Emotion Recognition Using EEG, Audio & Video Data

Merna Bibars [*1] **Danielle Dowe** [*1] **Khalid Khawaji** [*2] **Jeffrey Liu** [*2]

Abstract

Reliable emotion recognition is vital for affect-aware artificial intelligence, yet most electroencephalography (EEG)-based systems remain subject-dependent, requiring calibration for every new user. The EEG-Audio-Video (EAV) dataset from Lee et al. (2024) was investigated with replication of six established unimodal architectures (convolutional neural networks (CNNs) and transformers) and development of new random forest models. All models were retrained with 5-fold cross-validation and evaluated for within-subject and cross-subject performance. Re-implementation with cross-validation matched or modestly exceeded the EAV baselines for within-subject models; cross-subject performance dropped sharply for EEG but remained comparatively resilient for audio and video. Random forests demonstrated more robust performances on video data relative to CNNs and transformers. The study sets a quantitative baseline for future domain-adaptation and multimodal fusion work aimed at truly user-agnostic emotion-aware systems.

Code: https://github.com/kmak1001/StatMLEAV

Keywords: emotion recognition, EEG, audio, video, machine learning, convolutional neural networks, transformers, random forests

1. Introduction

Emotion recognition is a key capability for human-centric artificial intelligence (AI), enabling more natural and engaging interactions by allowing systems to adapt to users' affective states (Khare et al., 2024; Narimisaei et al., 2024). Traditionally, emotion recognition relies on observable cues such as facial expressions and speech intonation, which are relatively easy to capture and provide intuitive insight into a person's emotional state. However, emotions are complex and manifest differently across individuals, posing a challenge for models trained on one group of subjects to generalize to new people. Physiological signals like electroencephalography (EEG) have also been used to infer emotions directly from brain activity, potentially capturing affective processes not evident in outward behavior. Yet EEG signals are notoriously noisy and vary significantly between individuals, making it difficult to build emotion models that generalize across users (Ahuja & Sethia, 2025).

A key challenge in affective computing is to develop subject-independent models that remain accurate for unseen individuals. Many EEG-based emotion classifiers achieve high accuracy when trained and tested on the same subjects, but their performance drops markedly in cross-subject evaluations due to large inter-subject differences (Arjun et al., 2021). Thus, techniques for subject-independent modeling are critical for real-world deployments of emotion AI. The recently released EEG-Audio-Video (EAV) dataset (Lee et al., 2024) provides an excellent testbed for studying this problem. EAV contains synchronized EEG, audio, and video recordings from 42 participants in emotional conversational scenarios (covering five distinct emotion categories). Notably, it is the first dataset to encompass all three modalities in a unified conversational context, enabling direct comparisons of unimodal emotion recognition approaches. Lee et al. reported baseline emotion recognition models for each modality using deep neural networks, but these were trained in a subject-dependent manner. It remains unclear how well such unimodal models perform on entirely new participants—a gap we aim to address in this work.

In this paper, we systematically investigate subject-independent emotion recognition on the EAV dataset using each modality in isolation. We replicate representative convolutional neural network (CNN) and transformer architectures for each data modality and train them on data pooled from all subjects. We compare the performance of EEG-only, audio-only, and video-only models under this cross-subject setting and analyze which modalities are more robust to subject changes. To our knowledge, this is the first study to systematically evaluate subject-independent emotion recognition across EEG, audio, and video modalities on a common dataset. Our findings shed light on the challenges of cross-subject generalization in emotion recognition and

[*]Equal contribution [1]Machine Learning Center, Georgia Institute of Technology, Atlanta, GA, USA [2]Department of Electrical Engineering, Georgia Institute of Technology, Atlanta, GA, USA.

highlight the relative strengths of each modality for building more generalizable affective computing systems.

2. Methodology

2.1. Data

Data were taken from the EEG-Audio-Video (EAV) dataset (Lee et al., 2024). The EAV dataset included 42 professional actors portraying a given emotion (happiness, anger, sadness, calmness, and neutrality) while reading a 20-second script in a simulated conversation. For each subject, 100 trials, evenly distributed over the 5 emotions, were taken. Each trial was then segmented into 4 5-second sections, effectively creating 400 data points.

