the training data with replacement for each tree. Each tree is trained independently using a random subset of features (IBM, 2023). At each level, the data is split based on the best attribute, which is the attribute that provides the most information gain. The final decision is determined by the majority vote of the predictions from all decision trees. In the implementation described in this paper, 512 trees were used in the RF algorithm (Houssein 2021).

3. Support Vector Machine

Support Vector Machine (SVM) is a supervised learning method used for classification and regression analysis. This technique maps data into higher-dimensional spaces, allowing data that is not separable in its original space to be separated by a hyperplane learned from the algorithm. This approach maximizes the classifier's margin and its generalization capability. Additionally, various kernel functions can be utilized to separate data that is not linearly separable, such as the radial basis function (RBF) kernel, which projects input vectors into a Gaussian space (Hosseini 2021; Alhalaseh 2020). In this paper, SVM will be applied to the two datasets using both a linear and an RBF kernel. The regularization parameter will be set to 1, and the kernel coefficients will be set to 'scale'.

H. Deep Learning Models

Each deep learning model was trained for 200 epochs. Binary entropy loss was utilized, and the training was optimized using the Adam optimizer with a learning rate of 0.001 and a weight decay of 1e-4 training transformer model on the ICA-preprocessed DEAP dataset being an exception. An Stochastic Gradient Descent (SGD) with momentum of 0.75 and a learning rate of 0.01 were used. This measure was taken because the training was potentially stuck in a local minimum. The model was also trained for 300 epochs because of slow convergence. Dropouts of different values were implemented in the models to prevent overfitting and improve the generalization ability. The details of the models will be explained in this section.

1. Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a type of deep learning model designed to handle sequential data. In an RNN, the output is influenced not only by the current input but also by previous inputs. The current state of the network is computed using the equation:

$$h_t=\tanh(W_{hh}\ h_{t-1}+W_{Xh}\ X_t) \qquad (3)$$

where h_t and h_{t-1} represent the current and the previous states, respectively. W_{hh} stands for recurrent neuron weight and X_t is the input. The output of the network is computed by the equation:

$$Y_t=W_{hy}\ h_t \qquad (4)$$

where Y_t denotes the output and W_{hy} represents the weights of the output layer. It is important to note that RNNs are susceptible to the vanishing and exploding gradient problem (Kalpana Chowdary, 2022).

Figure 3 shows the architecture of the RNN used in this study. The number of units is determined by the number of channels in the dataset. For the DEAP dataset, which has 32 channels, the RNN consists of 64 units (twice the number of channels). Similarly, the DREAMER dataset, with 14 channels, was trained using an RNN with 28 units.

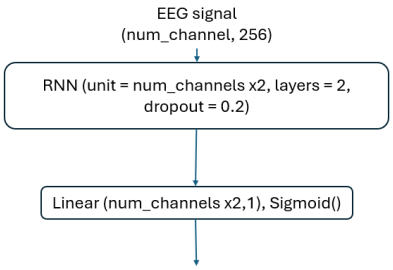


Fig.3. The architecture of the RNN implementation

2. Long-Short Term Memory

Long Short-Term Memory (LSTM) is an advanced variant of Recurrent Neural Networks (RNNs) designed to address the vanishing gradient problem commonly encountered by traditional RNNs. LSTM networks feature three gates: the forget gate, the input gate, and the output gate. These gates regulate whether data should be stored or discarded. Similar to RNNs, LSTM's short-term memory relies on the current hidden state, which is influenced by both the previous hidden state and the current input. In addition, LSTM introduces a long-term memory component known as the cell state. The cell state integrates new input with the previous hidden state and the previous cell state via the operations of the forget gate and input gate. The forget gate's function is represented by the equation:

$$ft = \sigma(X_t * U_f + H_{t-1} * W_f) \quad (5)$$

where W_f and U_f are the weights associated with the forget gate, X_t is the input at time t, and H_{t-1}represents the previous hidden state. The output f_t, which ranges between 0 and 1, determines how much information from the previous cell state will be retained. The cell state is updated according to the equation:

$$C_t = f_t * C_{t-1} + I_t * C\hat{}t \quad (6)$$

Here, I_t is the output of the input gate, calculated using the sigmoid function on the weighted sum of the input and the previous hidden state. C^t represents the tanh activation of the weighted sum of the input and hidden state (with different weights). The final output is computed at the output gate, which combines the result of the output gate equation, involving the sigmoid activation of the weighted sum of the input and previous hidden state, with the tanh activation of the cell state (Kalpana Chowdary, 2022).

The gating mechanism in LSTM differentiates it from standard RNNs, offering improved long-range dependency handling and better context representation. The architecture of the LSTM implemented in this study is shown in Figure 4. It follows the same design logic as the RNN implementation, with RNN units replaced by LSTM units.

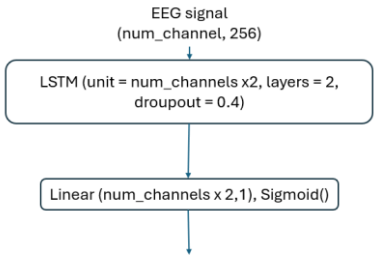


Fig.4. The architecture of the LSTM implementation

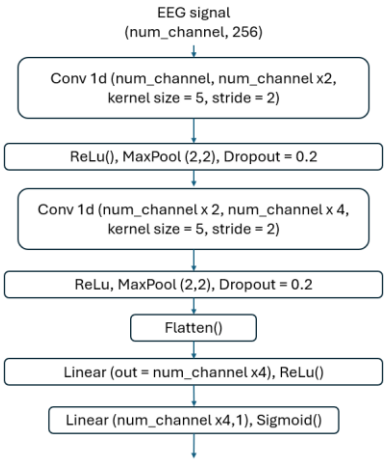


Fig. 5. The architecture of the CNN implementation

3. Convolutional Neural Networks

In this paper, a 1-D convolutional neural network (CNN) is implemented. In CNNs, convolutional kernels are applied to the input signal using a sliding window. The values of the neurons in the convolutional layer are computed by multiplying the input by the kernel weights and adding an offset bias. The output from the convolutional layer is then passed to the next layer. This layer-by-layer approach hierarchically extracts features, starting with low-level features from the input and progressively deriving higher-level features. The ability to extract abstract features and classify them makes CNNs well-suited for handling the non-linear and non-stationary nature of EEG signals (Lun,2020). Figure 5 illustrates the architecture of the CNN implemented in this study. The design includes two convolutional layers with ReLU activation, each followed by a max-pooling layer with a kernel size of 2. The final portion of the architecture consists of two fully connected linear layers, with sigmoid activation at the end for binary classification.

4. Transformers

The Transformer model, which is based on attention mechanisms, excels in handling sequential data by capturing long-range dependencies effectively. The multi-head attention mechanism allows the model to focus on specific parts of the input. The attention function maps the query and a set of key-value pairs to an output using the scaled-dot product, represented by the equation:

$$\text{Attention}(Q,K,V) = \text{softmax}(QK^T // \text{sqrt}(d_k))V \quad (7)$$

Scaled Dot Product Attention blocks are concatenated to form multi-head attention, as illustrated in Figure 6. Different weight metrics enable the exploration of patterns across various domains (Zeynali, 2023).

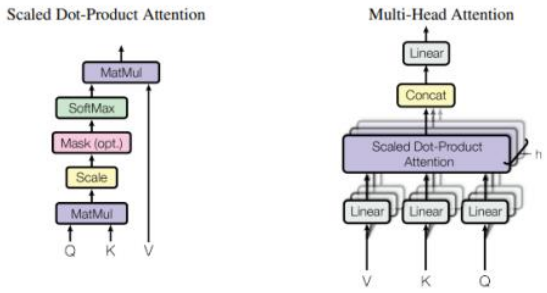


Fig. 6. A representation of the multi-head attention mechanism (Vaswani,2017)

In this paper, a multi-head attention block with 4 heads is used for the DEAP dataset and 2 heads for the DREAMER dataset. Each network comprises 4 Transformer blocks with a dropout rate of 0.25. The classification part of the network consists of 2 linear layers; the first layer includes ReLU activation and a dropout rate of 0.4, with an input size equal to the number of channels in the dataset and an output that doubles the input size. The final linear layer produces a binary output with sigmoid activation.

I. Evaluation Metrics

The evaluation metrics used in this paper are accuracy and F1 score. Accuracy is determined by the ratio of correct classifications to the total number of predictions. The F1 score, which ranges from 0 to 1, combines precision and recall, as shown in the following equation:

$$F1 = 2*precision*recall/(precision+recall) \quad (8)$$

In this equation, precision represents the rate of true positives among the positive predictions, while recall represents the ratio of true positives to the total number of positive examples.

IV. RESULTS

A. Machine Learning

Figures 7 and 8 present bar charts depicting the accuracies of the machine learning models trained on the DEAP and DREAMER datasets, respectively. Each figure includes the accuracies of 4 machine learning models. In each model, 4 groups of clustered bars representing the four combinations of preprocessing and feature extraction methods: STFT, ICA-STFT, DWT, and ICA-DWT. The classifiers achieved accuracies ranging from 49.10% to 97.73% (Extended Figures 1 and 2). The lowest accuracy was recorded by the RF valence classifier trained on the DEAP dataset using ICA for preprocessing and DWT for feature extraction, while the highest was achieved by the RF arousal classifier trained on the DEAP dataset using ICA for preprocessing and STFT for feature extraction.

Overall, the machine learning classifiers demonstrated satisfactory performance. A majority (36 out of 48) of the classifiers trained on the DEAP dataset exceeded 60% accuracy (Extended Figure 1), while most (30 out of 48) of those trained on the DREAMER dataset surpassed 80% accuracy (Extended Figure 2). Classifiers trained on the DREAMER dataset generally outperformed those trained on the DEAP dataset. Moreover, classifiers consistently performed better when data was preprocessed with ICA and features were extracted using STFT. Among the models trained on ICA-STFT data, RF classifiers outperformed others, achieving accuracies between 95.43% and 97.73% on the DEAP dataset and between 96.16% and 96.23% on the DREAMER dataset.

The linear SVM classifiers were the least performing models, with accuracies ranging from 55.91% to 60.77% on the DEAP dataset and 59.64% to 76.81% on the DREAMER dataset. Notably, the linear SVM consistently exhibited lower accuracies compared to its counterpart using the RBF kernel (Extended Figures 1 and 2).

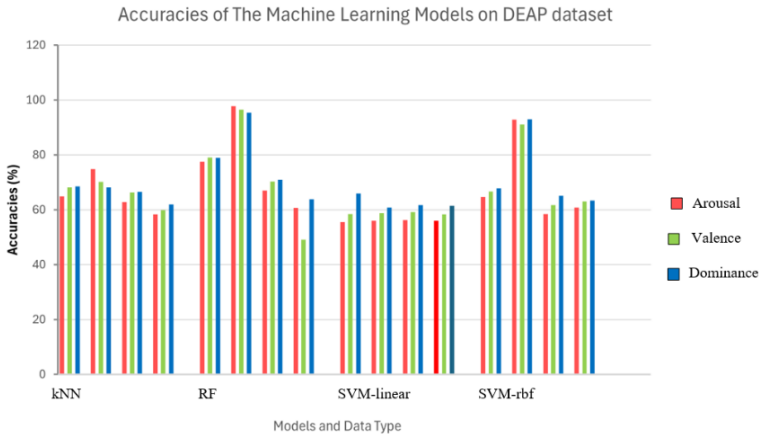


Fig.7. A bar chart illustrating the accuracies of the machine learning models on the DEAP dataset. Each model has 4 groups of clustered bars, referring to 4 combinations of preprocessing and feature extraction methods: STFT, ICA-STFT, DWT, and ICA-DWT.

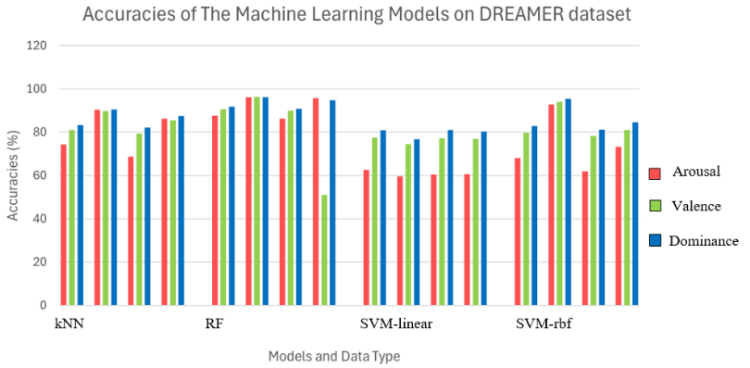


Fig.8. A bar chart illustrating the accuracies of the machine learning models on the DREAMER dataset. Each model has 4 groups of clustered bars, referring to 4 combinations of preprocessing and feature extraction methods: STFT, ICA-STFT, DWT, and ICA-DWT.

