

Evaluating Facial Expression Recognition for Emotion-Adaptive Music Recommendation: A Retrospective Analysis of a Mobile Prototype

Najwa Ayuni Jamaludin^{1,*} and Jumail Taliba²

¹Department of Software Engineering, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia

²Department of Emergent Computing, Universiti Teknologi Malaysia, Johor Bahru, Malaysia

*Corresponding Author

DOI: <https://doi.org/10.47772/IJRISS.2026.100900024>

Received: 03 September 2026; Accepted: 15 September 2026; Published: 28 September 2026

ABSTRACT

Facial expression recognition (FER) can provide affective input for adaptive applications, but its usefulness depends on whether the recognition component produces accurate and consistent outputs. This study retrospectively evaluates the FER component of Emoosic, an Android mobile prototype originally developed to select music according to recognized facial-expression categories. The analysis was restricted to 30 retained static images from the original evaluation, comprising 10 images each for happiness, sadness and anger. Each image had been processed twice using the original Affdex-based recognition setup. Outcomes were analyzed descriptively in terms of correct classification, incorrect classification, non-detection, detection coverage, accuracy among detected cases and agreement across repeated attempts. In the first attempt, 13 of 30 images were correctly classified (43.3%), four were incorrectly classified (13.3%) and 13 produced no emotion output (43.3%), resulting in 56.7% detection coverage. In the second attempt, correct classification increased to 17 images (56.7%), incorrect classification decreased to two (6.7%) and non-detection decreased to 11 images (36.7%), with 63.3% detection coverage. Happiness showed the strongest category-level performance, with 70.0% correct classification in both attempts. Sadness and anger showed lower and more variable recognition. Overall, 26 of 30 images produced identical outcomes across both attempts, corresponding to 86.7% repeated-attempt agreement. The findings indicate that the prototype could provide usable facial expression input for emotion-adaptive behaviour, but substantial non-detection and category-dependent variability limit its reliability. For adaptive systems, recognition performance should therefore be considered together with detection coverage and output consistency rather than accuracy among successfully detected cases alone. This is especially important when recognition output directly triggers a downstream recommendation.

Keywords: facial expression recognition, affective computing, emotion-adaptive systems, music recommendation, retrospective analysis

INTRODUCTION

Digital devices have become increasingly integrated into everyday activities and interactions, evolving from functional tools into technologies that users engage with continuously in different personal and social contexts. During these interactions, users may naturally display affective responses such as smiling when viewing enjoyable content, showing frustration when encountering difficulties or exhibiting other facial reactions to digital experiences. However, conventional interactive systems generally have limited awareness of these observable affective signals. This has motivated growing interest in affect-aware technologies that can incorporate information about users' responses to support more adaptive forms of human-computer interaction (Picard, 1997). This general motivation also formed the basis of the original Emoosic project, which explored the use of facial expressions as an input for adapting a mobile application's behaviour.

Facial expression recognition (FER) provides one approach for automatically extracting such information from visual input. FER systems analyse facial characteristics and classify observed expressions into predefined categories. While earlier approaches relied heavily on handcrafted features and conventional machine-learning methods, recent FER research increasingly employs deep-learning techniques and has expanded into areas including human-computer interaction, healthcare and intelligent systems (Sajjad et al., 2023; Ullah et al., 2024). Nevertheless, recognition performance remains affected by factors such as head pose, illumination, occlusion, image quality and expression intensity. These challenges become particularly important when FER is incorporated into practical applications where input conditions may be less controlled than those found in benchmark datasets.

At the same time, recognising a facial expression should not be equated directly with determining a person's internal emotional state. Facial movements provide observable information that may contribute to affective inference, but their interpretation can depend on context and individual variation. Barrett et al. (2019) cautioned against assuming that particular facial configurations provide universal and reliable indicators of specific emotions while Goel et al. (2024) demonstrated that contextual information can substantially influence emotion inference. Accordingly, the present study refers to the recognition of expected facial-expression categories rather than treating the system output as an objective measurement of a user's mood.

