2023 IEEE Symposium Series on Computational Intelligence (SSCI)
Mexico City, Mexico. December 5-8, 2023

Using contrastive learning to inject domain-knowledge into neural networks for recognizing emotions

Guido Gagliardi*[,1,2,3], Antonio Luca Alfeo[1,4], Vincenzo Catrambone[1,4], Mario G.C.A. Cimino[1,4], Maarten De Vos[2], and Gaetano Valenza[1,4]

***Abstract*—With application contexts ranging from psychophysiology to neuromarketing, electroencephalography (EEG)-based emotion recognition is a fundamental technology for affective computing. In this context, EEG signals can be processed via artificial neural networks (NNs) to achieve accurate recognition of users' emotions. Still, NNs are rarely employed in real-world decision-making processes, since their internal model works as a hardly trustable black box. A NN's reasoning can be explained in a human-comprehensible manner by exploring its latent space to understand if some domain knowledge is actually represented and exploited for the classification. Those approaches assume that a trained NN autonomously organizes its latent space according to some domain concepts to process the data via human-like reasoning. However, there is no guarantee that such an assumption holds, since the latent space is not built for this aim. On the other hand, forcing the organization of the latent space (e.g. via contrastive learning) can result in poor recognition performances due to information loss. To guarantee great recognition performances and provide a domain-knowledge-driven organization of NNs' latent space, we combine the well-known training procedure based on a categorical cross-entropy loss with a supervised contrastive learning approach for continuous values labels. The proposed approach (i) enables the explanation of NN's reasoning in terms of the importance of high-level domain concepts in the final classification, and (ii) results in a recognition performance comparable to or better than the one achieved via an approach based solely on maximizing recognition. The proposed approach is tested on the publicly available MAHNOB dataset.**

***Index Terms*—eXplainable Artificial Intelligence, Neural Networks, Affective Computing, Informed Machine Learning, Contrastive Learning**

I. Introduction

Affective computing is a comprehensive research domain focused on investigating emotional and mental states by analysing physiological signals. These approaches are increasingly required thanks to their effectiveness in identifying patterns associated with affective disorders (e.g., anxiety and depression) [1], and the adoption of user-friendly, non-intrusive, portable devices capable of collecting reliable physiological data [2]. Emotions can be modelled through a multidimensional space representation, known as the circumplex model of affect [3], comprising three dimensions: valence (ranging from positive to negative feelings), arousal (from mild to excited states), and dominance (from subtle to engaging emotions). Emotions can also be categorized via discrete basic emotions (joy, trust, fear, surprise, sadness, disgust, anger, and anticipation) [4]. Each of these categorical emotions is placed in a different space of the circumplex model of affect.

When it comes to emotion recognition tasks using non-invasive physiological data, electroencephalography (EEG) is commonly employed due to its favourable balance between temporal and spatial resolution [5]. EEG signals are typically collected by positioning a set of sensors (i.e. channels) on the patient's scalp, arranged according to standardized schemes such as the international standard pattern 10-20 [6]. According to a recent survey [7], the majority (89.4%) of EEG-based emotion recognition studies extract frequency domain features, e.g. using methods like Power Spectral Density (PSD). These location-specific features can be transformed into a tabular format, and associated with a label for each instance in the classification process. Recent studies have shown how, using a feature arrangement to generate an image, it is also possible to represent the very useful information of spatial proximity between electrodes in the analysis [8]. This results in superior recognition performances if combined with a classifier based on Convolutional Neural Networks [9].

Despite the impressive recognition performances, the results provided via NNs are hardly employed in decision-making real-world scenarios, since their internal model

[1] Department of Information Engineering, University of Pisa, Pisa, Italy.
[2] Dept. of Electrical Engineering, KU Leuven, Belgium
[3] Dept. of Information Engineering, University of Florence, Italy
[4] Bioengineering & Robotics Research Center E. Piaggio, School of Engineering, University of Pisa, Pisa, Italy.
∗ Correspondence to: guido.gagliardi@phd.unipi.it

