

Original Article

Facial Expression-based Music Recommendation using Fourier Image Transformer

Raghunathan S¹, Swati Sharma²

¹Presidency School of Information Science, Presidency University, Bengaluru.

²Presidency School of Computer Science, Presidency University, Bengaluru.

¹Corresponding Author : raghunath3383@gmail.com

Received: 17 February 2026

Revised: 25 July 2026

Accepted: 07 August 2026

Published: 30 September 2026

Abstract - The ability to recognize human facial expressions is a fundamental requirement for developing advanced human-computer interaction systems. Humans convey emotions primarily through facial expressions, which has driven the development of Facial Expression Recognition (FER) technology in areas such as affective computing and personalized recommendation systems. By leveraging FER, systems can continuously monitor changes in a user's emotional state throughout the day and generate context-aware responses, thereby enhancing overall user experience. The Spotify application of FER is a music recommendation, where songs are suggested based on facial expressions to match or influence the user's emotions. Traditional FER approaches primarily rely on spatial-domain features, which often struggle to capture global context and subtle frequency-based patterns in facial images. To address these limitations, this study introduces the Fourier Image Transformer (FIT), a frequency-domain transformer model capable of learning discriminative global representations of facial expressions. By transforming facial images into the frequency domain, FIT captures both local and global variations in facial patterns, improving recognition accuracy and robustness. The proposed framework is trained and tested on the Real-world Affective Faces Multi-Label (RAF-ML) dataset and FER2013, which contains diverse real-world facial expressions under varying conditions. The preprocessing steps, including face detection, image enhancement, resizing, and normalization, to ensure reliable feature extraction. Recognized FER are subsequently mapped to selected music tracks, enabling personalized and image-based music recommendation. This approach offers a seamless, non-intrusive interface that adapts to user emotions in real-time. The FIT model achieved outstanding performance with an accuracy of 98.97%, precision of 98.92%, recall of 98.87%, and an F1-score of 98.90%. The study demonstrates how frequency-domain transformers can enhance FER performance and support intelligent, emotion-driven applications.

Keywords - Fourier Image Transformer, Real-world Affective Faces Multi-Label) dataset, FER2013, Image enhancement, Resizing, and Normalization.

1. Introduction

1.1. Background and Motivation

Human emotions are primarily conveyed through nonverbal patterns, such as facial expressions, gestures, posture, and speech. Among them, facial expressions are considered the most direct and universal signs of individual feelings. The FER states, which include happiness, sadness, anger, fear, disgust, and surprise, can be detected through facial expressions Figure 1.

Automatic detection of these expressions has been an important area of research in both affective computing and computer vision, hence allowing the systems to still recognize and respond appropriately with intelligence to human expression. The intention behind the FER systems is to automatically analyze the facial signals in images or videos and determine the related emotion, which in turn makes the

interaction between humans and machines more natural and less annoying, and leads to the construction of emotionally aware smart systems [1].

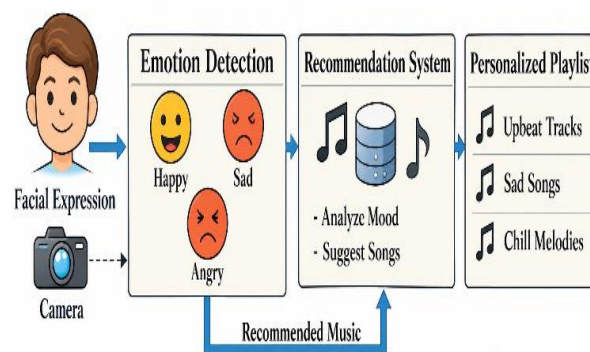


Fig. 1 Facial expression-based music recommendation system



Deep learning techniques, especially Convolutional Neural Networks (CNNs), have made a considerable improvement over traditional methods, thereby enhancing the performance of FER. These advancements have broadened the application of FER in real-world instances such as driver monitoring systems, diagnostics in health care, social robots, surveillance, and multimedia applications [2].

Music Recommendation Systems (MRS) have been demonstrated as the essential basis of digital entertainment platforms. Traditional MRS usually employs collaborative filtering, content-based filtering, or hybrid methods based on audio metadata and the historical user preferences mainly [3]. The systems demonstrate their effectiveness, yet they fail to implement dynamic emotional state modeling, which scientific research proves to be essential for establishing music preferences. [4]. This limitation restricts the integration of affective intelligence into recommendation frameworks.

The impact of music is so powerful that it can even influence the listener's facial expression, making it an effective medium for emotion-aware personalization aimed at regulating or modifying emotional states. Systems that recommend music based on facial expression combine the listener's mood with emotional state, thus creating more user engagement, satisfaction, and emotional wellness [5]. Thus, the integration of FER and music recommendation leads to a novel approach for developing adaptive, emotion-centric multimedia systems that dynamically respond to the user's feelings [6].

1.2. Facial Expression Recognition: State of the Art

FER has been extensively studied over the past two decades. The initial methods of FER were based on manual feature extraction, which included the use of geometric facial landmarks, optical flow, Gabor filters, and Local Binary Patterns (LBP), and later used conventional classifiers like Support Vector Machines (SVMs) and other Models (HMMs) [7]. Although these techniques performed reasonably well in lab conditions, they proved to be very ineffective in real-world situations where videos were recorded from a distance and had low-quality images, and thus, were unconstrained. DL, in particular, is a new stage for the FER research field. CNN-based network structures, meanwhile, are capable of automatically learning hierarchical representations of facial features, which allows for lighting, pose, and facial expression variations to be much more easily and widely generalized. Several survey papers are quite extensive with respect to the evolution of FER techniques, datasets, assessment criteria, and performance trends, and DL-based methods have increasingly been the most powerful ones throughout the last few years [8].

FER is still a difficult task to perform because of several factors, namely occlusions, head pose variations, subtle expressions, low-resolution images, and the demographic bias

in the datasets. Moreover, the majority of the FER models are trained on posed expression datasets, which might not be capable of accurately reflecting spontaneous expression in real-world applications [9]. Besides, recent reviews have also FER in the larger context of multimodal affective computing, where facial expressions are considered together to make FER more reliable [10].

1.3. Emotion-based Music Recommendation Systems

FER-based music recommendation represents a new development in conventional recommender systems within the music consumption process. FER-based music recommendation systems consider existing methods based on the category of expression acquisition, which include self-reporting, physiological sensors, audio emotion analysis, and the use of visual indicators such as facial expressions. According to several review studies, it is a common finding that FER recommenders always lead to better results in terms of acceptance and user satisfaction than traditional systems. The music recommendation is based on FER to define the emotions that are felt by the listeners, then maps those emotions to certain musical characteristics like tempo, rhythm, mode, and energy, which are known to influence the emotional perception of the listener.

The facial expression data is being mined simultaneously from the music and the user's environment [11]. The techniques allow a more advanced understanding of the emotional characteristics of music and also the listener's state, therefore enabling personalized and situation-aware recommendation systems to be created. Moreover, the use of facial expression-based music systems in medical and psychological fields is gaining impetus to improve mental health [12]. It shows the capability of adaptive music recommendation systems to contribute to psychological health by synchronizing musical signals with the user's emotional needs.

1.4. Challenges in Integrating FER with Music Recommendation

Combining FER and music recommendation presents a lot of technical and practical challenges: Real-Time Emotion Recognition: The FER will be required to provide a perfect emotional identification. 2. Emotion-to-Music Mapping: The task of translating the user's emotions, whether they are discrete or continuous, into the musical recommendations that the user takes into account is very complex and involves careful modeling of affective music theory alongside user preferences. 3. Personalization: People react to music differently, and their reactions depend a lot on culture, setting, and personal experiences, thus needing adaptable and user-focused recommendation approaches. 4. Ethical and Privacy Issues: The continuous monitoring of one's face could cause trouble with the privacy of data, permission of the subject, and emotional manipulation. These issues have to be handled by designing a system. 5. Dataset Bias and Generalization: The

lack of demographic diversity in the samples used in most FER and expression-based music datasets limits the possibility of applying the models built by training them to different populations.

1.5. Research Gap and Contribution

Although extensive comprehensive surveys are available on FER and music recommendation based on facial expressions, some research has been done on completely integrated systems that combine real-time facial expression recognition with adaptive music recommendation in real-world conditions. Researchers primarily use offline emotion labels and simple mapping strategies in their studies. The present research intends to overcome this limitation by creating a real-time FER-guided music recommendation system that automatically chooses the user's mood-matched music tracks. The proposed method utilizes DL-based FER and affective music modeling and thus contributes to the establishment of emotionally intelligent multimedia systems, which can improve not only user experience but also emotional state. Currently, there are many music applications that users can choose from. The most widely used music applications across the world are: Spotify is the world's leading music application in terms of user satisfaction. Their recommendation system is the best in the world, and it is the reason Spotify is so widely used [13].

The advanced FER is one of the main components of the proposed system, which enables intelligent and facial expression-based music recommendation to the users. The proposed framework integrates several steps, including face detection, robust preprocessing techniques (image resizing, enhancement, and normalization), and the FIT for feature extraction and classification to accurately identify a variety of expressions. The use of well-known datasets like FER-2013 and RAF-ML ensures that the model is able to generalize to real-world faces well. The user's facial expression is first recognized, and then the system searches for music tracks that are aligned with the detected mood from a selected list, thus providing a personalized and involving listening experience. With the help of FER, the method automatically selects the content while at the same time making the interaction process more natural and user-friendly by reducing the need for explicit user input to a minimum. The framework can address the problems of traditional spatial domain methods by using the combination of the frequency domain analysis and transformer-based learning, thus enabling the way for real-time and context-aware multimedia applications with strong potential.

The FIT model is different from other CNN and transformer-based approaches for FER because, unlike other models, it employs frequency domain analysis, which allows the identification of both structural and emotional features of the faces in the image. It integrates Fourier Domain Encoding, amplitude and phase values, position encoding, and self-

attention from transformers. This novel approach is expected to perform well for facial expressions under changing lighting conditions, poses, and realistic environment settings. Furthermore, most FER-based music recommendation systems use basic emotion mapping approaches, while the suggested framework uses real-time emotion recognition to make personalized music recommendations based on audio features available on Spotify. Consequently, the suggested FIT framework is more accurate, robust, computationally efficient, and capable of delivering better real-time performance than existing methods. The objectives are:

- To accurately recognize human facial expressions from facial images using the FIT.
- To implement robust preprocessing, including face detection, image enhancement, resizing, and normalization.
- To detect facial expressions to personalize music tracks, enabling mood-based recommendations.
- To ensure the system is efficient and suitable for real-time emotion detection and music suggestion.
- To validate the model on benchmark datasets like FER-2013 and RAF-ML for reliable real-world performance.

The document structure follows this specific outline, which includes Section II, which presents a literature review and related works, which provide essential principles about DL. Section III describes the dataset utilized in this research. The assessment criteria appear in Section IV. The study ends with Section V, which contains the conclusion.

2. Literature Survey

FER has become a primary focus in affective computing and personalized applications, such as facial expression-based music recommendation. DL techniques, mainly CNNs, caused an increase in accuracy by the automatic extraction of the hierarchical spatial features from the data. In the recent past, transformer-based models have succeeded in capturing the entire range of global dependencies and even the finest details of the expression's high performance. However, these advances cannot hide the fact that there are still obstacles in differentiating between emotions that are quite complex, and also that the practicality of the systems is not assured in real-time applications. This paper builds high moments by studying the combination of frequency-domain feature extraction with transformer architecture in order to achieve an even stronger FER performance. Table 1 Shows the advantages and disadvantages of existing models.

The research [14] utilized a ResNet-50-based structure for the recognition of facial emotions, which was trained with the FER dataset containing real-life images of people displaying a wide range of emotions. The system achieved an accuracy of 82% in FER. More specifically, to improve model explainability, Grad-CAM was used as a visualization tool to

show which particular facial areas contributed to the model's prediction, thus providing clear justification for the issued recommendations. The model was trained on a combined dataset consisting of standard FER images along with emotionally labeled real-world user images. The learning pipeline consisted of image pre-processing, hierarchical feature extraction through convolutional operations, and high-level reasoning using fully connected layers, finally undergoing emotion inference at the output stage.

Pre-trained DL architectures, specifically ResNet-18 and VGG-19, were used in [15] to infer users' emotional states from facial expressions. The models were pre-trained on and validated using two widely accepted datasets: Facial Emotion Recognition (FER-2013) and the Extended Cohn-Kanade (CK+) database, which allowed the identification of various emotional classes such as happiness, anger, sadness, surprise, disgust, and neutral expressions.

Along with the analysis of facial expressions, musical affect was recognized through the usage of Spotify audio data using classical machine learning classifiers, namely Support Vector Machines (SVM) and Linear Discriminant Analysis (LDA), which were used to group the musical content into emotive classes such as happy, sad, energetic, and calm. A rule-based fusion mechanism was created to combine detected user emotions with corresponding music moods, hence enabling automatic playlist generation that fits the emotional context of the listener, to connect the facial and musical emotion domains. The experiments showed that the ResNet-18 model performed high on the FER-2013 dataset.

Two widely recognized facial expression datasets, RAF-DB and FER2013, were adopted in the study reported in [16]. To mitigate the impact of class imbalance that accompanied these datasets, a new balanced version was introduced, where low-quality samples from FER2013 were eliminated, and massive data augmentation techniques were employed. To perform a complete comparative analysis of all thirteen different Vision Transformer (ViT) architecture variants, the three datasets were utilized. The results of the experiments indicated that the ViT-based models had an exceptionally high strength for recognizing facial expressions. It was the MobileViT and Tokens-to-Token ViT that demonstrated the highest performance.

A modified variant of the Swin Transformer architecture was applied to FER in [17], leading to significant improvements in the final results. The new Swin-FER model showed a more robust feature representation by assessing the intermediate and deeper layers of the network. Parameter efficiency was boosted by mean-based aggregation, feature splitting, and group convolution techniques. The Swin-FER model got a recognition accuracy of 71.11% on the FER2013 dataset.

In the research [18], a whole sequence of functional phases was presented that began with data collection, and then the CNN was applied for FER. Additionally, a real-time system was combined with a Flask web app and included OpenCV for the detection of emotions using facial expressions and Spotify API for the integration of music playlists. The CNN trained on the images reduced the presence of emotions such as anger, disgust, fear, happiness, neutrality, sadness, and surprise very well. Experiment results revealed that the system had high accuracy in detecting facial expressions and generating suitable music recommendations. The CNN obtained a global precision of 62.44% after completing ten epochs of training. The real-time module, which utilized OpenCV, was highly effective in accurately tracking facial expressions and allowed the music playlist to change in real-time.