One application area in which affective information has been explored is music recommendation. Emotion-aware music systems attempt to incorporate information about a user's affective state or the emotional characteristics of music when selecting suitable tracks. Recent approaches have continued to investigate the use of facial information together with affective representations such as valence and arousal for music recommendation. The original Emoosic prototype followed a simpler version of this concept. Developed as an Android mobile application, it used the Affdex SDK to classify facial images into happiness, sadness or anger and subsequently matched the recognised category with music using valence information obtained through the Spotify API.

However, the effectiveness of an emotion-adaptive application depends partly on the reliability of the recognition component providing its affective input. If an expression is incorrectly classified, the system may produce an inappropriate adaptive response. Similarly, if no expression is detected, the intended adaptation may not occur at all. Beyond classification accuracy, consistency is also relevant because repeated processing of the same input should ideally produce comparable outcomes. These issues were not examined systematically in the original project, which primarily focused on the development and functional implementation of the mobile prototype. Unlike conventional FER benchmarking studies that focus primarily on classification performance, this study examines recognition as an operational component within an adaptive application, with particular attention to non-detection and repeated-attempt consistency.

Therefore, the present study conducts a retrospective evaluation of the facial-expression recognition component of Emoosic using the retained image-level records from the original evaluation. The analysis focuses on 30 retained images representing happiness, sadness and anger, each of which was processed twice using the original recognition setup. Specifically, the study addresses the following research questions:

RQ1: How effectively did the facial-expression recognition component classify happy, sad and angry expressions in the retained test images?

RQ2: How consistent were the recognition outcomes across repeated classifications of the same images?

Related Work

Facial Expression Recognition

FER refers to the automatic identification and classification of facial expressions from visual information. FER systems generally involve processing facial information and extracting relevant features before classifying the observed expression. While earlier approaches relied extensively on handcrafted features and conventional machine-learning techniques, more recent methods increasingly employ deep-learning models that learn

discriminative facial representations directly from images (Li & Deng, 2022; Sajjad et al., 2023). Recent developments have also extended FER applications to areas such as human-computer interaction, healthcare and intelligent systems.

Despite advances in recognition techniques, FER remains sensitive to characteristics of both the facial expression and the input image. Variations in head pose, illumination, occlusion, image quality and facial appearance can affect recognition performance. Expression intensity can also influence whether sufficient facial cues are available for an expression to be distinguished. Recent reviews continue to identify variations in acquisition conditions, facial characteristics and datasets as important challenges affecting the robustness and generalisability of FER systems (Ko, 2018; Sajjad et al., 2023; Ullah et al., 2024). These issues are particularly relevant when FER is incorporated into real-world applications where input images may not exhibit the controlled conditions commonly found in benchmark datasets.

Consequently, evaluating FER within an application requires consideration not only of whether an incorrect expression is assigned but also whether the system is able to produce a usable recognition outcome. This distinction is important for adaptive applications because the recognition result often serves as input to a subsequent system response. In the present study, this issue is examined through the recognition of three facial-expression categories, namely happiness, sadness and anger, within an emotion-adaptive mobile prototype.

Facial Expressions in Affective Computing

Early approaches to facial-expression analysis were strongly influenced by the Facial Action Coding System (FACS), which describes observable facial movements through combinations of action units (Ekman & Friesen, 1978). Facial expressions have subsequently been widely used as observable signals in affective computing because they can provide information that supports more adaptive and responsive human-computer interaction. Systems capable of recognising facial expressions can use this information to modify their behaviour according to the perceived affective state of a user. However, recognising a particular facial configuration should not necessarily be interpreted as directly identifying a person's internal emotional state. Barrett et al. (2019), for example, reviewed evidence concerning the relationship between facial movements and emotion and cautioned against assuming that specific facial configurations consistently provide direct and universal indicators of particular emotions.

