scientific reports

Check for updates

OPEN

EEG-based emotional valence and emotion regulation classification: a data-centric and explainable approach

Linda Fiorini[1,4], Francesco Bossi[1,2,4] & Francesco Di Gruttola[1,3]✉

Emotion classification using electroencephalographic (EEG) data is a challenging task in the field of Artificial Intelligence. While many researchers have focused on finding the best model or feature extraction technique to achieve optimal results, few have attempted to select the best methodological steps for working with the dataset. In this study, we applied two different theoretical approaches based on the noise of the dataset: curriculum learning and confident learning. Curriculum learning involves presenting training examples to the model in a specific order, starting with easier examples and gradually increasing in difficulty. This approach has been shown to improve model performance. Confident learning is a method for identifying and correcting label errors in datasets. By identifying and correcting these errors, confident learning can improve the performance of machine learning models trained on noisy datasets. We then applied the Integrated Gradient technique in order to assess the explainability of each model. Our aim was to explore the impact of different models and methods on emotion classification performance using EEG data. We collected and used an EEG dataset in which participants rated the emotional valence of positive and negative pictures while performing an emotion regulation (ER) task, comparing a control condition (Look) with two ER strategies: cognitive reappraisal and expressive suppression. We performed a multilabel classification to identify emotional neutrality or polarization of emotional valence (both positive and negative) rated by participants and the emotion regulation strategy adopted during the task. We compared the performance of models trained on three datasets selected based on label noise and evaluated their suitability for this task. Our results suggest different patterns based on the architecture used for feature importance, highlighting both advantages and criticisms.

Affective computing and EEG

Artificial intelligence (AI) has made significant progress in recent years, enabling the development of systems that can accurately perform tasks such as speech and image recognition. One area of AI that has received particular attention is emotion recognition, which aims to identify and classify human emotional states from various input modalities, giving birth to a new branch of AI, i.e., affective computing[1].

Electroencephalography (EEG), a neurophysiological technique that measures the spontaneous electrical activity of the brain, has proven to be useful in the field of emotion recognition[2] due to its sensitivity to emotional changes[3]. For instance, studying emotions through physiological signals can be particularly useful for individuals who have difficulty expressing emotions through facial expression or speech, such as people with traits in the autism spectrum[4]. Thus, a tool that can translate emotions into feedback that is understandable by therapists or parents could be extremely helpful.

However, classifying EEG data can be challenging due to several factors. EEG signal contains both actual brain activity and noise and artifacts[5]. Additionally, the EEG signal is non-linear[6] meaning that linear equations may have limited effectiveness in modelling it. EEG signal is also non-stationary[6], meaning that its statistical properties vary over time. This can make it difficult for models trained on temporally limited EEG data to generalize at different times or for different people. Finally, there is high inter-subject variability in the EEG signal, which can drastically affect the performance of a model when evaluating different subjects[7].

[1]Molecular Mind Laboratory (MoMiLab), IMT School for Advanced Studies Lucca, Lucca, Italy. [2]Department of Information Engineering, University of Pisa, Pisa, Italy. [3]Department of Psychology 'Renzo Canestrari', University of Bologna, Bologna, Italy. [4]Linda Fiorini and Francesco Bossi contributed equally to this work. ✉email: francesco.digruttola@gmail.com

The consequence of these features, taken together, is represented by the struggle to have good results with machine or deep learning techniques.

One common approach to simplify the input data is feature extraction. This involves extracting relevant features from the data, such as time domain features like event-related potentials (ERPs)[8] or power spectral density [9–11]. In one recent paper[12], the authors extracted five frequency bands and used a Random Forest (RF) model on those extracted features, achieving an accuracy of 70% in classifying emotional valence. Further research reached even higher accuracy in classifying emotional valence and arousal by exploiting different methods based on spectral features [13–16]. Although feature extraction can be a promising approach for simplifying EEG data, another more intricate but promising method involves using raw data as input to the model. In this case, the architecture and composition of the model itself are used to extract the most important features. In the following lines, we are going to introduce the state-of-the-art neural networks used in literature. Convolutional neural networks (CNN) are often used for this purpose[17] as they are capable of initially extracting both local, low-level features and global, high-level features from raw input data[18]. Results achieved when using this approach in a binary classification of emotional valence are often considered good when the accuracy is slightly below 70%[17], even with feature extraction techniques. Another promising approach is classifying emotional valence using the combination of CNN and Recurrent Neural Networks (RNN). RNNs have an internal state that allows them to learn from long-term dependencies and temporal patterns in the data[19], which is extremely useful in the case of EEG data. In recent studies[20–23], the authors used a particular type of RNN, a Long Short Term memory network (LSTM)[24], particularly efficient to avoid the vanishing gradient problem, which occurs when the gradients of the loss function become very small for the weights of the earlier layers, and handling long sequences[25]. With this method, researchers reached an accuracy above 70% for the emotional valence.

Another promising approach that has been scarcely used in emotion classification could be using a CNN + GRU (Gated Recurrent Unit) network to combine the advantages of both methods.

Recently, researchers have also begun to use Transformer neural networks[26] for EEG analysis. Originally developed for natural language processing[27], these models have also been adapted to time series as input data [28] and, more recently, EEG data. Indeed, transformer models can capture the global contextual information[26] of the data, which can be very useful considering the discriminative spatial information deriving from each single electrode in EEG data [29].

Both feature extraction and deep learning methods applied to raw signals have advantages and disadvantages. On the one hand, feature extraction results are usually more accurate, but the signal is extensively processed, making it difficult to use for a future Brain-Computer Interface (BCI) application. BCI could be a great asset for some individuals in particular cases, such as interventions for people with autism[30] that can express their emotions in a misleading way for the therapist[31]. On the other hand, using the raw signal produces less accurate results, but this method could be more useful in this context, in which signal preprocessing is not possible.

Beyond the simple emotion recognition

In this paper, we aim to employ emotion recognition within a unique context where participants were trying to modify their emotions. To the best of our knowledge, this is the first study exploring the field of emotion regulation. Specifically, we used an EEG emotion regulation (ER) task to classify the perceived emotional valence of emotional pictures observed by participants and the ER strategy they adopted. We asked participants to assess the emotional valence of 60 images. For each picture, participants were instructed to adopt one out of two possible ER strategies (and a control condition). We used a novel approach based on identifying the optimal methodology to classify EEG data after minimal preprocessing.

ER refers to the "extrinsic and intrinsic processes responsible for monitoring, evaluating, and modifying emotional reactions, especially their intensity and duration"[32]. It plays an essential role in everyone's life: many studies highlighted the correlation between healthy ER strategies and social and affective adaptation[33,34], how it affects decision-making[35] and coping with stress[36] or the severity of symptoms in conditions such as Post-traumatic stress disorder (PTSD)[37] or Attention deficit hyperactivity disorder (ADHD)[38]. Also, the typical state that characterizes mood and anxiety disorders often depends on emotion dysregulation[39].

Two of the most studied ER strategies are cognitive reappraisal and expressive suppression[40]. Cognitive reappraisal is an antecedent-focused ER strategy, which refers to the attempt to reinterpret a situation eliciting emotions in a way that changes its meaning and emotional impact[34,41]. Expressive suppression, on the other hand, is a response-focused strategy and can be defined as the attempt to hide, inhibit, or reduce ongoing emotion-expressive behaviour (such as facial expressions, verbal utterances and gestures)[42].

We included this manipulation in our study for two reasons. Firstly, this represents the first effort in the literature to employ an AI model for classifying ER strategies. Secondly, and more significantly, ER strategies are employed by individuals on a daily basis. Therefore, a tool that can discriminate these strategies may be exceedingly valuable in research and clinical settings. For instance, it could facilitate investigations about how and when individuals tend to regulate their emotions or about comprehending typical and atypical coping mechanisms.

Data-centric strategies in EEG emotion recognition: confident and curriculum learning

As discussed above, emotion classification is still a challenging task for machine learning models, and it could become even more difficult considering the attempt to classify the ER strategies as well. For this reason, we believe that the most appropriate strategy for these tasks is focusing on the data quality, especially considering the intrinsic noise of EEG data already discussed above.

Some researchers in AI already focused on the so-called data-centric approaches. One of the most frequently used is Confident Learning, useful to identify and correct label errors in any dataset using any model[43]. Confident learning can help models avoid learning from unreliable or inconsistent labels, which can degrade

their accuracy and generalization. In the same vein, we also used curriculum learning. This is a training strategy that orders data samples from easy to hard, based on several criteria such as label noise or the ambiguity of the examples given[43]. The concept of curriculum learning was first introduced by Bengio and colleagues in a data-centric framework[44], who proposed a method for training neural networks by presenting training examples in a predefined order of difficulty. This approach was inspired by the idea of curriculum development in education, which involves designing a sequence of learning experiences that gradually build on one another. Since then, curriculum learning has been extensively studied in the machine learning literature, with researchers exploring a variety of different strategies for selecting the order in which training examples are presented[45]. Curriculum learning can help models focus on simpler concepts first and gradually progress to more complex ones, without being overwhelmed by noise or ambiguity.

To date, while some researchers have incorporated curriculum learning in emotion recognition studies[46], no EEG studies within affective computing have yet explored a data-centric approach, including curriculum or confident learning. Thus, we want to explore this methodology. Indeed, emotion recognition, with its inherent complex nature and consequent non-convex optimization challenges, presents itself as a prime candidate for curriculum learning in the development of classification models. This aligns with the principle that ideal problems for curriculum learning should involve non-convex optimization[44].

Also, we suggest that even the categorisation of emotional valence clusters (e.g. positive, neutral and negative) should be based on a data-driven approach. Thus, a promising way to approach this problem is to study not only the EEG variation, but also the noisiness of the data and the label assigned. For this reason, we believe that confident learning can be very beneficial for our work. Thus, the first aim of our study is to use a data-centric approach to feed the proposed models the best possible data based on input EEG data.

Integrated gradients for explainable AI

As this is a methodological, data-centric study, our focus extends beyond mere performance metrics to also include the strategies employed by the models. Our goal is not only to identify an effective strategy for fitting the models but also to comprehend the underlying reasons and mechanics of what makes certain approaches succeed or fail. Consequently, the second aim of this study is to use explainable AI (XAI) to study how different models make classification decisions with respect to emotion and ER strategy classification of raw EEG data. With regard to XAI, we employed the Integrated Gradient (IG) approach[47] to study the deep learning models we trained. IG requires no modification to the original network (Model Agnostic Approach) and is extremely simple to implement. Its objective is to assign an attribution score that underlines how each input feature is positively or negatively related to each prediction. It is based on two axioms: (i) sensitivity: if one feature change makes the classification output change, then that feature should have a non-zero attribution; (ii) implementation invariance: the attribution method result should not depend on the specificities of the neural network architecture [47].

