Deep Learning-Based Emotion-Driven Music Recommendation System

*

1st Dr. M. S. Maharajan
Associate Professor
Department of AI&DS
Panimalar Engineering College
maha84rajan@gmail.com

2nd Mrs. V. Rekha
Assistant Professor
Department of AI&DS
Panimalar Engineering College
rekhav20@gmail.com

3rd P. R. Gopikashree
Student
Department of AI&DS
Panimalar Engineering College
gopikassakipog@gmail.com

4th M. Devadarshini
Student
Department of AI&DS
Panimalar Engineering College
vasudevakrishnan171102@gmail.com

***Abstract*—People often find it hard to choose the perfect song that fits their present mood in the digital age. As a result, they waste unnecessary time looking for music that relates to their feelings. Song recommendation systems can greatly improve user experience by incorporating the latest developments in Deep Learning and Artificial Intelligence (AI). This study aims to create an emotion-based music recommendation system that automatically makes song recommendations by analysing a user's recorded facial expressions. When a user uploads a face image, the system uses a Convolutional Neural Network (CNN), specifically ResNet50V2, to detect emotions and do image pre-processing. Next, a suitable song is suggested based on the detected emotion's mapping to a related musical genre. A real-time and dynamic music selection process based just on the user's emotional state is provided by this method, in contrast to typical systems that rely on user input or preference history. This study shows how artificial intelligence (AI) may improve tailored entertainment experiences by giving consumers a simple and natural way to choose music. The technology guarantees a quick, interesting, and emotionally responsive music-recommendation experience by doing away with the necessity for manual searches.**

Index Terms—**Emotion recognition, Deep Learning, Music Recommendation, ResNet50V2, Transfer Learning, Facial Expression Recognition, Computer Vision.**

I. INTRODUCTION

The universal language is music. It has played a vital role in our lives from the dawn of mankind. Both on our terrible days and on our good days, we turn to music for solace. We are enlightened and inspired by music. Music can take many different forms, from the harp to electric guitar riffs, from drum rhythm to bird tweeting. No matter one's caste, creed, or religion, music unites people. It has a significant impact on our lives and unites people. Music has an impact on people's bodies and minds in addition to being an art form and a language. It engages our minds. Research indicates that music has therapeutic qualities, and programs that use music therapy can benefit people with anxiety, dementia, stress, and confidence issues [1].

A study in the journal of neuroscience suggests that customized music-based therapies are recommended for the treatment of brain illnesses linked to aberrant mood and emotion-related brain activity. The way people listen to music has evolved in the modern period, particularly with the explosive expansion of streaming services and apps like TikTok. Music is evaluated based on its popularity rather than its quality [2]. This makes it more difficult for people to hear great music from underappreciated musicians. Another kind of nonverbal communication is through facial expressions. They are the primary social communication mechanism used by humans, a majority of mammals, and several different kinds of animals.

Convolutional neural networks, or CNNs, are frequently used in facial recognition systems due to their superior image processing capabilities, particularly when it comes to interpreting facial expressions and features. CNN can recognize features in photos automatically as a type of deep learning system. CNN is used for facial recognition and may be trained to recognize certain facial characteristics, including the eyes, nose, and mouth, to find unique patterns for different people [3]. These patterns can then be used to identify and classify people. The following steps are involved in Recommending music through emotions:

- Image Capture: The user uploads a picture to supply an input image. The foundation for detecting and recommending emotions is this image.
- Image Pre-processing: Pre-processing methods including noise reduction, scaling, and grayscale

conversion are applied to the acquired image in order to improve feature extraction. Image quality and dimensions are guaranteed to be consistent through normalization.

- Facial Emotion Detection: A Convolutional Neural Network (CNN) model (ResNet50V2) analyses the pre-processed image to identify the user's emotional expression and identify face landmarks. Happy, sad, furious, shocked, neutral, and so on are some of the categories into which the model divides emotions.\
- Emotion Classification & Mapping: The identified emotion is assigned to a relevant music genre using predetermined emotion-to-genre relationships. Example:
  - Happy → Pop, Dance, Upbeat
  - Sad → Soft, Classical, Blues
  - Angry → Rock, Metal
  - Fear → Ambient, Instrumental, Chill
  - Surprise → Electronic, Experimental, Jazz
  - Neutral → Acoustic, Lo-Fi, Indie
- Music Recommendation Generation: The system retrieves a list of recommended songs from a curated database or an external API (e.g., YouTube, Spotify). The recommendations are based on the identified emotion and corresponding genre.

II. RELATED WORKS

This section contrasts the models that have been employed by current emotion-based music recommendation systems to accomplish their goals. By investigating various deep learning strategies, sentiment analysis techniques, and feature extraction methodologies, numerous research publications have advanced this topic. The effectiveness of various models has been better understood thanks to these studies, which have also helped in selecting the most suitable approach for our system [5].

A. Deep Learning for Emotion Recognition

S. Srinivasan et al. used deep convolutional neural networks to detect emotions. They developed two CNN models and combined them with pretrained architectures such as VGG19 and Xception to classify facial emotions [6]. Their models were trained on three distinct facial expression datasets, using ReLU activation functions for improved feature extraction. The study demonstrated that CNNs perform well in real-time applications when trained on extensive emotion datasets.

Premjith Ba et al. explored text-based sentiment analysis for music recommendation, testing CNN, LSTM, CNN-LSTM, BiLSTM, and CNN-BiLSTM models [7]. Their results showed that CNN-BiLSTM achieved the highest classification accuracy (83.21%), proving that CNNs are effective in extracting spatial features, while LSTMs and BiLSTMs are better suited for sequential data analysis, such as Chatbot interactions and user reviews.

Another study by J. James Anto Arnold et al. proposed a music recommendation system based on facial expressions. Their model processed webcam video recordings by extracting frames and applying the Facial Action Coding System (FACS) to categorize emotions into Happy, Angry, Surprise, and Sad [8]. This approach demonstrated the potential for real-time emotion-based music recommendations.

B. Music Recommendation Systems

Shakirova et al. investigated collaborative filtering methods for music recommendation and compared them with AI-powered models. Their research found that deep learning-based recommendation systems (CNNs and LSTMs) outperform conventional collaborative filtering algorithms, offering more personalized and accurate suggestions [9].

Chiang et al. introduced a music emotion classification system using KBCS, NWFE, and SVM for feature extraction and classification. Their study analysed 35 key auditory characteristics, such as rhythm, pitch, and timbre, achieving high classification accuracies of 86.94% and 92.33% for Happy, Tense, Sad, and Tranquil emotions [10]. This highlights the importance of feature selection in emotion-aware music recommendations.

To improve recommendation algorithms on streaming platforms like YouTube, Netflix, and Amazon, Fessahye et al. developed an enhanced T-RECSYS model, trained on the Spotify RecSys Challenge dataset. Their model achieved 88% precision by integrating deep learning algorithms with collaborative filtering techniques, proving that AI-based recommendation systems are highly scalable across digital platforms [11].

C. Facial Emotion Recognition for Music Recommendation

Zhang et al. enhanced music recommendations by integrating deep learning-based facial recognition. Their CNN-based model, trained on the LFW dataset, achieved an 88.56% facial expression recognition rate, validating the effectiveness of facial emotion detection in music selection [12].

