Journal of Computer Science

Original Research Paper

A Study on Emotion Analysis and Music Recommendation Using Transfer Learning

[1]Krishna Kumar Singh and [2]Payal Dembla

[1]*Department of Computer Science, Symbiosis Centre for Information Technology, Pune, India*
[2]*Department of Data Science and Data Analytics, Symbiosis Centre for Information Technology, India*

Article history
Received: 03-03-2023
Revised: 21-04-2023
Accepted: 02-05-2023

Corresponding Author:
Krishna Kumar Singh
Department of Computer Science, Symbiosis Centre for Information Technology, Pune, India
Email: krishnakumar@scit.edu

Abstract: As more and more people access and consume music through streaming platforms and digital services, music recommendation has grown in importance within the music industry. Given the abundance of music at our disposal, music recommendation algorithms are essential for guiding users toward new music and for creating individualized listening experiences. People frequently seek out music that fits their current emotional state or desired emotional state, which means that emotions can have a big impact on music recommendations. Emotions can be taken into account by music recommendation algorithms when deciding which songs or playlists to recommend to listeners. Face expressions are frequently used to gauge a person's mood. By using a webcam or any other external device, recognizable facial traits can now be extracted as inputs thanks to modern technology. Transfer learning is a method that is increasingly in demand for enhancing emotion recognition and music recommendation systems in the modern world. Transfer learning has evolved into a potent method for utilizing prior knowledge to enhance model performance and lessen the requirement for massive volumes of labeled data as a result of the data explosion and the availability of big pre-trained models. Hence, the objective of this study is to understand how transfer learning impacts the accuracy of detecting emotions from facial expressions and how the music recommendations can be personalized based on the detected emotions. This study aims at recommending songs by detecting the facial expressions of users using the FER2013 dataset for emotion recognition which is further extended by adding own images to the categories in the dataset from Google. A basic CNN, finetuned pre-trained ResNet50V2, finetuned pre-trained VGG16, and finetuned pre-trained EfficientNet50 B0 are trained on the dataset for emotion detection and compared. The music recommendation system is developed using the Spotify songs dataset extracted using Spotify web API. It uses k-means clustering for grouping tracks based on emotions and getting song recommendations based on the emotion predictions using finetuned ResNet50-V2 model with the highest training accuracy of 77.16% and validation accuracy of 69.04%. The findings reveal that using a transfer learning approach may effectively identify emotions from facial expressions and can have a potential impact on recommending music. It improves duties related to music recommendations and might be a useful method for assisting users in finding new music that fits the intended emotional state.

Keywords: Transfer Learning, Emotion Prediction, Music Recommendation System

Introduction

Regardless of age, geography, culture, or level of musical ability, music has always had a significant and extensive impact on human life. It is safe to state that one of the most well-liked pastimes is listening to music. The music industry has continued to show a lot of interest as a result. It is not surprising that over the past 15 years, digital and streaming music has gradually outpaced physical recordings in terms of revenue for the recorded music industry given how simple and affordable it is to listen to music (IFPI, 2018) music streaming services that are available online like "Spotify", "sound cloud" and "YouTube music" are drawing increased attention from the general public as a result of the expansion of the digital and streaming worldwide revenue of the music industry. Music service providers have derived to the conclusion that has to offer the customers not music tracks with a broad range of varieties, but also additional sophisticated and individualized new streaming services daily to survive in this competitive market, in which even the top companies in the music sector appear to be fighting for survival. As a result, the music industry unquestionably benefits from the application of data science disciplines like "machine learning", "data mining" and "recommender systems". The personalization of music services is a significant issue here (Chung *et al.*, 2009). Music has a profound link with people's mental worlds, serving as the quickest route to the soul. On the one hand, listening to music can help with basic mental and emotional needs. People frequently turn to music to unwind, reflect, or indulge in particular situations. On the other hand, music can affect the feelings and actions of listeners. Light music, for example, has the power to calm people's minds and serve as a form of mental therapy, but arousal music, on the other hand, might inspire people to take vengeful action. The influence that music has on people's thoughts and emotions makes it a two-edged sword that needs to be taken seriously (Liu *et al.*, 2023). Users should be able to select between personalized music playlists based on their emotional state, which will surely result in major improvements in the customization, enhancement, and knowledge of music streaming services as well as higher customer happiness. Numerous research from the last few years confirms that music affects how people feel and behave, as well as how their brains function. In one study of the reasons why individuals hear music, researchers found that the connection between arousal and mood was one of the most important functions of music. The potential of music to improve participants' moods and increase their level of self-awareness are two of its most crucial uses. It has been shown that a person's musical tastes are closely tied to their personality and mood (Shirwadkar *et al.*, 2022). Hence, emotions play a vital role in a person's selection of music.

People worry that machines will rule their lives and that humans' place in various fields will shrink as a result of the pervasiveness of the terms artificial intelligence and robotics. The truth, however, is very different from this picture. The ability to apply one's talents, skills, and abilities sets humans apart from other organisms. Additionally, people differ from one another in terms of quality and originality and those that stand out have intelligence. Human intelligence is defined as the capacity and talent for problem-solving. Artificial intelligence was developed primarily to simulate the human mind through the use of computer programs that can comprehend human behavior to investigate the behavior of human intelligence (Aggarwal *et al.*, 2022).

Due to the increasing popularity of artificial intelligence, deep learning, and machine learning, transfer learning has gained a lot of attention and popularity in recent years and has become a potent method for using previously learned models to do new tasks. Large volumes of data are necessary for deep learning algorithms to be taught efficiently, which can be difficult in many fields, including music. This problem is resolved by transfer learning, which enables models to use information gained from similar tasks or domains. In the world of music, transfer learning has been used in a variety of contexts, most notably in tasks involving music genre classification and music recommendation which also involves emotion prediction-based music recommendation tasks. Because emotions are subjective and can be impacted by a wide range of factors, including cultural background and personal experience, predicting emotions is a difficult process. A range of techniques, including Electroencephalography (EEG), body language, and vocal intonation, can be used to identify emotions (Revathy *et al.*, 2023). But observing facial expressions is a much easier and more useful approach. As it may be used to leverage knowledge obtained from related domains and minimizes the requirement for enormous amounts of labeled data, transfer learning is currently developing potential in the prediction of emotions from facial expressions. Transfer learning can be used to transfer knowledge from pre-trained models on audio processing or sentiment analysis tasks to increase the performance of emotion prediction models in the context of emotion prediction-based music recommendation systems.

The purpose of this study is to assess the efficiency and accuracy of personalized music recommendations based on projected emotions by using transfer learning techniques. In this research, we explore how pre-trained models, which are fine-tuned and re-trained on a variety of facial expression images, enhance the capability to recognize emotions in new images and generate more tailored and appropriate music recommendations based on the recognized emotions. The scope of this research is limited to the use of publically available facial expression images, however, it can be broadened in the future by incorporating real-time detection of emotions. This study aims at recommending songs by detecting the facial expressions of users using the FER2013 dataset for emotion recognition which is further extended by adding own images to the categories in the dataset from Google. A basic CNN, finetuned pre-trained ResNet50V2, finetuned pre-trained VGG16, and finetuned pre-trained

EfficientNet50 B0 are trained on the dataset for emotion detection and compared. The music recommendation system is developed using the Spotify songs dataset extracted using Spotify web API. It uses k-means clustering for grouping tracks based on emotions and getting song recommendations based on the emotion predictions using finetuned ResNet50-V2 model.

The main contributions of this study are: Expanding the FER2013 dataset by adding more images from Google to determine the impact on the accuracy of emotion prediction; Preprocessing the images by using image augmentation techniques; Fine-tuning pre-trained transfer learning models for emotion prediction and comparing existing work accuracies with the proposed model accuracies; and building a music recommendation system based on predictions from the model with the highest accuracy for emotion detection using Spotify music dataset. The paper is further divided into 6 sections including the background of the study in the literature review section; the detailed methodology in the proposed system design section; analysis of the datasets and finetuning of the models in the data analysis section; evaluating results and comparison with existing work in the results and discussion section; determining managerial and theoretical implications of the study in the implications: Theoretical and managerial section; and conclusions with future directions for the research in the conclusion and future direction section.

Singh (2023) Machine learning algorithms have significantly advanced the field of music recommendation in recent years. Emotion-based music recommendation systems are one subject that has drawn a lot of interest. Based on the feelings that songs elicit in the listener, these systems try to suggest music. Transfer learning is a well-liked machine learning method for modifying previously trained models for new tasks. The application of transfer learning in emotion-based music recommendation systems is investigated in this review of the literature.

Florence and Uma (2020) suggested a system that could recognize the user's facial expressions and extract facial landmarks from those expressions. These landmarks were then categorized to ascertain the user's specific emotion. Once the emotion was determined, the user would be shown songs that matched that emotion. It might let a user choose the music they should listen to, which would help them feel less stressed. The user would save time by not having to search for or look up tunes. The architecture that was proposed consisted of three sections: "Emotion extraction", "audio extraction" and "emotion audio extraction". Even so, it had several shortcomings, such as the difficulty of the suggested system to monitor and capture all the emotions with precision due to the paucity of images in the employed dataset of images. The image that the classifier gets must be taken in a well-lit setting for it to produce accurate findings. To effectively predict the user's sentiment, the image quality must be at least better than 320 p. In natural settings, the generalizability of handcrafted qualities is frequently insufficient.

James *et al.* (2019) proposed an "emotion-based music recommendation system" that attempted to scan and comprehend facial emotions, to make an acceptable playlist. The arduous task of classifying or organizing music into multiple lists manually was simplified by constructing a playlist that is appropriate based on a person's emotional traits. The proposed method focused on emotion recognition to build music players based on human feelings. A linear classifier was used for face detection. Based on the values of intensity of all point's pixels, "regression trees" trained using a "gradient boosting technique" were used to create a map of facial landmarks of a given face image. An SVM classifier with several classes was used to classify emotions. The four types of emotions were shocked, sad, furious, and pleased. The suggested technique was still unable to completely capture all the emotions because of the small number of photographs in the image dataset being used. There wasn't a range of feelings. In natural settings, the generalizability of handcrafted qualities was frequently insufficient.

Bhattarai and Lee (2019); IFPI (2018) suggested a deep learning method utilizing transfer learning and a Multilayer Perceptron (MLP) classifier for autonomous music mood recognition. To extract low-level audio aspects including timbre, pitch, and rhythm, this study leverages the million-song dataset, which contains audio attributes for more than one million songs. To separate the high-level characteristics from the low-level audio information, a pre-trained Convolutional Neural Network (CNN) called VGG16 is utilized as a feature extractor. VGG16 was initially developed for picture classification. To forecast the mood of the music, the collected features are then fed into a Multilayer Perceptron (MLP) classifier. The million-song dataset that the study employs may not accurately represent the range of musical genres and styles, which may limit the generalizability of the suggested approach to other datasets, which is one of the study's few shortcomings.

An innovative method for music recommendation systems that considers the listener's emotional state was introduced by Joshi *et al.* (2021) The suggested method makes music track recommendations based on the listener's emotional content by using deep learning techniques, specifically LSTM and CNN. The CNN model is used to extract pertinent characteristics from the music data while the LSTM model is utilized to model the temporal dependencies of the music. Additionally, transfer learning was used to boost the system's precision. The feature extractor for the authors' music recommendation system was a pre-trained VGG16 model. According to the findings, the suggested method performs better in terms of accuracy and personalization than several other existing music recommendation systems.

Sekaran *et al.* (2021) proposed a technique for AlexNet convolutional neural network transfer learning facial emotion recognition. On the FER2013 dataset, the authors improved the last few layers of the AlexNet network, which resulted in an accuracy of 62.54%. The authors also conducted tests to see what impact the quantity of the training dataset had on how well the facial emotion identification system worked. They discovered that the system's accuracy increases as the size of the training dataset does, which suggests that accumulating larger and more varied datasets may help face emotion recognition technology advance even further.

In the context of human-computer interaction, (Chowdary *et al.*, 2021) discussed the use of deep learning algorithms for facial expression recognition. On the CK + dataset, the authors applied transfer learning techniques employing pre-trained networks including ResNet50, VGG19, Inception V3, and MobileNet. The fully linked layers of the pre-trained ConvNets were removed by the authors and they added new fully connected layers that were more suited to the task. Only updating the weights may be trained on the newly added layers. For issues with emotion detection, the system had an average accuracy of 96%. A method using CNN for emotion recognition with Pygame and Tkinter for music recommendation was proposed by Athavle *et al.* (2021). On the FER2013 dataset, the authors trained the model. The writers used an in-built camera to record facial expressions and by recognizing the user's current feelings, an autonomously generated music playlist was created.

