EEG-Based Contrastive Learning Models For Object Perception Using Multisensory Image-Audio Stimuli

Xuan-The Tran
tran.xuanthe@uts.edu.au
GrapheneX-UTS Human-Centric Artificial Intelligence Centre, Faculty of Engineering and Information Technology, University of Technology Sydney
Sydney, NSW, Australia

Quoc-Toan Nguyen
quoctoan.nguyen@student.uts.edu.au
GrapheneX-UTS Human-Centric Artificial Intelligence Centre, Faculty of Engineering and Information Technology, University of Technology Sydney
Sydney, NSW, Australia

Linh Le
linh.le@uts.edu.au
GrapheneX-UTS Human-Centric Artificial Intelligence Centre, Faculty of Engineering and Information Technology, University of Technology Sydney
Sydney, NSW, Australia

Thomas Do
thomas.do@uts.edu.au
GrapheneX-UTS Human-Centric Artificial Intelligence Centre, Faculty of Engineering and Information Technology, University of Technology Sydney
Sydney, NSW, Australia

Chin-Teng Lin
chin-teng.lin@uts.edu.au
GrapheneX-UTS Human-Centric Artificial Intelligence Centre, Faculty of Engineering and Information Technology, University of Technology Sydney
Sydney, NSW, Australia

Abstract

Multimedia sources such as images and audio commonly activate human senses to perceive objects, but limited research has explored the combined effect of these stimuli on predicting semantic object perception. In this study, we compare the performance of EEG signals elicited by image and audio stimuli in classifying semantic objects, revealing that image stimuli are more discriminative than audio stimuli. Building on this, we developed a contrastive learning model that integrates image and audio stimuli, further enhancing classification performance. Our research makes several key contributions: it compares classifier performance with uni-sensory versus multisensory stimuli, demonstrates improved performance with contrastive learning models using EEG data from both image and audio stimuli, and introduces a novel method to generate positive and negative pairs for contrastive learning models using cross-sensory EEG data. These findings enhance our understanding of how humans perceive multimedia sources and highlight the potential of multisensory integration in EEG-based classification.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**.

Keywords

Multisensory BCI, Multimedia Stimuli, Object Perception, EEG, Contrastive Learning

BCIMM '24, October 28-November 1, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-1189-3/24/10
https://doi.org/10.1145/3688862.3689116

ACM Reference Format:
Xuan-The Tran, Quoc-Toan Nguyen, Linh Le, Thomas Do, and Chin-Teng Lin. 2024. EEG-Based Contrastive Learning Models For Object Perception Using Multisensory Image-Audio Stimuli. In *Proceedings of the 1st International Workshop on Brain-Computer Interfaces (BCI) for Multimedia Understanding (BCIMM '24), October 28-November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3688862.3689116

1 Introduction

Humans perceive the world through multiple senses, and understanding this perception is key in Brain-Computer Interface (BCI) research. Various experimental paradigms have been designed to engage specific senses and analyze brain dynamics. Common paradigms include SSVEP [5, 17, 32], which uses visual flicker to study visual perception; spatial hearing [7, 21], which examines auditory perception; motor imagery [1, 9]; which investigates motor functions; and overt or covert speech [14, 18], which explores communication abilities. These paradigms have significantly advanced our knowledge of brain function and human perception. However, a gap remains in BCI research concerning experiments that simultaneously engage multiple senses using multiple sources of stimuli. In real-life scenarios, humans often rely on multisensory inputs to perform tasks more effectively. Therefore, this study aims to evaluate the effectiveness of brain signals activated by image and audio stimuli in object perception. We also examine whether combining EEG signals from both stimuli types enhances the classification performance of machine learning models in identifying the objects perceived by participants.

Much research has utilized machine learning classifiers to classify perceived objects using EEG data. Traditional models such as support vector machines (SVM), random forest classifiers, and KNN models use transformed EEG features that capture crucial information in temporal, spectral, or spatial domains to train the model [24]. In contrast, deep learning classifiers like RNN and LSTM models

[22] are favoured for handling raw EEG data without requiring feature extraction [25]. Other CNN-based models, such as EEGNet [12] and DeepConvNet[20], have proven effective in object perception classification tasks. Furthermore, deep learning classifiers incorporating self-attention mechanisms have achieved commendable results [23]. Recently, self-supervised learning (SSL) methods have become popular for EEG signal processing, especially for generative tasks like image or speech generation from brain signals [2, 6, 16]. SSL methods, which include generative SSL, predictive SSL, and contrastive SSL [28], are particularly beneficial in scenarios with limited labelled data—a common constraint in BCI research due to the restricted number of trials.

In this study, we employ a contrastive learning model [11] to improve the classification of perceived objects using EEG data from image and audio stimuli, as well as combined signals from these stimuli. By using EEG data activated by both image and audio stimuli, we form positive and negative pairs from both uni-sensory and cross-sensory EEG data. Positive pairs are created when participants perceive the same object through both seeing and hearing, while negative pairs are formed when participants perceive different objects through these senses. Our method of creating positive and negative pairs without needing EEG data transformation, as is typically done in existing research, represents a novel contribution to contrastive learning.

We used an open dataset [29] that recorded EEG signals when participants saw and listened to objects belonging to three classes: guitar, penguin, and flower. We conducted five experiments using traditional machine-learning models, a simple CNN, and contrastive-learning deep-learning models. The results show that our contrastive-learning models improved classification performance when using uni-sensory EEG data. Additionally, classification performance was enhanced when cross-sensory EEG data from image and audio stimuli were combined. This highlights the potential of cross-sensory data to improve the effectiveness of contrastive learning models in classifying perceived objects.

