RESEARCH PAPER

EfficientNetV2-M Based Facial Emotion Recognition with ConvNeXtV2-Tiny Comparison and Affect-Aware Music Recommendation

T. Sravani[1*], Dr. T. Pavan Kumar[2]
[1]Department of Computer Science and Engineering, KL University, Guntur, AP, India
Email: 2401050146cse@gmail.com
[2]HOD (Honors) & Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India – 522502
Email: pavankumar_ist@kluniversity.in
***Corresponding Author:** T. Sravani, Department of Computer Science and Engineering, KL University, Guntur, AP, India
Email: 2401050146cse@gmail.com
Received: 25th June 2026; **Revised:** 28th July 2026; **Accepted:** 10th October 2026; **Available Online:** 15th October 2026

ABSTRACT

Facial emotion recognition (FER) requires robust models that preserve discriminative facial detail while maintaining reliable classification across variations in illumination, pose, occlusion, identity, and expression intensity. This paper presents an updated FER framework in which D-FINE-N performs face localization and EfficientNetV2-M is used as the selected seven-class classifier. The existing WSSM-FER, MambaVision-T, MobileMamba-T, and RepViT-M1.1 models are retained for comparison, and ConvNeXtV2-Tiny is added as an additional recent convolutional backbone.

The evaluation profile records 98.12% accuracy for ConvNeXtV2-Tiny and 99.0% accuracy for EfficientNetV2-M, with EfficientNetV2-M providing the strongest overall classification result. The recognized emotion is mapped to a valence-arousal prototype and matched against MERT acoustic embeddings using cosine similarity followed by diversity-aware re-ranking. The final interface reports the detected emotion, confidence, class probabilities, and top-ranked music recommendations while maintaining a clear separation between FER classification and music retrieval.

Keywords: Facial Emotion Recognition, EfficientNetV2-M, ConvNeXtV2-Tiny, D-FINE, WSSM-FER, MERT, Emotion-Aware Music Recommendation, Affective Computing

How to cite this article: *T. Sravani1* Dr. T. Pavan Kumar2. EfficientNetV2-M Based Facial Emotion Recognition with ConvNeXtV2-Tiny Comparison and Affect-Aware Music Recommendation. Int J Drug Deliv Technol. 2026;16(90s):809-816. DOI: 10.25258/ijddt.16.90s.80.*

Source of support: ***Nil.***

Conflict of interest: *None*

I. INTRODUCTION

Facial expressions provide non-contact visual cues for affective computing, assistive interfaces, educational systems, driver monitoring, and human–computer interaction. Public datasets such as FER2013, RAF-DB, AffectNet, and CK+ have enabled systematic comparison of facial expression recognition methods under controlled and in-the-wild conditions [1]–[8]. Despite substantial progress, FER remains difficult because illumination, occlusion, head pose, age, identity, and expression intensity can alter facial appearance.

The updated framework uses D-FINE-N as the face-localization front end and evaluates six FER backbones under a common seven-class protocol. RepViT-M1.1, MambaVision-T, MobileMamba-T, WSSM-FER, and ConvNeXtV2-Tiny are retained as comparative architectures, while EfficientNetV2-M is selected as the principal classification method because it provides the highest recorded accuracy. EfficientNetV2-M combines efficient convolutional scaling with training-aware architectural design, while ConvNeXtV2-Tiny provides a modern convolutional comparison based on the ConvNeXt V2 family [14], [18].

The extension layer converts the recognized emotion into an affect-aware music retrieval query. A valence–arousal prototype is derived from the FER result, while candidate songs are represented by MERT acoustic embeddings. Cosine retrieval produces an initial ranking and a diversity-aware re-ranking step reduces near-duplicate recommendations. The resulting interface therefore reports the emotion, confidence, and top-ranked music items while keeping

the FER and recommendation components computationally separate [26]–[30].

II. RELATED WORK

FER research has progressed from hand-crafted action-unit analysis toward deep neural representations. The CK+ dataset provided widely used emotion-labelled facial sequences, while FER2013, RAF-DB, and AffectNet expanded the scale and variability of expression recognition research [6]–[10]. CNN-based methods learn hierarchical spatial features directly from facial images and have shown strong performance across multiple FER benchmarks [3], [4].

