https://doi.org/10.48047/AFJBS.6.12.2024.1086-1101

AFJBS African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

Mood Based Music Recommendation System Using Deep Learning

Dr Rambabu Kusuma
CSE Department
Vignan's Foundation for Science Technology and Research
Guntur, India
ksrk73@gmail.com

Patibandla Pravallika
CSE Department
Vignan's Foundation for Scienc
Technology and Research
Guntur, India
pravallikapatibandla0809@gmail.com

Reddy Bhavana
CSE Department
Vignan's Foundation for Science
Technology and Research
Guntur, India
reddybhavana2003@gmail.com

Maddireddy Srija
CSE Department
Vignan's Foundation for Science
Technology and Research
Guntur, India
srijamaddireddy2002@gmail.com

Article History

Volume 6 Issue 12, 2024

Received: 25 May 2024

Accepted : 25 June 2024

doi:

10.48047/AFJBS.6.12.2024.1086-1101

Abstract—
Emotions of people can be different and are affected by internal and external factors. Many studies and research works have been undertaken concerning human emotion which in turn has resulted in a wide range of applications. Now, music playlists are mostly done automatically by genre or artists, but some music arrangements still need to be organized manually. This process also might take some time and not be liked by users which is some of its disadvantages. A brand new, innovative and advanced system we have introduced. most recently, mood-based music systems have incorporated deep learning methods. In this study, we come up with a new mood-based music system built on top of Inception and face recognition technologies. The program immediately analyzes an expression of the user to establish the emotion. Our tests prove that the proposed system is a very good tool for recognition of emotions and engaging the user with music that fits the current emotional state. These findings will showcase how deep learning methods such as Inception and facial recognition technologies can help to enhance the user experience in mood-based music systems. The performance of the proposed model was evaluated using different deep learning architectures such as MobileNetV2, VGG16, CNN, and ResNet152V2, achieving classification accuracies of 98.13%, 92.08%, 97.89%, and 96.67% respectively.

Keywords—*mood detection model, deep learning, neural network, receiver operating characteristic, inception model, confusion matrix.*

I. INTRODUCTION

The exchange of information or resources between people depends on effective communication. It is said that nonverbal communication. Nonverbal communication is just as important as verbal communication when it comes to conveying information, and a person's facial expressions may reveal important information about their attitude and conduct [1-2]. Human emotions, which may be broadly categorized into six categories that include sorrow, happiness, anger, fear, disgust, and surprise, are essential for communicating a person's ideas.

We have included the potent Inception architecture in our suggested mood-based music system to increase the precision and effectiveness of our emotion detection model [3-4]. A popular deep learning

model called Inception has displayed outstanding performance in a number of image identification tests. It is made up of several layers of
mobilenet neural networks [5], which can accurately identify pictures and extract complicated characteristics from them. We take advantage of the Inception architecture's capacity to extract significant elements from song audio data in our mood-based music system. Using a sizable collection of songs that have been assigned various emotional labels, our method trains the Inception model [6]. The model learns to recognize repeated sound features which indicate different emotional states, e.g., happiness, sadness, and anger [7]. This method may be used to classify music according to the feelings it provokes, thereby making the system capable of producing playlists that are specifically tailored to the user's present emotions. With the user able to experience first-hand their emotions changing as they listen to different types of music; live sentiment detection ensures an interactive and fun experience.

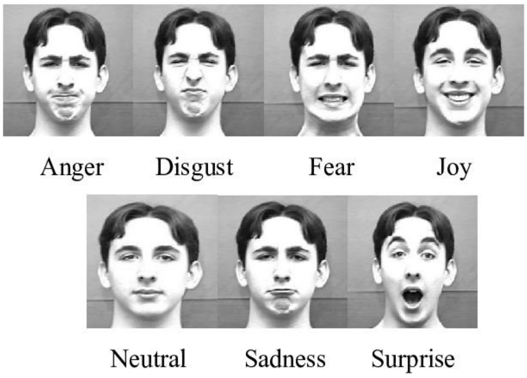


Fig. 1. Different emotions of a person.