Aim and hypothesis

Our hypothesis is that adopting those two novel approaches (namely data-centric approach and XAI) for emotion recognition while employing different emotion regulation strategies could lead to less biased networks and to understand how the predictions are made. Curriculum learning could help models focus on simpler concepts first and gradually progress to more complex ones, without being overwhelmed by noise or ambiguity. Confident learning could help models avoid learning from unreliable or inconsistent labels, which can degrade their accuracy and generalization. IG could enhance the understanding of how the predictions are made by different models.

In this study, we compared four different architectures: CNN, CNN + unidirectional RNN, CNN + bidirectional RNN, and Transformer. In summary, in this paper, we propose a novel approach that combines curriculum learning, confident learning and IG for emotion classification using different models while adopting two ER strategies, i.e., expressive suppression and cognitive reappraisal, and a neutral strategy that implies just looking at the stimuli.

Results

Behavioural results

In Fig. 1, participants' average responses are displayed. Table 1 shows the results of the mixed-effect linear model performed on the emotional valence ratings reported by participants. In particular, the estimated marginal means (emmean), standard errors (SE), asymptotic lower and upper confidence limits (asymp.LCL, asymp.UCL), z-ratio, and p-value for each of the comparisons are shown. The mixed-effects linear model on participants' ratings showed a significant interaction effect ($\chi^2 = 1106.7$, df = 40, $p = < 0.001$) between the mindset and the emotional valence. Post-hoc multiple comparisons (corrected according to Tukey's HSD) based on the different mindsets showed that looking at negative images was associated with lower (i.e., more negative) ratings than reappraising or suppressing them (see Table 1, statistically significant effects are in bold), while there was no significant difference between reappraising and suppressing negative images. On the other hand, looking at positive images was associated with higher emotional valence than reappraising or suppressing them, while there was no significant difference between reappraising and suppressing positive images. These results showed the overall effectiveness of the two emotional regulation strategies in modulating the emotional impact of the images.

Confident learning

The quality of data labels was evaluated using Random Forest (RF) as a benchmark model for the confident learning approach. Separate RF models were trained for emotional valence and ER strategy as target classes. All results reported refer to the dataset after applying the SMOTE algorithm, a data augmentation technique.

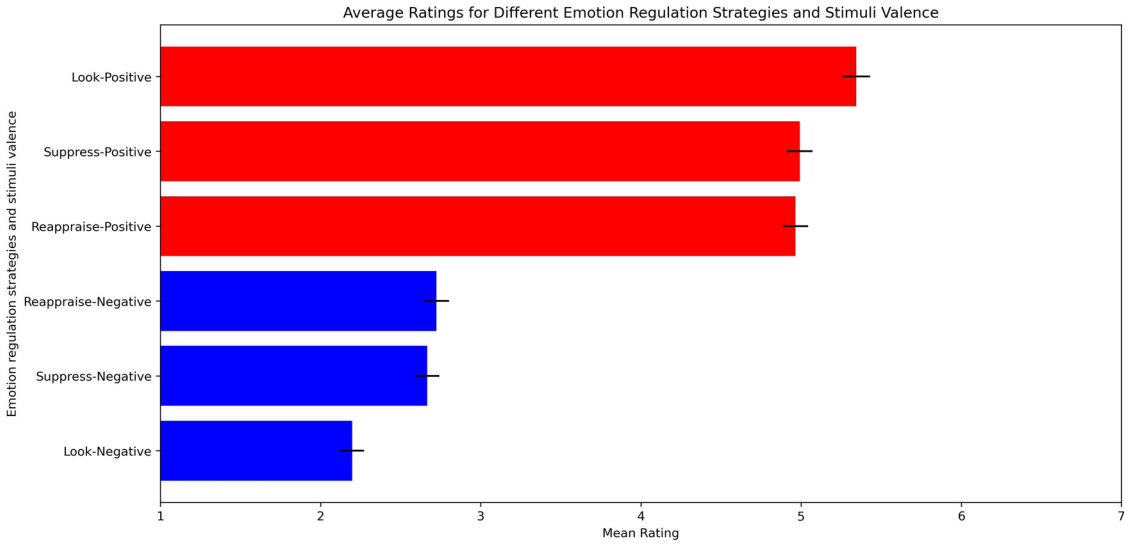

Fig. 1. Average and standard deviation of emotional valence ratings in the different ER strategies. In blue, pictures assessed a priori as negative, in red those that were considered positive.

CONTRAST	Marginal means differences	Standard Error	asymp. LCL	asymp. UCL	z-ratio	p-value
LOOK N—REAPPRAISE N	−0.527	0.106	−0.849	−0.204	−4.965	**< 0.0001**
LOOK N—SUPPRESS N	−0.469	0.118	−0.826	−0.112	−3.990	**0.0009**
REAPPRAISE N—SUPPRESS N	0.058	0.179	−0.486	0.602	0.323	0.9995
LOOK P—REAPPRAISE P	0.379	0.129	−0.014	0.771	2.932	**0.0395**
LOOK P—SUPPRESS P	0.352	0.087	0.088	0.616	4.052	**0.0007**
REAPPRAISE P—SUPPRESS P	−0.027	0.138	−0.446	0.392	−0.193	1

Table 1. Results of the post-hoc tests performed on the mindset * emotional valence two-way interaction effect in the mixed-effect linear model on the emotional valence ratings reported by participants.

	Precision	Recall	F1-score	N.examples
1	0.11	0.14	0.12	70
2	0.11	0.15	0.13	89
3	0.15	0.15	0.15	117
4	0.27	0.14	0.19	183
5	0.18	0.17	0.18	118
6	0.10	0.11	0.10	66
7	0.06	0.12	0.08	32
accuracy			0.14	675
macro avg	0.14	0.14	0.13	675
weighted avg	0.17	0.14	0.15	675

Table 2. Classification report on the validation set of random forest applied to the emotional valence of the pictures with seven classes (one for each possible rating, from 1 to 7).

Emotional valence

When classifying participants' ratings (from 1 to 7) in seven classes, the model performed with an F1-score of 0.15 (random chance = 0.14). (see Table 2). The labels that were best predicted were 4 and 5, with an F1-score of 0.19 and 0.18, followed by label 3 with an F1-score of 0.15. The labels that were least accurately predicted were the most positive ones, 6 and 7, with an F1-score of 0.10 and 0.08.

In order to simplify the classification problem and to balance the sample of each class, we then classified the emotional valence dividing it into three classes, i.e., negative (i.e., labels 1–2), neutral (3–4-5) and positive (6–7). The results reported in Table 3 of the random forest on the validation set show that the class with a higher F1-

	Precision	Recall	F1-score	N.examples
Negative	0.26	0.20	0.23	159
Neutral	0.65	0.62	0.64	418
Positive	0.17	0.26	0.20	98
accuracy			0.47	675
macro avg	0.36	0.36	0.36	675
weighted avg	0.49	0.47	0.48	675

Table 3. Classification report on the validation set of the Random Forest applied to emotional valence of the pictures, with three classes (Negative, Neutral and Positive).

	Precision	Recall	F1-score	N.examples
Polarised	0.40	0.38	0.39	257
Neutral	0.63	0.65	0.64	418
accuracy			0.55	675
macro avg	0.51	0.51	0.51	675
weighted avg	0.54	0.55	0.54	675

Table 4. Classification report of the validation set of random forest applied to the emotional valence of the pictures with two classes (Neutral and Polarised).

	Precision	Recall	F1-score	N.examples
Look	0.33	0.32	0.33	225
Reappraise	0.33	0.39	0.36	229
Suppress	0.33	0.28	0.30	221
accuracy			0.33	675
macro avg	0.33	0.33	0.33	675
weighted avg	0.33	0.33	0.33	675

Table 5. Classification report of the validation set of the Random Forest applied to ER Strategies.

score was the neutral one, with an F1-score of 0.64 and a general accuracy of 0.48. This may be the result of the unbalanced dataset, having a very represented neutral class and two other classes with fewer examples.

Given the unbalanced dataset, we then tried to classify data into two classes by merging positive (i.e., 6 and 7) and negative samples (i.e., 1 and 2) in a "polarized" class. Table 4 reports the results of a binary classifier that showed a more balanced result, having an F1-score of 0.39 for the polarised class and 0.64 for the neutral one. Even the precision was more balanced, having respectively 0.40 and 0.63 for the polarised and neutral classes.

Emotion regulation

For what concerns the ER strategy classification, results can be observed in Table 5. The overall accuracy is 0.33. The precision and F1-score were also similar for all classes, indicating that the model had a balanced performance across the classes. The model was slightly better at predicting the class Reappraise, which had the highest F1-score of 0.36. The model was slightly worse at predicting the class Suppress, which had the lowest F1-score of 0.30. The N.examples column shows that the validation set was balanced, with equal amounts of data for each class.

At this point, we began our analysis by training deep neural networks with different architectures using a curriculum learning approach. With this aim, we divided the dataset into three parts based on label noise computed with Cleanlab and for each deep learning architecture we created three deep learning models, trained, separately, with the easy part of the dataset, the easy and medium parts combined, and on the entire dataset. We then used an ensemble learning approach to combine the predictions of these three models to obtain the final result.

Convolutional neural networks

Table 6 shows the classification report for the MINI-VGG (Visual Geometry Group), a convolutive neural network with a shallow architecture.

When classifying valence and ER strategy the model obtained a weighted average F1-score of 0.42 on the test set. The model shows better performance in classifying emotional valence, with the highest F1-score for the Neutral class at 0.55, followed by the Polarised class with an F1-score of 0.51. The model performs poorly in

	Precision	Recall	F1-score	N.examples
Neutral	0.60	0.50	0.55	384
Polarised	0.44	0.60	0.51	232
Look	0.34	0.40	0.36	302
Reappraise	0.30	0.24	0.27	228
Suppress	0.27	0.32	0.29	226
weighted avg	0.42	0.43	0.42	1372

Table 6. Classification report on the test set of MINI-VGG.

	Precision	Recall	F1-score	N.examples
Neutral	0.54	0.61	0.57	384
Polarised	0.46	0.38	0.41	302
Look	0.30	0.32	0.31	232
Reappraise	0.33	0.48	0.39	228
Suppress	0.32	0.49	0.39	226
weighted avg	0.41	0.47	0.43	1372

Table 7. Classification report on test set of MINI-VGG with GRU.