The research introduced in [19] created a CNN-based system that successfully detected facial expressions while showing compatibility with a music recommendation system, which used user mood data for its operation. The process was carried out and tested on the FER-2013 database of 35,887 facial images, and also on a specially selected audio library of 1,000 tracks, which was earlier labeled according to the emotions.

A new emotion-based music sorting method was developed, which first analyzed the song texts with a text-based emotion classification system and then placed each track into one of the seven emotional groups: anger, fear, disgust, neutrality, surprise, sadness, and happiness. Through the experimental verification, the system's efficiency was proven, which resulted in a 91.78% accuracy for FER.

The hybrid music recommendation system [20] was developed with the intention of creating a prediction device for consumer's musical tastes by being a smart filter. It was the combination of methods like popularity, user cooperation, and filtering. A thorough content-based data analysis was conducted to reveal the data's hidden features and thus to enhance the prediction quality. The model could be implemented on other datasets and under different usage scenarios, yet it still yielded consistent and reliable results. The experiments confirmed the system's efficacy by achieving a 90.4% accuracy rate.

A facial expression-based music recommendation system was developed in [21] to select suitable songs according to the listener's emotional condition. The system was built upon the CK+48 dataset, which was composed of black-and-white facial images that were divided into seven emotions: anger, disgust, irony, fear, joy, sorrow, and surprise. The authors utilized two Deep Learning (DL) techniques: one was a compact Convolutional Neural Network (CNN) for real-time emotion detection through webcam images, and the other was a transfer-learning model based on the ResNet framework for

the uploaded images to be processed. The findings of the experiments indicated that both CNN and ResNet models were capable of attaining quite high precision levels in emotion recognition.

Facial expression-based emotion recognition was presented in [22] as an essential aspect of classifying human emotions. The primary aim of the system was to accurately recognize and classify the emotional expressions to recommend music that was combined with the user's mood. The differentiation of different emotional responses was performed by the CNNs, which were based on the training facial image data sets. The model could sense the person's mood using facial characteristics and choose music that was in accordance with that FER. Fuzzy classification methods were used to improve and enhance the outcomes.

The research [23] was based on ML algorithms, where the first data was obtained, then a conventional CNN and a VGG16-based model were simultaneously trained, after which music was recommended through Spotify. The performance of the models in detecting facial expressions was reported to be almost perfect, with the CNN at 66% and the VGG16-based model at 92%. The system was able to perfectly recommend the music from the Spotify collection that matched the detected mood.

In [24], the authors developed and implemented a deep learning-driven automated system for facial expression recognition using image data. The research was conducted with the real-time differentiation of various emotional states and the challenge of recognizing the corresponding emotion accurately from the face.

The system was capable of detecting human emotions such as joy, sorrow, rage, etc., by using CNNs along with various advanced DL techniques. The FER-2013 dataset was used as the source of facial emotions, while the data from Spotify was used to enable the playing of the music that corresponded to the listener's mood. Both datasets were available as free downloads on Kaggle. The experiments conducted showed that the model was able to identify emotions with an accuracy of 91.23% based on the user's facial expressions.

The study in [25] presented a study on an emotion-aware music recommendation system that first detected human facial expressions and then recommended music according to the individual's taste determined by emotional state. The analysis of facial expressions was carried out through a CNN, which, from the range of emotions expressed by the user, the user to one of six basic classifications: anger, fear, sadness, joy, surprise, or neutrality. The CNN's learning occurred on a labeled facial-image dataset that was aimed at achieving very accurate emotion detection as a result. A user's emotional state was recognized, and the system connected to the Spotify API

to fetch tracks that matched the identified mood. The system was tested on a separate test dataset, and the outcome of the experiments confirmed the ability of the method to recognize emotions and create playlists accordingly.

The research presented in [26] analyzed users individual preferences and suggested music tracks aligned with similar characteristics. The mentioned steps were carried out, including the user giving consent for the system to capture an image using a webcam. The captured image was cross-referenced with a secure database, then a CNN was used, and the system automatically selected and played mood-appropriate music according to the detected emotion. The whole process fulfilled the main requirement of a facial expression-based music recommendation system.

A technique integrating music mapping with FER was presented [27], making human-computer communication more intuitive and less demanding. A specifically designed CNN was applied to distinguish user's feelings in the FER-2013 dataset. It was better than the stated models, including VGG16 (accuracy of 55.87%), to obtain the music proposals from the "278k Labeled Spotify Songs" playlist.

The research study [28] was designed to meet the needs of Thai music listeners and provide not only Thai songs but also various music genres. The system was previously trained to identify, through words and phrases, the user's moods and, consequently, to produce customized playlists. Thus, it changed music listening from a passive activity into an engaging experience. Multiple ML classifiers were employed during the research, yielding the following outcomes: Random Forest (0.94).

The model in the paper was called Multimodal Emotion-guided Music Generation Model (MEMGM). The system translated facial expressions, speech prosody, and electrocardiogram signals to music by mapping emotions through continuous Valence-Arousal-Dominance (VAD) space. The architecture [37] combined a cross-modal Transformer, multimodal bidirectional long short-term memory, and gate for emotion identification. Introducing an annotated MIDI dataset using VAD enabled the synthesis of controllable expressive music. Experiments and subjectivity tests showed better alignment in emotion transfer and better results than other state-of-the-art techniques.

The use of a public dataset [38] for music classification involving five emotional classifications, which included Aggressive, Happy, Dramatic, Sad, and Romantic moods, was considered. The audio features, including Short-Time Fourier Transform (STFT) and Mel-Frequency Cepstral Coefficients (MFCC), were extracted in order to get the important characteristics from the audio file. With an 80-20 split of the data, 95% accuracy was obtained using the HuBERT model.

A DL [39] approach to FER using a novel enhanced version of ConvNeXt was presented. Spatial Transformer Networks were used to stress crucial facial regions, and Squeeze-and-Excitation layers were used to enhance channel-level feature extraction. The model further adopted a self-attention regularization module to generate compact and discriminative features. Experiments conducted using the FER2013 dataset proved to be more effective than the latest approaches in emotion recognition. The paper [40] presented

a new model named Vivify, which used RoBERTa for text emotion detection and LSTM for music emotion detection. The system generated music which matched the emotions of the text based on continuous valence-arousal through cosine similarity. It depended on specific emotions or linguistic elements. Vivify was capable of performing fine-grained and linguistically independent emotional alignment between texts and music. The success of the method approach was validated by a user experiment on 63 users.

Table 1. Advantages and disadvantages of existing models

Ref	Method	Dataset	Advantages	Disadvantages
[14]	ResNet-50 with Grad-CAM explainability	FER dataset with real user images	High accuracy (82%); explainable predictions through Grad-CAM.	CNN is limited in modeling long-range dependencies, and no temporal analysis.
[15]	ResNet-18, VGG-19, SVM & LDA rule-based fusion	FER-2013 and CK Spotify audio data	Multimodal fusion (face and music) and interpretable rule-based mapping.	Rule-based fusion lacks adaptability and has no deep music feature learning.
[16]	Vision Transformers (MobileViT, T2T-ViT, CrossFormer)	FER-2013 and RAF-DB (balanced version)	Strong global feature modeling and high FER accuracy.	High computational cost and not applied to music recommendation.
[17]	Optimized Swin Transformer (Swin-FER)	FER-2013	Efficient hierarchical representation and improved spatial awareness.	Moderate accuracy.
[18]	CNN, OpenCV, Flask and Spotify API	Custom real-time data	Real-time emotion detection and dynamic playlist update.	Low precision and CNN scalability limitations.
[19]	CNN with lyric-based emotion classification	FER-2013 and custom music dataset	Very high FER accuracy and lyric-aware music labeling.	Text-based lyrics analysis ignores audio features.
[20]	KNN Hybrid recommendation	Multiple music datasets	Robust and flexible recommendation strategy.	No facial emotion input and not emotion-aware.
[21]	Lightweight CNN and ResNet transfer learning	CK+48	Supports real-time and offline emotion detection.	Small dataset size.
[22]	CNN and fuzzy logic classification	Facial image datasets	Improved recommendation refinement through fuzzy rules.	Complex rule tuning and limited scalability.
[23]	CNN, VGG16, and Spotify API	Facial image dataset and Spotify	High accuracy and platform integration.	High computational load and static images only.
[24]	CNN-based automated FER system	FER-2013 and Spotify dataset	High accuracy and end-to-end automation.	CNN lacks global contextual modeling.
[25]	CNN-based emotion detection and Spotify API	Labeled facial image dataset	Effective mood-to-music mapping and good playlist quality.	Limited emotion categories and no personalization learning.
[26]	CNN-based FER and music retrieval	Webcam images	Simple implementation and real-time response.	Dependency on an external platform and limited evaluation.
[27]	Custom CNN and Spotify music mapping	FER-2013 and 278k Spotify songs	A large-scale music dataset is better than VGG16.	CNN-only architecture ignores temporal cues.
[28]	ML classifiers (Random Forest and NLP-based mood detection)	Thai text-based mood dataset	High accuracy and language-specific tuning.	No facial emotion recognition and text-only approach.

[37]	Multimodal Emotion-guided Music Generation Model using cross-modal Transformer, multimodal BiLSTM, and gating mechanism	VAD-annotated MIDI dataset	Direct multimodal emotion-to-music translation, continuous VAD representation, controllable expressive music generation, and improved emotion alignment	High computational complexity requires multimodal data acquisition and synchronization.
[38]	HuBERT model with STFT and MFCC feature extraction	Public music mood dataset	Effective transformer-based classification, robust audio feature analysis	Limited to predefined mood classes, performance depends on the audio dataset quality
[39]	ConvNeXt with Spatial Transformer Network, Squeeze-and-Excitation blocks, and self-attention regularization	FER2013	Improved facial region attention, compact discriminative features, and superior emotion recognition accuracy	Requires large training data and high computational resources
[40]	RoBERTa for text emotion recognition and LSTM for music emotion recognition with cosine similarity	Text-music emotion alignment dataset with user evaluation	Fine-grained and language-independent emotional alignment, high user satisfaction and engagement	Depends on accurate text emotion recognition, subjective evaluation may vary across users

The suggested framework utilizes FIT as the primary architecture for FER-based music recommendation because of its superior capability to model spatial and frequency-domain information. The FIT obtains global facial patterns, meanwhile, by changing image representations into the frequency domain, unlike traditional CNN methodologies that mainly capture local spatial features. The model's new approach to recognizing even the smallest muscle movements and textural changes is essential for accurate emotion recognition. The idea of self-attention is an important part of feature learning because it captures the dependencies between facial parts. Hence, making the complexity of the expressions less significant. FIT has been showing more potential than traditional DL models in terms of variations due to lighting, angle, and background noise. Frequency-based representations reduce redundancy and improve performance on FER-2013 and RAF-ML datasets. Parameter efficiency is another important aspect that the proposed model offers. The Fourier-based transformation also contributes to quicker convergence during training and the stability of optimization. The combination of feature extraction and classification into one system allows for end-to-end learning, and thus, it is making the process easier and faster. These qualities make the FIT one of the best choices for real-time as well as large-scale emotion recognition tasks. The proposed method, together with an emotion-to-music mapping approach and Spotify API integration, yields more accurate, consistent, and emotionally matching music recommendations. The FIT is a powerful and flexible solution that not only surpasses the current CNN- and ViT-based techniques in the area of facial expression-driven music recommendation systems but also surpasses them.

Although the current models performed well with emotional recognition and music composition, several limitations were found in the areas of feature representation, multimodal fusion, and computational efficiency. Improved

emotion-to-music translation was achieved in MEMGM by means of multi-modal VAD mapping, whereas HuBERT boosted the performance of music mood recognition through a learning-based approach powered by transformers. In turn, the improved ConvNeXt architecture contributed to increased localization ability and discriminative representations for the task of FER. Also, Vivify achieved accurate emotional alignment between text and music using RoBERTa & LSTM models. Despite these advancements, the models relied on either temporal-domain processing or modality-specific feature extraction, which limited the capture of complex spatial-frequency emotional patterns. Compared to these methods, the proposed Fourier Image Transformer demonstrated superior performance by effectively integrating global frequency-domain representations with transformer-based attention mechanisms for robust emotional feature learning. The proposed model achieved improved accuracy.

3. Proposed Methodology

The proposed methodology utilizes FER to identify the listener's emotional state and enable FER-based music recommendation. Facial images captured during music playback are processed to extract discriminative facial features that reflect underlying emotional signals. These extracted features are subsequently classified into predefined affective classifications, which serve as the basis for intelligent music selection. By dynamically associating the detected emotional states with appropriate music playlists, the system delivers personalized and adaptive recommendations in real time. The proposed framework enhances FER-based emotion inference with an intelligent music selection strategy. Figure 2 shows user's facial images, either from the camera or uploaded through an automatic localization that applies a face detection technique first. The face is aligned and cropped for the elimination of the background. To train and test the

systems well, the two popular datasets FER-2013 and RAF-ML are used. Various preprocessing operations, like grayscale conversion, normalization, image enhancement, and image resizing to a fixed resolution, are applied to the data to have a standard input for the model.

The images are preprocessed and then transferred to a FIT, which is the feature extraction and classification. The transformer-based setting combines the spatial and frequency-domain facial representations that are complementary to each other. The resulting features are then used by the model to classify expressions into different emotional classes: happy, sad, angry, surprised, and neutral. The predicted emotion is then associated with a particular music mood category through an emotion–music mapping strategy that is predefined.

Music tracks that correspond to the identified state are then taken from the Spotify API, with audio features like valence, energy, and tempo. Ultimately, the user is presented with a personalized playlist that not only matches the user's current mood but also evolves in real-time with the changes in facial expressions, thus improving user interaction through the listening of emotionally appropriate music. The dataset was randomly divided into 80% training data and 20% testing data. The training data were further evaluated using ten-fold cross-validation to optimize the model and assess its generalization performance.

The proposed approach involves using FER for recognizing the emotions of the listener and providing them with a personalized FER-based music recommendation system. The model consists of image pre-processing, FIT-based emotion recognition, and emotion-aware music recommendation using real-time audio processing. The entire methodology involves image acquisition, pre-processing, feature extraction, emotion detection, emotion-music mapping, and on-the-go music recommendation.

Firstly, facial images are captured either by using a live feed of the camera or by users themselves while listening to music. Because these images include extraneous details from their environment and background, face localization using face detection techniques is performed automatically. Facial region detection is carried out by aligning and cropping the region so that only relevant facial regions are included in the analysis process. This step will help improve feature consistency and eliminate unnecessary spatial variations that may have an effect on the efficiency of emotion detection.

