# A Hand-Crafted EEG Feature Bank with Stability-Based Selection for Movie Recall Detection

Hao-Tien Yu[1,*], Yi-En Dong[1]

[1]Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan

Abstract

Challenge 1.2 of the MediaEval 2026 *Predicting Movie and Commercial Memorability* task asks whether a viewer remembers a particular video clip from electroencephalography (EEG) alone. The EEG markers of recognition differ substantially across individuals, so we assemble a large feature bank, refine it through selection, and measure how far performance depends on a participant's own data. We clean each participant's raw recordings with independent component analysis (ICA) and AutoReject, removing ocular and muscular artifacts. From the cleaned epochs we compute about 11,000 features and reduce them to a small subset using stability selection and a forward search. We test five classifiers and their stacking ensemble under three training scopes (subject-specific, cross-subject, and all-subject). The best model in each scope reaches development AUC of 0.67, 0.60, and 0.59. Within each participant, however, decoding falls to near chance (mean AUC 0.53), so these aggregate scores mostly reflect stable differences between participants rather than a per-trial memory signal. Our best submitted run, a subject-specific SVM, reaches test AUC 0.607. The cross-subject run drops to 0.497, and every run scores lower on test than on development. Code and supplementary results are available at https://github.com/haotien91/2026-Predicting-movie-and-commercial-memorability_NTHU-CY.

## 1. Introduction and Related Work

Whether a viewer is familiar with a video reveals something about what makes media memorable, and it has uses in advertising and content design. Challenge 1.2 of the MediaEval 2026 *Predicting Movie and Commercial Memorability* task [1] asks this directly, "Is this person familiar with this video?", and the answer must come one trial at a time from EEG alone, with nothing read off the videos.

The Movie Memorability Dataset [2] records scalp EEG from 27 participants and labels each clip as remembered or not, giving 3,484 labeled epochs (2,122 not recognized, 1,362 remembered). Each participant's epochs are split into a labeled development portion, used for all training and model selection, and a held-out test portion scored by the organizers. The major difficulty is subject variability. EEG patterns that separate remembered from forgotten clips differ widely across people, so a model trained on one group transfers poorly to another. The 2025 results of Martín-Fernández et al. [3] show this directly, with AUC falling from 0.656 within subjects to 0.530 across them. It is also unclear which neural markers matter most, so we build a large feature bank and let data-driven selection prune it. To see how much performance depends on a participant's own data, we compare three *training scopes*, defined by whose data the model trains on relative to the person it is tested on. The *subject-specific* scope uses the person's own data, *cross-subject* uses leave-one-subject-out over the other 26 participants, and *all-subject* trains one model on all 27 at once.

*MediaEval'26: Multimedia Evaluation Workshop, June 15–16, 2026, Amsterdam, Netherlands and Online*
*Corresponding author.
charlessworknp@gmail.com (H. Yu); jddlake@gmail.com (Y. Dong)
0009-0008-9767-0428 (H. Yu); 0009-0006-1486-547X (Y. Dong)

On the 2025 edition of this task, Martín-Fernández et al. [3] used about 31 features across five functional domains (visual, auditory, semantic, emotion, and attention) with an ElasticNet classifier under leave-one-subject-out cross-validation. Their AUC was highest when the classifier saw each subject's own data (0.656, against a chance level of 0.50) and fell to 0.530 when it saw only the other subjects. The 2021 EEG pilot [4] reached within-subject AUC near 0.59, and both a coherence-map study [5] and a 2025 stacking-and-boosting entry [6] stayed near chance. Our pipeline cleans the raw recordings with ICA and AutoReject, then builds a feature bank of about 11,000 features and prunes it per classifier by stability-based selection. We run five classifiers and a stacking ensemble across the three training scopes.

## 2. Methods and Approach

### 2.1. Preprocessing

Since the dataset's pre-extracted ERP and ERSP features retain ocular and muscular artifacts that could obscure the weak recall signal, we extract features from the raw EEG instead. After re-referencing each recording to the common average, we clean the signals in two stages. First, we apply AutoReject [7], which learns per-channel peak-to-peak thresholds by cross-validation and repairs each epoch by interpolating its bad channels rather than dropping it. This cuts the mean epoch loss from 9.3% under a fixed peak-to-peak threshold to 1.3%. Second, we run extended-infomax ICA [8] with 15 components, label them with ICLabel [9], and remove any component classified as an eye-blink or muscle artifact above probability 0.6. Because AutoReject has already suppressed most artifacts, this step removes components in only 6 of the 27 participants, which avoids discarding genuine neural sources. We process each participant separately.