Kim et al. designed a personalized recommendation system using K-Means clustering on a self-collected dataset.

Their approach classified music based on user preferences, achieving 74% recommendation accuracy by clustering songs into rock (34%) and classical (40%) genres [13]. This highlights the role of unsupervised learning in music recommendation.

Ayata et al. explored physiological signal-based music recommendations, using GSR and PPG signals from the Multimodal Deep Emotion Dataset. Their model achieved 71.53% (GSR) and 70.76% (PPG) accuracy, proving that biological signals can effectively predict emotional states for personalized music selection [14].

D. Limitations of Existing Works

While the aforementioned studies present promising advancements in emotion-based music recommendation systems, they still have certain limitations that hinder their real-world effectiveness.

One major limitation is the dependence on pre-recorded data. Many existing models rely on static datasets rather than real-time facial emotion recognition, which significantly reduces dynamic user interaction. Since emotions are highly context-dependent and can change rapidly, models that fail to adapt to real-time variations may not provide an engaging and personalized music recommendation experience.

Another drawback is the limited number of emotion categories used in classification. Some studies only consider four or five primary emotions, which can oversimplify the complexity of human emotions. In reality, emotions are nuanced and can exist in multiple intensities, such as mild sadness versus deep sorrow. A restricted classification system may lead to less accurate music recommendations that do not fully capture the user's mood.

Moreover, many sentiment-based approaches suffer from a lack of contextual awareness. Several studies rely solely on textual reviews or physiological signals like heart rate and skin conductance to determine mood. While these factors can provide some insights, they may not be as accurate and expressive as facial expression analysis, which directly reflects a user’s emotions.

Dataset imbalance is another common issue affecting model performance. Many studies use small or imbalanced datasets, where some emotions have significantly fewer samples than others. This imbalance often leads to misclassification, particularly for neutral or complex emotions. As a result, models may favor certain emotions over others, reducing the overall accuracy of emotion-based recommendations.

Lastly, generalization challenges remain a critical concern. Many deep learning models achieve high accuracy on controlled datasets, but they struggle with real-world variations such as different lighting conditions, facial angles, occlusions, and background noise. These variations can negatively impact model performance, making it less reliable when deployed in real-time applications.

Addressing these limitations is crucial for developing more robust, adaptable, and user-centric emotion-based music recommendation systems.

E. How Our Approach Differs

Our proposed model effectively addresses these limitations by integrating real-time emotion detection, ensuring that users receive dynamic music recommendations based on their live facial expressions rather than relying on pre-recorded images. This enhances user interaction and personalization, making the system more responsive to emotional changes.

To improve accuracy, we employ deep feature extraction using ResNet50V2, a more advanced architecture compared to VGG19 or shallow CNNs used in prior studies. By leveraging transfer learning, our model can extract richer facial features and enhance classification performance, even with limited training data.

Additionally, our system expands emotion classification by detecting a wider range of emotions, including Happy, Sad, Angry, Neutral, Fear, and Surprise. This allows for a more context-aware recommendation process, as opposed to models that classify emotions into only a few basic categories.

Unlike traditional recommendation systems that rely purely on collaborative filtering or content-based filtering, our model introduces a hybrid recommendation strategy that directly maps facial expressions to predefined music genres. This approach ensures that song recommendations align more closely with the user's emotional state, rather than just historical listening patterns.

Furthermore, our model is designed for scalability and real-time deployment, making it lightweight and optimized for mobile and web applications. This allows seamless integration into various platforms, enhancing accessibility and usability.

By addressing these challenges, our model provides a highly personalized and interactive music recommendation experience, setting it apart from existing systems.

III. BACKGROUND DETAILS

A. Convolutional Neural Networks

In their study, Shaha et al. said that CNN is well-known for its ability to extract information and features. CNN is a deep learning-based neural network architecture that has many practical uses in data visualization and picture interpretation with the aid of artificial intelligence. An improved type of artificial neural network, CNN offers more precise visual characteristics for improved classification [4]. According to CNN, each incoming image is handled as a matrix. The necessary data is then recovered from the resultant matrix (output picture), which is created by performing mathematical operations over the several matrices (input image).

A convolutional neural network (CNN) with five layers—the input layer, convolutional layer, pooling layer, fully connected layer, and output layer—is shown in Figure 1.

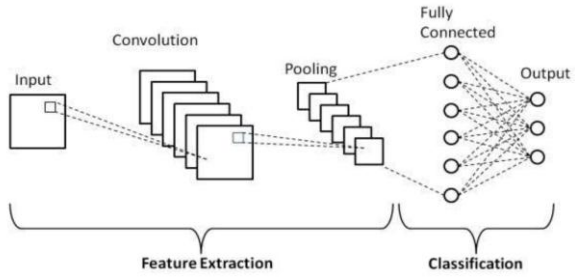

Figure. 1. CNN Architecture

1) Input layer: The input layer converts user-supplied pictures into matrices and forwards the matrix that results to the convolutional layer.

2) Convolutional Layer: The layer that appears following CNN's input layer is called the convolutional layer. In order to extract features from the supplied input images, mathematical operations are carried out in this initial layer. The convolution process is carried out over the input matrix, which consists of the filter (the AxB matrix, often referred to as the kernel, which is used to extract attributes) and matrix multiplication of the input picture. A feature map, which is created by multiplying the input by the kernel, contains the outcome. Two key terms that are used once the feature map is acquired are stride and padding.

3) Padding: In order to maintain spatial sizes during the convolution process, padding is the inclusion of extra pixels around the input image or feature map. It is essential to the construction and operation of convolutional neural networks and helps to reduce information loss at the edges.

4) Stride: As it interprets the picture, the computer needs to figure out how far the filter will have to travel across the picture in stride. It moves horizontally across the photo from the top left corner to the bottom right corner. Here, stride makes sure that the filter knows how many pixels (squares) it must skip in order to interpret the image. The number of features that the filter learns depends on the stride size. Due to extensive data extraction, more features are learned, as evidenced by the stride's lower size. Conversely, a larger stride size results in less data extraction and therefore fewer characteristics being learned.

$$\text{Size of Feature Map} = (B-A+1) \times (B-A+1)$$

Where,

B = The input matrix's row count
A = The input matrix's column count
$B \times B$ = The input matrix's size
$A \times A$ = the kernel/filter's size

5) Pooling Layer: In order to simplify mathematical calculations and save computation expenses, the pooling layer shrinks the matrix. It also helps to improve ConvNet's stability. In order to speed up calculation, the pooling layer helps to reduce the training parameters. In general, pooling operations can be divided into three categories: maximum, minimum, and average pooling. To summarize a chosen area of the image, max pooling selects the maximum feature values in that area. Min pooling summarizes the chosen area of the image by choosing the minimal feature values in that area. A region's average value is used in average pooling to calculate the sum of its characteristics.