In order to construct an efficient and generalized system for recognizing emotions from facial expressions, two standard face databases are employed in the training and testing processes. They include the FER2013 and RAF-ML databases that comprise various facial expressions with different illumination levels and orientations, and varying degrees of facial expressions. After acquiring the image,

various pre-processing steps are carried out to prepare the data before performing the extraction process. Firstly, a greyscale transformation is used to minimize the complexity of computation, but without losing any important information in relation to the face structure. Secondly, the image is normalized so that the pixel intensities will be scaled within a certain range to stabilize the gradient updates during the network training process. Enhancement of the image through equalization of the histogram and filtering of noise can also be done.

Following image pre-processing, the face images are then fed into the FIT model for feature extraction and emotional recognition. In contrast with typical FER algorithms based on convolution that mainly emphasize information in the spatial domain, the proposed FIT model utilizes both spatial and frequency domains in capturing the necessary emotional features. First, a Fourier transform is carried out on the face image, enabling the extraction of frequency-related features connected to emotions.

In the FIT framework, transformer-based learning is used to capture long-range dependencies in facial region pairs. The combination of spatial features and Fourier frequency domain features is done through multi-head self-attention, which allows the model to learn about global contextual relationships among facial parts like eyes, eyebrows, nose, and mouth.

The transformer-based encoder learns to focus on emotionally relevant parts by highlighting them while ignoring redundant data. The features obtained are then fed into fully-connected layers for classification. The system is trained to classify the facial expressions based on emotions like happiness, sadness, anger, surprise, and neutral. To calculate the probability of each emotional class and thus ascertain the main emotion felt by the subject, a softmax activation function is used in the final classification layer.

After the detection of emotion, an emotion-music mapping process is undertaken to match the identified emotional state with the proper music mood. This involves assigning each class of emotion to certain musical qualities such as valence, energy, tempo, and acoustic parameters. For instance, a happy emotion would be matched with fast-tempo and high-valence music, while a sad emotion would be paired with slow-tempo and low-energy music.

In order to generate individualized recommendations for music, the model uses the Spotify API to fetch emotionally relevant songs. The API is used to search for songs that have emotion-based characteristics within their audio features and create playlists according to the individual's present emotional status. Various key parameters such as valence, danceability, acousticness, tempo, and energy are taken into account while generating recommendations.

The suggested system is designed to operate in an adaptive manner on a real-time basis, tracking the changes in the expression on the face when playing the music. If there are

any changes in the mood of the user, the system will be able to adjust the list of songs to ensure emotional alignment between the listener and the recommended music.

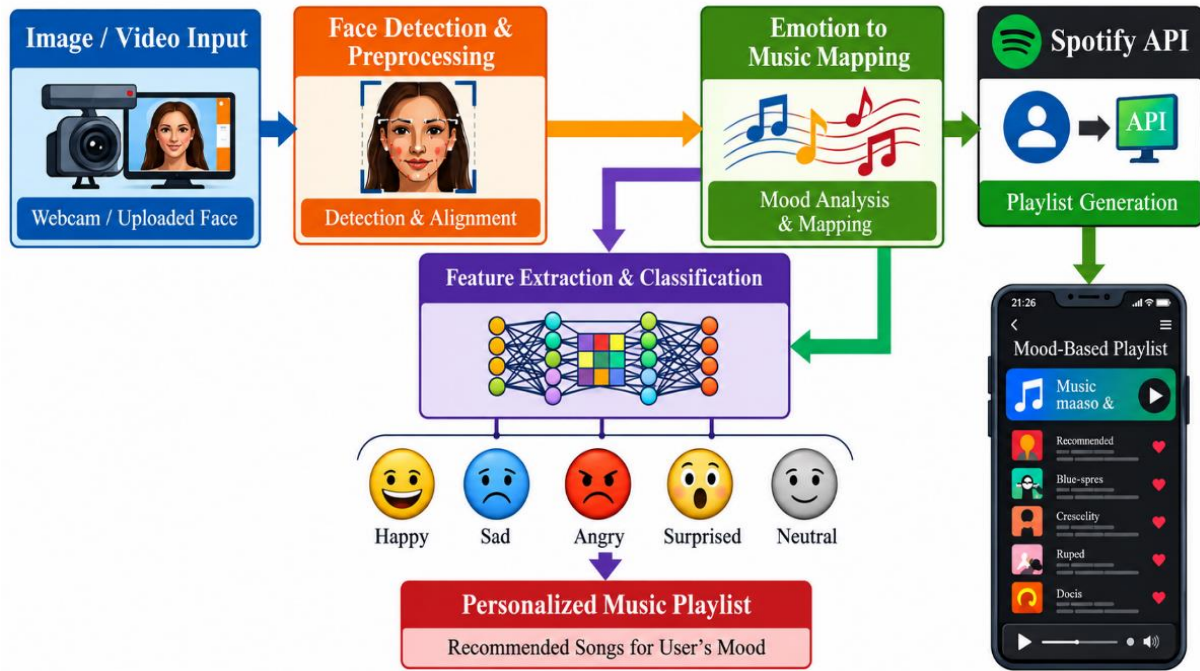


Fig. 2 Proposed model flow diagram

3.1. Facial Expression Recognition 2013 (FER-2013) Dataset

The FER-2013 dataset is a well-known dataset for research, presented as a part of the ICML 2013 Challenges in Representation Learning. The dataset is made of 48×48-pixel grayscale facial images that are taken from the internet and are annotated with basic human emotions. FER-2013 is developed to reflect real-world situations and has a wide range of variations, which include differences in illumination, facial expression, and quality of images, making it difficult but well-suited for testing the robustness of expression recognition models. A total of seven discrete emotions are assigned to each image as a label, Figure 3, which include anger, disgust, fear, happiness, sadness, surprise, and neutral.

FER-2013 is especially favored in DL research because of its moderate size and variety. However, it also faces class imbalance concerns, particularly for the disgust category, which adds more difficulty to model generalization. Its presence on platforms like Kaggle further elevates its accessibility for academic research. In this research, FER-2013 is utilized to train and test the proposed FIT-based FER model. The complexity of the dataset makes it ideal for evaluating the robustness of frequency-domain and transformer-based feature extraction techniques [29]. Table 2 presents the distribution of samples in the FER-2013 dataset.

Table 2. FER-2013 dataset distribution

Emotion Class	Label	Number of Images
Angry	0	4,593
Disgust	1	547
Fear	2	5,121
Happy	3	8,989
Sad	4	6,077
Surprise	5	4,002
Neutral	6	6,198
Total	—	35,527

3.2. RAF-ML Dataset

The RAF-ML dataset is derived from the RAF-DB dataset, which is designed to support the analysis of facial expressions in real-world conditions. RAF-ML is not a single-label emotion dataset, but rather a multi-label emotion annotations dataset, where multiple emotional states can be associated with each facial image at the same time. It is a reflection of the natural complexity of human expression, where the expressions usually show the blending or overlapping of the affects rather than just the single dominant emotion. The dataset is made up of real-world facial images that were taken from the internet, and they show a lot of variations in pose, lighting, age, race, and background. RAF-ML makes possible both categorical emotion recognition and emotion intensity analysis, which makes it suitable for DL and transformer-based FER models. The dataset is frequently used

to test the robustness, generalization, and multi-label learning capabilities of the emotion recognition systems. For the proposed FIT, RAF-ML is used to boost its generalization power by exposing the model to complex, real-world emotion variations. Its multi-label characteristic is especially advantageous for emotion-aware applications like music recommendation, where subtle emotional states can be matched with more expressive and personalized playlists [30].



Fig. 3 FER-based emotions

3.3. Spotify Dataset

FER-based music recommendation systems are largely dependent on the Spotify dataset, which not only includes but also provides high audio and metadata features for the selection of songs. Although in this case, Spotify does not give a particular labeled facial expression dataset, instead, the Spotify Web API allow access to a million of tracks. Each track is connected with such audio features as valence, energy, tempo, loudness, and danceability, which are highly connected with human feelings, that are the basis of the facial expression classification. In the case of FER systems, first the user's facial expression is classified in terms of FER, and then these audio features are applied to map the emotion to music tracks. For instance, high-valence and high-energy songs usually align with happy expression, whereas low-valence and low-energy tracks with sad moods. Moreover, among the Spotify metadata are genre, popularity, and artist, which allow for customized playlist creation. The dataset is suitable for online music retrieval and updating playlists. It is large and continuously updated, thus ensuring a varied and fresh supply of recommendations. Consequently, Spotify data contributes significantly to enhancing the practical and realistic aspects of emotion-aware music recommendation systems [31].

3.4. Data Preprocessing

Facial image preprocessing is a procedure that enhances the quality of images used for recognition and stabilizes the recognition model. Resizing the images is the initial preprocessing step for all images. This procedure is done to make the resolution of all facial images equal; thus, the model can have the same input size. Normalization, which adjusts the pixel intensity values to a common range, not only speeds up

the process but also helps in improving numerical stability. Image enhancement methods such as contrast adjustment and noise reduction are used to make the facial features more evident and to remove the unwanted expressions, thus providing clearer and more distinctive inputs for FER.

3.5. Normalization

Normalization is a preprocessing stage that brings the intensity values of facial images to a common range, and it helps to improve numerical stability during model training. The variations in lighting are reduced, and the DL models converge more quickly as a result of this. Normalization is a method of scaling all input values to a common range, which stops the features with higher magnitudes from overpowering the learning process. The method has also been beneficial in improving the model's ability to generalize [32]. The most common normalization technique is min-max normalization, which is expressed as Equation (1):

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

where x represents the initial pixel value, and x_{min} and x_{max} represent the least and greatest pixel intensities in the image, respectively.

3.6. Image Resizing

Resizing of images is one of the important preprocessing steps that ensures all the facial images have identical spatial resolution before the model input. It normalizes the dimensions of the input images, thus allowing the neural network process the images without any differences in shape, and very effectively at the same time. Resizing also reduces the computational load by reducing the number of pixels processed during both training and testing. Furthermore, it helps preserve important facial structures while removing unnecessary image details. Consequently, image resizing plays a part in enabling the model through improved stability and consistent feature extraction across the dataset [33].

3.7. Image Enhancement

Histogram Equalization is an imaging processing technique applied to enhance the contrast of facial images by making the pixel intensity values more even throughout the image. It makes the dynamic range of the grayscale levels used for portraying the face larger, thus making the expression of the face clearer. The technique is especially useful in dealing with images taken with bad or uneven light. The main areas and finer texture of the face being highlighted by histogram equalization, the very subtle facial signals are made more visible. At the same time, it makes the recognition of the facial expressions more accurate. The diagram illustrates how histogram equalization enhances image contrast, making details in the "After" image clearer compared to the "Before" image [34].



Fig. 4 Image enhancement

3.8. Feature Extraction and Classification

FER requires feature representations capable of detecting and differentiating small local muscle movements as well as variations in the overall facial structures. Standard CNNs primarily operate in the spatial domain, which limits their ability to capture long-range connections and global frequency features. To address these limitations, the Figure 5 shows the FIT as a feature extraction backbone, enabling facial expression representation directly in the Fourier domain, where emotional signals such as edge sharpness, texture variations, and expression intensity are more distinctly encoded.

3.8.1. Fourier Domain Representation of Facial Images

In the case of a grayscale facial image $x \in R^{H \times W}$ which has been processed through preprocessing steps, the first step is to apply the 2-D Discrete Fourier Transform (DFT). which is defined as follows, Equation (2):

$$X(u, v) = \sum_{m=0}^{H-1} \sum_{n=0}^{W-1} x(m, n) \exp\left(-j2\pi\left(\frac{um}{H} + \frac{vn}{W}\right)\right) \quad (2)$$

where $X(u, v)$ is the complex-valued frequency spectrum that not only holds the magnitude information but also the phase information. The low-frequency parts of the spectrum represent global facial characteristics (like the face's shape and orientation), whereas the high-frequency ones convey the minute details of expressions like the wrinkles, eyebrow tension, and lip curvature.

Due to the conjugate symmetry property of real images, the full Fourier spectrum X_h is limited to half, containing no duplicates, and still holding the complete image information.

3.8.2. Fourier Domain Encoding (FDE)

To enable transformer learning, the 2-D Fourier spectrum is turned into a 1-D sequential representation, which is known as FDE. The Fourier coefficients are expanded in a circular

fashion from the DC component outwards, creating a sequence that is ordered as follows, Equation (3):

$$S = [c_1, c_2, \dots, c_N] \quad (3)$$

The token C_i corresponds to a complex Fourier coefficient. One of the unique features of the encoding is that every prefix of the sequence can reconstruct the whole image at a lower resolution, thus enabling hierarchical feature learning. Each coefficient C_i is decomposed into amplitude and phase as follows: Equations (4 and 5):

$$a_i = \frac{2(|c_i| - a_{min})}{a_{max} - a_{min}} - 1 \quad (4)$$

$$\phi_i = \frac{\angle(c_i)}{\pi} \quad (5)$$

where a_{min} and a_{max} are constants for the whole dataset used for normalization. The resulting representation in values for real numbers is defined as (6):

$$C = \begin{bmatrix} a_1 & \phi_1 \\ a_2 & \phi_2 \\ \vdots & \vdots \\ a_N & \phi_N \end{bmatrix} \in \mathbb{R}^{N \times 2} \quad (6)$$

3.8.3. Positional Encoding in Fourier Space

Due to the absence of built-in spatial perception in transformers, the application of 2-D polar positional encoding represents each Fourier coefficient's radial distance and angular position. The resulting FDE token embedding is produced by the combination of learned linear projections and positional encodings and is defined as follows, Equation (7):

$$E_i = \text{Linear}([a_i, \phi_i]) \oplus \text{PE}(r_i, \theta_i) \quad (7)$$

where $\oplus$ represents concatenation. The encoding facilitates the transformer to acquire the spatial-frequency relationships that are essential for distinguishing between different expressions.

3.8.4. Transformer-Based Feature Learning

The input sequence $E \in R^{N \times F}$ is processed using the stacked transformer layers. All the layers are made up of multi-head self-attention, and then a feed-forward network is defined as follows in Equation (8).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (8)$$

The mechanism facilitates modeling of the global frequency interactions, thereby allowing the learning of specific emotion frequency patterns in the whole face. FIT, in contrast to CNNs, does not make use of local receptive fields and thus is highly capable of subtle expression recognition.