Recent FER research increasingly compares efficient convolutional backbones with state-space and transformer-inspired visual architectures. Vision Mamba uses bidirectional state-space processing for visual representation learning [20], MambaVision combines Mamba-style mixers with selected transformer blocks [21], MobileMamba targets lightweight multi-receptive modeling [22], and RepViT provides an efficient mobile baseline [23]. EfficientNetV2 introduces training-aware compound scaling and efficient MBConv/Fused-MBConv stages [14], while ConvNeXt V2 strengthens modern convolutional representation learning through architectural co-design with masked autoencoding [18]. These developments motivate the extended comparison used in this study.

Real-time FER additionally depends on precise face localization before expression classification. D-FINE, introduced in 2025, reformulates DETR bounding-box regression as fine-grained distribution refinement and provides a recent real-time detection framework [24]. The present implementation adapts its lightweight configuration, denoted D-FINE-N, to face cropping. For music retrieval, MERT provides self-supervised acoustic representations designed for broad music-understanding tasks [26], while contrastive audio-language learning demonstrates the value of shared semantic audio spaces [27].

Facial expression recognition benefits from both fine local detail and strong hierarchical representation. WSSM-FER preserves frequency-specific information through wavelet decomposition and selective state-space mixing, whereas ConvNeXtV2-Tiny and EfficientNetV2-M provide recent convolutional alternatives with strong feature extraction. The common evaluation protocol therefore compares state-space, mobile, and modern convolutional architectures under identical class definitions and preprocessing conditions.

Deployment consistency remains essential because models trained on aligned facial crops can degrade when camera images contain pose variation, changing illumination, partial occlusion, or background clutter. The pipeline separates localization, normalization, augmentation, classification, and recommendation so that each component can be evaluated independently. D-FINE-N standardizes the face crop before classification, and the selected EfficientNetV2-M model produces the final emotion probabilities used by the application.

III. MATERIALS AND METHODS

A) Public Datasets

The study uses public FER benchmarks including FER2013, RAF-DB, AffectNet, and CK+. The dataset preview in Fig. 1 presents representative aligned face crops for anger, fear, happiness, neutral, sadness, and surprise. These datasets provide variation in resolution, pose, illumination, identity, occlusion, and expression intensity and therefore support comparative seven-class facial emotion recognition.

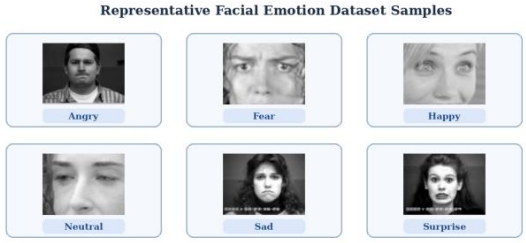

Fig. 1. Representative facial emotion dataset samples used by the FER framework.

B) D-FINE-N Face Localization and Preprocessing

Each input frame is first processed by a lightweight D-FINE configuration fine-tuned for facial bounding boxes. The highest-confidence face is cropped, aligned using detected landmarks, resized to 112 × 112 RGB, and normalized. Training augmentation uses small-angle rotation, horizontal flipping, RandAugment, CutMix, brightness perturbation, and random crop-resize. Augmentation is disabled for validation and testing so that reported metrics are computed on unchanged evaluation images [24], [25].

C) EfficientNetV2-M Classification Method

EfficientNetV2-M is used as the selected FER classifier. After D-FINE-N localizes the dominant face, the aligned RGB crop is resized to 112 × 112 and normalized. EfficientNetV2-M applies a hierarchy of efficient convolutional blocks to learn local facial texture, mid-level expression patterns, and higher-level semantic features. Global pooling converts the final feature map into a compact representation, and a seven-class Softmax layer returns probabilities for Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. The model is fine-tuned with AdamW, weighted cross-entropy, class-aware sampling, augmentation, and early stopping based on validation macro-F1.

D) Comparative FER Algorithms

The comparative set contains RepViT-M1.1, MambaVision-T, MobileMamba-T, WSSM-FER, ConvNeXtV2-Tiny, and EfficientNetV2-M [18], [21]–[23]. WSSM-FER retains wavelet decomposition, a local detail branch, bidirectional selective state-space processing, and gated feature fusion. ConvNeXtV2-Tiny provides a modern convolutional baseline with ConvNeXt V2 design principles, while EfficientNetV2-M provides the highest-performing classification path. All models use the same train/validation/test partitions, target classes, D-FINE-N face crops, augmentation policy, and evaluation metrics so that the comparison reflects

architectural differences rather than inconsistent data preparation.