The interpretation of facial expressions can also depend on information beyond the face itself. Goel et al. (2024) demonstrated that emotion inference can be influenced substantially by situational context and that facial cues alone may be insufficient for explaining how emotions are interpreted. Their findings further showed variation across emotion categories, with facial information playing a stronger role for some categories such as happiness. These observations are relevant to automated recognition because images categorised as representing a particular emotion may differ considerably in the clarity and intensity of the facial cues actually present.

Accordingly, the present study distinguishes between facial-expression recognition and the stronger claim of detecting a person's actual mood. Although the original prototype was developed as a mobile application for *mood detection*, its computational component classified observable facial information into predefined categories of happiness, sadness and anger. Therefore, the current retrospective analysis evaluates the system according to its ability to recognise these expected expression categories rather than treating the classifications as objective measurements of users' internal emotional states.

Emotion-Aware Music Recommendation

Music recommendation systems have increasingly incorporated affective information as an additional basis for personalisation. While conventional recommendation approaches may rely on listening history, user preferences or similarities between tracks, emotion-aware approaches attempt to consider the affective characteristics of both the listener and the music. A common approach represents affect using dimensions such as valence and arousal, allowing the affective characteristics estimated from the user to be compared with corresponding characteristics associated with music (Yang & Chen, 2012).

Recent studies demonstrate the continued relevance of integrating visual affect information with music recommendation. Tran et al. (2023) developed an emotion-aware music recommendation system that estimates affect from a live image of the user and represents the estimated state in terms of valence and arousal. These values are subsequently compared with corresponding values associated with tracks in a Spotify dataset to identify suitable recommendations. Similarly, Chheda et al. (2023) proposed a cross-modal recommendation approach in which images and music are represented within a common valence-arousal emotional space to generate music recommendations.

The Emoosic prototype examined in the present study followed a simpler categorical approach. Facial images were analysed using the Affect SDK and classified into happiness, sadness or anger. The recognised category was subsequently linked to the valence information of music obtained through the Spotify API, allowing music to be selected according to the recognised expression. Although the underlying technologies represent an earlier implementation, the general concept of connecting visual affect information with music characteristics remains evident in more recent emotion-aware recommendation research.

However, the effectiveness of such an adaptive process depends partly on the reliability of its initial facial-expression recognition stage. An incorrect classification may result in an inappropriate adaptive response while failure to detect an expression may prevent adaptation from occurring altogether. Therefore, rather than evaluating the music recommendation component itself, the present study retrospectively examines the performance and consistency of the facial-expression recognition component that supplied the affective input to the original application.

METHODOLOGY

Original Prototype

This study revisits the evaluation of Emoosic, a mobile application developed to provide music recommendations based on facial-expression recognition. The application accepted a facial image captured using the device camera or selected from the image gallery. The image was processed using the Affect SDK to identify one of three supported facial-expression categories: happy, sad or angry. The recognised category was subsequently passed to the music recommendation component, where the Spotify API was used to retrieve and play music corresponding to the detected expression. The intended user interaction with Emoosic is illustrated in Figure 1. The interface allowed users to provide a facial image, view the analysis process and subsequently access music associated with the recognised facial-expression category.

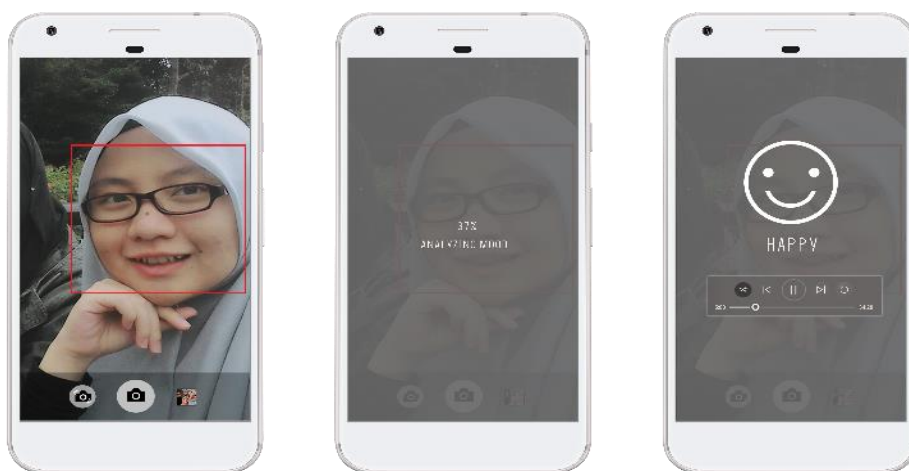


Figure 1: Emoosic prototype interfaces showing (a) facial image capture, (b) facial-expression analysis and (c) music playback following expression recognition.