In summary, our research makes several key contributions. Firstly, we compare classifier performance using uni-sensory stimuli (image and audio) versus multisensory stimuli (image and audio simultaneously). Secondly, our findings demonstrate that contrastive learning models improve classification performance when using EEG data from both image and audio stimuli. Thirdly, we introduce a novel method to generate positive and negative pairs for contrastive learning models using cross-sensory EEG data. These contributions enhance our understanding of how humans perceive multimedia sources and highlight the potential of multisensory integration in EEG-based classification.

2 Methodology

2.1 The Dataset

We used an open dataset [29] that recorded EEG signals from twelve participants with normal or corrected vision and hearing, primarily students from the University of Bath. The original study involved three paradigm variations for perception tasks across three sensory modalities: visual image, orthographic, and auditory comprehension. The semantic categories used were flower, penguin, and guitar. However, in this study, we focused on image and audio stimuli because we consider both image and orthographic as visual stimuli, and images may be the more direct method to visualize the object. Figure 1 illustrates the experiment paradigm related to the image and audio stimuli perception tasks, which provided the data for our models. In the visual image paradigm, participants perceived and imagined images from the three semantic categories, with the stimuli being coloured images against a black background. In the auditory comprehension paradigm, participants listened to recordings of the semantic categories spoken by different speakers.

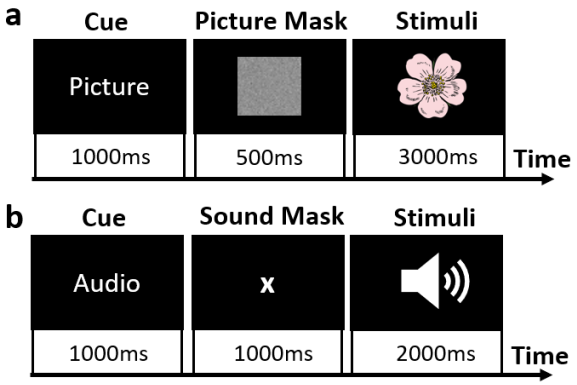

Figure 1: This figure illustrates the experimental paradigm used to present visual and auditory stimuli to participants. In the visual image paradigm (a), participants were shown images from one of three semantic categories (guitar, penguin, and flower), presented on a black background. The middle column shows a picture mask to make the participants stay focused. In the auditory paradigm (b), participants listened to recordings corresponding to the semantic categories (guitar, penguin, and flower). The "x" in the middle column indicates the sound mask. This setup allows for the comparison and combination of EEG responses to both visual and auditory stimuli for object perception tasks. This figure has been adapted from [29].

The experiment was conducted on a screen with a resolution of 1920 × 1080 pixels. EEG data was recorded using the ANT Neuro acquisition software 'eego', with triggers sent via a Lab Streaming Layer network to timestamp the stimuli and task information. A 128-channel ANT Neuro eego Mylab system with 124 EEG electrodes was used for data collection. The gel-based waveguard cap provided active shielding against environmental noise, and the data was sampled at 1024 Hz with a 24-bit resolution. Electrodes were positioned according to the five percent electrode system, an extension of the standard 10/20 layout. CPz served as the reference electrode, and the left mastoid was used as the ground. Impedance was maintained below 50 $k\omega$ for most electrodes to ensure high-quality data collection.

2.2 Data Processing

Both manual and automatic methods were employed to detect bad channels, and common average referencing in MNE was used

for re-referencing after each processing step. To ensure data quality, power-line noise at 50 Hz and its harmonics, as well as low-frequency drifts below 2 Hz, were filtered out. Independent Component Analysis (ICA) was applied to the raw pre-processed data to remove artifacts caused by eye movements and muscle activity. Artifact-related components were identified and rejected, resulting in cleaner EEG data for analysis.

For this research, we used the same epoch length of 2000ms for both image and audio stimuli EEG trials. We segmented the EEG data based on the semantic objects and sensory stimuli, resulting in six epoch groups: image-flower, image-penguin, image-guitar, audio-flower, audio-penguin, and audio-guitar (as shown in Figure 2). To prepare these epoch groups for our classifier model, we grouped epochs by image and audio stimuli, resulting in datasets with 1370 trials for each stimulus type. Each stimulus dataset was then split into training, validation, and test sets with a ratio of 60/20/20. The training, validation, and test sets are formatted as 3D arrays with the dimensions: number of EEG trials, number of EEG channels, and trial length. The shapes for the training, validation, and test sets are (822, 124, 2000), (274, 124, 2000), and (274, 124, 2000), respectively.

2.3 Classifiers

Machine learning models:

For this study, we employed three traditional machine learning models: Random Forest Classifier (RFC), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). The Random Forest Classifier was configured with 100 decision trees (n_estimators=100), using the Gini impurity criterion (criterion='gini') and a maximum depth of none (max_depth=None), allowing the trees to grow until all leaves are pure. The K-Nearest Neighbors model was set with 5 neighbours (n_neighbors=5), using the Euclidean distance metric (metric='euclidean') for calculating distances between data points. The Support Vector Machine was configured with a radial basis function kernel (kernel='rbf'), a regularization parameter of 1.0 (C=1.0), and a gamma value of 'scale' (gamma='scale'), which automatically adjusts the kernel coefficient based on the input data.

We train three machine learning models in image and audio stimuli EEG datasets extracted in temporal feature (mean), spectral feature (the power spectral density - PSD), and approximation entropy feature. We utilized a k-fold cross-validation approach with four folds and implemented an early stopping mechanism with the patience of 30 epochs to ensure that the model stopped training once performance on the validation set ceased to improve. The performance of the machine learning models is shown in Figure 2.