Work partially supported by (i) the EU Commission - Horizon 2020 Program under GA 101017727 of the project "EXPERIENCE"; (ii) the Italian Ministry of Education and Research (MIUR) in the framework of the FoReLab project (Departments of Excellence); (iii) the Italian Ministry of University and Research (MUR), in the framework of the "Reasoning" project, PRIN 2020 LS Programme, Project number 2493 04-11-2021; (iv) PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded by the European Commission under the NextGeneration EU programme, (v) PNRR - M4 C2 - Investimento 1.5, Creazione e rafforzamento di "ecosistemi dell'innovazione", costruzione di "leader territoriali di R&S", funded by the European Commission under the NextGeneration EU programme, project "THE" subproject 6; (vi) the University of Pisa, in the framework of the PRA 2022 101 project "Decision Support Systems for territorial networks for managing ecosystem services". This study uses the MAHNOB Database collected by Prof. Pantic and the iBUG group at Imperial College London and in part collected in collaboration with Prof. Pun and his team at University of Geneva, in the scope of the MAHNOB project financially supported by the ERC under the European Community's 7th Framework programme (FP7/2007-2013)/ERC starting GA 203143.

2023 IEEE Symposium Series on Computational Intelligence (SSCI) | 978-1-6654-3065-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/SSCI52147.2023.10371895

works as a black box [10]. To enable the trust in the outcomes of a NN it is essential to offer some explanations for its reasoning that can be easily understood by professionals and decision-makers who lack a background in machine learning (ML). Among the many explanation strategies, exploiting the domain knowledge can provide the most understandable explanations [11]. Indeed, this translates into the ability to understand if the NN is providing "the right answer for the right reasons" [12] on a human-like abstraction level, rather than via numerical representations or measurements. For example, in the case of emotions, it is possible to use concepts such as arousal, valence, and dominance to understand how a NN classifier distinguishes different emotions.

For this reason, the attention of the academic community has recently focused on examining if and how the NNs' reasoning can be tailored to resemble human decision-makers by representing some domain knowledge in its inner model [13]. To this aim, the latent space of the NNs can be explored to understand how the NN processes each sample to achieve the classification result [14]. For instance, in domains related to image recognition, the latent space of NNs in their first layers can represent low-level domain elements such as textures and edges [15], while those in the latest layers can represent higher-level domain concepts like specific objects [16].

However, according to different posthoc analyses on trained NNs [17], the information about a specific domain concept can be scattered throughout the whole network rather than triggering a specific part of the NN [18]. Given the difficulties related to understanding if and which NN nodes represent human-understandable concepts, the activation of a group of NN nodes can be linearly combined to represent some predefined higher-level concepts [19]. For instance, in [20] the authors propose a method to detect some Concept Activation Vectors, i.e. vectors in the latent space of an NN that are specifically chosen to align with predefined or automatically discovered concepts. Still, most of those approaches work under the assumption that a trained NN "places" each domain concept in one easy-to-classify portion of its latent space. However, since the latent space was not explicitly built to have this property, there is no reason to believe that the above assumption holds [17]. Rather than relying on these assumptions, the latent space of an NN can be constrained during its training to represent some domain concepts [17]. For instance, in [21], the authors constrain the NN model to represent some human-specified concepts; however, since each concept is represented via one single direction of the latent space, this may result in information loss.

To organize the latent space without sacrificing the recognition performance, in this study we propose an approach based on a combination of the well-known training procedure employing a categorical cross-entropy and a supervised contrastive learning approach for continuous values labels. Specifically, the NN is trained in an end-to-end fashion to aggregate (separate) instances belonging to similar (different) concepts (i.e. arousal, valence, or dominance values) in its latent space while keeping great recognition performance. The proposed approach (i) is specifically designed to organize the latent space of the NN in a domain-knowledge-driven manner, and (ii) allows measuring the importance of each domain concept for the classification as an explanation for its reasoning.

II. Materials and Methods

A. Experimental data

In this study, we used the publicly available MAHNOB dataset, which involved collecting physiological signals from a group of healthy volunteers who watched emotional videos. The evaluation of emotions was based on the circumplex model of affect, which defines emotions in a three-dimensional space: arousal, valence, and dominance. Arousal measures the intensity of the feeling, valence measures the pleasantness of the feeling, and dominance measures the perceived control over the feeling. The dataset consisted of data from 27 healthy participants, whose physiological signals were collected via a EEG helmet featuring 32-channel and 256 Hz sampling rate. Participants rated each video on the scales of valence, arousal, and dominance and assigned a label corresponding to the value of the categorical emotion perceived. All participants provided informed consent, as documented in the original paper that presented the dataset. The dataset can be accessed at *https://mahnob-db.eu/hci-tagging/*.

