National College of Ireland

Music Recommendation System Based On The Analysis Of The Images

MSc Research Project
MSc. In Data Analytics

Ramam Rajdev
Student ID: X22237216

School of Computing
National College of Ireland

Supervisor: Mr. Vikas Tomer

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: …….Ramam Rajdev……

Student ID: ……x22237216……..……

Programme: ……MSc.in Data Analytics…… **Year:** …2023-24…

Module: ……MSc. Research Project……

Supervisor: ……Mr. Vikas Tomer ……..……

Submission Due Date: ……16/09/2024……..……

Project Title:……Music Recommendation System Based On The Analysis Of The Images……

Word Count: ……9740…… **Page Count:** ……26……..……

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: …Ramam Rajdev……

Date: …16/09/24……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	☑
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☑
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☑

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Music Recommendation System Based On The Analysis Of The Images

Ramam Rajdev
X22237216

Abstract

Music recommendation system plays a significant role in the improvement of the user experience through recommender systems by enabling the listener to discover new artists/genre of music that they have an affinity towards. They are also helpful in the development of the music industry since they contribute to boosting activity, encouraging the subscription to streaming services, and introducing diverse music to the wide audience. The proposed music recommendation system in this research combines deep learning techniques and image analysis in order to provide improved user experience based on the books' recommendation and the corresponding music dependent on the derived emotional image sentiment. It uses Convolutional Neural Networks (CNNs) to identify emotion from image; models used include Visual Geometry Group 16 (VGG16), Visual Geometry Group 19 (VGG19), CNN & Residual Neural Network 50V2 (ResNet50V2). The highest selected performance accuracy was recorded in the scenario with the ResNet50V2 model at about 73%. While using the VGG19 model, the test set accuracy was about 63%; however, in the case of the custom CNN, the test set accuracy result was significant low of 2% and using VGG16 model similar to the custom CNN model, the test set accuracy was 0%. The identified emotions are matched with the moods and music tracks are suggested from a pool of songs which are maintained from a dataset. This approach does not follow the usual recommendation systems that do not incorporate a user's immediate emotional status in making recommendations, but rather provides a user-oriented approach.

1 Introduction

The use of online media platforms to distribute music could be amongst the most prominent and revolutionary changes in the lives of modern listeners of music. This change also clearly applies to the listening and producing of music. The latter continues to be somewhat limited to presenting only the users' preferences and tastes, but the emergence of complex multimodal recommendation systems is expected to result in significant shifts. These changes are quite contrasted with the pre-existing method that mainly involved the use of metadata in data procurement.

1.1 Background and Research Problem

While it has been comparatively well documented how people access information through computers, the way people listen to music has also substantially changed with the onset of the computer age. The rise of streaming services has sparked a meaningful change in the music industry by allowing its users instant access to an even larger selection of songs that exist in the market. However, due to the vast numbers of songs and artists available, users found that while there is a high possibility of finding songs they like, there is also a high likelihood that they would find it hard to find songs to fit their current mood or tone. Earlier, to manage this problem systems of recommending music have tried using the collaborative filtering system, a

system based on the content, and a system based on filtering both the content and collaboration (Ricci et al., 2011). These methods including the following do not adequately capture and/or respond to much of the user's current emotional condition, which is key for music choice.

1.2 Importance of Research

The invention of a music recommendation using the emotion analysis of images counteract several significant deficiencies in the existing recommendations systems. First, it offers a new form of input for users which marked as a new paradigm in contrast to the most well-known click interaction model that is based on the click path analysis and issued with the presence of visual data. This structured integration of contextual and affective features can notably improve the precision and appropriateness of suggestions acquired by exemplifying a broader view on the user's emotional condition (Cambria, et al., 2017).

Secondly, this research relates to the development of consumers' personalisation-based digital services. Since users demand highly targeted content with content marketing, the chance to feature music that fits the user's emotional condition increases user satisfaction (Said & Bellogín, 2014). This can result in higher user retention and satisfaction for music streaming service since the user will have the feeling that the service understands and cater to their needs i.e. designing a service that understands the user's current emotional state and based on it suggests the music which can improve the customer experience on the music platform.

1.3 Research Question

The primary research question is: How can the emotional sentiment derived from images be effectively integrated into music recommendation systems to enhance user experience?

1.4 Contributions and Benefits

This work is a contribution to the existing literature by integrating the emotion analysis of the images in music recommendation systems. Most current recommendation systems, based on users' previous activity and other attributes, do not consider the user's current mood. This paper contributes to the research by adroitly using deep learning to estimate moods from images and applying such emotions to present a better music selection that caters to the user's state of mind, making music recommendations relevant to the user's current feel.

On the commercial level, this study presents untapped prospects for music streaming applications and services, which might help to influence customers' behaviour and their habits when it comes to subscribing to or using these platforms. Mood sensitive suggestions can offer the competitive advantage, in terms of comfort and satisfaction rates among users. It can be used to market to the user, to create new products, and to improve user engagement, thus reflecting the user emotionally and practically for better result and relevance.

1.5 Structure of the Document

This chapter presents the research problem, importance, research questions, the Chapter 2 offers a detailed review of the current image-based emotion recognition and music recommendation systems. The third chapter of the report consists of comprehensive specification of technical tasks, like CNN implementation, data collection, resources required and annotation methods. The chapter 4 deals with the system implementation where in an in-depth discussion on the development of the models is discussed. The following chapter 5 discusses the results obtained from the experiments conducted in the study and finally chapter 6 provides the conclusion of the study including the extent to which the research question is answered and future scope for the study is also discussed.

All sections are an addition to the information provided in the previous sections, ensuring that the reader gets a comprehensive overview of the study and its importance for the field of music recommendation and affective computing.

2 Related Work

This section of the report discusses the state-of-the-art music recommendation systems based on the image sentiment analysis. It covers studies in the field from 2019-2024 with a special emphasis on the different advanced deep learning techniques for scene identification. The literature review is divided into 3 sections that include 1. Sentiment Analysis from Images, 2. Deep Learning in Scene Identification and 3. Music Recommendation Systems.

2.1 Sentiment Analysis in Images

Hsia et al. (2018) present a new representation learning framework to mitigate the heterogeneity divide between music and image datasets. The framework includes three main modules: as a CNN module, a network embedding (NE) module, and a retrieval module. The CNN module for image representation uses the VGG-19 model for the image representations. Specifically, the NE module learns representations from the heterogeneous data based on their neighbourhood proximity. The study adopts a dataset of 62,316 songs, seventy-two keywords and 33,459 images, the music dataset was sourced from KKBOX while images used were crawled using search engine. Hence, the proposed method provides a significant improvement over the suggested methods.

(Yang et al. 2021) also developed a model that integrates context understanding into emotion recognition through DL techniques. The proposed model aims at enhancing emotion recognition by combining the data from the image and context. HECO, EMOTIC, and Group Walk data sets were used. This is higher than the proposed model's average classification precision of 50 percent. 65% and their regression mean error rate was lower at 0. 7.

Wang et al. (2020) have put forward an approach to generate background music that aligns with the emotions elicited in virtual environment scenes. The method includes scene visualization and analysis, emotion recognition and generation of background music based on the trained neural network with the help of the chord sequences. The total volume of the dataset is three hundred pieces presenting different scenes connected with various emotional states (calm, happy, sad, angry, fearful). To conduct experiments, the nine different scenes were built into a virtual city. The background music synthesized was regarded for the emotional congruency, the quality of the transitions between music pieces, and the overall enriching of the VEnv experience. The proposed approach was faster in creation time in comparison to professional synthesis and professional qualitative experiments gave positive feedback.

For the same, the authors (Hoang et al. 2021) proposed the context-aware emotion recognition system using visual relationship detection. They employed a form of attention to include context related features and their attributes, these in combination with scene context and body features of the target subject for inferring emotion. The performance of the proposed system was evaluated on, CAER-S and EMOTIC datasets. The experimental results depicted that the proposed method achieved a better accuracy of 90% on CAER-S dataset. 49%, as well as competitive results on a benchmark – EMOTIC. The method properly introduced numerous factors from the scene context and proved to raise the model's recognition accuracy of emotion in an impressive manner.