Hung *et al.* (2019) proposed the framework of dense FaceLiveNet and offered a method to enhance the FaceLiveNet network with low and high accuracy in fundamental emotion recognition. This system was initially applied to the FER2013 basic emotion dataset using the very straightforward JAFFE and KDEF basic emotion detection model and it showed an accuracy of 70.02%. Second, the test accuracy rate reached 91.93% using the FER2013 basic emotion recognition model transferring to the learning emotion recognition model, which is 12.9% higher than the accuracy rate of 79.03% without using the transfer learning model. This result demonstrated that the use of transfer learning can effectively improve the recognition accuracy of the learning emotion recognition model. In addition, videos taken during class instruction by students from a national university in Taiwan were used as test data to assess the learning emotion recognition model's capacity for generalization. After being rebuilt, the model's recognition accuracy rate was 92.42%. Additionally, the model was rebuilt and achieved an accuracy rate of 84.59% after integrating the initial learning emotion database with all of the student image data. The outcome demonstrates that the learning emotion recognition model can process the unlearned image through transfer learning to attain excellent recognition accuracy (Singh, 2023; Mahapatra and Singh, 2022).

Different deep-learning algorithms for emotion recognition were described by (Yen and Li, 2022). ResNet50, Xception, EfficientNet B0, Inception, and DenseNet121 are the five models that were used in the study. The FER2013 and AffecNet datasets were used to train the models. According to the study's findings, using AffectNet boosted the accuracy of the ResNet-50, Xception, EfficientNet-B0, Inception, and DenseNet-121 models by 8.37, 10.45, 10.45, and 5.47%, respectively. These models' accuracy improved by 5.72, 2, 10.45, 5, and 9%, respectively, on FER2013. These findings demonstrated the reliability and benefits of the experiments conducted for this investigation. According to a specially created sentiment analysis framework, a method for mood improvement was proposed by Negre *et al.* (2022) The system trained a deep learning model for mood identification using the Fer2013 dataset and then matched the mood with a particular playlist. The songs were taken from the Spotify API and they are simply upgradeable to give the recommender system more relevant results. To enable user and system interaction and provide precise recommendations for efficient resource usage, a variety of technologies and computer languages were used during system development. The authors' goal was to create a system that, by playing appropriate music based on the user's mood, would lift it. The system validation showed a trustworthy and user-friendly system that successfully achieved a good accuracy of 60% for the trained model.

To recognize facial expressions, (Reddi and Krishna, 2023) proposed a method that combines transfer learning with Convolutional Neural Networks (CNNs) that have a limited number of parameters. The goal was to create a system that can detect the emotional state that, on average, most members of a group are experiencing in real time. On the FER2013, JAFFE, and CK + datasets, the suggested architecture was trained for real-time detection. The model could detect emotions such as joy, sadness, surprise, fear, anger, contempt, and neutrality. Nawaf and Jasim (2023) compared two models: The VGG16 model, which was reset and trained using the FER dataset, and a model created from scratch and trained exclusively on the FER dataset. To determine which model would be most effective at identifying human emotions, the models were tested using photographs from the internet. The outcomes demonstrate that the CNN model, which was created entirely from scratch, outperformed the VGG16 model with an accuracy of 87.133%. This study used CNN's capacity to extract features from images and classify them to improve the precision of identifying human emotions through facial expressions. Using a multi-class neural network, (Modran *et al.*, 2023) suggested a machine-learning model to categorize emotions into four groups and then forecast the results. The model seeks to forecast if a particular song will be helpful for a particular person. It takes into account a person's musical and emotional traits, but it is also taught to take into account solfeggio frequencies. A portion of the million datasets is used in this experiment. To enable the algorithm to forecast, the user chooses their preferred genre of music and their present state of mind. If the chosen song is incorrect, the app suggests an alternative kind of music that might be appropriate for that particular user.

The Music 4 All dataset is used in a study by Revathy *et al.* (2023) to assess the lyrical elements necessary for the detection of four significant human emotions: Joyful, furious, relaxed, and sad. Using emotionally important features and pre-trained word embeddings, the authors presented a unique method for labeling lyrics data. The BERT model for emotion prediction was used to retrain the dataset. The authors integrate transferred knowledge into their hybrid methodology, which includes BERT embeddings. It was found that the BERT model increases the model's 92% overall accuracy. Additionally, the author suggested a recommender system based on the Sentence Transformer Model. Meena *et al.* (2023) offered a study that uses the CK+, FER -2013, and Jaffe datasets along with the well-known deep convolutional neural network Inception V3 for emotion detection and categorization. The model was gradually fine-tuned by the authors using a pipeline training strategy to achieve high identification accuracy of 99.5%. The Inception V3 model had the highest accuracy, according to the authors, who compared it to other machine-learning methods. To perform a multiclass categorization of Bangla music genres including "Bangla Adhunik," "Bangla hip-hop," "Bangla band music," "Nazrulgeeti," "palligeet," and "Rabindra sangeet," (Hasib *et al.*, 2022) created a special technique called BMNet-5. This study employed 1742 pieces of Bangla music as its dataset. Based on a neural network created to identify music genres from auditory inputs, the proposed BMNet-5 was developed. The proposed model was tested for performance consistency using k-fold cross-validation with varied k-values and achieved an accuracy of 90.32%. Additionally, this model was utilized to train the SHAP model for all the genres in the Bangla music dataset. The existing work had several limitations or disadvantages that need to be addressed. The previous studies have only considered (IFPI, 2018).

According to the literature review, the existing work had several limitations and disadvantages. Most of the studies lack the generalizability of the dataset as they rely solely on publicly available facial expression images for emotion detection models. Also, few studies balanced the emotion dataset, but in the real-world scenario, the data might not be balanced. Hence, this study aims at overcoming the limitations by extending the already available FER2013 dataset by adding random images from Google for a more generalized dataset. This study also aims to provide a solution considering the real-world scenario by training models on the imbalanced dataset. The literature review also suggests that transfer learning is a potential strategy for enhancing the performance of music emotion detection and recommendation systems. To successfully extract features from audio signals and increase the classification accuracy of music emotion detection systems, transfer learning permits the use of pre-trained models on huge datasets.

Proposed Design and Methods

Proposed System Design

The objective of this research is to understand the effectiveness of transfer learning models on the prediction of emotions from facial expression images and to assess the efficiency and accuracy of personalized and appropriate music recommendations according to the predicted emotions. For this, the proposed architecture is broadly segregated into 4 sections. The first section involves using the publicly available FER2013 dataset (Verma, 2018) and extending the dataset by adding own images into each of the 7 categories from Google. The second section involves pre-processing the images by using image augmentation techniques so that the images are processed as per the model requirement. The third section involves fine-tuning the pre-trained transfer learning models and training on the processed images and predicting emotions. The fourth section involves using the model with the highest accuracy for emotion prediction and building a music recommendation system that would suggest appropriate music according to the predicted emotions. The proposed system architecture is shown in Fig. 1. The idea is to use "deep neural networks" to select the optimum feature abstraction.

Proposed System Architecture

Dataset loading (FER2013 + Google images): FER2013 dataset is used for detecting emotions. The dataset has been extended by downloading images from Google for the seven categories present in the dataset.

Data augmentation/preprocessing: The images are rescaled, flipped, rotated, and sharpened according to the model required to obtain better predictions.

Emotion detection: Four models are trained, that is: Basic CNN, Finetuned ResNet-50 V2, finetuned VGG-16, and Finetuned EfficientNet B0. The models are finetuned and trained after the augmentation of the dataset.

Song recommendation: The Spotify data, which includes the track names, ids, song features, artist names, etc., is used for song suggestions. K-means clustering is used to cluster song tracks and recommend songs based on the seven emotions in the FER dataset i.e., "angry", "happy", "sad", "fear", "neutral", "disgust" and "surprised". The best emotion model with the highest accuracy is used for making predictions. Recommendations are sorted based on the popularity of songs.

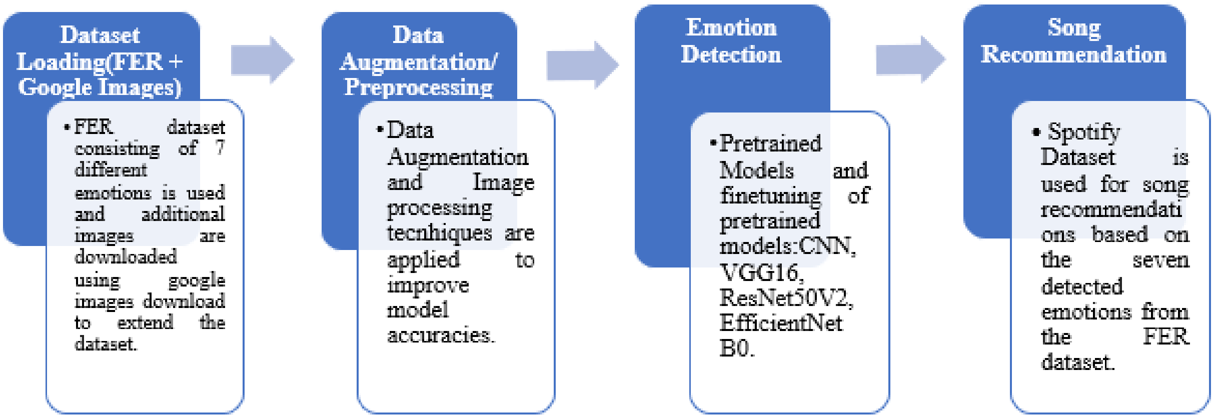

Dataset Loading(FER + Google Images)
•FER dataset consisting of 7 different emotions is used and additional images are downloaded using google images download to extend the dataset.
Data Augmentation/ Preprocessing
•Data Augmentation and Image processing tecnhiques are applied to improve model accuracies.
Emotion Detection
•Pretrained Models and finetuning of pretrained models:CNN, VGG16, ResNet50V2, EfficientNet B0.
Song Recommendation
• Spotify Dataset is used for song recommendations based on the seven detected emotions from the FER dataset.

Fig. 1: Proposed system flow

Table 1: Audio features provided by Spotify

Audio features	Description
duration_ms	The time in milliseconds for the track
Key	The predicted key of the completed song. In the absence of a key, the value is -1
Mode	The scale type from which a track's melodic content is formed is indicated by the track's modality (major or minor). 1 represents the major and 0 the minor
time_signature	The number of beats in each bar is indicated by the time signature, which is a measure
Acousticness	An indication of the track's acoustic nature, with a confidence level ranging from 0.0 to 1.0. The track's acoustic nature is highly likely, as indicated by the value of 1.0
Danceability	It is dependent on a number of musical factors, including “pace”, “rhythm stability”, “beat strength” and “overall regularity” describes a track's potential for dancing. The least danceable value is 0.0 and the most danceable value is 1.0
Energy	has a perceptual scale of 0.0 to 1.0 and measures intensity and activity
Instrument alness	The presence or absence of vocals in music. The instrumentality value should be as near to 1.0 as possible.
Liveness	A track is more likely to have been life if the liveness score is higher. The track is probably live if the value is larger than 0.8

Data Collection

The first section of the proposed architecture involves extending the FER2013 dataset by downloading images from Google for each of the 7 categories in the dataset. There are 2 datasets used for the proposed system, Facial Expression Recognition (FER) 2013 dataset for predicting emotions and the Spotify songs dataset for recommending appropriate music based on the detected emotions.

FER2013 dataset: This dataset, which comprises 48x48 pixel grayscale pictures of faces, was collected from Kaggle. Each face has been automatically registered such that it occupies the same space and is located in the same general area. There are seven categories in the dataset: Angry, disgusted, afraid, happy, sad, surprised, and neutral. This dataset has been extended by downloading Google images for all seven categories. Initially, there were 28,709 training images and 3,589 validation images. After adding images from Google, the resulting dataset consisted of 29,690 training images.

Spotify songs dataset: The Spotify dataset was extracted using Spotify web API. It consists of more than 160,000 songs from the year 1921. Each year consists of 100 songs with audio features. Table 1 displays the features for each music track which are part of the audio features offered by the Spotify API. All of Spotify's audio features are included in the table.

The dataset exploration, pre-processing, fine-tuning and model training, results, and comparison with existing work are explained in further sections.

Data Analysis

Facial Expression Recognition (FER2013) Dataset Exploration

The FER2013 dataset (Fig. 2) comprises photos that have been labeled with seven different emotions: "Angry," "disgust," "fear," "happy," "neutral," "sad," and "surprised." To extend this dataset, images were downloaded per category using google images download. The resultant dataset consisted of 29,690 training images and 2,589 validation images. Figure 3 shows the emotion distribution in FER2013 dataset, “happy” emotion is the highest in the dataset, and "disgusted" emotion is the lowest.