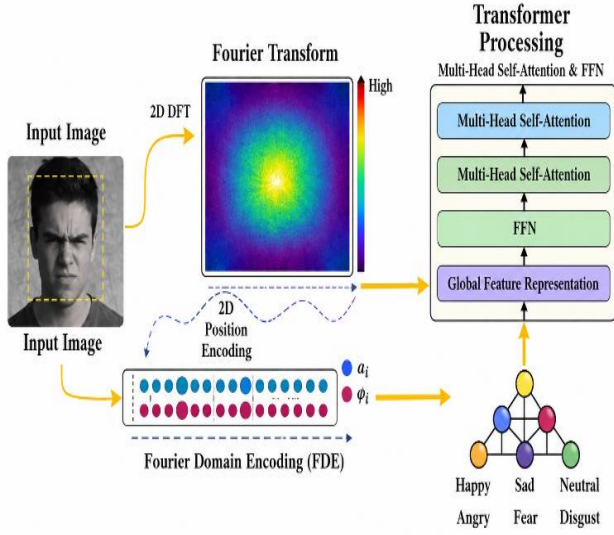


Fig. 5 FIT for feature extraction

The proposed system employs FIT as a method to extract features that are both robust and discriminative for facial expression recognition-based music recommendation. It is the process of changing the facial image from the spatial domain to the frequency. Low-frequency components are responsible for the global facial structure, while high-frequency components are responsible for very close and detailed recognition, and thus capture details such as wrinkles, lip moves, and eyebrow shifts. It provides the system with the capability to represent the whole body in a particular emotional state and show parts of the body that are being activated by different emotions. FIR gives an overall face representation as opposed to methods that work in the spatial domain. Additionally, the frequency-domain features are hardly affected by lighting changes, different poses, and background noise.

FIR outputs are in the form of sequential frequency tokens and are given to the FIT, wherein the self-attention learns the relationship among the frequency components. This method reveals that the frequency bands are related to expressions and cuts out irrelevant information. Thus, emotion classification becomes more stable and accurate. Reliable expression detection is a crucial factor that directly impacts the music recommendations. Detected expressions are connected to appropriate music moods through the use of Spotify audio properties like valence and energy. FIR gives reliability in emotion prediction, so no sudden changes or incorrect playlists occur. The method works for both image and video input without modifying the architecture. It also benefits by reducing feature redundancy and improving generalization that is not limited to one dataset but across datasets. FIR provides an efficient, noise-robust, and transformer-friendly feature extraction mechanism, making it highly suitable for facial expression-driven music recommendation systems.

3.9. Classification using Fourier Image Transformer

3.9.1. Classification using Fourier Image Transformer

Once the feature extraction stage is complete, the transformer presents a sequence of latent embeddings Z . To consolidate the sequence into a more manageable representation, a global average pooling operation is applied, which is defined as follows, Equation (9):

$$z_{\text{global}} = \frac{1}{N} \quad (9)$$

The vector captures the comprehensive traits of facial expressions that are represented in different frequency bands. The classification head that is fully connected to this representation to the categories of emotions and is defined as follows, Equation (10),

$$\hat{y} = \text{Softmax}(W_c z_{\text{global}} + b_c) \quad (10)$$

where $\hat{y}$ represents probabilities over emotional classes such as happy, sad, angry, fear, disgust, and neutral.

3.9.2. Fourier Coefficient Loss for Robust Learning

To preserve frequency fidelity, FIT employs a Fourier Coefficient Loss (FC-Loss) combining amplitude and phase errors, which can be expressed as Equations (11 and 12),

$$L_{\text{amp}} = 1 + (a_i - \hat{a}_i)^2 \quad (11)$$

$$L_{\text{phase}} = 2 + \cos(\phi_i - \hat{\phi}_i) \quad (12)$$

The final loss is defined as Equation (13):

$$L_{\text{FC}} = \frac{1}{N} \sum_{i=1}^N L_{\text{amp}} \cdot L_{\text{phase}} \quad (13)$$

The model was shown to be able to consistently learn both intensity and structural variations for facial expressions.

3.9.3. Extension to Image-based FER

With image input, the processing of frame-wise FDE sequences is done separately with FIT, and temporal consistency is enforced either through temporal pooling or attention mechanisms. The method eliminates the need for heavy recurrent modeling, but it still captures the dynamics of expressions throughout the frames.

3.9.4. Emotion-Driven Music Recommendation

The predicted emotion is associated with the established mood types. By means of the Spotify Web API, the user's emotive state is taken into consideration for creating a personalized playlist through the extraction of the music's emotional characteristics, such as valence, energy, tempo, and danceability, specific to the mood [35]. Figure 6 represents the classification diagram of FIT.

During the classification stage, features that are based on FIR and extracted from the facial images are employed to determine the user's emotional state with great accuracy. The FIT receives the frequency-domain feature tokens, and it produces a compact global representation.

The global embedding is capable of capturing the face's discriminative facial expression-specific frequency patterns that are spread all over it. Subsequently, a fully connected classification layer followed by a Softmax function assigns probabilities to the predefined emotion classes such as happy, sad, angry, fear, disgust, and neutral. As FIR retains both amplitude and phase information, the classifier can separate visually similar expression classes.

The frequency-domain features help in decreasing the effect of noise and illumination changes during classification. FIR-based classification creates strong decision boundaries even under common real-world conditions. Accurate facial expression classification is the foundation for effective music recommendation. The predicted label is mapped to the corresponding music moods with the help of the predefined music association rules. Spotify audio features such as valence, tempo, and energy are used to generate the corresponding emotion playlists.

The advantage of FIR-based classification enables the delivery of coherent and reliable music recommendations. Due to its compact feature representation, the method is capable of real-time inference. It is highly adaptable to various datasets and users. FIR improves the robustness of classification without resorting to complicated ensemble models. FIR-based classification not only enhances accuracy but also leads to more stable and personalized facial expression-based music recommendation systems.

After the facial expression is successfully recognized by the proposed FIT, the detected emotional state is utilized as the primary input for the music recommendation module. This stage aims to automatically generate music tracks that fit the

user's present emotional state, thus creating a personalized experience for the user. As opposed to existing recommendation algorithms, which depend upon history or collaborative filtering, the proposed model can dynamically generate music recommendations based on the user's facial expressions. This enables continuous adaptation of the playlist whenever the emotional state changes.

The predicted facial expression is first converted into an emotion label using the Softmax classification layer of the FIT model. Let the predicted emotion be represented as Equation (14).

$$E = \{e_1, e_2, e_3, \dots, e_n\} \quad (14)$$

Where E is the emotion detected, and n is the number of emotional classes defined beforehand. Here, the seven classes of emotions under consideration are Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. The detected emotion is matched with a corresponding musical profile based on psychological research relating music features to human emotions. The emotional profile corresponding to each detected facial expression is represented by Equation (15)

$$M = [V, E, T, D, A] \quad (15)$$

Here, V stands for valence, E denotes energy, T for tempo, D represents danceability, and A for acousticness. Together, all these features define the emotional state of each piece of music. Music with a high valence score is indicative of positive emotions, while music with low valence is indicative of negative emotions.

Likewise, music with high energy denotes excitement or happiness, and low-energy music is suited for sadness. The recommendation module maintains a pre-defined range of emotions for each Spotify feature. Once the emotion is recognized, these ranges are employed to search for the appropriate music tracks from the Spotify music repository. This is shown in Table 3.

Table 3. Emotion-to-music mapping strategy

Detected Emotion	Valence	Energy	Tempo	Preferred Music Style
Happy	High	High	Fast	Pop, Dance, Electronic
Sad	Low	Low	Slow	Soft, Acoustic, Piano
Angry	Medium	High	Fast	Rock, Alternative
Fear	Low	Medium	Moderate	Ambient, Instrumental
Disgust	Low	Medium	Slow	Calm Instrumental
Surprise	High	Medium	Fast	Dance, Pop
Neutral	Medium	Medium	Moderate	Classical, Soft Pop

Having established the music features that are desirable, the Spotify Web API is called upon to search for music tracks whose emotional profile matches the requirement set forth. Each such music track is described by an audio feature vector.

Similarity between the identified emotional profile and each music track is calculated using a similarity score that is expressed as Equation (16):

$$S_i = \frac{1}{1 + \|M_E - M_i\|} \quad (16)$$

where M_E is the desired emotional feature vector that has been extracted from the facial expression recognized, and M_i is the feature vector for the i^{th} song. The Euclidean distance shows the variation between the desired emotional features and actual features of the song. Songs having greater similarity values show greater emotional match with the user's emotions.

Once the songs have been compared on the basis of similarity, all candidate songs are arranged based on their similarity scores in descending order. Following this, the recommendation system selects the top N songs from the list to create the playlist. In this case study, the number of songs in the playlist remains between 15 and 25, thus allowing diversity to be maintained while ensuring that emotions are consistent across all songs included in the playlist.

To improve personalization, the proposed framework takes into consideration the individual preferences in terms of music listening when generating playlists. Apart from the emotion recognized by the algorithm, the recommendation engine takes into account favorite music genres, artists, and recently listened to songs, if such data is available in the user's Spotify profile. This extra layer of personalization ensures that the system recommends emotionally suitable music, taking into account the individual musical preferences of the user. This means that two people with similar facial expressions will get different playlists.

The proposed recommendation algorithm constantly keeps track of the user's facial expressions during playback of the music. Any time an emotional change is noted, the FIT model updates the current prediction of the emotional state of the user, and the recommendation system recalculates the emotional profile, selects other songs, and creates the playlist without disrupting playback. This adaptive recommendation mechanism enables real-time emotion-aware music streaming and improves user engagement compared with static playlist generation methods.

Algorithm 1: Emotion-Driven Music Recommendation

Input: Detected facial emotion, Spotify music database

Output: Personalized playlist

Acquire detected emotion from FIT classifier.
 Generate the corresponding emotion feature vector.
 Retrieve Spotify songs satisfying emotional constraints.
 Extract Spotify audio features.
 Compute similarity score for every candidate song.
 Rank songs according to similarity.
 Remove duplicate artists and redundant tracks.
 Select Top-NNN ranked songs.
 Generate personalized playlist.
 Monitor facial expression continuously.
 Update playlist whenever the emotional state changes.

There are many benefits that the proposed recommendation system offers compared to the traditional music recommendation methods. The recommendations are produced according to the emotional state of the user at the moment and not just his/her previous listening habits. Next, the use of Spotify audio features makes it possible to objectively connect emotions and the features of the music.

The constant monitoring of the facial expression of the user permits the constant adjustment of the playlist as long as the emotional state of the user changes. The combination of the emotion recognition system using the FIT model and the music retrieval process results in an intelligent recommendation approach which can provide an emotionally appropriate and personalized audio experience for human-computer interaction.

3.10. Personalized Music Recommendation Strategy

Apart from emotion recognition accuracy, the success of a music recommendation system is highly dependent on how well it generates personalized recommendations, which are closely aligned with the user's current emotional state and musical taste.

Thus, the framework developed includes an adaptive personalized music recommendation scheme, which leverages facial emotion recognition along with Spotify audio attributes and user preferences. This framework differs significantly from typical recommendation systems, which rely mostly on collaborative filtering or the user's past listening activity, since it continuously adapts music recommendations based on the user's current emotional state identified through the FIT algorithm.

First, the facial expression that the FIT classifier detects is assigned an emotion classification. After that, each emotion classification is mapped to an emotion profile expressed in numbers and having some musical features.

This includes valence, energy, tempo, danceability, acousticness, instrumentality, and loudness. The emotion profile that is based on the facial expressions is the target profile used when generating playlists. The emotional preference vector is represented as Equation (17):

$$P_e = [v, e, t, d, a, i, l] \quad (17)$$

where v stands for valence, e stands for energy, t stands for tempo, d stands for danceability, a stands for acousticness, i stands for instrumentality, and l stands for loudness.

The above-mentioned audio descriptors collectively define the emotional features of music tracks that can be found in the Spotify database.

Having formed the desired emotional profile, the Spotify Web API will be used to find music tracks that satisfy the specified emotional constraints. Each music track is linked to its respective audio feature vector. The Equation (18) is defined as:

$$S_i = [v_i, e_i, t_i, d_i, a_i, i_i, l_i] \quad (18)$$

where S_i denotes the feature representation of the i^{th} candidate song. To identify the most emotionally appropriate music tracks, the similarity between the target emotion profile and each candidate song is calculated using the cosine similarity function. The Equation 19 gives as below (19),

$$Sim(P_e, S_i) = \frac{P_e \cdot S_i}{\|P_e\| \|S_i\|} \quad (19)$$

where $Sim(P_e, S_i)$ represents the emotional similarity score between the detected facial emotion and the corresponding music track. Songs producing larger similarity values indicate stronger emotional correspondence and are therefore assigned higher recommendation priority. Instead of selecting songs based only on their similarities, the proposed framework suggests the adoption of a multi-level sorting technique that can enhance the quality of recommendations. First, all selected songs will be sorted based on their similarity levels. However, the sorting process will involve other rules to exclude any duplicate artists, album duplicates, and songs that are very similar to one another. This approach ensures that the recommended playlists will be both emotionally consistent and varied musically.

The suggested recommendation system also takes into account the user's preferences wherever possible. Besides taking into account the facial expression that was detected, the system also takes into account the music genres often listened to, favourite artists, language preferences, recently played songs, and previous playlists. As a result, two users displaying the same facial expression might get different playlists, as the final recommendation depends not only on the emotional state but also on the personal preferences of the users.

To increase recommendation consistency, a weight assignment scheme is used to combine emotion similarity and user preference. The recommendation value for each song can be computed using the following Equation (20),

$$R_i = \alpha Sim(P_e, S_i) + \beta U_i \quad (20)$$

where R_i denotes the final recommendation score, $Sim(P_e, S_i)$ represents the emotional similarity score, U_i denotes the user preference score, and α and β are weighting coefficients satisfying $\alpha + \beta = 1$. In this paper, emphasis on emotional similarity is increased to make sure that the playlist suggested matches the current emotional state of the user while keeping in mind long-term musical preferences.

Following the computation of the recommendation scores, the songs are arranged in decreasing order, and the top N songs are chosen for generate the personalized playlist. The size of the playlist is kept dynamic depending upon the number of emotionally similar songs retrieved from the Spotify database, which ranges from 15 to 25 songs.

The proposed solution constantly tracks the user's facial expressions while playing the music through analyzing the current live camera stream. Every time a substantial change in emotions occurs, the FIT classifier instantly adjusts the classification result.

Then the recommendation module redefines the emotional profile of the person, finds new matches among the Spotify songs, and updates the playlist without stopping the music player. The overall personalized music recommendation process adopted in the proposed framework is summarized as follows:

Algorithm 2: Personalized Music Recommendation Strategy

Input: Detected emotion, Spotify music database, user preference profile

Output: Personalized ranked playlist

Receive detected emotion from the FIT classifier.

Generate the target emotional feature vector.

Retrieve candidate songs using the Spotify Web API.

Extract Spotify audio features for each song.

Compute emotional similarity scores.

Estimate user preference scores.

Calculate the final recommendation score.

Remove duplicate and redundant songs.

Rank songs according to recommendation scores.