ME2M is a model put forward by (Zhang et al. 2020) for image sentiment analysis. This 'exploit cross-modal sentimental semantics mining method and sample — refinement strategies' to enhance the sentiment analysis, the model employs other classifiers such as KNN,

LR, RF, DT, NB, AdaBoost, GBDT, and XGBoost. The proposed model was evaluated on one of the two basic benchmark datasets which are Twitter 1 and FI. Specifically, Twitter 1 is a microblog newly collected from the social networking site, and the sentiment polarity of its entries has been manually labelled by AMT workers. The proposed ME2M model proved to work better compared to several other state-of-the-art baselines, especially in image sentiment analysis tasks, both in terms of classification accuracy and resistance to influence. The findings revealed that implementing the proposed model enabled the identification of cross-modal sentimental semantics and improvement of sample data refinement to reduce the identification of sentiment classification.

A mixed approach to real-time sentiment analysis, CNN, coupled with ARs was proposed by (Desai et al. 2020) to analyze both text and image components of the social media posts. In this paper, Local Binary Pattern is used to feature and Support Vector Machine for the image sentiment classification. The dataset is in the form of a collection of 7000 tweets most of which are in the form of text and images. Text constitutes 50% of the distribution while the images are only 20% of the distribution and clearly; it divided into three; positive images, negative images, and neutral images. The model showed a high degree of correct affective classification – more than 92% of the time in the multimodal case. The results depicted a higher success rate over traditional methods with efficient pros and cons of visual and textual sentiments.

(Chidambaram et al., 2021) conducted a detailed work towards a music recommendation system that included identifying the user's emotions using a VGG16 model for facial recognition. The system incorporates a webcam in capturing images then analyses the images to deduce emotions of the user and through an integration of the Spotify API, provides the appropriate music. The Affect Net was used to train the VGG16 model, and this dataset has more than one million facial images that are tagged per emotion and the valence and arousal strength. In accomplishing its purpose, the system achieved approximately 98% accuracy on the training set and 70% on the test set of detecting emotion and in providing appropriate recommendations for music that fits the user's current emotional context.

2.2 Deep Learning in Scene Identification

(Chen and Li 2020) proposed a new method of music emotion classification using features derived from both the audio and text. The integration of both CNN and LSTM networks is possible through a complex network that entails the use of Multiple Input Multiple Output (MIMO) technology. The proposed approach includes obtaining spectrograms' two-dimensional features with CNNs and sequences with LSTMs. Besides, the authors introduced one-dimensional features extracted from audio during the classification, namely Low-Level Descriptors (LLDs) while the textual features included word embeddings and chi-squared test vectors derived from text extracted from the lyrics.

Another crucial advancement in the study is the attempt to apply the stacking method which creates an ensemble learning of emotional features from both audio and lyrics. It can be used to effectively address the problem of information loss that is ordinarily linked to dimensionality reduction procedures. The Last was used in the conduct of the study. fm tag subset of the Million Song Dataset, encompassing two thousand songs categorized into four emotional tags: which consist of angry, happy, relaxed, and sad. The outcomes revealed that the researchers obtained considerable enhancements in classification gains, where the audio only technique had 68% gains, the lyrics only model scored 74% and the combined model obtained an overall impression of 78%.

However, the study also brings out the following limitations: the audio features are complex; high dimensionality of spectrograms and word embeddings; possibility of overfitting; and the model may not be easily scalable for other large data sets.

On similar lines, (Sakthi Priya et al. 2023) put forward a system of music recommendation based on facial emotions identified by the YOLOv5 algorithm. It analyses the images to determine the emotion you are in and treats this emotion as a search term to the music link in a selected website. By using 3152 images of facial expressions to emotions from EPFL-RLC dataset the system suggested music with reference to detected emotion showing much higher achievement towards recognition and mapping of the emotions and corresponding track list.

According to (Mukhopadhyay et al. 2020) study, he examined the factors of adaptation of the online learning systems to the learners' emotions and modes. This paper votes out a dynamic recognition and assessment model of learner emotion, with CNN. The above model categorizes the feelings and kinds of mentality of learners during online learning sessions. Instead of having the frame of images, which are individual broken at various somewhere with different range of emotions, the approach utilizes continuous frames of images. The dataset was built using real surveys while a CNN model was fitted using the Facial Expression Recognition (FER2013) dataset, composed of 28,709 samples labelled into seven basic emotions: Aligned with Ekman's list of basic emotions, which consists of: anger, disgust, fear, happy, sad, surprised, and neutral. Classification accuracy in emotions was 65% while the state of mind recognition was 62%. However, they list some limitations; one being that facial detection for learners is possible only when the surrounding is well illuminated and during least learner movement.

In a recent study, (Jadhav, and Kaur 2023) presented Mood Melody Mapper (MMM) which is a music recommendation system that relies on the image content as well as context. The system uses a combination of image caption generation, cosine similarity for text and lyric, and sentiment analysis to direct image emotions towards music sentiments. It consists of numbers and songs, and the VizWiz-Caption dataset for the training of image captioning model. Different Feature extraction techniques like pre-trained models such as ResNet152 and BERT were used. The evaluation showed that the proposed system can recommend music that is relevant to the content and mood of images, such an approach to the integration of content and sentiment analysis of visual content in music recommendations is feasible. The concept of the approach can be effectively used in social networks, photo gallery and event playlists.

(Wen, 2021) developed an intelligent music recommendation system that used deep learning and the IoT technology. For the feature extraction, Fast-RCNN algorithm is employed while SIFT algorithm is used for image processing before classification is conducted using SVM for recommending background music based on video content. This system was evaluated using several public multimedia databases and commercial background music collections. For the user-generated videos experiments, the authors also downloaded some ‘‘free’’ video clips with the background music. The intelligent music recommendation system was proved to have high performance in anti-interference, robustness, and recognition effect in the practical scene. The integration of the deep learning function and IoT further upgraded the system's capability to select suitable background music for each video content type.

2.3 Music Classification

(Elbir and Aydin, 2020), for music classification and recommendation, suggested MusicRecNet which is based on a CNN model. Features are extracted from the Mel spectrogram images of music data, then these features are used for classification of music and recommending the similar ones. The performance of the proposed model was evaluated on the GTZAN dataset which contains one thousand music pieces belonging to ten different genres. Using the MusicRecNet, the mean accuracy that has been realized was 81.8% for genre classification. This outperformed the previous classifiers with an accuracy of up to 97% being obtained by the model. Six percent accuracy while in conjunction with SVM. The

recommendation was good when the systems recommended music based on genre similarity; it exhibited the best accuracy in classifying classical music and the worst in classifying rock music.

(Gomathy and Geetha 2022) highlighted the classification of music genre using the XG-Boost classifier. This paper involved extensive feature extraction with the aid of Mel Frequency Cepstral Coefficients (MFCC) and other spectral characteristics. In the present study, the GTZAN dataset with one thousand audio tracks and 30 seconds each selected from 10 genres was employed. The last accuracy score of XG-Boost classifier was 0. 81, with XG-Boost being more accurate that Random Forest and KNN.

(Elbir et al. 2018) used feature extraction techniques based on spectrogram analysis and CNN for music genre classification and recommendation. Each of the feature extraction methods included zero-crossing rate, spectral centroid, spectral contrast, and Mel-frequency Cepstral Coefficients (MFCC). SVM and KNN were the two algorithms that were evaluated against each other. The GTZAN dataset has one thousand music tracks each 30 seconds long and the ten genres include – blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock. In the aspect of classification accuracy, the SVM algorithm performed the best. Although deep learning models like CNN did not show a much better performance than conventional machine learning techniques in this study.

2.4 Music Recommendation Systems

On similar lines, (Hemholz et al., 2019) introduced Moosic, an emotion-based music recommendation application which this paper proposes to compare with other applications. To assign the identified emotions into two-dimensional space it applies Russell's Circumplex Model of Affect and Thayer's model that include arousal and valence variables. Included in the previous design are physiological parameters such as heart rate and skin temperature with the accompanying input to assess the emotional state of the user to propagate the appropriate music list. The study aimed at using an experiment with forty-three participants (30 male and 13 female), aged 20-34 years old who frequently used music streaming services. The PowerPoint prototype demonstrated high user satisfaction, and the concept of using the input of emotions focused on selecting music was serious and constructive. This research proved that by employing biofeedback, the system's ability to recommend mood-specific music in real-time will be more deterministic.

(Fathollahi and Razzazi 2021) examined a music recommendation system that is based on CNNs for finding the degree of similarity between music and provide suggestions of tracks. It selects features from other layers of the ANN and the classification includes metrics like Euclidian distance and angle between the two classes using cosine similarity. The study utilized three datasets: It comprises of GTZAN, Emotify music dataset, and the Music Audio Benchmark Dataset (MABD). The CNN based system proved to be more accurate able to classify music genres and able to determine relationships between genres. It can be said that the outcomes of the system are satisfactory when the results are compared to various datasets with distinctive characteristics and features of songs in order to generate recommendations for music.