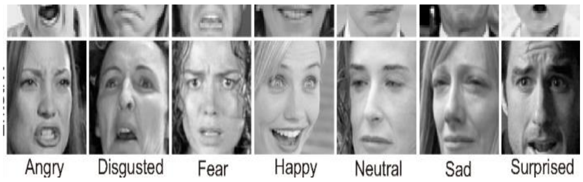


Fig. 2: Extended FER2013 dataset

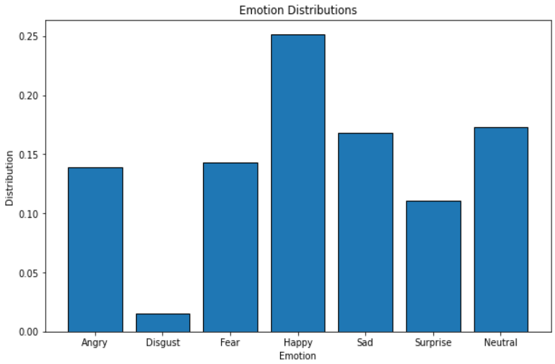


Fig. 3: Emotion distribution in extended FER2013 dataset

Figure 3, it can be visualized that the dataset is imbalanced. There are more images in the "happy" category. As this research represents a real-world problem of detecting facial expressions from images, it is highly possible to get an imbalanced dataset for the actual problem. Additionally, because transfer learning is being employed in this study, the pre-trained model can benefit from data from bigger, more diversified datasets to perform better on the unbalanced dataset. As a result, this dataset is kept unbalanced while taking the real-world situation into account. However, in future research, the transfer learning models could be trained by balancing the dataset and comparing the model accuracies with the accuracies of the model with an imbalanced dataset. Hasib *et al.* (2022) introduced a survey of methods that could be used for managing the classification with the imbalanced data and a combination of both sampling and deep learning methods to address the class imbalance problem. These studies could be used as references for the future direction.

Extended FER2013 Dataset Pre-Processing

Pre-trained models can also benefit from image augmentation to perform better. Pre-trained models are those that have been developed using huge datasets, like ImageNet, and then refined for a particular purpose using a smaller dataset. We can further diversify the training data set and enhance the performance of the pre-trained model by applying image augmentation techniques during the fine-tuning stage. Several image augmentation techniques are used in the study by the image data generator provided by Keras such as rotation to help the model recognize the image from different angles, zooming to help the model recognize zoomed images as well, random shifting horizontally and vertically to help the model to be more robust to different object positions within an image, flipping to flip the image horizontally or vertically for image recognition; scaling the images in the pixel range of [0,1] to help the model to recognize images that seem distant and shear transformation to help the model to simulate the effect of the object in an image viewed at different angles. Table 2 shows the different augmentation techniques used for the proposed models.

Basic CNN: The images used for training a Basic CNN model are rescaled so that the pixel values of each input image are between 0 and 1, zooming range is set to 0.2 so that each input image is randomly zoomed by a factor between 0.8 and 1.2, rotation range is set to 20 so that each input image is randomly rotated by a value between -20 and +20 degrees, width and height shift range is set to 0.1 so that the height and width of each input image will randomly shift up or down by a fraction of up to 10% of the total height and width and horizontal flip is set True so that each input image is randomly flipped horizontally with a probability of 0.5.

Fine-tuned ResNet50 V2: The images used for training a fine-tuned ResNet50 V2 model are rescaled so that the pixel values of each input image are between 0 and 1, zooming range is set to 0.2 so that each input image is randomly zoomed by a factor between 0.8 and 1.2, rotation range is set to 10 so that each input image is randomly rotated by a value between -10 and +10 degrees, width and height shift range is set to 0.1 so that the height and width of each input image will randomly shift up or down by a fraction of up to 10% of the total height and width and horizontal flip is set True so that each input image is randomly flipped horizontally with a probability of 0.5.

Fine-tuned VGG16: The images used for training a fine-tuned VGG16 model are rescaled so that the pixel values of each input image are between 0 and 1, zooming range is set to 0.2 so that each input image is randomly zoomed by a factor between 0.8 and 1.2, rotation range is set to 5 so that each input image is randomly rotated by a value between -5 and +5 degrees, width and height shift range is set to 0.2 so that the height and width of each input image will randomly shift up or down by a fraction of up to 20% of the total height and width and horizontal and vertical flip is set True so that each input image is randomly flipped horizontally and vertically with a probability of 0.5.

Fine-tuned EfficientNet B0: The images used for training a Basic CNN model are rescaled so that the pixel values of each input image are between 0 and 1, rotation range is set to 5 so that each input image is randomly rotated by a value between -5 and + degrees, the shear range is set to 0.2 so that each input image is randomly sheared by a degree chosen uniformly at random from the range (-0.2,0.2) and horizontal and vertical flip is set True so that each input image is randomly flipped horizontally and vertically with a probability of 0.5.

Table 2: Audio features provided by Spotify contd

Audio features	Description
Loudness	Loudness, or the characteristic of a sound, is the first cognitive indicator of tangible strength also known as magnitude. The typical range of values is “-60 to 0 dB”
Speechiness	Speechiness recognizes the presence of words that are spoken in a track. The recording is more entirely “speech-like” the attribute value is closer to 1.0. Tracks with values above 0.66 very certainly only include spoken words. Music and speech may both be included in tracks with values between 0.33 and 0.66, either separately or stacked together, as in rap music. Values below 0.33 are probably pieces of music or other non-speech tracks
Valence	An assessment of a track's musical positivity on a scale of 0.0 to 1.0
Tempo	The predicted beats per minute (bpm) pace of music (BPM)

Emotion Detection: Model Fine Tuning

This section comprises of preparation of models for training. For this study, 4 models were trained namely: A Basic CNN, Fine-tuned VGG16, finetuned ResNet50 V2, and Fine-tuned EfficientNet B0 as explained below (Fig. 4).

Convolutional Neural Network (CNN): A ConvNet is a “deep learning technique" that accepts an input image, assigns different elements and objects in order of the image importance that is biased and learnable weights, and can differentiate between them. In comparison to other classification methods, a "ConvNet" requires significantly less pre-processing. Unlike primitive approaches, where filters are engineered manually, "ConvNets” have the ability to learn these filters and their attributes. The organization of the “visual cortex" and the connection of the network of neurons in the human brain both of them have a significant influence on the design of a "ConvNet". Individual neurons only respond to the stimuli in this restricted region of the visual field which is known as the "Receptive Field". Several overlapping fields like this make up the total visual field. A CNN architecture was developed that consisted of 6 convolutional layers, 12 batch normalization layers, one conv_2D input layer, 3 max-pooling layers, 9 dropout layers with a dropout rate of 0.25, and 7 dense layers. Figure 5 shows the implemented CNN architecture.

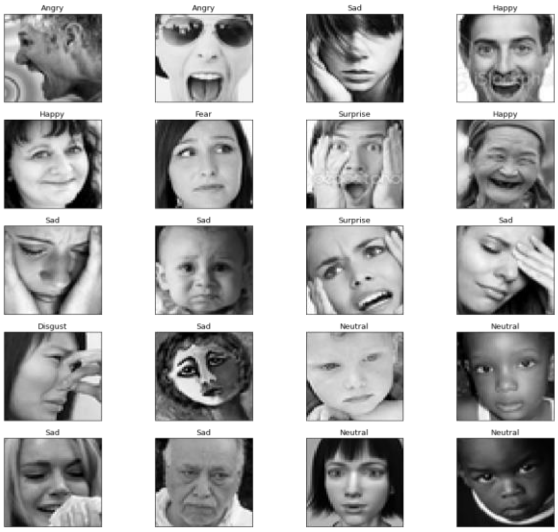


Fig. 4: Random images of different emotions in the dataset

ResNet50V2 Architecture: ResNet, one of the most effective deep neural networks, performed amazingly well in the ILSVRC classification challenge conducted in 2015. In the 2015 ILSVRC and COCO contests for ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation, ResNet won first place. On other identification tests, ResNet also displayed strong generalization performance. There are several versions of the ResNet architecture, each of which uses the same fundamental concept but with a different number of layers. ResNet models come in a variety of forms. The name ResNet followed by a number merely implies the ResNet architecture with a certain number of neural network layers. In this system, ResNet-50 version 2 is implemented, which uses imagenet weights and requires input shape of (224, 224, 3). The pre-trained ResNet 50 is finetuned by adding 7 layers namely dropout with a dropout rate of 0.25, 2 Batch Normalization layers, Flatten layer, 2 Dense layers with activation function as relu and softmax, and a dropout layer with a dropout rate 0.5. Figure 6 displays the layers added to finetune ResNet50V2 pre-trained model. Figure 7 shows the implemented ResNet50V2 architecture.

VGG-16 Model: Convolutional neural networks, a subset of artificial neural networks, are also referred to as ConvNets. A convolutional neural network is made up of an “input layer”, an “output layer” and “numerous hidden layers”. The “Convolutional Neural Network (CNN)” variation known as “VGG16” is known as one of the best computer vision models available today. The designers of this model analyzed the networks and increased the depth using an architecture with highly compact (3 3) convolution filters; the findings showed a significant improvement above the “state-of-the-art” settings. With the depth extended to 16-19 weight layers, approximately 138 trainable parameters were produced. VGG16 is used as an "object identification and classification algorithm" that, when used to segregate 1000 images into 1000 separate categories, obtains an accuracy of 92.7%. It is a renowned method for categorizing photographs and can be easily used with transfer learning. In this system, a pre-trained VGG-16 model, which has imagenet weights and requires input tensor shape as (48, 48, 3), is fine-tuned and trained on an extended FER dataset.

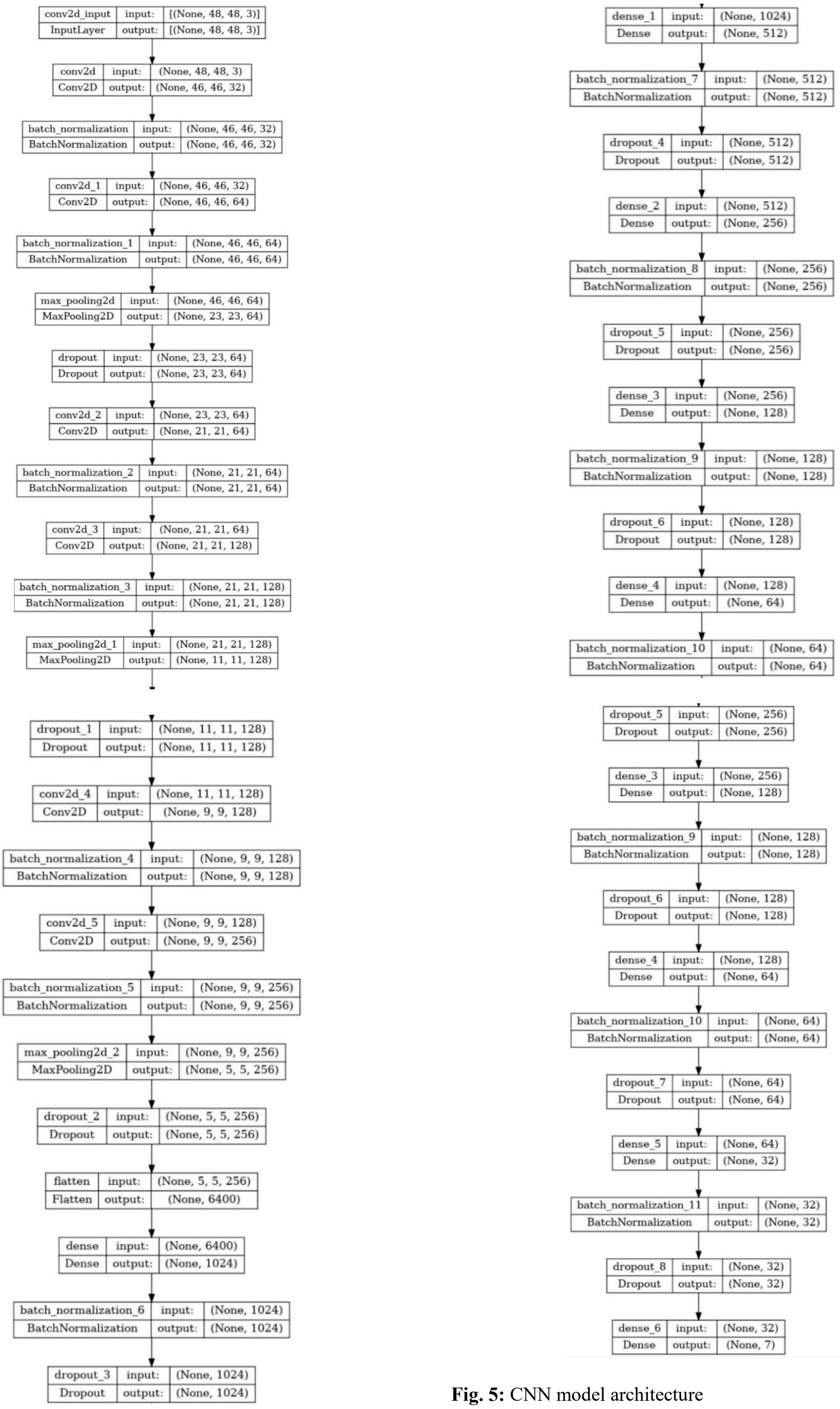
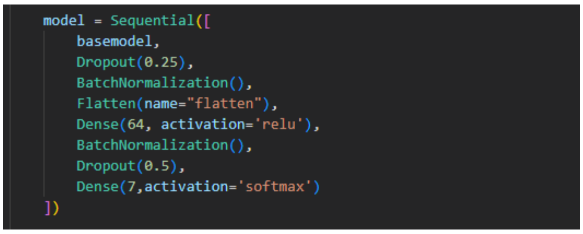

