

Evaluation of Multimodal Face and Voice Classification of Emotion-Derived Negative-Affect Labels

Chandra Prakash Singar¹ Puja Gupta² Meet Lahori³ Krishna Saxena⁴ Shaivi Barwe⁵ Neha Agrawal⁶

Department of Information Technology Shri G.S. Institute of Technology and Science, Indore , India

Abstract- This paper presents an end-to-end prototype that classifies emotion-derived negative-affect labels from facial video and speech audio. Clips from two acted-emotion corpora, CREMA-D and RAVDESS, are combined and mapped to a binary target: angry, disgust, and fearful clips are labeled 1, and happy, neutral, and calm clips are labeled 0. Because both corpora annotate acted expression rather than physiological or self-reported stress, the target is a proxy, and the results describe classification of that proxy. A facial branch (ResNet18 embeddings with a multilayer perceptron) and a voice branch (240 statistics derived from 40 MFCCs and their deltas, with logistic regression) are evaluated separately and under weighted late fusion (0.40 face, 0.60 voice). Splits are grouped by actor. On a held-out test set of 743 clips, late fusion achieved 78.33% accuracy, 82.85% F1-score, 86.44% recall, and 0.8522 ROC-AUC, compared with 76.04% accuracy for the face branch and 67.29% for the voice branch. The fusion gain over the face branch is clear on the CREMA-D test subset but reverses on the RAVDESS subset, and the accuracy difference between fusion and face-only is not established as significant by the analyses reported here. The system also displays outputs from two pretrained visible-expression models. These outputs are not stress evidence, and the application's expression-priority rule is not evaluated in this paper.

Keywords: multimodal learning; late fusion; facial features; speech features; MFCC; ResNet18; actor-independent evaluation; emotion-derived labels

I. INTRODUCTION

Emotional state can affect both vocal delivery and visible behavior, which motivates systems that combine more than one signal. Multimodal designs can exploit complementary information from video and audio, but the value of any combination depends on the labels, acquisition conditions, model design, and evaluation protocol. This study builds a prototype with a face encoder, an acoustic feature pipeline, two binary classifiers, and a fusion stage, and evaluates it under conditions chosen to limit leakage between training and testing.

Stress is the motivating concern, but the available labeled corpora do not measure it. CREMA-D [1] and RAVDESS [2] are datasets of acted emotional expression. Accordingly, the study is framed as classification of emotion-derived proxy labels. Its objectives are (i) to compare face-only, voice-only, and fused predictions under an actor-independent split, and

(ii) to report the uncertainty and error structure of these comparisons rather than only headline accuracy.

1. Contributions

- A two-branch baseline using ResNet18 facial embeddings and MFCC-based acoustic statistics, evaluated on an actor-grouped split across CREMA-D and RAVDESS.
- A late-fusion comparison against both unimodal baselines, with error analysis (false positives and false negatives) rather than a single summary metric.
- A description of how the expression-verification interface relates to, and is kept separate from, the measured classifiers.

2. Related Work

CREMA-D contains 7,442 audiovisual clips from 91 actors speaking 12 sentences, rendered in six emotional categories: anger, disgust, fear, happiness, neutral, and sadness [1]. The clips were rated by crowd-sourced

raters in audio-only, visual-only, and audiovisual conditions. RAVDESS contains 7,356 recordings from 24 professional actors, covering speech and song, with audiovisual and audio-only versions [2]. Both corpora were designed to study how emotion is expressed and perceived, not to measure stress.

FERPlus provides crowd-sourced relabelings of facial-expression images and was introduced to address label noise in facial-expression recognition [3, 4]. OpenCV Zoo distributes a seven-class facial-expression model based on a MobileFaceNet backbone [5]. These models classify visible-expression categories; neither measures stress. Performance figures for these pretrained models are not treated as results of the present experiments.

3. Background Of The Study

Emotion recognition using facial expressions and vocal characteristics has emerged as an important research area in affective computing and multimodal machine learning. Human emotions are expressed through multiple behavioral cues, including facial muscle movements, voice pitch, intensity, tone, and speech patterns. While facial expressions provide visual information about affective states, vocal features capture complementary acoustic characteristics. Consequently, combining facial and vocal information may improve emotion classification compared with relying on a single modality.