During these trials, EEG data were recorded on 30 electrodes at 500 Hz. The voltages were then band-pass filtered at 0.5 and 50 Hz and downsampled to 100 Hz. The filtered voltages in the time domain were used as the features. Audio of the speech was also recorded. Mel-frequency cepstral coefficients and chroma features were calculated. Video was taken of the facial features, recorded at 30 frames per second with 56×56 pixels. The videos were then downsampled to 5 frames per second.

2.2. Model Implementation

We followed a three-step process (Figure 1) to test and improve emotion recognition using the EAV dataset.

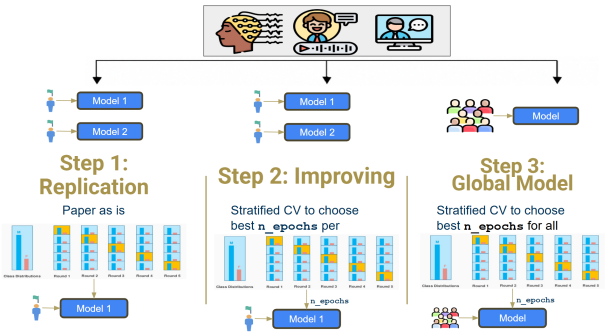

Figure 1. The three-step pipeline (Replication, Improving, and Global Model) was applied independently to each data modality: EEG, Audio, and Video.

2.2.1. Replication

In the first step, we repeated the same setup used in the original EAV paper. For each subject, the data were split into 70% for training and 30% for testing. The same models as the paper were used: EEGNet (Lawhern et al., 2018) and EEGformer (Wan et al., 2023) for EEG signals, Sequential Convolutional Neural Network (SCNN) (Yang et al., 2017) and Audio Spectrogram Transformer (AST) (Gong et al., 2021) for audio, and DeepFace (Taigman et al., 2014) and Video Vision Transformer (ViViT) (Arnab et al., 2021) for video. The average accuracy and F1 scores across all 42 subjects were taken and compared to Lee et al.'s results.

2.2.2. Improvement with Cross-Validation

Lee et al. did not use a separate validation set for any model selection, and our second step focused on improving performance by using 5-fold stratified cross-validation to pick the best number of training epochs for each subject. This helped make sure that each model trained just long enough to perform well without overfitting, which was found to be a problem in the original paper. We then retrained each model using these settings on the entire training set before evaluating on the held-out testing set.

2.2.3. Global Model

In the third step, we trained a single model for each modality using data from all subjects combined. This "global model" approach allowed us to test how well a shared model could generalize to different people. We again used cross-validation to decide how long to train, and then tested the final model on a global test set made by combining each subject's test data.

2.3. Random Forests

Random forest models (Breiman, 2001) were also developed to compare traditional machine learning with state-of-the-art deep learning methods. We applied the same 70%/30% data split and tuned the number of trees and depth using 5-fold stratified cross-validation. We tested this for EEG, audio, and video separately using the same features used in our deep learning experiments. This was performed for subject-dependent and subject-independent settings.

3. Results

We report performance in two stages. Section 3.1 analyses the subject-dependent setting, contrasting the originally published numbers ("Paper") with our *Replicated* re-implementation and the cross-validated *Improved* models (Tables 3). Section 3.2 examines subject-independent generalisation by comparing the best subject-dependent models (*Improved*) with a single *Global* model trained and evaluated on pooled data (Table 3, Fig.2).

3.1. Subject-Dependent Performance

Replication fidelity. Our re-implementation reproduced the modality ranking reported by Lee *et al.*: video > audio

	Paper		Replicated		Improved		Global	
Model	ACC	F1	ACC	F1	ACC	F1	ACC	F1
EEG								
EEGNet	59.5	58	47.7	46	**59.6**	**58**	42.3	35
EEGFormer	53.5	52	48.0	45	51.9	50	**50.3**	**50**
Rand. Forest	–	–	–	–	33.6	33	33.4	30
Audio								
SCNN	61.9	61	60.2	61	60.5	59	57.7	58
AST	62.7	62	56.6	56	**67.8**	**67**	**60.1**	**60**
Rand. Forest	–	–	–	–	35.5	35	39.0	38
Video								
DeepFace	71.4	70	67.0	65	70.7	65	71.6	71
ViViT	74.5	72	71.6	70	72.6	71	**74.6**	**74**
Rand. Forest	–	–	–	–	**77.8**	**78**	73.2	73