The article by (Wen and Wang 2020) proposed an Emotion-Aware Music Recommendation system that incorporated emotion classification and music preference representation. Information regarding a user is encoded in the form of images and the deep learning system employs both CNNs and RNNs for analysis. It studies the interactions of users and produced music that fits your mood at the time of the day. To achieve this, the study relied on open-source musical datasets and voluntary users' behavioural data originating from music streaming platforms. The collected datasets were features such as audio signals, tags/history and ratings given by users. Specifically, the proposed system proved to be effective at

identifying the subjects' emotional state as well as at suggesting appropriate music to them. The enhancement of emotion-aware components also brought profound positive effects on users' experience and satisfaction. This revealed the fact that the system had a higher level of efficiency when compared to the time-tested approach of giving recommendations.

(Sarin et al., 2022) created SentiSpotMusic – an intelligent music recommendation system that incorporates the sentiment analysis mechanism to help in the recommendation of music. In this case, the VADER sentiment intensity analyzer and the Musixmatch API are used to categorize the songs into those that possess a negative, neutral, or positive sentiment. This dataset was sourced from Kaggle, and the focus is to use 169000 songs from 1920 to 2020, key attributes include track ID, acousticness, danceability, popularity, valence and year of release, tempo, energy, explicit content, instrumental Ness, liveliness, duration, and much more. Therefore, this study featured exploratory data analysis as well as sentiment analysis using Tableau software. Specifically, the results of the study showed that the energy, danceability and valence varied between different songs with the ones frequently played more often possessing higher levels in all three aspects. Accordingly with the proposed system, sentiment-based song recommendation proved successful.

For instance, deep multimodal approach integrating content features of audio with the metadata of users and items for cold-start music recommendation was explored by (Oramas et al. 2017). This involved the use of the convolutional neural networks (CNNs) for extracting features and putting these features together alongside the collaborative filtering techniques. The study utilized data from Last which is a database by the name of Last.fm. Particularly, FM has two types of user-generated data: tags and play counts, and other metadata. The first key limitation of the proposed deep multimodal approach was that the incorporation of features from the different modes considerably enhanced recommendation precision compared to traditional CF methods, especially in the cold-start problem.

2.5 Summary Table

Table 1: Review Summary

Author (year)	Datasets used	Methodology	Model Used	Results obtained		Limitation
(Hsia et al. 2018)	62,316 songs from KKBOX, 33,459 images from a search engine, 72 keywords	Image resizing to 224x224, feature extraction using VGG-19, keyword extraction using Jieba toolkit, network embedding for feature integration.	CNN module (VGG-19), network embedding (NE) module, retrieval module	Significant improvement in bridging heterogeneity between music and image datasets		Limited to the datasets used; generalizability not tested.
(Yang et al. 2021)	HECO, EMOTIC, and Group Walk datasets	Image resizing and normalization, feature extraction using ResNet-50, integration of facial and	DL techniques integrating context understanding	Precision	65%	Model precision not high enough for certain real-world applications.
				Mean error rate	0.7	

		contextual features, emotion detection using pre-trained models.				
(Wang et al. 2020)	300 pieces of music, 9 scenes in a virtual city	Scene segmentation, emotion classification based on scene features, feature extraction from audio and visual data, mood mapping to music.	Neural network for scene visualization, emotion recognition, and background music generation	Faster music creation with positive feedback on emotional congruency		Limited scalability for larger and diverse datasets.
(Hoang et al. 2021)	CAER-S, EMOTIC datasets	Image cropping and resizing, feature extraction using ResNet-18, context detection via Faster R-CNN, emotion mapping based on contextual analysis.	Context-aware emotion recognition system using visual relationship detection	Accuracy	90%	Dependency on context features may limit model's flexibility.
(Zhang et al. 2020)	Twitter 1, FI datasets	Image resizing and normalization, feature extraction using traditional and deep learning methods, sentiment mapping using cross-modal semantics.	ME2M model, cross-modal sentimental semantics mining, sample refinement strategies	Improved classification accuracy and resistance to influence		Limited to image sentiment analysis; other modalities not explored.
(Desai et al. 2020)	7000 tweets with text and images	Image preprocessing including resizing and normalization, feature extraction using CNN, text analysis	CNN, ARs for sentiment analysis of text and image components	Accuracy	92%	Model performance on larger, more diverse datasets not tested.

		using sentiment classifiers.				
(Chidambaram et al. (2021)	Affect Net dataset (over 1 million facial images tagged by emotion, valence, and arousal strength)	Image acquisition and resizing, facial feature extraction using VGG16, emotion detection and mapping to Spotify music database.	VGG16 model for facial recognition, integrated with Spotify API	Training Set Accuracy	98%	Performance variability between training and test sets is significant.
				Testing Accuracy	70%	
(Chen and Li 2020)	Last.fm tag subset of Million Song Dataset (2000 songs)	Feature extraction from audio and lyrics using CNN and LSTM, normalization, integration of multimodal features using MIMO.	CNN, LSTM, MIMO, word embeddings, chi-squared test vectors	Audio Only Accuracy	68%	High dimensionality; overfitting risk; limited scalability
				Lyrics Only Accuracy	74%	
				Combined Accuracy	78%	
(Sakthi Priya et al. (2023)	EPFL-RLC dataset (3152 images of facial expressions)	Image acquisition, resizing, feature extraction using YOLOv5, emotion detection and mapping to music tracks.	YOLOv5 algorithm for facial emotion recognition, music recommendation	High achievement in emotion recognition and mapping to corresponding tracks		Performance on real-time recognition not evaluated.
(Mukhopadhyay et al. (2020)	FER2013 (28,709 samples)	Image preprocessing including resizing and normalization, emotion and state of mind classification using CNN.	CNN for learner emotion and state of mind recognition	Emotion classification accuracy	65%	Requires well-illuminated surroundings; limited to static conditions
				State of Mind Recognition Accuracy	62%	
(Jadhav & Kaur (2023)	VizWiz-Caption dataset	Image caption generation using CNN-RNN, sentiment analysis, and cosine similarity for music	Mood Melody Mapper (MMM) with image caption generation, cosine similarity,	Successfully recommended music relevant to image content and mood		Limited testing on diverse user populations and contexts.

		recommendation.	sentiment analysis			
(Wen, 2021)	Public multimedia databases, commercial background music collections	Feature extraction using Fast-RCNN, integration with IoT sensors, normalization of sensor and visual data.	Deep learning, IoT, Fast-RCNN for feature extraction, SVM for classification	High performance in anti-interference, robustness, and practical scene recognition effect		Practical implementation challenges in varying IoT environments.
(Elbir & Aydin 2020)	GTZAN dataset (1000 music pieces, 10 genres)	Audio preprocessing including Mel spectrogram conversion, feature extraction using CNN, normalization and augmentation for genre classification.	MusicRecNet (CNN) with Mel spectrogram image features	Accuracy	81.8%	Performance drops with less structured or different genre datasets.
(Gomathy & Geetha 2022)	GTZAN dataset (1000 audio tracks, 10 genres)	Audio feature extraction using MFCC, normalization and scaling of features, classification using XG-Boost.	XG-Boost classifier with MFCC features	Accuracy	81%	Feature extraction methods may not generalize to all music types.
(Elbir et al. 2018)	GTZAN dataset (1000 music tracks, 10 genres)	Audio preprocessing including spectrogram analysis, feature extraction using CNN and SVM, normalization of feature vectors.	Spectrogram analysis, CNN, SVM, KNN for genre classification	SVM Accuracy	81%	Conventional ML techniques showed comparable results to DL methods.
(Helmholz et al. 2019)	Experiment with 43 participants	Image and audio data acquisition, feature extraction using emotional modeling, integration of	emotion-based music recommendation using Russell's Circumplex Model	High user satisfaction, effective biofeedback for mood-specific music recommendation		Limited by small sample size and experimental scope.

		biofeedback mechanisms.			
(Fathollahi & Razzazi 2021)	GTZAN, Emotify music dataset, and Music Audio Benchmark Dataset (MABD)	Audio preprocessing including feature extraction using CNN, normalization and similarity metric calculation for genre classification.	CNN-based system for music similarity metrics	Accurate in classifying genres and determining relationships	Model dependency on specific dataset characteristics.
(Wen & Wang 2020)	Open-source musical datasets, user behavioral data from music streaming platforms	Feature extraction using CNN and RNN, normalization and integration with user behavioral data for music recommendation.	Emotion-Aware Music Recommendation using CNNs, RNNs	Effective in identifying emotional state and suggesting music, enhanced user experience	Practical scalability and real-time data processing concerns.
(Sarin et al. (2022)	Kaggle dataset (169000 songs from 1920 to 2020)	Sentiment analysis of lyrics using VADER, audio feature extraction, normalization, and mapping of sentiment to music tracks.	SentiSpotMusic with VADER sentiment intensity analyzer, Musixmatch API	Successful sentiment-based song recommendation, varied energy, danceability, and valence in songs	Model's ability to handle new, evolving music sentiment trends.
(Oramas et al. 2017)	Last.fm user-generated data: tags, play counts, and metadata	Feature extraction from audio and text using CNN, integration of multimodal data, normalization and collaborative filtering for recommendations.	Deep multimodal approach, CNN for feature extraction, collaborative filtering	Enhanced recommendation precision in cold-start problem, effective use of multimodal features	Complexity in integrating multiple data modalities for general recommendation.