B. *Deep Learning*

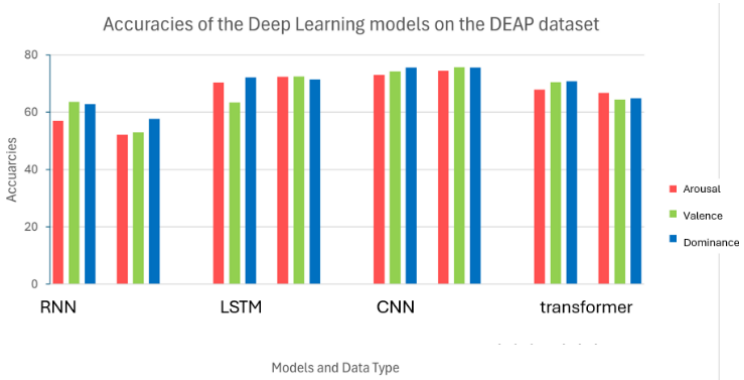


Fig.9. A bar chart illustrating the accuracies of the deep learning models on the DEAP dataset. Each model has 2 groups of clustered bars, referring to raw data and ICA-preprocessed data.

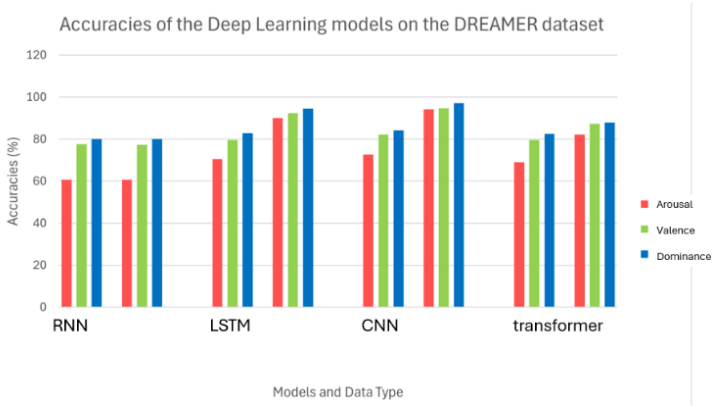


Fig.10. A bar chart illustrating the accuracies of the deep learning models on the DREAMER dataset. Each model has 2 groups of clustered bars, referring to raw data and ICA-preprocessed data

Figures 9 and 10 illustrate bar charts showing the accuracies of deep learning models trained on the DEAP and DREAMER datasets, respectively. Each model is represented by two groups of clustered bars, corresponding to raw data and ICA-preprocessed data. The deep learning classifiers achieved accuracies ranging from 56.94% to 97.14% (Extended Figures 19 and 20). The lowest accuracy was observed in the RNN arousal classifier trained on the raw DEAP dataset (Extended Figure 19), while the highest accuracy was recorded by the CNN valence classifier trained on the ICA-preprocessed DREAMER dataset (Extended Figure 20).

Similar to the machine learning classifiers, the deep learning models showed satisfactory performance. Excluding the RNN, which failed to exhibit learning behavior or improve accuracy over epochs, the accuracies of the deep learning models trained on the DEAP dataset ranged from 63.32% to 75.67%, while those trained on the DREAMER dataset ranged from 69.03% to 97.17%. Consistent with the machine learning models, classifiers trained on the DREAMER dataset generally performed better. The accuracy differences between raw and ICA-preprocessed data in the DEAP dataset varied depending on the model and the arousal. Valence and dominance domain, no clear trend was observed. However, classifiers trained on the ICA-preprocessed DREAMER dataset consistently outperformed those trained on the raw DREAMER dataset (Extended Figures 19 and 20).

Among the models trained on ICA-preprocessed data, CNN classifiers outperformed the others, achieving accuracies between 74.46% and 75.67% on the DEAP dataset and between 94.19% and 97.17% on the DREAMER dataset (Extended Figures 19 and 20).

V. DISCUSSION

A. *Comparsion of The Preprocessing and Feature Extraction Methods*

In the results section, it was noted that machine learning classifiers consistently performed better when the data was preprocessed with ICA and features were extracted using STFT. Similarly, deep learning classifiers, where feature extraction was handled by the network itself, also benefited from ICA preprocessing. This suggests that ICA is a robust preprocessing method for handling EEG data, aligning with findings from previous research (Pontifex. 2017).. However, the results of this study regarding feature extraction methods show that STFT outperformed DWT, which contradicts prior research that identified DWT as a more advanced feature extraction technique yielding superior results compared to STFT(Al-Fahoum 2014). A plausible explanation for this discrepancy is that the 2-second window size used for segmentation may have been insufficient for the DWT operation. As a result, boundary effects likely occurred in the data processed with DWT, limiting the performance of the classifiers trained on this data.

B. *Comparsion of The Performance of The Datasets*

Models trained on the DREAMER dataset outperformed those trained on the DEAP dataset. A possible explanation for this is that the DEAP dataset includes more channels, resulting in a larger number of features for machine learning and deep learning classification. This added complexity can make models more prone to overfitting, which limits their performance. To address this issue, feature selection methods can be employed, either through manual channel selection by experts or automatic techniques such as Principal Component Analysis (PCA). By focusing on the most effective channels for classification and discarding redundant information, the performance of the models can be improved.

C. *Performance of The Machine Learning and Deep Learning Models*

To evaluate and compare the models' performance more clearly, we focused on models trained using the best-performing preprocessing and feature extraction methods: ICA preprocessing combined with STFT for machine learning models and ICA preprocessing for deep learning models.

For the machine learning models, SVM with a linear kernel performed the worst. However, SVM with an RBF kernel showed strong performance across both datasets, trailing the top-performing RF models by only 2.57%-5.45% on the DEAP dataset and 0.87%-3.43% on the DREAMER dataset. This indicates that while SVM is effective, the choice of kernel is crucial. It also suggests that the EEG data may not be linearly separable, even in a higher-dimensional space. The RF model achieved the best accuracies, demonstrating its potential despite being less commonly used than SVM.

For deep learning models, RNN was the least performing. Although RNN achieved over 60% accuracy on the DREAMER dataset (Extended Figure 20), the training loss, accuracy, and F1 score plots revealed that the training losses were not decreasing over time, and the accuracies were not improving across epochs (Extended Figures 4 and 6). This suggests that the RNN networks were not learning effectively, likely due to the exploding and vanishing gradient problem, where extreme gradient values accumulate and hinder the network's learning. This explanation is supported by the fact that LSTM networks with similar architectures performed well on both the DEAP and DREAMER datasets, particularly on the DREAMER dataset, where accuracies across the arousal, valence, and dominance domains exceeded 90%. LSTMs address the exploding and vanishing gradient issues by incorporating long-short term memory, which helps maintain a better representation of context and filters out irrelevant information through the forget gate (Kalpana Chowdary, 2022).

Another model designed to handle sequential data, the transformer, performed adequately on both EEG datasets. Interestingly, the transformer model performed better on the raw DEAP dataset than on the ICA-preprocessed DEAP dataset. Initially, model accuracy did not improve over epochs with the original scheme using the Adam optimizer with a learning rate of 0.001. However, switching to an SGD optimizer with momentum of 0.75 and a learning rate of 0.01 over 300 epochs resolved the issue, suggesting that the model had been stuck in a local minimum. This indicates that there may still be room to optimize the transformer model.

CNN achieved the highest accuracies among the deep learning models, indicating that the hierarchical feature extraction process, layer by layer, works well for EEG data (Lun,2020). Compared to other deep learning models, the CNN in this study had a deeper architecture, suggesting that a hierarchical approach and deeper networks could further improve performance. While CNN did not perform as well as the top-performing RF model, it has the advantage of not requiring manual feature extraction. A potential improvement would be to combine CNN-extracted features with an RF classifier, creating a hybrid model that could push performance even further.

VI. FUTURE WORK AND CONCLUSION

In conclusion, effective preprocessing and feature extraction methods can significantly enhance the performance of classification models. This study demonstrated that ICA and STFT are effective as preprocessing and feature extraction methods, respectively. To prevent boundary effects in DWT, longer segment lengths should be used. RF and CNN were the best-performing machine learning and deep learning models. Testing a hybrid model combining these two approaches is recommended for future work. The scope of the study can also be expanded to gain a deeper understanding and provide further insights on the topic, including but not limited to introducing cross-validation for different hyperparameters, experimenting with different window sizes for segmentation, using feature selection methods such as PCA, and exploring more advanced hybrid models. Overall, this study provides a solid foundation for further research on EEG-based emotion classification.

ACKNOWNDEMENT

I would like to thank my incredible supervisor prof. Marcus Pearce for his support and insightful guidance.

REFERENCES

Soroush, M.Z., Maghooli, K., Setarehdan, S.K. and Nasrabadi, A.M., 2018. Emotion classification through nonlinear EEG analysis using machine learning methods. *Department of Biomedical Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran; Control and Intelligent Processing Centre of Excellence, School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran; Department of Biomedical Engineering, Faculty of Engineering, Shahed University, Tehran, Iran*.

Bazgir, O., Mohammadi, Z. and Habibi, S.A.H., 2018. Emotion recognition with machine learning using EEG signals. In: *2018 25th National and 3rd International Iranian Conference on Biomedical Engineering (ICBME)*, Qom, Iran, 2018, pp. 1-5. doi: 10.1109/ICBME.2018.8703559.

Alhalaseh, R. and Alasasfeh, S., 2020. Machine-Learning-Based Emotion Recognition System Using EEG Signals. *Computers*, 9(4), p.95.

Doma, V. and Pirouz, M., 2020. A comparative analysis of machine learning methods for emotion recognition using EEG and peripheral physiological signals. *Journal of Big Data*, 7(18).

Fink, A., Grabner, R.H., Neuper, C. and Neubauer, A.C., 2005. EEG alpha band dissociation with increasing task demands. *Research report*. Institute of Psychology, University of Graz, Universitaetsplatz 2/III, A-8010 Graz, Austria.

Koelstra, S., Muehl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I. (2012). DEAP: A Database for Emotion Analysis using Physiological Signals (PDF). IEEE Transactions on Affective Computing, 3(1), 18-31.

Katsigiannis, S., & Ramzan, N. (2018). 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), 98-107.

Teplan, M., 2002. Fundamentals of EEG measurement. *Measurement Science Review*, 2(2). Institute of Measurement Science, Slovak Academy of Sciences, Dúbravská cesta 9, 841 04 Bratislava, Slovakia.

Ullsperger, M., & Debener, S. (2010). Simultaneous EEG and fMRI: Recording, analysis, and application. *Using ICA for the analysis of Multichannel EEG data* (pp. 121-123). Oxford University Press.

Bhandari, A., Khare, V., Santhosh, J., & Anand, S. (2007). Wavelet based compression technique of Electro-oculogram signals. In A. Abu Osman, K. Ibrahim, & S. M. Naqshbandi (Eds.), *3rd Kuala Lumpur International Conference on Biomedical Engineering 2006* (pp. 440-443). IFMBE Proceedings, 15. doi:10.1007/978-3-540-68017-8_111

Zheng, W.L., & Lu, B.L. (2015). Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3), 162-175.

Duan, R.N., Zhu, J.Y., & Lu, B.L. (2013). Differential entropy feature for EEG-based emotion classification. *Proceedings of the 6th International IEEE EMBS Conference on Neural Engineering (NER)*, 81-84.

Islam, M.K., Rastegarnia, A. and Yang, Z., 2016. Methods for artifact detection and removal from scalp EEG: A review.

Yang, L., Ming, Z., & Longbin, J. (2009). Blind Source Separation Based on FastICA. *Ninth International Conference on Hybrid Intelligent Systems*, Shenyang, China, 2009, pp. 475-479. doi: 10.1109/HIS.2009.212.

Zabidi, A., Mansor, W., Lee, Y. K. & Che Wan Fadzal, C. W. N. F. (2012). Short-time Fourier Transform analysis of EEG signal generated during imagined writing. *2012 International Conference on System Engineering and Technology (ICSET)*, Bandung, Indonesia, 2012, pp. 1-4. doi: 10.1109/ICSEngT.2012.6339284.

Pontifex, M.B., Gwizdala, K.L., Parks, A.C., Billinger, M. & Brunner, C. (2017). Variability of ICA decomposition may impact EEG signals when used to remove eyeblink artifacts. *Psychophysiology*, 54(3), pp. 386-398. doi: 10.1111/psyp.12804. Epub 2016 Dec 27. PMID: 28026876; PMCID: PMC5584537.