Conceptually, the original prototype followed a sequential process in which an image was first acquired and analysed for facial expressions. The detected expression was then classified into one of the supported categories

and used to guide the subsequent music matching and playback process. Although the original system described this process as mood detection, the present study focuses specifically on the observable facial-expression recognition component and does not assume that the recognised facial configuration directly represents the user's internal emotional state. Furthermore, the present analysis does not re-evaluate the music recommendation component. Instead, it retrospectively examines the facial-expression recognition results retained from the original prototype evaluation.

Facial Expression Recognition Setup

Facial-expression recognition in Emoosic was implemented using the Affdex SDK within Unity (McDuff et al., 2016). The SDK provides facial analysis outputs as continuous scores ranging from 0 to 100 for several expressions and emotion-related categories. Although Affdex supports multiple outputs, the prototype was configured to use only three categories relevant to the application: joy, sadness and anger, which were represented in Emoosic as happy, sad and angry.

For camera-based input, frames were supplied to the Affdex detector using the *CameraInput* component. The application used a sampling rate of 10 frames per second with an image resolution of 320×240 pixels, and the front-facing camera was configured as the default camera. Detection results were obtained through the SDK's callback mechanism, and the resulting joy, sadness and anger scores were retrieved for subsequent classification.

The prototype applied threshold-based decision rules to determine whether an expression was sufficiently present to assign a category. A Joy score greater than 50 was used to assign the happy category based on preliminary testing during development. Lower thresholds were used for the other categories, with Sadness required to exceed 5 and Anger required to exceed 10. These lower thresholds were established during the original prototype testing because sadness and anger were found to be more difficult to detect unless their associated facial movements were sufficiently pronounced. The main configuration used for facial-expression recognition in the Emoosic prototype is summarised in Table 1.

Table 1: Facial-expression recognition configuration used in the Emoosic prototype

Configuration	Setting
Facial analysis SDK	Affdex SDK
Development environment	Unity
Facial-expression outputs used	Joy, sadness and anger
Emoosic categories	Happy, sad and angry
Camera sampling rate	10 frames per second
Input resolution	320×240 pixels
Default camera	Front-facing camera
Happy detection threshold	> 50
Sad detection threshold	> 5
Angry detection threshold	> 10

Retained Evaluation Dataset

The original evaluation of the Emoosic prototype involved 48 static facial images, comprising 16 images for each of the three target expression categories: happiness, sadness and anger. The images were originally

collected from stock-image websites including Pexels, Shutterstock, EveryPixel and 123RF using emotion-related search terms such as *happy*, *sad* and *angry*. The search category associated with each image was used as the expected emotion label during the original evaluation.

For the present retrospective analysis, complete image-level records were available for only 30 of the original 48 images. Accordingly, the analysis was restricted to these retained records rather than attempting to reconstruct the unavailable cases. The retained dataset contained 10 images for each expected expression category, resulting in a balanced set of 30 images. Each image had been processed twice using the original prototype, providing two repeated recognition outcomes per image. The possible system outputs were *Happy*, *Sad*, *Angry* or *No emotion detected*.

The original expected labels were retained for the present analysis to preserve the conditions under which the initial evaluation was conducted. These labels are treated as proxy expected labels rather than independently validated ground-truth emotional states because they were derived from the stock-image search categories rather than independent expert annotation. Therefore, the analysis evaluates the system's behaviour relative to the original expected categories rather than claiming objective recognition of users' internal emotional states.