1D-CNN Deep learning models:

We utilized a straightforward 1D Convolutional Neural Network (1D-CNN) to classify the EEG data. Additionally, this 1D-CNN model also served as the backbone for the contrastive learning models in this study. The model architecture comprises three convolutional layers, each followed by batch normalization and max pooling. The first convolutional layer has 64 filters, a kernel size of 3, a stride of 1, and a padding of 1. This is followed by a batch normalization layer and a max-pooling layer with a kernel size of 2 and a stride of 2. The second convolutional layer has 128 filters with the same kernel size, stride, and padding, followed by another batch normalization and max-pooling layer. The third convolutional layer consists of 256 filters, again with a kernel size of 3, stride of 1, and padding of 1, followed by batch normalization and max pooling. After the convolutional layers, the data is flattened and passed through a fully connected layer with 512 units, followed by a dropout layer with a dropout rate of 0.5 to prevent overfitting. The final output layer is a fully connected layer with the number of units equal to the number of classes, using a linear activation function. This architecture allows the model to learn complex temporal patterns in the EEG data, enhancing its ability to accurately classify the perceived objects. A four-fold cross-validation strategy, along with 30 epochs of training, was also applied to train the 1D-CNN model.

Contrastive learning models:

The constrastive learning model is trained with EEG data from image and audio stimuli and combines data from these two stimuli. For the single stimuli data, the positive pairs are generated from trials that use the same object (flower, penguin, guitar), the negative pair are generated from cross-trials that use a different object (flower-penguin, flower-guitar, and penguin-guitar). For the cross-stimuli data, the positive pairs are generated from trials that use the same object but belong to different stimuli (image/audio). For instance, a positive pair can be formed from trials of image and audio stimuli in the same trial_id (show the same object, e.g., a flower). Negative pairs are created from trials that use different objects in image and audio cross-stimuli. For example, a trial using the object "flower" image and another trial using the object "guitar" audio would form a negative pair.

The core of our model is an embedding network that transforms the input EEG signals into a lower-dimensional space. We use the same 1D-CNN deep learning model (described above) as the backbone for our contrastive learning models, which allow direct comparison of the performance of 1D-CNN and constrastive learning models. The 1D-CNN model processes the EEG data from each modality and outputs a fixed-size embedding vector.

To train our model, we utilize the contrastive loss function (Equation 1). This loss function is specifically designed to handle pairs of inputs, encouraging the network to bring embeddings of similar pairs closer and push embeddings of dissimilar pairs apart. For similar pairs (positive pairs with label 1), the loss is calculated as the squared Euclidean distance between the embeddings. For dissimilar pairs (negative pairs with label 0), the loss is computed as the squared margin minus the squared Euclidean distance, ensuring that the embeddings of dissimilar pairs are at least a specified margin apart. The total loss is the mean of all individual losses, balancing the model's efforts to cluster similar pairs and separate dissimilar pairs effectively.

$$L = \frac{1}{N}\sum_{i=1}^{N}\Big(y_i \cdot \|f(x_i^a) - f(x_i^b)\|^2 + (1-y_i)\cdot \max(0, m - \|f(x_i^a) - f(x_i^b)\|)^2\Big) \tag{1}$$

where N is the number of pairs, y_i is the label (1 for similar pairs and 0 for dissimilar pairs), $f(x_i^a)$ and $f(x_i^b)$ are the embeddings of the two samples in the pair, and m is the margin.

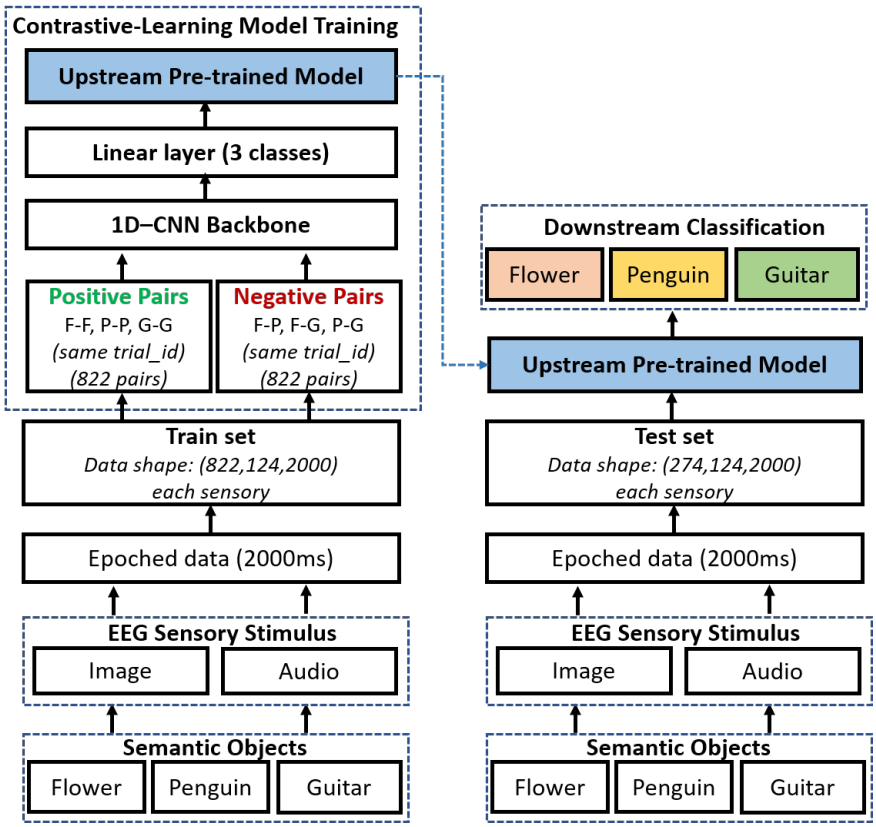

Figure 2: The figure illustrates the data processing and contrastive learning model training pipeline. The process begins with collecting EEG data from three semantic categories: Flower, Penguin, and Guitar. This data is then processed and labelled into positive pairs (same object class) and negative pairs (different object classes). The positive and negative pairs are fed into the contrastive learning model for training. The model leverages the upstream pre-trained models to refine its representations. The upstream pre-trained models were then used in downstream classification tasks for three semantic objects (flower, penguin, guitar).