conv2d_input | input: | [(None, 48, 48, 3)]
InputLayer | output: | [(None, 48, 48, 3)]
conv2d | input: | (None, 48, 48, 3)
Conv2D | output: | (None, 46, 46, 32)
batch_normalization | input: | (None, 46, 46, 32)
BatchNormalization | output: | (None, 46, 46, 32)
conv2d_1 | input: | (None, 46, 46, 32)
Conv2D | output: | (None, 46, 46, 64)
batch_normalization_1 | input: | (None, 46, 46, 64)
BatchNormalization | output: | (None, 46, 46, 64)
max_pooling2d | input: | (None, 46, 46, 64)
MaxPooling2D | output: | (None, 23, 23, 64)
dropout | input: | (None, 23, 23, 64)
Dropout | output: | (None, 23, 23, 64)
conv2d_2 | input: | (None, 23, 23, 64)
Conv2D | output: | (None, 21, 21, 64)
batch_normalization_2 | input: | (None, 21, 21, 64)
BatchNormalization | output: | (None, 21, 21, 64)
conv2d_3 | input: | (None, 21, 21, 64)
Conv2D | output: | (None, 21, 21, 128)
batch_normalization_3 | input: | (None, 21, 21, 128)
BatchNormalization | output: | (None, 21, 21, 128)
max_pooling2d_1 | input: | (None, 21, 21, 128)
MaxPooling2D | output: | (None, 11, 11, 128)
dropout_1 | input: | (None, 11, 11, 128)
Dropout | output: | (None, 11, 11, 128)
conv2d_4 | input: | (None, 11, 11, 128)
Conv2D | output: | (None, 9, 9, 128)
batch_normalization_4 | input: | (None, 9, 9, 128)
BatchNormalization | output: | (None, 9, 9, 128)
conv2d_5 | input: | (None, 9, 9, 128)
Conv2D | output: | (None, 9, 9, 256)
batch_normalization_5 | input: | (None, 9, 9, 256)
BatchNormalization | output: | (None, 9, 9, 256)
max_pooling2d_2 | input: | (None, 9, 9, 256)
MaxPooling2D | output: | (None, 5, 5, 256)
dropout_2 | input: | (None, 5, 5, 256)
Dropout | output: | (None, 5, 5, 256)
flatten | input: | (None, 5, 5, 256)
Flatten | output: | (None, 6400)
dense | input: | (None, 6400)
Dense | output: | (None, 1024)
batch_normalization_6 | input: | (None, 1024)
BatchNormalization | output: | (None, 1024)
dropout_3 | input: | (None, 1024)
Dropout | output: | (None, 1024)
dense_1 | input: | (None, 1024)
Dense | output: | (None, 512)
batch_normalization_7 | input: | (None, 512)
BatchNormalization | output: | (None, 512)
dropout_4 | input: | (None, 512)
Dropout | output: | (None, 512)
dense_2 | input: | (None, 512)
Dense | output: | (None, 256)
batch_normalization_8 | input: | (None, 256)
BatchNormalization | output: | (None, 256)
dropout_5 | input: | (None, 256)
Dropout | output: | (None, 256)
dense_3 | input: | (None, 256)
Dense | output: | (None, 128)
batch_normalization_9 | input: | (None, 128)
BatchNormalization | output: | (None, 128)
dropout_6 | input: | (None, 128)
Dropout | output: | (None, 128)
dense_4 | input: | (None, 128)
Dense | output: | (None, 64)
batch_normalization_10 | input: | (None, 64)
BatchNormalization | output: | (None, 64)
dropout_5 | input: | (None, 256)
Dropout | output: | (None, 256)
dense_3 | input: | (None, 256)
Dense | output: | (None, 128)
batch_normalization_9 | input: | (None, 128)
BatchNormalization | output: | (None, 128)
dropout_6 | input: | (None, 128)
Dropout | output: | (None, 128)
dense_4 | input: | (None, 128)
Dense | output: | (None, 64)
batch_normalization_10 | input: | (None, 64)
BatchNormalization | output: | (None, 64)
dropout_7 | input: | (None, 64)
Dropout | output: | (None, 64)
dense_5 | input: | (None, 64)
Dense | output: | (None, 32)
batch_normalization_11 | input: | (None, 32)
BatchNormalization | output: | (None, 32)
dropout_8 | input: | (None, 32)
Dropout | output: | (None, 32)
dense_6 | input: | (None, 32)
Dense | output: | (None, 7)

Fig. 5: CNN model architecture



```
model = Sequential([
    basemodel,
    Dropout(0.25),
    BatchNormalization(),
    Flatten(name="flatten"),
    Dense(64, activation='relu'),
    BatchNormalization(),
    Dropout(0.5),
    Dense(7,activation='softmax')
])
```

Fig. 6: ResNet50V2 finetuning

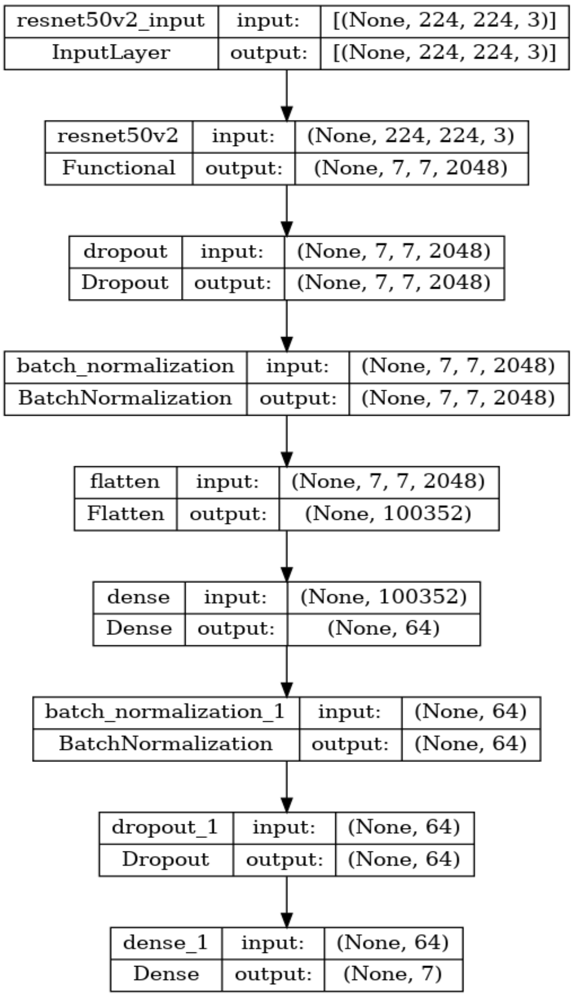
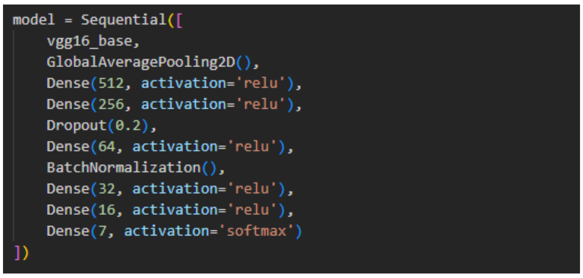

resnet50v2_input | input: | [(None, 224, 224, 3)]
InputLayer | output: | [(None, 224, 224, 3)]
resnet50v2 | input: | (None, 224, 224, 3)
Functional | output: | (None, 7, 7, 2048)
dropout | input: | (None, 7, 7, 2048)
Dropout | output: | (None, 7, 7, 2048)
batch_normalization | input: | (None, 7, 7, 2048)
BatchNormalization | output: | (None, 7, 7, 2048)
flatten | input: | (None, 7, 7, 2048)
Flatten | output: | (None, 100352)
dense | input: | (None, 100352)
Dense | output: | (None, 64)
batch_normalization_1 | input: | (None, 64)
BatchNormalization | output: | (None, 64)
dropout_1 | input: | (None, 64)
Dropout | output: | (None, 64)
dense_1 | input: | (None, 64)
Dense | output: | (None, 7)

Fig. 7: ResNet50V2 architecture



```
model = Sequential([
    vgg16_base,
    GlobalAveragePooling2D(),
    Dense(512, activation='relu'),
    Dense(256, activation='relu'),
    Dropout(0.2),
    Dense(64, activation='relu'),
    BatchNormalization(),
    Dense(32, activation='relu'),
    Dense(16, activation='relu'),
    Dense(7, activation='softmax')
])
```

Fig. 8: VGG16 finetuning

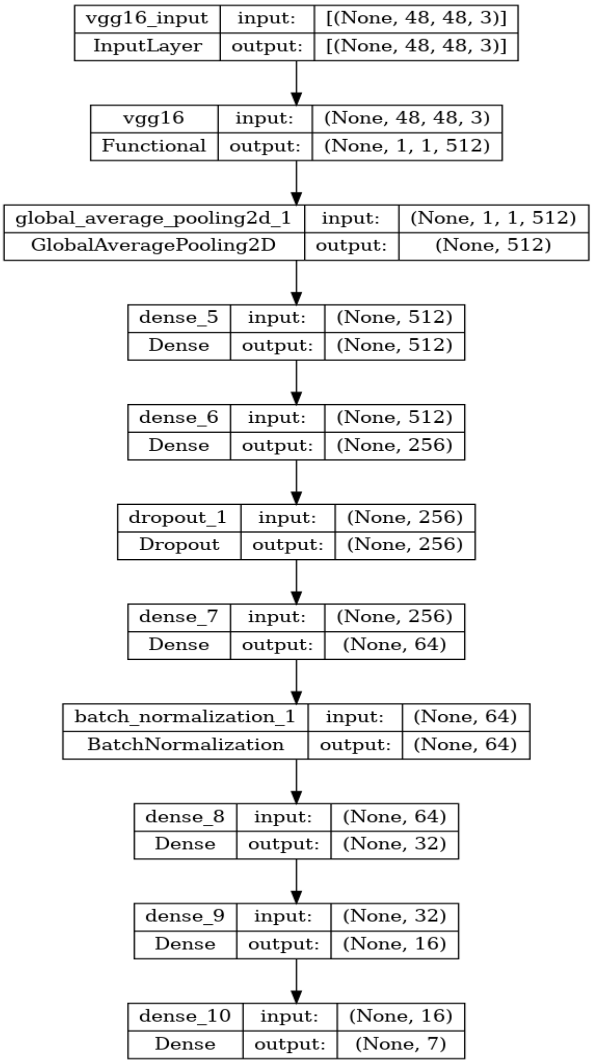

vgg16_input | input: | [(None, 48, 48, 3)]
InputLayer | output: | [(None, 48, 48, 3)]
vgg16 | input: | (None, 48, 48, 3)
Functional | output: | (None, 1, 1, 512)
global_average_pooling2d_1 | input: | (None, 1, 1, 512)
GlobalAveragePooling2D | output: | (None, 512)
dense_5 | input: | (None, 512)
Dense | output: | (None, 512)
dense_6 | input: | (None, 512)
Dense | output: | (None, 256)
dropout_1 | input: | (None, 256)
Dropout | output: | (None, 256)
dense_7 | input: | (None, 256)
Dense | output: | (None, 64)
batch_normalization_1 | input: | (None, 64)
BatchNormalization | output: | (None, 64)
dense_8 | input: | (None, 64)
Dense | output: | (None, 32)
dense_9 | input: | (None, 32)
Dense | output: | (None, 16)
dense_10 | input: | (None, 16)
Dense | output: | (None, 7)

Fig. 9: Finetuned VGG-16 architecture

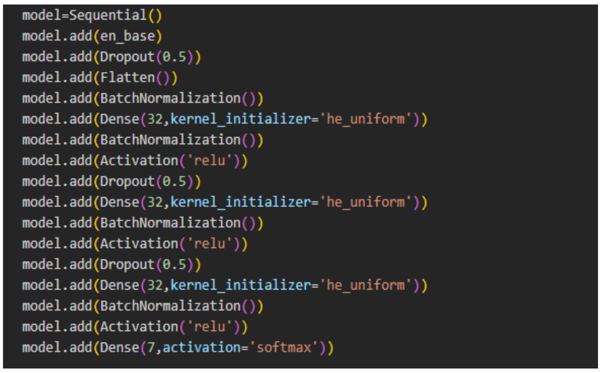
Figure 8 shows the layers that are added to fine-tune the pre-trained VGG-16 model. Except for the last 4, all the layers in the pre-trained VGG-16 model were freezed as a base model and on that 9 layers are added GlobalAveragePooling2D layer, 5 dense layers, a dropout layer with a dropout rate of 0.20 and a Batch normalization layer. Fig. 9 displayed the fine-tuned architecture of VGG-16 that was implemented.