3 Research Methodology

This study employs a modified Knowledge Discovery in Databases (KDD) methodology, which consists of six essential steps: understanding of the problem and data, data pre-

processing, knowledge generation, assessment of the obtained knowledge and application of the obtained knowledge. This structured approach offers a framework for building the Light Side recommendation system based on the emotions of individuals, which strengthens the study's credibility. Research methodology entails specific procedures of research design, data collection techniques, data pre-processing, feature selection, model training and assessment, and the integration of the modified KDD methodology. An overview of the study methodology is illustrated in the project flow of Figure 1.

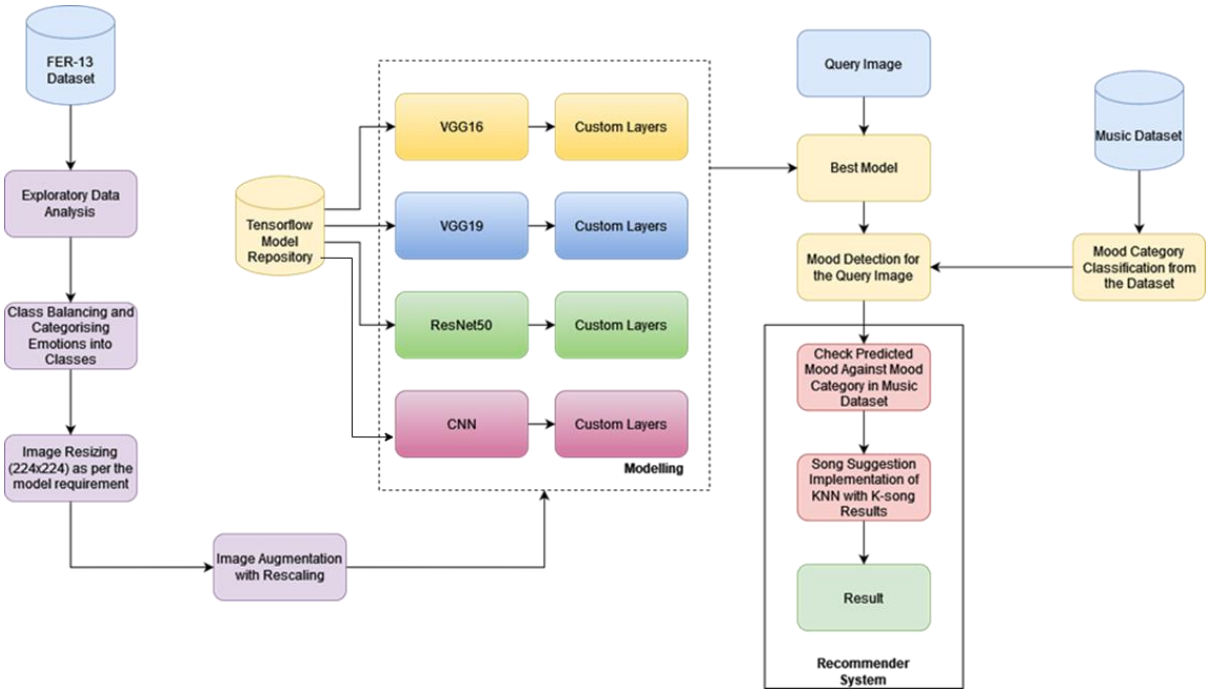


Figure 1: System Architecture

3.1 Data Collection

The data collection part in the study consists of two datasets viz. Facial Expression Recognition[1] Dataset and the Spotify[2] music dataset. Both datasets are obtained from the Kaggle repository.

3.1.1 FER Dataset

The datasets used in this study comprise two primary sources: The proposed approach uses the Facial Emotion Detection FER dataset and a music dataset with mood labels that can be found on Kaggle. The Emotion Detection FER dataset is a comprehensive collection of facial images labelled with seven distinct emotions: angry, disgusted, fear, happy, neutral, sad, and surprised and can be seen in Figure 2. The dataset is split into the training and testing data collection and every collection has subfolders of the seven emotions. This structure is quite useful in training and testing of machine learning models through the display of various possible facial expressions that are related to different feelings.

[1] https://www.kaggle.com/datasets/ananthu017/emotion-detection-fer
[2] https://www.kaggle.com/datasets/musicblogger/spotify-music-data-to-identify-the-moods

Figure 2: Seven Distinct Emotions from Dataset

3.1.2 Spotify Music To Identify the Moods Dataset

The dataset includes 686 samples with 19 attributes, which are connected to unique features of the audio tracks. These features include information regarding the authors, executors, genre, record label, publication date, and other criteria which help in detailed analysis and building of a strong recommendation system. The given data is in CSV format so that the data has a tabular structure for use it in analysis.
The list of attributes present in the dataset are enlisted in the table below.

Table 2: Music Moods Dataset

Attribute	*Description*
name	The title of the song.
album	The album in which the song is featured.
artist	The name of the artist or band who performed the song.
id	A unique identifier for the song within the Spotify database.
release_date	The date when the song or album was released.
popularity	A measure of the song's popularity, typically on a scale from 0 to 100.
length	The duration of the song in milliseconds.
danceability	A measure of how suitable the track is for dancing, based on tempo, rhythm stability, and beat.
acousticness	A confidence measure that indicates whether the track is acoustic.
energy	A measure of intensity and activity, representing how energetic a track feels.
instrumentalness	A measure predicting whether a track contains no vocals, with higher values indicating more instrumental content.
liveness	Detects the presence of an audience in the recording; higher values suggest a live performance.
valence	Describes the musical positiveness conveyed by the track, with higher values indicating more positive emotions.
loudness	The overall loudness of the track in decibels (dB).
speechiness	Measures the presence of spoken words in the track, with higher values indicating more speech-like content.
tempo	The speed or pace of the track, measured in beats per minute (BPM).
key	The key in which the track is composed.
time_signature	The time signature of the track, indicating the number of beats per bar.
mood	The categorised mood of the song, such as Happy, Sad, or Energetic.

3.2 Data Preparation

Data preparation is a critical phase that involves several sub-steps:

3.2.1 Data Balancing

To satisfy the necessity of the equal coverage of all the emotions and to make sure the data in the train and the test sets is balanced, the pictures from the Emotion Detection FER dataset are split into balanced train and test subdirectories (Batista, Prati and Monard, 2005). This includes applying transformations to each of the emotion categories to balance the numbers of images in the training and test sets.

3.2.2 Data Loading

The images are loaded and augmented with the help of ImagesDataGenerator class from the TensorFlow. This step involves downsizing the images and making batch from the train and test directories. Other approaches of data augmentation help improve the quality of the created dataset as augmentation, through transformations, helps to create more training example which in turn, helps to increase the ability of generalisation of the models.

3.2.3 Visualisation

To verify that images and labels from the training generator batch are loaded and categorized correctly, some of them are visualized. This step entails use of a matrix of images with their related emotion labels, which gives a confirmation that the required data has been rightly loaded and pre-processed.

Figure 3: Sample Data

3.3 Modelling

The data mining phase involves selecting appropriate models for emotion detection and training these models on the prepared dataset. The study employs four different models: VGG16, VGG19, CNN, and ResNet50V2.

3.3.1 VGG16 and VGG19 Models

VGG 16 and VGG19 are powerful deep convolutional neural networks. They have several convolutional layers that are succeeded by fully connected layers, and these make the network highly effective for image classification problems (Simonyan & Zisserman, 2014). VGG16 network has sixteen weight layers while the VGG19 has 19 which makes the network deeper to detect more features. The models used here are pre-trained on ImageNet, which means they can be brought in learned features for the purpose of Emotion Classification.

