Journal of Information Systems Engineering and Management
2025, 10 (55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Analysis of Deep-Learning-Based Emotion Predictors with Multi-Channel Multi-Label EEG Signals and SHAP XAI

Rishi Kumar Sharma[1*], Vivek Kumar[2], Rajendra Kumar[3]
[1*,2]Department of Computer Science Engineering, Quantum University, Roorkee, Uttarakhand
[3]Department of Computer Science & Engineering, Sharda School of Engineering & Technology, Sharda University, Greater Noida
[*]**Corresponding author:** Rishi Kumar Sharma
*Email: rishi.k.sharma@gmail.com

ARTICLE INFO

Received: 18 Dec 2024

Revised: 10 Feb 2025

Accepted: 28 Feb 2025

ABSTRACT

Human emotion recognition is a peculiar task. Humans express emotions or provide emotional responses via facial gestures, body temperature, and brain activity. Interestingly, brain activities can be observed via EEG recordings. Analysis of an individual's emotional state to stimulations such as video, music, or activity is vital to their behaviour. Deep learning (DL) models are popular and influential enough to predict emotions from EEG signals. Mapping predictions to different EEG channels or features would be critical to further understanding human behaviour. The study in this paper presents an analysis of various deep-learning model performances in predictions of emotions with multi-channel, multi-label EEG signals. In the context of DL, convolutional and recurrent deep neural network models are utilized for emotion recognition. The synergistic use of CNNs and RNNs to extract temporal, spatial, and spectral features from multi-modal physiological data is especially considered here. The DEAP dataset, being a rich source of multi-modal physiological signal representation data, is therefore used in this study. The DEAP data encapsulates a range of stimulated human emotions and is suitable for this study. Most importantly, emotion predictions from proposed DL models on the DEAP dataset are analyzed within an XAI framework. SHAP XAI framework is used to interpret the predictions from DL models and its mapping onto input different physiological signals within the DEAP dataset. Results from DL models indicate improved emotion recognition permeance and SHAP values from the XAI framework indicate the significance of the DL model architecture and its features in achieving this performance.

Index Terms – Emotion predictive analysis, deep-learning, Explainable AI learning, SHAP values.

1. INTRODUCTION

Emotion is a complex psychological state that arises in response to significant internal or external events, often involving a combination of subjective feelings, physiological changes, and behavioural expressions. It serves as a powerful motivator for action, guiding our decisions, interactions, and sense of well-being. An emotional response is the body's way of reacting to an emotional stimulus, encompassing a range of changes from heart rate and breathing adjustments to facial expressions and shifts in tone of voice [1], [2]. This response is not just an instinctual reaction but can also involve a conscious awareness of what one is feeling [3], [4], [5], providing insight into what we value, fear, or desire. Emotions play an essential role in human communication and connection, allowing us to relate to others and understand ourselves profoundly. 25 Before using inner expression data, text, facial expression, and speech were the most common methods to detect emotions [2], [6]. Using physiological data towards emotion recognition has become an appropriate

Journal of Information Systems Engineering and Management
2025, 10 (55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

alternative to external expression (facial expressions, text, and speech data). External expressions-based emotion detection and classification can be easily manipulated, which is why many recent studies have focused on physiological data [7]. Physiological data models can be utilized in unimodal or multi-modal approaches for emotion detection. However, unimodal and multi-modal emotion detection methods have pros and cons. The multi-modal method for emotion detection utilizes a combination of different physiological signals such as electrocardiograms (ECG), electromyogram (EMG), electroencephalogram (EEG), electrodermal activity(EDA), Photoplethysmogram (PPG), galvanic skin response (GSR), respiratory inductive plethysmograph (RIP), blood volume pressure (BVP) and temperature [8]. The multi-modal emotion detection method commonly gives better accuracy than the unimodal method; however, the multi-modal emotion detection method needs longer processing time and has a more complex data collection procedure than the unimodal method [8]. Deep learning has made significant strides in predicting emotion, particularly through the use of multi-modal data such as facial expressions, voice tone, body language, and even textual data. With advanced architectures like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and transformer models, deep learning systems can now process large and complex datasets to detect subtle cues related to emotion. This feature of the advanced DL models has applications in fields such as sentiment analysis, mental health support, personalized marketing, and human-computer interaction. Also, since DL methods can perform feature selection or extraction intrinsically, exclusive steps are not required to do so (311, 312). The DL architecture is suitable for scaling in terms of increased feature dimensions and increased number of samples. CNNs are often used to analyze facial expressions, where they can learn visual patterns associated with specific emotions (like happiness or anger) by focusing on micro-expressions, eye movement, or even posture. In audio processing, RNNs and transformer-based architectures such as BERT or GPT can be applied to recognize emotional tone and sentiment from speech or text data, identifying nuances in tone and context. More sophisticated models now integrate multi-modal inputs, combining visual, audio, and text data to create a more holistic analysis of emotion. Recently, numerous studies have used the DEAP dataset for DL- based emotional response prediction model development. Authors in [9] developed a CNN with a multi-scale kernels model for emotion recognition with DEAP data. They focussed on correlating different EEG signals with frequency, which helped develop multi-scale kernels for different frequency EEG signals. Their multi-scale kernels attempted to capture both local and global patterns. Their method: The proposed method achieved high average accuracies of 98.27% for Arousal and 98.36% for valence binary classification. The authors in [10] construct a 3-D feature that integrates spatial and spectral information from DEAP multi-channel EEG signals. It involves arranging power values from different frequency bands into a 3-D tensor. They
further proposed a framework, namely dilated bottleneck-based convolutional neural networks (DBCN), to process these 3-D features. This framework acts as a feature fusion module. Their model achieved high classification accuracies for participant- dependent strategies (89.67% for Arousal, 90.93% for Valence) and participant-independent (79.45% for Arousal, 83.98% for Valence). The authors proposed a novel method for emotion recognition with EEG signals from the DEAP dataset. Their method is based on multi-scale sample entropy, i.e. MSE and deep hybrid network that incorporates convolutional neural nets or CNNs and hidden Markov models or HMMs. One of the notable insights from their study is that it provides EEG channels with significant activations during classification. It helped them understand which features are contributing more and which are contributing less. Yu Chen et al. [11]combined CNNs with a Borderline-synthetic minority oversampling data augmentation technique to perform emotion classification with the DEAP dataset. At first, they extracted frequency domain features and then performed data augmentation to achieve balanced training data. This balanced dataset of frequency domain features is used to train a 1D-CNN-based DL model. They reported classification accuracies of 97.47% and 97.76% for Valence and Arousal, respectively. Baltatzis et al. in [12] studied bullying in schools by adopting a convolutional deep neural network model with EEG data. The EEG was recorded while making students watch specific videos. Their study identified whether a student is bullied or not. Tang et al. in proposed a method based on the deep 2D-CNNs to be used with single-trial MI EEG. Their feature extraction and classification method is based on

Journal of Information Systems Engineering and Management
2025, 10 (55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

the Spatiotemporal characteristics of EEG. In particular, the 2D-CNNs can extract features that are spatially correlated. They achieved accuracies of 0.6831 for Arousal and 0.6752 for valence classification. The 3D-CNN-based model by Salama, Elham S., et al. in [13] can capture spatial, spectral (channel) and temporal dependencies in EEG data. Their study was one of the preliminary studies that utilized 3D-CNNs for emotion recognition with multi-channel EEG data. They reported achieving recognition accuracies of 87.44% and 88.49% for valence and arousal classes, respectively. Qiao, Rui, et al in [14] proposed a novel model for multi-subject emotion classification. They used convolution strategies for feature abstraction. Their model is based on the principle that CNNs can, in fact, represent the correlations among information from multiple channels of an EEG. Consequently, this strategy helps construct discriminatory features. Their strategy achieved an accuracy of 87.27%, which was averaged over the 32 participants of the DEAP dataset.