The training involves feeding pairs of EEG signals through the embedding network to obtain their embeddings, calculating the contrastive loss, and updating the network parameters using back-propagation and an optimization algorithm (Adam optimizer). A four-fold cross-validation strategy, along with 30 epochs of training, was also applied to train the constrastive learning models. After training, the model is evaluated based on its ability to correctly identify positive and negative pairs. We used accuracy and confusion matrices to report the model's performance, as shown in Table 2 and Figure 4.

2.4 Experiments

In this study, we conducted five experiments to evaluate the effectiveness of various machine learning and deep learning models in classifying EEG data from different sensory modalities 1. These experiments were designed to compare the performance of traditional machine learning models, simple 1D-CNN, and contrastive learning models using EEG data from image and audio stimuli. By analyzing both uni-sensory and cross-sensory data, we aimed to determine which methods and combinations yield the highest classification accuracy and to explore the potential benefits of multisensory integration in EEG-based classification tasks.

Table 1: Overview of experiment setups in this study

Experiments	Training set	Testing set	Model
1	Image/Audio	Image/Audio	RFC, KNN, SVM
2	Image/Audio	Image/Audio	1D-CNN
3	Image	Image/Audio	CL
4	Audio	Image/Audio	CL
5	Image+Audio	Image/Audio	CL

2.4.1 Experiment 1: Train machine learning models with EEG data from image and audio stimulus. In the first experiment, we aimed to evaluate the performance of traditional machine learning models using various EEG feature sets for both audio and image stimuli. Specifically, we used Random Forest Classifier (RFC), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) models. The EEG feature sets considered in this experiment included temporal

mean features, power spectral density (PSD) features, and approximation entropy features. The goal was to determine which sensory EEG data (audio vs. image) yielded better performance in terms of classification accuracy.

2.4.2 Experiment 2: Train 1D-CNN Model with EEG data from image and audio stimulus. In the second experiment, we extended our evaluation by comparing the performance of a deep learning model, specifically a simple 1D-CNN. The objective was to assess whether the deep learning model could achieve higher classification accuracy than traditional machine learning models. Additionally, we aimed to confirm that the image-based EEG data continued to yield better performance than the audio-based EEG data when using a deep learning model.

2.4.3 Experiment 3 and 4: Contrastive learning model trained on uni-stimuli image and audio EEG dataset. The third experiment focused on developing a contrastive learning model specifically for audio stimuli within audio trials and for image stimuli within image trials. Contrastive learning is an effective technique for learning representations by contrasting positive pairs (same object class) against negative pairs (different object class). The goal was to determine which sensory EEG data (audio vs. image) yielded better performance in terms of classification accuracy and to assess whether contrastive learning models improve classification performance compared to traditional machine learning and deep learning models.

2.4.4 Experiment 5: Contrastive Learning model trained on image and audio cross-stimulus. The final experiment explored cross sensory contrastive learning by combining image and audio stimuli during the training phases. The Siamese Network was trained on EEG data from both image and audio stimuli, creating pairs that included cross-modal combinations (for example, an image trial paired with an audio trial). The model was evaluated on its ability to accurately classify and differentiate between mixed image and audio trials. This experiment aimed to leverage the rich and complementary information from both sensory modalities to enhance the overall performance of the contrastive learning model.

3 Results

3.1 Image stimuli yield higher object classification performance than audio stimuli

The comparative analysis of the test accuracy for different models across audio and image stimuli is presented in Figure 3. The Random Forest Classifier (RFC) achieved a test accuracy of 39.2% for audio stimuli and 51.2% for image stimuli. The K-Nearest Neighbors (KNN) model yielded test accuracies of 34.1% for audio and 44.5% for image stimuli. Similarly, the Support Vector Machine (SVM) model obtained accuracies of 32.0% for audio and 42.1% for image stimuli. The 1D Convolutional Neural Network (1D-CNN) model demonstrated test accuracies of 35.3% for audio and 48.6% for image stimuli. Across all models, the performance was consistently higher for image stimuli than audio stimuli, highlighting the more distinct neural responses elicited by visual stimuli. The inclusion of standard deviation bars provides an indication of the variability in model performance, with the image stimuli generally showing less variability and higher accuracy, emphasizing the effectiveness of visual stimuli in eliciting more robust and classifiable EEG responses.

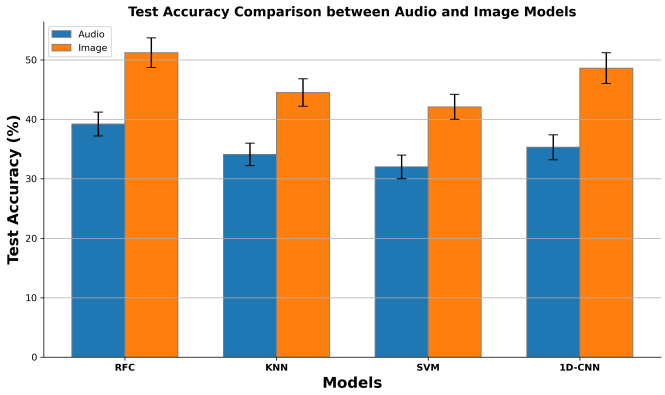

Figure 3: The figure compares the accuracy of different models for audio and image stimuli. Across all models, the accuracy was consistently higher for image stimuli than for audio stimuli, indicating that visual stimuli elicit more distinct neural responses. Standard deviation bars illustrate the variability in model performance.

3.2 Contrastive learning models improve object classification in both EEG image and audio stimulus datasets

The results presented in Table 2 highlight the performance improvements achieved by using Contrastive learning models compared to the 1D-CNN backbone for object classification in both EEG image and audio stimulus datasets. The 1D-CNN model achieved an accuracy of 48.6% (±3.12) for image stimuli and 35.3% (±2.05) for audio stimuli. In contrast, the contrastive learning model trained on image stimuli (CL_image) significantly improved the accuracy to 70.2% (±4.25) for image stimuli and 55.8% (±4.16) for audio stimuli. Similarly, the contrastive learning model trained on audio stimuli (CL_audio) achieved accuracies of 58.6% (±4.89) for image stimuli and 63.7% (±3.47) for audio stimuli. These results indicate that contrastive learning models can effectively leverage the underlying data structures in both EEG image and audio datasets, leading to significant improvements in classification performance compared to the traditional 1D-CNN approach.