EfficientNet B0 Model: Using a compound coefficient, the convolutional neural network construction and scaling method EfficientNet uniformly scales all depths, span, and pixel density dimensions. Unlike traditional practice, which scales these parameters freely, the EfficientNet scalability method uniformly adjusts network width, depth, and clarity using a set of predefined scalability coefficients. The reasoning behind the complex scaling method is that larger input images need more channels to catch more perfectly alright patterns on the larger picture and more layers to increase the network's receptive field. The fundamental EfficientNet-B0 network is composed of squeeze-and-excitation blocks as well as MobileNetV2 reversed bottleneck residual blocks. The fundamental EfficientNet-B0 network is composed of squeeze-and-excitation blocks as well as MobileNetV2 reversed bottleneck residual blocks. EfficientNets also transfer well and reach state-of-the-art accuracy despite employing orders of magnitude fewer parameters for the "cifar-100 (91.7%), Flowers (98.8%)" and 3 additional transfer learning datasets. In this system, a pre-trained EfficientNet B0 model is finetuned by freezing all the layers except the last 4 layers of the base model. Over the base model, 3 dropout layers with a dropout rate of 0.5, a flattened layer, 4 batch normalization layers, 3 dense layers with 32 hidden units and kernal initializer as "he_uniform", a dense layer with 7 hidden units and activation function as "softmax" and 3 activation layers with activation function "relu" are added. Fig. 10 displayers the layers added to finetune the pre-trained EfficientNet B0 model.

Figure 11 displays the implemented EfficientNet B0 architecture.

Emotion Detection: Model Training

The extended FER2013 dataset is split into training and validation images in the ratio of 90:10, in which there are 29,690 training images and 3,589 validation images. The number of layers and fine-tuning parameters is explained in the "emotion detection: Model fine-tuning" section. Table 3 shows the number of layers of the model, augmentation technique used, number of epochs for the model to be trained, learning rate, optimizer, loss function, and metrics used for the model. All the models were trained using Kaggle GPU Accelerator.



```
model=Sequential()
model.add(en_base)
model.add(Dropout(0.5))
model.add(Flatten())
model.add(BatchNormalization())
model.add(Dense(32,kernel_initializer='he_uniform'))
model.add(BatchNormalization())
model.add(Activation('relu'))
model.add(Dropout(0.5))
model.add(Dense(32,kernel_initializer='he_uniform'))
model.add(BatchNormalization())
model.add(Activation('relu'))
model.add(Dropout(0.5))
model.add(Dense(32,kernel_initializer='he_uniform'))
model.add(BatchNormalization())
model.add(Activation('relu'))
model.add(Dense(7,activation='softmax'))
```

Fig. 10: EfficientNet B0 finetuning

Table 3: Models of fine-tuning and training

Proposed models	Image augmentation techniques	Number of finetuned layers	Training Epochs	Learning rate	Optimizer	Loss function	Metrics
Basic CNN	Rescale = 1/255; rotation range =20; zoom_range = 0.2; width_shift_range = 0.1; height_shift_range = 0.1; horizontal_flip = True	39	50	Default, 0.001	Adam	Categorical_crossentropy	Accuracy
Finetuned ResNet-50 V2	Rescale = 1/255; rotation range = 10; zoom_range = 0.2; width_shift_range = 0.1; height_shift_range = 0.1; horizontal_flip = True	7	30	Default, 0.001	Adam	Categorical_crossentropy	Accuracy
Finetuned VGG-16	Rescale = 1/255; rotation range = 5; zoom_range = 0.2; width_shift_range = 0.2; height_shift_range = 0.2; horizontal_flip = True; vertical_flip = True	9	30	Default, 0.001	Adam	Categorical_crossentropy	Accuracy
Finetuned EfficientNet B0	Rescale = 1/255; rotation range = 5; shear_range = 0.2; horizontal_flip = True; vertical_flip = True	15	21	Default, 0.001	Adam	Categorical_crossentropy	Accuracy

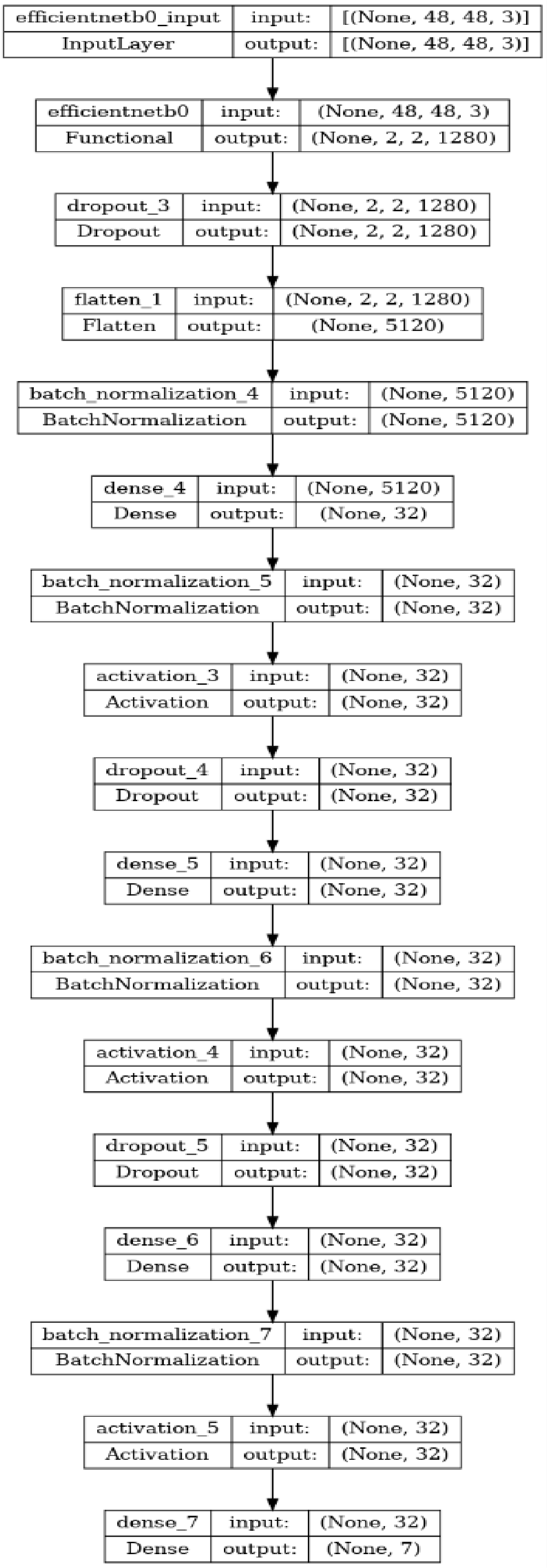

efficientnetb0_input | input: | [(None, 48, 48, 3)]
InputLayer | output: | [(None, 48, 48, 3)]
efficientnetb0 | input: | (None, 48, 48, 3)
Functional | output: | (None, 2, 2, 1280)
dropout_3 | input: | (None, 2, 2, 1280)
Dropout | output: | (None, 2, 2, 1280)
flatten_1 | input: | (None, 2, 2, 1280)
Flatten | output: | (None, 5120)
batch_normalization_4 | input: | (None, 5120)
BatchNormalization | output: | (None, 5120)
dense_4 | input: | (None, 5120)
Dense | output: | (None, 32)
batch_normalization_5 | input: | (None, 32)
BatchNormalization | output: | (None, 32)
activation_3 | input: | (None, 32)
Activation | output: | (None, 32)
dropout_4 | input: | (None, 32)
Dropout | output: | (None, 32)
dense_5 | input: | (None, 32)
Dense | output: | (None, 32)
batch_normalization_6 | input: | (None, 32)
BatchNormalization | output: | (None, 32)
activation_4 | input: | (None, 32)
Activation | output: | (None, 32)
dropout_5 | input: | (None, 32)
Dropout | output: | (None, 32)
dense_6 | input: | (None, 32)
Dense | output: | (None, 32)
batch_normalization_7 | input: | (None, 32)
BatchNormalization | output: | (None, 32)
activation_5 | input: | (None, 32)
Activation | output: | (None, 32)
dense_7 | input: | (None, 32)
Dense | output: | (None, 7)

Fig. 11: Finetuned EfficientNet B0 architecture

Music Recommendation System

The music recommendation system is developed using Spotify Dataset that was extracted using Spotify web API. The dataset consisted of ID, artists, song name, popularity, song duration, year, release_date, and Spotify audio features. The first step before building a music recommendation system is to explore the song dataset extracted using Spotify web API.

Figure 12 shows the histogram plot of the features of music tracks in the Spotify dataset. It can be analyzed that features such as energy, year, and valence have high variability and features such as liveness, danceability, tempo, and loudness have relatively lower variability and are concentrated in some ranges. The correlation matrix in Fig. 13 shows that danceability and valence have a high correlation, loudness has a high correlation with energy, accousticness has a negative correlation with energy, loudness, popularity, and year, popularity is highly positively correlated with year. Also, loudness and energy seem to have a strong correlation with popularity.

According to the year, the plot of audio features is displayed in Fig. 14. It can be analyzed that some characteristics decline over the years, e.g., accousticness, and instrumental Ness whereas others such as energy show an increase over the years. Features such as valence are more or less the same over the years.

After the song data is analyzed, the model with the highest accuracy is used to predict the emotion of the image and recommend songs accordingly.

For detecting emotions, four models are trained on an extended FER dataset and compared, which are the basic CNN model, finetuned ResNet50 V2 model, finetuned VGG-16, and finetuned EfficientNet B0. Out of the four models, the highest training accuracy 77.16%, and validation accuracy of 69.04% was obtained by the ResNet50 V2 model. Hence, this model was used for predicting emotions from images. Figure 15 shows the predicted classes of emotions using finetuned ResNet50 V2 Model. Emotions such as happiness, angry, surprise, neutral, fear, and sad were predicted correctly for most of the images in the dataset.

The next step is to recommend music based on the predicted emotions. The Spotify dataset is used for song recommendation. The main audio features such as acoustics, danceability, energy, instrumental Ness, liveness, loudness, popularity, speechiness, tempo, and valence are used for recommending songs. All these features are fit-transformed using MinMaxScaler from the sklearn library. The number of clusters is defined as 5. The data is split in the ratio of 75:25 using the train_test_split function from the sklearn library.

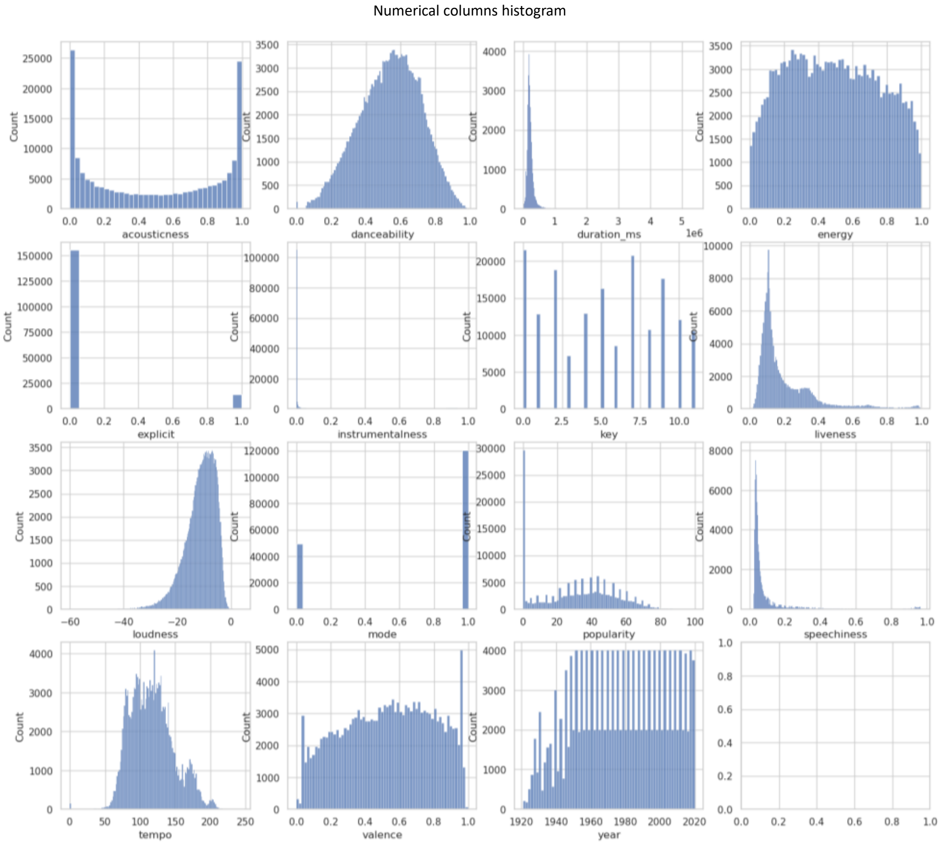
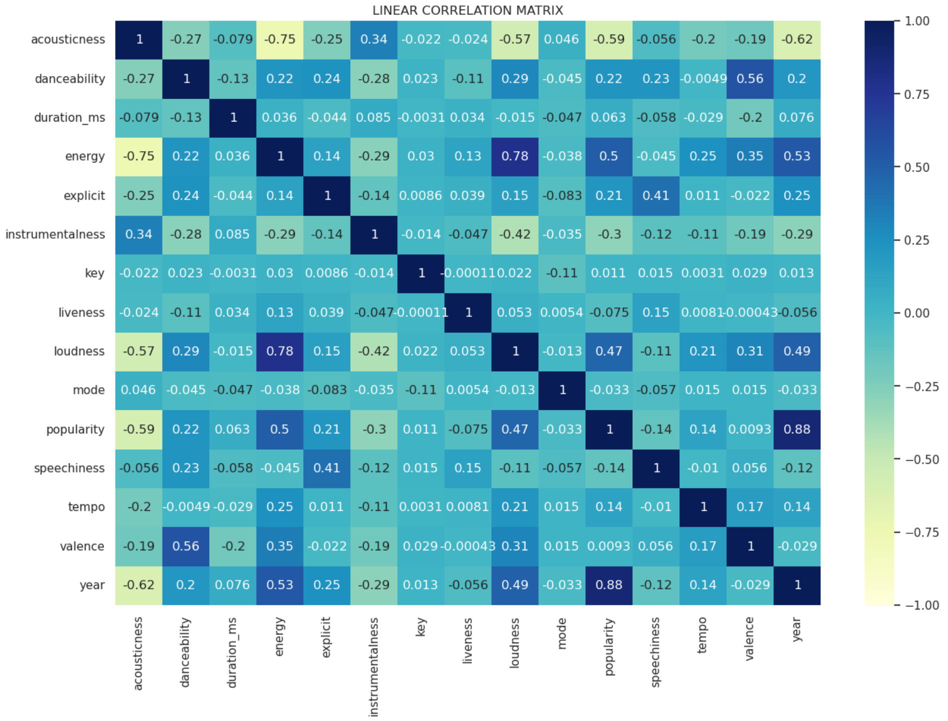