	Precision	Recall	F1-score	N.examples
Neutral	0.57	0.34	0.43	384
Polarised	0.44	0.58	0.50	302
Look	0.32	0.70	0.44	232
Reappraise	0.35	0.41	0.38	228
Suppress	0.31	0.41	0.36	226
weighted avg	0.42	0.48	0.43	1372

Table 8. Classification report on test set of MINI-VGG with bidirectional GRU.

recognizing ER strategies. Indeed, the class Look is the best classified, with an F1-score of 0.36, while Reappraise and Suppress classes showed an F1-score of 0.27 and 0.29 respectively.

Convolutional neural networks + recurrent neural networks

Table 7 shows the results of the MINI-VGG with GRU layers. The model showed a weighted average F1-score of 0.43. Once again, the highest F1-score is for the Neutral class at 0.57, while the Polarised class showed an F1-score of 0.41. The ER classes totalized an F1-score of 0.31, 0.39 and 0.39 for Look, Reappraise and Suppress respectively.

Table 8 shows results of the MINI-VGG with bidirectional GRU layers. The model showed a weighted average F1-score of 0.43. In this case, the best performance concerned the Polarised class, having 0.50 points of F1-score, followed by Neutral with 0.43 points. The ER classes showed a 0.30, 0.43 and 0.42 F1-score for Look, Reappraise and Suppress respectively.

The ER strategies performed better than every other model tested, having an F1-score of 0.44, 0.38 and 0.36 for Look, Reappraise and Suppress, respectively.

Transformers

Table 9 shows the results of the Transformer architecture. The model showed a weighted average F1-score of 0.28. In this case, the best performance regards the Polarised class, having 0.56 points of F1-score, followed by neutral with 0.53 points. Each of the ER classes showed an F1-score of 0.02.

Explainability (Integrated gradient)

In Fig. 2, we reported an example of a matrix of the IG attribution score for each EEG feature (channels x time) in every deep learning model considered. All figures for other classes and classifications (i.e., True Negative, False Positive, False Negative) are reported in the supplementary materials. We find that each model showed a similar pattern of explainability for every predicted class. The CNN architecture reported a scattered and fuzzy attribution importance pattern across all features. Similarly, for both GRU and bidirectional GRU models, a scattered layout limited to the initial or final time parts of the data was underlined. On the other hand, the Transformer architecture evidenced a defined pattern, with some channels showing a more stable and consistent influence over time.

	Precision	Recall	F1-score	N.examples
Neutral	0.57	0.50	0.53	384
Polarised	0.46	0.70	0.56	302
Look	0.33	0.01	0.02	232
Reappraise	0.23	0.01	0.02	228
Suppress	0.20	0.01	0.02	226
weighted avg	0.39	0.30	0.28	1372

Table 9. Classification report on the test set of Transformer model.

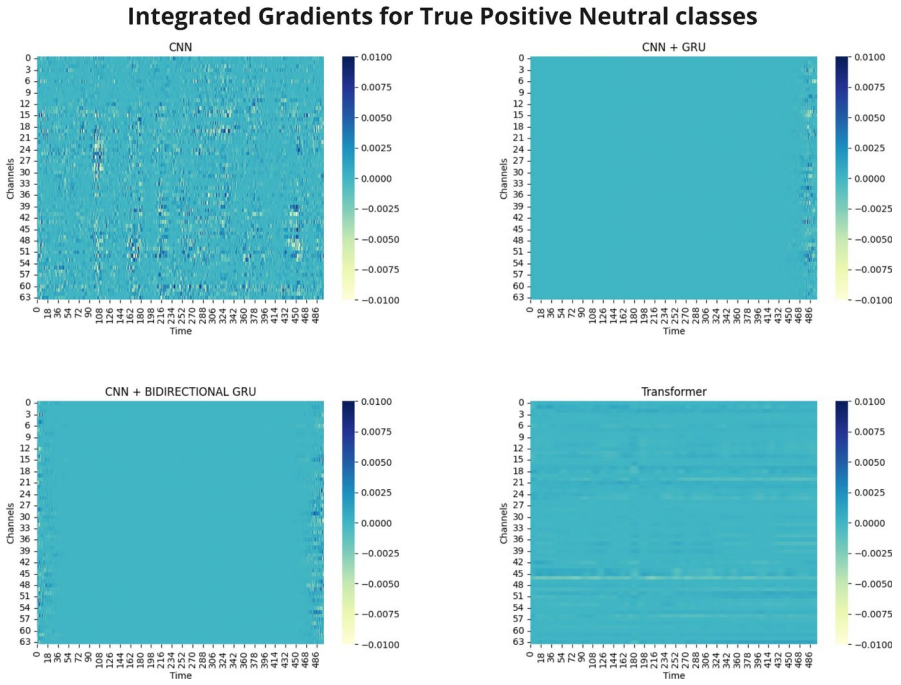

Fig. 2. Graphic representation of IG attribution score for each model tested. Here we report as an example the condition when each model correctly classified the Neutral class (True Positive). The attribution value represents the relation intensity between each feature (time x channel) and the target class. The colour scale ranges from dark blue to yellow, representing positive and negative attribution scores (feature importance) respectively. All figures for other classes and classifications (i.e., True Negative, False Positive, False Negative) are reported in the supplementary materials.

To test the observed differences in the IG data on a quantitative basis, we employed four random-intercept linear mixed-effect models (LMMs). The results of these analyses are summarized in Table 10 and the data are represented in Fig. 3.

In the CNN model, post-hoc comparisons performed on the Cluster main effect showed that the central cluster presented significantly higher IG scores than the posterior cluster ($z = 8.18$, $p < 0.001$), which in turn presented greater values than the anterior cluster ($z = 2.7$, $p = 0.019$). Comparisons based on time showed that, independently of different channel clusters, 150-200 ms was the time bin with the highest IG value, as it was significantly higher than 0-50 ms, 50-100 ms, 100-150 ms, and 250-300 ms time bins (all zs > 3.7, all ps < 0.01). The 450-500 ms time bin also showed greater IG values than the 50-100 ms time bin ($z = 3.9$, $p = 0.004$).

The unidirectional GRU showed the greatest importance of the central cluster again, as it showed higher values than both the anterior ($z = 4.7$, $p < 0.001$) and posterior ($z = 2.7$, $p = 0.021$) clusters, which, in reverse, did not differ significantly ($z = 2.0$, $p = 0.111$). The time main effect, with the highest effect size, showed that 400-450 ms (all zs > 5.7, all ps < 0.001) and 450-500 ms time bins (all zs > 220, all ps < 0.001) presented significantly higher values than any other time bins. The post-hoc comparisons performed on the interaction effect showed that differences between clusters were present only in the 450-500 ms time bin, with significantly higher values in the central cluster than in the posterior one ($z = 8.2$, $p < 0.001$), which in turn was higher than the anterior one ($z = 6.2$, $p < 0.001$).

The bidirectional GRU showed the same pattern across clusters as the unidirectional GRU, with the central cluster showing higher values than both the anterior ($z = 5.9$, $p < 0.001$) and posterior ($z = 5.4$, $p < 0.001$) clusters, which, in reverse, did not differ significantly ($z = 0.5$, $p = 0.875$). The time main effect showed that 0-50 ms

Effect	Num DF	Den DF	F	p-value	p-value code	f^2 marginal	f^2 conditional
CNN							
Cluster	2	20547	64.1	< .001	***	5.70E-03	6.20E-03
Time	9	20547	4.64	< .001	***	1.80E-03	1.90E-03
Cluster*Time	18	20547	0.39	0.99		3.10E-04	2.60E-04
GRU-unidirectional							
Cluster	2	20547	11	< .001	***	9.70E-04	9.80E-04
Time	9	20547	10421	< .001	***	4.5	4.5
Cluster * Time	18	20547	10.4	< .001	***	8.30E-03	8.40E-03
GRU-bidirectional							
Cluster	2	20547	21.4	< .001	***	2.00E-03	2.00E-03
Time	9	20547	10216	< .001	***	4.4	4.4
Cluster * Time	18	20547	10.5	< .001	***	8.30E-03	8.50E-03
Transformer							
Cluster	2	20547	529	< .001	***	5.00E-02	5.10E-02
Time	9	20547	1.35	0.204		5.50E-04	5.60E-04
Cluster * Time	18	20547	0.31	0.998		2.20E-04	2.10E-04

Table 10. Integrated Gradients statistical comparisons.