Al-Fahoum, A. S., & Al-Fraihat, A. A. (2014). Methods of EEG signal features extraction using linear analysis in frequency and time-frequency domains. *International Scholarly Research Notices*, https://doi.org/10.1155/2014/730218.

Shaker, M. M. (2007). EEG waves classifier using wavelet transform and Fourier transform. *World Academy of Science, Engineering and Technology, International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering*, 1(3).

Hazarika, N., Chen, J. Z., Tsoi, A. C., & Sergejew, A. (1997). Classification of EEG signals using the wavelet transform. *Signal Processing*, 59, 61-72.

Stergiadis, C., Kostaridou, V.-D. and Klados, M.A., 2022. Which BSS method separates better the EEG Signals? A comparison of five different algorithms. *Biomedical Signal Processing and Control*, 72(Part A), p.103292. Available at: https://doi.org/10.1016/j.bspc.2021.103292

Albera, L., Kachenoura, A., Comon, P., Karfoul, A., Wendling, F., et al., 2012. ICA-based EEG denoising: a comparative analysis of fifteen methods. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 60(3 Special issue on Data Mining in Bioengineering), pp.407-418. Available at: https://doi.org/10.2478/v10175-012-0052-3.

Al-Fahoum, A.S. and Al-Fraihat, A.A., 2014. Methods of EEG signal features extraction using linear analysis in frequency and time-frequency domains. *International Scholarly Research Notices*, [online] Available at: https://doi.org/10.1155/2014/730218

Al-Qazzaz, N.K., Bin Mohd Ali, S.H., Ahmad, S.A., Islam, M.S. & Escudero, J. (2015) 'Selection of Mother Wavelet Functions for Multi-Channel EEG Signal Analysis during a Working Memory Task', *Sensors (Basel)*, 15(11), pp. 29015-29035. doi: 10.3390/s151129015. PMID: 26593918; PMCID: PMC4701319.

Wang, J. & Wang, M. (2021) 'Review of the emotional feature extraction and classification using EEG signals', *Cognitive Research: Principles and Implications*. doi: 10.1016/j.cogr.2021.04.001. Available at: https://doi.org/10.1016/j.cogr.2021.04.001

Hosseini, M.-P., Hosseini, A., & Ahi, K. (2021). A Review on Machine Learning for EEG Signal Processing in Bioengineering. *IEEE Reviews in Biomedical Engineering, 14*, 204-218. doi:10.1109/RBME.2021.3066181

Saunders, D., 2022 Domain Adaptation for Neural Machine Translation: A Survey. University of Cambridge, Engineering Department, Cambridge, United Kingdom

Houssein, E.H., Hammad, A. & Ali, A.A. Human emotion recognition from EEG-based brain–computer interface using machine learning: a comprehensive review. *Neural Comput & Applic* **34**, 12527–12557 (2022). https://doi.org/10.1007/s00521-022-07292-4

IBM (2023). *What is Random Forest? | IBM*. [online] www.ibm.com. Available at: https://www.ibm.com/topics/random-forest.

Kalpana Chowdary, M., Anitha, J., & Hemanth, D. J. (2022). Emotion recognition from EEG signals using recurrent neural networks. *Electronics*, *11*(15), 2387. https://doi.org/10.3390/electronics11152387

Lun, X., Yu, Z., Chen, T., Wang, F., & Hou, Y. (2020). A simplified CNN classification method for MI-EEG via the electrode pairs signals. *Frontiers in Human Neuroscience*, 14, Article 338. https://doi.org/10.3389/fnhum.2020.00338

Zeynali, M., Seyedarabi, H. and Afrouzian, R., 2023. Classification of EEG signals using Transformer-based deep learning and ensemble models. Biomedical Signal Processing and Control, 86(Part A), p.105130. Available at: https://doi.org/10.1016/j.bspc.2023.105130

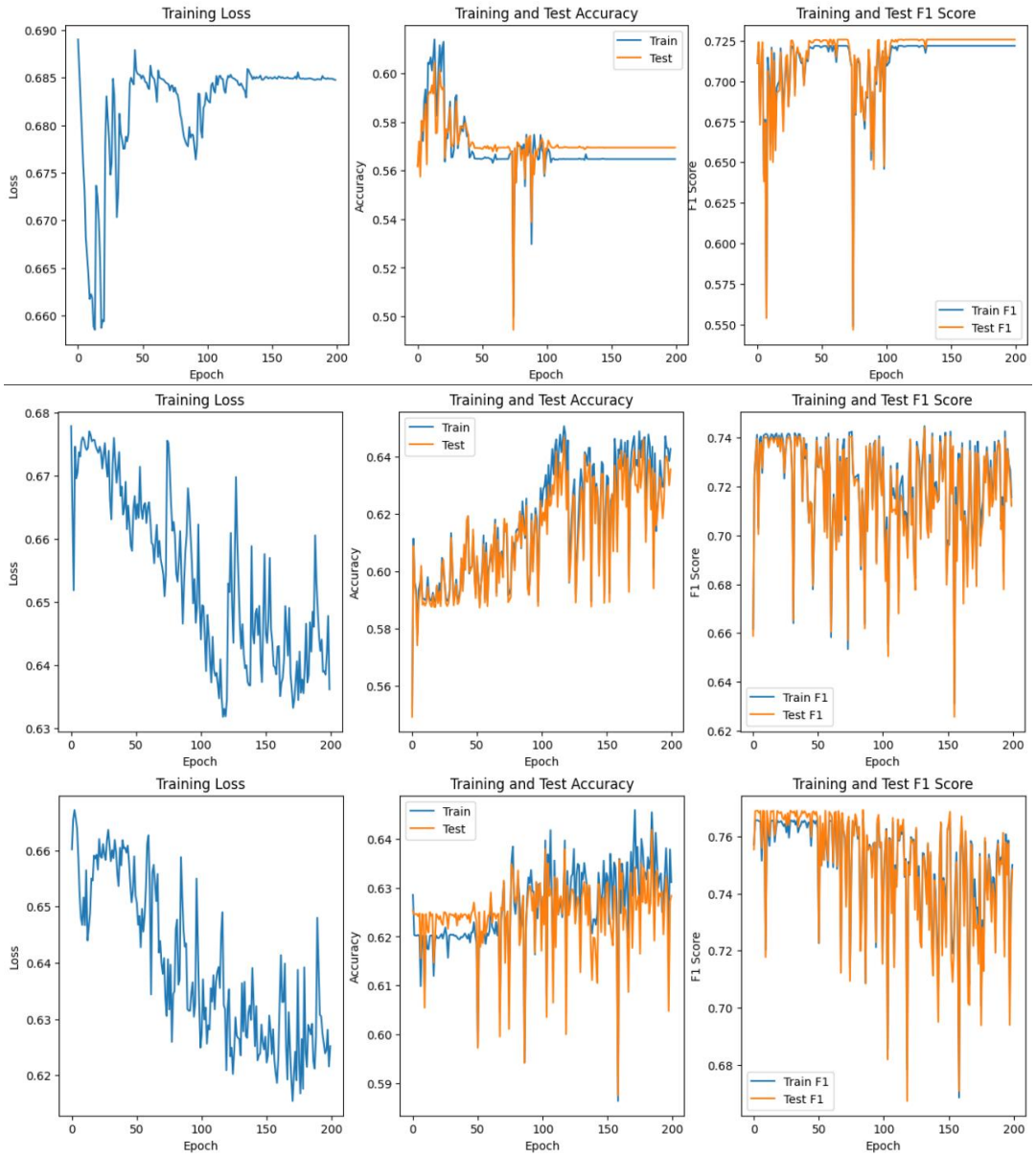
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I., 2017. Attention Is All You Need. arXiv preprint arXiv:1706.03762 Available at: https://doi.org/10.48550/arXiv.1706.03762

model	data type	emotion	accuracy (%)	f1-score
kNN	STFT	Arousal	64.91	0.64
kNN	STFT	Valence	68.11	0.67
kNN	STFT	Dominance	68.43	0.67
kNN	ICA-SFST	Arousal	74.76	0.73
kNN	ICA-SFST	Valence	70.07	0.66
kNN	ICA-SFST	Dominance	68.18	0.60
kNN	DWT	Arousal	62.70	0.62
kNN	DWT	Valence	66.20	0.65
kNN	DWT	Dominance	66.52	0.65
kNN	ICA-DWT	Arousal	58.28	0.48
kNN	ICA-DWT	Valence	59.88	0.47
kNN	ICA-DWT	Dominance	61.96	0.48
RF	STFT	Arousal	77.49	0.77
RF	STFT	Valence	78.98	0.79
RF	STFT	Dominance	78.87	0.79
RF	ICA-SFST	Arousal	97.73	0.98
RF	ICA-SFST	Valence	96.46	0.96
RF	ICA-SFST	Dominance	95.43	0.95
RF	DWT	Arousal	66.97	0.66
RF	DWT	Valence	70.21	0.69
RF	DWT	Dominance	70.94	0.70
RF	ICA-DWT	Arousal	60.70	0.52
RF	ICA-DWT	Valence	49.10	0.49
RF	ICA-DWT	Dominance	63.83	0.53
SVM-linear	STFT	Arousal	55.54	0.48
SVM-rbf	STFT	Arousal	64.66	0.63
SVM-linear	STFT	Valence	58.41	0.59
SVM-rbf	STFT	Valence	66.55	0.65
SVM-linear	STFT	Dominance	65.93	0.63
SVM-rbf	STFT	Dominance	67.78	0.66
SVM-linear	ICA-SFST	Arousal	55.91	0.43
SVM-rbf	ICA-SFST	Arousal	92.76	0.93
SVM-linear	ICA-SFST	Valence	58.78	0.46
SVM-rbf	ICA-SFST	Valence	91.01	0.91
SVM-linear	ICA-SFST	Dominance	60.77	0.51
SVM-rbf	ICA-SFST	Dominance	92.86	0.93
SVM-linear	DWT	Arousal	56.16	0.42
SVM-rbf	DWT	Arousal	58.47	0.5
SVM-linear	DWT	Valence	59.16	0.46
SVM-rbf	DWT	Valence	61.66	0.55
SVM-linear	DWT	Dominance	61.71	0.49
SVM-rbf	DWT	Dominance	65.13	0.58
SVM-linear	ICA-DWT	Arousal	56.01	0.42
SVM-rbf	ICA-DWT	Arousal	60.76	0.54
SVM-linear	ICA-DWT	Valence	58.27	0.42
SVM-rbf	ICA-DWT	Valence	63.01	0.55
SVM-linear	ICA-DWT	Dominance	61.42	0.48
SVM-rbf	ICA-DWT	Dominance	63.28	0.51

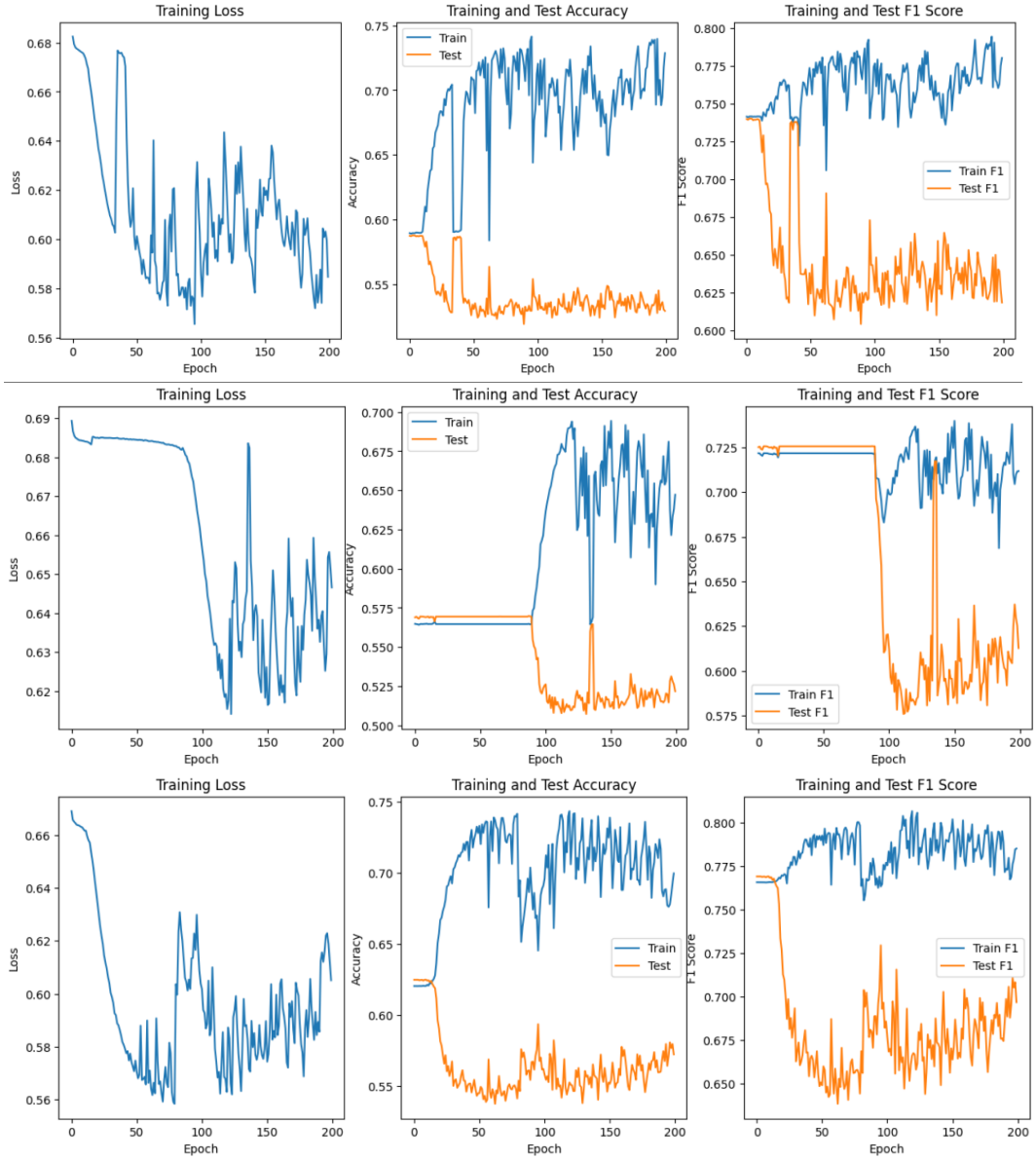
Extended Figure 1. Accuracies and F1-scores of the Machine Learning Classification Models on the DEAP dataset

