International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

An Xai-Based CNN-RNN Model for Emotion Recognition with Multi-Channel Signals

Rishi Kumar Sharma[1*], Vivek Kumar[2], Rajendra Kumar[3]

Submitted: 15/03/2024 **Revised**: 29/04/2024 **Accepted**: 07/05/2024

Abstract – Human emotion recognition is a peculiar task. Humans express emotions via facial gestures, body temperature, and brain activity. Interestingly, brain activities can be observed via EEG recordings. The DEAP dataset is a rich source of multimodal physiological signal representation data including EEG recording encapsulating range of stimulated human emotions. In this paper, a novel CNN-RNN architecture is reported to recognize human emotions using the DEAP data. The CNN-RNN model utilizes the concepts of two-dimensional convolutions in a time-distributed fashion at first and later, recurrence exploits the temporal information. The trained model addresses the issues related to synergistic exploitation of temporal and multi-channel information in high- dimensional feature spaces and attempts to improve the recognition performance. The impact of the model in classification performance is explained via the concepts of SHAP explainable AI (XAI) approach. Results indicate improved classification accuracy and SHAP values from the XAI framework indicate the significance of the architecture in achieving satisfactory performance.

Index Terms *- Emotion detection, time-distribute convolutions, CNN-RNN, Explainable AI learning.*

1. INTRODUCTION

Emotion is a reaction to an event based on conscious and subjective experience such as seeing a video, listening to music or sound, reading a book, or expressing by any particular biological response. Emotions show an individual physiological status which is helpful in some medical diagnoses, suspect identification in crime, games, marking/gamming, education etc. [1], [2], [3]. These reactions may be internal or external biological responses [4], [5]. The reactions such as voice change, facial expression and body language are identified as external reactions however many inner expressions (physiological expressions) are also used for emotion detection and identification.

Before using inner expression data, text, facial expression and speech were the most common methods to detect emotions [5], [6]. Using physiological data towards emotion recognition has become an appropriate alternative to external expression (facial expressions, text, and speech data). Because external expressions-based emotion detection and classification can be easily manipulated. That is why many recent researches have focused on physiological data [7]. Physiological data models can be utilized in unimodal or multimodal approaches for emotion detection.

However, both unimodal and multimodal emotion detection methods have their pros and cons. The multimodal 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 multimodal emotion detection method commonly gives better accuracy, compared to the unimodal method however multimodal emotion detection method needs longer processing time and has complex data collection procedure compared to the unimodal method [8].

Among other methods, deep learning methods have gained popularity in emotion identification with physiological signals. In particular, the use of CNNs and RNNs in this regard is worth mentioning. However, recently, the combined use of CNNs and RNNs for emotion identification is gaining attention. Li et al. [9] applied wavelet features to train CNN combined with LSTM and the binary classification accuracy reached 72%. Roy et al. [10] concentrated on natural and abnormal brain activities and suggested four different DL architectures based on CNN-GRUs. The proposed ChronoNet model achieved 90.60% and 86.57% training and test accuracies, respectively. Supratak et al. [11] 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 automatically from EEG epochs.

[1*]*Research Scholar, Department of Computer Science Engineering, Quantum University. Roorkee, Uttarakhand*
[2]*Professor, Department of Computer Science Engineering, Quantum University. Roorkee, Uttarakhand*
[3]*Associate Professor, 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*

This approach achieved an accuracy of 90%. Bashivan et al. [12] proposed a novel approach for learning representations from multi-channel EEG time series, and demonstrated its advantages in the context of the mental load classification task. They train a deep recurrent CNN inspired by state-of-the-art video classification to learn robust representations from the sequence of images. The proposed approach is designed to preserve the spatial, spectral, and temporal structure of EEG which leads to finding features that are less sensitive to variations and distortions within each dimension. They achieved an overall accuracy. From the above literature, this current study finds motivation to explore and utilize CNN-RNN models for emotion identification with multi-channel physiological signals including EEGs.