In addition, the confusion matrices in Figure 4a,b provide a detailed comparison of the classification performance for different models and stimuli types. Figure 4a matrix shows the performance of the CL_image model on image stimuli. This model achieved a high accuracy, with clear distinctions between the classes. The Flower class (red) has a true positive count of 63, with 12 and 17 instances misclassified as Penguin and Guitar, respectively. The Penguin class (blue) shows 59 true positives, with 15 and 17 misclassifications as Flower and Guitar, respectively. The Guitar class (green) has the highest true positive count of 71, with minimal misclassifications, indicating a strong performance in classifying images.

The Figure 4b matrix represents the CL_audio model's performance on audio stimuli. This model shows a slightly lower accuracy than the image stimuli, with the Flower class having 53 true positives and higher misclassifications (25 as Penguin and 14 as Guitar). The Penguin class shows 60 true positives, with 15 and 16 instances misclassified as Flower and Guitar, respectively. The Guitar class has 62 true positives, with 15 misclassifications as Flower and 14 as Penguin. This highlights the challenges of distinguishing between classes when using audio stimuli, as evidenced by the higher misclassification rates.

Table 2: Performance of classifiers in our experiments

Models	Image Stimulus	Audio Stimulus
1D-CNN *(backbone)*	48.6 ± 3.12	35.3 ± 2.05
CL_image	*70.2 ± 4.25*	55.8 ± 4.16
CL_audio	58.6 ± 4.89	*63.7 ± 3.47*
CL_image_audio_equal	67.3 ± 4.63	60.9 ± 4.15
CL_image_audio_combine	**73.6 ± 3.52**	**66.2 ± 4.31**

3.3 The cross-stimulus image and audio improves contrastive learning models performance

Further analysis in Table 2 demonstrates that combining cross-stimuli image and audio data significantly enhances the performance of contrastive learning models. The contrastive learning model trained on a cross-sensory dataset (with the total number of positive and negative pairs equal to the number of these pairs in the uni-stimulus datasets)(CL_image_audio_equal) achieved an accuracy of 67.3% (±4.63) for image stimuli and 60.9% (±4.15) for audio stimuli. On the other hand, when combining image and audio stimuli, which doubled the number of training pairs, the model achieved the highest accuracies of 73.6% (±3.52) for image stimuli and 66.2% (±4.31) for audio stimuli. This combined cross-stimulus approach leverages the complementary information in both data types, resulting in improved model robustness and classification accuracy, highlighting the benefits of using a multisensory approach in EEG-based classification tasks.

The CL_image_audio_combine model has been shown in Figure 4c,d matrices demonstrate the performance of both image and audio stimuli. For image stimuli (Figure 4c), the model demonstrates superior performance with many true positives for each class: 136 for Flower, 139 for Penguin, and 129 for Guitar. Misclassifications are significantly reduced compared to the other models, indicating the effectiveness of combining image and audio data in improving classification accuracy. For audio stimuli (Figure 4d), the model also shows improved performance with true positive counts of 112 for Flower, 127 for Penguin, and 124 for Guitar. Although there are still some misclassifications, the combined model outperforms the audio-only model, suggesting that leveraging multiple sensory inputs enhances the model's robustness and accuracy in classifying EEG data.

3.4 Contrastive learning models performance explanation

To visualize how the contrastive learning model classifies objects in the representation space, we applied t-SNE visualization [26] for the best performance model (CL_image_audio_combine).

The CL_image_audio_combine model's performance on EEG data was shown in Figure 5, which illustrates the t-SNE visualization of three classes: Flower, Penguin, and Guitar. Figure 5a represents the results for image stimuli, while Figure 5b corresponds to audio stimuli.

In the image stimuli visualization, the clusters for each class (red for Flower, blue for Penguin, and green for Guitar) are more distinct and well-formed. Most data points for each class are tightly grouped around their respective centres (black 'x' markers), indicating fewer misclassifications. The Flower and Guitar clusters exhibit some overlap, but overall, the image stimuli data demonstrates a clear separation between the classes.

In contrast, the audio stimuli visualization shows more overlap between the clusters, particularly between the Flower and Guitar classes. Although the Penguin class remains relatively distinct, there is a noticeable increase in misclassified points compared to the image stimuli. The centres of the clusters are still visible, but the spread of points around these centres is larger, indicating a higher degree of classification uncertainty for audio stimuli.

This comparison highlights the model's performance in distinguishing between classes when using image stimuli compared to audio stimuli, as evidenced by the tighter and more distinct clusters in the t-SNE plot for image stimuli.

4 Discussion

In this paper, our primary objective was to explore the impact of multisensory integration on the performance of EEG-based classification models. We aimed to compare classifier performance using uni-sensory stimuli (image and audio) versus multisensory stimuli (image and audio simultaneously). Additionally, we introduced a novel method to generate positive and negative pairs for contrastive learning models using cross-sensory EEG data.

The results of our study demonstrate the significant advantages of multisensory integration in EEG-based classification. The constrastive learning model trained on multisensory EEG data consistently outperformed models that relied solely on either image or audio stimuli. The t-SNE visualizations revealed more distinct and compact clusters for the combined stimuli, particularly for image stimuli, indicating a higher accuracy and fewer misclassifications. The confusion matrices further corroborated these findings, showing a higher number of true positives and reduced misclassifications for the combined model. These results suggest that integrating multiple sensory inputs can enhance the discriminative power of EEG-based classification models.