model	data type	emotion	accuracy (%)	f1-score
kNN	STFT	Arousal	74.23	0.73
kNN	STFT	Valence	80.96	0.77
kNN	STFT	Dominance	83.26	0.80
kNN	ICA-SFST	Arousal	90.41	0.90
kNN	ICA-SFST	Valence	89.68	0.89
kNN	ICA-SFST	Dominance	90.43	0.89
kNN	DWT	Arousal	68.63	0.66
kNN	DWT	Valence	79.29	0.74
kNN	DWT	Dominance	82.12	0.77
kNN	ICA-DWT	Arousal	86.20	0.86
kNN	ICA-DWT	Valence	85.47	0.83
kNN	ICA-DWT	Dominance	87.38	0.85
RF	STFT	Arousal	87.64	0.87
RF	STFT	Valence	90.71	0.90
RF	STFT	Dominance	91.67	0.91
RF	ICA-SFST	Arousal	96.16	0.96
RF	ICA-SFST	Valence	96.27	0.96
RF	ICA-SFST	Dominance	96.18	0.96
RF	DWT	Arousal	86.15	0.86
RF	DWT	Valence	89.85	0.89
RF	DWT	Dominance	90.78	0.9
RF	ICA-DWT	Arousal	95.70	0.96
RF	ICA-DWT	Valence	50.95	0.55
RF	ICA-DWT	Dominance	94.81	0.95
SVM-linear	STFT	Arousal	62.59	0.53
SVM-rbf	STFT	Arousal	68.08	0.64
SVM-linear	STFT	Valence	77.57	0.68
SVM-rbf	STFT	Valence	79.60	0.74
SVM-linear	STFT	Dominance	80.94	0.73
SVM-rbf	STFT	Dominance	82.88	0.78
SVM-linear	ICA-SFST	Arousal	59.64	0.50
SVM-rbf	ICA-SFST	Arousal	92.73	0.93
SVM-linear	ICA-SFST	Valence	74.51	0.69
SVM-rbf	ICA-SFST	Valence	93.96	0.94
SVM-linear	ICA-SFST	Dominance	76.81	0.72
SVM-rbf	ICA-SFST	Dominance	95.31	0.95
SVM-linear	DWT	Arousal	60.36	0.46
SVM-rbf	DWT	Arousal	61.94	0.49
SVM-linear	DWT	Valence	77.20	0.68
SVM-rbf	DWT	Valence	78.27	0.70
SVM-linear	DWT	Dominance	81.06	0.73
SVM-rbf	DWT	Dominance	81.19	0.74
SVM-linear	ICA-DWT	Arousal	60.62	0.46
SVM-rbf	ICA-DWT	Arousal	73.20	0.70
SVM-linear	ICA-DWT	Valence	76.84	0.67
SVM-rbf	ICA-DWT	Valence	81.04	0.76
SVM-linear	ICA-DWT	Dominance	80.31	0.72
SVM-rbf	ICA-DWT	Dominance	84.60	0.80

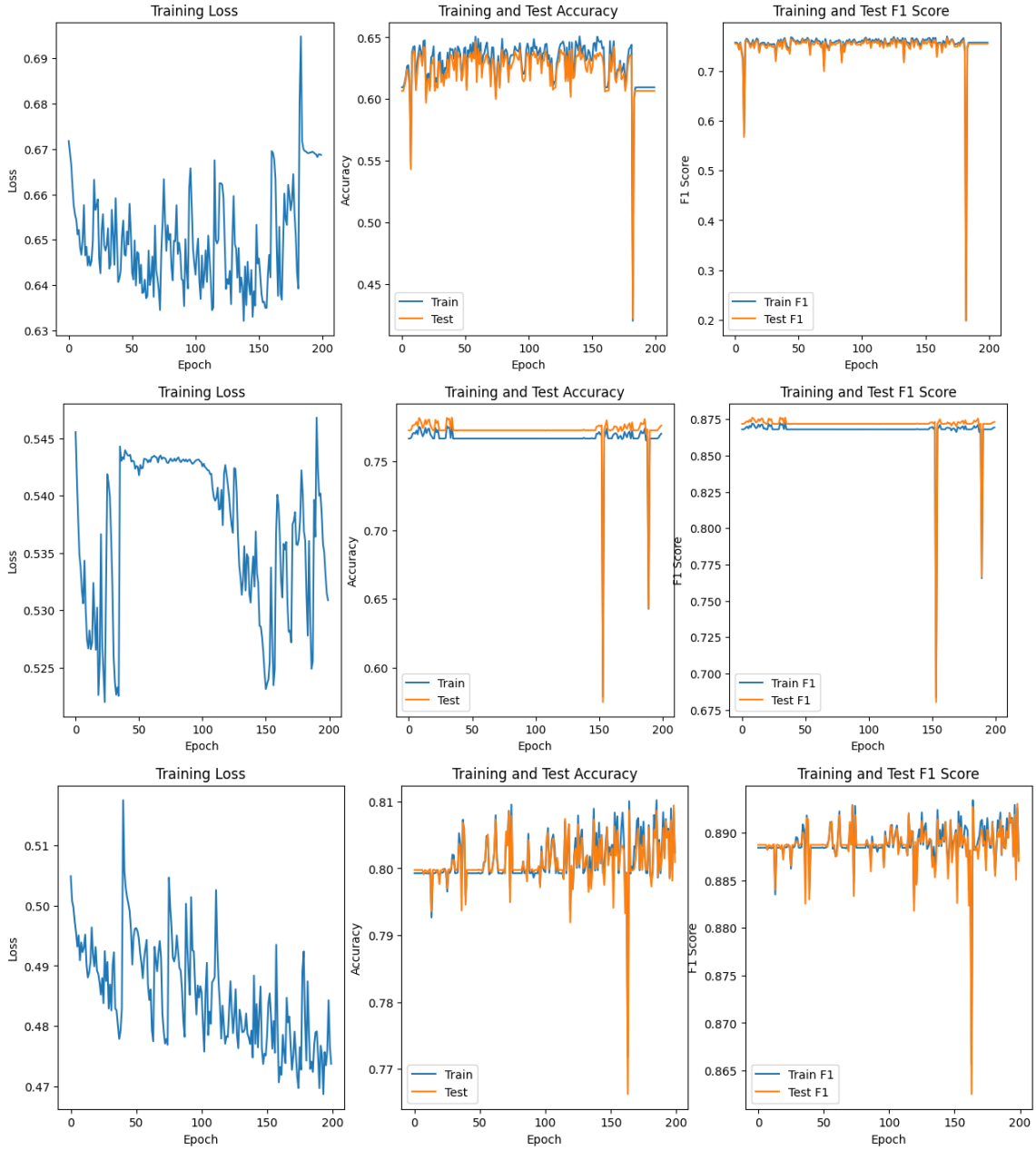
Extended Figure 2. Accuracies and F1-scores of the Machine Learning Classification Models on the DREAMER dataset



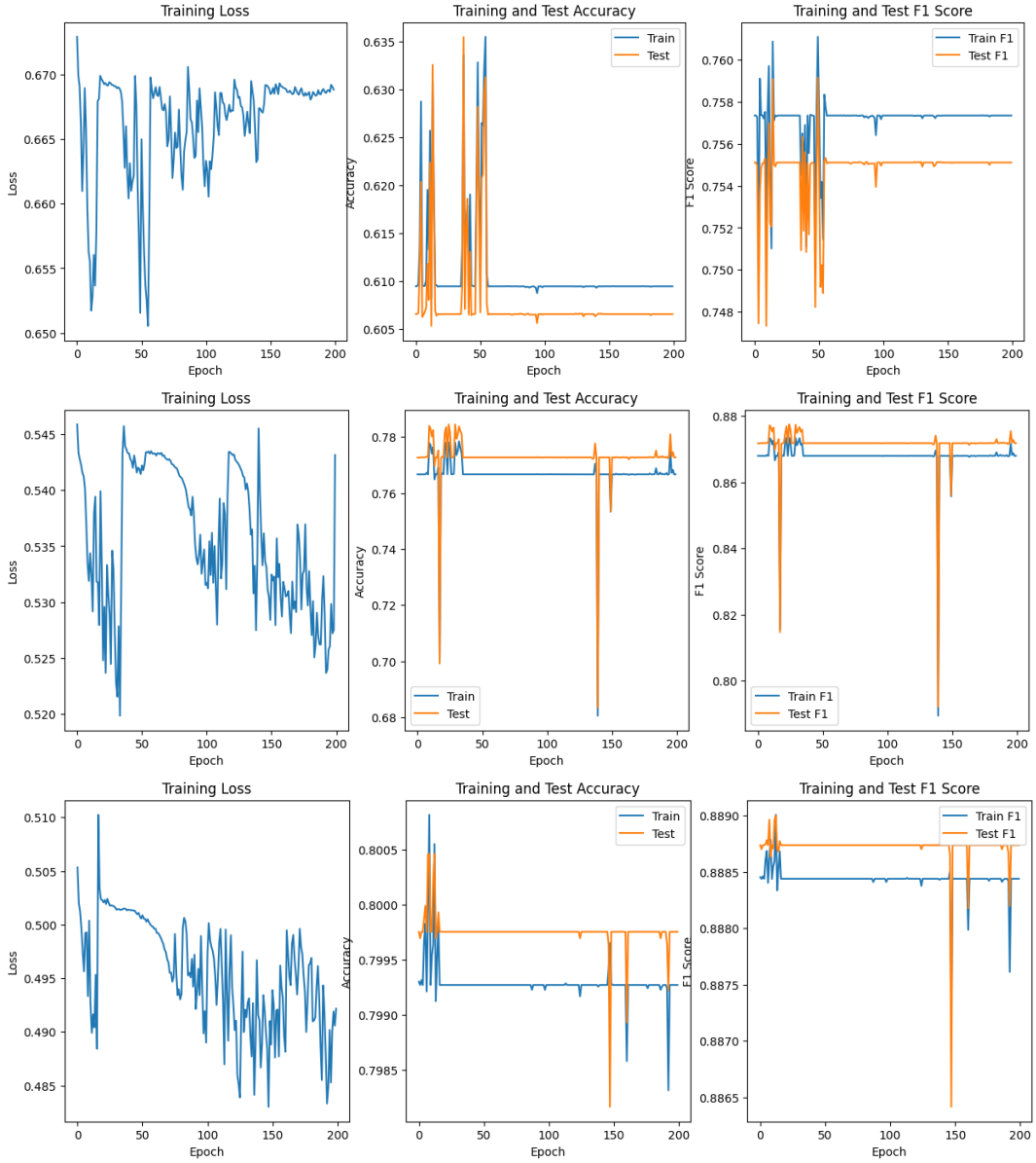
Extended Figure 3. Training loss, training and test accuracy, and training and test F1 score of RNN arousal, valence and dominance classifier on DEAP dataset.