### 2.2. Feature Bank

Since it is unclear which signals carry the memory trace, we favor a broad bank over a hand-picked set. We compute roughly 11,000 features per epoch across ten families: band power (3528 base features), band-power ratios (60), alpha-adapted power in bands set by each participant's individual alpha (162), ERP components and differences (29), signal complexity (378), nonlinear dynamics (216), phase-amplitude coupling (486), phase locking (420), and the weighted phase lag index (240). Each is computed over several channels, regions, and time windows. EEG amplitudes carry large offsets between participants from differences in skull and electrode placement, so we add a z-scored copy of every feature, normalized with each participant's own mean and standard deviation over all of that participant's epochs. The normalization is unsupervised and reads no labels. It doubles the 5,519 base features to 11,038. No family is assumed useful in advance. Selection decides what to keep.

### 2.3. Feature Selection

With far more features than epochs, selection matters, and we run it separately for each classifier. We first rank features by stability selection [10], which keeps the features chosen consistently across resampled data rather than those that fit one split by chance. Across 30 subsamples that each hold a random half of the participants, we fit the classifier and record how often each feature ranks among that model's most important. We then go down this ranking with a forward search, adding features in order and scoring subset sizes $k$ from 1 to 500 by five-fold cross-validation over participants. We keep the $k$ that maximizes balanced accuracy. Balanced

**Table 1**
Submitted runs: held-out test AUC and the development-set AUC used to select the runs. Stacking (all models) combines all five classifiers, and Stacking (subset) a per-participant subset of them. The development-set value is higher than the test value for every run.

| Model | Training scope | Test AUC | Dev. AUC |
|---|---|---|---|
| SVM | subject-specific | 0.607 | 0.640 |
| Stacking (all models) | subject-specific | 0.584 | 0.658 |
| CatBoost | subject-specific | 0.569 | 0.619 |
| Stacking (subset) | subject-specific | 0.564 | 0.670 |
| XGBoost | cross-subject | 0.497 | 0.600 |

accuracy is only the internal criterion for choosing $k$ here, not a reported outcome. The score peaks at small $k$, around 30 to 50 features for the tree models, then declines as more are added, so the search also controls model complexity.

### 2.4. Classifiers and Stacking

We use five classifiers spanning linear, kernel, and tree-based families: ElasticNet logistic regression, an SVM with a radial basis function kernel, LightGBM, XGBoost, and CatBoost. Each is paired with its own selected feature subset, so the five draw on different feature families, and since no single classifier is best across scopes we combine them by stacking rather than committing to one. A second-level logistic regression, the *meta-classifier*, takes the five base models' predicted probabilities and returns the final probability. We report two stacking variants, one using all five base models and one that selects a subset of them by cross-validation. The meta-classifier is evaluated out-of-fold with folds that keep each participant on one side only, a 5-fold StratifiedKFold within participants for subject-specific stacking and a participant-keyed GroupKFold for all-subject stacking. This matches the leave-one-subject-out scheme used for cross-subject training.

## 3. Results and Analysis

### 3.1. Performance

We report an *aggregate AUC*, scoring all participants' trials together as one set instead of averaging a separate AUC for each participant. Apart from the test AUC in Table 1, all results are development-set estimates from cross-validation.

Every run scored lower on test than on development, by 0.033 to 0.106. The development ordering did not carry over. The run with the highest development AUC (0.670) placed only fourth on test. The subject-specific SVM reached the highest test AUC (0.607) with the smallest drop, and it beat both stacking runs, so the meta-classifier did not help on the test set. The cross-subject run, by contrast, fell to chance (0.497). Across the development set, no single classifier dominated the way ElasticNet did for the 2025 entry [3], and SVM in particular was sensitive to scope, dropping below chance under all-subject training.

The aggregate AUC overstates the per-trial signal. Within each participant, a separate AUC averages 0.49 to 0.56 across the subject-specific models, with standard deviations of 0.07 to 0.09 over participants, and stays close to the 0.50 chance level. For SVM it sits well below the aggregate value of 0.640. The simplest reading is that the aggregate AUC mostly tracks how often each participant remembered clips, not a per-trial memory signal. Per-participant AUC

ranges from 0.31 to 0.73, but with only 75 to 136 trials each, much of that spread is small-sample noise.

### 3.2. Which Features Are Selected

We analyze the four tree and linear classifiers. SVM gives no per-feature importance, so we leave it out. An abundant category is selected often even when nothing favors it, so raw counts mislead. We instead report an *enrichment ratio* for each category $c$,

$$E_c = \frac{s_c/S}{b_c/B},$$

where $s_c$ and $S$ are the numbers of selected features in $c$ and in total, and $b_c$ and $B$ the corresponding counts in the whole bank. A value of one means selection in proportion to availability.

Selection concentrates on a few categories. Alpha-adapted power, the band power recomputed around each participant's own alpha peak, is selected at 4.4 times its share of the bank, whereas plain band power is selected in proportion to its size (0.99). The alpha family and low beta are also favored (2.2 and 1.8), along with the early post-stimulus window (2.0) and left temporal sites (2.3 to 2.9). Theta, delta, gamma, the later windows, and right temporal sites all fall at or below 1. These preferences hold in every model-and-scope combination we examined.