B. EEG Signal preprocessing

Extensive details on the performed EEG signal preprocessing steps can be found in [8]. Briefly, to obtain clean EEG signals for the classification task, we implemented a series of processing steps, including frequency filtering, artefact rejection, removal of eye and cardiac artefacts, interpolation of contaminated channels, and average re-referencing. The EEG power spectral density was extracted using Welch's method with a Hanning window. The PSD time series were integrated into four frequency bands: theta (θ), alpha (α), beta (β), and gamma (γ).

C. Arrangement of spatial features into an image

To organize the input features representing each frequency band and EEG channel, a sparse matrix was created using the 10-20 EEG electrode placement representation. This arrangement scheme has been effective in allowing CNN models to leverage the spatial relationship between sensors, resulting in improved performance in emotion recognition tasks (see further details in [8]).

D. NN-Based Architecture

In this section, we describe a novel neural network architecture for emotion recognition. Our architecture is designed to effectively extract meaningful features from input images

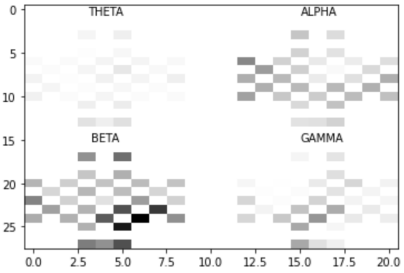

Fig. 1: Representation of the 2×2 block matrix as a greyscale image.

exploiting a supervised contrastive learning methodology and leverage them for accurate classification.

The architecture starts with two convolutional layers, each followed by a Rectified Linear Unit (ReLU) activation function to introduce non-linearity.

The convolutional layers are followed by a max-pooling layer to downsample the feature maps and reduce the spatial dimensions. Next, we employ a flattening layer to convert the 2D feature maps into a 1D vector.

This allows us to feed the extracted features into a fully connected layer. The subsequent layer, which we refer to as the Contrastive Layer (CL), serves two purposes: it applies a contrastive learning approach, which encourages the model to learn an input representation resembling the circumplex model of affect in an Informed Machine Learning fashion; it performs normalization on the output elements to ensure that they are in a normalized form.

The CL forces the network to exploit the circumflex space related information while predicting categorical emotions. In this way the network can be viewed as separated in two parts: in the first it learns a new representation of the data in which the enhanced with the arousal, valence and dominance information; in the second the network use this new representation to classify the target emotion. Hence, it first represents the input data in a new space where its position resembles the one expressed in the circumplex space, and then it uses this new representation to classify human emotions. This procedure simulate the human reasoning while classifying emotions inside a neural network, enhancing its robustness and understandability.

Following the contrastive layer, we introduce two fully connected layers for classification purposes. The first fully connected layer uses a ReLU activation function, and the second instead utilizes a Softmax activation function, providing probabilities for each categorical emotion (e.g. happy, angry, etc.). The final layer of the architecture has a number of neurons equal to the number of emotion categories in the MAHNOB dataset, which is 9. This layer allows the model to output the predicted probabilities for each emotion category.

To train our model, we employ a training strategy that involves training for 1000 epochs with early stopping and checkpointing. We evaluate the best classification performance on the validation set using a nested Monte Carlo 5-fold cross-validation approach, ensuring robust evaluation and preventing overfitting.

E. Loss Function: CCL and SCL

The loss function used in our neural network architecture combines Categorical Crossentropy Loss (CCL) and Supervised Contrastive Learning (SCL). The framework was implemented in PyTorch.

1) CCL: The CCL measures the discrepancy between the predicted output and the ground truth categorical labels. Let y_i represent the ground truth categorical label for input i, and let p_i represent the predicted output probabilities for each emotion category. The CCL is computed as follows:

$$\text{CCL}(y_i, p_i) = -\sum_j y_{i,j} \log(p_{i,j})$$

2) SCL: The SCL aims to learn a representation in the output of the Contrastive Layer that preserves the distances between samples in the circumplex space. In the circumplex model of affect, emotions are represented in a three-dimensional space: arousal, valence, and dominance. We want the projection of samples i in the Contrastive Layer, represented by c_i, to preserve the distances between samples in the circumplex space, represented by $y_i = (a_i, v_i, d_i)$.