The synergistic use of CNNs and RNNs in this regard is worth mentioning. However, the combined use of CNNs and RNNs for emotion identification has recently gained attention. Li et al. [15] applied wavelet features to train CNN combined with LSTM, and the binary classification accuracy reached 72%. Roy et al. in [16] concentrated on segregating brain activities into natural or abnormal with DL. They analyzed four different DL architectures that were based on convolutional networks and recurrent networks, especially using GRUs. One of their proposed models, namely ChronoNet, claimed to have achieved 90.60% training and 86.57% testing accuracies. Supratak et al. [17] proposed a deep learning model named DeepSleepNet for automatic sleep stage scoring based on raw single-channel EEG. They utilize CNN to extract time-invariant features and Bi-LSTM to learn transition rules among sleep stages from EEG epochs automatically. This approach achieved an accuracy of 90%. Bashivan et al. in [18] proposed a novel DL model that could learn representations from time series signals from a multi-channel EEG dataset. They demonstrated that the model is efficient in mental load classification among participants. They trained a deep recurrent CNN that was inspired by state-of-the-art models in the image and video processing domains. Their model aimed at learning robust representations from the multi-channel time-series EEG signals. Their designed model is aimed to preserve and maybe capture the spatial, spectral, and temporal structure of input EEG that is variation- insensitive distortion-insensitive. They achieved TPRs and TNRs higher than 60% for seven participants among the thirteen participants they considered. J. Chen et al. in [] proposed a convolutional-recurrent layers '-based hybrid neural network model for learning spatiotemporal EEG representations from multi-channel EEG data. They transformed 1D representations into 2D meshes. Then, they segmented these 2D meshes into equal parts. A combination of parallel and cascaded convolutional-recurrent architecture is used to extract features from these segments. They reported classification accuracies of approx. 93% for both Valence and Arousal. This current study finds motivation to explore and utilize CNN-RNN models for emotion identification with multi-channel physiological signals, including EEGs, from the above literature.

Despite progress, challenges remain. Emotions are highly subjective, context-sensitive, and culturally influenced, making generalization difficult. Furthermore, emotions are dynamic and can change rapidly, posing additional challenges for real-time analysis. Researchers are actively working on refining models to better handle these complexities, and with continual improvements in data availability, computational power, and model design, the field is moving closer to highly accurate and context-aware emotion recognition systems. In particular, it is also notable that although studies have shown that EEG signal classification via deep learning models can achieve high prediction accuracy [19] [20]. However, these models are still considered “black boxes,” lacking interpretability and immediate understanding ability for healthcare professionals. In recent years, explainable AI or XAI, has become an increasingly significant tool in the AI world because of its application in understanding critical decisions and the fact that regulators hold businesses responsible for their AI models' judgments. Its rapid growth suggests that in the days to come, real-time AI deployment and perception may change dramatically. An XAI framework's module typically consists of two parts: the interpretability module and the explainability module. [21]. Explaining the black-box model's decision output is the primary goal of the explainability model.

Journal of Information Systems Engineering and Management
2025, 10 (55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

Explainability tries to answer the '*why an algorithm produces a particular response*' question. Therefore, it considers issues like the weighting of each variable inside the model to evaluate the relative value of each variable in answering the question. Although the procedure that takes place within the model may continue to be a mystery, we are aware of the reasons why the response has been delivered. In the context of understanding analytical models and algorithms, interpretability refers to the process of identifying how the model or algorithm arrived at its results. For example, when a model is interpretable, it is easy to comprehend the inputs and processes utilized to arrive at its predictions. Frameworks like GradCAM [22], [23], Local Interpretable Model-Agnostic Explanation (LIME) [24], [25], Shapley Additive explanations (SHAP) [26] [27], Layer-wise Relevance Propagation (LRP) [28], [29], and others fall under explainability models. In order to train an interpretable model that is based on the predictions of black-box models, the well-known Local Interpretable Model-agnostic Explanation (LIME) was developed. Under normal circumstances, the LIME can rapidly produce superior local explanations for any black-box model. Game-theoretic elements were included in the Shapley additive explanation (SHAP), which resulted in an improvement to the LIME model. It attributes characteristic elements of the data to the measurement results that are significant for making predictions. A more comprehensive explanation of learning models is provided by the SHAP, which contributes to an overall improvement in comprehension. Among the many methods that are capable of producing visual explanations for the decisions that CNN-based models make, the Grad-CAM approach is yet another example.

Therefore, this paper presents a study in which novel human emotion recognition models are developed and utilized. These models are based on the concepts of CNNs, RNNs, and XAI. The study proposes convolutional-recurrent architecture-based models for use in human emotion-response prediction and also proposes an XAI framework to interpret the performance of these models. At first, a multi-channel, multi-label signal is fragmented into equal parts across the time dimension. The CNNs act on these time-wise fragments of the entire signal to extract discriminatory features. In contrast, the RNNs act on these temporally distributed discriminatory features obtained as an output from the CNNs. The Shapley Additive Explanations, more commonly termed the SHAP XAI method, is used to explain the significance and contributions of the model, especially discriminatory features from CNNs and temporal-information-based features from RNNs in emotional response prediction applications.

The rest of this paper is divided as follows. Section 1 provides a premise for human emotion recognition using physiological signals, various sensors used so far to record and represent these emotions or plausible indicators, popular datasets in the field, and CNN-RNN-based models and methods popularly reported so far. In section 2, under materials and methods, at first, the DEAP dataset considered for study is discussed. Further, the proposed methodology describes DL model configurations and the SHAP XAI framework. Section 3 discusses DL model training and testing processes, classification performance evaluations, and model performance interpretability within the SHAP XAI framework. Finally, section 4concludes the study.

2. Material and Methods

1. Dataset

The DEAP dataset is a database for emotion analysis using physiological signals [30]. It included EEG signals and other specific physiological signals from 32 participants. These signals are recorded while they watch 1-minute music videos. Overall, each participant watched 40 such videos, and the corresponding physiological responses were stored at a rate of 1280 unique observations. A total of 44 sensors are used to record 48 different physiological responses. The raw recordings are down-sampled to 512 Hz. The sensors consisted of 32 EEG sensors, 12 peripheral sensors, and one status signal channel. The dataset description is briefly summarized in Table 1. For model development in this study, all sensory data except the face videos are considered. Each participant's reported emotion can be majorly decomposed into Arousal, Valence, and Dominance, as shown in Figure 1.

Journal of Information Systems Engineering and Management
2025, 10 (55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

Table 1 Summary of the DEAP dataset containing multi-channel physiological signals.

Dataset	DEAP
Participants	32
Sensory input signals: EEGs, EMGs, EOGs, GSR, RR, Plethy, Temperature	44 32, 4, 4, 1, 1, 1, 1
Stimulation	Music video clips
Stimulation duration	1-minute
Number of stimulations per participant	40
Emotions	40 emotions based on the *Arousal-Valence* map
Supplementary data	Face videos

155

2. Methodology

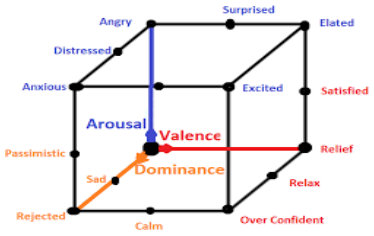


Figure 1 Emotional response decomposition 3D map.

The overall methodology for emotion recognition in this study is as follows.

1. The multi-modal and multi-label physiological time-series signal dataset, i.e. DEAP dataset, is preprocessed first.
2. Various models based on CNN, RNN, and C-RNN architectures are proposed and developed for emotion recognition.
3. An XAI framework, i.e. SHAP, is utilized to analyze how well and why the DL models perform on the emotion recognition problem.

Each step is discussed separately here under the following sections and sub-sections. A flowchart to represent this overall methodology is shown in Figure 2. The overall idea of this methodology is to understand the functioning of DL models in emotion recognition tasks given a multi-source, multi-modal, and multi-label physiological signal.

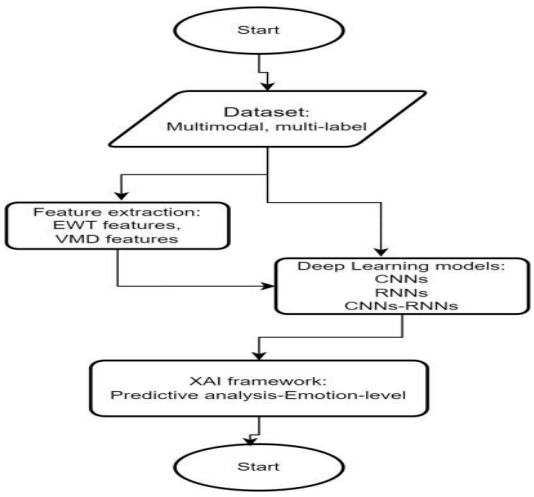


Figure 2 Flowchart summarizing proposed methodology for the study.