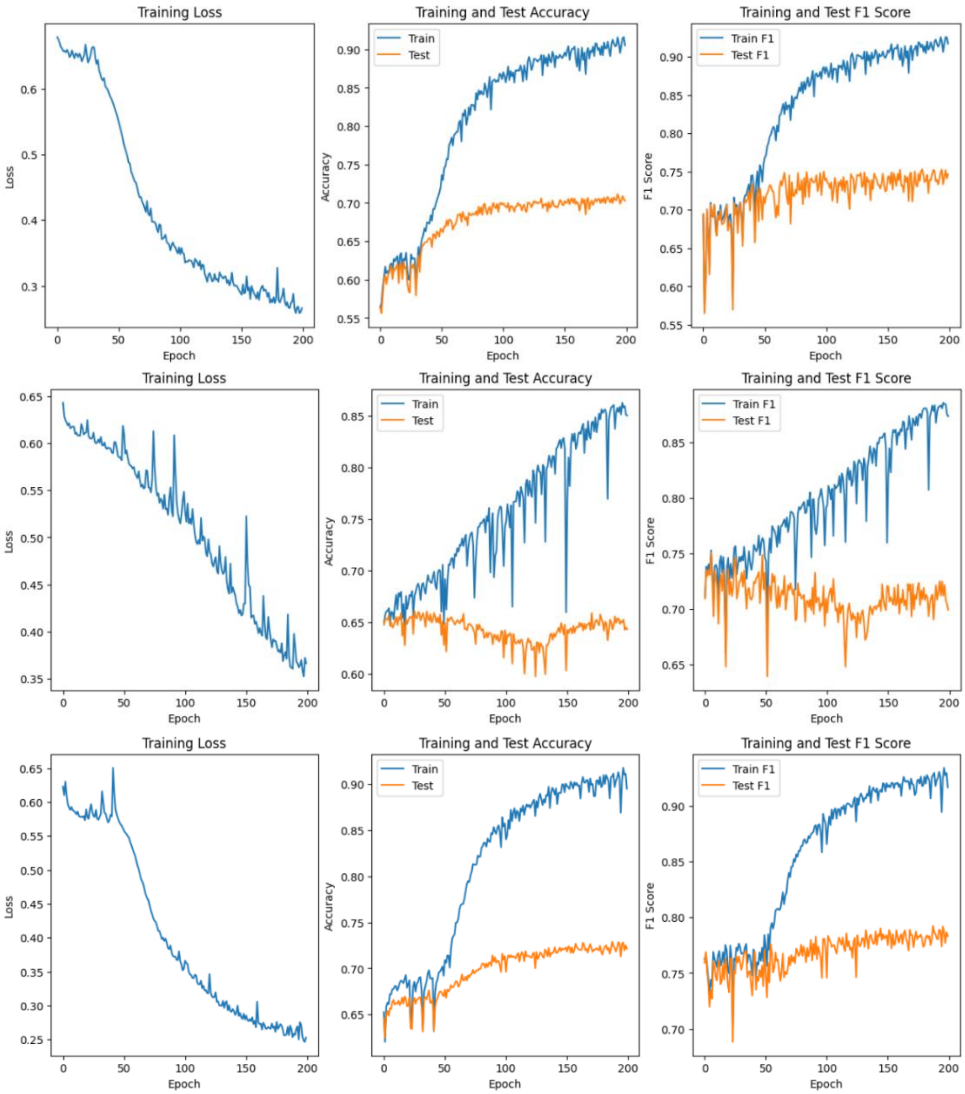
Extended Figure 4. Training loss, training and test accuracy, and training and test F1 score of RNN arousal, valence and dominance classifier on the ICA-preprocessed DEAP dataset.



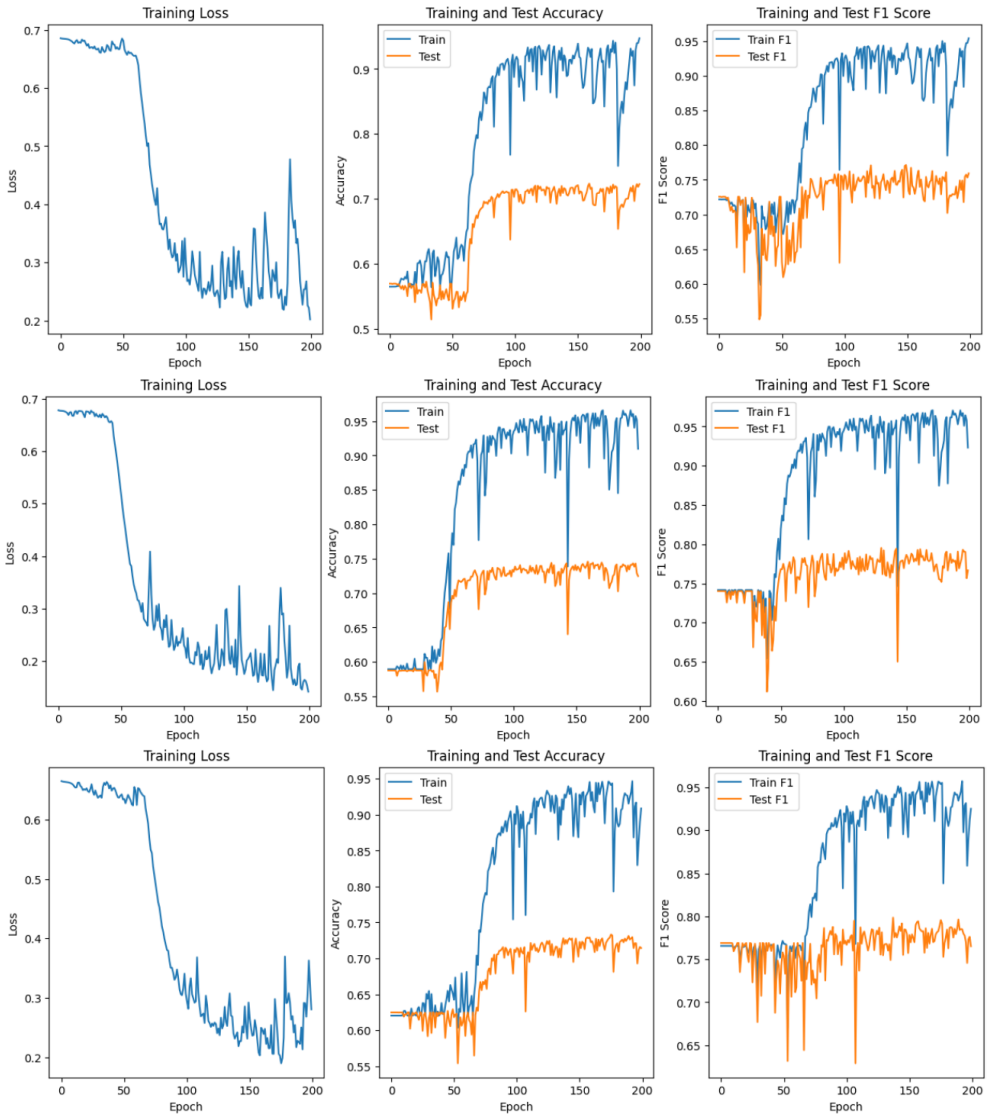
Extended Figure 5. Training loss, training and test accuracy, and training and test F1 score of RNN arousal, valence and dominance classifier on the DREAMER dataset.



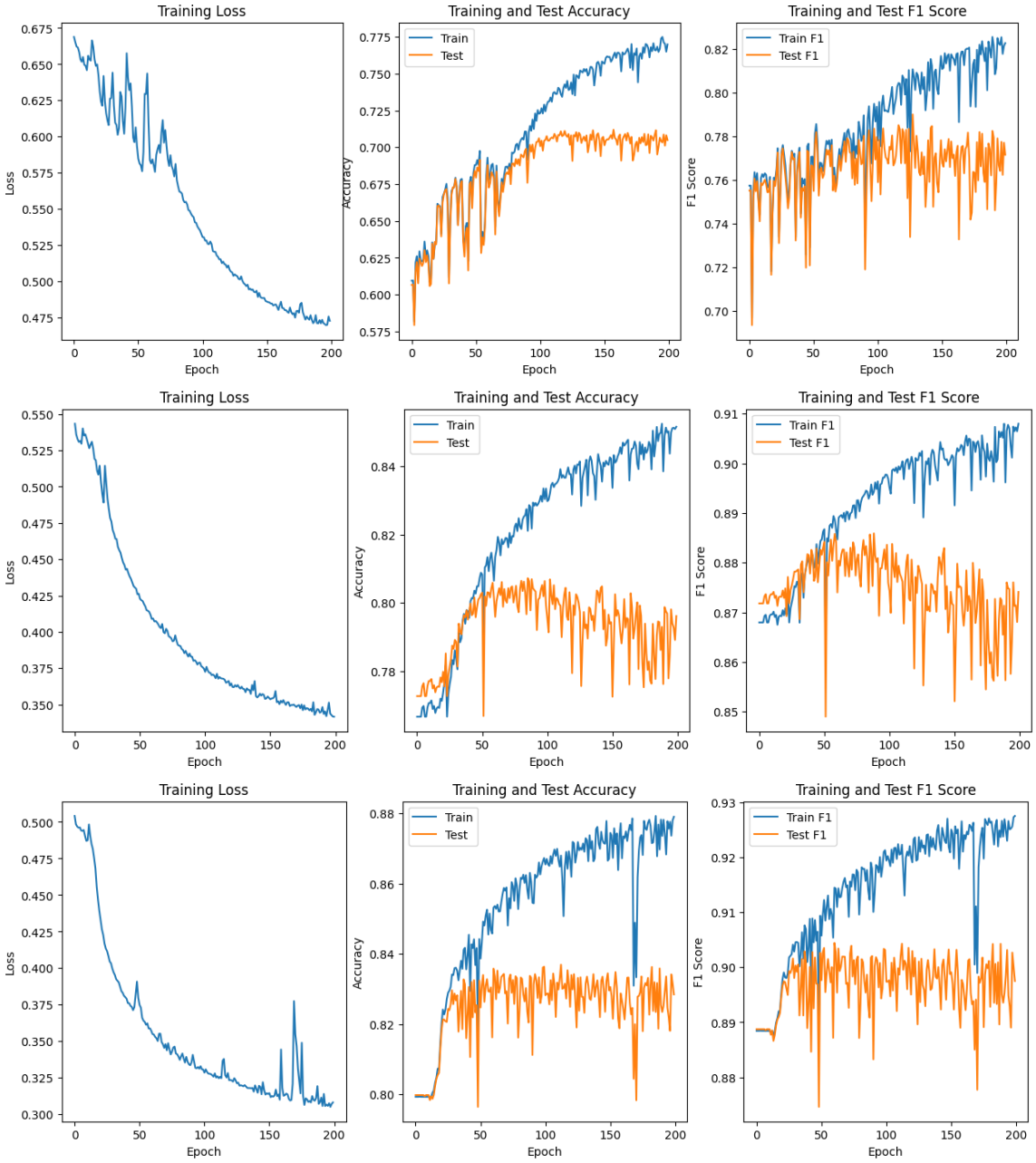
Extended Figure 6. Training loss, training and test accuracy, and training and test F1 score of RNN arousal, valence and dominance classifier on the ICA-preprocessed DREAMER dataset.