To calculate the SCL, we define an anchor point a, and select two samples i and j from the batch for mining. The loss is given by:

$$\text{SCL}(a, i, j) = \left(\log\left(\frac{d(c_a, c_i)}{d(c_a, c_j)}\right) - \log\left(\frac{d(y_a, y_i)}{d(y_a, y_j)}\right)\right)^2$$

Here, c_a represents the projection of the anchor point a, $d(c_a, c_i)$ is the Euclidean distance between the projections of samples a and i, and $d(y_a, y_i)$ is the Euclidean distance between the circumplex representations of anchor a and sample i.

Mining strategies can be employed to select the samples i and j. Mining is one of the most important aspect of contrastive learning and metric learning [22], which is the process of finding the best samples to train on. The easiest way to determine train samples in contrastive learning is by means of randomly chosen positive and negative pairs of objects [23], i.e. positive elements refers to element of the same class of the anchor point a and negative to sample of different class. When it comes from contrastive learning for continual labels, the problem is even more severe, since positive elements can be not properly represented in the training batch. For this reason, authors in [24] propose to mine the 2-nearest samples of the anchor point considering their actual distance in the latent space, i.e $d(c_a, c_i), d(c_a, c_j)$. However, this approach is highly sensitive to the ϵ parameter, which ensure that the denominator $d(c_a, c_j)$ does not become 0. Moreover, by focusing on the nearest sample, the approach tends to enhance the formation of micro-cluster in the latent space instead of building a homogeneous space representation. In our approach, we use the farthest samples for mining, aiming to encourage more global coherence in the space.

By selecting distant samples, we ensure that the network focuses on capturing broader relationships and patterns in the circumplex space rather than fine-grained clusters.

The overall loss function is obtained by combining the CCL and SCL, weighted by parameters α and β, which are both set to 0.5 to equally balance the importance of each component. The combined loss is calculated as:

$$\text{CombinedLoss} = \alpha\text{CCL} + \beta\text{SCL}$$

By optimizing this combined loss function during training, our model learns to accurately classify emotions while preserving the distances in the circumplex space, leading to improved emotion recognition performance.

III. Results and Discussion

In this section, we detail the experimental setups and the results obtained from our architecture to demonstrate the effectiveness of our multi-objective loss function in simultaneously optimizing the conceptual representation of the latent space and the model's performance. Our goals are twofold:

(1) Injecting (conceptual-)knowledge into the model: We aim to align the data representation with the circumplex space, which should either increase or maintain classification performance compared to training the model solely with CCL. To evaluate this, we compared the accuracy of our proposed model with a baseline model obtained by replicating the proposed neural network architecture but trained only with CCL. Similarly, we also compared our approach with two state-of-the-art methods for conceptual representation in latent spaces: an updated version of Conceptual-Space-Embedding (CSE) and another contrastive learning approach for continuous label representation. The latter approach, unlike our proposed method, utilizes a mining mechanism based on the nearest samples rather than the farthest samples.

(2) Assessing the effectiveness of different architectures in representing and utilizing concept-related information during classification: inspired by Concept Activation Vector (CAV) approach, we extract information on how well a projection in the latent space can be recognized as representing high or low arousal, valence, or dominance. The CAV score is computed using a linear classifier with softmax activation and two output neurons, where one predicts the "high" class and the other predicts the "low" class. For each concept (e.g., arousal, valence, dominance), we compute the CAV score by taking the absolute difference between the probability scores for the "high" and "low" classes. This represents the separability of the concept in the latent space. Higher CAV scores indicate a more distinct separation between the classes, i.e. arousal, valence and dominance, in the latent space.

The CAV scores at the CL (Contrastive Layer) level represent the effectiveness of the architecture in capturing and representing the underlying concept in the latent space. These scores indicate how well the model has learned to encode the information related to the concept of interest (e.g., arousal, valence, dominance) in the first part of the network.

On the other hand, the CAV scores at the DENSE and CLASS layers provide insights into how effectively the architecture utilizes the concept-related information during the classification task. These scores indicate to what extent the model takes advantage of the encoded concept information in the subsequent layers of the architecture, leading to improved classification performance.

For example, if we observe high CAV scores at the CL level but relatively lower scores at the DENSE and CLASS levels, it suggests that the architecture captures the concept well in the latent space but it is not exploiting that information during the classification process. Conversely, high CAV scores at both the CL and DENSE/CLASS levels indicate that the architecture not only captures the concept in the latent space but also that it use the concept information during classification.