6) Fully Connected Layer: This layer multiplies the weight matrix and input together and adds a bias vector. This layer does the job of linking neurons from the previous layer and fully connected layer. The calculation formula of a fully connected layer is indicated in Equation 1.

$$y_{jk}(x) = f\left(\sum_{i=1}^{nH} w_{jk} x_i + w_{j0}\right) \quad (1)$$

In this case,

W = Weight matrix
W_0 = Bias vector
X = Input matrix
Y = Output matrix

7) Output Layer: It is the final layer of the Convolutional Neural Network, whose task is to predict the final result by

projecting the features learned from input images. The output from this layer classifies the emotion of a user, which may be any one of the emotions for which the machine is trained.

B. ResNet50V2

Residual Network 50 Version 2, or ResNet50V2, is a deep convolutional neural network (CNN) architecture that uses residual connections to improve feature extraction capabilities. ResNet designs were developed by He et al. to address the issue of vanishing gradients in deep networks, which frequently impede efficient training. With the help of a pre-activation residual block, ResNet50V2 enhances gradient flow and makes learning in very deep networks more effective than ResNet50.

ResNet50V2 maintains good accuracy while lowering computing complexity with its bottleneck architecture. It is especially useful for computer vision tasks like facial recognition and image classification because of its enhanced design.

The input layer, convolutional layers, residual blocks, global average pooling, and the output layer are the five primary parts of the ResNet50V2 architecture.

1) Input Layer: The input layer processes user-provided images and converts them into numerical matrices. These matrices are passed to the initial convolutional layers for feature extraction.

2) Convolutional Layers**:** Similar to standard CNNs, ResNet50V2 applies convolutional filters over the input matrix to extract important spatial and texture-based features. However, it enhances this process with pre-activation residual blocks, ensuring stable gradient propagation.

3) Residual Blocks**:** The core innovation in ResNet50V2 is the use of residual connections, allowing feature information to bypass layers and preventing degradation in performance. Each residual block consists of:

- 1×1 Convolution (dimensionality reduction)
- 3×3 Convolution (feature extraction)
- 1×1 Convolution (dimensionality restoration)
- Skip connection (identity mapping of the original input to the output of the residual block)

The advantage of residual blocks is that they help deep networks learn efficiently by allowing gradients to flow directly through the network during backpropagation.

4) Batch Normalization and Activation**:** Unlike ResNet50, where activation (ReLU) and batch normalization are applied after the convolutional layers, ResNet50V2 applies batch normalization and ReLU before the convolution operation. This modification leads to improved training stability and convergence.

5) Global Average Pooling (GAP) and Fully Connected Layer**:** Instead of using fully connected layers with high parameters, ResNet50V2 applies Global Average Pooling (GAP) to reduce overfitting and computational complexity. The GAP layer summarizes spatial feature maps, creating a low-dimensional feature vector, which is then passed through a fully connected layer for classification.

6) Output Layer: The output layer in ResNet50V2 classifies the input image into one of the predefined categories. In the context of facial emotion recognition, this layer maps facial expressions to specific emotion classes such as happy, sad, angry, surprised, and neutral. The softmax function is typically used to compute the final probability distribution across these categories.

The concept of residual learning in ResNet50V2 is mathematically represented as follows:

$$y=F(x,Wi)+x$$

where:

- x is the input to the residual block,
- Wi represents the convolutional layer weights,
- F(x,Wi) is the learned transformation (convolution, batch normalization, and activation),
- y is the final output of the residual block after summation.

This formulation enables efficient gradient flow, ensuring deeper networks do not suffer from the vanishing gradient problem.

IV. PROPOSED WORKS

The approach to creating the emotion-based song recommendation system commences with the collection and pre-processing of data. The accompanying flowchart (Figure 2) illustrates the comprehensive workflow of the proposed project. Essential libraries were imported to manage datasets, conduct visualizations, and execute deep learning models. The dataset, which includes images labelled with emotions, was loaded in

conjunction with a distinct music dataset that correlates moods with specific songs. To maintain data integrity, the distributions of classes were visualized, and sample images were presented. Subsequently, the dataset was divided into training and testing subsets, typically following an 80:20 ratio [14]. Pre-processing procedures, such as resizing images to 224x224 for ResNet50V2, normalizing pixel values (scaling them between 0 and 1), and applying data augmentation techniques (including rotation, flipping, and zooming), were implemented to enhance generalization.

In the course of model development, two distinct deep learning architectures were employed: a Convolutional Neural Network (CNN) and ResNet50V2. The CNN was designed with convolutional layers dedicated to feature extraction, pooling layers aimed at reducing dimensionality, and fully connected layers responsible for classification tasks. The model was compiled utilizing a suitable optimizer, such as Adam, along with a categorical cross-entropy loss function. To mitigate the risk of overfitting, call backs like Early Stopping and ReduceLROnPlateau were incorporated, and the training process was conducted over 50 epochs [15]. In a similar vein, the ResNet50V2 model, which is pre-trained, underwent fine-tuning by freezing its initial layers and training solely the upper layers, adhering to the same pre-processing, compilation, and training protocols as the CNN.

In the context of model evaluation and performance assessment, both the test loss and accuracy were analysed for the CNN and ResNet50V2 architectures. To gain deeper insights into their performance, graphical representations were created to illustrate the relationship between test loss and epochs, as well as test accuracy and epochs. Additionally, a Confusion Matrix was constructed to evaluate the effectiveness of the models in classifying various emotions.

After classifying emotions, the system established a mechanism for recommending songs based on these emotional insights. The music dataset underwent processing, allowing for the predicted emotions to be aligned with songs that correspond to specific moods [16]. Subsequently, the five most popular songs, arranged in descending order of their popularity, were presented in relation to the identified emotion.

To facilitate real-time inference, the system incorporated a feature for immediate predictions. New images were uploaded, pre-processed, and formatted for input into the model, utilizing both CNN and ResNet50V2 to ascertain emotional states.

The resulting labels and their corresponding confidence scores were presented. Ultimately, both models were preserved for subsequent use and deployment, guaranteeing their readiness for real-time applications. This organized methodology guarantees an effective emotion recognition system that is seamlessly integrated with a customized music recommendation feature [17].