Many contrastive learning models have been implemented with EEG data, such as Contrastive Predictive Coding [3, 4], Transformation Contrastive Learning [15], Non-negative EEG Contrastive Learning [8, 30], Spatial Contrastive Learning [13], and Graph Contrastive Methods [10, 31], all of which have demonstrated potential in enhancing model performance in various experiment tasks. In this study, we introduce a novel contrastive learning method that

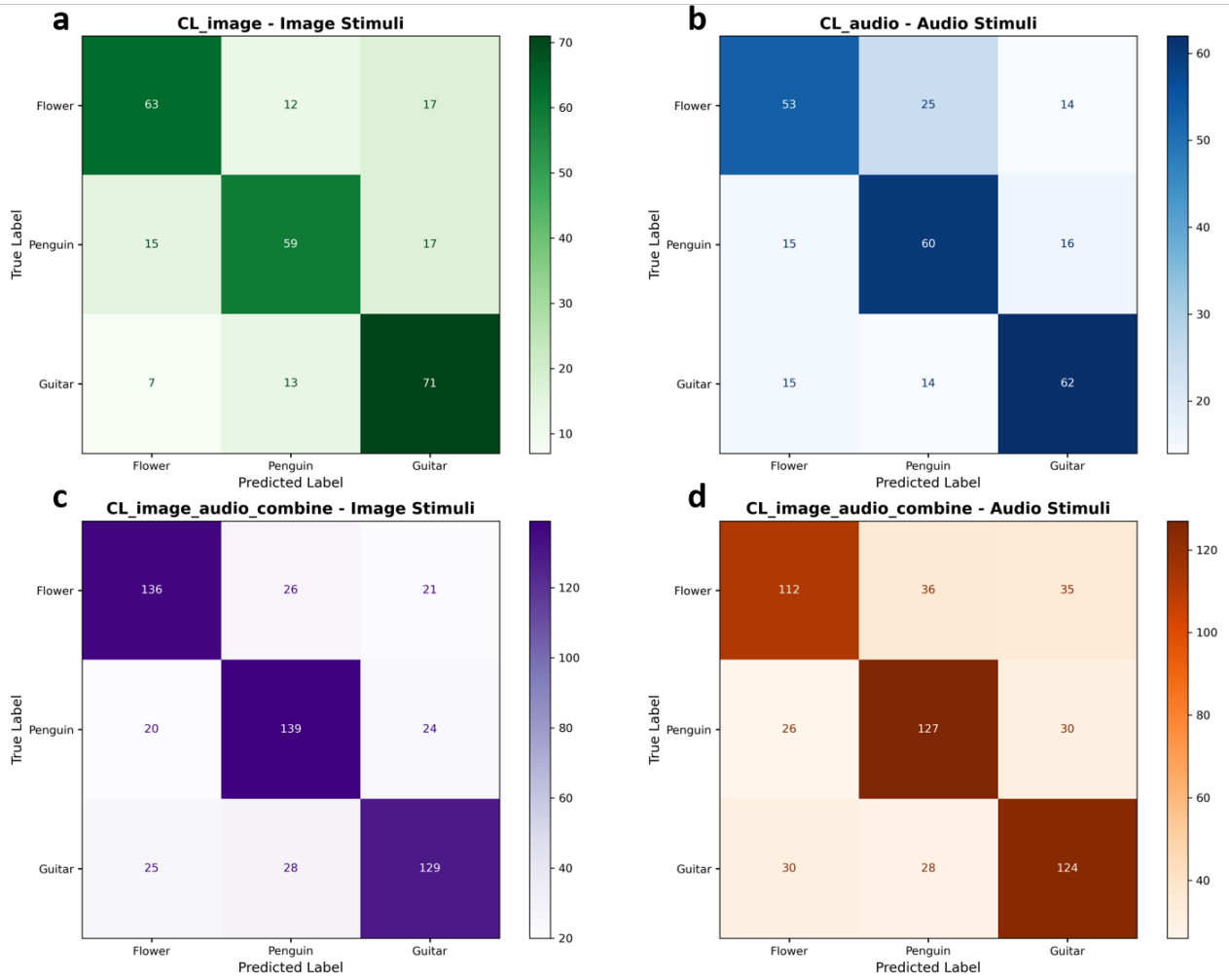

Figure 4: This figure shows four confusion matrices representing the classification performance of different contrastive learning models for image and audio stimuli. Subplot (a) displays the confusion matrix for the CL_image model using image stimuli, while subplot (b) shows the confusion matrix for the CL_audio model using audio stimuli. Subplot (c) illustrates the confusion matrix for the CL_image_audio_combine model using image stimuli, and subplot (d) presents the confusion matrix for the CL_image_audio_combine model using audio stimuli.

forms positive and negative pairs from cross-sensory EEG data. Our model underscores the potential of multisensory integration in improving the accuracy and robustness of EEG-based classification systems. The enhanced performance observed in the combined model can be attributed to the richer and more diverse information captured from both image and audio stimuli. This multisensory approach may align with how humans perceive and process information, providing a more comprehensive understanding of brain activity.

Our study builds upon previous research that has explored the use of EEG data for classification tasks [24]. However, unlike prior studies that predominantly focused on single-modal stimuli, this work highlights the benefits of a multisensory approach. Previous studies have shown that individual sensory modalities can provide valuable insights into brain activity, but they often suffer from limitations in accuracy and generalizability. By integrating both image and audio stimuli, our study addresses these limitations and demonstrates a significant improvement in classification performance. This approach not only enhances the model's accuracy but also provides a more holistic understanding of the underlying neural processes.