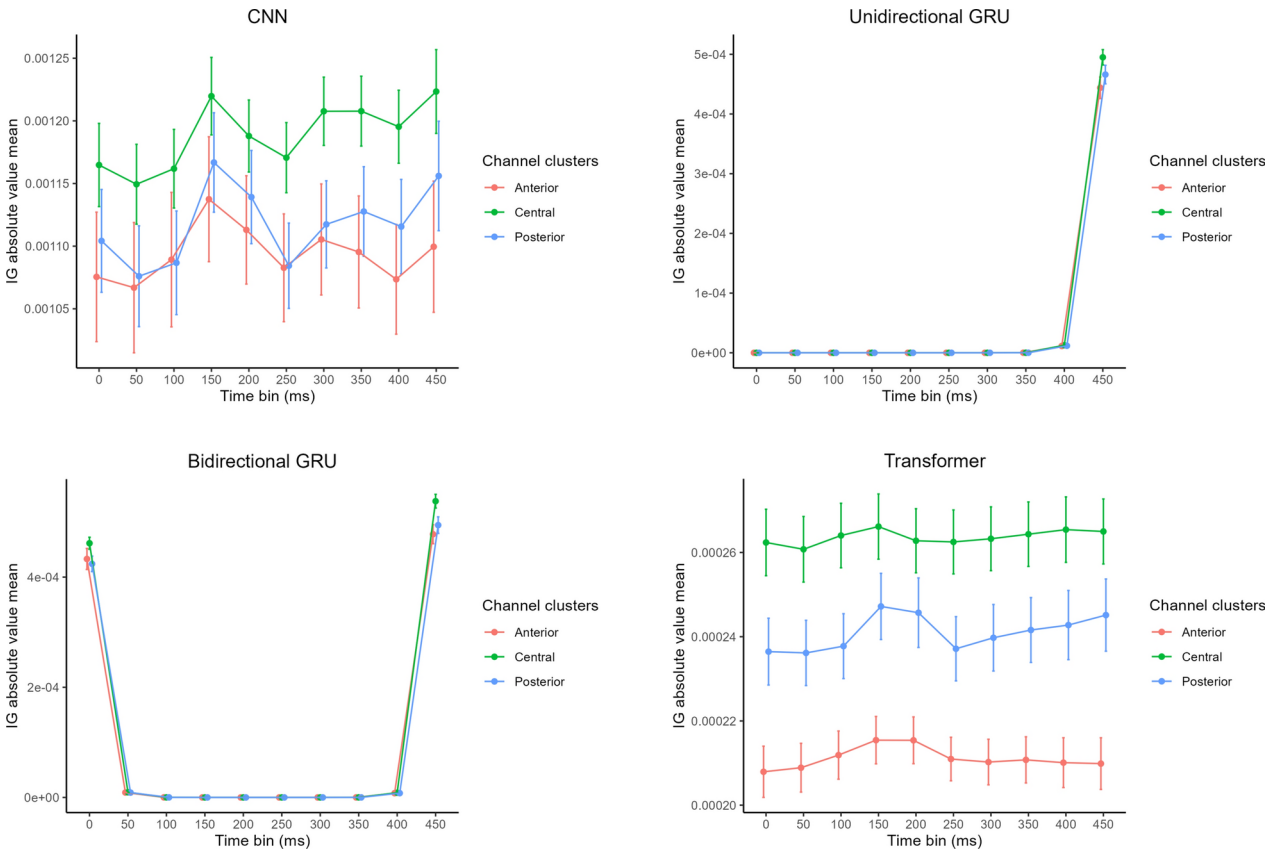

Fig. 3. Graphic representation of results of the statistical comparisons of the IG attribution score for each model tested. Time bins of 50 ms are represented on the x-axis, while the mean of the IG absolute value is represented on the y-axis. The three colored lines represent three channel clusters (i.e., anterior, central and posterior). Error bars represent 95% confidence intervals.

(all $zs > 155$, all $ps < 0.001$), 50-100 ms (all $zs > 3.2$, all $ps < 0.045$), and 450-500 ms time bins (all $zs > 178$, all $ps < 0.001$) presented significantly higher values than any other time bins. Concerning the interaction effect, the comparisons based on time showed that, when considered separately for each channel cluster, only the 0–50 ms (all $zs > 86$, all $ps < 0.001$) and 450-500 ms time bins (all $zs > 97$, all $ps < 0.001$) showed significantly higher values than any other time bins. The comparisons based on the cluster showed that, in the 0-50 ms time bin, the central cluster presented greater IG values than both the posterior ($z = 7.8$, $p < 0.001$) and anterior clusters ($z = 5.9$, $p < 0.001$), while these latter two did not differ significantly between each other ($z = 1.9$, $p = 0.154$); in the 450-500 ms time bin, the central cluster presented greater IG values than the posterior one ($z = 9.0$, $p < 0.001$), which in turn was significantly greater than the anterior one ($z = 3.4$, $p = 0.002$).

The only statistically significant effect in the Transformer model was the Cluster main effect. This showed that the central cluster presented greater IG values than the anterior ($z = 32.4$, $p < 0.001$) and posterior clusters ($z = 14.0$, $p < 0.001$), the latest being also significantly higher than the anterior one ($z = 18.4$, $p < 0.001$).

Discussion

In this study, we employed a multilabel classification approach to classify the emotional valence of visual stimuli and emotion regulation strategies utilized by participants while looking at those visual stimuli, based on participants' EEG signals. Our study aimed to provide a methodological framework to handle raw EEG data and provide reliable results in this context. Therefore, the first aim was to prepare the dataset for the deep learning models training via a data-centric approach. To reach this objective, we combined curriculum learning[44] and confident learning [43] techniques. Subsequently, in the second aim we compared how different deep learning models took classifying decisions by using an XAI technique, namely IG [47].

First of all, we selected the proper labels based on the confident learning approach, thus resulting in two classes for emotional valence classification (i.e., neutral vs. polarised) and three classes for ER strategy (i.e., Look, Reappraise and Suppress).

To adopt the curriculum learning approach, we used an ensemble approach with three models having the same architecture. The only change was the training dataset, giving the three different models different training based on the dataset's difficulties.

For the multilabel classification with 5 classes and 6 different combinations (i.e., 2 valence * 3 ER strategy classes) we chose four different models: a MINI-VGG, a MINI-VGG with GRU, a MINI-VGG with bidirectional GRU and a Transformer. The MINI-VGG[48] is a simplified version of the VGG network[49], which is a Convolutional Neural Network (CNN) designed to process data -in this case EEG data- by applying multiple convolutional layers to extract spatial features. The MINI-VGG with GRU model combines the MINI-VGG with a Gated Recurrent Unit (GRU), which is a type of Recurrent Neural Network (RNN) that processes sequential data by maintaining hidden states to capture temporal dependencies. The MINI-VGG with bidirectional GRU model further enhances the GRU by using bidirectional GRUs, which process the input sequence in both forward and backward directions to capture context from both past and future time steps. The Transformer model utilizes self-attention mechanisms to handle sequential data by allowing each position in the sequence to attend to all other positions, effectively capturing long-range dependencies without relying on recurrent structures. In general, the best classes' performance was neutral and polarised. Regarding the ER strategies, the models still cannot predict those strategies properly.

Our results are interesting even considering that combining CNN with RNN, especially uni- and bidirectional GRU, is quite an unexplored approach in the domain of raw EEG data classification [50].

Our intuition was that the neurons of the long-term memory recurrent neural network have the benefit of memorizing both the long-term and short-term emotional information present in the EEG signal, facilitating the recognition of emotions.

Considering the architectures' interpretability, (i.e., studying the attributions of the gradients of each model) we realised that RNN tends to memorize shorter sequences and that an array with 64×500 time points is too long to be learned.

In the statistical comparisons it can be seen that unidirectional GRU tend to consider only the last 100 ms of the input, while the bidirectional GRU only the first and last 100 ms. Further studies combining Convolutional and Recurrent architectures should therefore test shorter time series compared to our dataset (e.g., 50–100 ms). The CNNs, instead, seem to consider the whole data, but at the same time, the pattern of the importance of the feature is scattered, and it makes the interpretation very difficult. It looks like there are no specific channels or time points to influence the outcome of the classification. Based on IG, the Transformer is the model showing the most interpretable patterns. This model showed patterns of specific channels that can be positively or negatively correlated with the prediction of the model over time. For the Transformer model, the IG pattern is constant over time, thus showing that this architecture can learn from the whole epoch, unlike the GRU models. It is also interesting to notice that the Transformer model has the best score for the emotional valence domain and the worst in the ER strategy one.

Regarding the interpretability of the results, we interpreted the findings from our statistical comparisons (Fig. 3). In particular, differences between clusters were found consistently in all models. However, we can only interpret these differences in a speculative way given that we performed all analyses at sensor level. We found that central sensors were the most important in classifying polarised vs. neutral emotional valence, followed by posterior and anterior ones. It is well known that the insula and cingulate cortex (CC) are closely interconnected structures within a mesolimbic network, which is essential for generating and perceiving the motor and autonomic changes that occur during emotional experiences[51]. As affective experience intensifies, activity in the mesolimbic network also increases [52–55]. The signal from at least part of these regions can be identified in EEG channels in this central cluster. Moreover, neuroscience literature considers the Late Positive Potential (LPP) as one of the most important ERP associated with emotion processing and it is observed in the

time interval from 400 to 500 ms averaged over sensors in centro-parietal regions[56] having its source in the CC[57]. Therefore, these findings provide information about the validity of the models used. It appears that they focused on the electrodes showing the activity of areas most commonly involved in emotional experience, based on neuroscientific research.

We believe that our work is valuable for both research and applied purposes for two main reasons. Firstly, while most research in the literature focuses primarily on finding the best model by evaluating classic performance measures[50] (e.g., F1-score, accuracy), we trust that research should focus on data first. In this vein, we have tested a robust methodological solution based on a data-centric approach for dealing with raw EEG data, which can be particularly challenging[6] since artifacts in the signal can affect deep learning models.

Secondly, classic performance measures, which are widely used in the literature to evaluate deep learning models, do not reveal how these black box models make decisions. For this reason, we believe that it is always appropriate to adopt an XAI technique to turn models into explainable white boxes, rather than solely discussing their performance. This approach is valuable not only in choosing the best model for a classification problem but also in refining the architecture accordingly. For instance, in using the CNN + GRU architecture, we observed a temporal limit in the time series memorization only thanks to the XAI technique. Each step (i.e., data-centric approach and XAI) in our methodological framework could be useful to researchers and practitioners to study a specific dataset in other fields and build a personalised model that could be used for a real-time BCI. Also, it should be considered that there are many reasons for this result, despite the strict methodological and theoretical background of our work.

Indeed, choosing to divide the dataset into polarised and neutral is, from our point of view, an interesting approach. As shown in Fig. 1, several ratings for every condition range between 3 and 5, so the dataset itself hinted at studying the neutral values. Besides that, our model may be better suited for a particular kind of emotion recognition, meaning that we do not aim to recognize positive or negative emotions as many studies do[17,58,59].

Those findings suggest that even after considering possible issues with label noise and the best training for the models, there is still work to do to improve those deep learning models, especially for the emotion regulation strategies adopted. By the way, the performance of emotional valence classification is above the chance level, especially in Transformer models, when considering that data were processed only for filtering the signal and rejecting bad epochs to approximate a possible BCI approach and that we were using a multilabel classifier[60].

The interesting advantage of this approach is that it would be even more useful in everyday life, since a great part of the stimuli do not necessarily have strong emotional valence: everyday objects carry subtle affective valences, defined by some authors[61] as "micro-valences", which are intrinsic to their perceptual representations.

Besides, the novel aspect of this study is also the classification of the ER strategy, which has never been investigated in the field of artificial intelligence. However, despite all methodological checks, classification performance is poor in ER strategies: there are several reasons for these results. We first need to point out that we did not choose the train, validation and test set by randomly selecting epochs from all the participants, but we split the data based on random subjects. Kamrud[62] and colleagues have mathematically shown that cross-participant models, where samples are randomly taken from any subject, tend to have underestimated error rates between 35 and 3900%, thus overestimating the model's performance. This can explain some apparently good results in the domain of raw EEG emotion classification. This is related to several reasons: first, EEG varies across participants due to non-stationarity and individual differences[63]. Thus, splitting the data based on subjects ensures that the training, validation, and test sets are independent and reduces the risk of overfitting. Randomly selecting epochs from all participants implies that data from the same subject could appear in both the training and test sets. This could lead to biased performance, as the model may learn to recognize specific characteristics of individual subjects rather than generalizing to new data. Second, splitting the data based on subjects allows us to evaluate the generalizability of our model to new subjects. By training our model on data from one set of subjects and testing it on data from a different set of subjects, we can assess how well our model can classify emotional valence in new individuals.