Deep learning techniques have enabled the extraction of meaningful facial representations from images and video frames. Convolutional neural networks, such as ResNet18, can generate facial embeddings that support the classification of emotion-related patterns. Similarly, Mel-frequency cepstral coefficients (MFCCs) are widely used to represent the spectral characteristics of speech signals and provide useful acoustic features for speech and emotion analysis. Integrating these visual and acoustic representations through multimodal fusion offers a potential approach to capturing complementary information from facial and vocal expressions.

Despite these developments, multimodal emotion classification presents several challenges. Facial expressions and vocal characteristics may convey inconsistent emotional cues, and model performance

can vary across datasets, speakers, and recording conditions. Furthermore, differences in dataset composition and annotation procedures may limit the generalizability of trained models. Evaluation strategies that separate speakers between training and testing are therefore important for assessing performance on previously unseen individuals. Another important consideration is the distinction between emotion recognition and stress detection. Emotion-related labels may represent negative affect, but they do not directly measure psychological stress. Treating emotion-derived categories as proxies requires careful interpretation to avoid unsupported conclusions.

Against this background, the present study evaluates a multimodal face-and-voice classification framework using ResNet18-based facial embeddings, MFCC-based acoustic features, and weighted late fusion. Using emotion-derived negative-affect labels, the study compares facial, vocal, and fused models to examine their classification performance and the potential benefits of multimodal integration. The findings aim to clarify the effectiveness and limitations of combining visual and acoustic information while identifying methodological requirements for improved generalization and reliable interpretation.

II. DATA AND LABEL CONSTRUCTION

1. Datasets

After removing ambiguous labels, the master dataset contains 7,227 clips: 6,171 from CREMA-D and 1,056 from RAVDESS. Across the two corpora, the proxy labels contain 4,389 positive and 2,838 negative samples from 115 actors (91 in CREMA-D and 24 in RAVDESS). Table 1 summarizes the split structure

Table 1. Clip counts by split and corpus

Split	Total	CREMA-D	RAVDESS
Train	5,740	4,948	792
Validation	744	612	132
Test	743	611	132
Total	7,227	6,171	1,056

1. Emotion-To-Proxy Mapping

Angry, disgust, and fearful clips are mapped to label 1. Happy, neutral, and calm clips are mapped to label 0.

Sad and surprise are excluded from the primary binary evaluation because their assignment to either class is ambiguous under this scheme. The mapping is an analytic choice made for this study. It is not a validated psychological or clinical definition of stress. A person may be stressed while appearing happy or neutral, and an angry or fearful expression does not establish stress.

2. Actor-Grouped Split

Splits are grouped by actor, so no actor contributes clips to more than one partition. The recorded clip counts are 5,740 for training, 744 for validation, and 743 for testing; corpus-level clip counts are reported in Table 1. Exact actor counts per partition are not reproduced in the available paper record and should be recovered from the split metadata before an independent replication. The validation set was used for model selection and for choosing the fusion weight. The test set was held out for final evaluation.

The test set contains 450 positive and 293 negative clips, so the positive class makes up 60.6% of the test set. A classifier that predicts positive for every clip would therefore score 0.606 accuracy. This majority baseline is the reference point for all accuracy figures in Section 6.

V. METHODS

1. Facial Branch

YuNet [6] detects faces in sampled video frames. Face crops are normalized with ImageNet statistics [8] and passed to a ResNet18 encoder [7] pretrained on ImageNet, with its final classification layer replaced by an identity mapping. The encoder produces a 512-dimensional embedding per frame; embeddings are mean-pooled across the sampled frames. A multilayer perceptron classifies the pooled vector, and feature scaling is fitted on the training partition only. The exact number of frames sampled per clip and the crop input resolution are not specified in the available experiment record; these implementation details must be recorded from the source code for exact replication.