Table 1. **Mean classification accuracy (ACC) and macro-averaged F1-score (%) for EEG, Audio, and Video models under four training regimes.** "Paper" refers to the originally published results of Lee et al. (Lee et al., 2024), obtained via per-subject 70/30 train/test splits without cross-validation. *Replication* denotes our reimplementation using the same subject-dependent protocol and splits. *Improved* reports performance after introducing 5-fold stratified cross-validation on the 70% training set, still in a subject-dependent setting. *Global* corresponds to subject-independent training and evaluation, where models were trained with cross-validation on pooled data from all 42 subjects and tested on a pooled 30% global test set. Bold values indicate the best performance within each modality, and gray font highlights the largest drop in performance between the cross-validated subject-dependent and global evaluations.

> EEG. Absolute accuracies, however, fell short by 2–12 % depending on architecture (Table 3). The gap was smallest for the audio SCNN (–1.7 % in ACC), and largest for EEGNet (–11.8 %), indicating that the vision pipeline in the original study is more robust to implementation details than the EEG pipeline.

Cross-validated improvement. Introducing 5-fold stratified cross-validation on the 70 % training split yielded consistent gains (Table 3). The strongest lift occurred in the audio AST model, whose accuracy jumped from 56.6 % to 67.8 % (+11.2 %), surpassing the original paper's benchmark by 5.1 pp and narrowing the video–audio gap to within 5 %. EEGNet fully recovered to the published 59.5 % (+11.9 %) and eliminated the deficit to EEGFormer. Vision models were already near ceiling: ViViT rose modestly (+1.0 %), while DeepFace remained within 1 % of the original score. Random-forest baselines highlight feature strength rather than network depth: the video RF attained 77.8 % accuracy—highest overall in the subject-dependent setting—whereas RFs on EEG and audio stayed near 35 %.

Cross-validation rectifies much of the under-performance observed in the naïve replication, with the largest relative benefit for EEG and audio. Yet the modality hierarchy persists: rich visual cues in facial video remain decisively more informative than audio or EEG signals when training and testing occur on the same individual.

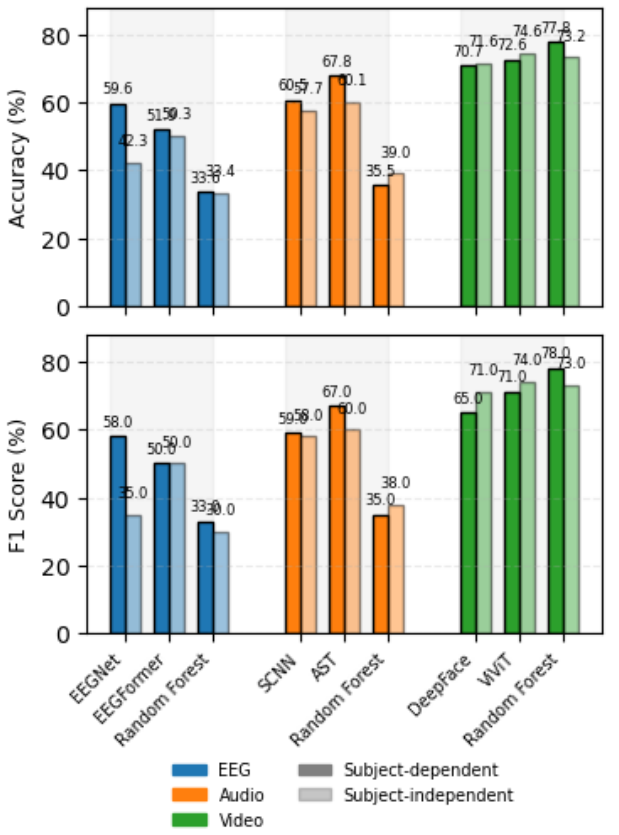

Figure 2. **Performance comparison of subject-dependent versus subject-independent models across EEG, audio, and video modalities.** The top panel shows accuracy (%) and the bottom panel shows F1 score (%) for each modality. Each pair of adjacent bars corresponds to one modality: the solid bar is subject-dependent performance and the transparent bar is subject-independent performance.