Still, the main limitation of this work is that our models do not outperform the current state-of-the-art, but we have adopted a challenging and specific dataset (EEG raw data), without any real benchmark in model performance. Since each example is potentially affected by artifacts, it is reasonable to obtain lower performance compared to a pre-processed EEG dataset. However, we obtained fairly good performance in emotional valence classification and this research line is potentially more useful in creating models classifying real-time data with minimal preprocessing (i.e., BCI).

Future directions could consist of applying and comparing the effectiveness of this data-centric and XAI approach to different classification problems using EEG raw data. For example, this work could be a seminal step in finding solutions when dealing with large artifacts such as movement.

Additionally, future studies might explore applying these data-centric techniques to different kinds of data within the field of affective computing. While curriculum learning has been experimented with in speech emotion recognition[46], to the best of our knowledge the use of confident learning to assess the difficulty of examples remains unexplored. We suggest that integrating these methods could enhance model performance, particularly with data types where curriculum learning has already shown promising results. Moreover, using a dataset with more observations, or trying transfer learning from another dataset could increase the models' performance. We also want to point out that we did not make use of synthetic (i.e., simulated) data, which could lead to risks such as bias in data diversity [64].

We also think that the state-of-the-art literature may show biased results considering the problem of cross-participant training, validation, and test split. Indeed, out of six EEG deep learning models currently used in research, five are cross-participant models[50]. However, only one of these five models follows proper dataset partitioning methods to ensure that the model is tested on data from participants it has not evaluated before[50].

It must also be considered that a model's performance can surpass the state-of-the-art literature. However, if there is no way to understand why or what it is classifying, the model cannot be deemed reliable. This is especially true in fields like healthcare, representing a possible application for BCIs, where decisions have significant consequences. Interpretable models allow us to gain insights into the model's decision-making process, detect potential biases, and ensure that the model aligns with our expectations and requirements.

For these reasons, we believe that, despite the improvable performance, this work represents an important advance in the field of emotion classification using EEG, allowing a new methodological approach to disentangle the common issues of label noise and overestimated performance.

Methods

Participants

We collected data from thirty paid participants (17 females, mean age 26 ± 6.12). Our exclusion criteria comprised the presence of a history of any neurological or psychiatric disease, use of active drugs, abuse of any drugs (including nicotine within 2 h preceding the study and alcohol within 24 h preceding the study), as well as being informed about the aim of the study. Written informed consent was obtained from all participants according to the declaration of Helsinki; the IMT Ethical Committee approved the project.

Materials

We selected sixty stimuli from the Oasis database[65]. In this database, 900 images are present, rated by 822 participants. Every image was rated via a Likert scale from 1 to 7 according to their valence and arousal.

Our aim was to have a small dataset with positive emotional valence for half of the stimuli and negative for the other half. We selected images that respected three criteria: (i) choosing images with no sensitive contents (i.e., sexual or violent scenes) based on the indications of the Ethical Committee; (ii) selecting images with high arousal values (> 4) to clearly elicit emotions; (iii) including images with the highest or lowest validation valence values to use strongly polarized positive and negative stimuli.

Overall, we included 60 stimuli, i.e., 30 positive images with mean $\pm$ sd valence and arousal respectively of 5.7 ± 1.06 and 4.7 ± 1.6, and 30 negative images with mean $\pm$ sd valence and arousal values of 2.1 ± 1.1 and 4.7 ± 1.8, respectively.

In order to check that the stimuli had the same valence polarization for our participants, we compared valence scores for all images between the validation dataset and scores assigned by participants in this study. Even if 21 images showed a significant difference between the means in independent samples t-tests, it is also shown in Fig. 4 that none of these stimuli showed an inversion of positive vs. negative valence, i.e., positive stimuli with mean valence ratings above 4 (=neutral valence) in the validation dataset showed mean valence ratings above 4 also in our dataset, and vice-versa for negative stimuli below 4.

Procedure

After signing the informed consent module, participants were asked to sit on a comfortable chair in an electrically shielded and soundproof room while not crossing their legs or arms. A 19″ monitor was positioned in front of them, 1 mt distant.

The EEG cap was then prepared on the participant's head. Resting state activity was then recorded for five minutes, but it will not be used in this study. Experimenters explained to the subjects the experimental procedure

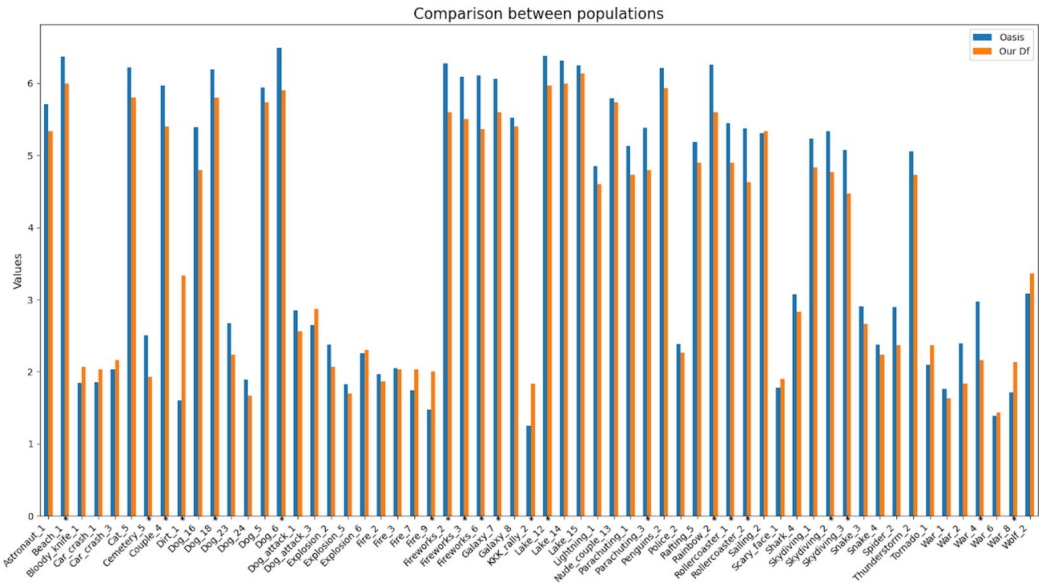

Fig. 4. The mean valence of each stimulus assessed by our subjects during the experiment is in red compared with the mean valence rated in the Oasis dataset, in blue.

before the experiment started. Moreover, written instructions were shown on the monitor to give them the information they needed.

The experimental design is shown in Fig. 5. The experiment consisted of three blocks in which participants were asked to assess the emotional valence of 60 pictures shown on the monitor while adopting three different emotion regulation strategies: 'Look', 'Reappraise' and 'Suppress', for a total of 180 stimuli seen during the experiment.

Each mindset was explained by a short training by experimenters and with written instructions at the beginning of each of the three blocks, which were shown in a counterbalanced order. Participants were free to ask the experimenters information about the different ER strategies at any moment via a microphone.

Once participants understood what mindset they had to adopt, they could press a key to begin the visualization and rating of the stimuli. The mindset was also reminded during the whole experimental block by showing a small text (i.e., 'LOOK', 'REAPPRAISE' and 'SUPPRESS') above the pictures.

Each stimulus, which was shown for 1000 ms, was preceded by a fixation cross of 1000 ms. After the stimulus visualization, participants were asked to rate the emotional valence of the picture, according to the mindset they adopted in that block, on a Likert scale ranging from 1 (absolutely negative) to 7 (absolutely positive). The response was self-paced.

The dataset generated during the current study is available in the Emotion Regulation Task repository, https://osf.io/yv468/.

EEG recording and preprocessing

The neurophysiological activity of each participant was recorded with a 64-channel EGI EEG system. Electrodes were positioned according to the 10–20 International System. The online reference was Cz and the sampling frequency was 1000 Hz with an impedance below 50KΩ.

All the pre-processing was performed using the MNE-Python package.

The raw signal was filtered (1–48 Hz bandpass with a 50 Hz notch filter). We used the RANSAC (random sample consensus) method[66] to detect bad channels. The signal was then epoched: each epoch began 500 ms before the stimulus onset lasting 1000 ms. We then normalized the epoched signal considering a baseline from 500 to 200 ms before the stimulus onset and cropped the signal before the appearance of the stimulus. At this point, each epoch lasted 500 ms.

We chose to use shorter epochs for our algorithm to be precise for real-time emotion recognition. Other studies in the literature used shorter window size[67] and neuroscience literature offers evidence about early neural correlates linked to emotions[68]. The total was 180 epochs for each subject. We used the Autoreject algorithm[69] to identify and reject bad epochs. We set the consensus from 0.2 to 0.5 in steps of 0.1. The consensus corresponds to the proportion of bad channels that are allowed in order to accept the epoch. If the number of bad channels allowed exceeds the real number of bad channels, the algorithm rejects the epoch, otherwise, it can interpolate the signal from 1 to 8 bad electrodes. The algorithm chooses the best value possible for each subject. The average number of rejected epochs was 9.16 ± 9.97 per participant. Then the signal was re-referenced offline to the average of all channels.