One retained image in the Angry category contained two individuals, despite the original application being designed for single-face input. The image was retained because it formed part of the original evaluation dataset, but it was recorded as a protocol deviation for transparency.

Evaluation Measures

The retained evaluation records were analysed descriptively to examine both recognition performance and consistency across repeated attempts. Each recognition outcome was classified into one of three categories: correct classification, incorrect classification or no emotion detected. A correct classification occurred when the system output matched the expected expression label, while an incorrect classification occurred when a different emotion category was produced. Cases in which the system did not return any of the three target expressions were recorded as *No emotion detected*.

Recognition performance was examined separately for the first and second attempts. The correct classification rate was calculated as the proportion of images correctly classified out of all 30 retained images. Similarly, the incorrect classification rate and no-detection rate were calculated relative to the full retained dataset. In addition, detection coverage was calculated as the proportion of images for which the system produced any emotion classification, regardless of whether the classification was correct.

Because non-detection represented a substantial system outcome, accuracy was also examined among cases where a classification was successfully produced. This measure, referred to as accuracy among detected cases, was calculated as the number of correct classifications divided by the total number of detected cases. However, this measure was treated as secondary to the overall correct classification rate because it excludes images for which the system failed to produce an emotion output.

Performance was also examined separately for happiness, sadness and anger to identify differences across the three expected expression categories. For repeated-attempt consistency, the outputs from the first and second attempts were compared for each image. An outcome was considered consistent when both attempts produced the same result, including cases where both attempts resulted in *No emotion detected*. The proportion of identical outcomes across the 30 images was used to summarise recognition consistency.

No inferential statistical tests were applied because the retained dataset was small and consisted of purposively selected stock images rather than a random sample intended for population-level inference.

RESULTS

Overall Recognition Outcomes

Table 2 summarises the overall facial-expression recognition outcomes across the 30 retained images for both evaluation attempts. In the first attempt, 13 images were correctly classified, corresponding to an overall correct classification rate of 43.3%. Four images (13.3%) were assigned an incorrect expression category while no emotion was detected for 13 images (43.3%). Consequently, the system produced an emotion classification for 17 of the 30 images, resulting in a detection coverage of 56.7%.

In the second attempt, the number of correct classifications increased to 17 images (56.7%). Incorrect classifications decreased to two images (6.7%) while 11 images (36.7%) resulted in no emotion being detected. Detection coverage therefore increased to 63.3%, with the system producing an emotion classification for 19 of the 30 images.

When only cases for which the system produced an emotion classification were considered, 13 of 17 detected cases were correctly classified in the first attempt, giving an accuracy among detected cases of 76.5%. In the second attempt, 17 of 19 detected cases were correct, corresponding to 89.5%. However, this measure was treated as secondary because it excludes images for which the system did not return an emotion classification.

Table 2: Overall recognition outcomes across the two attempts

Outcome	Attempt 1	Attempt 2
Correct classification	13/30 (43.3%)	17/30 (56.7%)
Incorrect classification	4/30 (13.3%)	2/30 (6.7%)
No emotion detected	13/30 (43.3%)	11/30 (36.7%)
Detection coverage	17/30 (56.7%)	19/30 (63.3%)
Accuracy among detected cases	13/17 (76.5%)	17/19 (89.5%)

Recognition Performance by Expression Category

Recognition performance differed across the three expected expression categories, as shown in Table 3. Happiness produced the highest correct classification rate in both attempts. Seven of the 10 happy images were correctly classified in each attempt, corresponding to a recognition rate of 70.0%. One image was incorrectly classified as angry while two images resulted in no emotion being detected. The outcomes for the happy images were identical across both attempts.

Sadness showed the lowest correct classification rate in the first attempt, with only two of the 10 images correctly classified (20.0%). Two images were incorrectly classified as angry while six resulted in no emotion being detected. In the second attempt, the number of correctly classified sad images increased to five (50.0%). No incorrect emotion classification occurred in this attempt, although five images still resulted in no emotion being detected.