Fig. 12: Histogram plot of audio features



	acousticness	danceability	duration_ms	energy	explicit	instrumentalness	key	liveness	loudness	mode	popularity	speechiness	tempo	valence	year
acousticness	1	-0.27	-0.079	-0.75	-0.25	0.34	-0.022	-0.024	-0.57	0.046	-0.59	-0.056	-0.2	-0.19	-0.62
danceability	-0.27	1	-0.13	0.22	0.24	-0.28	0.023	-0.11	0.29	-0.045	0.22	0.23	-0.0049	0.56	0.2
duration_ms	-0.079	-0.13	1	0.036	-0.044	0.085	-0.0031	0.034	-0.015	-0.047	0.063	-0.058	-0.029	-0.2	0.076
energy	-0.75	0.22	0.036	1	0.14	-0.29	0.03	0.13	0.78	-0.038	0.5	-0.045	0.25	0.35	0.53
explicit	-0.25	0.24	-0.044	0.14	1	-0.14	0.0086	0.039	0.15	-0.083	0.21	0.41	0.011	-0.022	0.25
instrumentalness	0.34	-0.28	0.085	-0.29	-0.14	1	-0.014	-0.047	-0.42	-0.035	-0.3	-0.12	-0.11	-0.19	-0.29
key	-0.022	0.023	-0.0031	0.03	0.0086	-0.014	1	-0.00011	0.022	-0.11	0.011	0.015	0.0031	0.029	0.013
liveness	-0.024	-0.11	0.034	0.13	0.039	-0.047	-0.00011	1	0.053	0.0054	-0.075	0.15	0.0081	-0.00043	-0.056
loudness	-0.57	0.29	-0.015	0.78	0.15	-0.42	0.022	0.053	1	-0.013	0.47	-0.11	0.21	0.31	0.49
mode	0.046	-0.045	-0.047	-0.038	-0.083	-0.035	-0.11	0.0054	-0.013	1	-0.033	-0.057	0.015	0.015	-0.033
popularity	-0.59	0.22	0.063	0.5	0.21	-0.3	0.011	-0.075	0.47	-0.033	1	-0.14	0.14	0.0093	0.88
speechiness	-0.056	0.23	-0.058	-0.045	0.41	-0.12	0.015	0.15	-0.11	-0.057	-0.14	1	-0.01	0.056	-0.12
tempo	-0.2	-0.0049	-0.029	0.25	0.011	-0.11	0.0031	0.0081	0.21	0.015	0.14	-0.01	1	0.17	0.14
valence	-0.19	0.56	-0.2	0.35	-0.022	-0.19	0.029	-0.00043	0.31	0.015	0.0093	0.056	0.17	1	-0.029
year	-0.62	0.2	0.076	0.53	0.25	-0.29	0.013	-0.056	0.49	-0.033	0.88	-0.12	0.14	-0.029	1

Fig. 13: Correlation matrix

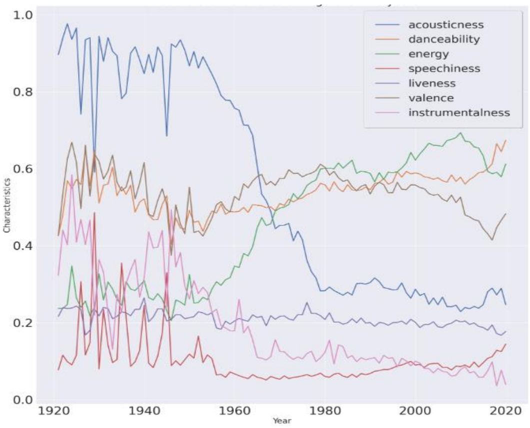


Fig. 14: Audio features over the years

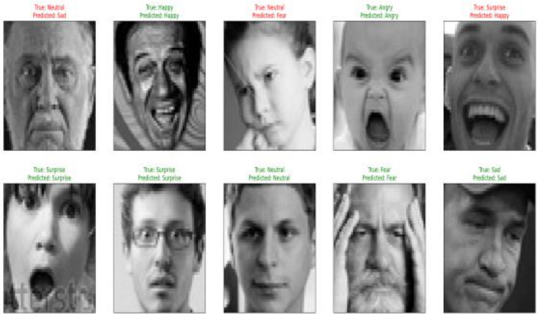


Fig. 15: ResNet50 V2 predictions

Results and Discussion

A lot of research had already been conducted on Emotion detection using facial expression recognition datasets. The previous studies have used the FER dataset, CK+ dataset, etc., for implementing emotion recognition models. For this study, the FER dataset has been extended by downloading images from Google using the google image download package in Python for each category of the FER2013 dataset. There are previous studies that used transfer learning models such as ResNet50, inception, ensemble, etc. for emotion detection. Table 4 shows the models and respective accuracies that were implemented in previous studies for facial expression recognition. For this study, 4 models were trained on an extended FER dataset, namely, CNN, finetuned ResNet50 V2, finetuned VGG-16, and finetuned EfficientNet B0. The results of the four models are explained below.

39-Layered Basic CNN Model

The input images to the basic CNN model were of size 48*48 which were obtained by data augmentation using an image data generator that produced rescaled, flipped, rotated, and sharp images. The basic CNN gave the output of predicting the class of emotion as per the FER dataset. A CNN architecture was developed that consisted of 6 convolutional layers, 12 batch normalization layers, one conv_2D input layer, 3 max-pooling layers, 9 dropout layers with a dropout rate of 0.25, and 7 dense layers. Figure 5 shows the implemented CNN architecture. The model was compiled using the "categorical cross entropy" loss function and "Adam optimizer". "Keras callback ReduceLROnPlateau" was used to reduce the learning rate for every 10 epochs if the validation accuracy does not increase. The model training was for 50 epochs and a batch size of 64 was used. The CNN with data augmentation resulted in a training accuracy of 69.13% and a validation accuracy of 66.17%. Figure 16 shows the loss vs epochs and accuracy vs epochs plot for training and validation.

Figure 17 shows the confusion matrix of implemented CNN architecture.

Finetuned ResNet50-V2 Model

The input images to the ResNet-50 V2 model were of size 224*224 which were obtained by data augmentation using an image data generator that produced rescaled, flipped rotated, and sharpened images. The ResNet-50 V2 model gave the output of predicting the class of emotion as per the FER dataset. In this system, ResNet-50 version 2 is implemented, which uses imagenet weights and requires input shape of (224, 224, 3). The pre-trained ResNet 50 is finetuned by adding 7 layers namely dropout with a dropout rate of 0.25, 2 batch normalization layers, flatten layer, 2 Dense layers with activation function as relu and softmax, and a dropout layer with a dropout rate of 0.5. Figure 7 shows the implemented ResNet50V2 architecture.

Table 4: Comparison of existing work accuracies and proposed model accuracies

Studies conducted	Datasets used	Models	Accuracy %
Khanzada *et al*. (2020)	FER2013, CK+, JAFFE	ResNet50	73.2
		SeNet50	70.0
		VGG-16	69.5
		Ensemble	74.8
Pramerdorfer and Kampel (2016)	FER2013	VGG	72.7
		Inception	71.6
		ResNet	72.4
Chowdary *et al.* (2021)	CK+	ResNet50	94.59
		VGG-19	89.19
		Inception V3	87.16
		MobileNet	96.62
Athavle *et al.* (2021)	FER2013	SVM	66
		ELM	62
		CNN	95

The model was compiled using the "categorical cross entropy" loss function and "Adam optimizer". "Keras callback ReduceLROnPlateau" was used to reduce the learning rate if the validation accuracy does not increase. The model training was for 30 epochs and a batch size of 64 was used. The ResNet-50 V2 model with data augmentation resulted in a training accuracy of 77.16% and a validation accuracy of 69.04%. Figure 18 shows the loss vs epochs and accuracy vs epochs plot for training and validation.

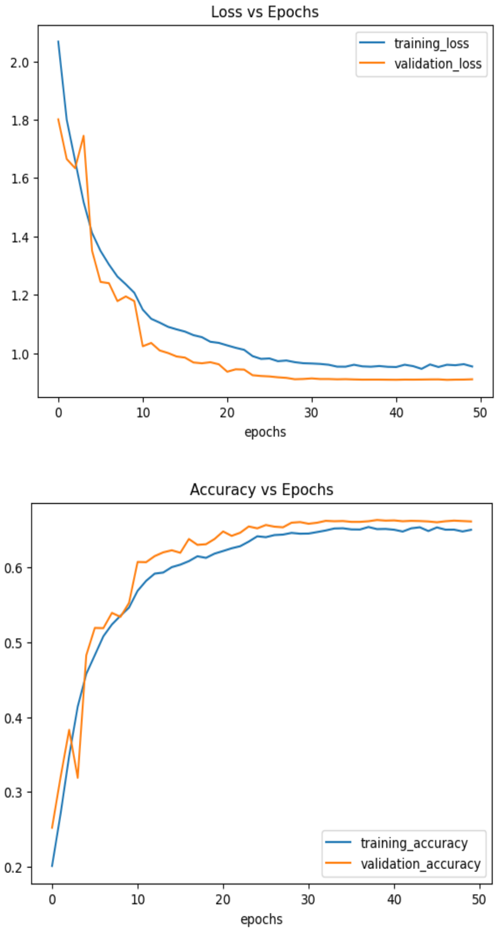


Fig. 16: CNN accuracy and loss plots

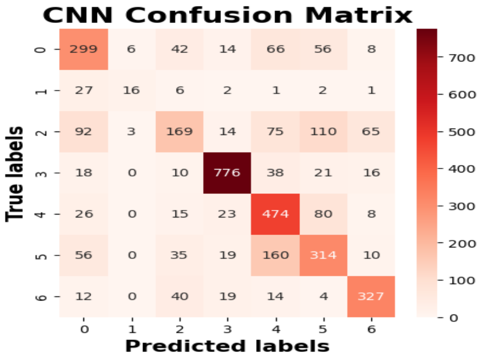


Fig. 17: CNN confusion matrix

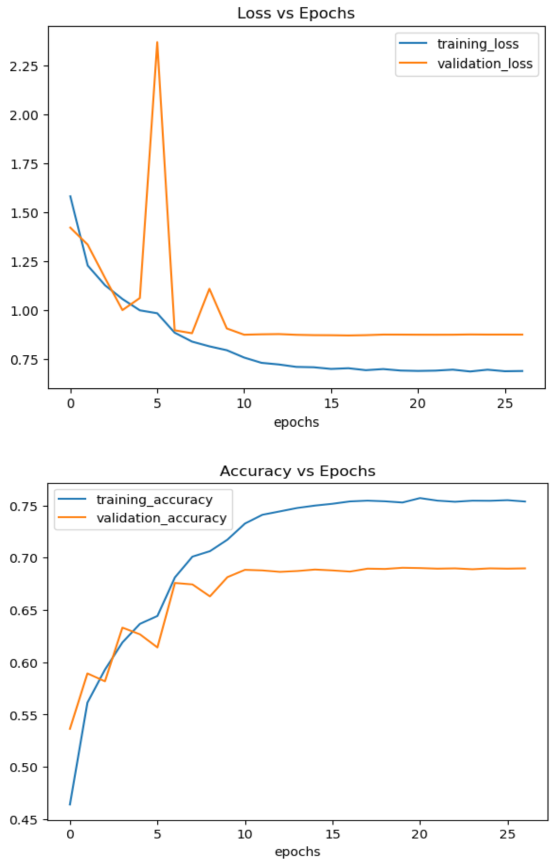


Fig. 18: ResNet50-V2 loss and accuracy plots

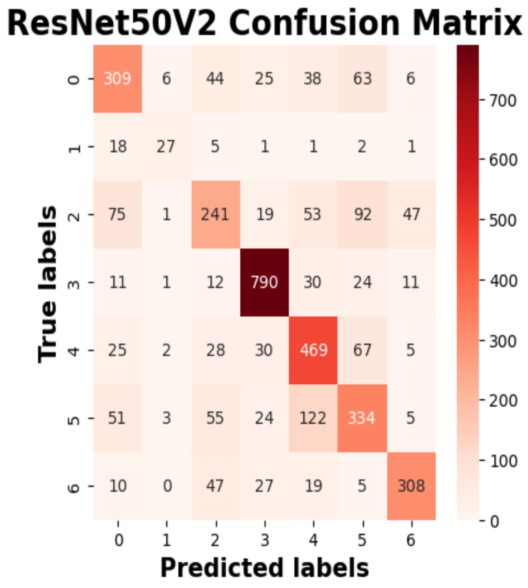


Fig. 19: ResNet 50 V2 confusion matrix

Figure 19 shows the confusion matrix of implemented ResNet50 V2 architecture.