3.3.2 Convolutional Neural Network (CNN)

For this study, a Custom CNN is implemented using the following layers namely Convolutional, MaxPooling, Dropout, Flatten, Dense and Batch Normalization. Common to

most CNNs for image recognition are that convolutions permit capturing hierarchical features in an image. The design thus emphasizes factors such as complexity while at the same time. striking a balance that would allow it to properly detect the various emotions it has been designed to recognize without necessarily requiring much from a computer. The mathematical operations (Tariq, 2023) happening at different layers are detailed below.

1. *The convolution operation:* The convolutional layer of the model is reasonable for the convolution that happens between the input data and the filter kernel (K). The filter kernel used in the implementation is of 3x3 size. This convolution generates a feature map of extracted features from the image I. The equation for the feature map output is given below.

$$F[i,j] = (I * K)_{[i,j]}$$

Equation 1

The ij-th entry for the feature map is given by equation 2.

$$f[i,j] = \sum_{x}^{m1} \sum_{y}^{m2} K_{[x,y]} I_{[i-x,j+y-1]}$$

Equation 2

2. *Pooling Operation*: The pooling operation in a CNN can be performed using multiple methods such as max pooling, sum pooling, average pooling etc. In the pooling operation (MaxPooling) maximum entry from a 2x2 grid of the feature map is used to create a new feature map. This is visualised in Figure 4 (Tariq, 2023) below.

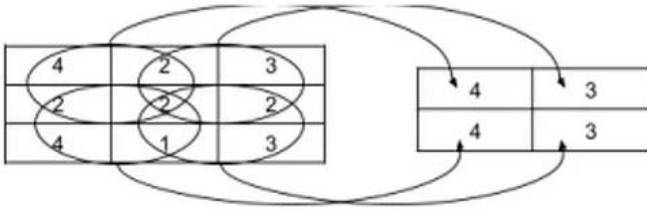


Figure 4: MaxPooling operation on the Feature Map

3. *Dense Layer Operation:* The dense layer in the model takes in the flattened data and converts it to another vector which can be the output or the input to the output neuron. The operation at the Dense Layer follows the equation (Tariq, 2023) given below.

$$X = \sum_{i}^{\square} w_i P_i + b_i$$

Equation 3

Where the output of the dense layer obtained through an activation function which will be based on the position of the dense layer. If the layer is not the output layer a ReLU activation function is used which follows the equation given below.

$$ReLU(X) = \max(0, X)$$

Equation 4

Where X is the input to the activation. If the dense layer is the output layer of the model, then the Sigmoid activation function is used for multi-class classification. Wherein the output of the Sigmoid activation function is obtained through equation below.

$$Sigmoid(X) = \frac{1}{1 + e^{-X}}$$

Equation 5

The final output of the Dense Layer then can be written as the equation (Tariq, 2023) given below.

$$z = g(X)$$

Equation 6

Where g is the activation function.

3.3.3 ResNet50V2 Model

ResNet50V2 is a residual network that uses the shortcut connections, which help to solve the vanishing gradient problems, thus making it easy to train very deep networks (He, 2016). This model also evolved on the ImageNet dataset and is recognized for its stability and performance in the image categorization tasks. The structure of ResNet50V2 consists of residual blocks so that the network can learn residual functions, which makes training easier and increases the level of the necessary changes.

3.3.4 Evaluation of the Models

The models implemented in the study are evaluated based on the accuracy the models achieve in detecting the emotions for the unknown data obtained from the test set of the dataset. The accuracy will be calculated by comparing the detected emotion by the implemented models and the actual emotion respective to the images in the test dataset.

3.4 Recommendation System

The final process of the methodology is to use the knowledge obtained in the data mining and evaluation phases to build the music recommendation system. This system makes use of the mood data found in the music dataset and applies the use of machine learning to create personalized recommendations for music depending on the mood identified.

3.4.1 Data Preprocessing for Music Recommendation System

To enhance the compatibility of the dataset features for music, some records are normalised including danceability, acousticness, energy, valence, and tempo. This is particularly important to keep the features on a similar scale to allow for a good model training and recommendations. The normalisation process of the feature values can be done through techniques that set the expected value of the feature equal to zero and makes the standard deviation equal to one.

3.4.2 Mood Mapping

Correspondence between the identified emotions within the images and the moods coming from the music database is created. It is essential for connecting the emotional state detected by the emotion detection model to the mood categories existing in the dataset of musical works. For instance, general moods in the music data set could include energetic or upbeat to which ‘happy’ and ‘surprised’ emotions might be assigned accordingly.

3.4.3 K-Nearest Neighbours (KNN) Model

Of all the algorithms, KNN is used in the recommendation of music tracks. To implement KNN for each of the moods, a model is trained that searches for the nearest neighbours in the feature space of the given music dataset. This helps in the determination of other songs which are in harmony with the detected emotion and thus recommend them to the system. The KNN model employs Euclidean distance to calculate distance and, thereby, identify the nearest tracks in terms of the features.

3.4.4 Recommendation Process

Finally using a mapping function if an emotion gets detected from the given facial image, then it is converted in terms of mood category. Subsequently, with the aid of the trained KNN model of this mood category, the nearest music tracks are searched for. The top recommendations are arrived at based on the distances which are nearer to each in the feature space. This way, the songs that recommend are in harmony with the user's current feeling, thereby increasing the chances of accepting the recommendations.

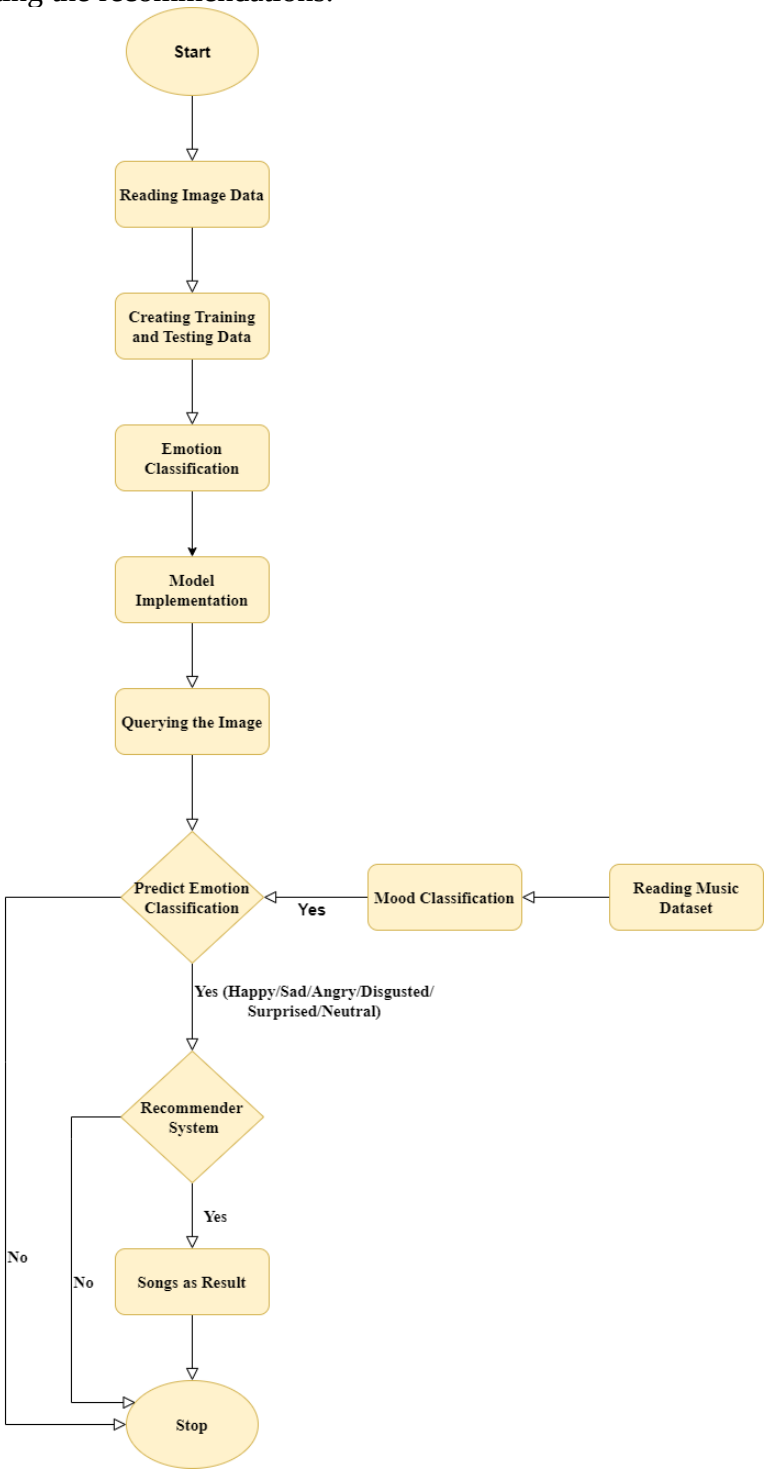


Figure 5: Flow Chart

3.4.5 Evaluation

The performance of the recommendation system in terms of recommended music that corresponds to the detected emotions is used in the assessment. Evaluations such as user feedback and accuracy are used for determining the system's effectiveness and drawbacks. The evaluation includes obtaining the precise performance of the system through a visual inspection of the recommendation provided by the model.