Table I reports the accuracy values of the analyzed models and their corresponding CAV scores for arousal, valence, and dominance at three levels where the contrastive loss is applied (CL, DENSE, CLASS).

Our approach achieved a f1-score of 0.89, demonstrating its strong classification performance also in comparison with the state-of-the-art literature [9]. The CAV scores at the CL level indicate that our architecture effectively captures the concept information, with high arousal, valence, and dominance scores of 0.94, 0.95, and 0.96, respectively. These results are consistent with the other approaches, which exhibit lower classification performances in terms of f1-score but similar conceptual representation metrics at the CL layer. This suggests that our approach successfully incorporates the circumplex space information and utilizes it effectively to improve classification performance. In contrast, both the CSE and Nearest Mining approaches yield lower classification performances compared to the baseline model, which lacks concept information. This implies that the inclusion of concept information in these cases decreases the classification performance.

The CAV score for the DENSE and CLASS layers shows the utilization factor of concept information for the different models. Comparing these scores with the baseline provides valuable insights into how the network incorporates concepts for classification tasks, i.e. since the baseline network is not informed with concepts information its concepts utilization factor should be taken as lower bound or lower reference point.

Hence, we can define the utilization factor $CUF_{i_j}(\cdot)$ of the network i at layer j of the concept $(\cdot)$ as:

$$CUF_{i_j}(\cdot) = \frac{CAV_{i_j}(\cdot)}{CAV_{baseline_j}(\cdot)} \tag{1}$$

CUF is higher than 1 if the utilization of the concept is higher than the one exploited by the baseline model. Table II shows the concept utilization factors of the different models at the DENSE and CLASS levels.

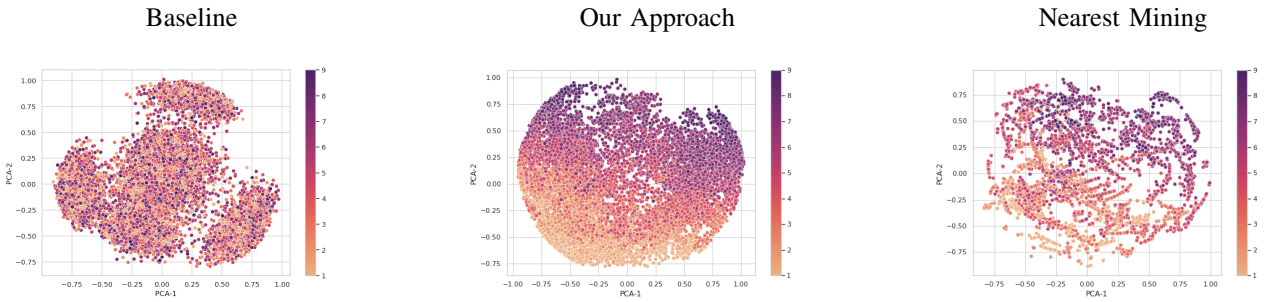

Fig. 2: Principal components plot of the contrastive layer of the baseline, our method, and nearest mining, for the Arousal concept. The colours indicate the arousal intensity from 1 to 9.

TABLE I: Average accuracy and CAV scores for different models at different levels

	Accuracy	CAV Arousal			CAV Valence			CAV Dominance		
	F1-Score	CL	DENSE	CLASS	CL	DENSE	CLASS	CL	DENSE	CLASS
Our Approach	**0.89**	0.94	0.85	0.68	0.95	0.91	0.89	0.94	0.83	0.68
Baseline	0.88	0.86	0.7	0.68	0.92	0.89	0.89	0.87	0.79	0.68
Nearest Mining	0.87	0.97	0.93	0.68	0.98	0.96	0.89	0.96	0.93	0.68
CSE	0.77	0.97	0.98	0.97	0.98	0.98	0.98	0.98	0.98	0.97

TABLE II: Concepts Utilization Factors for different models at different levels i.e. DENSE (D) and CLASS (C)

	Arousal		Valence		Dominance	
	D	C	D	C	D	C
Our Approach	1.2	1.0	1.0	1.0	1.1	1.0
Nearest Mining	1.3	1.0	1.1	1.0	1.2	1.0
CSE	1.4	1.4	1.1	1.0	1.2	1.4

As demonstrated by Table II and Table I, there exists a trade-off between concept utilization and classification performance. Analysing the CLASS layer, we can observe that the CUF of the CSE model consistently surpasses that of the other approaches. However, it achieves lower recognition performance in terms of F1-score. This suggests that the CSE model excessively focuses on concept information, leading to a decline in recognition performance. This trade-off is one of the primary limitations discussed by the authors when proposing the CSE model [11]. Similar considerations can be made when considering the DENSE layer and comparing the CSE model and the Nearest Mining model with our approach.