Finetuned VGG-16 Model

The input images to the VGG-16 model were of size 48*48 which were obtained by data augmentation using Image Data Generator that produced rescaled, flipped rotated, and sharpened images. The VGG-16 model gave the output of predicting the class of emotion as per the FER dataset. In this system, VGG-16 is implemented, which uses imagenet weights and requires input shape of (48, 48, 3). The pre-trained VGG-16 is finetuned by adding 9 layers namely dropout with dropout rate 0.20, a batch normalization layer, 6 dense layers with activation function as relu and softmax, and a global average pooling layer. Figure 9 shows the implemented VGG-16 architecture. The model was compiled using the "categorical cross entropy" loss function and "Adam optimizer". "Keras callback ReduceLROnPlateau" was used to reduce the learning rate if the validation accuracy does not increase. The model training was for 30 epochs and a batch size of 64 was used. The finetuned VGG-16 model with data augmentation resulted in a training accuracy of 51.40% and a validation accuracy of 51.57%. Fig. 20 shows the loss vs epochs and accuracy vs Epochs plot for training and validation.

Figure 21 shows the confusion matrix of implemented ResNet50 V2 architecture.

Finetuned EfficientNet B0 Model

The input images to the EfficientNet B0 model were of size 48*48 which were obtained by data augmentation using an image data generator that produced rescaled, flipped rotated, and sharpened images. The EfficientNet B0 model gave the output of predicting the class of emotion as per the FER dataset. In this system, EfficientNet B0 is implemented, which uses imagenet weights and requires an input shape of (48, 48, 3). In this system, a pre-trained EfficientNet B0 model is finetuned by freezing all the layers except the last 4 layers of the base model. Over the base model, 3 dropout layers with a dropout rate of 0.5, a flattened layer, 4 batch normalization layers, 3 dense layers with 32 hidden units and kernal initializer as "he_uniform", a dense layer with 7 hidden units and activation function as "softmax" and 3 Activation layers with activation function "relu" are added. Figure 11 shows the implemented EfficientNet B0 architecture. The model was compiled using the "categorical cross entropy" loss function and "Adam optimizer". "Keras callback ReduceLROnPlateau" was used to reduce the learning rate if the validation accuracy does not increase. The model training was for 21 epochs and a batch size of 64 was used. The finetuned EfficientNet B0 model with Data Augmentation resulted in a training accuracy of 25.00% and a validation accuracy of 24.49%. Figure 22 shows the loss vs epochs and accuracy vs epochs plot for training and validation.

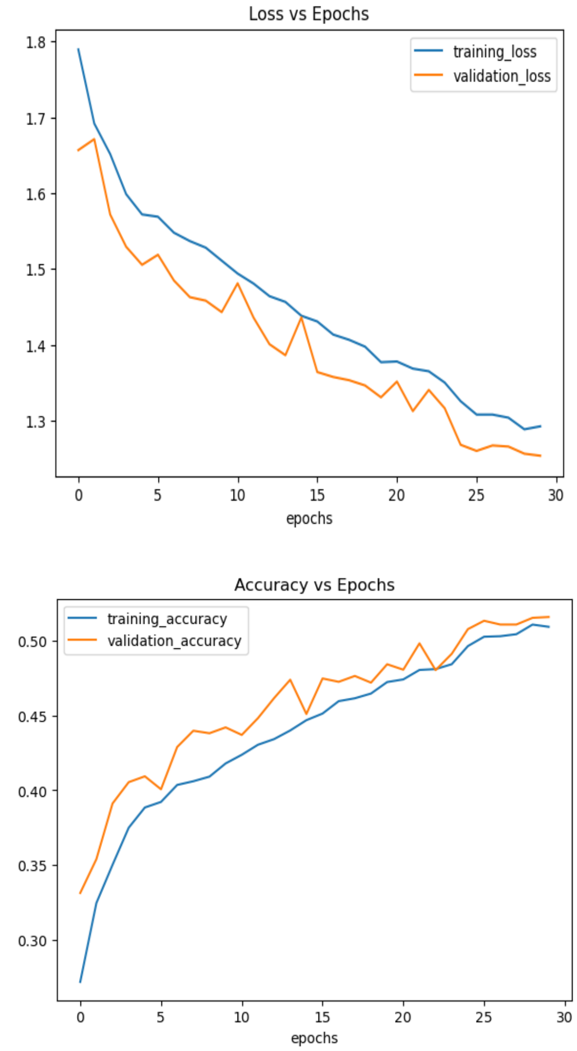


Fig. 20: VGG-16 loss and accuracy plots

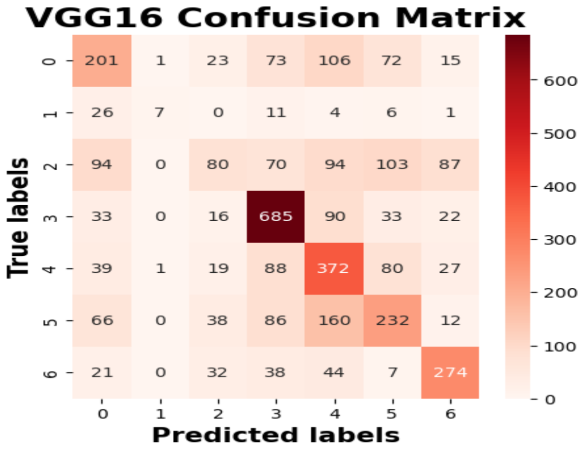


Fig. 21: VGG-16 confusion matrix

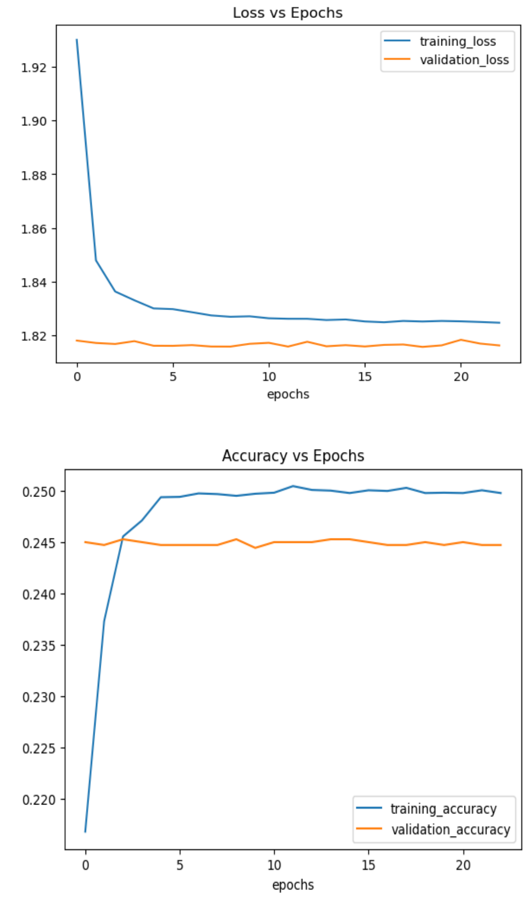


Fig. 22: EfficientNet B0 loss and accuracy plots

Figure 23 shows the confusion matrix of implemented ResNet50 V2 architecture.

Out of all the implemented models, the highest accuracy was obtained by the ResNet50 V2 model, which is 77.16% training accuracy and 69.04% validation accuracy. Finetuned EfficientNet B0 model performed the poorest with a training accuracy of 25.00% and validation accuracy of 24.49% which can be summarized in Table 5.

Recommendation Results

Comparing the four models, the highest accuracy was obtained by ResNet-50 V2 Model. Hence, it was used to predict the emotion for the recommendation of songs. K-means clustering is used for classifying the Spotify song dataset into clusters of various categories like "Angry", "sad", “surprised”, “neutral”, “disgust”, “fear” and “happy”. The recommended songs are sorted based on popularity. The recommendation results are displayed in Figs. 24-25. In recommendation 1, the model predicted the emotion as “happy”, so the recommender system suggested songs that could maintain the predicted emotion in a happy state. Whereas, in recommendation 2, the predicted emotion is “sad”, so the recommender systems suggested songs that could lighten up the predicted mood into a happy state.

Findings of the Study

This study aims at understanding the effectiveness of transfer learning in detecting emotions and its influence on personalized music recommendations. Table 5 shows the comparison of the previous research that has been conducted and the proposed study. It is essential to mention that the FER2013 dataset has been widely used in previous studies, but for this study, considering the generalizability of the dataset in a real-world situation, the dataset is extended by adding random images to emotion categories. Also, several transfer learning models have already been implemented in existing research. After fine-tuning the pre-trained transfer learning models and re-training them on the extended FER2013 dataset, it is found that the ResNet50 V2 model has the highest accuracy of 77.16%. The fine-tuned ResNet50 V2 model outperformed all the existing ResNet models used in the existing studies as mentioned in Table 5. As visualized in Fig. 19, the confusion matrix of ResNet50 V2, because of the imbalanced classes, the accuracy for emotion is also different. According to previous studies, it can be analyzed that compared to traditional machine learning algorithms, transfer learning models are more effective in predicting emotions from facial expressions. This study proves that transfer learning can impact the prediction of emotions by choosing the appropriate pre-trained model and tuning the hyperparameters to improve the accuracy of the model on the new dataset.

Table 5: Comparison of existing work accuracies and proposed model accuracies. Contd

Studies Conducted	Datasets used	Models	Accuracy %
Yen and Li (2022)	FER2013, AffecNet	ResNet50	70
		Xception	65
		EfficientNet B0	68
		Inception	67
Proposed models	FER2013 + google images	DenseNet121	71
		Fine-tuned 7-layered ResNet-50 V2	Training Acc: 77.16 Validation Acc: 69.04
		Fine-tuned 9-layered VGG-16	Training Acc: 51.40 Validation Acc: 51.57
		Fine-tuned 15-layered EfficientNet B0	Training Acc: 25.00 Validation Acc: 24.49

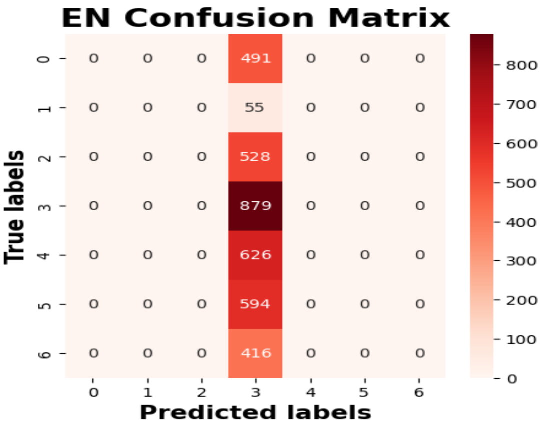
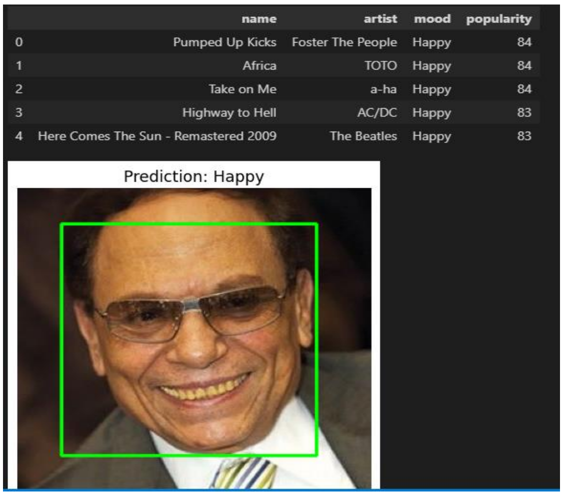