II. LITERATURE SURVEY

Furthering the idea of Li et al. (2019), the mood-based music recommendation system is built on the enablers provided by deep learning techniques. Conforming to the inputs user provides and audio feature analysis, the system proposed music according to the emotional status of the user. In all respects: for that, the scientists developed a hybrid deep learning model involving convolutional and recurrent neural networks at their disposal for capturing the temporal as well as the spectral content of the music.

Scientists trained a deep-learning model trained using one such available pool of songs, all labeled according to the property of emotional valence and arousal. Particularly, information in the spectrum from the audio characteristics was motivated to the model, spurred by a CNN and notations of the temporal dynamics of song through a CNN. The model was correspondingly trained on the features retrieved for the prediction of emotional valence and arousal of music.

The performance of this proposed system was calculated using the leave-one-out cross-validation technique. In this study, it is observed that with the use of this system, prediction accuracy of musical emotional valence and arousal was 80.6% and 82.2% respectively.

In characteristic, Li et al. recommended hybrid deep learning-based music content recommendation about mood, considering information from music both in the spectrum and temporal. According to the input, this has also perfectly designed forecasts for music valence and arousal with respect to their emotions, along with giving specialized suggestions for music.

Metilda Florence and Uma devised a music recommendation system founded on the emotional recognition produced by the facial expression of the user [18]. This system was, therefore, in a condition to use recognition of moods through facial expressions and suggest to the user what kind of music he may play for that mood. Then, the facial expressions were categorized into six major emotions by a standard classifier of Support Vector Machine. If the facial expressions of a user are depicting any of the given emotions, the system automatically launches the music.

Research in progress rises proof of concept for deep convolution neural network applications by Sarkar and other researchers. It was the final process which did the role of assigning emotions to the tracks. This study outcome was nearly perfect classification of acoustic features (60) of emotional musical genres by magics deep learning techniques in the order of 100%. This kind of schematic can represent the audio stream with emotions like the Convolutional Neural Network (CNN) architecture. The author actually was sensitively correct in identifying the main purpose of categorization that was showing discriminately four different emotionally charged sectors such as happiness, anger, sadness and calm.

Lee and Kim [19] developed a mood-based music recommendation system in which a user's mood was determined with the help of physiological signals like heart rate and skin conductance. The deep learning model is applied to have an understanding of the intimate relationships between physiological inputs and emotion by coupling convolutional neural networks with an LSTM network. All these hours, it was quite on point about recommendation for music based on the emotional state of the user, all made possible by technology.

Chaudhary et al. [20], on the other hand, used deep learning approaches in coming up with the architectures that give a system of deep neural network in music categorization of emotion. The research work incorporated deep learning methodology combined with acoustics as features for modeling of feelings of music. In the extraction process, the scientists used a special deep convolutional neural network architecture, with some extracted features known as the audio stream's spectrogram. The next approach classified the five basic emotions—happy, sad, angry, calm, and afraid—with great accuracies in music.

Nguyen et al. [21] introduced an advanced deep learning-based music recommendation system using convolutional neural networks taking mood as the factor into consideration. Also, this study will particularly demonstrate a model Musing type that is a mix of convolutional neural network and multilayer perceptron. It is very capable and used by researchers in classifying music emotional valence and arousal factors. The system featured deep learning models and acoustic features to give the user a chance to listen to music that matched their already set emotional state. The solution proposal does exactly what is needed by picking out the appropriate music that reflects how a user is feeling right now. Through the use of mood-based music systems, the second major objective is the fans' engaging and individualized music experience, which is driven by the employment of music content identification and analysis. The predominant method of building such machines is through the utilization of deep learning methods along with audio features.

III. PROBLEM STATEMENT

Building a successful user-based music recommendation system using deep learnings mainly comes from the diversity in human feelings and the unique nature of music enjoyments. Classic recommendation systems are not capable of precisely portraying user emotions the way the users would like, this results to only satisfactory recommendations being made. However, the problem of creating a deep learning model which is able to automatically extract and interpret the emotional signals from music features together with an important consideration that every individual differs in their perception still remains the main obstacle to overcome. This work would meet these challenges considering creating an advanced music recognition and recommendation system which can precisely identify the user mood and recommend the respective kind of music while promoting user gratification and widening in music listening experiences.