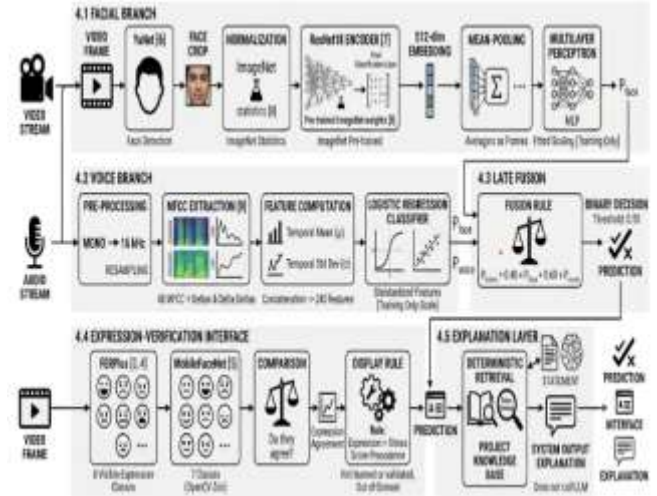


Figure 1: Proposed system working

2. Voice Branch

Audio is converted to mono and resampled to 16 kHz. The pipeline extracts 40 MFCC coefficients [9] together with first-order and second-order deltas. For each of the three 40-dimensional sequences, the temporal mean and standard deviation are computed and concatenated, yielding 240 features. A logistic regression classifier is trained on standardized features, with scaling fitted on the training partition only.

3. Late Fusion

The two branches produce independent probabilities, which are combined as

$$P_{fusion} = 0.40 \times P_{face} + 0.60 \times P_{voice}.$$

On the validation partition, face weights from 0.0 to 1.0 were compared, and 0.40 was selected. The offline binary decision threshold is 0.50.

4. Expression-Verification Interface

The application also runs two pretrained expression models: FERPlus [3, 4], which predicts eight visible-expression classes, and the OpenCV Zoo MobileFaceNet model [5], which predicts seven. The interface displays both outputs and whether they agree. The application contains a display rule that gives clear expression classes precedence over the trained stress-proxy score in certain cases. That rule is an engineering choice made for the interface. It is not a learned or validated classifier, and this paper reports no

performance measurements for it. The expression models were trained on different datasets from the one evaluated here, so their outputs on these clips are out-of-domain.

5. Explanation Layer

A deterministic retrieval component selects statements from a project-specific knowledge base using token overlap. It does not call a generative language model. It explains the system's output and does not alter predictions.

6. Experimental Setup

Feature pipelines were checked for duplicate clip identifiers, non-finite values, metadata mismatches, and split mismatches. Feature records matched all 7,227 expected clips. The selected models were retrained on

the training partition and evaluated once on the 743-clip test set. Reported metrics are accuracy, precision, recall, F1-score, ROC-AUC, average precision, and confusion matrices. For accuracy and recall, approximate 95% confidence intervals are given using the normal (Wald) approximation for a binomial proportion, treating clips as independent. This approximation ignores actor-level correlation and is therefore optimistic. It is reported as an indication of sampling uncertainty, not a formal significance test.

VI. RESULTS

1. Overall Test Performance

Table 2 summarizes overall test performance. Fusion has the highest accuracy and F1-score of the three evaluated models.

Table 2. Overall test results on 743 clips. Majority-baseline values are derived from the class counts; AUC and average precision are not defined for a constant classifier in the same way as for the others.

Model	Accuracy (95% CI)	Precision	Recall	Specificity	F1	ROC-AUC	Avg. Precision
Majority (all positive)	0.606	0.606	1.000	0.000	0.755	0.500	–
Face only	0.760 (0.730–0.791)	0.809	0.791	0.713	0.800	0.837	0.891
Voice only	0.673 (0.639–0.707)	0.704	0.793	0.488	0.746	0.708	0.772
Face + voice	0.783 (0.754–0.813)	0.796	0.864	0.659	0.829	0.852	0.893

Relative to the face branch, fusion increased F1 by 0.0285 and recall by 0.0733. False negatives fell from 94 to 61. However, fusion produced 100 false positives, compared with 84 for the face branch and 150 for the voice branch. Fusion's recall gain was therefore accompanied by more false alarms than the face branch produced, and the two metrics should be read together.

The average-precision difference between fusion and face-only (0.8925 versus 0.8912) is 0.0013. This difference is too small to interpret. The accuracy difference between fusion and face-only is 0.023. The intervals in Table 2 overlap, and the paired comparison that would settle the question requires per-clip predictions, which are not included in the confusion matrices. The overall fusion advantage should therefore be treated as suggestive. Table 3 reports results on the two test subsets separately.

2. Results By Corpus

Table 3. Test results by corpus. Bold marks the best model within each subset.