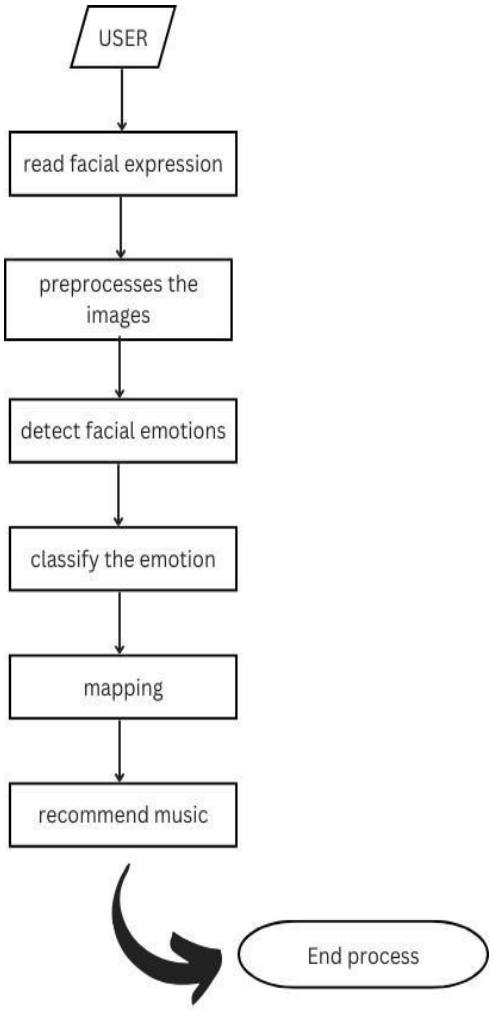

Figure. 2. Detailed Working of the proposed model

V. IMPLEMENTATION

A. Dataset

FER2013 is a popular standard collection of facial expression recognition that was used to train the model. 7 emotions are included in the dataset, as indicated in Table I: anger, disgust, fear, happiness, neutrality, sadness, and surprise. Thousands of 48x48 pixel grayscale photos taken from actual situations are included in each emotion category. The model's capacity for generalization and detection rate are enhanced by learning it using the FER (2013) dataset, which features an abundance of expressions on faces gathered in randomized settings [18]. The program can efficiently categorize sensations in realistic-time by employing the data at hand, which makes it resilient and flexible for emotion-based applications in the real world.

TABLE I
EMOTIONS DETAILS

Emotions	Training Images	Testing Images
Happy	7215	1774
Sad	4830	1247
Neutral	4965	1233
Angry	3995	958
Fear	4097	1024
Disgust	436	111
Surprise	3171	831

B. Steps for Data Collection

Step 1: Importing Required Libraries

To efficiently handle data and visualization, essential libraries such as NumPy, Pandas, Matplotlib, and Seaborn are imported. NumPy and Pandas facilitate numerical computations and data manipulation, while Matplotlib and Seaborn enhance data visualization. For image processing tasks, OpenCV (cv2) and ImageDataGenerator are utilized, enabling real-time data augmentation and preprocessing. TensorFlow and Keras serve as the backbone for implementing deep learning models, providing a robust framework for training and evaluation. Additionally, optimization techniques like EarlyStopping, ModelCheckpoint, and ReduceLROnPlateau are incorporated to prevent overfitting, save the best model, and dynamically adjust the learning rate, ensuring improved model performance.

Step 2: Loading the Dataset

The emotion-based picture dataset, which contains tagged images for different emotions, is loaded initially. The emotion recognition model will be trained using the dataset, which consists of facial expressions categorized into several emotion classes. In order to allow mood-based song suggestion, the music dataset is also loaded, mapping different moods to corresponding songs. After the datasets have been loaded, a summary of their makeup is shown, including the number of emotion classes, sample size for each class, and overall dataset distribution. By doing this, any uncertainty about the data before pre-processing and model training is eliminated.

Step 3: Visualizing Dataset

A bar graph is created to display the quantity of photos in each class of emotion in order to visualize the dataset's distribution. This makes it easier to identify class disparities and ensure that training data is appropriately represented. To verify that the annotations are accurate, a few example photographs are also randomly selected from the collection and displayed with the appropriate labels. This procedure guarantees that the dataset is appropriately organized and prepared for the emotion recognition model to be trained successfully.

Step 4: Splitting the Dataset

For effective model training and testing, the dataset is split into training and testing sets in an 80:20 ratio (or a preferred ratio). While the testing set is used to evaluate the deep learning model's performance on fresh data, the training set is used to train the model. An apparent picture of the data distribution between training and testing sets is provided by concatenating the number of samples in each set after splitting and presenting the results in a well-organized table manner. A well-prepared and balanced dataset for model building is guaranteed by this procedure.

Step 5: Data Pre-processing

Pre-processing the training data involves resizing all photos to a standard size (e.g., 224x224 for ResNet50V2) to ensure consistency across the dataset. To help the model converge, pixel values are normalized by scaling them between 0 and 1. Additionally, data augmentation techniques like flipping, zooming, and rotation are employed to fictitiously expand the dataset's size and boost the model's capacity to generalize to data that hasn't been seen before. Images are downsized to a consistent fixed size with the training data for pre-processed test data in order to make them compatible with the model. Normalizing pixel values improves forecast accuracy and consistency. For the best model performance and efficient training and testing, all of these pre-processing operations are important.

Step 6: Loading Processed Data

Images are first converted into NumPy arrays to enable TensorFlow/Keras compatibility and efficient computation, thereby preparing the dataset for deep learning model training. To make sure they are prepared for model training and testing, both the training and test datasets are loaded into memory. Finally, to make sure the dataset has been processed correctly and has the desired dimensions for picture inputs and labels, the morphologies of the training and testing data are examined. This is to guarantee that the data is appropriately organized prior to being incorporated into the deep learning model.

Step 7: Model Training Data Collection

The pre-processed images and labels are retained in variables X_train and y_train, the input features and target labels used for training. A validation split of the training data is set up to track model performance during training, which can prevent overfitting and facilitate generalization.

Following data preparation, a dataset summary is printed out with the class distribution and shape of X_train and y_train as well as the validation set. This is done to ensure that the dataset is properly structured and ready for training deep learning models.

Step 8: Training CNN and ResNet50V2 Models

In constructing the CNN model, there are five implemented layers, i.e., three convolutional blocks and two fully connected layers:

- ❖ CNN1 - First Convolutional Block: It has a convolutional layer, activation function (ReLU), and max-pooling in order to learn low-level features.
- ❖ CNN2 - Second Convolutional Block: Again, extracts complex features by employing a different set of convolutional and pooling layers.
- ❖ CNN3 - Third Convolutional Block: Expands the feature extraction procedure to make it deeper so the model can extract sophisticated patterns.
- ❖ Fully Connected Layer: Flattens the extracted features and feeds them through dense layers to classify.
- ❖ Output Layer: Employing a softmax activation function to classify images into various emotion categories.

The model is then compiled with a suitable loss function (e.g., categorical cross-entropy), optimizer (e.g., Adam), and evaluation metrics (e.g., accuracy). Call-backs like EarlyStopping, ModelCheckpoint, and ReduceLROnPlateau are used to avoid overfitting and improve model performance. The CNN model is trained for 50 epochs and the training history is saved in order to see how the performance of the models varies. Lastly, ResNet50V2 is utilized employing transfer learning by unfreezing the topmost layers of the pre-trained model and retraining it on the emotion dataset. This method applies the strength of deep feature extraction but enables specific tuning for the task of emotion classification.

Step 9: Model Evaluation & Performance Analysis

Both the CNN model and the ResNet50V2 are trained and evaluated using the test data in order to gauge the models' performance. To determine whether the models can generalize to the unseen data, crucial metrics like accuracy and loss are computed. The following graphs are created in order to visually analyse the performance:

- Test Loss vs. Epoch: Shows how the loss changes over the epochs, pointing out trends of under fitting or overfitting.
- Test correctness vs. Epoch: Displays a trend in model correctness that reflects stability and convergence.

In addition, the categorization performance is examined by calculating and displaying a Confusion Matrix. Class-wise accuracy, misclassifications, and prediction bias are all disclosed by the confusion matrix. It guarantees a thorough examination of the models before they are used..