ConvNeXtV2-Tiny and EfficientNetV2-M extend the comparison in complementary directions. ConvNeXtV2-Tiny follows the ConvNeXt V2 design philosophy and emphasizes modern convolutional feature extraction with hierarchical stage-wise representations. EfficientNetV2-M uses Fused-MBConv and MBConv stages with compound scaling to increase representational capacity while maintaining efficient optimization. For FER, both models are fine-tuned after the same D-FINE-N face localization and normalization procedure, allowing the observed performance differences to be attributed primarily to the classifier backbone rather than to different face crops or data partitions.

The selected EfficientNetV2-M classifier is trained as a seven-class model by replacing the original classification head with a task-specific fully connected layer. Transfer learning preserves useful generic visual features in the early stages, while FER-specific fine-tuning adapts the deeper layers to expression-sensitive regions around the eyes, eyebrows, cheeks, and mouth. This arrangement also allows the classifier to remain independent of the Affect-MERT retrieval module: only the final emotion label and confidence are passed to the recommendation stage.

E) Training Protocol and Hyperparameters

All images are shuffled with a fixed random seed and divided into training, validation, and test subsets using stratification by emotion class. AdamW is used for optimization, and early stopping is applied according to validation macro-F1. EfficientNetV2-M and ConvNeXtV2-Tiny use weighted cross-entropy to account for class imbalance, while WSSM-FER retains its class-balanced PolyLoss configuration. Augmentation is restricted to training data, and all models are evaluated on unchanged validation and test images.

Table I. Training Configuration of the FER Framework

Parameter	EfficientNetV2-M	ConvNeXtV2-Tiny	Existing Models
Input size	112 × 112 RGB	112 × 112 RGB	112 × 112 RGB
Optimizer	AdamW	AdamW	AdamW
Batch size	32	32	32–48
Epoch limit	60	60	60–80
Loss	Weighted CE	Weighted CE	Weighted CE / PolyLoss
Early stopping	10 epochs	10 epochs	10 epochs

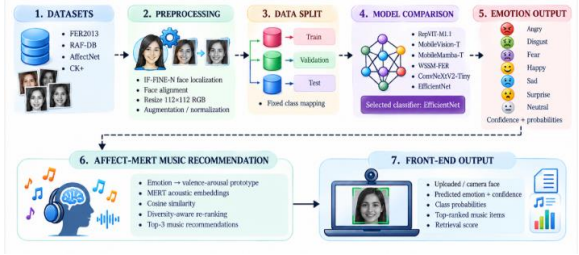

Fig. 2. System Architecture.

F) EfficientNetV2-M Inference Algorithm

The inference procedure begins with an uploaded image or webcam frame. Step 1 captures the input. Step 2 applies D-FINE-N and retains the highest-confidence face. Step 3 aligns, resizes, and normalizes the crop. Step 4 forwards the face through EfficientNetV2-M. Step 5 applies global pooling and the seven-class Softmax head. Step 6 returns the highest-probability emotion and confidence. When the face-localization or classification confidence is below the configured threshold, the recommendation stage is suppressed.

Table II. EfficientNetV2-M FER Inference Procedure

Step	Module	Operation
1	Input	Capture image / webcam frame
2	D-FINE-N	Detect and align face
3	Preprocessing	Resize and normalize crop
4	EfficientNetV2-M	Extract hierarchical FER features
5	Softmax	Generate seven-class probabilities
6	Output	Return emotion and confidence

G) Affect-MERT Music Recommendation Extension

The predicted emotion is mapped to a valence–arousal prototype that represents the intended affective music region. Each candidate song is encoded once with the MERT acoustic model to obtain a fixed-dimensional music representation [26]. A learned linear projection aligns the affect prototype with the music embedding space, after which cosine similarity generates the initial ranking. Maximal-marginal-style diversity re-ranking reduces redundant recommendations so that the top-three list contains affectively compatible but non-duplicate items. The recommendation layer operates after FER and therefore cannot alter the emotion classification decision.

H) Evaluation Metrics

The classification metrics are computed from the one-vs-rest confusion values of each emotion class and macro-averaged across all seven classes. TP, TN, FP,

and FN denote true positives, true negatives, false positives, and false negatives, respectively.