What gets selected still depends on the classifier and the scope. Each model keeps 150 features, and any two models share far more of them than the roughly two expected by chance. Even so, the linear and tree models overlap the least, and one model keeps only a third to a half of its features when the scope changes. There is a common core, but no single feature set. Selection frequency is also not the same as contribution. Added in stability order, the cross-validated score climbs from 0.560 at one feature to 0.652 at 150 and then stays flat, so accuracy builds across more than a hundred features rather than a few. The selected features are associated with recall, but their rank does not measure their importance.

### 3.3. Ablation Study of Preprocessing

We re-ran the subject-specific stacking pipeline under three lighter cleaning settings. Aggregate AUC stayed between 0.676 and 0.682, and the full pipeline reached 0.670, with within-participant AUC between 0.53 and 0.59 throughout. The spread across participants was larger than any gap between settings, so the cleaning steps made no difference we can separate from noise.

We keep ICA and AutoReject anyway. They remove clearly ocular and muscular components and make recordings more comparable across participants, which the AUC does not capture. The limit here is the weak per-trial signal, not residual artifact.

## 4. Conclusion

We built a large hand-crafted EEG feature bank with per-classifier stability selection and stacking, and evaluated it under three training scopes. Our best run reached test AUC 0.607 with a subject-specific SVM, while cross-subject training fell to chance. The within-participant analysis is the main lesson, since decoding stays near chance inside each participant and the aggregate AUC mostly reflects differences between participants rather than a per-trial memory signal. Progress on this task depends on strengthening that within-participant signal, not on further feature or model engineering.

## Declaration on Generative AI

During the preparation of this work, the author(s) used Claude Opus 4.8 in order to: Grammar and spelling check, Paraphrase and reword. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

## References


[1] I. Martín-Fernández, A. Ganesh, M. G. Constantin, C.-H. Demarty, M. Gil-Martín, S. Halder, B. Ionescu, A. Matran-Fernandez, R. Savran Kiziltepe, A. García Seco de Herrera, Overview of the mediaeval 2026 predicting movie and commercial memorability task, in: Proc. of the MediaEval 2026 Workshop, Amsterdam, The Netherlands and Online, 2026.

[2] R. Cohendet, K. Yadati, N. Q. K. Duong, C.-H. Demarty, Annotating, understanding, and predicting long-term video memorability, in: Proceedings of the 2018 ACM International Conference on Multimedia Retrieval (ICMR), Yokohama, Japan, 2018, pp. 178–186.

[3] I. Martín-Fernández, M. Lobo-Alonso, S. Esteban-Romero, M. Gil-Martín, F. Fernández-Martínez, Exploring movie recall prediction using functional descriptors of the eeg signal, in: Working Notes Proceedings of the MediaEval 2025 Workshop, CEUR Workshop Proceedings, 2025. URL: https://2025.multimediaeval.com/paper4.pdf.

[4] L. Sweeney, A. Matran-Fernandez, S. Halder, A. García Seco de Herrera, A. Smeaton, G. Healy, Overview of the eeg pilot subtask at mediaeval 2021: Predicting media memorability, in: Working Notes Proceedings of the MediaEval 2021 Workshop, volume 3181 of *CEUR Workshop Proceedings*, 2021. URL: https://ceur-ws.org/Vol-3181/paper16.pdf.

[5] R. Kleinlein, E. Rodríguez Sebastián, F. Fernández-Martínez, Understanding media memorability from event-related potential records and visual semantics, in: Working Notes Proceedings of the MediaEval 2022 Workshop, volume 3583 of *CEUR Workshop Proceedings*, 2023. URL: https://ceur-ws.org/Vol-3583/paper35.pdf.

[6] S. M. T. Mariappan, M. Ramasamy, B. Arul, MediaEval 2025: A multimodal approach for predicting movie and commercial memorability using stacking and gradient boosting, in: Working Notes Proceedings of the MediaEval 2025 Workshop, CEUR Workshop Proceedings, 2025. URL: https://2025.multimediaeval.com/paper7.pdf.

[7] M. Jas, D. A. Engemann, Y. Bekhti, F. Raimondo, A. Gramfort, Autoreject: Automated artifact rejection for MEG and EEG data, NeuroImage 159 (2017) 417–429. doi:10.1016/j.neuroimage.2017.06.030.

[8] T.-W. Lee, M. Girolami, T. J. Sejnowski, Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources, Neural Computation 11 (1999) 417–441. doi:10.1162/089976699300016719.

[9] L. Pion-Tonachini, K. Kreutz-Delgado, S. Makeig, ICLabel: An automated electroencephalographic independent component classifier, dataset, and website, NeuroImage 198 (2019) 181–197. doi:10.1016/j.neuroimage.2019.05.026.

[10] N. Meinshausen, P. Bühlmann, Stability selection, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72 (2010) 417–473. doi:10.1111/j.1467-9868.2010.00740.x.