Our approach outperforms the other models in terms of concept utilization and classification performance. By comparing the utilization factors in Table II and the recognition scores in Table I, it is evident that our approach strikes a better balance between concept utilization and classification accuracy.

In the CLASS layer, our approach achieves comparable concept utilization factors compared to the CSE and Nearest Mining models. However, it consistently outperforms them in terms of recognition performance, as indicated by higher F1-scores. This indicates that our approach effectively leverages the concept information without compromising classification accuracy, unlike the other models.

Similarly, when considering the DENSE layer, our approach achieves comparable concept utilization factors while maintaining or surpassing their recognition scores. This highlights the effectiveness of our approach in effectively utilizing concept information at different levels of the network architecture.

Overall, our approach strikes a favourable balance between concept utilization and classification performance, making it a preferable choice over the other models. It maximizes the utilization of concept information while maintaining or even improving the recognition accuracy, providing a more robust and efficient solution for the task at hand.

Figure 2 presents a grid of 3 subfigures, each figure represents results obtained with a different model: the baseline model trained solely with the CCL loss, the model based on nearest mining, and the proposed model. Each subfigure displays a scatter plot of the first two principal components of the features extracted by the Contrastive Layer. The colours of the points in the plots correspond to the scale value for the Arousal concept. Clearly, the baseline model, lacking any conceptual information about the circumplex space, does not impose any order on the data, resulting in a random distribution for all classes. On the other hand, both approaches that leverage contrastive learning successfully provide the model with conceptual information. Our approach demonstrates a more homogeneous and coherent ordering, closely aligned with the circumplex space, as it prevents the formation of micro-clusters in the data, which instead can be found in the Nearest Mining plot. Also, our approach results in a smooth and progressive representation of the different grades of Arousal, Valence, and Dominance.

In summary, Figure 2 visually illustrates the effectiveness of our proposed approach and the comparative performance of the different models in leveraging conceptual information from the circumplex space. The plots clearly demonstrate that our approach successfully incorporates this information, resulting in a more consistent and ordered data representation. This provides valuable insights into the impact of our method on the representation and utilization of conceptual

information during the classification process.

IV. Conclusion

In this study, we have introduced a novel loss function aimed at incorporating domain knowledge, specifically the circumplex space-derived information, into neural networks for a multiclass categorical emotion recognition task. The proposed methodology empowers the network to extract emotional information related to arousal, valence, and dominance dimensions within its structure, enabling the recognition of the perceived emotional category by the subject and, thus, simulating human reasoning.

The proposed approach has demonstrated significant improvements in emotional recognition performance compared to all other tested methods. Moreover, it effectively learns the most optimal latent space representation that embeds information from the circumplex space, as evidenced by both the PCA components plots and the CAV scores associated with the concept distribution. Furthermore, this study introduces a new metric, the Concept Utilization Factor (CUF), which quantifies the extent to which concept-derived information is extracted by a concept-aware neural network within its layers, in comparison to a baseline model that lacks concepts-awareness. Our approach achieves the best trade-off between CUF and target-label recognition performance.

Overall, this research presents a promising direction for emotion recognition tasks by effectively leveraging domain knowledge and enriching the neural network's understanding of emotions through the circumplex space-derived information.

Moreover, it is essential to emphasize how the adoption of such a latent space ordering approach can significantly enhance explainable artificial intelligence (XAI) techniques applied to these types of neural networks. By incorporating domain knowledge and leveraging the circumplex space-derived information, our methodology not only improves emotion recognition performance but also enhances the interpretability of the model's decisions. The ability to trace emotional representations back to their underlying concepts opens new possibilities for understanding the neural network's decision-making process in emotional recognition tasks. We believe that this work lays the foundation for future studies in XAI, encouraging researchers to explore and refine explainability methods within the context of concept-aware neural networks.

References

[1] R. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: analysis of affective physiological state," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001.