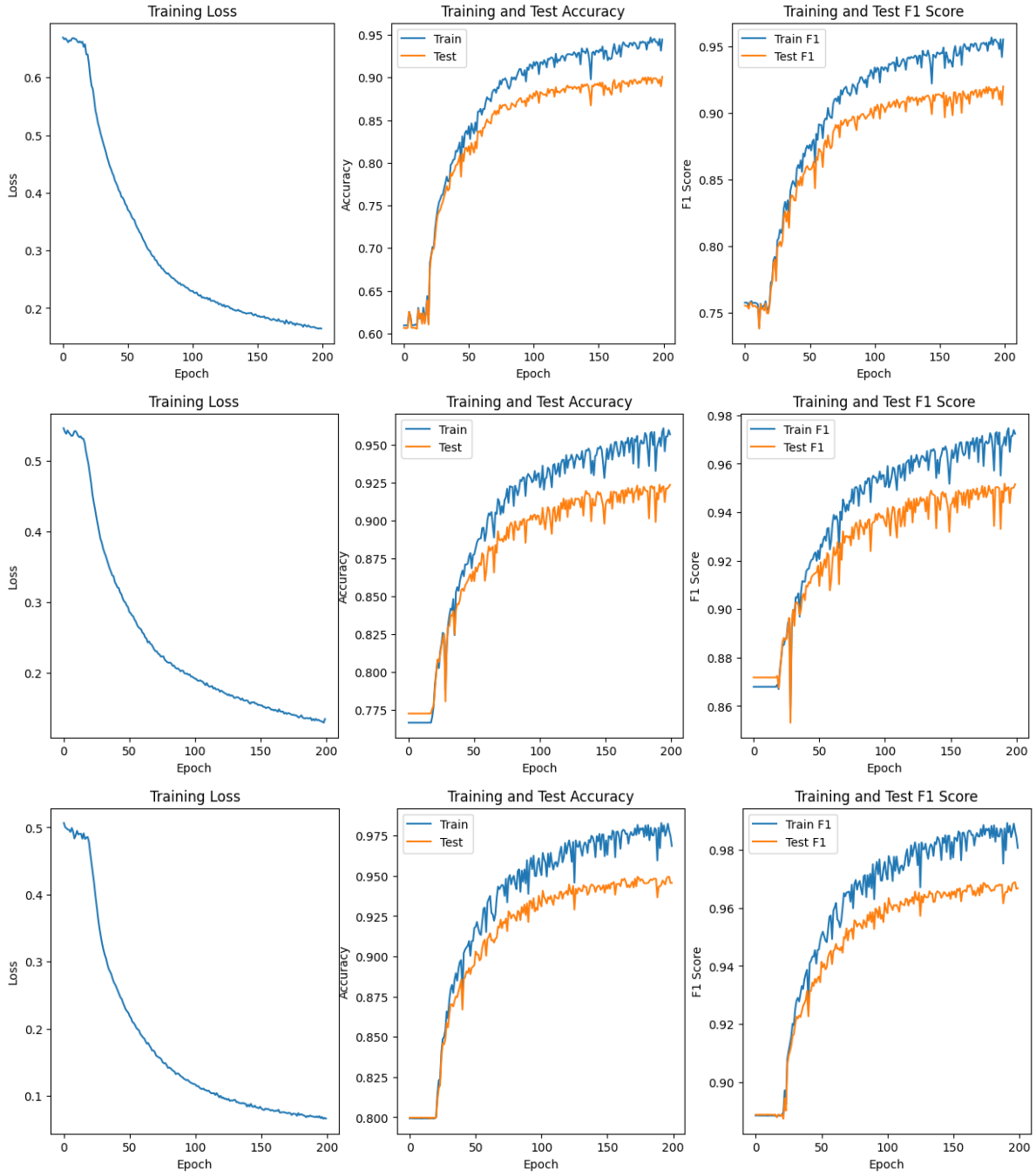
Extended Figure 7. Training loss, training and test accuracy, and training and test F1 score of LSTM arousal, valence and dominance classifier on the DEAP dataset.



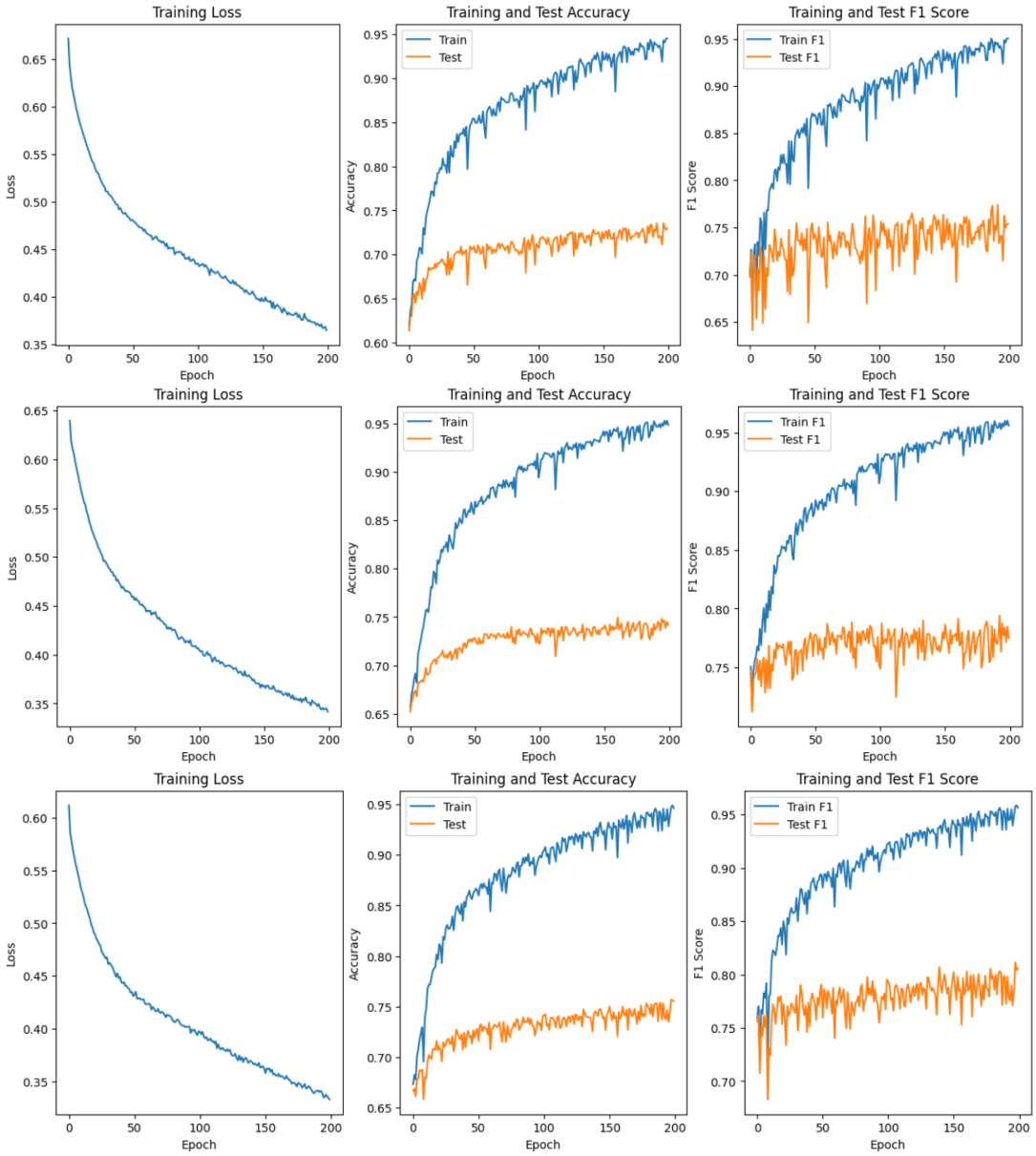
Extended Figure 8. Training loss, training and test accuracy, and training and test F1 score of LSTM arousal, valence and dominance classifier on the ICA-preprocessed DEAP dataset.



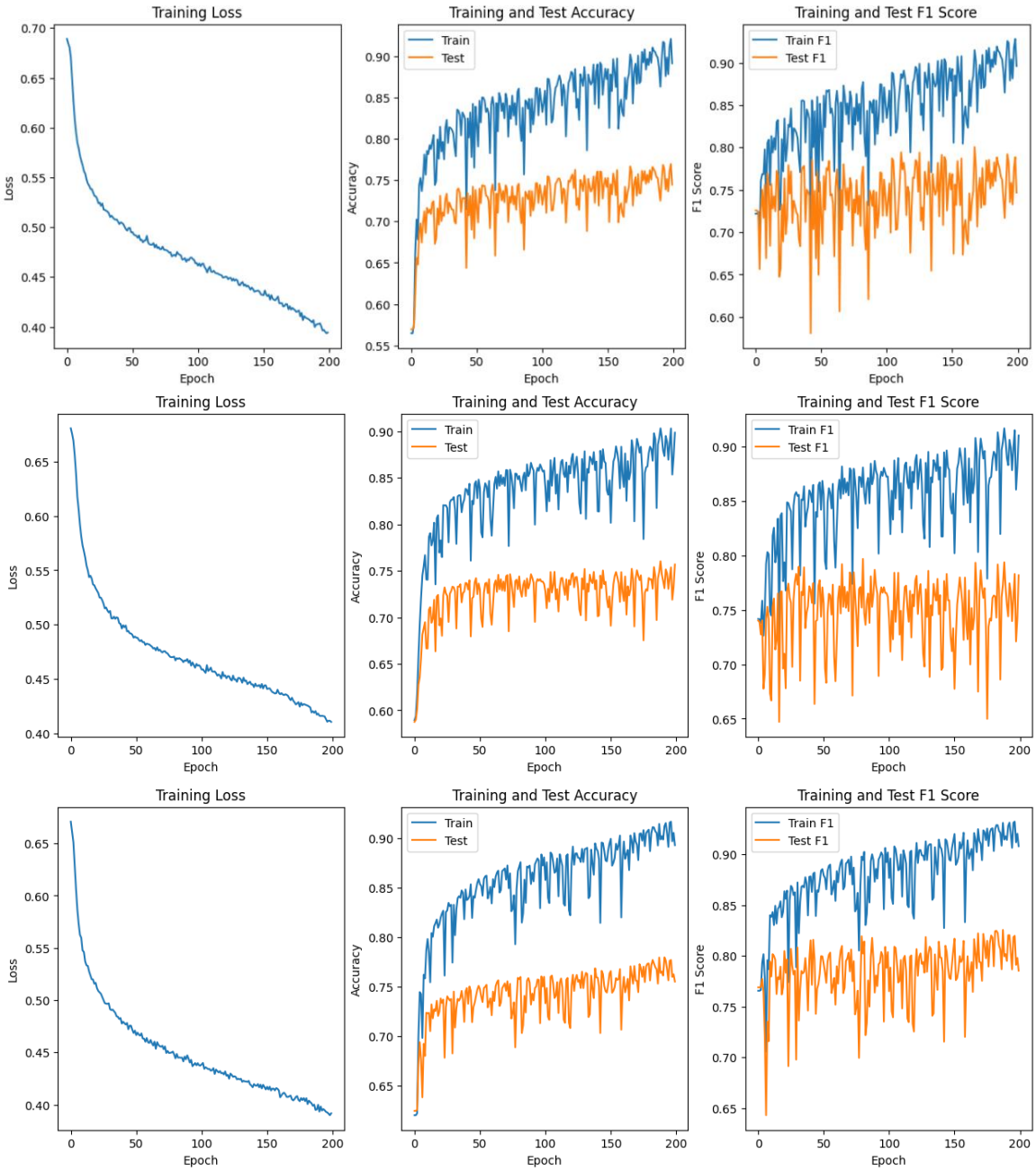
Extended Figure 9. Training loss, training and test accuracy, and training and test F1 score of LSTM arousal, valence and dominance classifier on the DREAMER dataset.