Data preprocessing

To prepare the data for analysis, we randomly assigned data from 22 participants to the training set, 4 participants to the validation set, and 4 to the test set. We also applied Principal Component Analysis (PCA)[70] to reduce the dimensionality of the training set and balance the classes. We selected 100 principal components that explained 87% of the total variance. We verified that the inverse PCA matrix was highly correlated with the original matrix (i.e., distance correlation coefficient[71] = 0.99). We then used SMOTE[72](Synthetic Minority Over-sampling

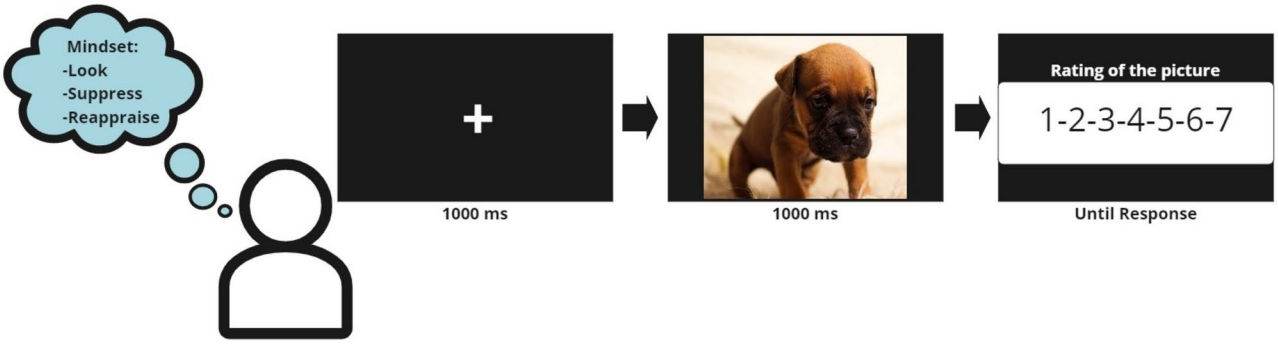

Fig. 5. Experimental Procedure: After being instructed about the ER strategies and their meaning, participants were asked to adopt two possible ER strategies (i.e., expressive suppression or cognitive reappraisal), or a control condition (i.e., "Look") in three separate counterbalanced blocks. In each of the 60 trials per block, a fixation cross was displayed on the monitor for 1000 ms, followed by either a positive or a negative stimulus for 1000 ms. Participants then rated the emotional valence of the stimulus on a Likert scale from 1 (absolutely negative) to 7 (absolutely positive) with their mouse.

Technique) to augment the data in the training set based on the principal components and reconstructed the original matrix shape by applying inverse PCA.

SMOTE works by selecting examples that are close to the feature space. For each instance in the minority class, SMOTE calculates the k-nearest neighbors. It then randomly selects one or more of these nearest neighbors and generates synthetic examples by interpolating between the selected neighbor and the original instance. In this function, the k parameter has to be set because, when a random example from the minority class is first chosen, k nearest neighbors are selected in order to decide where the algorithm can draw a line between the examples in the feature space and generating a new sample at a point along that line. We chose k = 5 because, in the literature, it is a frequently used default value[73].

After preprocessing the data, we included 64 channels and a time window of 500 ms. With a sample frequency of 1000 Hz, the shape of each example was (1, 64, 500).

Data analysis

Behavioural

Statistical behavioral tests were conducted using R[74]. We adopted the standard 0.05 Alpha significance level to test against the null hypotheses. In particular, we used Linear Mixed-Effects Models (lmer) package[75]. We fitted a linear mixed-effects model testing how the rating of the images differs between mindsets (baseline, reappraisal or suppression) and emotional valence assessed a priori (negative or positive) while accounting for the variation across subjects and file images. The model has three fixed effects (mindset, emotional valence of the stimulus, and their two-way interaction) and two random effects (subject and file image). We then used Estimated Marginal Means (emmean) package[76] to calculate the contrasts between each condition and correct for multiple comparisons using the Tukey method[77].

Data-centric approach

In order to adopt a data-centric approach with confident and curriculum Learning, we used random forest (RF) algorithms.

A random forest (RF) is an ensemble learning method that consists of a collection of decision trees: it is called a "forest" because it is made up of many decision trees[78].

A decision tree's fundamental goal is to classify or predict data by iteratively splitting the data based on the values of particular attributes[79]. For instance, in this dataset, the random forest could split the data for each electrode or time point. The model incorporates randomness by randomly choosing the splitter and each tree from various random subsamples of data. Every node in the tree indicates a split in the data, and the leaf node is where the ultimate prediction is formed.

Random forests are often used in the literature because they reduce overfitting since every decision tree is trained on different subsets of the data and then the predictions are averaged together. We used this algorithm as a benchmark for deep learning to calculate the sanity of the labels (according to the confident learning approach) for several reasons: (i) as mentioned above, the model is very good at reducing overfitting; (ii) the random forest is one of the few algorithms that calculates the class weights, and it usually works properly with non-linear data such as EEG. In particular, we used the Cleanlab package. Cleanlab is an open-source framework for machine learning and analytics with noisy data. It provides methods to identify, quantify, and correct errors in datasets, measure and track dataset quality, and train reliable models with noisy labels. Cleanlab is based on rigorous theoretical foundations based on the work of Northcutt and colleagues[43].

To create different datasets based on both the difficulty of the samples (curriculum learning) and the correct classification classes (confident learning), we trained two kinds of random forests (RFs). One type of RF was used to classify the emotional valence of the images and the other to classify the ER strategies.

We used three criteria to prepare our dataset in the most methodologically sound way: (i) ensuring that the dataset was as balanced as possible, (ii) making theoretical assumptions about both emotional valence and emotion regulation strategy, and (iii) simplifying the classification problem.

To achieve this, we trained the first RF for emotional valence with 7 classes, corresponding to the 7-point Likert scale used for assessment. The second RF had 3 classes (positive, neutral, and negative) and the third RF had just two classes (neutral and polarised). The other type of RF was used to classify the ER strategy of the participants into 3 classes, corresponding to the 3 emotion regulation strategies. We then calculated the class overlap for each model to determine the best labels for our dataset for multilabel classification of both emotional valence and emotion regulation. We first selected the best label match for emotional valence and ER strategy based on the above-mentioned criteria. Thus, we decided to use a binary classification for emotional valence (polarized and neutral), while maintaining the three original classes for emotion regulation strategy. Then, we measured the label quality for each example and averaged it with emotional valence and mindset. We divided the dataset into three parts based on the 33rd and 66th percentiles of the distribution: easy, intermediate, and difficult.

Deep learning architectures

In this study, we used four different ensemble deep learning architectures for our multilabel classification with five classes (Neutral, Polarised, Look, Reappraise, Suppress): (1) MINI-VGG, (2) MINI-VGG with GRU, (3) MINI-VGG with bidirectional GRU and (4) A Transformer model.

Multilabel classification involves N non-exclusive labels (in our case, N = 5). Each label is considered as a binary classification problem whose predicted probability is independent with respect to the other classes.

This multilabel classification approach allows us to avoid using three separate models (one for each ER strategy) and, therefore, training our data on emotional valence on a dataset split into three parts. Additionally,

given our experimental protocol, it was necessary for two labels to be present simultaneously (i.e., one label for emotional valence and one for ER strategy).

To make predictions, we used an ensemble majority voting classifier and trained three different models based on the difficulty of the training data. Each model was trained using the principles of curriculum learning. We started by training one model on easy examples, another one on easy and medium examples, and the third one trained on the entire dataset. We used this approach, as Bengio suggested[44], because easy data are more easily learned by the model. Indeed, by training three models on datasets of increasing difficulty, we aimed to provide the ensemble models with a broader view of EEG signals, avoiding biased predictions.

In an ensemble majority voting classifier, the prediction with the highest number of votes was selected, thus a prediction was confirmed if at least two out of the three classifiers voted for it. We used a Sigmoid function in the last layer of each model, thus obtaining in output a prediction for each label independently. For each prediction, we set a threshold of 0.5 to classify the presence (i.e., probability > 0.5) or absence of the variable.

Regarding the specific deep learning architectures, we used a variant with 1D CNN of the MINI-VGG of Kranthi and colleagues [48] because some previous works[50] demonstrated that shallow deep learning architectures achieve better performance on EEG data as features for classification tasks. The MINI-VGG we used consisted of two 1D convolutional blocks followed by a classification block. The convolutional blocks were composed of two convolutional layers, each with ReLU activation. In the first block, the convolutional layers used 32 filters with a kernel size of 3, while in the second block, 64 filters were used with a kernel size of 3. At the end of each block a pooling layer performed max pooling operation over a window of 2. There was always a batch normalization after the convolutional one. After each pooling a dropout layer was used. The dropout value was taken as 0.15. Finally, the output was passed through a Dense layer with sigmoid activation function to produce the final predictions.

With regards to the MINI-VGG + GRU and MINI-VGG + bidirectional GRU, we refer to the modification of the MINI-VGG by including three GRU (unidirectional or bidirectional) layers with 128 units and a *tanh* activation function after the second convolutional block. Instead of dropout, we used Monte Carlo Dropout (MCDropout). The output of the GRU blocks was then passed through a Batch Normalization layer and another MCDropout with a dropout percentage of 0.15. Finally, the output was passed through a Dense layer with sigmoid activation function to produce the final predictions.

For the Transformer, the model consisted of two encoders and a classification block. Each encoder block applies layer normalization and multi-head attention with 64 head size and 2 heads to the inputs, followed by a dropout with a rate of 0.4 and a normalization. Then, two convolutional layers with 64 filters with a kernel with size of 1 are applied. Only the first convolutional layer used a ReLU activation function. Between the two convolutional layers, a dropout with a dropout percentage of 0.15 was applied. The output of the final transformer block is then passed through a global average pooling layer and a series of dense layers with 128 units, ReLU activation, and a dropout rate of 0.4. The final output is produced by a dense layer with 5 units and a sigmoid activation function.

For every model, we used Adam optimizer[80] with a learning rate of 1e-06.

For a graphical representation of each architecture employed, see Supplementary Materials.

We trained the model with a default of 20,000 epochs, but then we also used the early stopping callback from Keras[81] (with patience of 20 epochs), taking as a benchmark the loss functions of the validation set.

Convolutional neural networks

Convolutional neural networks (CNN), which are a type of neural network used for multidimensional data[82]. The main component of the CNN are the convolutional layers.

A convolutional layer applies a set of filters (also called kernels or weights) to the input data. Every filter is a matrix applied to the data to produce a new feature map of the local region of the data to which has been applied. The process of applying a filter to the input data is called convolution. The input data is usually a multidimensional array. The filters are also multi-dimensional arrays smaller in size than the input data[83].

The convolutional layer applies each filter to the input data by sliding it over the input data and performing an element-wise multiplication with the values at each position. It then sums the results and stores the result in a new feature map. This process is repeated for each filter, producing multiple feature maps.

The output of the convolutional layer is a multi-dimensional array called a feature map, which has the same number of dimensions as the input data (e.g., height, width, and depth) but the size of the feature map is usually smaller than the input data because the filters are smaller and do not overlap.

The convolutional layer may also have parameters called stride and padding, which control how filters are applied to the input data. Stride controls the step size with which the filters are applied, and padding controls how the input data is padded with additional values around the border before the filters are applied.

One of the main issues with convolutional layers is that the feature map generated by the filter is dependent on its location. As a result, during training, convolutional neural networks learn to associate the presence of specific features with their location in the input features.