Step 10: Prediction & Mapping Indices to Emotion Classes

The following actions are taken in order to apply the models on fresh, unseen images:

- The new photos should be loaded, resized to 224x224, and normalized in accordance with the training data format.
- Make Forecasts: Both the CNN model and ResNet50V2 predict the emotions of the new photos.
- Display Results: To allow for a direct comparison of model performance on unseen data, the predicted labels are displayed alongside the actual class labels.

This procedure helps assess how well the models generalize in practical situations.

Step 11: Mapping Emotion Predictions to Music Dataset

To combine the music recommendation system with the emotion recognition model, the dataset for mood-based song recommendations is loaded and pre-processed. The dataset consists of song names, artists, and popularity scores corresponding to various emotions. Depending on the predicted emotion, the system pulls the top five songs based on popularity in descending order, guaranteeing the most appropriate recommendations. Lastly, the suggested songs for the identified mood appear, enabling users to benefit from customized music recommendations compatible with their mood.

Step 12: Saving the Model for Future Use

The trained ResNet50V2 and CNN model are saved for deployment so that they can be utilized for real-time inference. The models are saved in the HDF5 (.h5) format or the TensorFlow SavedModel format, maintaining the learned weights and architecture. The saved models are then loaded to check their integrity to ensure they work properly for real-time emotion detection and recommendation of music. This is an important step for successful deployment and real-world application of the system.

C. Proposed Steps

1. Import libraries required

NumPy, Pandas, Matplotlib, and Seaborn are essential Python libraries for data manipulation, analysis, and visualization. NumPy provides support for large, multi-dimensional arrays and matrices, while Pandas simplifies data handling with its powerful DataFrame structure. Matplotlib and Seaborn enable comprehensive data visualization, making it easier to interpret trends and patterns. OpenCV is widely used for image processing and computer vision tasks, allowing efficient manipulation and analysis of images. TensorFlow and Keras facilitate deep learning model development, offering robust tools for neural network training and deployment. Additionally, ImageDataGenerator aids in data augmentation to enhance model performance, while EarlyStopping and ModelCheckpoint optimize training by preventing overfitting and saving the best-performing model, respectively.

2. Data Pre-processing and Visualization

To begin with, we load the dataset and print the train-test split statistics to understand the distribution of data. Next, we visualize the class distribution of emotions to check for any imbalances in the dataset. Finally, we display sample images from the dataset to gain insights into the data quality and structure before proceeding with model training.

3. Prepare Data for Model Training

To prepare the dataset for model training, we first perform image pre-processing, including resizing the images to a uniform shape and normalizing pixel values for better model performance. Next, we convert the images into arrays and apply label encoding to transform categorical labels into numerical values. The dataset is then split into training and testing sets to evaluate model performance effectively. Finally, we load and preprocess the images to ensure they are in the correct format and ready for input into the deep learning model.

4. Build and Train a CNN Model

To build the Convolutional Neural Network (CNN) for image classification, we start by defining the architecture. The model consists of three convolutional blocks:

- **Conv Block 1:** The first convolutional layer extracts low-level features such as edges and textures.
- **Conv Block 2:** The second convolutional layer captures more complex patterns and spatial hierarchies.
- **Conv Block 3:** The third convolutional layer further refines feature extraction.
- **Fully Connected Layers:** These layers process the extracted features and produce the final classification output.

Next, we compile the model by selecting an appropriate optimizer, loss function, and evaluation metrics while setting key hyper parameters such as the learning rate and batch size. To enhance model performance and prevent overfitting, we implement callback functions like EarlyStopping (to halt training when no improvement is observed) and ModelCheckpoint (to save the best model). The CNN model is then trained for 50 epochs using the pre-processed dataset. Finally, we evaluate the model's performance on the test set by analysing test loss and test accuracy to assess its generalization capability.

5. Visualize Model Performance

To analyse the model's performance, we start by plotting loss vs. epochs and accuracy vs. epochs to visualize the training and validation trends over time. These plots help identify issues like overfitting or under fitting by showing how the loss decreases and accuracy improves during training. Next, we generate a confusion matrix to evaluate the model's emotion classification performance. The confusion matrix provides detailed insights into the model's predictions, highlighting the number of correct and incorrect classifications for each emotion category. This allows us to assess class-wise performance and identify any misclassifications that may need further improvements in the model.

6. Implement ResNet50V2 for Transfer Learning

To enhance emotion classification performance, we utilize ResNet50V2 as a feature extractor, leveraging its pre-trained weights to extract deep features from images. Initially, all layers except the last 50 layers are frozen to retain learned features while allowing fine-tuning on the target dataset. Next, we modify the ResNet50V2 architecture by adding a custom classification head tailored for emotion recognition. This involves adding fully connected layers and an activation function suited for multi-class classification. After modifying the architecture, we fine-tune the model by training it for 30 epochs, adjusting only the unfrozen layers to learn task-specific features. Finally, we evaluate the model's test accuracy to measure its effectiveness in emotion classification, comparing it to the CNN model to assess performance improvements.

7. Make Predictions on New Images

For inference, we start by loading the processed images, ensuring they are pre-processed in the same way as the training data (resized, normalized, and converted into arrays). Next, we perform predictions using the trained CNN model, feeding the images into the model to obtain output probabilities for different emotion classes. Finally, we predict new emotion-based classes by selecting the class with the highest probability for each image. This step allows the model to classify unseen images into corresponding emotions, enabling real-world application of emotion recognition.

8. Music Recommendation Based on Predicted Emotion

Once the model detects an emotion, we proceed to display the top 5 song recommendations based on the predicted emotion. These recommendations are selected from a curated playlist that aligns with the detected mood, ensuring a personalized music experience. Next, we fetch relevant music dataset previews, which may include song titles, artists, and album covers, providing users with a visual and textual overview of the recommended tracks. This enhances the user experience by making the emotion-based music recommendation system more interactive and engaging.

9. Model Saving

To preserve the trained model for future use, we **save it in** .h5 **format** using TensorFlow/Keras. This allows the model to be reloaded later for inference or further fine-tuning without retraining from scratch. Saving the model ensures that all learned weights, architectures, and configurations are stored

efficiently, making it easy to deploy the emotion-based music recommendation system whenever needed.

VI. RESULT ANALYSIS

A. Training and Testing Accuracy & Loss

The accuracy and loss metrics for both training and testing phases offer a comprehensive understanding of the model's performance and its learning capabilities. Throughout the training process, there was a noticeable increase in accuracy alongside a steady decline in loss, suggesting that the model successfully grasped the dataset's characteristics.

In the testing phase, the model's proficiency in generalizing to new, unseen data was assessed, with high test accuracy and low test loss serving as indicators of its reliability. The visual representations of loss versus epochs and accuracy versus epochs illustrated a consistent learning trajectory, which minimized the risk of overfitting [19]. Additionally, the implementation of call-backs and regularization strategies contributed to enhancing the model's performance, rendering it particularly effective for emotion-based music recommendation tasks.