[2] N. S. Suhaimi, J. Mountstephens, and J. Teo, "Eeg-based emotion recognition: A state-of-the-art review of current trends and opportunities," *Computational intelligence and neuroscience*, vol. 2020, 2020.

[3] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and psychopathology*, vol. 17, no. 3, pp. 715–734, 2005.

[4] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *American Scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[5] R. Abiri, S. Borhani, E. W. Sellers, Y. Jiang, and X. Zhao, "A comprehensive review of eeg-based brain–computer interface paradigms," *Journal of neural engineering*, vol. 16, no. 1, p. 011001, 2019.

[6] R. W. Homan, J. Herman, and P. Purdy, "Cerebral location of international 10–20 system electrode placement," *Electroencephalography and clinical neurophysiology*, vol. 66, no. 4, pp. 376–382, 1987.

[7] S. M. Alarcão and M. J. Fonseca, "Emotions recognition using eeg signals: A survey," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 374–393, 2019.

[8] G. Gagliardi, A. L. Alfeo, V. Catrambone, D. Candia-Rivera, M. G. C. A. Cimino, and G. Valenza, "Improving emotion recognition systems by exploiting the spatial information of eeg sensors," *IEEE Access*, vol. 11, pp. 39544–39554, 2023.

[9] G. Gagliardi, A. L. Alfeo, V. Catrambone, M. G. Cimino, M. De Vos, and G. Valenzal, "Fine-grained emotion recognition using brain-heart interplay measurements and explainable convolutional neural networks," in *2023 11th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 1–6, 2023.

[10] A. L. Alfeo, A. G. Zippo, V. Catrambone, M. G. Cimino, N. Toschi, and G. Valenza, "From local counterfactuals to global feature importance: efficient, robust, and model-agnostic explanations for brain connectivity networks," *Computer Methods and Programs in Biomedicine*, p. 107550, 2023.

[11] A. L. Alfeo, M. G. Cimino, and G. Gagliardi, "Concept-wise granular computing for explainable artificial intelligence," *Granular Computing*, vol. 8, no. 4, pp. 827–838, 2023.

[12] Y. Gao, S. Gu, J. Jiang, S. R. Hong, D. Yu, and L. Zhao, "Going beyond xai: A systematic survey for explanation-guided learning," *arXiv preprint arXiv:2212.03954*, 2022.

[13] I. C. Kaadoud, L. Fahed, and P. Lenca, "Explainable ai: a narrative review at the crossroad of knowledge discovery, knowledge representation and representation learning," in *MRC 2021: Twelfth International Workshop Modelling and Reasoning in Context*, vol. 2995, pp. 28–40, ceur-ws. org, 2021.

[14] D. Bau, J.-Y. Zhu, H. Strobelt, A. Lapedriza, B. Zhou, and A. Torralba, "Understanding the role of individual units in a deep neural network," *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30071–30078, 2020.

[15] N. Díaz-Rodríguez, A. Lamas, J. Sanchez, G. Franchi, I. Donadello, S. Tabik, D. Filliat, P. Cruz, R. Montes, and F. Herrera, "Explainable neural-symbolic learning (x-nesyl) methodology to fuse deep learning representations with expert knowledge graphs: The monumai cultural heritage use case," *Information Fusion*, vol. 79, pp. 58–83, 2022.

[16] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene cnns.," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[17] Z. Chen, Y. Bei, and C. Rudin, "Concept whitening for interpretable image recognition," *Nature Machine Intelligence*, vol. 2, no. 12, pp. 772–782, 2020.

[18] B. Zhou, D. Bau, A. Oliva, and A. Torralba, "Interpreting deep visual representations via network dissection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 9, pp. 2131–2145, 2018.

[19] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *International conference on machine learning*, pp. 2668–2677, PMLR, 2018.

[20] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim, "Towards automatic concept-based explanations," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[21] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept bottleneck models," in *International Conference on Machine Learning*, pp. 5338–5348, PMLR, 2020.

[22] K. Musgrave, S. Belongie, and S.-N. Lim, "Pytorch metric learning," 2020.

[23] M. Kaya and H. Ş. Bilge, "Deep metric learning: A survey," *Symmetry*, vol. 11, no. 9, p. 1066, 2019.

[24] S. Kim, M. Seo, I. Laptev, M. Cho, and S. Kwak, "Deep metric learning beyond binary supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2288–2297, 2019.