Journal of Information Systems Engineering and Management
2025, 10 (55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

CNN Model: Configuration

At first, the input signal is transformed via a 2-layer series of convolutions and non-linear activations functions. The first layer has 32 one-dimensional convolutional filters, each of which exploits the information encapsulated in a 1000-step long multi-channel signal segment. This transformed signal segment is then down-sampled via the *max-pooling* strategy. The down-sampled signal segment is further transformed via another set of 64 convolutional filters and further down-sampled again via the same pooling strategy as earlier. The output from this layer is vectorized and fed into a series of fully connected layers to accommodate the scaling of the signal. The non-linear action functions keep the non-linear relationship mapped through these transformations. Finally, a classification layer with a S*oftMax* activation is employed to achieve a binary classification output. The overall model is termed the *1D-CNN-EEG* model, and its architecture description is listed in Table 2. 180

RNN Model: Configuration

The development of an RNN-based model is a bit tricky. This is because the signal length in time is 7860, a massive amount for an RNN to process and memorize. Therefore, the fragmentation of the signals into smaller temporal-length signals is critical to enable the model to learn faster. In this model, at first, the input signal is fragmented into 16 500-step signal segments. These 16 segments are then fed into 16 separate RNNs simultaneously, with four filters in each RNN. These 16 RNNs provide 16 outputs, which are then fed into another RNN, which takes a 16-step long signal.

The outputs from the previous RNNs are arranged chronologically. This means that the out from the RNN, which takes the first 500-step signal
segment, is the first element in the 16-step long signal input to the second stage RNN. This keeps the temporal relationship intact up to an extent. The output from this second stage RNN is then fed into a fully connected layer. The non-linear action functions keep the non-linear relationship mapped through these transformations. Finally, a classification layer with a S*oftMax* activation is employed to achieve a binary classification output. The overall model is termed the *RNN-EEG* model, and its architecture description is listed in Table 3.

CNN-RNN Model: Configuration

In the RNN-EEG model development, it was clear that significantly large-length signals are challenging to process directly with RNNs. The fragmentation and temporally-localized processing of the signal help, but the 500-step signal segment is still significant for RNNs to process. A similar transformation of these temporally localised signal segments could retain their temporal dependency with respect to each other. These transformations can be achieved with CNNs. Therefore, the input signal is initially fragmented into 16 500-step signal segments in this model. These signal segments are then transformed via a 2-layer series of convolutions and non-linear activations functions. The first layer has 32 one-dimensional convolutional filters, each exploiting the information encapsulated in a 500-step-long multi-channel signal segment. This transformed signal segment is then down-sampled via the *max-pooling* strategy. The down-sampled signal segment is further transformed via another set of 64 convolutional filters and further down-sampled again via the same pooling strategy as earlier. Sixteen of these 2-layer transformations are simultaneously applied to the 16 signal fragments, proving that 16 transformed signal elements are temporally correlated. This 16-step long signal is then fed into an RNN layer. The output from this RNN is then fed into a fully connected layer. The non-linear action functions keep the non-linear relationship mapped through these transformations. Finally, a classification layer with a S*oftMax* activation is employed to achieve a binary classification output. The overall model is termed the C-*RNN-EEG* model, and its architecture description is listed in Table 4.
Variants of these CNN, RNN, and C-RNN models are developed, trained and tested for emotion detection and recognition.

Journal of Information Systems Engineering and Management
2025, 10 (55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 2 CNN model configuration.

Model	Layer				
	Input	Convolution		Fully connected (FC_{first}, FC_{seond})	Classification
		First	Csecond		
CNN-EEG	Data dimensions = 1□□□□□□□□□□ □□ (samples, time, channels) = 32 participants, 40 videos, 60 seconds, 40 channels	Convolution type = one-dimensional Filter count = 32 Kernel size =1000 Pooling = □ Activation =Tanh Dropout fraction = 0.15	Convolution type = one-dimensional Filter count = 64 Kernel size =2000 Activation =Tanh Pooling = □ Dropout fraction = 0.15	Node count = 64, 32 Activation =Tanh, Tanh Dropout fraction = 0.15, 0.15	Label count = 2 Activation = SoftMax

Table 3 RNN model configuration

		Layer				
Model	Input	Fragmentation	Recurrence		Fully connected (FC_{first},	Classification
			Stage 1	Stage 2		
RNN-EEG	Data dimensions = 1□□□□ □□□□□ □□ (samples, time, channels) = 32 participants, 40 videos, 60 seconds, 40 channels	Fragment the input signal into 16 small segments, with each segment being 500 steps long. Number of fragments = 16 (with padded input signal) Fragmented signal length = 500	16 RNNs, one for each signal segment Recurrence cell type = LSTM Cell length = 500 Number of filters = 4 Recurrence activation = Sigmoid Output activation = Tanh Signal flow: Bidirectional. Architecture strategy = Many to one	1 RNN to combine the previous stage 16 outputs from individual RNNs Recurrence cell type = LSTM Cell length = 16 Number of filters = 10 Recurrence activation = Sigmoid Output activation = Tanh Signal flow: Bidirectional. Architecture strategy = Many to one	Node count = 32 Activation =Tanh Dropout fraction = 0.15	Number of labels: 2 Activation: Softmax

Table 4 CNN-RNN model configuration

Model	Layer					
	Input	Convolution stage 1	Convolution stage 2	Concatenation stage	Recurrence stage:	Classification stage
C-RNN-EEG	Data dimensions = 1□□□□ □□□□□ □□ (samples, time, channels) = 32 participants, 40 videos, 60 seconds, 40 channels 16 fragments are created, each with a length of 500, i.e One fragment = 1□□□ □□□□□ □□	1D convolution (separately for each of the 16 fragments) Filter shape: 199 Number of filters: 32 Pooling size: 2 Activation =ReLU Dropout fraction = 0.15	1D convolution (separately for each of the ten fragments) Filter shape: 151 Number of filters: 64 Pooling size: 2 Activation =ReLU Dropout fraction = 0.15	Concatenate all 16 fragments in chronological order Shape: 16 '64 (16 represent time-stamps and 64 represent features)	Recurrence cell type = LSTM Cell length = 10 Number of filters = 12 Recurrence activation = Sigmoid Output activation = Tanh Signal flow: Bidirectional. Architecture strategy = Many to one	Number of labels: 2 Activation: Softmax

Journal of Information Systems Engineering and Management
2025, 10 (55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

Table 5 CNN-EEG, RNN-EEG, and C-RNN-EEG model training parameter settings

Hyperparameter settings	L o s s

SHAP Explainable AI Framework

The SHAP (Shapley Additive exPlanations) framework, a part of the Shapley explainable AI framework, is based on cooperative game theory and aims to provide insights into how each feature contributes to a model's predictions. SHAP leverages Shapley values, which are a method from game theory that can fairly allocate the "payout" (here, the model output or prediction) among the "players" (features), allowing us to quantify and interpret each feature's contribution to an individual prediction. Here is a detailed breakdown of the foundations of SHAP, including its mathematical basis. Lloyd Shapley developed Shapley values in the context of cooperative game theory. The idea is to determine each player's contribution in a game where players work together to achieve an expected outcome.

Given, A set of players $N = \{1,2,\ldots\ldots\ldots\ldots N\}$. A function $v : 2^N \rightarrow R$ that assigns a "payout" to each subset of players, representing the total value (or contribution) that any subset can achieve together. For any player iii in the game, the Shapley value $\phi_i(v)$ is given by:

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S)) \quad (1)$$

Where, S is any subset of players that does not include player i, |S| is the size of subset S, and $v(S \cup \{i\}) - v(S)$is the marginal contribution of player i to the subset S. The Shapley value thus represents an average of the marginal contributions of player iii across all possible subsets SSS of other players.

Application of Shapley Values to Model Interpretability: In the SHAP framework, each feature in the model corresponds to a **player** in the cooperative game, and the model output is considered the total "payout" that we want to distribute amongthe features.

Let

- $f(x)$ represents the model's prediction for a specific input x.
- $f(x')$ denotes the prediction if features are missing or replaced with their baseline values,
- $f(x) - E[f(x)]$ is the "payout" we want to distribute, where $E[f(x)]$ is the expected model output over all inputs.

Then, for a given prediction, the SHAP value $\square_i$ for the feature i can be viewed as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} (f(S \cup \{i\}) - f(S)) \quad (2)$$

Where,

- F is the set of all features.
- $S \subseteq F\setminus\{i\}$ represents a subset of features excluding i,
- $f(S \cup \{i\}) - f(S)$ represents the marginal contribution of features i to the subset S.