For anger, four of the 10 images were correctly classified in the first attempt (40.0%) while one was incorrectly classified as happy and five produced no emotion classification. In the second attempt, five images were correctly classified (50.0%). One image remained incorrectly classified as happy while four resulted in no emotion being detected.

Table 3: Recognition outcomes by expected expression category

Expected expression	Attempt	Correct	Incorrect	No emotion detected
Happiness	1	7 (70.0%)	1 (10.0%)	2 (20.0%)
	2	7 (70.0%)	1 (10.0%)	2 (20.0%)
Sadness	1	2 (20.0%)	2 (20.0%)	6 (60.0%)
	2	5 (50.0%)	0 (0.0%)	5 (50.0%)
Anger	1	4 (40.0%)	1 (10.0%)	5 (50.0%)
	2	5 (50.0%)	1 (10.0%)	4 (40.0%)

Consistency Across Repeated Attempts

Recognition outcomes were identical across the two attempts for 26 of the 30 retained images, corresponding to an overall repeated-attempt agreement of 86.7%. Consistency was highest for happiness, where all 10 images produced the same outcome in both attempts. For anger, nine of the 10 images produced identical outcomes, while sadness showed the lowest consistency, with seven of the 10 images producing the same result across both attempts.

Four images produced different outcomes between the first and second attempts. Two sad images that were initially classified as angry were correctly classified as sad in the second attempt. A third sad image changed from *No emotion detected* to a correct sad classification, while one angry image changed from *No emotion detected* to a correct angry classification. Thus, all four changes observed across repeated attempts resulted in improved recognition outcomes during the second attempt. The consistency results are tabulated in Table 4 below.

Table 4: Consistency of recognition outcomes across repeated attempts

Expected Expression	Identical Outcomes	Changed Outcomes	Agreement
Happiness	10/10	0/10	100.0%
Sadness	7/10	3/10	70.0%
Anger	9/10	1/10	90.0%
Overall	26/30	4/30	86.7%

DISCUSSION

Recognition Performance and Non-Detection

The results indicate that the facial-expression recognition component was able to produce correct classifications for a proportion of the retained images, although its overall performance was limited by a relatively high number of non-detections. The correct classification rate increased from 43.3% in the first attempt to 56.7% in the second attempt. However, no emotion was detected for 43.3% and 36.7% of the images respectively. This suggests that failure to produce a recognition outcome represented a substantial limitation of the system.

When only images for which an emotion was detected were considered, the proportion of correct classifications was considerably higher, reaching 76.5% in the first attempt and 89.5% in the second. These values should

nevertheless be interpreted cautiously because they exclude cases where the system was unable to return an emotion category. In an adaptive application, non-detection is itself an important outcome because the subsequent adaptation depends on receiving a usable recognition result. Therefore, recognition performance should be considered together with detection coverage rather than relying solely on accuracy among successfully detected cases.

This finding is particularly relevant to emotion-adaptive applications such as Emoosic. The facial-expression recognition stage functions as the input to the subsequent music-selection process. Consequently, a failure to detect an expression prevents the application from proceeding with the intended affect-based adaptation while an incorrect classification may lead to a response that does not correspond to the expected expression.

Differences Across Expression Categories

Recognition performance differed noticeably across happiness, sadness and anger. Happiness showed the strongest and most stable performance, with seven of the ten images correctly classified in both attempts. In comparison, sadness achieved only 20.0% correct classification in the first attempt, although this increased to 50.0% in the second attempt. Anger showed a similar but less pronounced improvement, increasing from 40.0% to 50.0%.

The stronger recognition of happiness is consistent with the possibility that some facial expressions provide more distinctive observable cues than others. Previous work has shown that the contribution of facial information to emotion inference varies across emotion categories, with happiness often associated with comparatively distinctive facial cues (Goel et al., 2024). At the same time, the present results should not be interpreted as evidence that happiness is inherently easier to recognise in all FER systems because the current evaluation involved only a small set of stock images.