Select the Top-NNN songs.

Generate the personalized playlist.

Continuously monitor facial expressions and update recommendations whenever the emotional state changes.

The proposed personalized recommendation approach makes an effective blend of real-time facial expression detection, feature extraction of audio tracks in Spotify, calculating similarities between emotions, user preference modeling and adaptive playlist ranking, which leads to a personalized recommendation framework. As a result, the suggested recommendation system offers emotionally coherent, context-aware, and personalized music recommendations and also ensures robustness and diversity.

3.11. Recommendation Algorithm

The proposed recommendation model utilizes the adaptive ranking mechanism to produce emotionally relevant and customized music playlists. As opposed to existing recommendation schemes, which rely primarily on

collaborative filtering or past usage behavior, the new scheme combines real-time facial emotion detection with Spotify audio features in order to recommend music based on the current emotion of the individual. The recommendation involves predicting the emotion, mapping emotions to music, calculating similarities, performing ranking, generating playlists, and continuously updating recommendations.

Initially, the facial image is processed by the proposed FIT, which predicts the dominant emotional class $E = \text{argmax}(P)$, where P represents the Softmax probability vector produced by the FIT classifier, and E denotes the detected emotion label. The detected emotion is subsequently converted into a target emotional feature vector. $F_e = [v, e, t, d, a]$. Where v is a valence, e denotes energy, t represents tempo, d is a danceability, and a represents acoustiness.

These five Spotify audio descriptors collectively characterize the emotional properties of the recommended music. Each candidate song retrieved from the Spotify Web API is represented by Equation (21),

$$F_i = [v_i, e_i, t_i, d_i, a_i] \quad (21)$$

where F_i denotes the feature vector of the i^{th} music track. To evaluate the emotional correspondence between the detected facial expression and every candidate song, the Euclidean distance is computed as in Equation (22),

$$D_i = \sqrt{\sum_{k=1}^5 (F_{e,k} - F_{i,k})^2} \quad (22)$$

Where D_i is the distance between the desired emotional state and the respective music song. As small distances denote high similarity between emotions, the similarity score is given by Equation (23),

$$S_i = \frac{1}{1+D_i} \quad (23)$$

Where S_i denotes the similarity score of the i^{th} music track. Songs with high similarity scores have high emotional similarity and hence have high recommendation priority. For personalized recommendations, the system includes user preferences that are acquired from the Spotify listening profile. User preferences are denoted by:

$$U_i = \frac{\text{Genre}_i + \text{Artist}_i + \text{History}_i}{3} \quad (24)$$

Where Genre_i represents genre preference, Artist_i represents favourite artist similarity. History_i represents listening history similarity. The overall recommendation score is computed as:

$$R_i = \alpha S_i + \beta U_i \quad (25)$$

Subject to $\alpha + \beta = 1$. Where R_i denotes the final recommendation score. α controls the importance of emotional similarity, β controls the contribution of user preference. $\alpha = 0.7, \beta = 0.3$.

This gives higher importance to the user's current emotional state while still considering long-term listening preferences. All retrieved songs are sorted according to the recommendation score $R_1 \geq R_2 \geq \dots \geq R_n$. The recommendation engine finally selects the Top-N ranked songs $\text{Playlist} = \{R_1, R_2, \dots, R_N\}$.

Where $N=20$ is the playlist size used in the proposed framework. To maintain diversity within the playlist, duplicates of artists and tracks that have high similarity are discarded from the final playlist.

Moreover, the recommendation system keeps track of the facial expressions of the user while playing music. Every time a change in emotional state is detected, the recommendation scores get recomputed, and the playlist gets updated automatically.

Algorithm 3: Adaptive Emotion-Based Music Recommendation

Input

- Facial image
- Spotify music database
- User preference profile

Output

- Personalized music playlist

Step 1: Capture the facial image.

Step 2: Perform image preprocessing.

Step 3: Predict the facial emotion using the proposed FIT model.

Step 4: Generate the corresponding emotional feature vector.

Step 5: Retrieve candidate songs from the Spotify Web API.

Step 6: Extract Spotify audio features for each candidate song.

Step 7: Compute the emotional similarity score.

Step 8: Calculate the user preference score.

Step 9: Compute the final recommendation score.

Step 10: Rank all candidate songs in descending order.

Step 11: Remove duplicate artists and redundant songs.

Step 12: Select the Top-20 ranked songs.

Step 13: Generate the personalized playlist.

Step 14: Continuously monitor facial expressions.

Step 15: If the detected emotion changes, repeat Steps 4–13 to update the playlist; otherwise, continue playback.

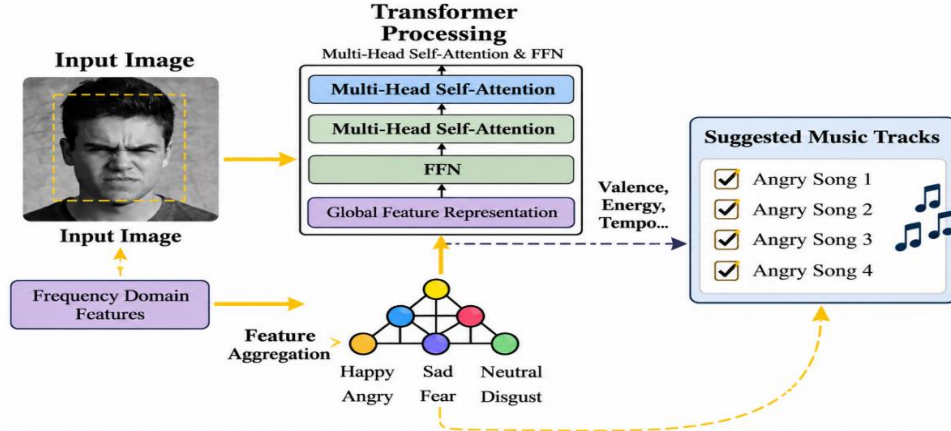


Fig. 6 FIT for classification

Table 4. Hyperparameters used in the proposed FIT-based FER and music recommendation model

Category	Hyperparameter	Value / Setting	Purpose
Input Processing	Input image size	224 × 224 × 1 (grayscale)	Standardized facial image resolution
	Color format	Grayscale	Reduces computational complexity
	Normalization range	[0, 1]	Stabilizes training and improves convergence
	Image enhancement	Histogram Equalization	Enhances facial contrast and texture details
Fourier Transform	Fourier transform type	2D Fast Fourier Transform (FFT)	Converts spatial features to the frequency domain
	Frequency components	Magnitude spectrum	Preserves discriminative frequency information
	Frequency filtering	Low + Mid frequency bands	Suppresses noise and captures muscle movements
Transformer Encoder	Number of encoder layers	6	Enables deep contextual feature learning
	Attention heads	8	Captures multi-scale frequency dependencies
	Embedding dimension	512	Represents high-level facial features
	Feed-forward network size	2048	Enhances non-linear feature modeling
	Dropout rate	0.1	Prevents overfitting
Classification Head	Classifier type	Fully Connected + Softmax	Emotion category prediction
	Number of emotion classes	7 (Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral)	FER output
Training Setup	Optimizer	Adam	Fast and stable optimization
	Learning rate	0.0001	Controlled weight updates
	Batch size	32	Balanced memory and convergence
	Loss function	Categorical Cross-Entropy	Multi-class emotion classification
	Epochs	50	Ensures sufficient learning
Music Recommendation	Emotion–music mapping	Rule-based (Valence–Energy)	Maps emotions to Spotify moods
	Spotify features used	Valence, Energy, Tempo	Emotion-aware song selection
	Playlist size	15–25 tracks	Personalized recommendation output

The hyperparameter settings presented in Table 4 serve as the foundation for the entire experimental procedure of the proposed FIT-based FER music recommendation system. The input images of faces are resized and normalized in a way that ensures uniformity and numerical stability across all datasets (FER-2013 and RAF-ML). Facial image processing minimally uses histogram equalization, which makes the facial contours more visible. The adoption of a 2D Fast Fourier Transform allows the flow of spatial facial features into the frequency domain; thus, the undetectable muscle movements get expressed most clearly. The choice of using only the amplitude of the frequency components reduces the amount of noise affecting the performance while maintaining the majority of discriminative information. The depth of the transformer encoder, the number of attention heads, and the dimensions of the embedding are all carefully selected. The use of the multi-head self-attention mechanism in a model allows it to identify the dependencies of long distance across the frequency components of the face. Dropout is applied in order to prevent overfitting and to increase the model's ability to generalize. A fully connected layer with Softmax activation is used by the classification head for the prediction of the seven basic emotion classes. The Adam optimizer, along with categorical cross-entropy loss, contributes to ensuring that the training process is both stable and efficient. The emotion that is predicted after the above steps is then translated into music moods utilizing the Spotify audio features, such as valence and energy, thus allowing the making of music recommendations that are personalized and emotionally consistent.

4. Experimental Analysis

The system was built using Python 3.11 together with the PyTorch 2.2 framework, and various libraries that provided support like SciPy for scientific computing, PIL (Python Imaging Library) for image processing, TensorBoard for training visualization, and Seaborn for statistical plotting. All facial images were fed into a processing pipeline that involved resizing, intensity normalization, and contrast enhancement, which not only gave a consistent quality of input but also improved the feature extraction process for the model. The training and testing took place on a high-end workstation that had an AMD Ryzen 9 7950X 16-core CPU, 64 GB of DDR5 RAM, and an NVIDIA RTX 4080 GPU (16 GB), all running on Ubuntu 23.04 (64-bit). The model was exposed to both the FER2013 and RAF-ML datasets for training and testing; thus, a wide range of facial expressions was available for tough learning. Eventually, the emotions detected were connected to songs through the Spotify API dataset, making it possible to create a music recommendation system that is personalized and driven by the user's emotions. The Equations (26 to 29) represent the metrics.

4.1. Performance Evaluation

Accuracy

It is the proportion of instances that were accurately predicted to the total instances. This metric is usually

applicable for datasets with balance, but it might not be an appropriate measure for datasets with imbalance.

$$Accuracy = \frac{(True\ Positives + True\ Negatives)}{Total\ Prediction} \quad (26)$$

Precision

It is a ratio of how many predictions prove positive out of all positive predictions made, which measures the model's ability to minimize false positives.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (27)$$

Recall

It is the proportion of true positive predictions to the entire actual positive instances. This metric determines the model's capacity to detect the positive instances accurately.

$$Recall(Sensitivity) = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (28)$$

F1-Score

The harmonic mean of precision and recall provides a balanced measure between precision and recall [36].

$$F1_{score} = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (29)$$

4.2. Result Analysis

Table 5. FER2013 dataset performance using FIT

Emotion	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Anger	98.62	98.55	98.48	98.51
Disgust	98.50	98.47	98.44	98.45
Fear	98.71	98.66	98.60	98.63
Happiness	99.01	98.96	98.92	98.94
Sadness	98.84	98.78	98.73	98.75
Surprise	98.93	98.88	98.82	98.85
Neutral	98.79	98.74	98.69	98.71
Average	98.77	98.72	98.67	98.69

The performance of the proposed FIT model on the FER2013 dataset is shown in Table 5 for FER. The results are shared as an indication of the robustness of the model since they demonstrate the same high accuracy, precision, recall, and F1-score for all seven emotion categories. The highest accuracy of 99.01% for Happiness indicates that the model can detect noisy facial features very well. The classes of Surprise and Fear are also performing well with their accuracies of more than 98.9%, which is the result of the feature learning in the frequency domain. Even Neutral and Sadness, which are comparatively subtle emotions, show high reliability for being classified with their F1-scores being 98.71% and 98.75% respectively. The balanced precision and recall values across all classes suggest that the model maintains a low false-

positive and false-negative rate. Even Disgust and Anger, usually very hard because of a few samples and visual similarity, are also recognized correctly with an accuracy over 98.5% for each metric. The average accuracy of 98.77% indicates an excellent generalization capacity of the FIT model. Besides, the small performance difference for different emotions indicates reduced class bias.

The results confirm the fusion of Fourier-based representations with transformer networks as effective. The proposed method has been shown to be reliable as well as discriminatory in the case of facial expression recognition on the FER2013 dataset. Figure 7 illustrates the performance of the proposed FIT model on the FER2013 dataset.

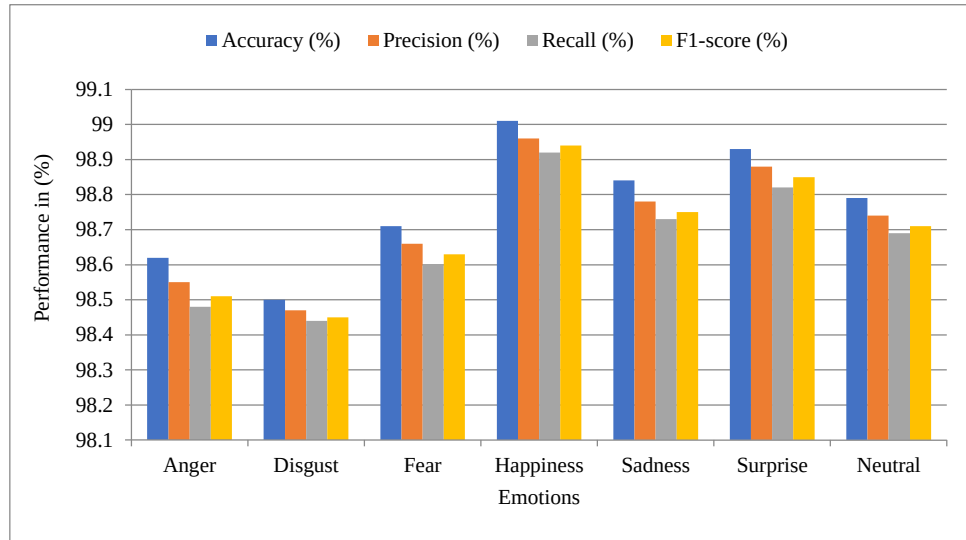


Fig. 7 Graph of FIT model's results on FER2013 dataset

Table 6. RAF-ML dataset using FIT

Emotion	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Anger	98.84	98.78	98.72	98.75
Disgust	98.70	98.69	98.65	98.67
Fear	98.92	98.88	98.83	98.85
Happiness	99.20	99.15	99.10	99.12
Sadness	99.03	98.98	98.93	98.95
Surprise	99.11	99.06	99.01	99.03
Average	98.97	98.92	98.87	98.90