This approach distributes the prediction difference $f(x) - E[f(x)]$ across all features, creating an additive feature attribution as:

Journal of Information Systems Engineering and Management
2025, 10 (55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

$$f(x) = E[f(x)] + \sum_{i=1}^{n} \phi_i \quad (3)$$

Where, each ϕ_i is the SHAP value for feature i and represents its contribution to the difference between $f(x)$ and $E[f(x)]$.

Gradient-Based Attribution

At the core, the Gradient explainer relies on gradients to approximate the contribution of each input feature to the model's This gradient tells us the sensitivity of $f(x)$ to small changes in x_i, and can give a measure of $x^i s$ local influence on the model$i$ output. *Integrated Gradients*: This method addresses the problem of using only the local gradient to approximate feature importance. Integrated gradient calculates the cumulative effect of each feature along a straight-line path from a baseline input (often zero or a neutral state) to the actual input. For a single feature x_i in the input x, the integrated gradient is calculated as: output. In general, if $f(x)$ is the model's output given input x, the gradient of f with respect to each input feature x is $\frac{\partial f}{}$.

$$IG_i(x) = (x_i - x_i^{\prime}) \int_{\alpha=0}^{1} \frac{\partial f(x^{\prime} + \alpha(x_i - x^{\prime}))}{\partial x_i} d\alpha \quad (4)$$

Where x is the actual input, $x^{\prime}$ is a baseline input (typically zero or some other reference point) and α varies from 0 to 1, moving along the path from $x^{\prime}$ to x.

Approximating Shapley Values with Integrated Gradients

The Shapley value from the SHAP XAI module (ref) for a feature in a model represents the average ontribution of that feature across all possible feature combinations (or "coalitions"). Calculating the exact Shapley value is computationally expensive, especially for deep networks with many features. The SHAP XAI gradient explainer method approximates the Shapley value using expected integrated gradients over different baseline samples. It averages integrated gradients across multiple baselines rather than calculating the contributions from every possible coalition. For an input feature x_i, the Shapley approximation in Gradient explainer is:

$$\phi_i(x) \approx E_{x^{\prime} \sim baseline} \left[(x_i - x_i^{\prime}) \int_{\alpha=0}^{1} \frac{\partial f(x^{\prime} + \alpha(x_i - x^{\prime}))}{\partial x_i} d\alpha \right] \quad (5)$$

Where $E_{x^{\prime} \sim baseline}$ represents an expectation over a set of baseline values $x^{\prime}$.

1. Results and Discussion

This section covers a detailed discussion of results obtained with the variants of the proposed DNN models. At first, a generic tabulation of model performance via overall accuracy as an indicator is presented. Different versions of the proposed CNN- EEG, RNN-EEG, and C-RNN-EEG are developed for different emotions. Next, a SHAP XAI framework-based interpretation of the C-RNN-EEG models is made, and corresponding interpretations are discussed.

1. Model Training and Hyperparameter Settings

A total of 12 models, 3 for Arousal, 3 for Valence, and 3 for Dominance, have been developed and trained. The three models for each emotion, say Arousal, are based on the CNN-EEG, RNN-EEG, and C-RNN-EEG

Journal of Information Systems Engineering and Management
2025, 10 (55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

models discussed in Section 2. Each model is trained and tested on 1280 samples. An 87/6.5/6.5 (%) ratio is opted for training, validation, and testing respectively. Table 6 tabulates the distribution of samples for each model. Each model attempts to classify a baseline emotion, say Arousal, into three categories: *Low*, *Medium*, and *High*. *Categorical Cross Entropy* is chosen as the loss function, and the *Adam* method is considered for optimization. A batch size of 6 is set as the total sample size is low. The model is trained for over 200 epochs, and performance saturation is achieved around the 100th epoch, as shown in Figure 3. The model keeps these weights as the difference between training and validation accuracy and loss is minimal at this point. 293

Table 6 DEAP dataset sample distribution for training and testing.

Model	Emotion	Classes	Samples (overall: 1280)	
			Training and validation	testing
CNN-EEG-V	Valence	Low, Medium, High	1200	80
CNN-EEG-A	Arousal	Low, Medium, High	1200	80
CNN-EEG-D	Dominance	Low, Medium, High	1200	80
RNN-EEG-V	Valence	Low, Medium, High	1200	80
RNN-EEG-A	Arousal	Low, Medium, High	1200	80
RNN-EEG-D	Dominance	Low, Medium, High	1200	80
C-RNN-EEG-V	Valence	Low, Medium, High	1200	80
C-RNN-EEG-A	Arousal	Low, Medium, High	1200	80
C-RNN-EEG-D	Dominance	Low, Medium, High	1200	80

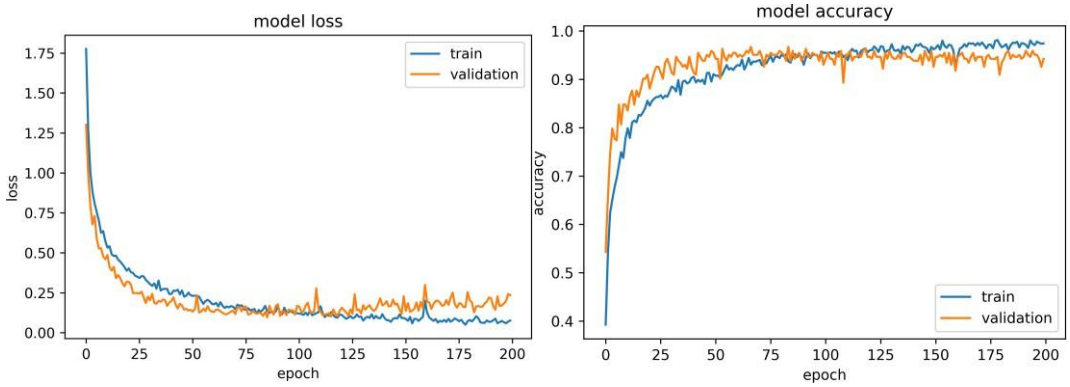


Figure 3 Training progress plots for C-RNN-EEG model; (a) Categorical cross-entropy loss, and (b) Overall accuracy.

2. Model performance: Quantitative analysis

Table 7 reports the classification performance of the different DNN models on emotion recognition with the DEAP dataset. Since DNN models have matured to be efficient, robust, and reliable classification models with physiological signals, it is unnecessary to include results from other classical machine learning methods here. It is clear from Table 7 that the proposed C-RNN-EEG model variants, i.e., C-RNN-EEG-A, C-RNN-EEG-V, and C-RNN-EEG-D, have performed better than their CNN-only and RNN-only counterparts. The C-RNN-EEG-D performs the best, with 76.67% overall accuracy.

To further understand the predictions up to interpretable emotional responses (see Figure 1), Table 8 and Table 9 present the emotional response predicted by the C-RNN-EEG variants when prompted with different testing samples. For example, the C- RNN-EEG-A responded with *High-Arousal* for samples s02-4, 5, and 11. Here, s02 corresponds to the participant ID, and 4, 5, and 11 correspond to the music video ID shown to the participant. In contrast, the C-RNN-EEG-V predicted *Low* Valence for samples s02-11 and *High* Valence for s02- 4 and 5. Whereas the C-RNN-EEG-D predicted *Low* Dominance for samples s02-11 and *Medium*

Journal of Information Systems Engineering and Management
2025, 10 (55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

Dominance for s02- 4 and 5. Mapping the *High-Arousal*, *Low-Valence*, and *Low-Dominance* response from participant s02 to the emotion 3D-map shown in Figure 1 reveals *Angry* emotion.

Table 7 Classification accuracies of popular algorithms in human emotion classification with DEAP dataset [31].

Emotion	Class	Model	Overall accuracy (%)
Arousal	Low	CNN-EEG-A	69.10
	Medium	RNN-EEG-A	68.32
	High	C-RNN-EEG-A	73.12
Valence	Low	CNN-EEG-V	68.20
	Medium	RNN-EEG-V	67.11
	High	C-RNN-EEG-V	76.25
Dominance	Low	CNN-EEG-D	69.5
	Medium	RNN-EEG-D	68.5
	High	C-RNN-EEG-D	76.67

313

Table 8 Participant (id s02) emotional response to sample music videos.

Emotion	Class	Patient ID; music video ID
Valence	Low	S02; 1, 7, 11
	Medium	S02; 3, 8, 9
	High	S02; 4, 5, 6
Arousal	Low	S02; 6, 8, 9
	Medium	S02; 1, 3, 7
	High	S02; 4, 5, 11
Dominance	Low	S02; 6, 7, 11
	Medium	S02; 4, 5, 9
	High	S02; 1, 3, 8

Table 9 Consolidated tabulation of predicted motions against sample music videos for the participant (id s02).