To avoid this problem the pooling layers are generally used in the convolutional networks[84]. The CNN we are going to use is a variant with 1D of the MINI-VGG that derives from the VGG that is a convolutional neural network architecture that was introduced by the Visual Geometry Group (VGG) at the University of Oxford in 2014 [49].

Recurrent neural networks

Recurrent neural networks (RNNs) are a type of artificial neural network that can handle sequential data or time series data, such as speech, text, or video[85]. Unlike feedforward neural networks, which assume that the input and output data are independent of each other, RNNs have feedback connections that allow them to store and

reuse information from previous computations. This means that RNNs have connections between their hidden units that form loops so that the output of a unit can influence its own input in the next time step[19,85]. This allows RNNs to store information from previous inputs in their hidden state, which can be used to process the current input. This gives RNNs the ability to learn from long-term dependencies and temporal patterns in the data. However, RNNs also face some challenges, such as vanishing or exploding gradients and difficulty in handling long sequences[86]. To address these issues, several variants of RNNs have been proposed, such as long short-term memory (LSTM)[87] or gated recurrent unit (GRU)[88]. They both use a gating mechanism to control the flow of information and avoid the vanishing gradient problem that affects standard RNNs. The main difference between them is that GRU has two gates (reset and update) while LSTM has three gates (input, output and forget). GRU is simpler and faster than LSTM, but LSTM has more flexibility and accuracy on larger datasets[89].

Also, GRUs can be unidirectional or bidirectional: a unidirectional GRU processes the input sequence in one direction, while a bidirectional GRU processes the input sequence in both forward and backward directions. This allows the model to capture both past and future context when making predictions.

Transformer networks

The Transformer network[26] is a type of neural network that utilizes a self-attention mechanism to achieve high-quality results and lower computational requirements for language translation tasks compared to recurrent and convolutional models. This mechanism allows the model to directly compare all parts of the input, regardless of their position, and assign an attention score to each part based on its relevance to the current task. The Transformer network is composed of multiple encoders and decoders. The encoders are stacked on top of each other in the model, with each encoder containing two sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. Residual connections are present around both sub-layers, followed by a normalization layer. Since we are working on time series, the decoder is not used in this case, as in other cases in literature[90].

Unlike recurrent neural networks (RNNs) and LSTMs, Transformer networks do not have an inherent way to capture the relative positions of the input elements. To provide this contextual information, positional encoding is used in conjunction with each input vector. Positional encoding is not part of the model architecture itself but rather a pre-processing step. A positional encoding vector is generated for each input element and added to its corresponding embedding vector. This allows the model to learn spatial information from the "injected" pattern in the embedding vector.

Integrated gradients

In our implementation, we used the Integrated Gradients (IG) method to explain the predictions of our model. IG is an interpretability technique method originally proposed in Sundararajan [47] for deep neural networks that visualizes the importance of input features in relation to the model's predictions. In the domain of XAI (Explainable Artificial Intelligence), interpretability emphasizes comprehending the internal mechanisms of models, thus understanding and detailing their internal functions, whereas explainability centers on clarifying the decision-making process. As a result, interpretability requires a higher degree of granularity compared to explainability[91].

In particular, integrated gradients define an attribution value for each feature by considering the integral of the gradients taken along a straight path from a baseline instance x' to the input instance x. Since we are using classifiers, the gradient usually refers to the output corresponding to the true class or to the class predicted by the model.

We used Alibi[92], which is an open-source library in Python. We used a null baseline and the number of steps we used was 25.

We used the "explain" method of this instance to compute the attributions for our selected examples, passing in the examples, baselines, and target as arguments. The attributions were then obtained from the "attributions" attribute of the returned explanation object.

To explain the IG results on a quantitative basis, we tested the differences between channels and timepoints employing statistical inferential models. To reduce the probability of Type I errors, we clustered the gradient values of each trial in the test set in 10 time bins of 50 ms each and all EEG channels in 3 clusters (i.e., anterior – 23 channels, central – 23 channels, and posterior – 18 channels). The absolute value of the IG score was averaged across each of these 30 clusters (i.e., 10 time bins × 3 channel clusters). We then compared these clusters using four random-intercept linear mixed-effects models[75] (LMMs) (i.e., one model for the CNN, one for the unidirectional GRU, one for the bidirectional GRU and one for the Transformer). Each model included the main effects of time and channels and their interaction as fixed factors. When encountering a statistically significant fixed effect, it was probed through post-hoc comparisons with Tukey's correction for multiple comparisons. For each effect, the f^2 effect size index was computed, according to the following formula:

$$f^2 = \frac{R^2_{AB} - R^2_A}{1 - R^2_{AB}}$$

where R^2_{AB} represents the coefficient of determination of the full model (i.e., with the effect of interest) and R^2_A represents the coefficient of determination of the null model (i.e., without the effect of interest). LMMs present two possible methods to compute R^2, i.e., marginal R^2 (without the contribution of random effects) and conditional R^2 (with the contribution of random effects). Therefore, two effect sizes were computed for each effect, i.e., marginal f^2 and conditional f^2.

To test how our methodological framework can be extended to different datasets, we replicated the same methodology using the DEAP dataset[93]. This is a comprehensive multimodal dataset designed for the analysis of

human affective states, validated and used in previous literature. Detailed results and discussion can be found in the Supplementary Materials (https://osf.io/yv468/).

Data availability

The dataset generated during the current study is available in the Emotion Regulation Task repository, https://osf.io/yv468/.

Received: 16 November 2023; Accepted: 3 October 2024
Published online: 14 October 2024

References

1. Picard, R. W. *Affective Computing*. (MIT press, 2000).
2. Bos, D. O. EEG-based emotion recognition. *The influence of visual and auditory stimuli* **56**, 1–17 (2006).
3. Apicella, A., Arpaia, P., Mastrati, G. & Moccaldi, N. EEG-based detection of emotional valence towards a reproducible measurement of emotions. *Sci. Rep.* **11**, 21615 (2021).
4. Yirmiya, N., Kasari, C., Sigman, M. & Mundy, P. Facial expressions of affect in autistic, mentally retarded and normal children. *J. Child Psychol. Psychiatry* **30**, 725–735 (1989).
5. Lai, C. Q. et al. Artifacts and noise removal for electroencephalogram (EEG): A literature review. in *2018 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)* 326–332 (IEEE, 2018).
6. Klonowski, W. Everything you wanted to ask about EEG but were afraid to get the right answer. *Nonlinear biomedical physics* **3**, 1–5 (2009).
7. Clerc, M., Bougrain, L. & Lotte, F. *Brain-Computer Interfaces 1: Methods and Perspectives*. (John Wiley & Sons, 2016).
8. Jenke, R., Peer, A. & Buss, M. Feature extraction and selection for emotion recognition from EEG. *IEEE Trans. Affect. Comput.* **5**, 327–339 (2014).
9. Frantzidis, C. A. et al. Toward emotion aware computing: an integrated approach using multichannel neurophysiological recordings and affective visual stimuli. *IEEE Trans. Inf Technol. Biomed.* **14**, 589–597 (2010).
10. Ding, R., Li, P., Wang, W. & Luo, W. Emotion processing by ERP combined with development and plasticity. *Neural plasticity* (2017).
11. Atkinson, J. & Campos, D. Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers. *Expert Syst. Appl.* **47**, 35–41 (2016).
12. Torres, E. P., Torres, E. A., Hernández-Álvarez, M. & Yoo, S. G. Emotion recognition related to stock trading using machine learning algorithms with feature selection. *Ieee Access* **8**, 199719–199732 (2020).
13. Galvão, F., Alarcão, S. M. & Fonseca, M. J. Predicting exact valence and arousal values from EEG. *Sensors* **21**, 3414 (2021).
14. Liu, Z.-T. et al. Electroencephalogram emotion recognition based on empirical mode decomposition and optimal feature selection. *IEEE Transactions on Cognitive and Developmental Systems* **11**, 517–526 (2018).
15. Liu, J. et al. Emotion detection from EEG recordings based on supervised and unsupervised dimension reduction. *Concurrency and Computation: Practice and Experience* **30**, e4446 (2018).
16. Yin, Z., Wang, Y., Liu, L., Zhang, W. & Zhang, J. Cross-subject EEG feature selection for emotion recognition using transfer recursive feature elimination. *Frontiers in neurorobotics* **11**, 19 (2017).
17. Hu, J. et al. ScalingNet: extracting features from raw EEG data for emotion recognition. *Neurocomputing* **463**, 177–184 (2021).
18. Albawi, S., Mohammed, T. A. & Al-Zawi, S. Understanding of a convolutional neural network. in *2017 international conference on engineering and technology (ICET)* 1–6 (Ieee, 2017).
19. Medsker, L. R. & Jain, L. C. *Recurrent neural networks. Design and Applications* **5**, 64–67 (2001).
20. Nakisa, B., Rastgoo, M. N., Rakotonirainy, A., Maire, F. & Chandran, V. Automatic emotion recognition using temporal multimodal deep learning. *IEEE Access* **8**, 225463–225474 (2020).
21. Kim, B. H. & Jo, S. Deep physiological affect network for the recognition of human emotions. *IEEE Transactions on Affective Computing* **11**, 230–243 (2018).
22. Kang, J.-S., Kavuri, S. & Lee, M. ICA-evolution based data augmentation with ensemble deep neural networks using time and frequency kernels for emotion recognition from EEG-data. *IEEE Transactions on Affective Computing* **13**, 616–627 (2019).
23. Li, Y., Huang, J., Zhou, H. & Zhong, N. Human emotion recognition with electroencephalographic multidimensional features by hybrid deep neural networks. *Applied Sciences* **7**, 1060 (2017).
24. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural computation* **9**, 1735–1780 (1997).
25. Graves, A. & Graves, A. Long short-term memory. *Supervised sequence labelling with recurrent neural networks* 37–45 (2012).
26. Vaswani, A. et al. Attention is all you need. *Advances in neural information processing systems* **30**, (2017).
27. Chernyavskiy, A., Ilvovsky, D. & Nakov, P. Transformers:"the end of history" for natural language processing? in *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21* 677–693 (Springer, 2021).
28. Liu, M. et al. Gated transformer networks for multivariate time series classification. *arXiv preprint* arXiv:2103.14438 (2021).
29. Wei, Y. et al. TC-Net: A Transformer Capsule Network for EEG-based emotion recognition. *Computers in Biology and Medicine* **152**, 106463 (2023).
30. Fan, J. et al. A Step towards EEG-based brain computer interface for autism intervention. in *2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC)* 3767–3770 (IEEE, 2015).
31. Eack, S. M., Mazefsky, C. A. & Minshew, N. J. Misinterpretation of facial expressions of emotion in verbal adults with autism spectrum disorder. *Autism* **19**, 308–315 (2015).
32. Thompson, R. A. Emotion regulation: A theme in search of definition. *Monographs of the society for research in child development* 25–52 (1994).
33. Gross, J. J. Emotion regulation in adulthood: Timing is everything. *Current directions in psychological science* **10**, 214–219 (2001).
34. John, O. P. & Gross, J. J. Healthy and unhealthy emotion regulation: Personality processes, individual differences, and life span development. *Journal of personality* **72**, 1301–1334 (2004).
35. Heilman, R. M., Crişan, L. G., Houser, D., Miclea, M. & Miu, A. C. Emotion regulation and decision making under risk and uncertainty. *Emotion* **10**, 257 (2010).
36. Wang, M. & Saudino, K. J. Emotion regulation and stress. *Journal of Adult Development* **18**, 95–103 (2011).
37. Ehring, T. & Quack, D. Emotion regulation difficulties in trauma survivors: The role of trauma type and PTSD symptom severity. *Behavior therapy* **41**, 587–598 (2010).
38. Seymour, K. E. et al. Emotion regulation mediates the relationship between ADHD and depressive symptoms in youth. *Journal of abnormal child psychology* **40**, 595–606 (2012).
39. Thompson, R. A., Lewis, M. D. & Calkins, S. D. Reassessing emotion regulation. *Child Development Perspectives* **2**, 124–131 (2008).
40. Cutuli, D. Cognitive reappraisal and expressive suppression strategies role in the emotion regulation: an overview on their modulatory effects and neural correlates. *Frontiers in systems neuroscience* 175 (2014).