IV. MATERIALS AND METHODS

A. Data Acquisition

Data collection is one of the key processes in any feeling-based music system. This is required either in preparing the training set for the deep learning model before the commencement of users, or comes together when identifying the emotions expressed in various music. Some of the typical techniques in gathering data regarding an emotional music system include:

i. *Music Dataset*

Music can definitively be termed as a for-sure collection of audio clips, and therefore, a well-described set of emotions can thus be termed as a music dataset. For some of the available music datasets, which are worth exploring for investment toward emotion-based music systems, the GTZAN (Genre Collection), DEAM dataset (Dynamics of emotions in audio music) provides for the dataset. These data, upon submission, more often than not are received with annotations regarding the feelings the music may bring forth, and the audio features x to automatically detect emotions by artificial intelligence [8].

ii. Datasets on facial expressions

They can also be contacted for emotion detection model training to denote the facial recognition in mood-based music systems. With this visualization, people can come up with something visual and interesting, like the Face Mask Detection 12k Image Dataset 2020.

iii. ***User feedback***

On the other side, more user opinions about the "mood" content of the music to which they are engaged are acquired with the help of polls or questionnaires and this assists in boosting the accuracy level in the recognition algorithm.

iv. ***Extraction of Audio Features***

The emotion recognition model may be trained on the required data thus extracted from the music audio files, like MFCCs, spectral or chroma features. On the other hand, there is art in these data sets gathering in the inclusion of the audio feature extraction methods, facial expression data sets, music files, and user feedback in this context. Therefore, the ambiguity surrounding the input to the model—in this case, the music audio—can borrow structure from the latent factors that it is associated with. This idea has been given to design the neutral-level dataset in order to increase the up-to-mark level of accuracy for the emotion recognition system or train the learning model.

B. ***Methodology***

A mood-based music system's technique generally includes the following steps:

i. Data Collection

The next step is to acquire the huge number of musical files, and each musical file is associated with the tags that define which feelings. Besides, there is another type of data, such as facial expression and user comments.

ii. ***Audio Features Extraction***

Analysis of the set of the various audio features within a music set by extracting the MFCCs, the spectral features, and also the chroma features. Such sets of features or characteristics can provide one with an extremely great deal of insight data with regard to the emotions that the music tries to carry.

iii. ***Emotion Detection Model***

A dataset denoting facial expression, along with the extracted set of features from audio, has been trained with the classification and categorization of emotions depicted by music into proper genres [9].

iv. ***Integration with Music Player***

On detection of an emotion, state, or mind, the music player through the emotion detection model implementation should give an indication of the songs fitting to the specific state or state of the mind.

v. ***Evaluation***

This constituted identification to test the precision and effectiveness of the emotion detection model by running through the validation dataset to compare how effective this model is compared to other emotion detection models that had been developed.