Participant ID, Music video ID	Emotional response		
	Valence	Arousal	Dominance
S02, 1	Low	Medium	High
S02, 3	Medium	Medium	High
S02, 4	High	High	Medium
S02, 5	High	High	Medium
S02, 6	High	Low	Low
S02, 7	Low	Medium	Low
S02, 8	Medium	Low	High
S02, 9	Medium	Low	medium
S02, 11	Low	High	Low

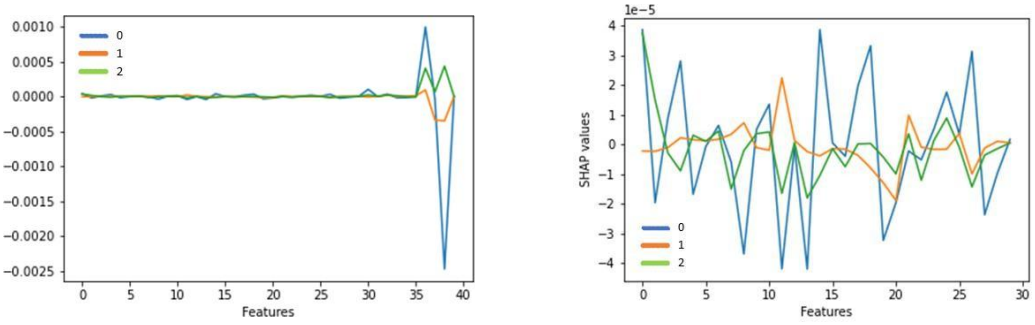
3. Model interpretability with SHAP XAI framework

Although the C-RNN-EEG model variants are performing better than their counterparts, it is essential to understand why they are able to do so. In order to understand how a model can perform well on the DEAP

Journal of Information Systems Engineering and Management
2025, 10 (55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

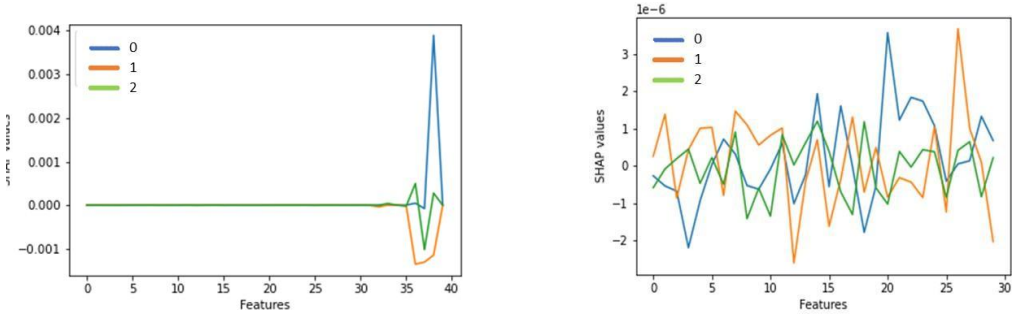
Research Article

dataset, an XAI framework is opted here. Shapley additive explanations, commonly known as the SHAP XAI framework, are considered. A mathematical foundation on the SHAP XAI framework is discussed in section 2. The SHAP XAI framework attempts to explain individual predictions via game theoretically optimal SHAP values. SHAP values are computed for testing samples of the dataset and are reported here. Figures 4(a), 5(a), and 6(a) present the SHAP values from the C-RNN-EEG-A model for all the 40 features (physiological sensory inputs, see Table 1) for participant 's02' watching music video (id: 20). It is evident from these figures that *GSR*, *RR*, *Plethy*, *Temperature* are contributing more towards prediction relative to other features. Therefore, to understand the contributions of EEG signals towards classification output, Figures 4(b), 5(b), and 6(b) present the SHAP values for only EEG signal features for participant 's02' watching music video (id: 20). These figures indicate how different EEG signals contribute to different levels of Arousal state (classes- *Low*, *Medium*, and *High*). It is still challenging to map which parts of the signal (temporally) contribute to the emotional state. Future studies are needed to help us map this. This study restricts itself to mapping features with emotional states only.



(a) **(b)**

Figure 4 SHAP values summary plot for C-RNN-EEG-A model; (a) Patient id: s01-20 for all 40 features, (b) Patient id: s0120 for all EEG features only.

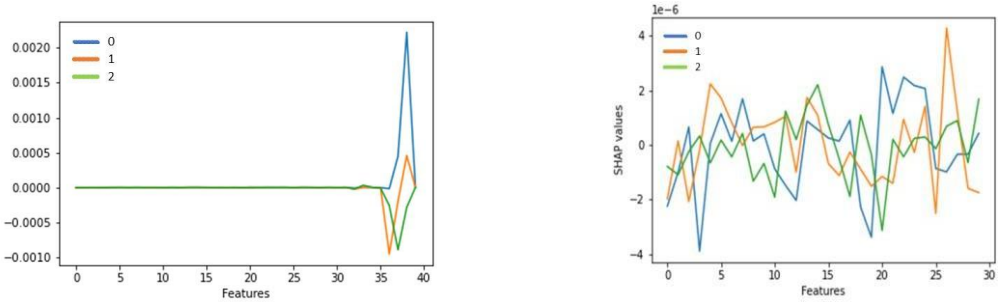


(a) **(b)**

Figure 5 SHAP values summary plot for C-RNN-EEG-V model; (a) Patient id: s02-20 for all 40 features, (b) Patient id: s02-20 for all EEG features only.

Journal of Information Systems Engineering and Management
2025, 10 (55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article



(a) **(b)**

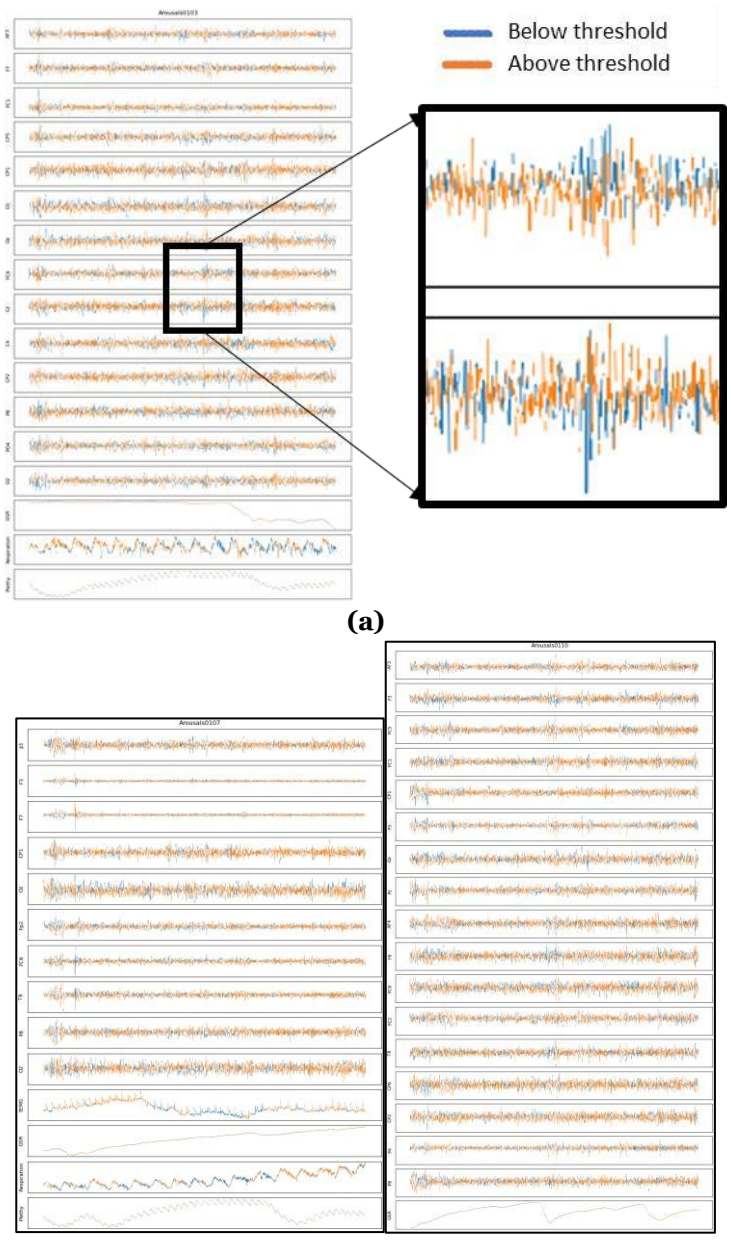
Figure 6 SHAP values summary plot for C-RNN-EEG-D model; (a) Patient id: s02-20 for all 40 features, (b) Patient id: s02-20 for all EEG features only.