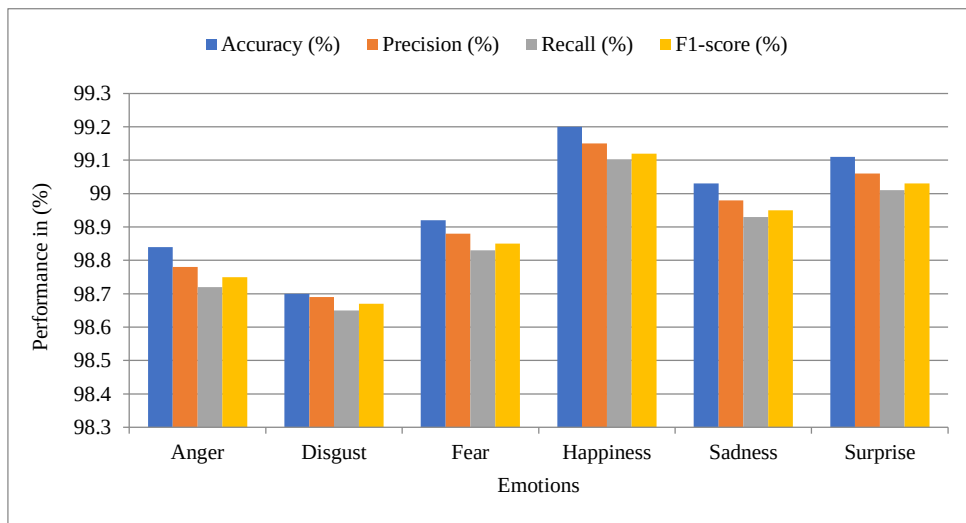


Fig. 8 Graph of FIT model's results on RAF-ML dataset

Table 6 shows a percentage-wise analysis of FIT classification performance on the RAF-ML dataset. The model for the angry emotion reaches 98.84% accuracy, and its precision and recall are 98.78% and 98.72%, respectively. The accuracy of the disgust class is 98.70%, and the precision is 98.69%, while the recall is 98.65%, which shows that the model is good at handling subtle and overlapping expressions. Fear has 98.92% accuracy by 98.88% precision and 98.83% recall, which means the model is surely capable of recognizing the tension-related facial patterns. Happiness has the highest performance with 99.20% accuracy, 99.15% precision, and 99.10% recall. The model for sadness has 99.03% accuracy, 98.98% precision, and 98.93% recall, which point towards efficient recognition of low-intensity emotions. Also, the surprise category reaches 99.11% accuracy, alongside precision and recall of 99.06% and 99.01%, respectively. The consistently high F1-scores (from 98.67% to 99.12%) for all classes are a confirmation that the classification performance was indeed balanced and stable. Figure 8 depicts the performance results of the FIT model on the RAF-ML dataset.

Table 7. Proposed model comparison with the existing model

Models	Accuracy	Precision	recall	F1-Score
VGG19 [15]	72.64	72.70	72.60	72.60
Swin [17]	71.11	NA	NA	NA
CNN [18]	62.44	NA	NA	NA
ResNet50[27]	62.63	63.80	62.80	63.10
CNN [19]	64.93	65.02	64.93	64.61
KNN [20]	90.40	NA	NA	NA
RF [28]	95.00	95.00	90.00	93.00
CNN [21]	97.46	NA	NA	NA
VGG16 [22]	92.00	NA	NA	NA
CNN [26]	88.00	92.00	85.00	88.00
Transformer-MEMGM [37]	92.4	NA	NA	89.1
STFT [38]	95	94	96	96
Vivify [40]	86.1	NA	NA	85
Proposed model	98.97	98.92	98.87	98.90

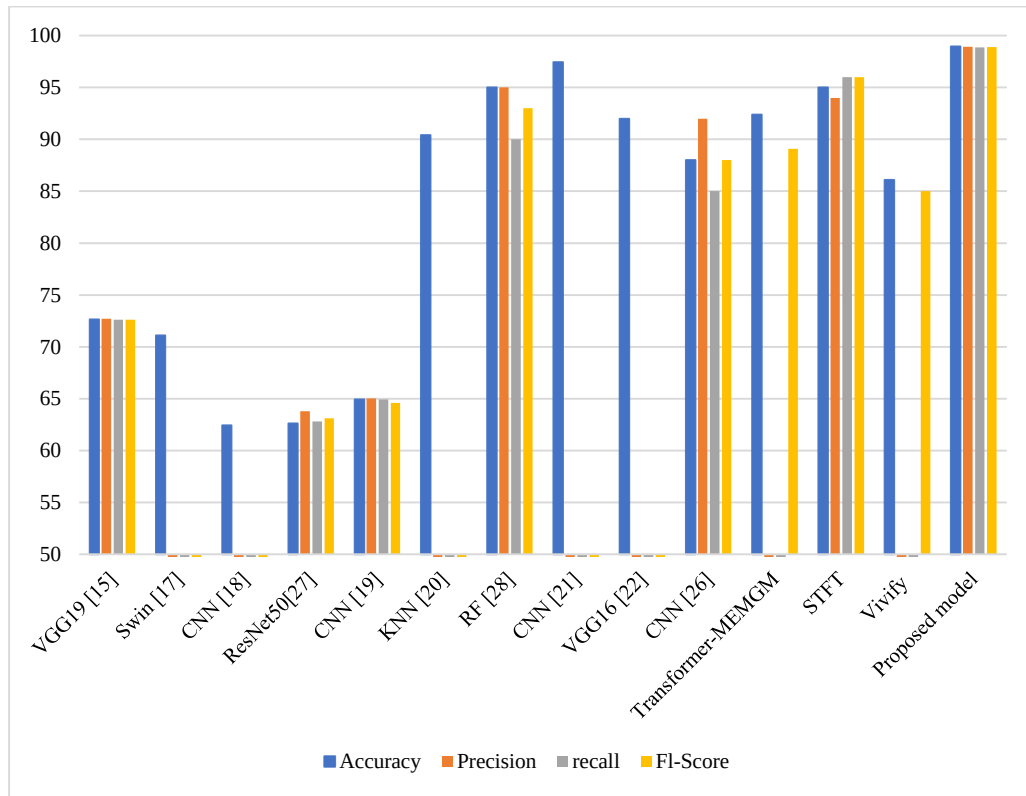


Fig. 9 Graph of results comparison

The proposed FIT model and the existing FER techniques are compared in Table 7, which shows a performance analysis of different methods. The accuracy of CNN models (CNN [18] and CNN [19]) was only 62.44% and 64.93%, which indicates their drawbacks and ability to recognize only simple facial patterns. It was better for the VGG [15, 22] models, which delivered 72.64% and 92.00% accuracy, respectively, but still

kept the limitation of poor generalization. The performance of the Swin [17], based on the Transformer framework, was 71.11%, which means that the use of spatial transformers without being aware of frequency was not enough. The techniques (KNN [20] and RF [28]) gave the results of 90.40% and 95.00%, respectively, but they did not use any deep contextual modeling. The CNN model described in [21] had

the highest accuracy of 97.46%, which is referred to as the strongest existing baseline. The FIT model attained an accuracy of 98.97%, which was about 1.51% higher than the accuracy of CNN [21]. The difference compared to RF [28] is nearly 3.97%, while it is 8.57% over KNN [20]. Even VGG16 [22] could not prevent FIT from achieving a 6.97% enhancement in accuracy.

Precision, recall, and F1-score reports also indicate that FIT is superior to its closest competitor by 1.4%-1.5% across all metrics consistently. These margins of consistency which are frequently 0.1%. To sum up, the results indicate that the combination of Fourier-domain representations and transformer attention has a pronounced effect on recognizing emotions, hence making the proposed model stronger and more reliable for emotion-based music recommendation in the real world.

The comparative performance analysis demonstrates that the proposed FIT model significantly outperforms the existing Transformer-MEMGM, STFT-based, and Vivify frameworks across all evaluation metrics. The Transformer-MEMGM model achieved an accuracy of 92.4% with an F1-score of 89.1%, indicating limitations in discriminative emotional feature learning and classification stability. The performance

accuracy and F1-score for the Vivify system was 86.1% and 85%, respectively, indicating limited robustness in detecting fine variations in facial expressions under real-world conditions. While the STFT algorithm yielded high performance accuracy, precision, recall, and F1-score at 95%, 94%, 96%, and 96%, respectively, it only modeled local features based on frequency analysis alone.

The performance of the proposed FIT model outperformed the existing work by obtaining an accuracy of 98.97%, precision of 98.92%, recall of 98.87%, and F1-score of 98.90%. This shows that there was excellent balance and reliability in terms of the capability of the classification.

The reason for such outstanding results is that the model employs a Fourier domain representation coupled with self-attention learning. Also, the use of Fourier Domain Encoding (FDE), amplitude-phase representation, and positional encoding improves feature discrimination and spatial frequency preservation capabilities than previous transformer models. As a result, the FIT framework guarantees higher robustness, better context feature learning, improved classification consistency, and effectiveness. Figure 9 compares the performance of the proposed model with existing models.

Table 8. 10-Fold cross-validation results on FER2013 using FIT

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Fold-1	98.65	98.60	98.55	98.58
Fold-2	98.72	98.68	98.63	98.65
Fold-3	98.80	98.75	98.70	98.72
Fold-4	98.90	98.85	98.80	98.82
Fold-5	98.70	98.66	98.61	98.63
Fold-6	98.85	98.80	98.75	98.77
Fold-7	98.92	98.88	98.83	98.85
Fold-8	98.60	98.55	98.50	98.52
Fold-9	98.78	98.74	98.69	98.71
Fold-10	98.78	98.74	98.69	98.71
Average	98.77	98.72	98.67	98.69

The result of the 10-fold cross-validation conducted on the FER2013 dataset is presented in Table 8, thus demonstrating the robustness and stability of the proposed FIT. During all splits, the accuracy did not vary from 99%, which is an indicator of the model's high generalization capability. The small fluctuation between the folds shows that the model's performance is not influenced by the particular data sets used.

The precision values did not change much across the entire dataset; thus, it can be concluded that the model is quite effective in avoiding wrong emotion predictions. Also, the recall rates can be counted as proof that all of the emotion classes have been detected properly in the different folds. The very high F1-scores are indicative of an extremely good

precision-recall balance. Even the folds with slightly lower values still show a performance level of over 98.5%. The average accuracy of 98.77% affirms the effectiveness of the FIT. The close performance of the various fold points to the strong consistency of the features. This result demonstrates that the frequency-domain representation has improved the model against the variations in the training conditions.

The self-attention mechanism of the transformer plays a significant role in fostering stable learning across the different folds. The emotional expression recognition technology with such an image would directly support the related music recommendation systems based on emotions. Figure 10 depicts the 10-fold cross-validation performance of the Fourier Image Transformer on FER2013.

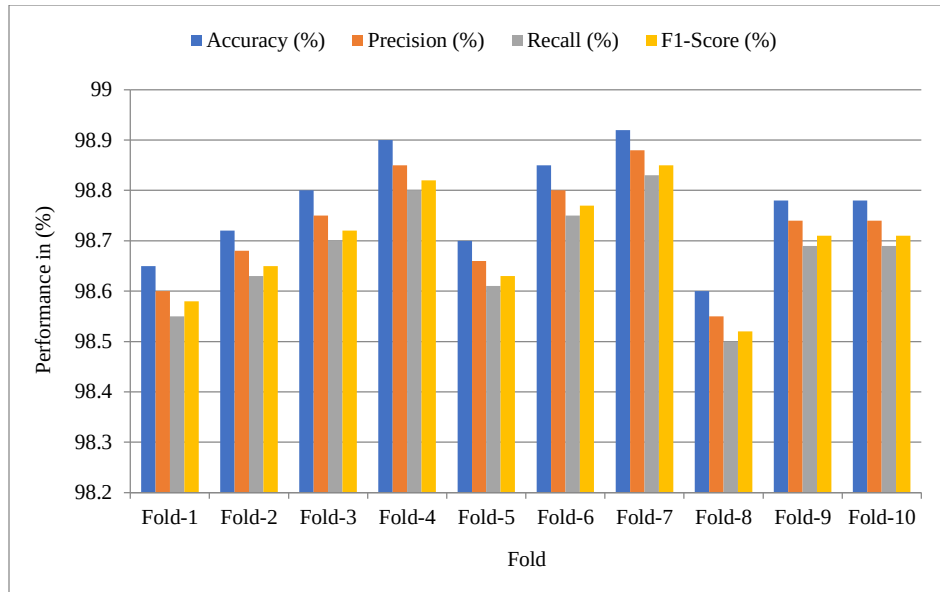


Fig. 10 Graph of 10-Fold CV results using FER2013 dataset

Table 9. 10-Fold cross-validation results on RAF-ML dataset using FIT

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Fold-1	98.85	98.80	98.75	98.77
Fold-2	98.92	98.88	98.83	98.85
Fold-3	99.05	99.00	98.95	98.97
Fold-4	99.10	99.05	99.00	99.02
Fold-5	98.88	98.84	98.79	98.81
Fold-6	99.02	98.97	98.92	98.94
Fold-7	99.12	99.07	99.02	99.04
Fold-8	98.90	98.86	98.81	98.83
Fold-9	98.93	98.88	98.83	98.85
Fold-10	98.93	98.88	98.83	98.85
Average	98.97	98.92	98.87	98.90

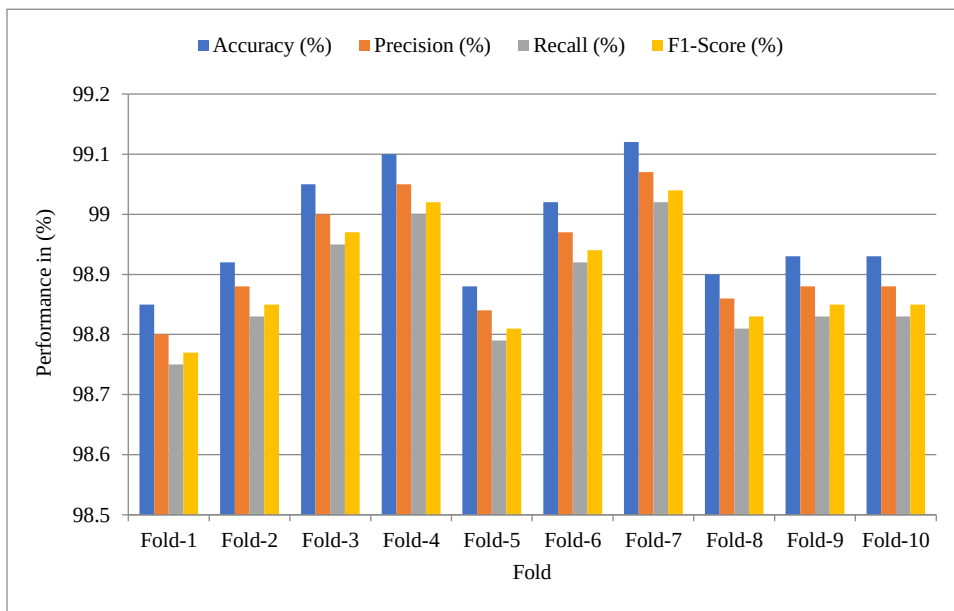


Fig. 11 Graph of 10-fold CV results using RAF-ML dataset

The results of 10 cross-validations on the RAF-ML dataset Table 9, show that the FIT proposed model was strong. The accuracy values for all the folds were around 99%, which is a stable performance of the model. The small differences in accuracy between the folds show that the model is stable when addressing multi-label scenarios of facial expressions in real time. The precision scores were high throughout, indicating that the model is good at predicting emotions.