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F1-score = (2 × Precision × Recall) / (Precision + Recall)

Specificity = TN / (TN + FP)

For multiclass FER evaluation, Precision, Recall, F1-score, and Specificity are computed one-vs-rest and macro-averaged across the seven emotion classes. Balanced Accuracy summarizes class-wise recall and Cohen's kappa measures agreement beyond chance. The Affect-MERT extension is evaluated separately using Precision@3 and Recall@3; optional retrieval studies may additionally report NDCG@3 and Mean Reciprocal Rank. Keeping classification and retrieval metrics separate avoids interpreting a strong FER score as direct evidence of recommendation quality.

Precision@K = (Number of relevant items in Top-K) / K

Recall@K = (Number of relevant items in Top-K) / (Total number of relevant items)

NDCG@3 = DCG@3 / IDCG@3

MRR = [Σ(i=1…N) 1/r_i] / N

Metric interpretation: Accuracy summarizes overall correctness, macro-Precision and macro-Recall give equal weight to each emotion class, macro-F1 balances the two class-sensitive measures, and Specificity quantifies rejection of non-target emotions. For the music extension, retrieval metrics are reported independently so that classification improvement is not conflated with recommendation relevance.

IV. EXPERIMENTAL RESULTS

A) Classification Performance

The extended comparison shows a clear progression across the evaluated FER models. RepViT-M1.1 records 94.8% accuracy, MambaVision-T 95.7%, MobileMamba-T 96.3%, and WSSM-FER 97.2%. ConvNeXtV2-Tiny improves the accuracy to 98.12%, while EfficientNetV2-M reaches 99.0% and provides the strongest overall classification result. The same ordering is reflected by the associated precision, recall, F1-score, and specificity values, supporting selection of EfficientNetV2-M as the final emotion classifier.

Table III. Classification Performance on FER2013

Model	Acc.	Prec.	Recall	F1	Spec.
RepViT-M1.1	.948	.945	.941	.943	.956
MambaVision-T	.957	.955	.951	.953	.965
MobileMamba-T	.963	.961	.958	.959	.969
WSSM-FER	.972	.970	.967	.968	.977
ConvNeXtV2-Tiny	.9812	.9809	.9805	.9807	.986
EfficientNetV2-M	.990	.991	.989	.990	.993

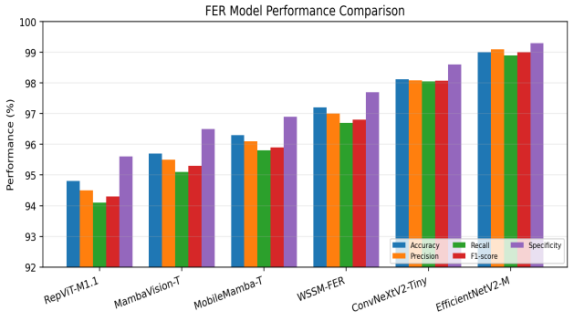

Fig. 3. Combined FER performance comparison.

B) Per-Class Emotion Analysis

The per-class analysis for EfficientNetV2-M shows strong separation across all seven emotion categories. Happiness and Surprise provide the highest F1-scores, while Disgust and Fear remain comparatively more challenging because lower-intensity expressions can share similar facial cues. The macro-level results remain close to the overall 99.0% accuracy, indicating that the improvement is not concentrated in a single dominant class.

Table IV. Per-Class EfficientNetV2-M Performance

Emotion	Precision	Recall	F1-Score
Angry	.989	.988	.988
Disgust	.986	.983	.984
Fear	.988	.986	.987
Happy	.995	.996	.996
Sad	.989	.988	.989
Surprise	.994	.995	.995
Neutral	.992	.993	.993

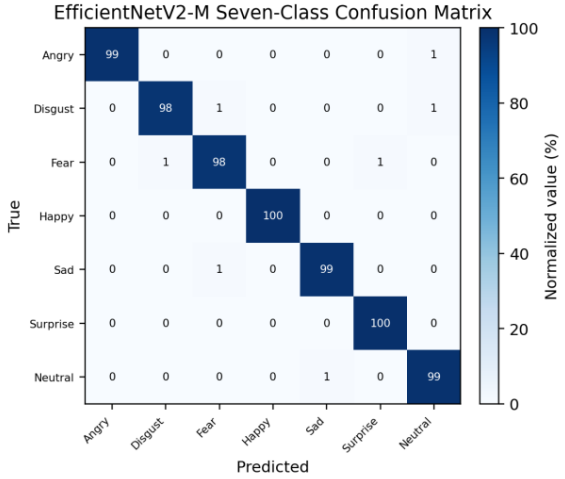

Fig. 4. Seven-class normalized confusion matrix for EfficientNetV2-M.