4. Gradient-based attributions for time-series signal analysis:

Since the features are time-series signals, it is important to map which parts or segments of the time-series signals are contributing to a particular emotional state. The gradient-based attribution method of the SHAP XAI framework relies on gradients to approximate the contribution of each input feature to the model's output. Model's output to small changes in feature(s) and can give a measure of the feature's local influence on the model output. The SHAP XAI framework builds upon the Integrated Gradients (IG) method (refer to equations 1-5), which addresses the problem of using only the local gradient to approximate feature importance. Integrated gradients calculate the cumulative effect of each feature along a straight-line path from a baseline input (often zero or a neutral state) to the actual input. More details of this method are provided in section 2 (refer to equation 5). This can be achieved by napping the SHAP values over the time-series signals for features contributing over a threshold value. Figure 7(a, b, and c) depicts SHAP values overlaid on selected feature time-series data for participant s01 watching music video id 1, 3, and 7, respectively. All three samples are classified as *Medium-Arousal*. The segments highlighted in orange are time segments contributing towards classification. It can be observed from these figures and the zoomed section that the SHAP values are significant around the same time-stamp in all features. The set of features contributed above the threshold is {*AF3, F7, FC1, CP5, CP1, O1, O2, FC6, C2, C4, CP2, P8, PO4, O2, GSR, Perspiration, Plethy*}.Figure 8(a, b, and c) depicts SHAP values overlaid on selected feature time-series data for participant s01 watching music video id 6, 8, and 9, respectively. All three samples are classified as *Low-Arousal*. The segments highlighted in orange are time segments contributing towards classification. Figure 9(a, b, and c) depicts SHAP values overlaid on selected feature time-series data for participant s01 watching music video id 4, 5, and 11, respectively. All three samples are classified as *High-Arousal*. The segments highlighted in orange are time segments contributing towards classification.

Journal of Information Systems Engineering and Management
2025, 10 (55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

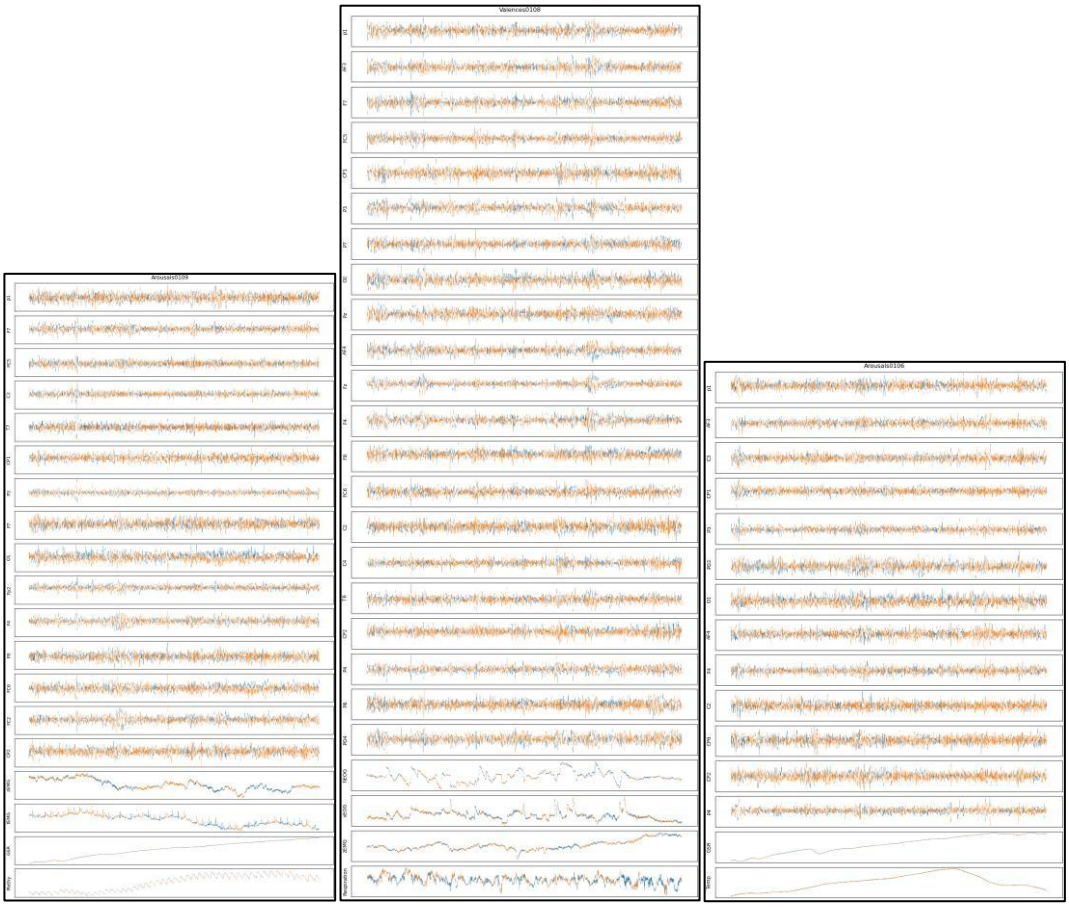


(a) **(b)** **(c)**

Figure 7 SHAP values or integrated gradients overlaid on selected feature time-series data for participant s02 watching music video id (a) 1, (b) 3, and (c) 7. The class label is *Medium - Arousal*. The classification model is C-RNN-EEG-A.

Journal of Information Systems Engineering and Management
2025, 10 (55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

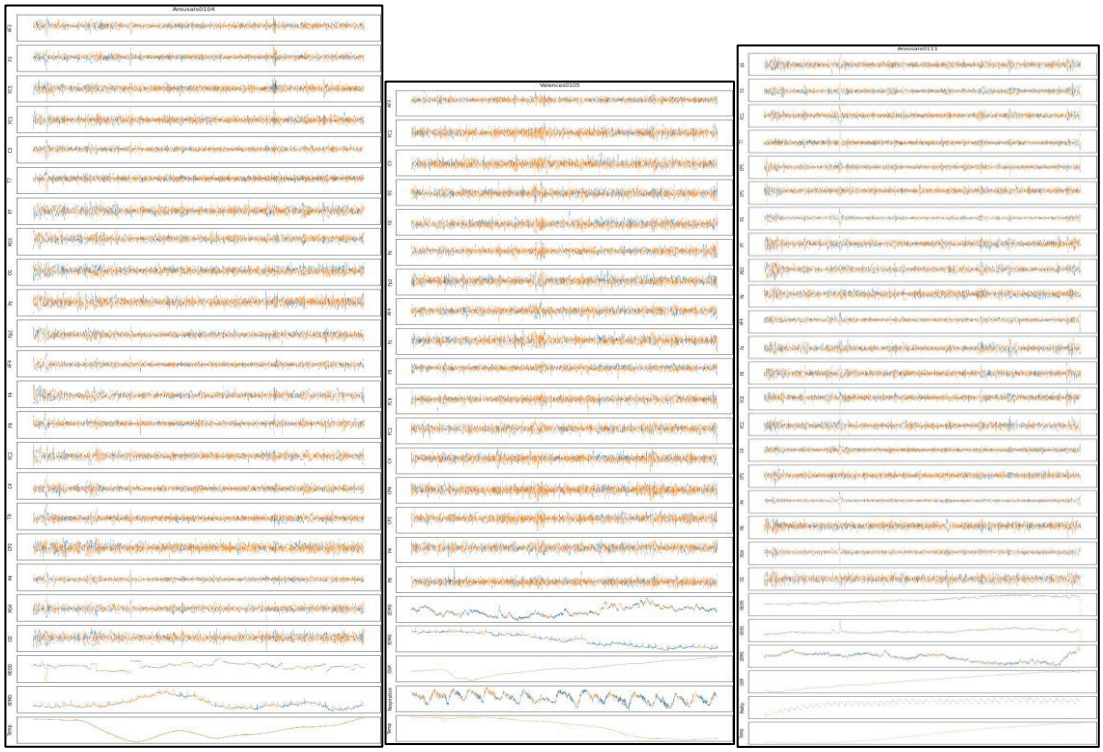


(a) **(b)** **(c)**

Figure 8 SHAP values or integrated gradients overlaid on selected feature time-series data for participant s02 watching music video id (a) 6, (b) 8, and (c) 9. The class label is *Low -Arousal*. The classification model is C-RNN-EEG-A.