Subset	Model	Accuracy	F1	ROC-AUC
CREMA-D (611)	Face	0.728	0.774	0.795
CREMA-D (611)	Voice	0.676	0.747	0.722
CREMA-D (611)	Fusion	0.777	0.824	0.836
RAVDESS (132)	Face	0.909	0.923	0.996
RAVDESS (132)	Voice	0.659	0.740	0.695
RAVDESS (132)	Fusion	0.811	0.850	0.952

On the CREMA-D subset, fusion outperforms both unimodal branches on accuracy, F1, and ROC-AUC. On

the RAVDESS subset, the face branch outperforms fusion on all three metrics, and fusion outperforms only the voice branch. The RAVDESS subset contains 132 clips from three test actors. Its face-only results are therefore unstable, and the RAVDESS figures should not be generalized to other populations. The reversal on RAVDESS is an important negative result: the fusion advantage observed overall depends on the CREMA-D subset.

3. Error Structure

Table 4. Confusion matrices on the 743-clip test set.
Rows are true labels (0, 1); columns are predicted labels (0, 1).

Model	Confusion matrix (TN, FP; FN, TP)
Face only	[[209, 84], [94, 356]]
Voice only	[[143, 150], [93, 357]]
Face + voice	[[193, 100], [61, 389]]

Voice-only has the lowest specificity (0.488), so it misclassifies roughly half of the negative clips. Fusion's specificity (0.659) is lower than the face branch's (0.713), reflecting the additional false positives discussed above.

4. Discussion

The results indicate that, on the CREMA-D test subset, facial appearance and acoustic characteristics carry partly complementary information for the proxy classification task. Voice-only performance is lower than face-only on F1 and ROC-AUC. Fusion improves recall and F1 overall, and it does so on CREMA-D, but not on RAVDESS. The evidence for a general fusion benefit is therefore mixed, and the study does not establish it. Actor-grouped splitting reduces the risk of identity leakage, which occurs when clips from the same speaker appear in both training and test partitions. Evaluating on two corpora provides a partial check on generalization, but both corpora share an acted-emotion design. Performance on naturally occurring stress would require different data and labels. The expression-verification interface makes visible-expression predictions inspectable. It does not make them valid indicators of stress. A rule of the form "happy or neutral means no stress" is not scientifically supported, because expression is context-dependent and the expression models recognize appearance

categories rather than physiological or psychological states. Accordingly, the interface is presented as a display and quality-control aid, and its decision rule is excluded from the paper's claims.

5. Limitations And Responsible Use

Reproducibility details. The manuscript record does not specify the exact frame-sampling count, face-crop input resolution, actor counts per split, software versions, random seeds, or model hashes. These details should be recovered from the implementation and experiment logs before an exact replication is attempted.

- Proxy labels. The targets are derived from acted-emotion categories. They are not physiological measurements, validated questionnaires, or stress annotations.
- Expression-to-stress mismatch. Visible emotion cannot establish internal stress. The interface's display rule is an engineering choice and is not clinical evidence.
- Generalization. Acted speech, a limited number of actors, recording conditions, language, culture, lighting, pose, and demographic composition may all affect performance. The study does not include cross-corpus training and testing, which would be the most direct check of generalization.
- Small RAVDESS test subset. Its 132 clips from three actors limit confidence in any dataset-specific conclusion.
- Uncertainty. The confidence intervals assume independent clips and are optimistic. Results from a single split and a single random seed do not capture the variability across splits.
- Calibration. Model scores are not calibrated and should not be interpreted as probabilities of stress.
- No clinical validation. This prototype must not be used for diagnosis, employment decisions, academic discipline, or other consequential decisions.
- Privacy and consent. Any use of recorded faces or voices requires informed consent, minimal retention, secure storage, and clear disclosure of what is processed.
- Licensing. The licenses of the datasets and pretrained models carry attribution and redistribution conditions that must be checked before release.

VII. CONCLUSION AND FUTURE WORK

This research evaluated a multimodal face-and-voice framework integrating ResNet18 facial embeddings, MFCC-based acoustic features, and weighted late fusion for emotion-based classification. On an actor-grouped test set of 743 clips, the fusion model achieved 78.33% accuracy, 82.85% F1-score, 86.44% recall, and a ROC-AUC of 0.8522, yielding the strongest overall results among the three evaluated models. However, its performance varied across datasets, outperforming the face-only model on CREMA-D but underperforming on RAVDESS. Therefore, the results do not establish a statistically significant overall improvement over the face-only approach.