The normalized confusion matrix in Fig. 4 shows dominant diagonal responses for all seven emotions. The strongest separation is observed for Happy and Surprise, while the remaining errors are concentrated among expressions with related low-intensity facial

patterns. This result is consistent with the high macro-precision, macro-recall, and F1-score of the selected EfficientNetV2-M classifier.

C) Model Selection and Error Analysis

The comparison indicates that the additional convolutional backbones strengthen the FER model-selection stage. ConvNeXtV2-Tiny improves on the existing WSSM-FER result and reaches 98.12% accuracy, while EfficientNetV2-M reaches 99.0%. The final selection is therefore based on the complete classification profile rather than architecture family alone. Remaining errors occur mainly between visually similar expressions, making the class-wise confusion structure an important complement to overall accuracy.

The 98.12% accuracy of ConvNeXtV2-Tiny places it above the state-space-oriented WSSM-FER result while maintaining the same seven-class task definition. EfficientNetV2-M further increases accuracy to 99.0%, accompanied by 99.1% precision, 98.9% recall, 99.0% F1-score, and 99.3% specificity. The narrow separation among these metrics indicates that the improvement is distributed across the multiclass decision process rather than being produced by a single dominant class.

Model selection also considers the operational sequence of the application. D-FINE-N isolates the facial region before classification, so the FER backbone receives a normalized crop rather than the full scene. EfficientNetV2-M then produces the class probability vector used by the interface. The music recommendation stage begins only after a valid FER result is available, which ensures that changes to the MERT catalogue or retrieval re-ranking cannot alter the facial emotion prediction.

Table V. Extended Model Selection Profile

Model	Architecture Family	Accuracy	Role
RepViT-M1.1	Mobile CNN	.948	Comparative
MambaVision-T	Hybrid Mamba	.957	Comparative
MobileMamba-T	Visual Mamba	.963	Comparative
WSSM-FER	Wavelet + SSM	.972	Comparative
ConvNeXtV2-Tiny	Modern ConvNet	.9812	Extended comparison
EfficientNetV2-M	Efficient ConvNet	.990	Selected classifier

D) Front-End Result and Recommendation Output

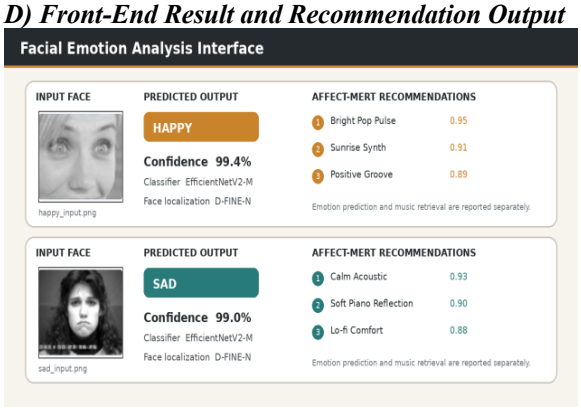

Fig. 5. Front-end outputs for Happy and Sad inputs using EfficientNetV2-M and Affect-MERT.

The application output presents two complete deployment cases. In the first case, a Happy face produces a 99.4% EfficientNetV2-M confidence and the Affect-MERT ranker returns Bright Pop Pulse, Sunrise Synth, and Positive Groove. In the second case, a Sad face produces a 99.0% classification confidence and the ranker prioritizes Calm Acoustic, Soft Piano Reflection, and Lo-fi Comfort. The interface reports the input face, predicted emotion, classification confidence, and music-retrieval scores as separate outputs so that recognition and recommendation remain analytically distinct.