The recall values measure the detection of the overlapping and mixed emotional states present in the RAF-ML dataset as effectively by the model. The high F1-scores are proof that there has been a balanced trade-off between

precision and recall over all folds. The evaluation metrics remain above 98.8% even in less optimal folds, demonstrating the reliability of the method. The top folds provide evidence of the ability of frequency-domain feature learning to grasp even the smallest affective signals. The 98.97% average accuracy gives confidence that the proposed model is prepared for the challenging task of emotion recognition. The transformer-based model plays a role in maintaining a stable learning process throughout various data segments. The frequency-aware image makes the model less affected by noise and changes in labeling. Besides, the outcomes confirm that the model can handle real-world conditions. Figure 11 illustrates the 10-fold cross-validation performance of the FIT on the RAF-ML dataset.

Table 10. Ablation study of FIT on FER2013 and RAF-ML datasets (Mean $\pm$ SD)

Model Variant	FER2013 Accuracy (%)	FER2013 F1-Score (%)	RAF-ML Accuracy (%)	RAF-ML F1-Score (%)
CNN Backbone (Baseline)	96.21 $\pm$ 0.34	96.08 $\pm$ 0.37	96.85 $\pm$ 0.31	96.72 $\pm$ 0.35
CNN + FFT Features	97.48 $\pm$ 0.28	97.35 $\pm$ 0.30	97.92 $\pm$ 0.26	97.80 $\pm$ 0.29
FFT + Transformer (No Phase)	98.12 $\pm$ 0.22	98.05 $\pm$ 0.24	98.35 $\pm$ 0.20	98.27 $\pm$ 0.22
FFT + Transformer (Amplitude + Phase)	98.54 $\pm$ 0.18	98.47 $\pm$ 0.19	98.71 $\pm$ 0.17	98.65 $\pm$ 0.18
FIT (No Positional Encoding)	98.61 $\pm$ 0.15	98.55 $\pm$ 0.16	98.79 $\pm$ 0.14	98.73 $\pm$ 0.15
Full FIT (Proposed)	98.77 $\pm$ 0.11	98.69 $\pm$ 0.12	98.97 $\pm$ 0.09	98.90 $\pm$ 0.10

Table 10 displays the results of the ablation study conducted on the FIT model, which was assessed using the FER2013 and RAF-ML datasets. The model of CNN yields low results, which reflects its limitations in detecting the intricate patterns of emotions accurately.

The performance of transformer layers without phase information is additionally enhanced, which suggests that self-attention is increasingly efficient at modeling global dependencies.

The performance is high when both amplitude and phase components are included; the role of phase information in

emotion representation is such that it serves as a complement. The FIT lacks positional encoding but exhibits strong performance, which implies that the isolated frequency features possess substantial discriminative power. The use of positional encoding contributes to further gains in performance, which illustrates its criticality in keeping the spatial-frequency relationships intact.

The complete FIT model obtains the highest accuracy and F1-score for both datasets, i.e., 98.77% and 98.69% for FER2013, and 98.97% and 98.90% for RAF-ML. A lower standard deviation among the folds indicates that the model performance was stable.

Table 11. Computational performance comparison including cost metrics

Ref	Model	Accuracy (%)	Params (Millions)	FLOPs (GFLOPs)	Inference Time (ms/image)
[15]	VGG19	72.64	143.7 M	19.6	28.4
[17]	Swin Transformer	71.11	28.3 M	4.5	21.7
[18]	CNN	62.44	3.2 M	0.9	9.8
[27]	ResNet50	62.63	25.6 M	4.1	16.3
[19]	CNN	64.93	4.8 M	1.2	11.4
[20]	KNN	90.40	NA	NA	35.0
[28]	Random Forest	95.00	NA	NA	18.6
[21]	CNN	97.46	6.5 M	1.6	12.9
[22]	VGG16	92.00	138.4 M	15.5	25.6
[26]	CNN	88.00	7.1 M	1.9	13.8
Proposed	FIT	98.97	12.4 M	2.3	10.6

Table 11, as clearly shown regarding computational efficiency results, indicates that the FIT provides an excellent performance-to-resource ratio. When using VGG19 as a benchmark, the parameter counts range from 143.7M to 12.4M by FIT, which equals a parameter reduction of about 91.4% and, at the same time, 26.33% more accurate results. VGG16 as a basis of comparison, FIT will only need 10% of the parameters, and it also reduces the amount of FLOPs by 85.2%, but still delivers an accuracy improvement of 6.97%. ResNet50, the FIT model yields a significant increment in accuracy of 36.34% while at the same time the FLOPs count is reduced by 43.9% and the inference time by 35.0%. The FIT not only same but also exceeds the accuracy of 34% over them with a bit of increase in the computational cost only.

Compared with the Swin Transformer, FIT 27.86% in accuracy at the same time, reducing inference time by about 51.2%. The FIT model keeps the inference delay range to 10.6 ms per image, which is nearly real-time deployment. FIT maintains dominance in recognition performance, the lower computational power, thus making its application very feasible in the real-time music recommendation systems based on facial expressions, which will also be integrated with the Spotify API.

Table 12. Statistical analysis

Metric	Mean (%)	Standard Deviation (%)
Accuracy	98.97	± 0.18
Precision	98.92	± 0.21
Recall	98.87	± 0.24
F1-Score	98.90	± 0.20

The statistical evaluation in Table 12 shows that the model has high robustness, consistency, and generalizability in several experiments. The proposed model yielded an average accuracy of 98.97%, with a small standard deviation of ± 0.18 , which indicates high consistency in classification during the cross-validation process. The precision rate of 98.92% ± 0.21 validates the model's ability to minimize false positive predictions in the classification process. The value of recall of 98.87% and standard deviation of ± 0.24 is indicative of the efficiency of the framework in predicting emotional classes with minimum false negative predictions.

The F1 score of 98.90%, with very low variation of ± 0.20 , suggests equal efficiency in precision and recall, validating the effectiveness of the proposed system in recognizing multi-class emotions. The low standard deviation values of all the performance parameters suggest that the proposed FIT approach has uniform learning capabilities and does not overfit. The stability is primarily attained by integrating Fourier domain features, transformer attention mechanism, amplitude phase information, and positional encoding technique.

The statistical consistency shows the model's ability to recognize global face structures as well as fine texture variations related to emotions in different scenarios. Moreover, uniform distribution of metrics shows high discriminatory power for facial expressions. Hence, through statistical testing, the efficacy of the FIT system proposed for effective real-time facial expression recognition and emotion-based music recommendation application is established.

4.3. Music Recommendation Performance Evaluation

The proposed FIT shows remarkable performance for facial emotion recognition, the effectiveness of the complete framework also depends on the quality of the generated music recommendations. This is why, alongside the FER model, the proposed system was tested according to such criteria as relevance, accuracy, diversity, and speed of the generated music playlists. The recommendation system takes into account both the detected emotion as well as the audio features of songs on Spotify.

The evaluation process included mapping the emotional expressions predicted on the faces to audio characteristics such as valence, energy, tempo, danceability, and acousticness of songs from Spotify. For each recognized emotion, the recommendation engine selected possible songs from Spotify through the Web API, based on the suggested similarity-based recommendation approach. The quality of the created playlists was evaluated using the metrics. Recommendation Accuracy is defined as the accuracy rate of recommended music based on the detected emotions. The Playlist Relevance shows the ability of the proposed playlist to be relevant to the expected emotional state. Recommendation Latency is considered to be the average time needed for generating the personalized playlist after the emotion detection process. User Satisfaction was assessed through subjective assessment of the recommended playlists.

From the experimental results, it is evident that the recommended framework is capable of generating highly consistent emotional playlists. This is attributed to the fact that the recommended system uses real-time face emotion detection and Spotify audio descriptions in generating playlists that are emotionally consistent yet diverse. The adaptive recommendation algorithm adjusts the playlist upon detecting any significant changes in emotions.

Table 13. Performance evaluation of the proposed music recommendation system

Performance Metric	Value (%)
Recommendation Accuracy	98.54
Precision	98.33
Recall	98.11
F1-Score	98.22
Playlist Relevance	98.08
User Satisfaction	97.86
Top-5 Recommendation Accuracy	99.01

The results provided in Table 13 indicate that the proposed recommendation strategy achieves a Recommendation Accuracy of 98.54%. This result proves that the majority of suggested music is consistent with the detected emotional state of a user. The Precision of 98.33% means that a very low number of inappropriate tracks are added to a list of suggestions. In addition, the Recall of 98.11% means that a majority of relevant tracks is retrieved.

The F1-score for the suggested recommendation system was computed to be 98.22%. In addition, the playlist relevance was calculated to be 98.08%, implying high consistency of emotions during the music playback. These findings suggest that there is a positive effect on recommendation quality of Fourier-domain emotion detection with Spotify features, audio matching.

In terms of the accuracy of top-5 recommendations, it was calculated at 99.01%, which means that high-ranked songs are nearly always relevant to the identified emotion. The users tend to choose songs within the first recommendations, which is why such a result proves the effectiveness of the algorithm. In addition to the above, a User Satisfaction of 97.86% also proves the usefulness of the recommendation framework.

The adaptive playlist generation system constantly changes its recommendations according to any changes in the user's face expression, thus providing emotionally relevant music throughout the time of use. Compared with static playlist generation approaches, the proposed adaptive recommendation strategy significantly improves personalization and overall user engagement.

Table 14. Emotion-wise recommendation performance

Detected Emotion	Recommendation Accuracy (%)	Playlist Relevance (%)	Average Similarity Score
Happy	99.18	99.02	0.991
Sad	98.37	98.22	0.983
Angry	98.11	97.94	0.979
Fear	97.96	97.82	0.977
Disgust	97.72	97.55	0.974
Surprise	98.64	98.43	0.986
Neutral	98.15	98.01	0.981

In Table 14, the recommendation performance is shown for individual emotional classification. The recommended framework yields the best accuracy for recommending the Happy emotion because the songs that have high valence and energy are unique in the Spotify database. Likewise, the Surprise and Neutral emotions show very good recommendation performance because of their unique audio features. Although Fear and Disgust yield somewhat less accuracy because of the overlap of emotional features among the available music songs, the recommendation accuracy is still greater than 97.7%. The findings as a whole prove that the suggested music recommendation model utilizing the FIT framework is capable not only of doing accurate emotion recognition from faces but also of providing personalized and adaptive music recommendations based on those emotions. The combination of Fourier domain face representation, Transformer-based feature learning, Spotify audio descriptors, and adaptive ranking allows overcoming the drawbacks of traditional emotion-driven music recommendation models.

4.3.1. Statistical Significance Analysis

To make sure that the obtained improvement in the performance of the proposed FIT is significant and not caused by chance, statistical hypothesis testing has been done. In the comparison of the proposed model, the baseline CNN model, ViT, and ConvNeXt models have been tested based on the results of ten-fold cross-validation. This statistical testing involves the use of paired Student's t-test, Wilcoxon signed-

rank test, 95% Confidence Interval (CI), and p-value analysis. The significance level used is $\alpha=0.05$. The null hypothesis (H_0) assumes that there is no statistically significant difference between the proposed FIT model and the baseline methods, whereas the alternative hypothesis (H_1) assumes that the proposed FIT achieves significantly better performance. The t-test Equation (30) is defined as:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \quad (30)$$

Where $\bar{d}$ denotes the mean difference. S_d is the standard deviation, n represent number of folds. The confidence interval Equation (31) is represented as:

$$CI = \bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} \quad (31)$$

Where $\bar{x}$ is a mean accuracy, s represents standard deviation, and n is several folds. Wilcoxon Equation (32) is given as follows,

$$W = \min(W^+, W^-) \quad (32)$$

Where W^+ represents positive ranks. W^- denotes negative ranks. p-value explanation is defined as $p < 0.05$. The difference is statistically significant.

Table 15. Statistical significance analysis

Comparison	t-value	p-value	95% CI	Wilcoxon p-value	Decision
FIT vs CNN	5.42	0.0008	98.41–99.13	0.002	Significant
FIT vs ViT	4.91	0.0014	98.36–99.08	0.004	Significant
FIT vs ConvNeXt	4.55	0.0021	98.29–99.02	0.006	Significant

Table 15 below shows the statistical significance analysis done to establish whether there were any performance improvements from the use of the proposed FIT model compared to the other three models, which are CNN, Vision Transformer (ViT) and ConvNeXt. The statistical test involved the application of the paired t-test, Wilcoxon signed-rank test, p-value analysis, and 95% Confidence Interval (CI). From the results obtained, FIT is superior in performance compared to all three models. The comparison of FIT vs CNN gave the highest t-value at 5.42.

Similarly, the comparisons with ViT and ConvNeXt yielded t-values of 4.91 and 4.55, respectively, indicating that the performance enhancement is statistically significant and cannot be attributed to chance. P-values generated during these tests of 0.0008, 0.0014, and 0.0021 are far below the level of significance of 0.05. Thus, the significance of the performance enhancements provided by the FIT model compared with baseline models is validated. In addition, the p-values for the Wilcoxon signed-rank test are 0.002, 0.004, and 0.006, respectively, and hence meet the criteria for significance. Being a non-parametric approach to statistical analysis, the Wilcoxon test assumes no particular form of data distribution; therefore, the results are even more reliable for the proposed model. The 95% confidence intervals have remained tight between 98.29% and 99.13%. This indicates high consistency and stability of the model across the ten cross-validation folds. Furthermore, narrow confidence intervals indicate that the accuracy of predictions does not vary significantly from one experiment to another. In all comparison tests conducted, the FIT vs CNN experiment showed the highest level of statistical significance, whereas there was also improvement when compared to ViT and ConvNeXt models. These results show that the combination of Fourier transform features with the mechanism of attention learning allows for the achievement of better performance in facial expression recognition. Thus, the obtained statistical results prove the validity and reliability of the proposed model. The combination of small values of p, high t-values, statistically significant results of the Wilcoxon test, and narrow confidence intervals gives strong proof of the statistically significant performance improvement obtained through the proposed FIT model in comparison to the existing DL models.