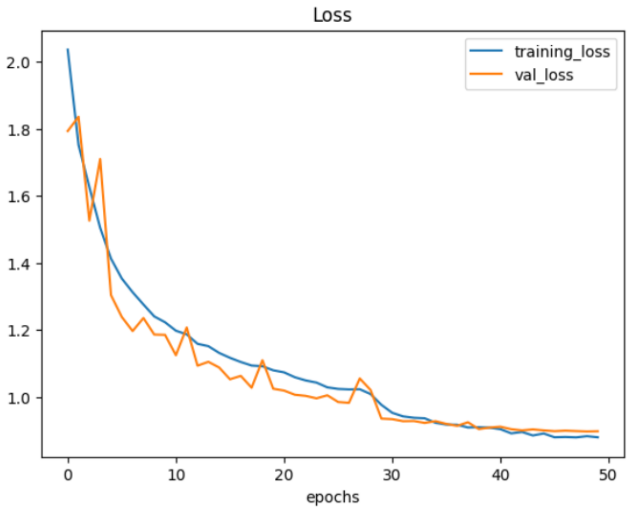

Figure. 3. CNN Model Loss

Convolutional neural networks (CNNs) and their model loss are depicted in Figure 3, which also compares the model's training and validation losses. The figure's X-axis represents the epochs, while the Y-axis shows the values of the training loss. Successful learning and little overfitting are indicated by the CNN model's loss curves, which show a steady decreasing trend.

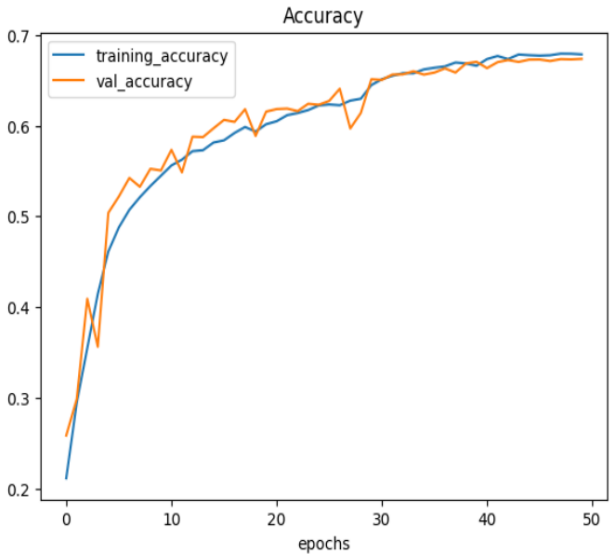

Figure. 4. CNN Model Accuracy

By comparing the training and validation accuracy metrics, Figure 4 shows the convolutional neural network's (CNN) accuracy. The figure's Y-axis shows the training accuracy, while the X-axis represents the epochs. The CNN model's accuracy trajectories exhibit a steady rising trend over the course of the epochs, suggesting that it is capable of learning and generalizing. The model's reliability in identifying emotions for music recommendation is confirmed by the close proximity of the training and validation accuracy curves, which suggests a low degree of overfitting.

Figure 5 illustrates the loss analysis for the ResNet50V2 model, detailing the training and validation loss metrics across various epochs. The horizontal axis represents the epochs, while the vertical axis indicates the loss values. At the outset, there is a significant decline in loss, reflecting a phase of rapid learning, which is subsequently followed by a more gradual decrease. The similarity between the training and validation loss indicates that the model is not experiencing overfitting, thereby demonstrating a well-balanced learning process that contributes to its reliability in predicting facial emotions for music recommendation.

Figure 6 presents the accuracy metrics of the ResNet50V2 model, showcasing a comparison between training and validation accuracy across various epochs. The horizontal axis denotes the number of epochs, while the vertical axis indicates the accuracy levels for both training and validation datasets. The accuracy trajectories reveal a consistent learning pattern, with training accuracy showing a steady upward trend.

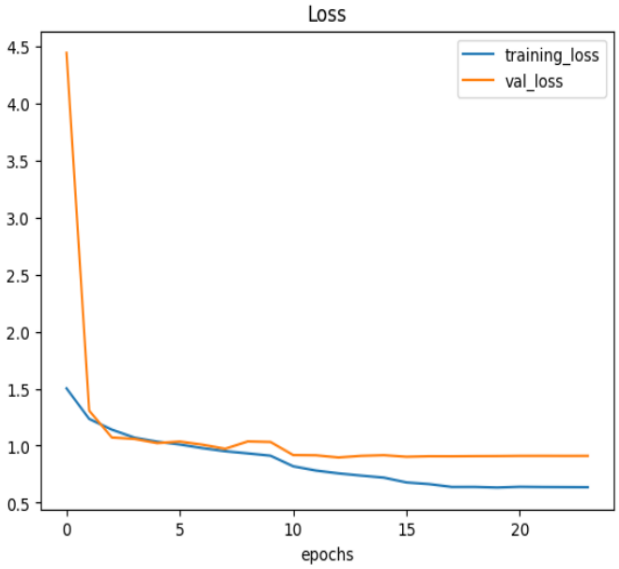

Figure. 5. ResNet50v2 Model Loss

Meanwhile, the validation accuracy reaches a plateau after a certain threshold, indicating the model's proficiency in generalizing to new, unseen data. The narrow margin between training and validation accuracy points to minimal overfitting, underscoring the model's capability in effectively classifying emotions within the music recommendation framework.

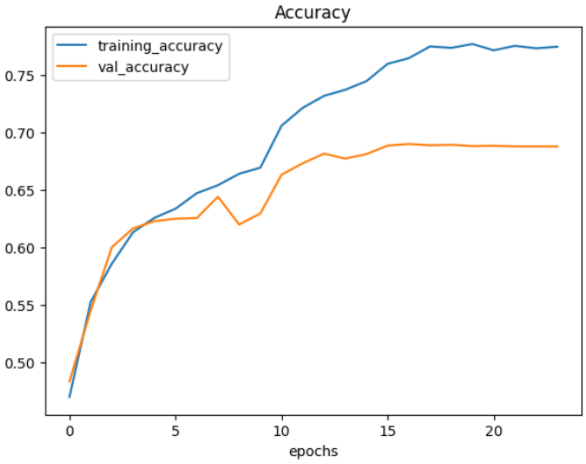

Figure. 6. ResNet50v2 Model Accuracy

B. Confusion Matrix

Figure 7 displays the confusion matrix associated with the CNN model, which demonstrates its classification efficacy across various emotion categories. The horizontal axis denotes the predicted labels, while the vertical axis reflects the actual labels. Values along the diagonal represent instances that have been accurately classified for each category, whereas the off-diagonal entries indicate instances of misclassification.

A greater concentration of values along the diagonal suggests that the model performs well in identifying specific emotions, while the existence of off-diagonal values points to potential areas for enhancement. The color gradient within the matrix illustrates the distribution of predictions, with lighter hues indicating higher frequencies. This confusion matrix offers valuable insights into the model's capabilities and highlights the challenges it faces in terms of misclassification.