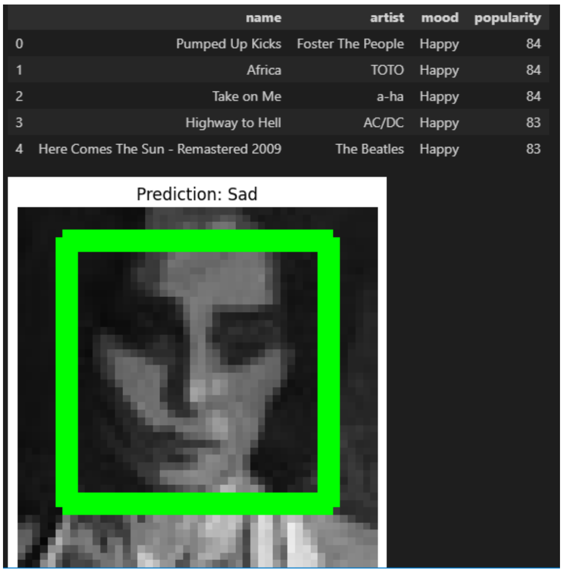
Fig. 23: EfficientNet B0 confusion matrix



	name	artist	mood	popularity
0	Pumped Up Kicks	Foster The People	Happy	84
1	Africa	TOTO	Happy	84
2	Take on Me	a-ha	Happy	84
3	Highway to Hell	AC/DC	Happy	83
4	Here Comes The Sun - Remastered 2009	The Beatles	Happy	83

Prediction: Happy

Fig. 24: Recommendation 1



	name	artist	mood	popularity
0	Pumped Up Kicks	Foster The People	Happy	84
1	Africa	TOTO	Happy	84
2	Take on Me	a-ha	Happy	84
3	Highway to Hell	AC/DC	Happy	83
4	Here Comes The Sun - Remastered 2009	The Beatles	Happy	83

Prediction: Sad

Fig. 25: Recommendation 2

Previous studies have also found that certain musical genres can stimulate particular emotional reactions, which can be seen in our facial expressions. The way we perceive and interpret music can also be influenced by facial expressions. Hence, this study also explains how the predicted emotions affect the music recommendations as displayed in Figs. 24-25. It can be justified that transfer learning affects the prediction of emotions, thereby affecting the recommendations of music.

Implications Theoretical and Managerial

For businesses and organizations that deal with music data, such as music streaming services, music recommendation systems, music analysis companies, and the fields of machine learning and music psychology, the impact of transfer learning on emotion prediction and music suggestion has various managerial and theoretical implications. These are a few of the implications.

Importance of emotional context: The success of models for music suggestion and prediction based on emotions demonstrates that emotional context affects how people perceive and prefer music. This emphasizes the need for a deeper comprehension of the psychological processes underpinning musically induced emotional responses.

Improving the accuracy of emotion prediction: For music data, transfer learning can dramatically increase the precision of emotion prediction models. Businesses that depend on precise emotion prediction for their products or services, like music recommendation systems or playlists based on mood, may find this to be useful.

Leveraging cross-domain transfer learning: Businesses may be able to use pre-trained models from other domains to increase the precision of their music analysis or recommendation systems by utilizing cross-domain transfer learning. For instance, a model that has already been trained to identify images can be modified to identify musical genres.

Customer retention: Since it offers users a more unique and engaging experience, emotion-based music suggestions can increase customer retention rates. For music streaming services that depend on subscription-based business models, this can be very helpful.

Interdisciplinary research: The use of transfer learning to anticipate emotions and make music recommendations demonstrates the potential advantages of interdisciplinary research. Combining machine learning and music psychology knowledge can produce fresh ideas and advancements in both disciplines.

Theoretical and managerial ramifications of emotion prediction-based music recommendation systems go beyond the confines of the music business. They offer insightful data on user activity, can enhance the overall user experience, and offer marketing opportunities and information on how new music is made.

Conclusion

In the present world, transfer learning is a technique that is increasingly in demand for improving emotion recognition and music recommendation systems. Due to the data explosion and the availability of large pre-trained models, transfer learning has developed into a powerful technique for utilizing existing knowledge to improve model performance and reduce the need for enormous amounts of labeled data. The goal of this study is to comprehend how transfer learning affects the precision of facial expression emotion detection and how personalized music recommendations can be made based on the identified emotions. This study examined the transfer learning of 4 models for understanding their efficiency and accuracy in predicting emotions. The models that are trained for this study are 39-layered CNN, finetuned ResNet-50 V2, Finetuned VGG-16, and finetuned EfficientNet B0. The experiment was conducted using the FER2013 dataset which was extended by adding random facial expression images from Google to get a generalized dataset. The dataset was kept imbalanced considering the unavailability of balanced datasets in the real-world scenario. All the input images were pre-processed using suitable image augmentation techniques. It was concluded that the fine-tuned ResNet 50 V2 model outperformed with the highest accuracy of predicting emotions. The model accuracies were compared to previous studies and it was observed that the ResNet 50 V2 that was fine-tuned for this study performed better. Further, a music recommendation system was developed that recommended songs using the Spotify song database, based on the predicted emotions.

From this study, it was observed that transfer learning could become one of the most promising approaches for detecting emotions from facial expressions due to its ease of using pre-trained models that are already learned using large databases. The already learned features could make the model train easily on new features and could also improve the accuracy of the model owing to its generalization capability. It is also concluded that it could affect music recommendations due to the prediction of emotions as understanding the emotional state of a person and recommending appropriate songs could enhance user engagement and increase user satisfaction.

The future direction of this research would be improving the accuracies of the model by adding or changing more hyperparameters and comparing the results. Real-time image capturing could be implemented to get the real-time emotional state of the user for more accurate recommendations. More diverse facial expression datasets with real-time images could be built for more accurate predictions. Other transfer learning models or ensemble learning techniques could be attempted in the future to see the difference in accuracy.

Future research may also focus on multimodal deep learning, which integrates data from many sources, including text, audio, and video, to increase the precision and interpretability of emotion recognition and music recommendation systems. To make sure that this system is created and utilized properly, future research should also look into the ethical implications of this system, including concerns about privacy, bias, and fairness.

Acknowledgment

The authors deeply acknowledge the insightful comments and suggestions made by the editors and reviewers of this manuscript.

Funding Information

This manuscript was not awarded funds from public, commercial, or non-profit funding organizations.

Author’s Contributions

Krishna Kumar Singh: Conception and designed, Analysis, and interpretation of data.

Payal Dembla: Conception and designed and acquisition of data.

Ethics

The authors confirm that there are no ethical issues or conflicts of interest.

References

Aggarwal, K., Mijwil, M. M., Al-Mistarehi, A. H., Alomari, S., Gök, M., Alaabdin, A. M. Z., & Abdulrhman, S. H. (2022). Has the future started? The current growth of artificial intelligence, machine learning and deep learning. *Iraqi Journal for Computer Science and Mathematics*, *3*(1), 115-123. https://doi.org/10.52866/ijcsm.2022.01.01.013

Athavle, M., Mudale, D., Shrivastav, U., & Gupta, M. (2021). Music Recommendation Based on Face Emotion Recognition. *Journal of Informatics Electrical and Electronics Engineering (JIEEE)*, *2*(2), 1-11. https://doi.org/10.54060/JIEEE/002.02.018

Bhattarai, B., & Lee, J. (2019). Automatic music mood detection using transfer learning and multilayer perceptron. *International Journal of Fuzzy Logic and Intelligent Systems*, *19*(2), 88-96. https://doi.org/10.5391/IJFIS.2019.19.2.88

Chowdary, M. K., Nguyen, T. N., & Hemanth, D. J. (2021). Deep learning-based facial emotion recognition for human computer interaction applications. *Neural Computing and Applications*, 1-18. https://doi.org/10.1007/s00521-021-06012-8

Chung, T. S., Rust, R. T., & Wedel, M. (2009). My mobile music: An adaptive personalization system for digital audio players. *Marketing Science*, *28*(1), 52-68. https://doi.org/10.1287/mksc.1080.0371

Florence, S. M., & Uma, M. (2020, August). Emotional detection and music recommendation system based on user facial expression. In *IOP Conference Series: Materials Science and Engineering* (Vol. *912*, No. 6, p. 062007). IOP Publishing. https://doi.org/10.1088/1757-899X/912/6/062007

Hasib, K. M., Tanzim, A., Shin, J., Faruk, K. O., Al Mahmud, J., & Mridha, M. F. (2022). BMNet-5: A novel approach of neural network to classify the genre of Bengali music based on audio features. *IEEE Access*, *10*, 108545-108563. https://ieeexplore.ieee.org/abstract/document/9916245/

Hung, J. C., Lin, K. C., & Lai, N. X. (2019). Recognizing learning emotion based on convolutional neural networks and transfer learning. *Applied Soft Computing*, *84*, 105724. https://doi.org/10.1016/j.asoc.2019.105724

IFPI. (2018). Global Music Report - Annual state of the industry. Retrieved from IFPI. *https://www.ifpi.org/ifpi-global-music-report-2018/ [Accessed 15 April 2023]*.

James, H. I., Arnold, J. J. A., Ruban, J. M. M., Tamilarasan, M., & Saranya, R. (2019). Emotion based music recommendation system. *Emotion*, *6*(03).

Joshi, S., Jain, T., & Nair, N. (2021, July). Emotion based music recommendation system using LSTM-CNN architecture. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 01-06). IEEE. https://ieeexplore.ieee.org/abstract/document/9579813

Khanzada, A., Bai, C., & Celepcikay, F. T. (2020). Facial expression recognition with deep learning. *arXiv preprint arXiv:2004.11823*. https://doi.org/10.48550/arXiv.2004.11823

Liu, Z., Xu, W., Zhang, W., & Jiang, Q. (2023). An emotion-based personalized music recommendation framework for emotion improvement. *Information Processing & Management*, *60*(3), 103256. https://doi.org/10.1016/j.ipm.2022.103256

Mahapatra, M., & Singh, K. K. (2022). Prediction of causes and effects of obesity in India by supervise learning approaches. *Obesity Medicine*, *34*, 100436. https://doi.org/10.1016/j.obmed.2022.100436

Meena, G., Mohbey, K. K., & Kumar, S. (2023). Sentiment analysis on images using convolutional neural networks-based Inception-V3 transfer learning approach. *International Journal of Information Management Data Insights*, *3*(1), 100174. https://doi.org/10.1016/j.jjimei.2023.100174

Modran, H. A., Chamunorwa, T., Ursuțiu, D., Samoilă, C., & Hedeşiu, H. (2023). Using Deep Learning to recognize Therapeutic Effects of Music Based on Emotions. *Sensors*, *23*(2), 986. https://doi.org/10.3390/s23020986

Nawaf, A. Y., & Jasim, W. M. (2023). A pre-trained model vs dedicated convolution neural networks for emotion recognition. *International Journal of Electrical & Computer Engineering (2088-8708)*, *13*(1).

Negre, M. A., Popescu, P.-S., Mocanu, M., & Mihaescu, M. C. (2022). MusicBud: A Music Recommendation System Based on Deep Learning algorithms. *Proceedings of RoCHI*, 130-136. http://rochi.utcluj.ro/articole/10/RoCHI2022-Negret.pdf

Pramerdorfer, C., & Kampel, M. (2016). Facial expression recognition using convolutional neural networks: State of the art. *arXiv preprint arXiv:1612.02903*. https://doi.org/10.48550/arXiv.1612.02903

Reddi, P. S., & Krishna, A. S. (2023). CNN Implementing Transfer Learning for Facial Emotion Recognition. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(4s), 35-45. https://www.ijisae.org/index.php/IJISAE/article/view/2569

Revathy, V. R., Pillai, A. S., & Daneshfar, F. (2023). LyEmoBERT: Classification of lyrics' emotion and recommendation using a pre-trained model. *Procedia Computer Science*, *218*, 1196-1208. https://doi.org/10.1016/j.procs.2023.01.098

Sekaran, S. A. R., Lee, C. P., & Lim, K. M. (2021, August). Facial emotion recognition using transfer learning of AlexNet. In *2021 9th International Conference on Information and Communication Technology (ICoICT),* (pp. 170-174). IEEE. https://ieeexplore.ieee.org/abstract/document/9527512

Shirwadkar, A., Shinde, P., Desai, S., & Jacob, S. (2022). Emotion Based Music Recommendation System. *International Journal for Research in Applied Science and Engineering Technology (iJRASET), 10*(XII), 690-694. https://doi.org/10.22214/ijraset.2022.47996

Singh, K. K. (2023). Study of Early Risks of Depression by Analysing Social Media Posts, *IIMS Journal of Management Science*, *14*(01), 2023. https://doi.org/10.1177/0976030X221112529

Verma, R. (2018). *Kaggle*. https://www.kaggle.com/datasets/deadskull7/fer2013

Yen, C. T., & Li, K. H. (2022). Discussions of Different Deep Transfer Learning Models for Emotion Recognitions. *IEEE Access*, *10*, 102860-102875. https://ieeexplore.ieee.org/abstract/document/9903451