A key limitation is that the target labels represent emotion-derived proxies rather than directly measured stress. Consequently, the findings support emotion-related classification but do not validate stress detection. Future research should prioritize cross-corpus evaluation, actor-grouped cross-validation with repeated random seeds, and statistical significance testing using McNemar's test and actor-level bootstrap confidence intervals. Comparisons with stronger unimodal baselines and sensitivity analyses of alternative emotion-to-proxy mappings are also necessary. Furthermore, expression-based decision rules should be validated on independently labeled datasets. Ultimately, reliable stress-related conclusions require datasets containing independently annotated stress labels, physiological measurements, or validated self-report assessments. These improvements would strengthen the robustness, generalizability, and scientific validity of the proposed multimodal framework.

1. Acknowledgments

No funding or additional acknowledgments information was supplied for this draft. The authors should amend this statement if financial support, institutional support, or other contributions need to be acknowledged.

REFERENCES

1. H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014. doi: 10.1109/TAFFC.2014.2336244.
2. S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS ONE*, vol. 13, no. 5, e0196391, 2018. doi: 10.1371/journal.pone.0196391.
3. E. Barsoum, C. Zhang, C. Canton Ferrer, and Z. Zhang, "Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution," in *Proc. ACM International Conference on Multimodal Interaction (ICMI)*, 2016. arXiv:1608.01041.
4. Microsoft, "FERPlus: New label annotations for the Emotion FER dataset," GitHub repository, <https://github.com/microsoft/FERPlus> (accessed October 2026).
5. OpenCV, "OpenCV Zoo: Facial Expression Recognition" (Progressive Teacher model, MobileFaceNet backbone), model repository, https://github.com/opencv/opencv_zoo (accessed October 2026).
6. W. Wu, H. Peng, and S. Yu, "YuNet: A Tiny Millisecond-level Face Detector," *Machine Intelligence Research*, vol. 20, no. 5, pp. 656–665, 2023.
7. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
8. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proc. IEEE CVPR*, 2009.
9. S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

10. Gupta, P., Sharma, V. and Varma, S., 2021. People detection and counting using YOLOv3 and SSD models. *Materials Today: Proceedings*.
11. Gupta, P., Shukla, M., Arya, N., Singh, U. and Mishra, K., 2022. Let the Blind See: An AlloT-Based Device for Real-Time Object Recognition with the Voice Conversion. In *Machine Learning for Critical Internet of Medical Things* (pp. 177-198). Springer, Cham.
12. Kushwaha, U., Gupta, P., Airen, S. and Kuliha, M., 2022, December. Analysis of CNN Model with Traditional Approach and Cloud AI based Approach. In *2022 International Conference on Automation, Computing and Renewable Systems (ICACRS)* (pp. 835-842). IEEE
13. Gupta, P., Sharma, V. and Varma, S., 2022. A novel algorithm for mask detection and recognizing actions of human. *Expert Systems with Applications*, 198, p.116823.
14. Gupta, P., Sharma, V. and Varma, S., 2022, September. An Algorithm for Counting People using Dense Nets and Feature Fusion. In *2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA)* (pp. 1248-1253). IEEE.
15. Gupta, P. and Kulkarni, N., 2013. An introduction of soft computing approach over hard computing. *International Journal of Latest Trends in Engineering and Technology (IJLTET)*, 3(1), pp.254-258
16. Gupta, P., Kulkarni, A. and Sarda, A., 2013. An embedded health care supervisory systems. *International Journal of Latest Trends in Engineering and Technology (IJLTET)*, 3, pp.379-386
17. Jain Monarch, Singar Chandra Prakash , Gupta P, Toxic Speech Classification Using Deep Learning and Machine Learning: A Survey. *Journal of Propulsion Technology* article 10400
18. Agrawal Neha, Gupta Puja, Barwe Shaivi, Singar Chandra Parakash, Agrawal Deepesh, ESP32-S3 AI Voice Assistant: A Cloud-Augmented Embedded System for Real-Time Voice Interaction. *International Journal of Leading Research Publication* 2203.
19. Gupta, P, N. Rathore Comparison of cloud and edge computing based face recognition for security enhancement 2025 *Journal of Discrete Mathematical Sciences and Cryptography*