3.2. Subject-Independent Generalisation

Figure 2 juxtaposes each *Improved* (subject-dependent) model with its *Global* (subject-independent) counterpart; numerical values are in Table 3. Generalization to unseen subjects incurs a performance penalty across all modalities, but the magnitude is modality-specific.

EEG. EEGNet suffers the steepest decline (–17.3 % in ACC, highlighted in grey in Fig.2), highlighting the notorious inter-individual variability of neural signals. Even the

more expressive EEGFormer retains only 50.3 % accuracy (–1.6 % relative to EEGNet in the global setting).

Audio. The audio AST model, while robust within the subject, loses 7. 7 % when evaluated globally (67. 8 % → 60. 1 %). However, it remains the best performing audio approach and outperforms all EEG models with subject independence.

Video. Vision models exhibit the smallest degradation. ViViT drops a mere 2.0 % (72.6 % → 74.6 %), retaining state-of-the-art performance. The video RF also remains competitive (73.2 %), confirming that facial features generalize better between individuals than biosignals. The cross-subject penalty scales with signal idoidiosyncrasy: EEG $\geq$ audio $>$ video. Although cross-validation equalizes subject-dependent performance, it cannot overcome the inherent variability of EEG, nor fully stabilize audio models. In contrast, video-based approaches exhibit resilience, suggesting that they are the most viable backbone for population-level affective computing systems.

4. Discussion

Our experiments establish a clear hierarchy of modality effectiveness: video-based classifiers achieve the highest emotion-recognition accuracy, followed by audio and then EEG. On our dataset, baseline results on our multimodal dataset showed mean accuracies of approximately 71.4% for video, 61.9% for audio, and 60.0% for EEG. Comparable studies report the same ordering, e.g., approximately 67.2% for video, 58.2% for audio, and 53.5% for EEG. This consistent pattern arises from fundamental differences in information content and noise across modalities. There is a clear disparity in the number of extracted features across modes. Visual streams deliver high-dimensional, spatiotemporal information—about 235 000 features for a one-second clip (25 frames × 3 channels × 56 × 56 pixels)—that directly encodes facial expressions and gestures. Audio contributes roughly 80 000 features (Mel-frequency cepstral coefficients and chroma descriptors) capturing prosody, while EEG offers the sparsest representation, around 15 000 features (30 channels at 100 Hz over 5 s) and is most susceptible to artefacts and inter-subject variability. Accordingly, rich visual cues dominate classification performance, audio adds moderate discriminative power, and EEG provides the least.

A likely reason EEG underperformed in these experiments is the absence of spectral-feature extraction; prior work shows that converting time-domain voltages to the frequency domain and examining band-specific power improves accuracy (Zhang et al., 2023). Decomposing the signal into the canonical into canonical delta, theta, alpha, beta, and gamma bands and computing power or spectral entropy captures rhythms that often correlate with emotional states. Fast Fourier, wavelet, and related transforms thus yield substantially richer representations than time-domain samples alone, and routinely drive higher classification accuracy. To narrow the modality gap, future multimodal emotion-recognition studies should attempt other methods of feature extraction, such as frame differencing for video data (Li et al., 2025) and evoked potentials for EEGs (Nie & Ku, 2023).

One plausible explanation for the performance gap is that the dataset was recorded with actors deliberately portraying each emotion through scripted, cue-guided dialogues. Acted expressions are typically more stereotyped—and therefore easier to recognize—than spontaneous ones. Prior work shows that feature importance shifts between the two settings: pitch-related cues dominate acted speech, whereas MFCCs are more informative for natural speech. Consequently, models trained on our cue-driven sessions may overestimate real-world accuracy. The effect is modality-specific: exaggerated facial and vocal cues can inflate video and audio scores, while EEG signals, which reflect genuine affective states, may change little. Our findings—particularly the strong performance of visual models—should therefore be interpreted in light of this collection paradigm. Future work should test whether the same modality hierarchy persists when emotions are elicited naturally and without conscious control, where EEG differences may become more pronounced.