The weaker performance for sadness and anger may also reflect variation in the clarity and intensity of the facial expressions contained in the retained images. Several images originally labelled as sad or angry did not display highly prototypical facial configurations and some conveyed affect partly through posture or situational context. This is relevant because facial-expression recognition operates primarily on observable facial information rather than the broader meaning of the image. Barrett et al. (2019) similarly caution against assuming a direct and universal correspondence between particular facial configurations and internal emotional states. Accordingly, the observed category-level performance may reflect both limitations of the recognition system and uncertainty in the proxy expected labels.

Consistency Across Repeated Recognition Attempts

The two recognition attempts produced identical outcomes for 26 of the 30 images, corresponding to an overall agreement of 86.7%. Happiness showed complete consistency across both attempts while anger produced 90.0% agreement. Sadness showed the greatest variation, with agreement observed for only 70.0% of the images.

Interestingly, all four cases that changed between attempts resulted in improved recognition during the second attempt. Two sad images changed from an incorrect anger classification to the expected sad classification while another sad image and one angry image changed from non-detection to the expected category. However, the available records do not provide sufficient information to determine why these outputs changed between repeated processing. Therefore, the improvement observed in the second attempt should not be interpreted as evidence of learning or systematic improvement by the recognition system.

Instead, the variation indicates that the same input image did not always produce an identical recognition outcome under the original application setup. This has practical implications for systems in which a single recognition result is used to trigger an adaptive response because inconsistent classification may lead to different system behaviours for the same input.

Implications for Emotion-Adaptive Applications

The findings highlight the importance of considering uncertainty at the recognition stage when facial-expression information is used to drive adaptive behaviour. In the original Emoosic prototype, recognised expressions were subsequently mapped to music characteristics using Spotify valence information. This creates a sequential dependency in which the quality of the music adaptation depends partly on the reliability of the preceding facial-expression classification.

Recent emotion-aware music recommendation research continues to explore similar pipelines in which affective information derived from users is connected to music characteristics, including representations based on valence and arousal. However, the present findings indicate that such systems should not assume that facial-expression recognition will always provide a reliable input. Mechanisms for handling uncertain classifications or cases where no expression is detected may therefore be important when designing affect-adaptive systems. Potential mechanisms include confidence-based thresholds, an explicit uncertainty or abstention state or multimodal signals that combine facial and contextual cues.

Contemporary FER research increasingly uses deep-learning architectures evaluated on benchmark datasets (Li & Deng, 2022; Sajjad et al., 2023; Ullah et al., 2024). Direct numerical comparison is not appropriate because this study uses a small retained stock-image set and a legacy Affdex implementation under different evaluation conditions. Future work should compare modern FER models using a common dataset and protocol.

Overall, the retrospective analysis suggests that the original prototype demonstrated the feasibility of incorporating facial-expression recognition into an emotion-adaptive mobile application, but also exposed limitations associated with detection coverage, differences between expression categories and consistency across repeated classifications. These findings support a more cautious interpretation of facial-expression recognition as one source of affective information rather than a direct measurement of a user's internal mood.

CONCLUSION

This study retrospectively evaluated the facial-expression recognition component of Emoosic, a mobile prototype originally developed to support emotion-adaptive music selection. The analysis focused on 30 retained test images representing happiness, sadness and anger, with each image processed twice using the original recognition setup. Rather than treating facial-expression classification as direct measurement of a user's internal mood, the present study examined the system according to its ability to recognise the expected expression categories and to produce consistent outcomes across repeated attempts.

The findings showed that recognition performance varied across both attempts and expression categories. The overall correct classification rate increased from 43.3% in the first attempt to 56.7% in the second while non-detection remained substantial at 43.3% and 36.7% respectively. Happiness showed the strongest performance, with 70.0% of images correctly classified in both attempts. Recognition of sadness and anger was lower and more variable. Across repeated attempts, 26 of the 30 images produced identical outcomes, corresponding to an agreement of 86.7%. However, the observed variation in several cases indicates that repeated processing of the same image did not always result in the same recognition outcome.