3.5 Conclusion

This chapter has carefully described the research method used in this study, which includes the modified KDD methodology which is used to define data mining roles; method of data collection; and pre-processing method, as well as the method used in the training of models. Thus, this methodology forms the basis for the systematic approach to building the emotion based music recommendation system suggested in the context of this study while meeting the identified objectives. In the implementation chapter, the authors will move away from laying down frameworks of theoretical concepts to deploying the strategies or models identified in order to arrive at the stated objectives.

4 Implementation

The implementation section involves the theoretical framework derived from the methodology to the realisation of these ideas, to evidence how the music recommendation system based on the recognition of emotions is formulated. First, it describes the setting up environment of tools and related libraries and platforms followed by integrating the preprocessed data into the selected machine learning models, which are VGG16, VGG19, CNN, and ResNet50V2, with focus on its implemented configurations and parameters. The chapter also provides an expansion of the actual construction of the music recommendation system, concentrating on the aforementioned conversion of emotions to moods alongside an explanation of the use of the K-Nearest Neighbours algorithm to generate personalised recommendations of music.

4.1 Environmental Setup

Table 3: Hardware requirements for the Experiments

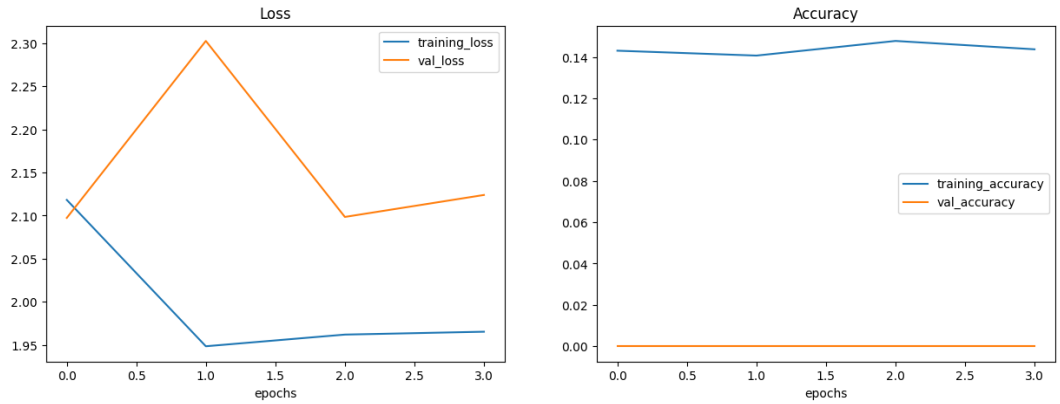
IDE	Google Colab (Cloud-based)
Computation	Python 3 Compute Backend
CPU	Intel Xeon
RAM	16 GB
Programming language	Python
Framework	Tensorflow
Modelling library	Sklearn, Imblearn, Numpy, Pandas, Matplotlib and Keras

4.2 Experiment 1: Implementation of the VGG16 for Emotion Recognition

The first model for emotion detection is based on the VGG16 model. This model has the input layer that takes images with the size of 224 x 224 pixels and 3 layers of RGB color. The model itself has its foundation in VGG16 – convolutional neural network pre-trained on ImageNet data; this made it possible to use transfer learning and rely on efficient features learned from a vast amount of data from the same source (yan & Zisserman, 2015). The top layer of VGG16 is then eliminated to suit the needs of emotion recognition since transfer learning should be

applied in the required project as per the recommendation by Yosinski et al., 2014. A flatten layer converts the output from the convolutional layers in the form of matrix of 2D into 1D vector While, dense layer with 1024 ReLU activation units captures complex patterns in the data as recommended by Nair & Hinton (2010). The last fully connected layer with the number of units as the number of emotion classes and ReLU activation function yields the predictions.

For model compilation, categorical cross-entropy loss is used, appropriate for multi class problems (Zhang & Sabuncu, 2018) and Adam optimizer is applied for efficient training of deep networks (Kingma & Ba, 2015). With regards to measurement, accuracy is adopted as the means of determining the ability of the model in the classification of emotion classes.
Figure 6 below depicts the model performance for the VGG16 in terms of loss 6(a) and accuracy 6(b).



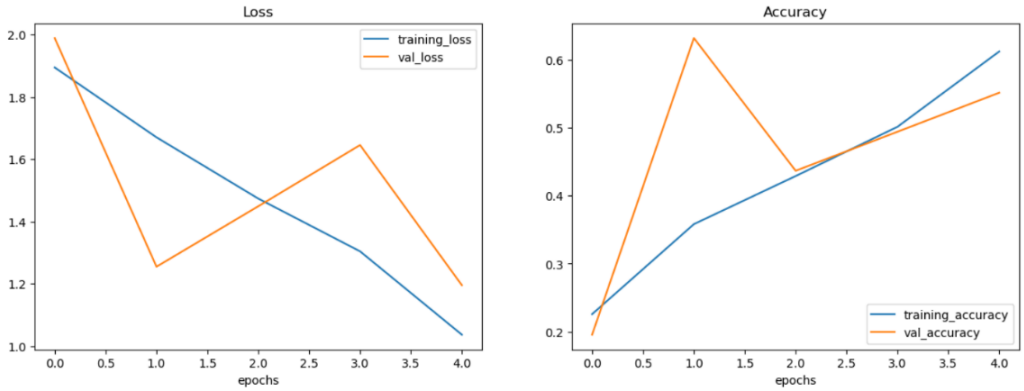
a) b)

Figure 6: VGG16 model Performance Curve

4.3 Experiment 2: VGG19 Implementation for Emotion Recognition

The second architecture has an input layer that takes images of 224 x 224 pixels and with 3 colour channels, RGB. This model's base is VGG19, another model of the CNN family, similar to ImageNet and augmenting the network's depth and feature extraction capabilities over the VGG16 (Simonyan & Zisserman, 2015). As in the previous architecture, the top classification layer of VGG19 is dispensable in order to insert layers suitable for emotion recognition Using the propensity of transfer learning (Yosinski et al., 2014). After the VGG19 base, a flatten layer, a fully connected layer with 1024 neurons and ReLU activation is added. In order to increase the process of regularisation and eliminate the problem of overfitting the network architecture a dropout layer with the dropout rate of 0.2 is introduced (Srivastava et al., 2014). Following this, there is another dense layer with 512 neurons and ReLU activation function and another dropout layer of drop rate 0.2. Before that, a dense layer having a number of neurons as same as the emotion classes and SoftMax activation function is used to obtain the output. The model is trained using the categorical cross-entropy loss function this is appropriate in multi-class problems, and the Adam optimizer at a reduced learning rate of 1e-4 for fine tuning during training (Kingma & Ba, 2015).