It is also notable that although studies have shown that EEG signal classification via deep learning models can achieve high prediction accuracy [13] [14] but, 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 increasingly significant tool in the AI world because of its application in understanding critical decisions as well as the fact that regulators hold businesses responsible for the judgments their AI models make. 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 model and the explainability model. [15]. Explaining the black-box model's decision output is the main goal of the explainability model. Explainability tries to answer the '*why an algorithm produces a particular response*' question. Therefore, it takes into consideration issues like as the weighting of each variable inside the model in order 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 that the model utilized in order to arrive at its predictions. Frameworks like GradCAM [16], [17], Local Interpretable Model-Agnostic Explanation (LIME) [18], [19], Shapley Additive explanations (SHAP) [20] [21], Layer-wise Relevance Propagation (LRP) [22], [23], 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 is capable of rapidly producing 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 tends to attribute 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 are made by CNN-based models, the Grad-CAM approach is yet another example.

This paper highlights the study in which a novel human emotion classification model is developed based on the concepts of CNNs, RNNs, and XAI. The study proposes a novel model for use in human emotion classification. In particular, at first, a complete multi-channel/multi-lead signal is fragmented in time dimension. The CNNs act on these time-wise fragments of the entire signal whereas the RNNS act on temporally distributed transformed features obtained from the CNNs. The Shapley Additive Explanations or more commonly termed as the SHAP method is used for explainability of the significance of the model for the application. 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 is discussed describing model configuration and impact. Section 3 highlights classification performance results obtained using the proposed CNN-RNN model on the popular DEAP dataset and its comparison with other recent results. This section also covers the discussion on the explainability of the proposed model performance. Finally, section 4 concludes the study.

2. MATERIAL AND METHODS

Dataset

The DEAP database is a Database for Emotion Analysis using the Physiological signals dataset [24]. It includes EEG signals and certain other physiological signals of 32 participants. These signals are recorded while they watched 1-minute music videos. Overall, each participant watched 40 such videos and corresponding physiological responses are stored amounting to 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 emotional responses majorly include arousal, valence, like dislike, dominance and familiarity. Preliminary analytics are made on the data. Two-dimensional histograms are used to visualize sample distribution for each major emotional response and are shown in Figure 1. For example, it is clear in Figure 1(a) that participant number 15 showed a 'high-valence' response to approximately 15 videos, a 'high-arousal' response to approximately 10 videos, a 'high-dominance' response to only 2 videos and, a 'high-liking' response to 25 videos. Also, participant 15 was not familiar with any of the music videos.

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

Dataset	DEAP
Participants	32
Signals	44
EEGs, EMGs, EOGs, GSR, RR, Plethy, Temperature	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

Methodology

A multimodal emotion classification process has a few key steps. Based on those steps, a methodology as shown in Figure 2 is opted here in this study. The First step is a data acquisition/collection step followed by its preprocessing step which removes unwanted noise from data. Next is an extraction or selection of features step and the final step is the classification of one of the target labels. Here a CNN-RNN based model is proposed for emotion classification.

CNN-RNN Model: Configuration, training and validation

A classification model is developed using CNNs, RNNs, and Dense layers. The CNN layers act on time-wise fragments of the entire signal. Each time-wise fragment acts as a time stamp. CNN layers extract more complex features from the high dimensional input feature space of 40 features as discussed in the dataset section above. The extracted features from CNNs at any particular fragment act as one time-stamp signal for RNNs. The RNNs exploit this 10 time-stamp complex signal and attempt to extract temporal information from the signal. The model gives the advantage of both CNNs which are excellent feature extractors and RNNs which can represent the entire past in one activation. Based on this notion, a CNN-RNN model is developed. Its configuration, training, and hyperparameter settings are discussed here. Table 2 summarizes the model configuration. At first, the input signal with 60 seconds and 40 features information is fragmented into signals of 10 seconds each. These 10 second signals are then exploited using several convolutional layers and pooling layers. Two-dimensional convolutional filters are used for this purpose. Once the features are transformed, these are concatenated to be put into a recurrence layer for temporal information extraction. Each 10-second information acts as a time node. The recurrent layer provides extracted features that are finally connected to a classification layer with 'SoftMax' via a fully connected mode.