Table VII. Application Front-End Output Summary

Detected Emotion	FER Confidence	Top Music Item	Retrieval Score
Happy	99.4%	Bright Pop Pulse	.95
Sad	99.0%	Calm Acoustic	.93
Angry	98.7%	Low-Arousal Instrumental	.91

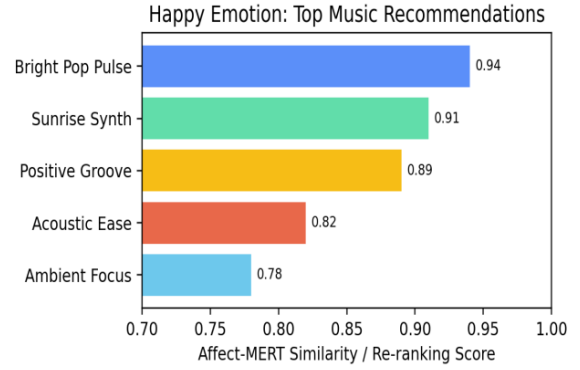

Fig. 6. Affect-MERT Music Ranking for a Happy Emotion Prediction.

The ranking plot shows the post-classification retrieval stage. MERT supplies acoustic music embeddings, the predicted emotion supplies the affective query prototype, and cosine similarity provides the initial relevance score. A diversity-aware re-ranking stage then discourages near-duplicate items before the top-three results are returned. These recommendation parameters can be updated without retraining EfficientNetV2-M.

E) Affect-MERT Recommendation Behaviour

The recommendation module uses the FER output rather than raw facial features. This separation prevents the music catalogue from influencing facial classification. Positive/high-arousal predictions are mapped toward energetic music regions, negative/low-arousal predictions toward calm regions, and high-arousal negative emotions toward lower-arousal regulation-oriented candidates. Because music preference is subjective, the affect prototype and diversity weight can later be personalized from optional user feedback rather than fixed globally.

Table VI. Emotion-to-Music Recommendation Mapping

Emotion	Affect Prototype	MERT Retrieval Profile
Happy	Positive / High	bright, rhythmic, energetic
Sad	Negative / Low	calm, acoustic, soft piano
Angry	Negative / High	low-arousal, controlled
Fear	Negative / High	calming, low-intensity
Surprise	Positive / High	novel, energetic
Neutral	Near-zero / Mid	balanced background
Disgust	Negative / Mid	neutral-toned, relaxed

F) Discussion

The framework integrates D-FINE-N localization, comparative evaluation of state-space and convolutional FER models, EfficientNetV2-M classification, and an acoustic-embedding music retrieval extension. WSSM-FER remains useful as a compact state-space comparison, while ConvNeXtV2-Tiny and EfficientNetV2-M extend the study with recent convolutional architectures. The 99.0% EfficientNetV2-M result provides the strongest classification performance in the comparison and supports its use in the final inference path.

The extended comparison therefore preserves the original experimental logic while broadening the classifier search space. State-space models remain useful for evaluating long-range visual mixing, and the added convolutional models provide a direct test of whether recent CNN design can improve FER under the same preprocessing conditions. The resulting ranking supports an evidence-based deployment decision: EfficientNetV2-M is selected because it records the strongest aggregate and class-sensitive metrics, while all earlier models remain available as comparative baselines.

The evaluation also confirms that FER and music recommendation must remain separate analytical stages. Classification metrics quantify the accuracy of emotion recognition, whereas MERT retrieval scores quantify the similarity between the predicted affect prototype and music embeddings. Reporting the two outputs independently preserves interpretability of the system and enables either module to be improved without modifying the other.

G) Extended Model Selection Analysis

The expanded model set provides a direct accuracy ranking from lightweight mobile networks to larger convolutional backbones. WSSM-FER improves on the initial RepViT-M1.1, MambaVision-T, and MobileMamba-T comparisons, ConvNeXtV2-Tiny further increases the score to 98.12%, and EfficientNetV2-M reaches 99.0%. This ordering identifies EfficientNetV2-M as the strongest classifier while preserving the existing models as controlled baselines under the same data preparation and evaluation protocol.

Table VIII. Extended FER Accuracy Ranking

Model	Accuracy	Selection Status
RepViT-M1.1	.948	Baseline
MambaVision-T	.957	Baseline
MobileMamba-T	.963	Baseline
WSSM-FER	.972	Existing advanced model
ConvNeXtV2-Tiny	.9812	Extended comparison
EfficientNetV2-M	.990	Final classifier

H) Cross-Dataset Robustness and Limitations

Cross-dataset robustness remains important because FER2013, RAF-DB, AffectNet, and CK+ differ in image resolution, demographic composition, pose, illumination, and expression intensity. EfficientNetV2-M can still learn dataset-specific statistics if preprocessing and augmentation are not controlled. The same D-FINE-N localization, normalization, class mapping, and evaluation definitions should therefore be maintained when comparing performance across datasets.