41. Lazarus, R. S. & Alfert, E. Short-circuiting of threat by experimentally altering cognitive appraisal. *The Journal of Abnormal and Social Psychology* **69**, 195 (1964).
42. Gross, J. J. & Levenson, R. W. Emotional suppression: physiology, self-report, and expressive behavior. *Journal of personality and social psychology* **64**, 970 (1993).
43. Northcutt, C., Jiang, L. & Chuang, I. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research* **70**, 1373–1411 (2021).
44. Bengio, Y., Louradour, J., Collobert, R. & Weston, J. Curriculum learning. in *Proceedings of the 26th annual international conference on machine learning* 41–48 (2009).
45. Soviany, P., Ionescu, R. T., Rota, P. & Sebe, N. Curriculum learning: A survey. *International Journal of Computer Vision* **130**, 1526–1565 (2022).
46. Lotfian, R. & Busso, C. Curriculum learning for speech emotion recognition from crowdsourced labels. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **27**, 815–826 (2019).
47. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. in *International conference on machine learning* 3319–3328 (PMLR, 2017).
48. Kranthi Kumar, K., Bharadwaj, R., Ch, S. & Sujana, S. Effective deep learning approach based on VGG-mini architecture for iris recognition. *Annals of the Romanian Society for Cell Biology* 4718–4726 (2021).
49. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint* arXiv:1409.1556 (2014).
50. Roy, Y. et al. Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering* **16**, 051001 (2019).
51. Gu, X., Hof, P. R., Friston, K. J. & Fan, J. Anterior insular cortex and emotional awareness. *Journal of Comparative Neurology* **521**, 3371–3388 (2013).
52. Zerna, J., Strobel, A. & Scheffel, C. EEG microstate analysis of emotion regulation reveals no sequential processing of valence and emotional arousal. *Scientific Reports* **11**, 21277 (2021).
53. Bijanzadeh, M. et al. Decoding naturalistic affective behaviour from spectro-spatial features in multiday human iEEG. *Nature human behaviour* **6**, 823–836 (2022).
54. Ochsner, K. N., Silvers, J. A. & Buhle, J. T. Functional imaging studies of emotion regulation: a synthetic review and evolving model of the cognitive control of emotion. *Annals of the New York Academy of Sciences* **1251**, (2012).
55. Barrett, L. F., Mesquita, B., Ochsner, K. N. & Gross, J. J. The Experience of Emotion. *Annu. Rev. Psychol.* **58**, 373–403 (2007).
56. Schupp, H. T., Junghöfer, M., Weike, A. I. & Hamm, A. O. The selective processing of briefly presented affective pictures: An ERP analysis. *Psychophysiology* **41**, 441–449 (2004).
57. Calbi, M. et al. How context influences the interpretation of facial expressions: a source localization high-density EEG study on the "Kuleshov effect". *Scientific reports* **9**, 2107 (2019).
58. Kasuga, Y., Shin, J., Hasan, M. A. M., Okuyama, Y. & Tomioka, Y. EEG-based Positive-Negative Emotion Classification Using Machine Learning Techniques. in *2021 IEEE 14th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC)* 135–139 (IEEE, 2021).
59. Stikic, M., Johnson, R. R., Tan, V. & Berka, C. EEG-based classification of positive and negative affective states. *Brain-Computer Interfaces* **1**, 99–112 (2014).
60. Aggarwal, S. & Chugh, N. Review of machine learning techniques for EEG based brain computer interface. *Archives of Computational Methods in Engineering* 1–20 (2022).
61. Lebrecht, S., Bar, M., Barrett, L. F. & Tarr, M. J. Micro-valences: perceiving affective valence in everyday objects. *Frontiers in psychology* **3**, 107 (2012).
62. Kamrud, A., Borghetti, B. & Schubert Kabban, C. The effects of individual differences, non-stationarity, and the importance of data partitioning decisions for training and testing of EEG cross-participant models. *Sensors* **21**, 3225 (2021).
63. Cohen, M. X. *Analyzing Neural Time Series Data: Theory and Practice*. (MIT press, 2014).
64. Whitney, C. D. & Norman, J. Real Risks of Fake Data: Synthetic Data, Diversity-Washing and Consent Circumvention. in *The 2024 ACM Conference on Fairness, Accountability, and Transparency* 1733–1744 (2024).
65. Kurdi, B., Lozano, S. & Banaji, M. R. Introducing the open affective standardized image set (OASIS). *Behavior research methods* **49**, 457–470 (2017).
66. Derpanis, K. G. Overview of the RANSAC Algorithm. *Image Rochester NY* **4**, 2–3 (2010).
67. Xiangkun Yu , Zhengjie Li, Zhibang Zang, Yinhua Lin. Real-Time EEG-Based Emotion Recognition. *MDPI, Sensors* (2023).
68. Olofsson, J. K., Nordin, S., Sequeira, H. & Polich, J. Affective picture processing: an integrative review of ERP findings. *Biological psychology* **77**, 247–265 (2008).
69. Jas, M., Engemann, D. A., Bekhti, Y., Raimondo, F. & Gramfort, A. Autoreject: Automated artifact rejection for MEG and EEG data. *NeuroImage* **159**, 417–429 (2017).
70. Wold, S., Esbensen, K. & Geladi, P. Principal component analysis. *Chemometrics and intelligent laboratory systems* **2**, 37–52 (1987).
71. Edelmann, D., Móri, T. F. & Székely, G. J. On relationships between the Pearson and the distance correlation coefficients. *Statistics & probability letters* **169**, 108960 (2021).
72. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* **16**, 321–357 (2002).
73. Bennin, K. E., Keung, J. W. & Monden, A. On the relative value of data resampling approaches for software defect prediction. *Empirical Software Engineering* **24**, 602–636 (2019).
74. R Core Team, R. R: A language and environment for statistical computing. (2013).
75. Bates, D. Fitting linear mixed models in R. *R news* **5**, 27–30 (2005).
76. Lenth, R., Singmann, H., Love, J., Buerkner, P. & Herve, M. *Package 'Emmeans'*. (2019).
77. Keselman, H. J. & Rogan, J. C. The Tukey multiple comparison test: 1953–1976. *Psychological Bulletin* **84**, 1050 (1977).
78. Breiman, L. *Random forests. Machine learning* **45**, 5–32 (2001).
79. Safavian, S. R. & Landgrebe, D. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics* **21**, 660–674 (1991).
80. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint* arXiv:1412.6980 (2014).
81. Gulli, A. & Pal, S. *Deep Learning with Keras*. (Packt Publishing Ltd, 2017).
82. O'Shea, K. & Nash, R. An introduction to convolutional neural networks. *arXiv preprint* arXiv:1511.08458 (2015).
83. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 2818–2826 (2016).
84. Bailer, C., Habtegebrial, T. & Stricker, D. Fast feature extraction with CNNs with pooling layers. *arXiv preprint* arXiv:1805.03096 (2018).
85. Schuster, M. & Paliwal, K. K. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing* **45**, 2673–2681 (1997).
86. Bengio, Y., Simard, P. & Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks* **5**, 157–166 (1994).
87. Yu, Y., Si, X., Hu, C. & Zhang, J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural computation* **31**, 1235–1270 (2019).

88. Cho, K. et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint* arXiv:1406.1078 (2014).
89. Mateus, B. C., Mendes, M., Farinha, J. T., Assis, R. & Cardoso, A. M. Comparing LSTM and GRU models to predict the condition of a pulp paper press. *Energies* **14**, 6958 (2021).
90. Siddhad, G., Gupta, A., Dogra, D. P. & Roy, P. P. Efficacy of transformer networks for classification of raw EEG data. *arXiv preprint* arXiv:2202.05170 (2022).
91. Rojat, T. et al. Explainable artificial intelligence (xai) on timeseries data: A survey. *arXiv preprint* arXiv:2104.00950 (2021).
92. Klaise, J., Van Looveren, A., Vacanti, G. & Coca, A. Alibi explain: Algorithms for explaining machine learning models. *The Journal of Machine Learning Research* **22**, 8194–8200 (2021).
93. Koelstra, S. et al. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing* **3**, 18–31 (2011).

Author contributions

F.B. and F.D.G., contributed to the design and the conception of the research. F.B., L.F. and F.D.G. contributed to the implementation and the analysis of the results. F.B., F.D.G. and L.F. contributed to data collection. L.F. contributed to writing the manuscript. F.B., and F.D.G. contributed to the manuscript revision. L.F. prepared Figs. 1, 4, 5, F.D.G. prepared Figs. 2 and Supplementary Figs. 1, 2, 3, 4 and F.B. prepared Fig. 3. All authors contributed to the article, read and approved the submitted version.

Funding

This work was financially supported by Intesa Sanpaolo Innovation Center S.p.A.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-75263-x. Further supplementary materials can be found in the OSF Repository: https://osf.io/yv468/.

Correspondence and requests for materials should be addressed to F.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.