Model Training and Hyperparameter Settings

The model is trained and tested on 1280 samples. An 80/10/10 ratio has opted for training, validation, and testing respectively.

Categorical Cross Entropy is chosen as the loss function and Adam method is considered for optimization. A batch size of 16 is set as the total sample size is low. The model is trained for over 70 epochs and performance saturation is achieved.

contributing more in comparison to other parts towards classification.

3. Results and Discussion

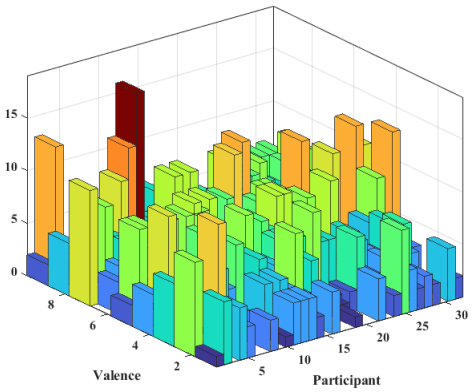
The CNN-RNN model summarized in Table 2 is applied to the DEAP dataset discussed under the *Dataset* section. Classification performance over different emotions is listed in Table 3. Classification performances of other popular human emotion classification methods reported in the literature [25], [26], [27]are also listed in Table 3.

It is clear from the table that the proposed CNN-RNN model has performed better than support vector machines (SVMs), logistic regression (LR), decision trees (DT), K-nearest neighbors (KNN), and linear discriminant analysis (LDA).

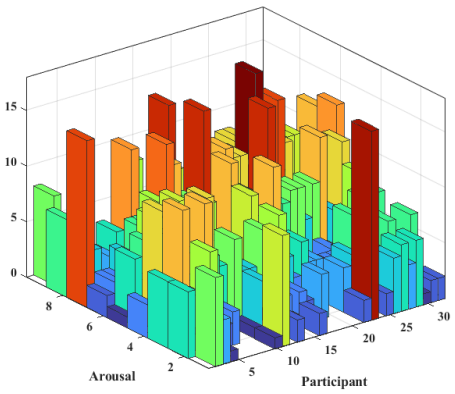
In order to understand how the proposed can perform well on the DEAP dataset, an XAI framework is opted for. Shapley Additive Explanations or more commonly known as SHAP method are considered. This method tries to explain individual predictions via game theoretically optimal SHAP values. SHAP values are computed for testing samples of the dataset are reported here. Figure 3 depicts the SHAP values for three different class labels. It is clear that for the 'valence' class (in blue), feature 22 is most significant whereas for 'Liking', feature 15 seems more significant. In contrast to feature significance, SHAP method also identifies signal segments in time that are contributing more towards the classification of a particular signal. Figure 4 depicts parts of an arbitrary signal of the 'valence' class. The segments highlighted in orange are time segments

Conclusion

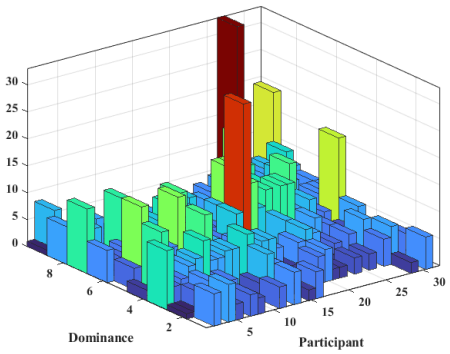
The study presented here establishes the potential of synergistic exploitation of CNNs and RNNs in achieving improved emotion classification performance with multi-modality physiological signals. In order to use the best of both models, the complete signal is first fragmented in the time dimension. From the 60 - second recorded signal, 10 fragments of 1 second each are made. The CNNs are firstly deployed to extract complex features from high- dimensional multi-physiological DEAP data. Two-layer architecture is 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 the features. Two-dimensional convolutions in the convolutional layers and LSTM units in the recurrence layer are used. The CNN-RNN model proposed here achieved satisfactory classification accuracy for all major emotions namely valence, arousal, liking, and dominance classes. The SHAP method-based feature importance plots are drawn from the CNN-RNN model. The SHAP values and analytics helped explain the better performance of the proposed model.