Figure 7 below depicts the training performance in terms of loss and accuracy for the VGG19 model.



a) b)

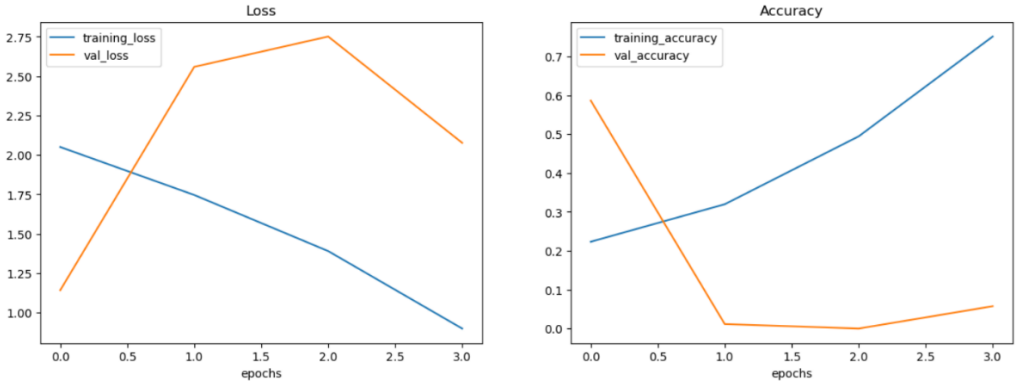
Figure 7: VGG 19 model Performance Curve

4.4 Experiment 3: Custom CNN model for Emotion Recognition

The third architecture also has the same input dimensions: images of 224 x 224 pixels and 3 bands (RGB). However, this model differs from the previously described approaches by using a specific CNN architecture based on the authors' design, rather than fine-tuning an already existing architecture such as VGG16 or VGG19. The convolution layers start with the Conv2D layer with 32 filters and the ReLU activation function with the kernel size of 3 x 3 to enable the model obtain basic features such as edges and textures from the input images (LeCun et al., 1998). This is followed by another layer known as MaxPooling2D layer with a pool size of 2 and padding set to be 'same'; This layer is productive in lessening the spatial dimensions of the formulated feature maps, and thus, minimizes the computational complexity in addition to preventing overfitting. A dropout layer with dropout rate of 0.25 is added.

After that, the feature maps are flattened and passed to a dense layer with 64 units which applies ReLU as activation function. In order to stabilise the model and speed up training and a Batch Normalization layer is added to normalise the activations of the previous layer to mitigate internal covariate shift (Ioffe & Szegedy, 2015). The architecture ends with a dense layer having a 7 units as the emotion classes and SoftMax activation.

The model is trained using categorical cross-entropy as loss function, since it is commonly used for multi-class problems, and the Adam optimizer with learning rate set to 1e-4, chosen because of its good performance in deep learning models' training (Kingma & Ba, 2015). Figure 8 below shows the training performance for the Custom CNN model.



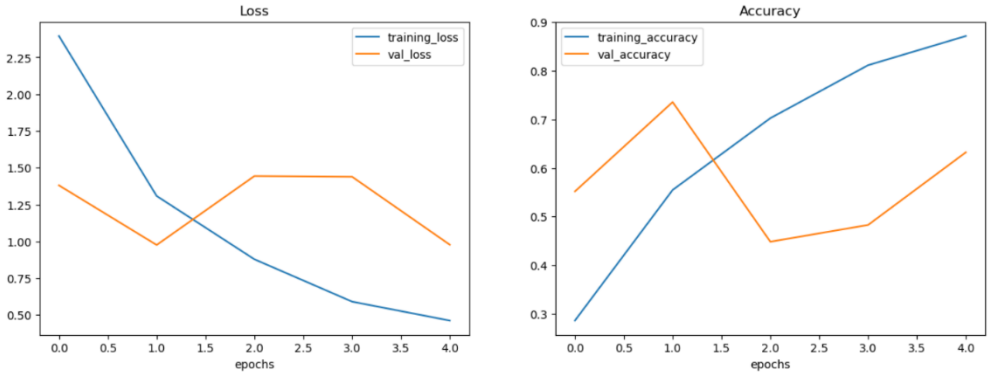
a) b)

Figure 8: CNN Performance Curve

4.5 Experiment 4: ResNet50V2 implementation for Emotion Recognition

The fourth architecture introduces a modified input with the shape 224x224x3 as RGB and uses ResNet50V2 as a base. ResNet50V2 belongs to ResNet family pre-trained on ImageNet, which is characterized by the great depth along with the effectiveness of feature extraction through the use of residual learning that reduces the effect of vanishing gradients experienced in deep networks (He et al., 2016). As with the previous architectures, the final classification layer of ResNet50V2 is omitted to enable future adaptations to the specifics of the emotion recognition tasks yet inherit the strong extracted features from ImageNet (Yosinski et al., 2014). In relation to the base model further layers added are a dropout layer, set at a rate of 0.25 is used for dropout layer. This is followed by a batch normalization layer. A flatten layer then follows. Next, there is dense layer with 64 units and ReLU activation. Then another batch normalization layer comes into the pipelines before the dropout layer with a dropout rate of 0.5. The architecture then ends with a dense layer containing the number of emotion classes where and SoftMax activation.

The model is compiled using categorical cross entropy loss which is suitable for multi-class classification problem and Adam optimizer with learning rate set at 1e-4 since it performs very well in the cases where the gradients are sparse and it adapts during training (Kingma & Ba, 2014). Promisingly, the accuracy metric remains the most important measure of the model's capacity to classify emotions. Figure 9 below shows the training performance for the ResNet50V2 model.



a) b)

Figure 9: ResNet performance curve

5 Evaluation

The obtained model's accuracy is examined on the test set to estimate the model's ability to classify emotions successfully. In the evaluation's section, the true labels are compared to the predicted labels with the intention of measuring the kind of emotion's identification that has been achieved. The findings are then discussed to pinpoint out the best model, which in its turn will be employed for the next step – the implementation of the music recommendation system. The models implemented in the study are evaluated based on the accuracy metrics for comparison.

5.1 Model Performance Evaluation

The performance of four different models (VGG16, VGG19, CNN, and ResNet50V2) was evaluated using accuracy as the primary metric. The models were trained on the emotion dataset and tested on a separate test set. The training and validation loss and accuracy were plotted for each model to visualize their learning curves.

Table 4: Accuracy

Models	Accuracy
VGG16	0.0000
VGG19	0.6321
ResNet50	0.7356
CNN	0.5862

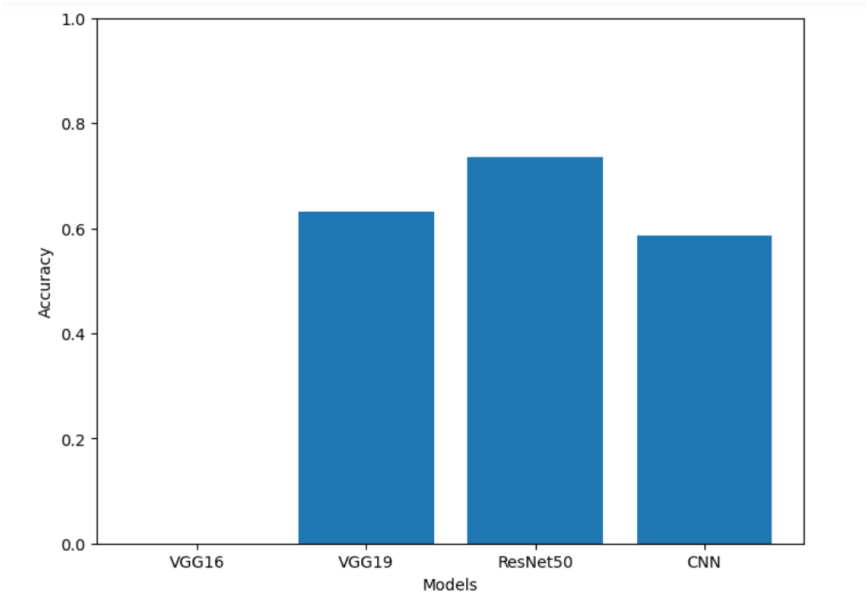


Figure 10: Model Comparison

5.2 A sample recommendation

A query image shown in Figure 11 is given to the best model for predicting the emotion.

	name	artist	mood	popularity
5	A Burden to Bear	Emmanuelle Rimbaud	Calm	27
636	While the Rest of the World Sleeps	Josh Kramer	Calm	35
28	Anaviosi	Alexi Musnitsky	Calm	38
43	Back+	Ton Snijders	Calm	26
418	Pierre	Hicham Chahidi	Calm	35

Figure 11: A sample recommendation for Angry Prediction

The model (ResNet50) predicted emotion to be angry which is correct. For the predicted emotion, five different songs are recommended using the KNN model as used in the study.

5.3 Discussion

Table 5 below compares the outcomes of this study with the past literature.

Table 5: Comparison with Existing work

Model	Score in This Study	Relevant Literature	Literature Performance	Comparison
VGG16	0.000000	(Chidambaram et al. 2021) - VGG16 for facial recognition in a music recommendation system	98% accuracy on training set; 70% on test set	VGG16 underperformed significantly in this study, contrasting with the high accuracy reported by Chidambaram et al. (2021). Further tuning might be needed.
VGG19	0.6321	(Hsia et al. 2018) - VGG19 for image representation in a heterogeneous data framework	Significant improvement in data representation task	VGG19 performed moderately in this study, aligning somewhat with its established efficacy in the literature but with less impactful results.
ResNet50	0.7356	(Wang et al. 2020) - ResNet used in emotional congruency for background music generation	Praised for stability and high performance in image categorization	ResNet50 outperformed other models in this study, consistent with its strong performance in related tasks as reported in the literature.
CNN	0.5862	(Desai et al. 2020) - CNN for affective classification in social media sentiment analysis	Over 92% accuracy	CNN performed moderately in this study compared to its reported high accuracy in the literature, indicating a possible need for further optimisation.