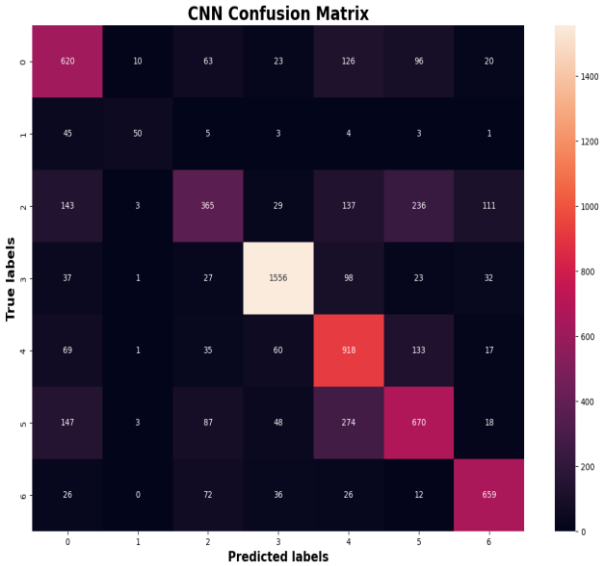

Figure. 7. CNN Confusion Matrix

Figure 8 illustrates the confusion matrix associated with the ResNet50V2 model, highlighting its effectiveness in classifying various emotion categories. The horizontal axis denotes the predicted labels, while the vertical axis reflects the actual labels. Correct classifications are represented by the diagonal elements, whereas the off-diagonal elements indicate instances of misclassification.

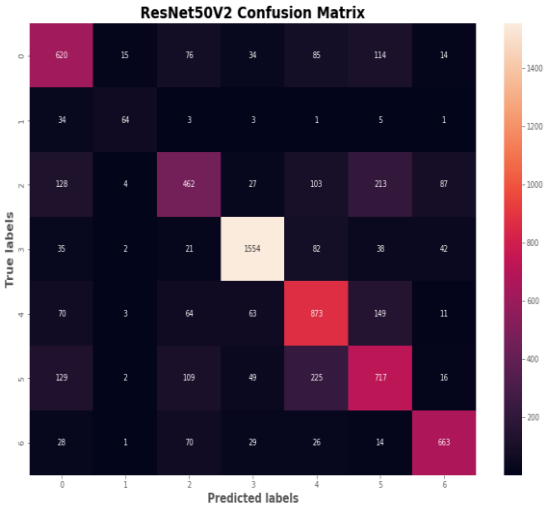

Figure. 8. ResNet50V2 Confusion Matrix

The colour intensity within the matrix emphasizes the frequency of predictions, with lighter hues signifying greater occurrences. This analysis indicates that while the model excels

in recognizing certain emotions, it still experiences some misclassifications. Such a visualization offers critical insights into the capabilities and limitations of the ResNet50V2 model in the context of emotion classification for music recommendation.

C. Music Prediction

Figure 9 illustrates a music recommendation system that utilizes facial emotion detection technology. In the image, the identified face is marked with a green bounding box, and the system has classified the emotion as ”Angry.” Above the image, a table presents five song recommendations that are linked to a calm mood, indicating that the system’s objective is to create an emotional equilibrium by suggesting music that counteracts the observed emotional state.

	name	artist	mood	popularity
0	Lost	Annelie	Calm	64
1	Curiosity	Beau Projet	Calm	60
2	Escaping Time	Benjamin Martins	Calm	60
3	Just Look at You	369	Calm	59
4	Vague	Amaranth Cove	Calm	59

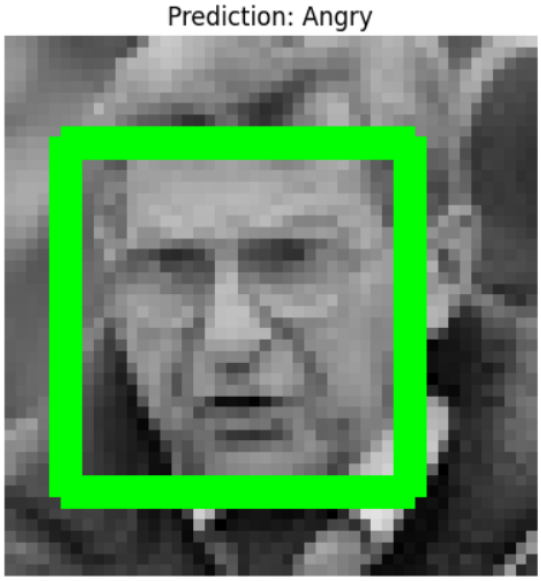

Figure. 9. ResNet50V2 Music Recommendation Prediction

Figure 10 illustrates a comparable configuration in which the system has identified the emotion as ”Happy.” Once more, the recognized face is highlighted with a green bounding box. The preceding table enumerates five songs that correspond to a joyful mood, suggesting that the system proposes music that is in harmony with the user’s present emotional condition.

	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

Figure. 10. CNN Music Recommendation Prediction

VII. COMPARATIVE ANALYSIS

In this segment, we analyze the efficacy of Convolutional Neural Network (CNN) and ResNet50V2 in the context of facial emotion recognition, focusing on their predictive outcomes. The findings indicate that ResNet50V2 outperformed CNN in terms of accuracy when identifying facial expressions. A significant pattern of misclassification was noted, particularly with neutral faces frequently being incorrectly identified as sad or fearful, which suggests that the model struggles with differentiating between nuanced emotional cues [19]. Furthermore, instances of surprise were occasionally confused with happiness, likely due to the similarities in facial features associated with these emotions. Conversely, both models exhibited commendable performance in accurately recognizing anger and fear, demonstrating their capability in detecting more pronounced emotional expressions [20].

A more in-depth examination indicates that CNN encountered greater difficulties in accurately identifying neutral, sad, and fearful emotions, resulting in a higher rate of misclassifications. In contrast, ResNet50V2 demonstrated a lower error rate, which implies superior feature extraction and

generalization capabilities attributed to its more complex architecture. Both models excelled in detecting happy and angry expressions, as these emotions are characterized by clearly defined facial features. Ultimately, ResNet50V2 surpasses CNN in performance, owing to its deeper architecture and the advantages of pre-trained weights that facilitate enhanced feature extraction and classification [21]. Conversely, CNN's elevated error rate in recognizing nuanced emotions highlights the necessity for improved training methodologies or the implementation of data augmentation strategies.

This comparative study emphasizes the benefits of employing ResNet50V2 in contrast to a conventional CNN, thereby underscoring its effectiveness for facial emotion recognition. Additionally, a comprehensive confusion matrix or performance table could further substantiate these conclusions [22].

VIII. CONCLUSION

The research effectively executed and contrasted a Convolutional Neural Network (CNN) with ResNet50V2 for the purpose of recommending music based on facial emotion recognition. The dataset underwent preprocessing, and both models were subjected to rigorous training and testing protocols, which encompassed data augmentation, model compilation, and the use of callbacks to improve performance and mitigate overfitting [23]. The evaluation of the models was carried out through the analysis of test accuracy, loss graphs, and confusion matrices to evaluate their classification effectiveness. The CNN model, while relatively straightforward, yielded encouraging outcomes, showcasing its proficiency in effectively identifying facial features [24]. In contrast, ResNet50V2, which is a more complex and pre-trained architecture, displayed superior performance attributed to its enhanced feature extraction abilities. Analyzing the predictions from both models revealed their respective advantages and limitations. A music recommendation system was developed by examining emotion predictions and correlating them with a dataset that includes various song attributes. This system effectively identified and suggested the five most suitable songs based on their mood and popularity, showcasing its practical utility. Additionally, the trained models were preserved for future use, facilitating smooth deployment in real-world scenarios [25]. This project underscores the efficacy of utilizing deep learning techniques for facial emotion recognition in the development of personalized music recommendation systems. Prospective improvements may involve refining the models to achieve greater accuracy, broadening the dataset to enhance generalization capabilities, and incorporating real-time emotion recognition to facilitate dynamic recommendations [26].