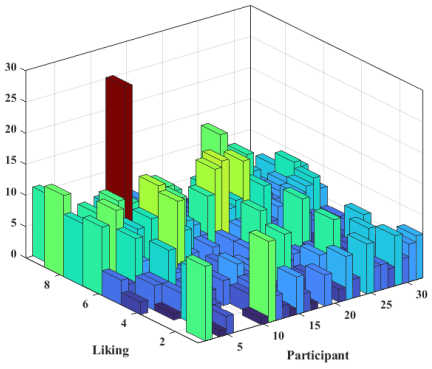
(a)



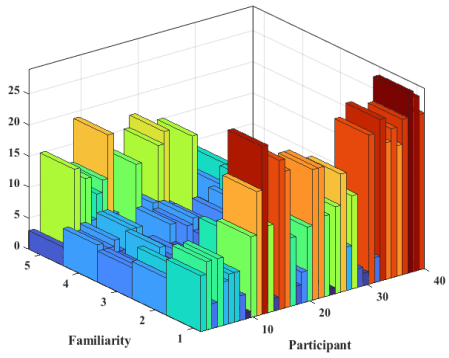
(b)



(c)



(d)



(e)

Figure 1 Participant ratings two-dimensional histogram: (a). Participant -Valence, (b). Participant-Arousal, (c). Participant-Dominance, (d). Participant-Liking, and (e). Participant-Familiarity.

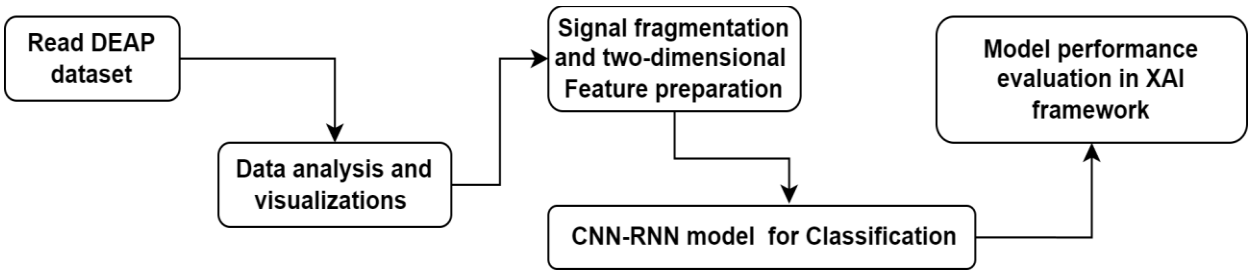


Figure 2 Flowchart for the proposed CNN-RNN based emotion classification model

TABLE 2 CNN-RNN model configuration

Input signal shape	Fragmentation stage	Convolution stage 1: 2D convolution + 2D Pooling (For each of the 10 fragments)		Convolution stage 2: 2D convolution (For each of the 10 fragments)		Concatenation Concatenate all 10 fragments	Recurrence stage: LSTMs		Classification stage
		Configuration	Output shape	Configuration	Output shape	Shape: 10*40 (10 being time stamps and 49 being features)	Configuration	Output	Number of labels: 4 • Activation: Softmax
8064X40	10 fragments in time creating each fragment of shape 806X40	Filter shape: 130*10 Number of filters: 20 Pooling size: 2*2	337*15*20	Filter shape: 337*15 Number of filters: 40	1*40		Number of filters: 20 Recurrence shape: 10 Signal flow: Bidirectional. Strategy = Many to one	1*20	
Hyperparameter settings	Loss: categorical cross entropy Batch size: 46 Number of epochs: 40 Optimizer Adam								