The goal of this paper was to determine how the inclusion of the emotional content seconded from images, in music recommendation systems, can improve the experience of users. The study proposed and tested several CNN-based models namely VGG16, VGG19, a proposed CNN, and ResNet50V2 for classifying emotions from captured images using deep learning. The identified emotions were matched with moods that correspond to the music data set which would facilitate the recommendation of the music tracks based on the user's mood.

To some extent, it was possible to achieve different levels of success in the models that were formulated. The efficiency of ResNet50V2 became apparent by revealing the highest accuracy of the emotion detection conforming the model's stable performance in the image categorization. As for the VGG16 model, it performs worse than others considerably, which also proves the need for improving the choice of models and their fine-tuning in deep learning

tasks. In doing this, the study was able to show that through image-based emotion detection, then stronger music recommendation systems could be created whereby the music recommended to the user is relevant to their current emotional state.

5.4 Future Work

The future work may consider directly the limitations arise in this study, and at the same time the analysis might broaden the range of the research. First, more enhancement of the suboptimal models, such as VGG 16 and Custom CNN, could be attempted to enhance the base models' capacity of detect emotions accurately. This may entail further adjustments to hyperparameters, data preprocessing such as data augmentation and consideration of different architectures or model combinations.

Moreover, expanding the set of images and associated emotions that are used in the research would increase the generalization capability of the models. Further adding the real-time identification of the user's emotional state and the ability to make recommendations based on it, using wearables or social media may also be an interesting avenue for the future research. The expansion of these future directions would extend knowledge and improvement to the suggested music recommendation system, which benefits the areas of affective computing and individualized digital environments.

References

Cambria, E., Poria, S., Gelbukh, A. and Thelwall, M., 2017. Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, *32*(6), pp.74-80.

Chen, C. and Li, Q., 2020. A multimodal music emotion classification method based on multifeature combined network classifier. *Mathematical Problems in Engineering*, *2020*(1), p.4606027.

Chidambaram, G., Ram, A.D., Kiran, G., Karthic, P.S. and Kaiyum, A., 2021. Music recommendation system using emotion recognition. *International Research Journal of Engineering and Technology (IRJET)*, *8*(07), pp.2395-0056.

Desai, N., Venkatramana, S. and Sekhar, B.V.D.S., 2020. Automatic visual sentiment analysis with convolution neural network. *International Journal of Industrial Engineering & Production Research*, *31*(3), pp.351-360.

Elbir, A. and Aydin, N., 2020. Music genre classification and music recommendation by using deep learning. *Electronics Letters*, *56*(12), pp.627-629.

Elbir, A., Çam, H.B., Iyican, M.E., Öztürk, B. and Aydin, N., 2018, October. Music genre classification and recommendation by using machine learning techniques. In *2018 Innovations in intelligent systems and applications conference (ASYU)* (pp. 1-5). IEEE.

G. Sakthi Priya, A. Evangelin Blessy, S. Jeya Aravinth, M. Vignesh Prabhu, R. VijayaSarathy, "Recommendation of Music Based on Facial Emotion using Machine Learning Technique", *Advances in Computational Intelligence in Materials Science*, pp. 102- 110.

Gomathy, C.K. and Geetha, V., 2022. Music Classification Management System. *International Journal of Early Childhood Special Education (INTJECSE) Doi*, *1*

Goodfellow, I., Bengio, Y. and Courville, A., 2016. Deep Learning. MIT Press.

He, K., Zhang, X., Ren, S. and Sun, J., 2016. Identity mappings in deep residual networks. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14 (pp. 630-645). Springer International Publishing.

Helmholz, P., Meyer, M. and Robra-Bissantz, S., 2019. Feel the moosic: emotion-based music selection and recommendation.

Hoang, M.H., Kim, S.H., Yang, H.J. and Lee, G.S., 2021. Context-aware emotion recognition based on visual relationship detection. *IEEE Access*, *9*, pp.90465-90474.

Hsia, C.C., Lai, K.H., Chen, Y., Wang, C.J. and Tsai, M.F., 2018. Representation learning for image-based music recommendation. *arXiv preprint arXiv:1808.09198*.

Ioffe, S. and Szegedy, C., 2015, June. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning (pp. 448-456). pmlr.

Jadhav, A.J. and Kaur, I., 2023. The Melodies of an Image: Exploring Music Recommendations Based on an Image's Content and Context. *Fifth Knowledge-aware and Conversational Recommender Systems (KaRS) Workshop @ RecSys 2023, pp. 1-4.*

Kingma, D.P., 2014. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Mukhopadhyay, M., Pal, S., Nayyar, A., Pramanik, P.K.D., Dasgupta, N. and Choudhury, P., 2020, February. Facial emotion detection to assess Learner's State of mind in an online learning system. In *Proceedings of the 2020 5th International Conference on Intelligent Information Technology* (pp. 107-115).

Nair, V. and Hinton, G.E., 2010. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th international conference on machine learning (ICML-10) (pp. 807-814).

Oramas, S., Nieto, O., Sordo, M. and Serra, X., 2017, August. A deep multimodal approach for cold-start music recommendation. In *Proceedings of the 2nd workshop on deep learning for recommender systems* (pp. 32-37).

Ricci, F., Rokach, L. and Shapira, B., 2010. Introduction to recommender systems handbook. In *Recommender systems handbook* (pp. 1-35). Boston, MA: springer US.

Said, A. and Bellogín, A., 2014, October. Comparative recommender system evaluation: benchmarking recommendation frameworks. In *Proceedings of the 8th ACM Conference on Recommender systems* (pp. 129-136).

Sarin, E., Vashishtha, S. and Kaur, S., 2022, February. SentiSpotMusic: a music recommendation system based on sentiment analysis. In *2021 4th International Conference on Recent Trends in Computer Science and Technology (ICRTCST)* (pp. 373-378). IEEE.

Sheikh Fathollahi, M. and Razzazi, F., 2021. Music similarity measurement and recommendation system using convolutional neural networks. *International Journal of Multimedia Information Retrieval*, *10*, pp.43-53.

Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), pp.1929-1958.

Wang, Y., Liang, W., Li, W., Li, D. and Yu, L.F., 2020, October. Scene-aware background music synthesis. In *Proceedings of the 28th ACM International Conference on Multimedia* (pp. 1162-1170).

Wang, D., Zhang, X., Yu, D., Xu, G. and Deng, S., 2020. Came: Content-and context-aware music embedding for recommendation. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(3), pp.1375-1388.

Wen, X., 2021. Using deep learning approach and IoT architecture to build the intelligent music recommendation system. *Soft Computing*, *25*(4), pp.3087-3096.

Werner, A., 2020. Organizing music, organizing gender: algorithmic culture and Spotify recommendations. *Popular Communication*, *18*(1), pp.78-90.

Wu, B., Zhong, E., Horner, A. and Yang, Q., 2014, November. Music emotion recognition by multi-label multi-layer multi-instance multi-view learning. In *Proceedings of the 22nd ACM international conference on Multimedia* (pp. 117-126).

Yang, D., Huang, S., Wang, S., Liu, Y., Zhai, P., Su, L., Li, M. and Zhang, L., 2022, October. Emotion recognition for multiple context awareness. In *European Conference on Computer Vision* (pp. 144-162). Cham: Springer Nature Switzerland.

Yosinski, J., Clune, J., Bengio, Y. and Lipson, H., 2014. How transferable are features in deep neural networks?. Advances in neural information processing systems, 27.

sYu, Y., Shen, Z. and Zimmermann, R., 2012, October. Automatic music soundtrack generation for outdoor videos from contextual sensor information. In *Proceedings of the 20th ACM international conference on Multimedia* (pp. 1377-1378).

Zhang, H., Wu, J., Shi, H., Jiang, Z., Ji, D., Yuan, T. and Li, G., 2020. Multidimensional extra evidence mining for image sentiment analysis. *IEEE Access*, *8*, pp.103619-103634.

Tariq, F., 2023. Breaking Down the Mathematics Behind CNN Models: A Comprehensive Guide. https://medium.com/@beingfarina/breaking-down-the-mathematics-behind-cnn-models-a-comprehensive-guide-1853aa6b011e.