These findings suggest that facial-expression recognition can provide useful input for emotion-adaptive applications, but its limitations should be considered when recognition outputs are used to trigger subsequent system behaviour. In the original prototype, the detected expression was used to guide music selection through Spotify valence information. Therefore, non-detection may prevent the intended adaptation while incorrect or inconsistent classification may influence the suitability of the resulting system response.

The study is limited by the small retained dataset, the use of stock-image search categories as expected labels and the evaluation of a legacy recognition implementation. Accordingly, the findings should be interpreted as specific to the retained image set and the original Affdex-based implementation and should not be generalised to contemporary FER systems or broader populations of facial images. Future evaluations should use independently annotated labels and larger, more diverse datasets spanning age, ethnicity, lighting, facial pose

and real-time capture. Nevertheless, the retrospective analysis highlights the importance of considering detection coverage, category-specific performance and consistency alongside conventional recognition accuracy. Future emotion-adaptive systems may benefit from mechanisms that account for uncertain or unavailable facial-expression classifications rather than assuming that a valid recognition outcome will always be produced.

Ethical Approval

Ethical approval was not applicable to the present retrospective analysis because no new participants were recruited and no new human-subject data were collected. The study analysed retained system-output records generated during the original prototype evaluation using third-party stock images.

Conflict Of Interest

The authors declare no conflict of interest.

Data Availability

The derived evaluation data supporting the findings of this study are available from the corresponding author on reasonable request. The underlying facial images are not redistributed because they originated from third-party stock-image sources and may be subject to licensing restrictions.

REFERENCES

1. Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1), 1–68. <https://doi.org/10.1177/1529100619832930>
2. Chheda, R., Bohara, D., Shetty, R., Trivedi, S., & Karani, R. (2023). Music recommendation based on affective image content analysis. *Procedia Computer Science*, 218, 383–392. <https://doi.org/10.1016/j.procs.2023.01.021>
3. Ekman, P., & Friesen, W. V. (1978). *Facial action coding system: A technique for the measurement of facial movement*. Consulting Psychologists Press.
4. Goel, S., Jara-Ettinger, J., Ong, D. C., & Gendron, M. (2024). Face and context integration in emotion inference is limited and variable across categories and individuals. *Nature Communications*, 15(1), 2443. <https://doi.org/10.1038/s41467-024-46670-5>
5. Ko, B. C. (2018). A brief review of facial emotion recognition based on visual information. *Sensors*, 18(2), 401. <https://doi.org/10.3390/s18020401>
6. Li, S., & Deng, W. (2022). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 13(3), 1195–1215. <https://doi.org/10.1109/TAFFC.2020.2981446>
7. McDuff, D., Mahmoud, A., Mavadati, M., Amr, M., Turcot, J., & El Kaliouby, R. (2016). AFFDEX SDK: A cross-platform real-time multi-face expression recognition toolkit. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 3723–3726). ACM. <https://doi.org/10.1145/2851581.2890247>
8. Picard, R. W. (1997). *Affective computing*. MIT Press. <https://doi.org/10.7551/mitpress/1140.001.0001>
9. Sajjad, M., Ullah, F. U. M., Ullah, M., Christodoulou, G., Alaya Cheikh, F., Hijji, M., Muhammad, K., & Rodrigues, J. J. P. C. (2023). A comprehensive survey on deep facial expression recognition: Challenges, applications, and future guidelines. *Alexandria Engineering Journal*, 68, 817–840. <https://doi.org/10.1016/j.aej.2023.01.017>
10. Tran, H., Le, T., Do, A., Vu, T., Bogaerts, S., & Howard, B. (2023). Emotion-aware music recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13), 16087–16095. <https://doi.org/10.1609/aaai.v37i13.26911>
11. Ullah, S., Ou, J., Xie, Y., & Tian, W. (2024). Facial expression recognition (FER) survey: A vision, architectural elements, and future directions. *PeerJ Computer Science*, 10, e2024. <https://doi.org/10.7717/peerj-cs.2024>

-
12. Yang, Y.-H., & Chen, H. H. (2012). Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology*, 3(3), Article 40. <https://doi.org/10.1145/2168752.2168754>