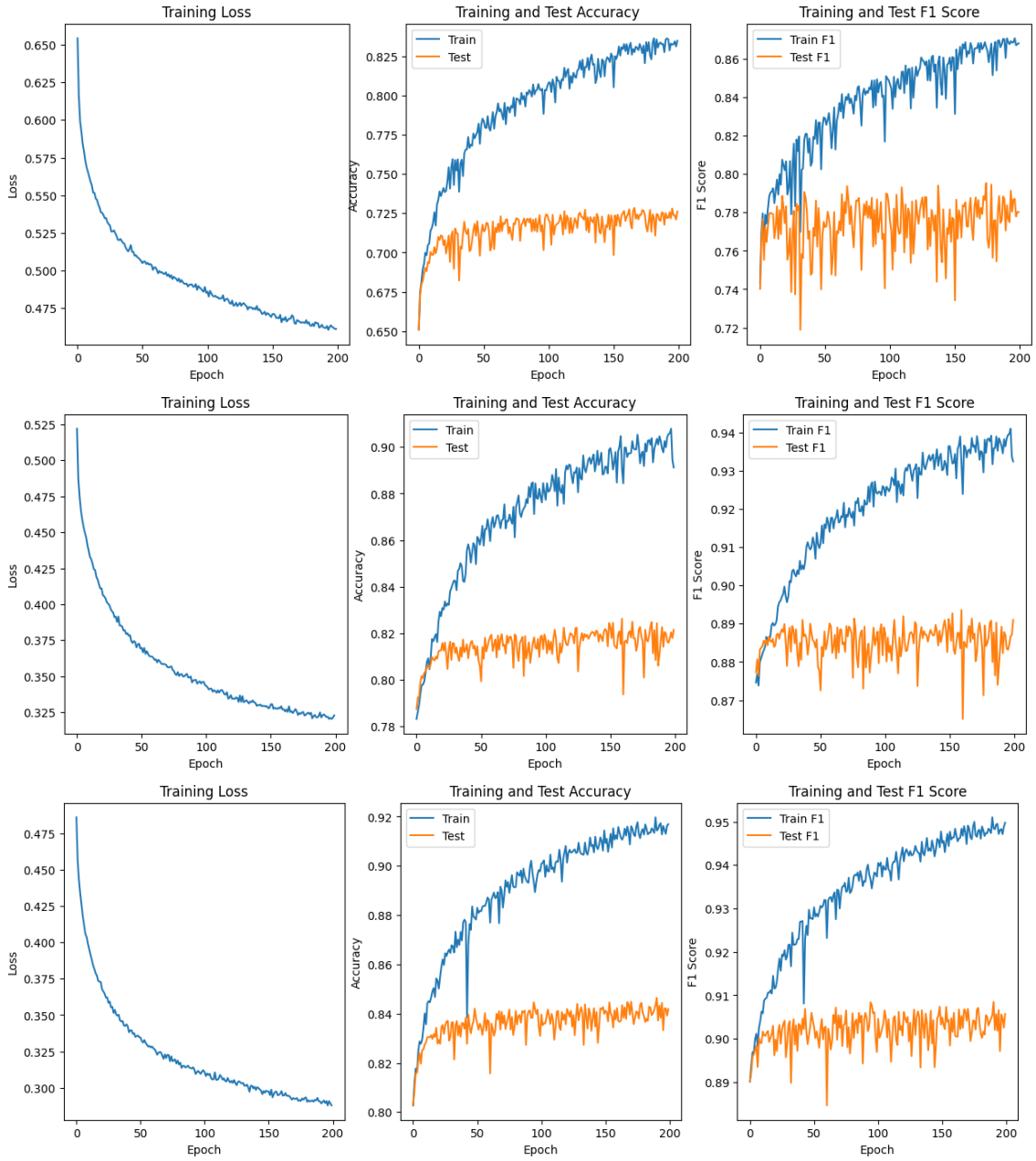
Extended Figure 10. Training loss, training and test accuracy, and training and test F1 score of LSTM arousal, valence and dominance classifier on the ICA-preprocessed DREAMER dataset.



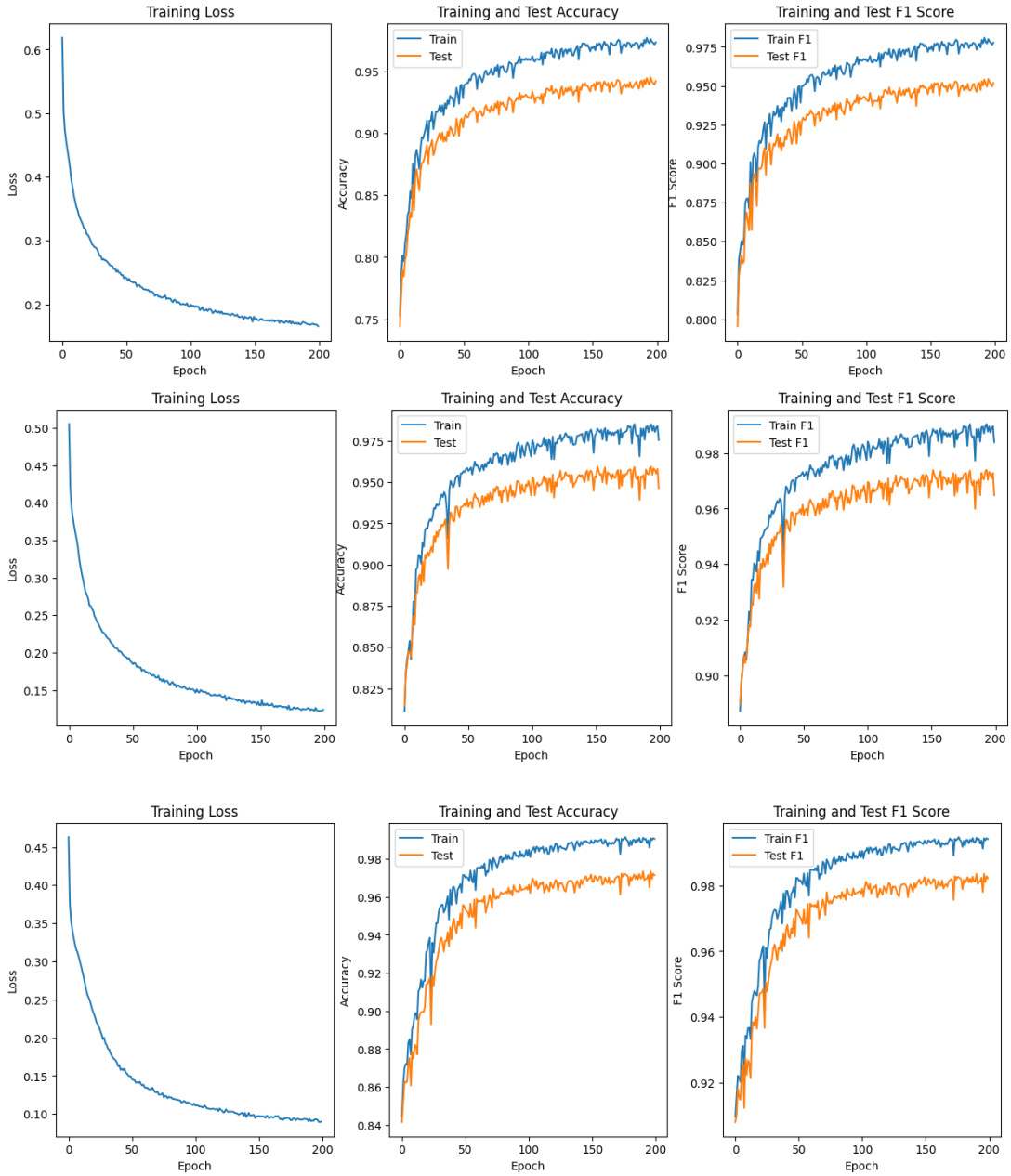
Extended Figure 11. Training loss, training and test accuracy, and training and test F1 score of CNN arousal, valence and dominance classifier on the DEAP dataset.



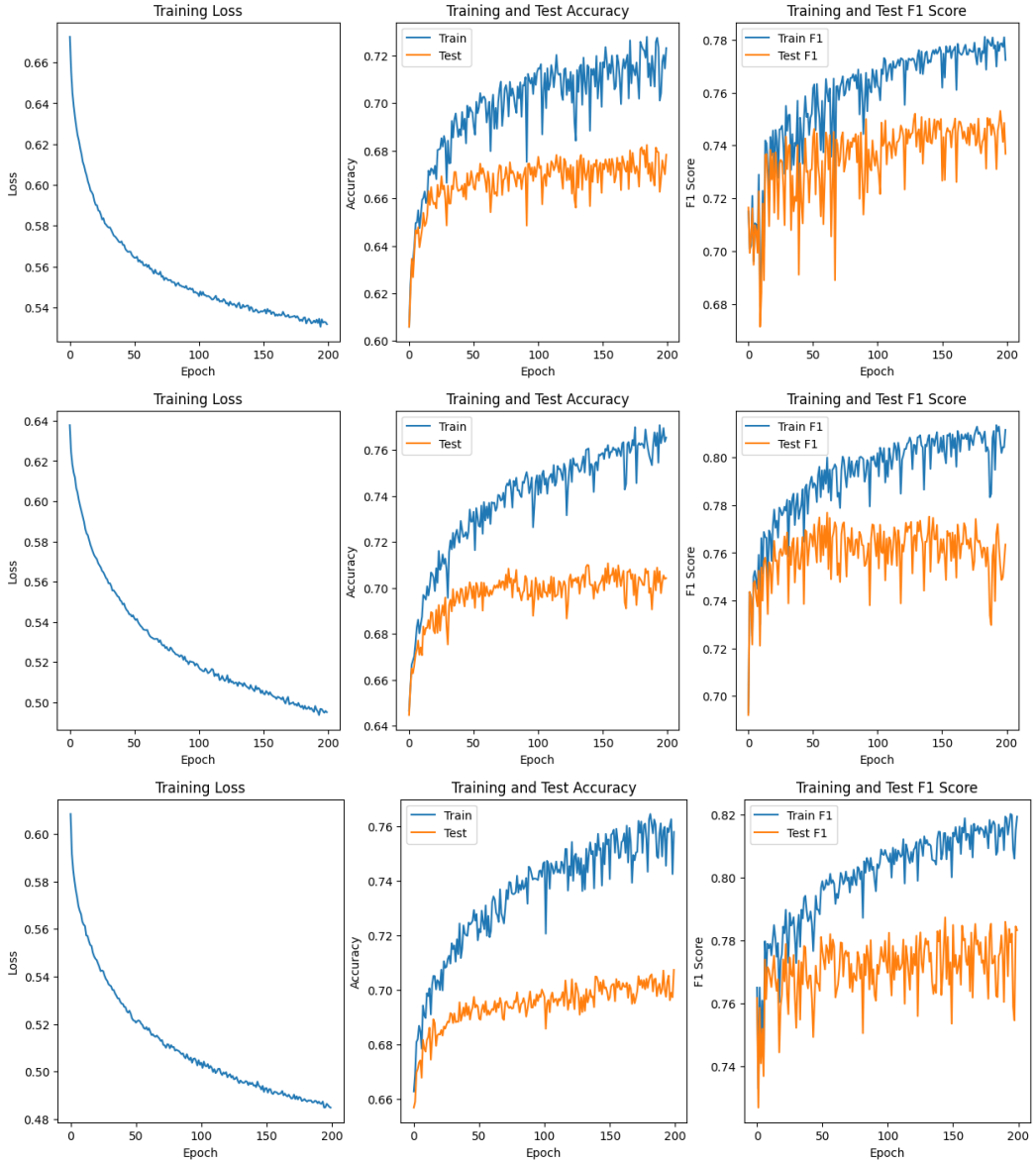
Extended Figure 12. Training loss, training and test accuracy, and training and test F1 score of CNN arousal, valence and dominance classifier on the ICA-preprocessed DEAP dataset.



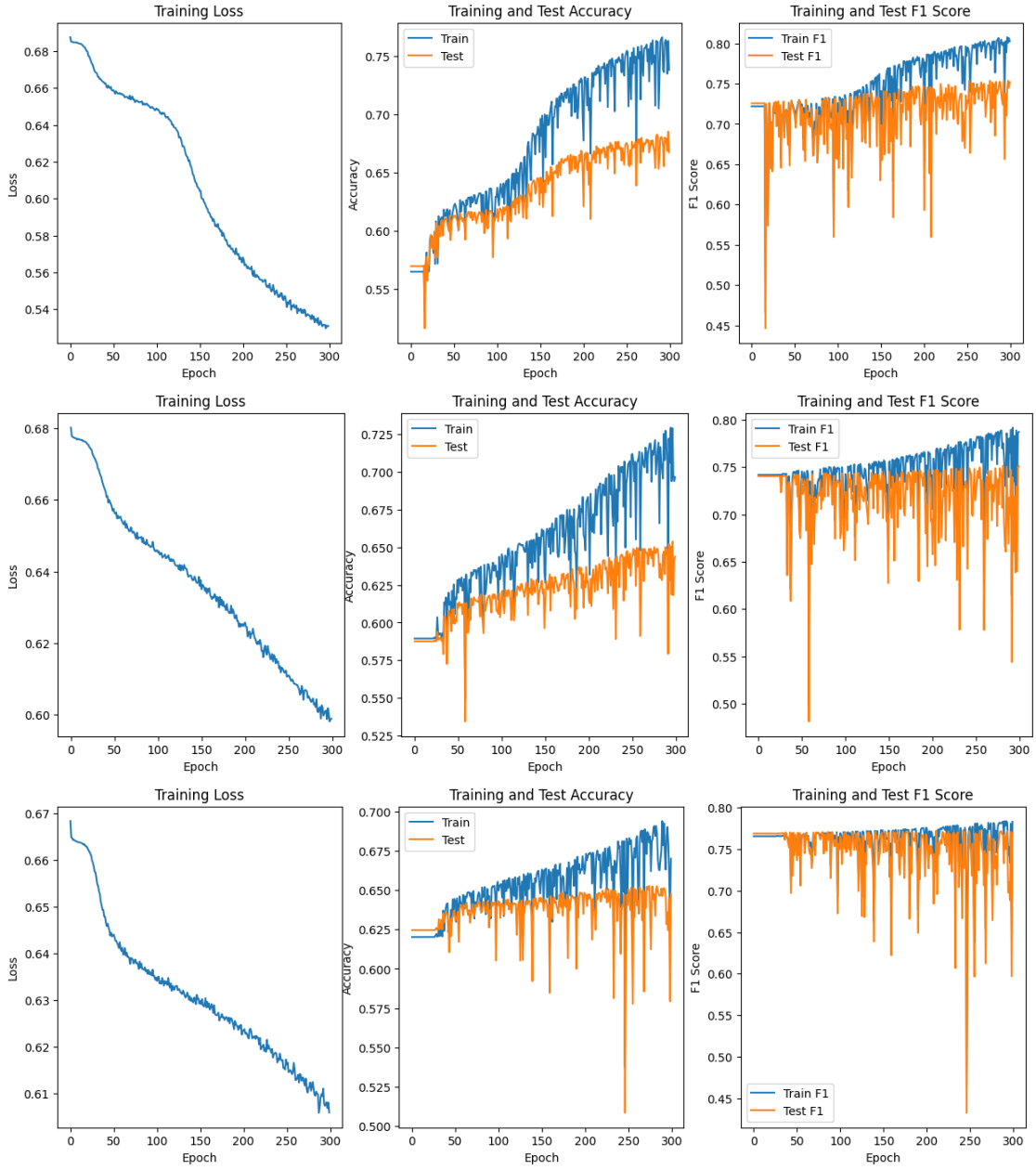
Extended Figure 13. Training loss, training and test accuracy, and training and test F1 score of CNN arousal, valence and dominance classifier on the DREAMER dataset.