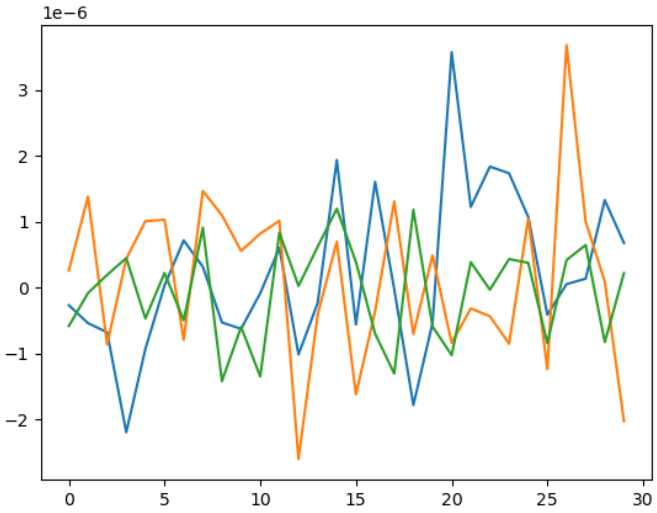


Figure 3 SHAP values for each feature for an arbitrary sample.

Table 3 Classification accuracies of popular algorithms in human emotion classification with DEAP dataset [27].

Class	Method	Accuracy (%)
Arousal	CNN-RNN	69.10
	Perm-Feature-CNN	67.10
	LDA	63.12
	LR	64.82
	DT	64.53
	KNN	63.35
	SVM	63.86
Liking	CNN-RNN	74.25
	Perm-Feature-CNN	71.25
	LDA	67.34
	LR	68.08
	DT	67.78
	KNN	63.12

	SVM	67.18
Valence	CNN-RNN	69.54
	Perm-Feature-CNN	68.54
	LDA	56.25
	LR	63.12
	DT	64.6
	KNN	57.68
	SVM	63.59
Dominance	CNN-RNN	68.5
	Perm-Feature-CNN	69.5
	LDA	66.67
	LR	67.73
	DT	71.73
	KNN	63.12
	SVM	66.25