Journal of Information Systems Engineering and Management
2025, 10 (55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article



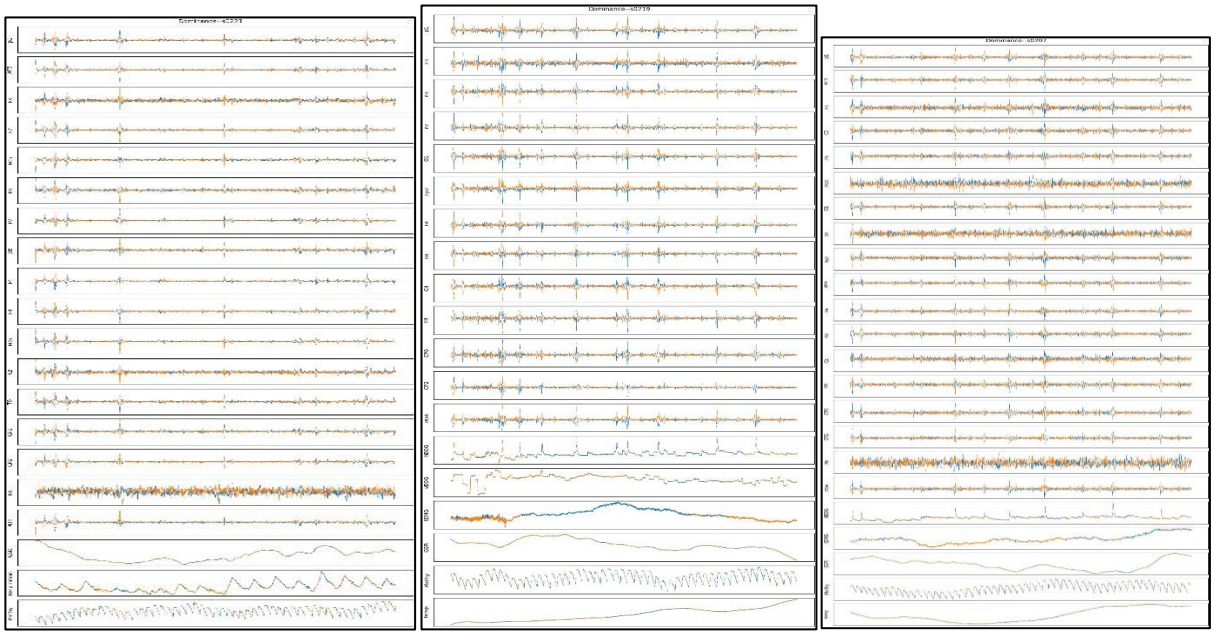
(a) **(b)** **(c)**

Figure 9 SHAP values or integrated gradients overlaid on selected feature time-series data for participant s02 watching music video id (a) 4 (*High-Arousal*), (b) 5 (*High-Arousal*), and (c) 11 (*High-Arousal*). The classification model is C-RNN-EEG-A.

Figure 10 (a, b, and c) depicts SHAP values overlaid on selected feature time-series data for participant s02 watching music video id 7, 19, and 21, respectively. The three samples are classified as *Low-, Medium-,* and *High- Dominance,* respectively. The segments highlighted in orange are time segments contributing towards classification. Finally, Figure 11 (a, b, and c) depicts SHAP values overlaid on selected feature time-series data for participant s02 watching music video id 1, 18, and 20, respectively. The three samples are classified as *Low-, Medium-,* and *High- Valence,* respectively. The segments highlighted in orange are time segments contributing towards classification.Similar studies can be done for other DNN models, such as CNN-only or RNN-only models; however, this study does not cover these.

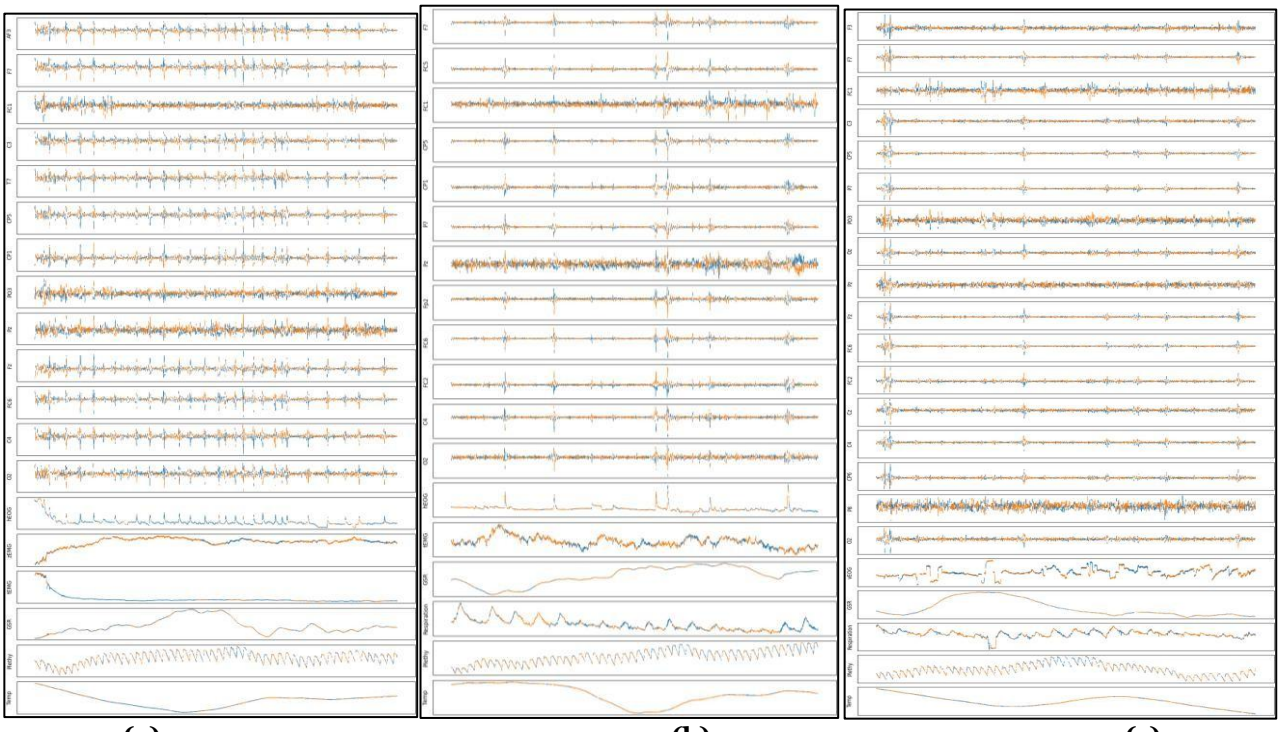
Journal of Information Systems Engineering and Management
2025, 10 (55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article



(a) **(b)** **(c)**

Figure 10 SHAP values or integrated gradients overlaid on selected feature time-series data for participant s04 watching music video id (a) 7 (*Low -Dominance*), (b) 19 (*Medium -Dominance*), and (c) 21 (*High -Dominance*). The classification model is C-RNN-EEG-D.



(a) **(b)** **(c)**

Figure 11 SHAP values or integrated gradients overlaid on selected feature time-series data for participant s04 watching music video id (a) 1 (*Low -Valence*), (b) 18 (*Medium -Valence*), and (c) 21 (*High -Valence*). The classification model is C-RNN-EEG-V.

Journal of Information Systems Engineering and Management
2025, 10 (55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Conclusion

The study presented here establishes the potential of synergistic exploitation of CNNs and RNNs in achieving improved emotion recognition performance with multi-modality, multi-source, and multi-label physiological signals. In order to use the best of both models, the complete signal is first fragmented in the time dimension. From the 60-second full-length recorded signal, ten fragments of 1-second each are made. The CNNs are first employed to extract complex features from high- dimensional, multi-modal, multi-label physiological DEAP data. Two convolutional layers are employed to obtain transformed feature space. Every 1-second transformed feature further acted as a time node for a recurrence layer to exploit temporal information underlying within the features. A single recurrence layer is employed to extract this temporal information. Considerately, a *C-RNN-EEG* model is realized. SHAP XAI framework is used to interpret the performance of the *C-RNN- EEG* model. SHAP values approximating integrated gradients are used to indicate the contributions of features. SHAP-value- based interpretations reveal portions of time-series physiological signals that are contributing to emotion recognition. Participant-wise analysis of features contributing to emotion recognition is also presented. The study reveals the crucial importance of model performance interpretation for a detailed understanding of how and why models are able to perform well and what directions need to be improved. However, more studies aligned towards this objective are beneficial to strengthen the methodology

Statements and Declarations Conflict of Interest

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Funding

The authors did not receive support from any organization for the submitted work.