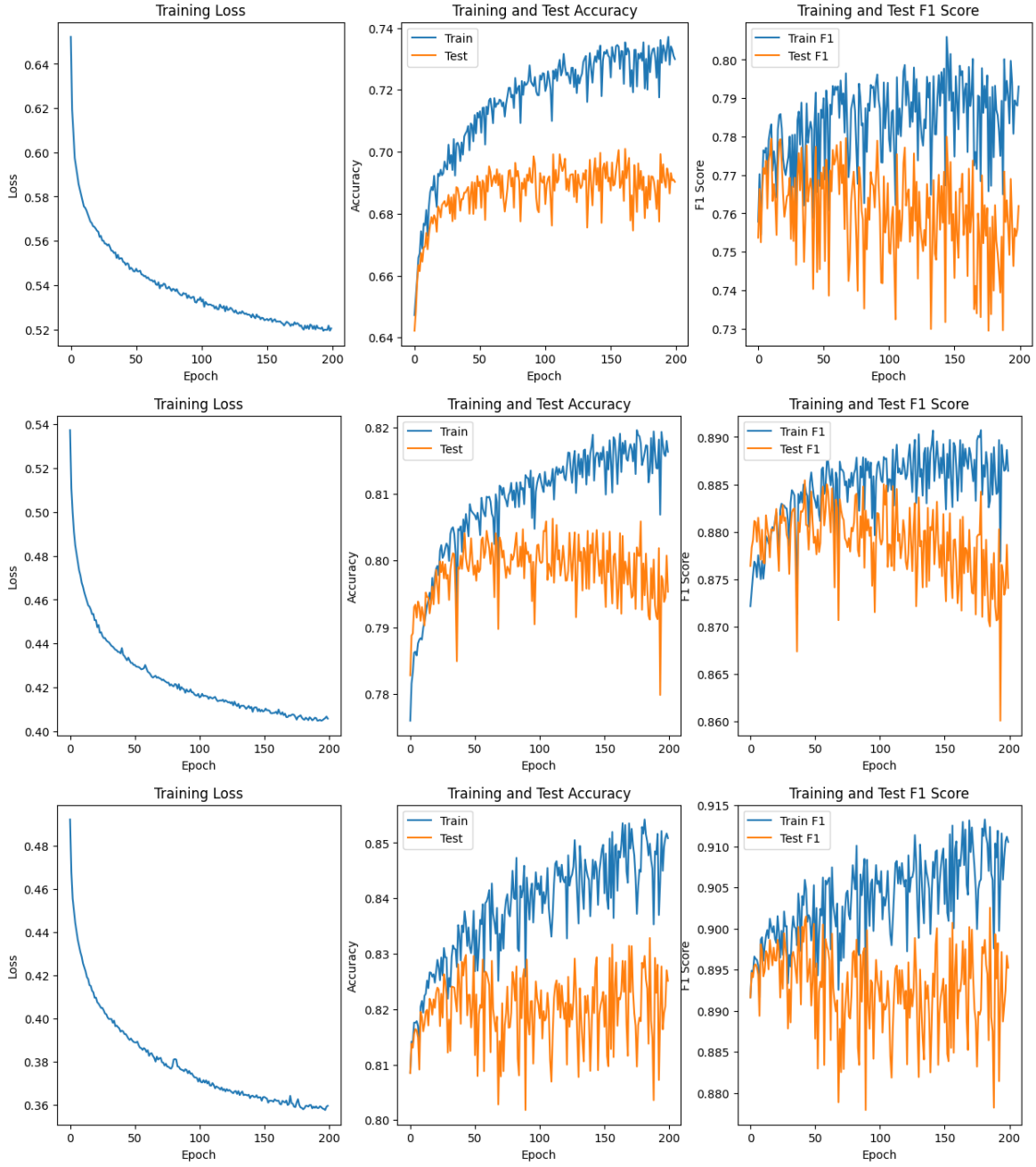
Extended Figure 14. Training loss, training and test accuracy, and training and test F1 score of CNN arousal, valence and dominance classifier on the ICA-preprocessed DREAMER dataset.



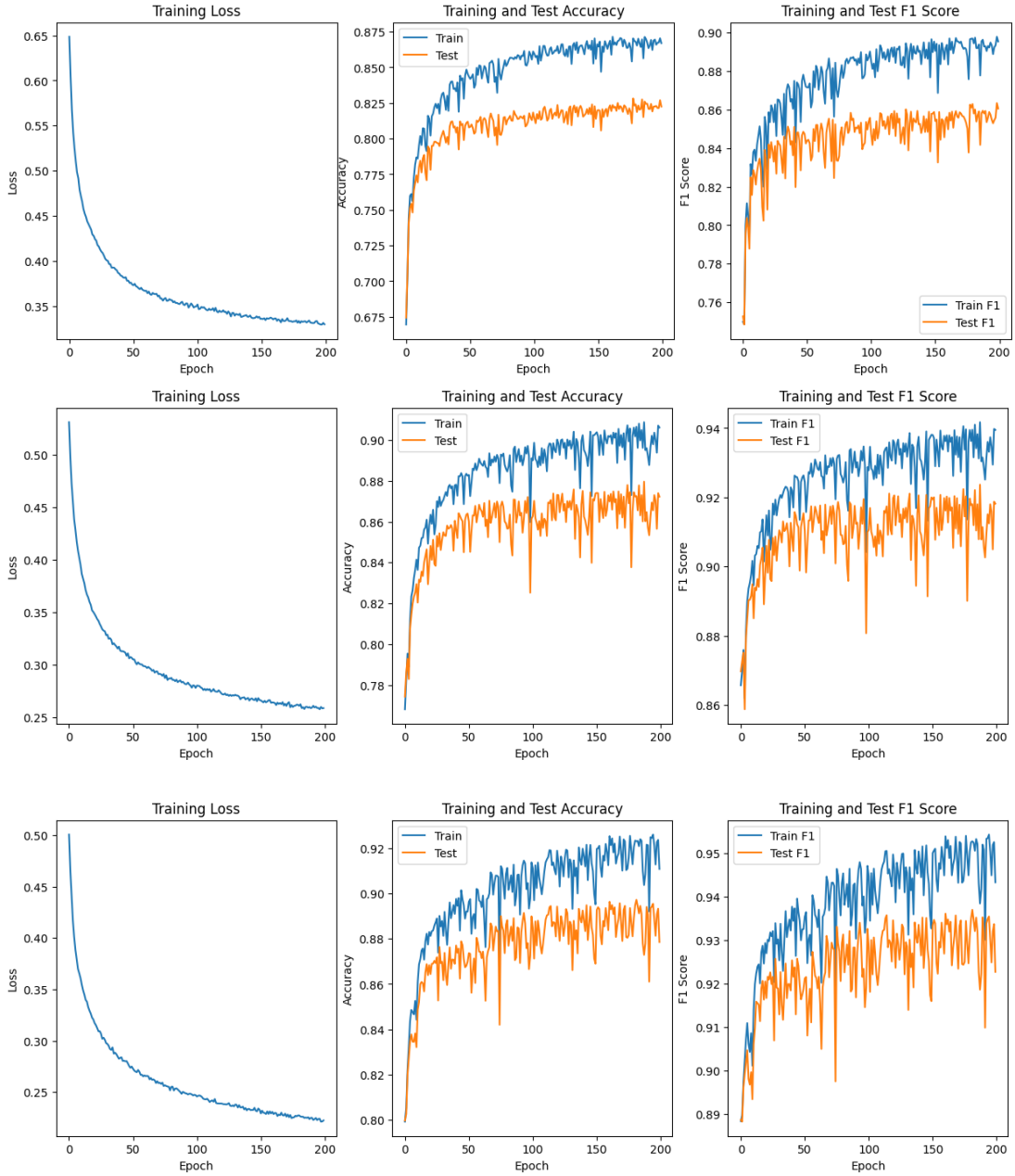
Extended Figure 15. Training loss, training and test accuracy, and training and test F1 score of transformer arousal, valence and dominance classifier on the DEAP dataset.



Extended Figure 16. Training loss, training and test accuracy, and training and test F1 score of transformer arousal, valence and dominance classifier on the ICA-preprocessed DEAP dataset.



Extended Figure 17. Training loss, training and test accuracy, and training and test F1 score of transformer arousal, valence and dominance classifier on the DREAMER dataset.



Extended Figure 18. Training loss, training and test accuracy, and training and test F1 score of transformer arousal, valence and dominance classifier on the ICA-preprocessed DREAMER dataset.

model	data type	emotion	accuracy (%)	f1-score
RNN	raw	Arousal	56.94	0.73
RNN	raw	Valence	63.54	0.71
RNN	raw	Dominance	62.83	0.75
RNN	raw-ICA	Arousal	52.18	0.61
RNN	raw-ICA	Valence	52.93	0.62
RNN	raw-ICA	Dominance	57.62	0.7
LSTM	raw	Arousal	70.29	0.75
LSTM	raw	Valence	63.32	0.7
LSTM	raw	Dominance	72.14	0.78
LSTM	raw-ICA	Arousal	72.27	0.76
LSTM	raw-ICA	Valence	72.47	0.76
LSTM	raw-ICA	Dominance	71.43	0.76
CNN	raw	Arousal	72.95	0.76
CNN	raw	Valence	74.22	0.77
CNN	raw	Dominance	75.56	0.81
CNN	raw-ICA	Arousal	74.46	0.75
CNN	raw-ICA	Valence	75.67	0.78
CNN	raw-ICA	Dominance	75.51	0.79
transformer	raw	Arousal	67.83	0.74
transformer	raw	Valence	70.42	0.76
transformer	raw	Dominance	70.74	0.87
transformer	raw-ICA	Arousal	66.71	0.75
transformer	raw-ICA	Valence	64.38	0.75
transformer	raw-ICA	Dominance	64.77	0.77

Extended Figure 19. Accuracies and F1-scores of the Deep Learning Classification Models on the DEAP dataset.

model	data type	emotion	Accuracy (%)	f1-score
RNN	raw	Arousal	60.66	0.75
RNN	raw	Valence	77.63	0.87
RNN	raw	Dominance	80.09	0.89
RNN	raw-ICA	Arousal	60.66	0.75
RNN	raw-ICA	Valence	77.27	0.87
RNN	raw-ICA	Dominance	79.98	0.89
LSTM	raw	Arousal	70.50	0.77
LSTM	raw	Valence	79.62	0.87
LSTM	raw	Dominance	82.84	0.9
LSTM	raw-ICA	Arousal	90.08	0.92
LSTM	raw-ICA	Valence	92.36	0.95
LSTM	raw-ICA	Dominance	94.57	0.75
CNN	raw	Arousal	72.62	0.78
CNN	raw	Valence	82.15	0.89
CNN	raw	Dominance	84.19	0.9
CNN	raw-ICA	Arousal	94.19	0.95
CNN	raw-ICA	Arousal	94.64	0.96
CNN	raw-ICA	Valence	97.14	0.98
transformer	raw	Arousal	69.03	0.76
transformer	raw	Valence	79.53	0.87
transformer	raw	Dominance	82.51	0.9
transformer	raw-ICA	Arousal	82.25	0.86
transformer	raw-ICA	Arousal	87.23	0.92
transformer	raw-ICA	Valence	87.85	0.92

Extended Figure 20. Accuracies and F1-scores of the Deep Learning Classification Models on the DREAMER dataset.