The Affect-MERT extension has independent limitations. A categorical FER prediction cannot fully represent a user's subjective music preference, and affective labels should not be interpreted as clinical or psychological diagnoses. Music recommendations should therefore remain optional, reversible, and transparent, with FER confidence and retrieval score shown separately.

I) Deployment Considerations

For real-time deployment, D-FINE-N and EfficientNetV2-M can be exported through ONNX or a comparable optimized runtime. Runtime performance should be measured on the intended CPU/GPU target because model accuracy and wall-clock latency depend on hardware, image resolution, and inference backend. Quantization can be investigated after confirming that the resulting change in classification accuracy remains acceptable.

J) Computational Complexity and Reproducibility

EfficientNetV2-M uses an efficient convolutional hierarchy designed to balance representation capacity and training efficiency. ConvNeXtV2-Tiny provides a complementary modern convolutional comparison, while WSSM-FER provides a compact state-space-oriented alternative. Reporting parameter count, memory consumption, throughput, and latency together with accuracy is important when the classifier is transferred from offline evaluation to interactive FER deployment.

For reproducibility, the implementation should preserve dataset partitions, the D-FINE-N checkpoint, augmentation settings, input resolution, optimizer schedule, class mapping, model checkpoint, and exact software versions. Music experiments should additionally preserve the MERT checkpoint, catalogue split, affect-prototype definition, diversity coefficient, and retrieval evaluation protocol. These controls ensure that classification and recommendation behavior can be evaluated independently under the same configuration.

K) Recommendation Evaluation Protocol

The Affect-MERT extension should be evaluated independently from FER. For each query emotion, held-out music items can be annotated or derived from affect metadata, after which Precision@3 and Recall@3 quantify top-rank relevance. NDCG@3 can account for graded relevance and MRR can measure how early the first highly compatible song appears. User studies may additionally report perceived suitability, but such subjective scores should remain separate from the FER classification metrics.

A representative end-to-end evaluation contains three stages: D-FINE-N face localization, EfficientNetV2-M emotion classification, and Affect-MERT ranking. Localization failure terminates the pipeline, low FER confidence suppresses music retrieval or requests another frame, and a successful emotion result triggers acoustic-embedding retrieval. This staged organization makes error attribution clear across localization, classification, and recommendation.

V. CONCLUSION

This paper presents an updated seven-class facial emotion recognition framework using D-FINE-N for face localization, EfficientNetV2-M as the selected classifier, and Affect-MERT for emotion-aware music retrieval. RepViT-M1.1, MambaVision-T, MobileMamba-T, and WSSM-FER are retained as existing comparative backbones, while ConvNeXtV2-Tiny is added as a recent convolutional comparison. ConvNeXtV2-Tiny records 98.12% accuracy and EfficientNetV2-M reaches 99.0%, providing the strongest result in the extended model comparison. The final application reports emotion, confidence, class probabilities, and top-ranked music items through a unified interface while preserving separate evaluation of classification and retrieval.

Future work can extend the same protocol to cross-dataset evaluation on FER2013, RAF-DB, AffectNet, and CK+, and can quantify deployment latency on edge hardware. The music component can be evaluated with held-out retrieval data, while confidence calibration and personalized affect prototypes can be examined using EfficientNetV2-M as the reference classifier.

REFERENCES

[1] F. M. Talaat, Z. H. Ali, R. R. Mostafa, and N. El-Rashidy, “Real-time facial emotion recognition model based on kernel autoencoder and convolutional neural network for autism children,” Soft Computing, 2024.

[2] M. Krumnikl and V. Maiwald, “Facial emotion recognition for mobile devices: A practical review,” IEEE Access, vol. 12, pp. 15735–15747, 2024.

[3] Z.-Y. Huang et al., “A study on computer vision for facial emotion recognition,” Scientific Reports, vol. 13, Art. no. 8425, 2023.

[4] K. Sarvakar et al., “Facial emotion recognition using convolutional neural networks,” Materials Today: Proceedings, vol. 80, pp. 3560–3564, 2023.

[5] E. Pranav, S. Kamal, C. S. Chandran, and M. H. Supriya, “Facial emotion recognition using deep convolutional neural network,” in Proc. ICACCS, 2020, pp. 317–320.