Financial interests

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Data Availability

The data used in this work is the DEAP: Database for Emotion Analysis using Physiological Signals dataset [30]. The dataset used here is under the Open Access Policy.

Research Involving Human and /or Animals

No humans or animals were experimented upon during this study.

Informed Consent

No consent is required since the data is publicly available and all the required consents have already been obtained by the data publisher.

References

[1] J. L. McGaugh, *Emotions and bodily responses: A psychophysiological approach*. Academic Press, 2013.
[2] S. Z. Li, A. K. Jain, Y.-L. Tian, T. Kanade, and J. F. Cohn, "Facial expression analysis," *Handbook of face recognition*, pp. 247–275, 2005.
[3] K. Oatley, *Best laid schemes: The psychology of the emotions*. Cambridge University Press, 1992.
[4] P. Thagard, *Mind: Introduction to cognitive science*. MIT press, 2005.
[5] M. C. Nussbaum, *Upheavals of Thought: The Intelligence of Emotions*. Cambridge University Press, 2001. doi: 10.1017/CBO9780511840715.

[6] K. Sailunaz, M. Dhaliwal, J. Rokne, and R. Alhajj, "Emotion detection from text and speech: a survey," *Soc Netw Anal Min*, vol. 8, pp. 1–26, 2018.
[7] Y. Wang *et al.*, "A systematic review on affective computing: Emotion models, databases, and recent advances,"*Information Fusion*, vol. 83, pp. 19–52, 2022.
[8] M. Egger, M. Ley, and S. Hanke, "Emotion recognition from physiological signal analysis: A review," *Electron Notes Theor Comput Sci*, vol. 343, pp. 35–55, 2019.
[9] T. D. T. Phan, S. H. Kim, H. J. Yang, and G. S. Lee, "EEG-based emotion recognition by convolutional neural network with multi-scale kernels," *Sensors*, vol. 21, no. 15, Aug. 2021, doi: 10.3390/s21155092.
[10] Q. Yao, H. Gu, S. Wang, and X. Li, "A Feature-Fused Convolutional Neural Network for Emotion Recognition from Multi-channel EEG Signals," *IEEE Sens J*, vol. 22, no. 12, pp. 11954–11964, Jun. 2022, doi: 10.1109/JSEN.2022.3172133.
[11] Y. Chen, R. Chang, and J. Guo, "Effects of Data Augmentation Method Borderline-SMOTE on Emotion Recognition of EEG Signals Based on Convolutional Neural Network," *IEEE Access*, vol. 9, pp. 47491–47502, 2021, doi: 10.1109/ACCESS.2021.3068316.
[12] V. Baltatzis, K. M. Bintsi, G. K. Apostolidis, and L. J. Hadjileontiadis, "Bullying incidences identification within an immersive environment using HD EEG-based analysis: A swarm decomposition and deep learning approach," *Sci Rep*, vol. 7, no. 1, 2017, doi: 10.1038/s41598-017-17562-0.
[13] E. S. Salama, R. A. El-Khoribi, M. E. Shoman, and M. A. Wahby Shalaby, "EEG-Based Emotion Recognition using 3D Convolutional Neural Networks," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 8, pp. 329–337, 2018, doi: 10.14569/IJACSA.2018.090843.
[14] R. Qiao, C. Qing, T. Zhang, X. Xing, and X. Xu, "A novel deep-learning based framework for multi-subject emotion recognition," in *2017 4th International Conference on Information, Cybernetics and Computational Social Systems (ICCSS)*, 2017, pp. 181–185.
[15] X. Li, D. Song, P. Zhang, G. Yu, Y. Hou, and B. Hu, "Emotion recognition from multi-channel EEG data through convolutional recurrent neural network," in *2016 IEEE international conference on bioinformatics and biomedicine (BIBM)*, 2016, pp. 352–359.
[16] S. Roy, I. Kiral-Kornek, and S. Harrer, "ChronoNet: A deep recurrent neural network for abnormal EEG identification," in *Artificial Intelligence in Medicine: 17th Conference on Artificial Intelligence in Medicine, AIME 2019, Poznan, Poland, June 26–29, 2019, Proceedings 17*, 2019, pp. 47–56.
[17] A. Supratak, H. Dong, C. Wu, and Y. Guo, "DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 11, pp. 1998–2008, 2017.
[18] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from EEG with deep recurrent-convolutional neural networks," *arXiv preprint arXiv:1511.06448*, 2015.
[19] J. Thomas, L. Comoretto, J. Jin, J. Dauwels, S. S. Cash, and M. Brandon, "EEG Classification via Convolutional Neural Network-Based Interictal Epileptiform Event Detection," in *Conf Proc IEEE Eng Med Biol Soc.*, 2019, pp. 1–13. doi: 10.1109/EMBC.2018.8512930.EEG.
[20] M. Husken and P. Stagge, "Recurrent neural networks for time series classiÿcation," *Neurocomputing*, vol. 50, pp. 223–235, 2003.
[21] S. Ali *et al.*, "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence," *Information Fusion*, vol. 99, Nov. 2023, doi: 10.1016/j.inffus.2023.101805.
[22] H. Taniguchi, T. Takata, M. Takechi, and A. Furukawa, "Explainable Artificial Intelligence Model for Diagnosis of Atrial Fibrillation Using Holter Electrocardiogram Waveforms," 2021. doi: 10.1536/ihj.21-094.
[23] M. Ganeshkumar, V. Ravi, V. Sowmya, E. A. Gopalakrishnan, and K. P. Soman, "Explainable Deep Learning-Based Approach for Multilabel Classification of Electrocardiogram," *IEEE Trans Eng Manag*, vol. 70, no. 8, pp. 2787–2799, 2023, doi: 10.1109/TEM.2021.3104751.
[24] M. S. Islam, I. Hussain, M. M. Rahman, S. J. Park, and M. A. Hossain, "Explainable Artificial Intelligence Model for Stroke Prediction Using EEG Signal," *Sensors*, vol. 22, no. 24, Dec. 2022, doi: 10.3390/s22249859.
[25] I. Hussain *et al.*, "An Explainable EEG-Based Human Activity Recognition Model Using Machine-Learning Approach and LIME," *Sensors*, vol. 23, no. 17, Sep. 2023, doi: 10.3390/s23177452.
[26] H. Alsuradi, W. Park, and M. Eid, "Explainable Classification of EEG Data for an Active Touch Task Using Shapley Values," in *Human-Computer Interaction*, 2020. doi: 10.1007/978-3-030-60117-1.
[27] K. Zhao, G. S. Member, D. Xu, K. He, and G. Peng, "Interpretable Emotion Classification Using Multidomain Feature of EEG Signals," *IEEE Sens J*, vol. 23, no. 11, pp. 11879–11891, 2023, doi: 10.1109/JSEN.2023.3266322.
[28] J. Manuel, M. Torres, S. Medina-devilliers, T. Clarkson, M. D. Lerner, and G. Riccardi, "Evaluation of interpretability for deep learning algorithms in EEG emotion recognition: A case study in autism," *Artif Intell Med*, vol. 143, no. May, p. 102545, 2023, doi: 10.1016/j.artmed.2023.102545.
[29] C. A. Ellis, D. A. Carbajal, R. L. Miller, V. D. Calhoun, and M. D. Wang, "An Explainable Deep Learning Approach for Multimodal Electrophysiology Classification," in *bioRxiv*, IEEE, 2021, pp. 12–15.
[30] S. Koelstra *et al.*, "DEAP: A database for emotion analysis; Using physiological signals," *IEEE Trans Affect Comput*, vol. 3, no. 1, pp. 18–31, Jan. 2012, doi: 10.1109/T-AFFC.2011.15.
[31] A. Tripathi and T. Choudhury, "Permuted layer-based CNN for Emotion Detection with Multi-Modality Physiological Signals," in *2023 IEEE International Conference on Contemporary Computing and Communications (InC4)*, 2023, pp. 1–5. doi: 10.1109/InC457730.2023.10263176.