Despite the promising results, our study has several limitations that warrant further investigation. Firstly, we used 1D-CNN as the backbone for our contrastive learning model. More advanced deep learning models, such as Transformer [27] or U-NET [19], may be considered to improve the accuracy of the contrastive learning method. Secondly, a custom loss function could be designed to further optimize model performance. Additionally, our study focused

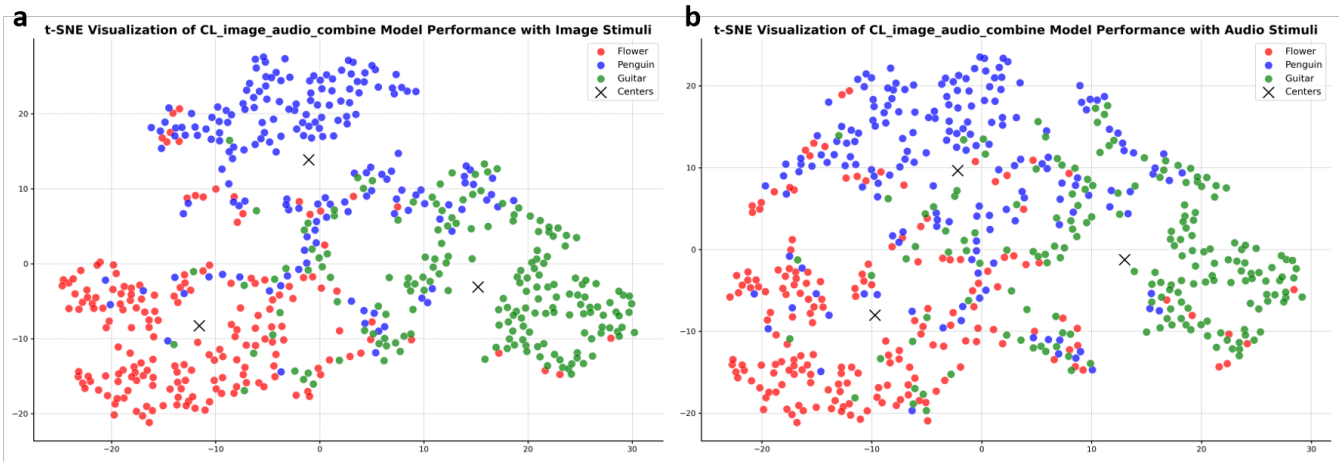

Figure 5: The figure illustrates t-SNE visualizations of the contrastive learning model (CL_image_audio_combine) performance with image and audio stimuli. Subplot (a) represents the t-SNE visualization for image stimuli, while subplot (b) shows the t-SNE visualization for audio stimuli. Each point corresponds to an EEG trial, with colours indicating different object classes: red for Flower, blue for Penguin, and green for Guitar. The black 'X' marks denote the cluster centers. The t-SNE plots reveal more distinct and compact clusters for image stimuli than audio stimuli, indicating better classification performance with visual inputs.

on three specific classes (Flower, Penguin, Guitar), and it is essential to evaluate the generalizability of the multisensory approach to other categories and contexts. Another limitation is that combining image and audio stimuli data effectively doubles the training size, leading to better model classification performance. However, when using an equal-size training set (between multisensory and each unisensory dataset), the CL_image_audio_equal model did not outperform the CL_image and CL_audio models in their respective testing stimuli. Future research should build upon these findings and address these limitations to further advance the field of EEG-based brain-computer interfaces and cognitive neuroscience.

5 Conclusion

Our study demonstrates that multisensory integration enhances the performance of EEG-based classification models. Combining image and audio stimuli, the contrastive learning model achieved superior accuracy and robustness compared to uni-sensory models. The t-SNE visualizations and confusion matrices revealed more distinct and compact clusters with fewer misclassifications, particularly for image stimuli. Our findings underscore the potential of leveraging multiple sensory inputs to improve the discriminative power and reliability of EEG-based classification systems. Additionally, our research demonstrates improved performance with contrastive learning models using EEG data from both image and audio stimuli and introduces a novel method to generate positive and negative pairs for contrastive learning models using cross-sensory EEG data. These contributions enhance our understanding of how humans perceive multimedia sources and highlight the potential of multisensory integration in EEG-based classification.

Acknowledgements

This work was partly supported by the Australian Research Council (ARC) under discovery grants DP210101093 and DP220100803 and the UTS Human-Centric AI Centre funding sponsored by GrapheneX (2023-2031). Research was also partially sponsored by the Australia Defence Innovation Hub under Contract No. P18-650825, Australian Cooperative Research Centres Projects (CRC-P) Round 11 CRCPXI000007, US Office of Naval Research Global under Cooperative Agreement Number ONRG - NICOP - N62909-19-1-2058, and AFOSR – DST Australian Autonomy Initiative agreement ID10134. We also thank the NSW Defence Innovation Network and the NSW State Government of Australia for financial support in part of this research through grant DINPP2019 S1-03/09 and PP21-22.03.02. Xuan-The Tran would like to thank the support of the Science and Technology Scholarship Program for Overseas Study for Master's and Ph.D. degrees at VinUniversity, Vingroup, Vietnam. Corresponding authors: Thomas Do and Chin-Teng Lin (Emails: thomas.do@uts.edu.au and chin-teng.lin@uts.edu.au)

References

[1] Minkyu Ahn and Sung Chan Jun. 2015. Performance variation in motor imagery brain–computer interface: a brief review. *Journal of neuroscience methods* 243 (2015), 103–110.

[2] Miguel Angrick, Maarten C Ottenhoff, Lorenz Diener, Darius Ivucic, Gabriel Ivucic, Sophocles Goulis, Jeremy Saal, Albert J Colon, Louis Wagner, Dean J Krusienski, et al. 2021. Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity. *Communications biology* 4, 1 (2021), 1055.

[3] Anahit Babayan, Miray Erbey, Deniz Kumral, Janis D Reinelt, Andrea MF Reiter, Josefin Röbbig, H Lina Schaare, Marie Uhlig, Alfred Anwander, Pierre-Louis Bazin, et al. 2019. A mind-brain-body dataset of MRI, EEG, cognition, emotion, and peripheral physiology in young and old adults. *Scientific data* 6, 1 (2019), 1–21.