REFERENCES

[1] Hongli Zhang, Alireza Jolfaei, and Mamoun Alazab, "A Face Emotion Recognition Method Using Convolutional Neural Network and Image Edge Computing," IEEE Access, vol. 7, pp. 159081-159089, 2019.

[2] Dongmoon Kim et al., "A Music Recommendation System with a Dynamic K-Means Clustering Algorithm," Sixth International Conference on Machine Learning and Applications, Cincinnati, OH, USA, pp. 399403, 2007.

[3] Deger Ayata, Yusuf Yaslan, and Mustafa E. Kamasak, "Emotion Based Music Recommendation System Using Wearable Physiological Sensors," IEEE Transactions on Consumer Electronics, vol. 64, no. 2, pp. 196-203, 2018.

[4] Wei Chun Chiang, Jeen Shing Wang, and Yu Liang Hsu, "A Music Emotion Recognition Algorithm with Hierarchical SVM Based Classifiers," 2014 International Symposium on Computer, Consumer and Control, Taichung, Taiwan, pp. 1249-1252, 2014.

[5] M P, Sunil & ., Hariprasad S A. (2023). Facial Emotion Recognition using a Modified Deep Convolutional Neural Network Based on the Concatenation of XCEPTION and RESNET50 V2. International Journal of Electrical and Electronics Engineering Research. 10. 94-105. 10.14445/23488379/IJEEE-V10I6P110.

[6] Sriraj Katkuri, Mahitha Chegoor, Dr. K. C. Sreedhar, M. Sathyanarayana, 2023, Emotion Based Music Recommendation System, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 12, Issue 05 (May 2023)

[7] S. Madderi, S. Ponnaiyan, M. Subramanian, and K. Thulasingam, "A new mining and decoding framework to predict expression of opinion on social media emoji's using machine learning models," IAES International Journal of Artificial Intelligence, vol. 13, no. 4, pp. 5005–5012, Dec. 2024.

[8] N. K. E. and J. K., "Class based dynamic feature centric data deduplication scheme for efficient mitigation of side channel attack in cloud," J. Electr. Eng. Technol., vol. 19, no. 3, pp. 1933–1942, Mar. 2024.

[9] D. D. G., N. Kumar, J. M., and R. V., "Innovative brain tumor detection: Stacked random support vector-based hybrid gazelle coati algorithm," Biomed. Signal Process. Control, vol. 101, p. 107156, 2025.

[10] R. V., J. S. Manoharan, R. Hemalatha, and D. Saravanan, "Deep learning models for multiple face mask detection under a complex big data environment," Procedia Comput. Sci., vol. 215, pp. 706–712, 2022.

[11] J. Venkatesh, K. S. Kumari, V. Rekha, N. Geethanjali, S. K. Rajesh Kanna, and K. Sivakumar, "Transformer models; Capsule neural networks; Electric networks; Wavelet transforms; Error rates," Library of Progress-

Library Sci. Inf. Technol. Comput., vol. 44, no. 3, p. 12685, 2024.

[12] K.M. Aswin et al., ”HERS:Human Emotion Recognition System,” 2016 International Conference on Information Science, Kochi, India, pp. 176179, 2016.

[13] Shlok Gilda et al., ”Smart Music Player Integrating Facial Emotion Recognition and Music Mood Recommendation,” 2017 International Conference on Wireless Communications, Signal Processing and Networking, Chennai, India, pp. 154-158, 2017.

[14] Ashish Tripathi, Abhijat Mishra, Rajnesh Singh, Bhoopendra Dwivedy, Amit Kumar, Kuldeep Singh, ”Facial Emotion-Based Song Recommender System Using CNN,” International Journal of Engineering Trends and Technology, vol. 72, no. 6, pp. 315-327, 2024.

[15] R Prasanna, M Jenath, M Vinoth, J Joseph Ignatious, M S Maharajan, P Banu Priya, “Enhanced blood prothrombin time detection deploying flexible substrate UWB antenna from artifacts removed pure plasma through statistical multiple regression modelling” , Computers and Electrical Engineering, Volume 122, 2025, 109963, ISSN 0045-7906.

[16] Manali Shaha, and Meenakshi Pawar, ”Transfer Learning for Image Classification,” 2018 Second International Conference on Electronics, Communication and Aerospace Technology, Coimbatore, India, pp. 656660, 2018.

[17] K.S. Krupa et al., ”Emotion Aware Smart Music Recommender System Using Two Level CNN,” 2020 Third International Conference on Smart Systems and Inventive Technology, Tirunelveli, India, pp. 1322-1327, 2020.

[18] Jamdar, A., Abraham, J., Khanna, K., & Dubey, R. (2015). Emotion Analysis of Songs Based on Lyrical and Audio Features. International Journal of Artificial Intelligence & Applications, 6(3), 35–50.

[19] Pettijohn, T. F., Williams, G. M., & Carter, T. C. (2010). Music for the Seasons: Seasonal Music Preferences in College Students. Current Psychology, 29(4), 328–345.

[20] E. Shakirova, ”Collaborative filtering for music recommender system,” 2017 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), 2017, pp. 548-550,

[21] T Tulasi Sasidhar, Premjith B, Soman K P, ”Emotion Detection in Hinglish(Hindi+English) Code-Mixed Social Media Text”, Procedia Computer Science,Volume 171,2020, Pages 1346-1352,ISSN 1877-0509.

[22] Davis Moswedi, and Ritesh Ajoodha, ”Music Classification Using Fourier Transform and Support Vector Machines,” 2022 International Conference on Engineering and Emerging Technologies, Kuala Lumpur, Malaysia, pp. 1-4, 2022.

[23] E. Jing, Y. Liu, Y. Chai, S. Yu, L. Liu, Y. Jiang, and Y. Wang, "Emotion-Aware Personalized Music Recommendation with a Heterogeneity-Aware Deep Bayesian Network," 2024, arXiv:2406.14090. [Online]. Available: https://arxiv.org/abs/2406.14090

[24] X. Chang, X. Zhang, H. Zhang, and Y. Ran, "Music Emotion Prediction Using Recurrent Neural Networks," 2024, arXiv:2405.06747. [Online]. Available: https://arxiv.org/abs/2405.06747

[25] T. Babu, R. R. Nair, and G. A., "Emotion-Aware Music Recommendation System: Enhancing User Experience Through Real-Time Emotional Context," 2023, arXiv:2311.10796. [Online]. Available: https://arxiv.org/abs/2311.10796