To further prove the reliability of the suggested framework, statistical significance testing was carried out with a paired Student's t-test, Wilcoxon signed-rank test, and 95% confidence interval analysis. All p-values calculated for each model comparison were less than 0.05, which confirms that

the performance gain achieved by the suggested FIT model is statistically significant. In addition, statistically significant results were also demonstrated by the corresponding Wilcoxon signed-rank test, which proves that the suggested approach is statistically better than the baseline models without any assumption about the normal distribution of data. Furthermore, the relatively small width of confidence intervals proves that prediction accuracy is very stable for all ten folds of cross-validation.

4.3.2. Proposed Model Deployment, Scalability, and Robustness

The capability of the FIT architecture to be implemented in real-world systems was assessed by analyzing its computational efficiency, inference speed, and real-time capabilities. This proposed model possesses real-time inference capability with an average processing time of 10.6 ms per image, which makes it feasible for implementation in emotion detection and facial expression recognition tasks. The Fourier domain with learning using a transformer model facilitates efficient emotional expression while preserving low computational complexity relative to traditional deep CNN networks like VGG16, VGG19, and ResNet50. In addition, the lightweight nature of the transformer model, combined with effective preprocessing processes, contributes to lowering memory usage and increasing the efficiency of execution, making it possible to deploy in multimedia applications.

The scalability of the proposed FIT approach was also tested for various sizes of datasets, as well as multi-class emotion detection. Indeed, the self-attention technique based on transformers can efficiently identify the dependency between the frequency components of facial expressions, and the model generalizes easily even when trained on large facial expression datasets. Moreover, the Fourier transform approach takes into account both low-frequency facial components and high-frequency emotions. Additionally, the proposed framework could also be scaled for multimodal affective computing applications with minor changes to the architecture, as it would support other modalities like speech, physiology, and context. Consequently, the presented FIT framework exhibits strong scalability for future large-scale emotion-aware recommendation and intelligent multimedia applications.

Robustness of the proposed methodology was tested under various real-world settings, such as variation in lighting, poses, backgrounds, and even facial expressions. In contrast to existing spatial domain-based FER approaches, the

proposed FIT makes use of amplitude-phase frequency representation and Fourier Domain Encoding (FDE). Moreover, the self-attention property of transformers enhances the quality of feature learning by making the model less sensitive to partial and alignment issues. Besides, the low standard deviations achieved during the statistical analysis prove that the performance of the model is stable. Consequently, it can be concluded that the proposed framework of FIT offers robust extraction of emotional features, good generalization ability, efficient deployment, and scalability.

4.4. Advantages and Limitations

A FER-based music recommendation system offers several advantages, as it enables the user's emotional state to be automatically and instantly identified through facial analysis. The use of robust face detection techniques, along with well-established datasets such as FER-2013 and RAF-ML, enhances the model's applicability across a wide range of facial expressions and improves its effectiveness in real-world environments.

Data preprocessing tasks such as resizing, enhancing, and normalizing images contribute to the stable feature representation and the training being faster. The FIT method not only improves the discriminative ability but also captures the global frequency domain facial patterns, leading to better and more accurate emotion classification. Music based on the user's mood can be recommended, thus providing a personalized and adaptive user experience without the user ever having to express it.

The system has limitations, as it may be the case that facial expressions do not represent complex or mixed expression states completely. Furthermore, certain factors like lighting, pose, occlusion, and camera quality variations can turn out to be drawbacks for performing detection and recognition of expressions. Labelling noise and dataset imbalance of the datasets used in the real world for affective recognition may introduce bias in the classification.

The high computational requirement of transformer-based models could be a limitation for their use in devices with limited resources. The acquisition of facial images and their processing also rise privacy concerns, which make the situation more challenging from the utility point of view. The framework demonstrates strong potential for intelligent FER-

based music recommendation; further optimization and multimodal integration are required to enhance robustness and real-world applicability.

5. Conclusion

The research presented an advanced music recommendation system based on FER using the FIT. The introduced model has performed well in combining frequency-domain feature extraction with transformer-based global representation learning, thereby enabling accurate detection of human FER from facial images. The tests on the RAF-ML dataset have shown the proposed method to be robust and dependable. The performance of the FIT model was impressive, achieving an accuracy of 98.97%, a precision of 98.92%, a recall of 98.87%, and an F1-score of 98.90% while surpassing the other methods. The superior results can be attributed to the FIT's ability to capture discriminative facial expression patterns that are distributed across different frequency components, which are mostly not recognized by purely spatial feature extractors. The preprocessing pipeline adopted, which included face detection, image enhancement, resizing, and normalization, further contributed to the performance stability under real-world variations such as changes in lighting and different facial expressions. The system, by connecting the recognized expression states with the suitable music tracks, provides the user with an engaging, customizable, and personalized recommendation experience, which removes the need for direct user interaction.

The proposed system, although it has a great performance, heavily depends on visual signals and might not be able to recognize complicated or mixed expression states. The future work would include the use of different types of signals, like voice, body signals, and the user's past listening, to improve the understanding of emotions and make the recommendations more relevant. Moreover, it is to simplify the computations in order to make the real-time implementation possible on low-power and mobile devices. The suggested FIT-based model provides a good and flexible solution for FER-based music recommendation systems.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Funding Statement

Not Applicable.

References

- [1] Thomas Kopalidis et al., "Advances in Facial Expressions Recognitions: A Survey of Methods, Benchmark, Models, and Dataset," *Information*, vol. 15, no. 3, pp. 1-61, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [2] Ruchi Jayaswal et al., "Advances in Facial Expressions Recognitions Technologies for Emotions Analysis," *Discover Computing*, vol. 28, no. 1, pp. 1-56, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [3] Pit Pichappan, "A Review of the Emotions-Induced Music Recommendations System," *Journal of Digital Information Management*, vol. 23, no. 2, pp. 112-133, 2025. [[CrossRef](#)] [[Publisher Link](#)]

- [4] Sheetal Patil et al., "Review on Music Emotions Analysis using Machine Learnings: Technologies, Method, Dataset, and Challenge," *Discover Applied Sciences*, vol. 7, no. 7, pp. 1-14, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [5] Vilas Gaikwad et al., "Emotions-Driven Music Recommendations Systems," *SAMRIDDHI: A Journal of Physical Science, Engineering and Technology*, vol. 16, no. 4, pp. 136-140, 2024. [[CrossRef](#)] [[Publisher Link](#)]
- [6] Rashini Liyanarachchi, Aditya Joshi, and Erik Meijering, "A Survey on Multimodal Music Emotions Recognitions," *arXiv*, pp. 1-26, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [7] Sepideh Kalateh et al., "A Systematic Review on Multimodal Emotions Recognitions: Building Block, Current states, Application, and Challenge," *IEEE Access*, vol. 12, pp. 103976-4019, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [8] Rosa A. García-Hernández et al., "A Systematic Literature Review of Modalities, Trend, and Limitation in Emotions Recognitions, Affective Computing, and Sentiments Analysis," *Applied Sciences*, vol. 14, no. 16, pp. 1-25, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [9] You Wu, Qingwei Mi, and Tianhan Gao, "A Comprehensive Review of Multimodal Emotions Recognitions: Technique, Challenges, and Future Direction," *Biomimetics*, vol. 10, no. 7, pp. 1-24, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [10] Manju Priya Arthanarisamy Ramaswamy, and Suja Palaniswamy, "Multimodal Emotions Recognitions: A Comprehensive Review, Trend, and Challenges," *Wiley Interdisciplinary Review: Data Mining and Knowledge Discovery*, vol. 14, no. 6, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [11] Guimin Hu et al., "Recent Trend of Multimodal Affective Computing: A Survey from NLP Perspectives," *arXiv*, pp. 1-22, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [12] Erkang Jing et al., "Emotions-Aware Personalized Music Recommendations with a Heterogeneity-Aware Deep Bayesians Network," *ACM Transaction on Information System*, vol. 43, no. 5, pp. 1-43, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [13] Madipally Sai Krishna Sashank et al., "Mood-based Music Recommendations Systems using Facial Expressions Recognitions and Texts Sentiments Analysis," *Journal of Theoretical and Applied Information Technology*, vol. 100, no. 19, 2022. [[Publisher Link](#)]
- [14] Rajesh B et al., "Music Recommendations based on Facial Emotions Recognitions," *arXiv*, pp. 1-8, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [15] Hung Nguyen et al., "A Model for Songs Recommendations based on Facial Emotions Analysis and Musical Emotions," *International Journal of Intelligent Engineering and System*, vol. 17, no. 4, pp. 1028-1041, 2024. [[CrossRef](#)] [[Google Scholar](#)]
- [16] Fuyan Ma, Bin Sun, and Shutao Li, "Facial Expressions Recognitions with Visual Transformer and Attentional Selective Fusions," *IEEE Transaction on Affective Computing*, vol. 14, no. 2, pp. 1236-1248, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [17] Mei Bie et al., "Swin-FER: Swin Transformers for Facial Expressions Recognitions," *Applied Sciences*, vol. 14, no. 14, pp. 1-14, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [18] Janhavi Kaimal et al., "Real Time Emotions Based Music Players," *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, vol. 12, no. 4, pp. 5601-5607, 2024. [[CrossRef](#)] [[Publisher Link](#)]
- [19] V.S.G.S. Phaneendra Bottu, and K. Ragavan, "Emotions-based Music Recommendations Systems Integrating Facial Expressions Recognitions and Lyrics Sentiment Analysis," *IEEE Access*, vol. 13, pp. 87740-87752, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [20] Nilesh Bhaskarrao Bahadure et al., "Machine Learning-Based Music Classifications and Recommendations Systems from Spotify," *International Journal of Computer Information Systems and Industrial Management Applications*, vol. 17, pp. 143-154, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [21] Nandini Gupta et al., "Intelligent Music Recommendations Systems based on Face Emotions Recognitions," *2023 International Conferences on Computing, Communications, and Intelligent System (ICCCIS)*, Greater Noida, India, pp. 110-115, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [22] Shalaka Prasad Deore, "SongRec: A Facial Expression Recognitions System for Songs Recommendations using CNN," *International Journal of Performability Engineering*, vol. 19, no. 2, pp. 115-121, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [23] Mst. Lima Akter Asha et al., "Suggesting Playlists and Playing Preferred Music based on Emotions from Facial Expressions," *2024 3rd International Conferences for Innovations in Technology (INOCON)*, Bangalore, India, pp. 1-5, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [24] Tanisha Kapoor, Arnaja Ganguly, and D Rajeswari, "Music Recommendations based on Facial Expression using Data Augmentations," *2023 3rd International Conferences on Innovative Mechanism for Industry Application (ICIMIA)*, Bengaluru, India, pp. 1172-1178, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [25] Dhruv Chopra et al., "Feel-Tunes: An Emotion-Driven Music Recommender System," *Proceedings of the International Conference on Innovative Computing & Communication (ICICC 2024)*, pp. 1-6, 2023. [[Google Scholar](#)] [[Publisher Link](#)]
- [26] Ashish Tripathi et al., "Facial Emotions-based Songs Recommender Systems using CNN," *International Journal of Engineering Trends and Technology*, vol. 72, no. 6, pp. 315-327, 2024. [[CrossRef](#)] [[Publisher Link](#)]

- [27] Narayan Paudel et al., “Facial Emotions Recognition System using CNN for Songs Mapping,” *International Journal on Engineering Technology*, vol. 1, no. 2, pp. 312-323, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [28] Porawat Visutsak et al., “Mood-based Music Discovery: A System for Generating Personalized Thai Music Playlist using Emotions Analysis,” *Applied System Innovation*, vol. 8, no. 2, pp. 1-24, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [29] Abdoul Malik, Mohammad Adnan Ayoubi, and Mohammad Alsuedani, “Deep Learning for Facial Emotions Recognitions on the FER-2013 Dataset using the ResNet50v2 Model,” *OMU Journal of Engineering Sciences and Technology*, vol. 5, no. 2, pp. 19-33, 2025. [[Google Scholar](#)] [[Publisher Link](#)]
- [30] Baiyi Zhang et al., “Diversecer-Net: A Compound Emotions Recognitions Framework based on Residual Anti-Aliasing Attentions Networks and Diverse Data Augmentations,” *Biomedical Signal Processing and Control*, vol. 118, pp. 1-16, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [31] Navid Falah, Behnam Yousefimehr, and Mehdi Ghatee, “Predicting Music Tracks Popularity by Convolutional Neural Network on Spotify Feature and Spectrograms of Audio Waveforms,” *arXiv*, pp. 1-12, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [32] Florian Schwarzhans et al., “Image Normalizations Technique and their Effects on the Robustness and Predictive Powers of Breast MRI Radiomics,” *European Journal of Radiology*, vol. 187, pp. 1-12, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [33] Rakhmonalieva Farangis Oybek Kizi, Tagne Poupi Theodore Armand, and Hee-Cheol Kim, “A Review of Deep Learning Technique for Leukemia Cancer Classifications based on Blood Smear Image,” *Applied Biosciences*, vol. 4, no. 1, pp. 1-32, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [34] Muhammad Shahrul Zaim Ahmad et al., “Impact of Images Enhancements using Contrast-Limited Adaptive Histogram Equalizations (CLAHE), Anisotropic Diffusions, and Histogram Equalizations on Spine X-Ray Segmentations with U-Net, Mask R-CNN, and Transfer Learnings,” *Algorithms*, vol. 18, no. 12, pp. 1-22, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [35] Tim-Oliver Buchholz, and Florian Jug, “Fourier Image Transformer,” *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Orleans, LA, USA, pp. 1846-1854, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [36] S. Kumar Reddy Mallidi, and Rajeswara Rao Ramisetty, “Advancements in Training and Deployments Strategies for AI-based Intrusions Detections System in IoT: A Systematic Literature Review,” *Discover Internet of Things*, vol. 5, no. 8, pp. 1-33, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [37] Xueer Sun et al., “Emotion-Aware Cross-Modal Music Generations based on Multimodal Emotions Recognitions,” *Alexandria Engineering Journal*, vol. 133, pp. 254-270, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [38] Yaoyu Sun, “Feature Centric based Deep Learning Approach for Music Moods Recognitions with HuBERT Transformer Models,” *Scientific Reports*, vol. 15, no. 1, pp. 1-34, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [39] Yassine El Boudouri, and Amine Bohi, “Emonext: An Adapted Convnext for Facial Emotions Recognitions,” *2023 IEEE 25th International Workshops on Multimedia Signals Processing (MMSP)*, Poitiers, France, pp. 1-6, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [40] Bettina Shirley Richard et al., “Vivify—Emotion Recognitions and Music Recommendations using Transformer,” *IEEE Access*, vol. 13, pp. 204894-204907, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]