[4] Hubert Banville, Omar Chehab, Aapo Hyvärinen, Denis-Alexander Engemann, and Alexandre Gramfort. 2021. Uncovering the structure of clinical EEG signals with self-supervised learning. *Journal of Neural Engineering* 18, 4 (2021), 046020.

[5] Fabrizio Beverina, Giorgio Palmas, Stefano Silvoni, Francesco Piccione, Silvio Giove, et al. 2003. User adaptive BCIs: SSVEP and P300 based interfaces. *PsychNology J.* 1, 4 (2003), 331–354.

[6] Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. 2023. Seeing beyond the brain: Masked modeling conditioned diffusion model for human vision decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[7] Francesco Ferracuti, Alessandro Freddi, Sabrina Iarlori, Sauro Longhi, and Paolo Peretti. 2013. Auditory paradigm for a P300 BCI system using spatial hearing. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 871–876.

[8] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems* 33 (2020), 21271–21284.

[9] Mahyar Hamedi, Sh-Hussain Salleh, and Alias Mohd Noor. 2016. Electroencephalographic motor imagery brain connectivity analysis for BCI: a review. *Neural computation* 28, 6 (2016), 999–1041.

[10] Thi Kieu Khanh Ho and Narges Armanfard. 2023. Self-supervised learning for anomalous channel detection in EEG graphs: Application to seizure analysis. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 7866–7874.

[11] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems* 33 (2020), 18661–18673.

[12] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. 2018. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of neural engineering* 15, 5 (2018), 056013.

[13] Yang Li, Ji Chen, Fu Li, Boxun Fu, Hao Wu, Youshuo Ji, Yijin Zhou, Yi Niu, Guangming Shi, and Wenming Zheng. 2022. GMSS: Graph-based multi-task self-supervised learning for EEG emotion recognition. *IEEE Transactions on Affective Computing* 14, 3 (2022), 2512–2525.

[14] Diego Lopez-Bernal, David Balderas, Pedro Ponce, and Arturo Molina. 2022. A state-of-the-art review of EEG-based imagined speech decoding. *Frontiers in human neuroscience* 16 (2022), 867281.

[15] Mostafa Neo Mohsenvand, Mohammad Rasool Izadi, and Pattie Maes. 2020. Contrastive representation learning for electroencephalogram classification. In *Machine Learning for Health*. PMLR, 238–253.

[16] David A Moses, Matthew K Leonard, Joseph G Makin, and Edward F Chang. 2019. Real-time decoding of question-and-answer speech dialogue using human cortical activity. *Nature communications* 10, 1 (2019), 3096.

[17] Liang Ou, Thomas Do, Xuan-The Tran, Daniel Leong, Yu-Cheng Chang, Yu-Kai Wang, and Chin-Teng Lin. 2023. Improving CCA Algorithms on SSVEP Classification with Reinforcement Learning Based Temporal Filtering. In *Australasian Joint Conference on Artificial Intelligence*. Springer, 376–386.

[18] Jerrin Thomas Panachakel and Angarai Ganesan Ramakrishnan. 2021. Decoding covert speech from EEG-a comprehensive review. *Frontiers in Neuroscience* 15 (2021), 642251.

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 234–241.

[20] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. 2017. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human brain mapping* 38, 11 (2017), 5391–5420.

[21] Martijn Schreuder, Benjamin Blankertz, and Michael Tangermann. 2010. A new auditory multi-class brain-computer interface paradigm: spatial hearing as an informative cue. *PloS one* 5, 4 (2010), e9813.

[22] Alex Sherstinsky. 2020. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena* 404 (2020), 132306.

[23] Jiayao Sun, Jin Xie, and Huihui Zhou. 2021. EEG classification with transformer-based models. In *2021 ieee 3rd global conference on life sciences and technologies (lifetech)*. IEEE, 92–93.

[24] Xuan-The Tran, Thomas Do, Nikhil R Pal, Tzyy-Ping Jung, and Chin-Teng Lin. 2024. Multimodal fusion for anticipating human decision performance. *Scientific Reports* 14, 1 (2024), 13217.

[25] Xuan-The Tran, Thomas Tien-Thong Do, and Chin-Teng Lin. 2023. Early Detection of Human Decision-Making in Concealed Object Visual Searching Tasks: An EEG-BiLSTM Study. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 1–4.

[26] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[28] Weining Weng, Yang Gu, Shuai Guo, Yuan Ma, Zhaohua Yang, Yuchen Liu, and Yiqiang Chen. 2024. Self-supervised Learning for Electroencephalogram: A Systematic Survey. *arXiv preprint arXiv:2401.05446* (2024).

[29] Holly Wilson, Mohammad Golbabaee, Michael J Proulx, Stephen Charles, and Eamonn O'Neill. 2023. EEG-based BCI dataset of semantic concepts for imagination and perception tasks. *Scientific Data* 10, 1 (2023), 386.

[30] Chaoqi Yang, Cao Xiao, M Brandon Westover, Jimeng Sun, et al. 2023. Self-supervised electroencephalogram representation learning for automatic sleep staging: model development and evaluation study. *JMIR AI* 2, 1 (2023), e46769.

[31] Weishan Ye, Zhiguo Zhang, Min Zhang, Fei Teng, Li Zhang, Linling Li, Gan Huang, Jianhong Wang, Dong Ni, and Zhen Liang. 2023. Semi-supervised dual-stream self-attentive adversarial graph contrastive learning for cross-subject eeg-based emotion recognition. *arXiv preprint arXiv:2308.11635* (2023).

[32] Danhua Zhu, Jordi Bieger, Gary Garcia Molina, and Ronald M Aarts. 2010. A survey of stimulation methods used in SSVEP-based BCIs. *Computational intelligence and neuroscience* 2010, 1 (2010), 702357.