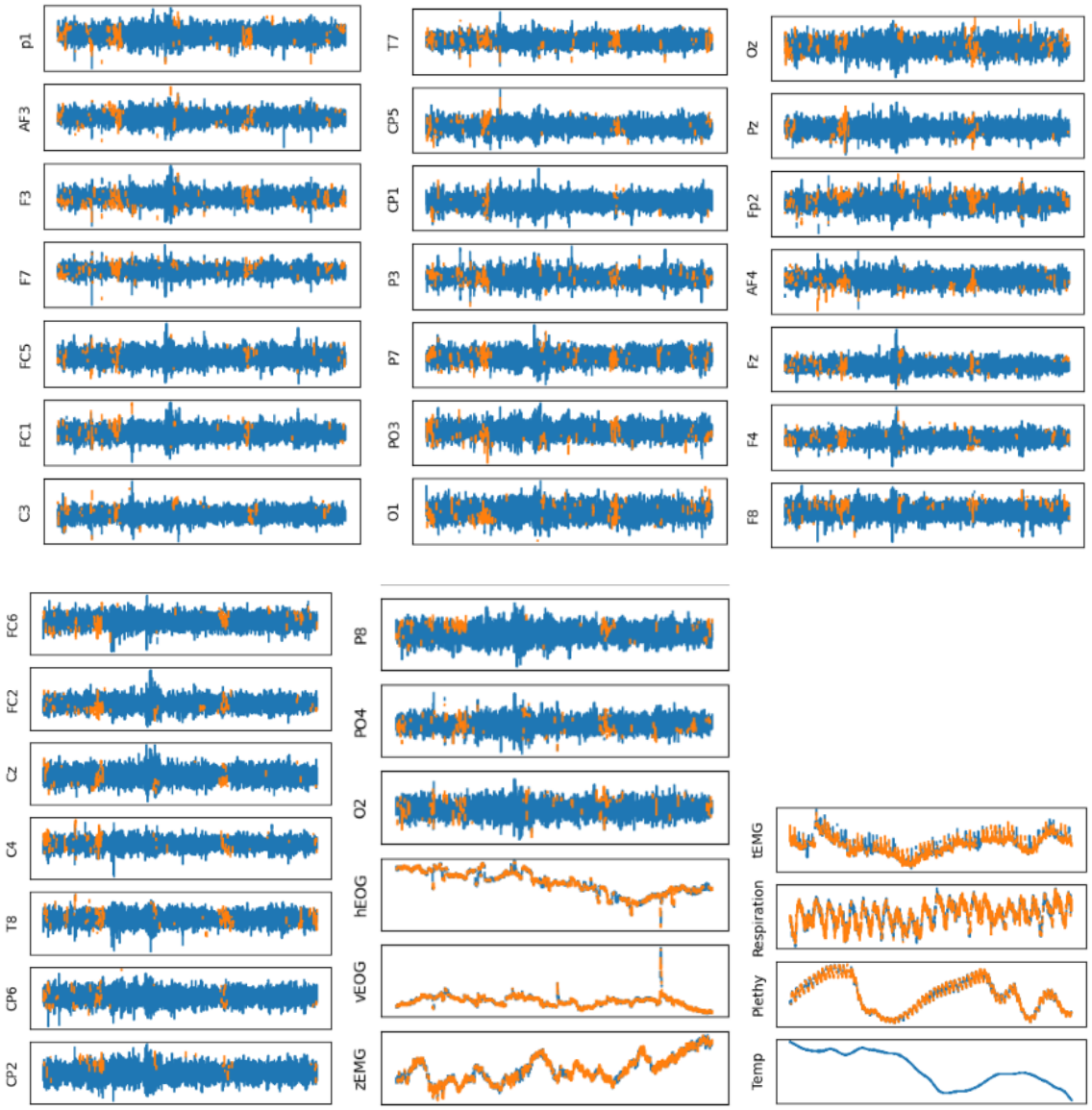


Figure 4 EEG signal segments in time from each feature, contributing to the classification of the 'Valence' class.

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 [24]. The dataset used here is under Open Access Policy.

Authors' contributions

Rishi Kumar Sharma is the lead on the study. His contributions include - data selection, algorithm implementation and paper writing. Manish Sharma contributed towards the ideation of the approach and analytical discussion on the results. Rajendra Kumar contributed to defining the problem statement, algorithm implementation, and paper writing.

Research Involving Human and /or Animals

No Humans or Animals have been experimented upon during this study.

Informed Consent

No consent is required since the data is publicly available and all the required consents are already taken by the data publisher.

References

[1] K. Oatley, *Best laid schemes: The psychology of the emotions*. Cambridge University Press, 1992.

[2] P. Thagard, *Mind: Introduction to cognitive science*. MIT press, 2005.

[3] M. C. Nussbaum, *Upheavals of Thought: The Intelligence of Emotions*. Cambridge University Press, 2001. doi: 10.1017/CBO9780511840715.

[4] J. L. McGaugh, *Emotions and bodily responses: A psychophysiological approach*. Academic Press, 2013.

[5] 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.

[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] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] M. Husken and P. Stagge, "Recurrent neural networks for time series classiÿcation," *Neurocomputing*, vol. 50, pp. 223–235, 2003.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Information Fusion*, vol. 59, pp. 103–126, Jul. 2020, doi: 10.1016/J.INFFUS.2020.01.011.

[26] V. Doma and M. Pirouz, "A comparative analysis of machine learning methods for emotion recognition using EEG and peripheral physiological signals," *J Big Data*, vol. 7, no. 1, Dec. 2020, doi: 10.1186/s40537-020-00289-7.

[27] 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.