As expected, the performance of the global (subject-independent) models was slightly worse than that of the subject-dependent models. Introducing inter-subject variability—and recognizing that intra-subject variability differs across individuals—makes generalization more difficult. For instance, one person may express anger with furrowed brows, while another bares their teeth. However, the models were remarkably resilient, with most global classifiers only decreasing by 1–7 % in accuracy. The notable exception was EEGNet, which experienced a 17.3 % decline. Additional testing on data from new subjects is needed to fully evaluate robustness. Finally, a logical next step is to develop a multimodal fusion model that integrates all three modalities. Fusion can occur at the feature level (early fusion) or by combining the outputs of each modality-specific model (late fusion). Our aim is to determine how much multimodal integration improves emotion-prediction performance—and whether the accuracy gains justify the additional computational cost of training and inference.

5. Conclusion

Replication of the average accuracy and F1 scores from the EEG-audio-video paper (Lee et al., 2024) could not be achieved; however, upon application of 5-fold cross vali-

dation, metrics approached that of the authors'. The video modality performed best, followed by audio, then EEG. Global models were implemented for each modality and saw a slight decrease in performance relative to all-subject models. Random forests proved more accurate than transformers or CNNs for video data; nonetheless, accuracy was significantly worse for EEG and audio data. Future analyses should examine the effectiveness of models combining all three modalities and other feature extraction methods.

6. Author Contributions

- Merna Bibars: Audio experiments
- Danielle Dowe: EEG experiments
- Khalid Khawaji: Video experiments
- Jeffrey Liu: Random forest experiments

References

Ahuja, C. and Sethia, D. Ss-emerge - self-supervised enhancement for multidimension emotion recognition using gnns for eeg. *Scientific Reports*, 15(14254), 2025.

Arjun, A., Singh Rajpoot, M., and Raveendranatha Panicker, M. Subject independent emotion recognition using eeg signals employing attention driven neural networks. *arXiv preprint arXiv:2106.03461*, 2021. URL https://arxiv.org/abs/2106.03461. Submitted on 7 Jun 2021, last revised 21 Dec 2021.

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., and Schmid, C. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6836–6846, 2021.

Breiman, L. Random forests. *Machine Learning*, 45(1): 5–32, 2001.

Gong, Y., Chung, Y.-A., and Glass, J. Ast: Audio spectrogram transformer. In *Proceedings of the Interspeech 2021*, pp. 571–575, 2021.

Khare, S. K., Blanes-Vidal, V., Nadimi, E. S., and Acharya, U. R. Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Information Fusion*, 102:102019, 2024. ISSN 1566-2535.

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., and Lance, B. J. Eegnet: A compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of Neural Engineering*, 15 (5):056013, 2018.

Lee, M.-H., Shomanov, A., Begim, B., Kabidenova, Z., Nyssanbay, A., Yazici, A., and Lee, S.-W. Eav: Eeg-audio-video dataset for emotion recognition in conversational contexts. *Scientific Data*, 11(1026), 2024.

Li, J., Zhou, H., Qian, Y., Dong, Z., and Wang, S.-J. Micro-expression recognition using dual-view self-supervised contrastive learning with intensity perception. *Neurocomputing*, 619:129142, 2025. ISSN 0925-2312.

Narimisaei, J., Naeim, M., Imannezhad, S., Samian, P., and Sobhani, M. Exploring emotional intelligence in artificial intelligence systems: a comprehensive analysis of emotion recognition and response mechanisms. *Annals of Medicine and Surgery*, 86(8):4657–4663, 2024.

Nie, L. and Ku, Y. Decoding emotion from high-frequency steady state visual evoked potential (ssvep). *Journal of Neuroscience Methods*, 395(109919), 2023.

Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1701–1708, 2014.

Wan, Z., Li, M., Liu, S., Huang, J., Tan, H., and Duan, W. Eegformer: A transformer-based brain activity classification method using eeg signal. *Frontiers in Neuroscience*, 17:1148855, 2023.

Yang, H., Yuan, C., Xing, J., and Hu, W. Scnn: Sequential convolutional neural network for human action recognition in videos. In *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 355–359, 2017.

Zhang, H., Zhou, Q.-Q., Chen, H., Hu, X.-Q., Li, W.-G., Bai, Y., Han, J.-X., Wang, Y., Liang, Z.-H., Chen, D., Cong, F.-Y., Yan, J.-Q., and Li, X.-L. The applied principles of eeg analysis methods in neuroscience and clinical neurology. *Military Medical Research*, 10(1):67, 2023.