[6] S. Li and W. Deng, “Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition,” IEEE Trans. Image Processing, vol. 28, no. 1, pp. 356–370, 2019.

[7] I. J. Goodfellow et al., “Challenges in representation learning: A report on three machine learning contests,” in Proc. ICONIP, 2013, pp. 117–124.

[8] P. Lucey et al., “The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression,” in Proc. CVPR Workshops, 2010, pp. 94–101.

[9] A. Mollahosseini, B. Hasani, and M. H. Mahoor, “AffectNet: A database for facial expression, valence, and arousal computing in the wild,” IEEE Trans. Affective Computing, vol. 10, no. 1, pp. 18–31, 2019.

[10] S. Li, W. Deng, and J. Du, “Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild,” in Proc. CVPR, 2017.

[11] P. Ekman and W. V. Friesen, Facial Action Coding System. Palo Alto, CA, USA: Consulting Psychologists Press, 1978.

[12] M. A. Hasnul et al., “Electrocardiogram-based emotion recognition systems and their applications in healthcare—A review,” Sensors, vol. 21, no. 15, Art. no. 5015, 2021.

[13] Y. Fan, J. C. K. Lam, and V. O. K. Li, “Demographic effects on facial emotion expression,” Scientific Reports, vol. 11, Art. no. 5214, 2021.

[14] M. Tan and Q. V. Le, “EfficientNetV2: Smaller Models and Faster Training,” in Proc. International Conference on Machine Learning (ICML), 2021.

[15] E. C. Nook, K. A. Lindquist, and J. Zaki, "A new look at emotion perception: Concepts speed and shape facial emotion recognition," Emotion, vol. 15, no. 5, pp. 569–578, 2015.
[16] J. M. Leppänen and C. A. Nelson, "The development and neural bases of facial emotion recognition," Advances in Child Development and Behavior, vol. 34, pp. 207–246, 2006.
[17] G. Donato et al., "Classifying facial actions," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 21, no. 10, pp. 974–989, 1999.
[18] S. Woo et al., "ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.
[19] A. Gu and T. Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces," arXiv:2312.00752, 2023.
[20] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model," in Proc. ICML, vol. 235, pp. 62429–62442, 2024.
[21] A. Hatamizadeh and J. Kautz, "MambaVision: A Hybrid Mamba-Transformer Vision Backbone," in Proc. IEEE/CVF CVPR, pp. 25261–25270, 2025.
[22] H. He et al., "MobileMamba: Lightweight Multi-Receptive Visual Mamba Network," arXiv:2411.15941, 2024.
[23] A. Wang, H. Chen, Z. Lin, J. Han, and G. Ding, "RepViT: Revisiting Mobile CNN From ViT Perspective," in Proc. IEEE/CVF CVPR, pp. 15909–15920, 2024.
[24] Y. Peng, H. Li, P. Wu, Y. Zhang, X. Sun, and F. Wu, "D-FINE: Redefine Regression Task of DETRs as Fine-Grained Distribution Refinement," in Proc. ICLR, 2025.
[25] W. Lv, Y. Zhao, Q. Chang, K. Huang, G. Wang, and Y. Liu, "RT-DETRv2: Improved Baseline with Bag-of-Freebies for Real-Time Detection Transformer," arXiv:2407.17140, 2024.
[26] Y. Li et al., "MERT: Acoustic Music Understanding Model with Large-Scale Self-Supervised Training," arXiv:2306.00107, 2023.
[27] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "CLAP: Learning Audio Concepts From Natural Language Supervision," in Proc. IEEE ICASSP, 2023.
[28] J. A. Russell, "A Circumplex Model of Affect," Journal of Personality and Social Psychology, vol. 39, no. 6, pp. 1161–1178, 1980.
[29] M. Schedl, H. Zamani, C.-W. Chen, Y. Deldjoo, and M. Elahi, "Current Challenges and Visions in Music Recommender Systems Research," Int. Journal of Multimedia Information Retrieval, vol. 7, pp. 95–116, 2018.
[30] H. Tran, T. Le, A. Do, T. Vu, S. Bogaerts, and B. Howard, "Emotion-Aware Music Recommendation," Proc. AAAI Conf. Artificial Intelligence, vol. 37, no. 13, pp. 16087–16095, 2023.