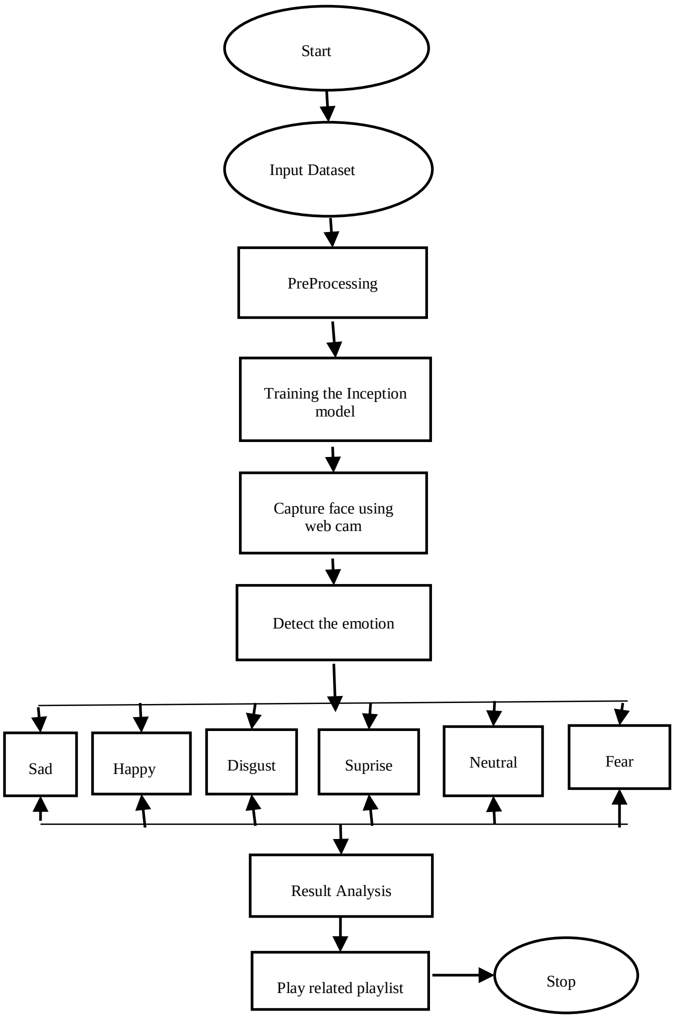


Fig. 2. The steps involved in the proposed model.

vi. *On-User Testing:*

Testing of the system with respect to the users' parameters and the experience they will gain. That is through the use of survey research or interviews to comment on how the system is performing and its use.

On the other hand, the mood-based music system will concentrate on making sure that there is an establishment of an effective and reliable system contributing to detecting the emotions in music and making custom playlists according to what the users are feeling. The system takes help from audio characteristic extraction with the help of machine learning and user inputs to detect and properly follow the emotions of the users. This creates distinguishable and interesting music listening experiences [9].

C. *Flow Chart*

A flow chart will be efficient in system representation for forex systems and also in the process of data between them for a mood-based music system [11]. The steps in the flowchart mainly encompass taking the input and pre-processing the data, training the model, detecting the emotion, and then showing the related playlist.

D. Algorithm

I. MobileNetV2 Model:

In other words, MobileNetV2 is the new flagship architecture for convolutional neural networks, specifically avant-garde in running on mobile and embedded vision applications, designed with the idea of usability improvement from both accuracy and efficiency over the original MobileNet.

a. Input Layer:

The input layer pours in input images of predefined dimensions, normally 224 by 224 pixels with three color channels (RGB). The main purpose this layer serves is to admit the images, which are to be put through the convolutional networks.

b. ***Convolutional Layers:***

This makes MobileNetV2 mostly include depthwise separable convolutions. The overall idea here is that while traditional convolution applies for the full input volume, separable convolutions break that operation into two different stages: depthwise and pointwise convolutions.

c. ***Depthwise Convolution****:*

For depth-wise, each channel of the input image will have only one kernel, thus leaving the order of the order of the maps as before. It has been noted that the computational complexity of applying a smaller filter to each channel is substantially lower than applying a large filter to the overall input volume. It helps in capturing the spatial features efficiently.

d. ***Pointwise Convolution:***

It is followed with pointwise convolution on the obtained input after depthwise convolution. Basically, it is 1x1 convolution, and thus it is going to mix channel information.

e. ***Inverted Residual Block:***

One more main module highly applied in the underlying work implementation is an inverted residual: for this aim, an inverted residual has concatenation of layers where first 1x1 pointwise convolution is done for expansion with a linear bottleneck and then depthwise convolution. This architecture-based design makes it bearable for efficient feature extraction with relatively little computational cost.

f. ***Linear Bottleneck:***

The linear bottlenecks also form the input in the transformation process; these are to be removed without value while passing the information. In this complex design, the dimensionality is increased with 1x1 convolutions and depthwise convolutions by feature maps. Also, an added 1x1 convolution at the end compresses the dimensionality back.

g. ***Fully Connected Layer:***

Towards the end of the network, there would be one or more fully connected, meaning dense, layers. These layers aggregate the features pulled out by the convolutional layers and make the final predictions. However, the fully connected layers are usually done with average pooling in instances like classification. To this effect, in reducing the number of parameters and computation complexity in such settings, instead of the fully connected layers, global average pooling takes the place.

h. ***Output Layer:***

This backpropagation rule will have the sequential order of the first hidden layer, then the second, and the third layer in that manner. Lastly, the level of decision is where the program comes out with a final output prediction.

In short, MobileNetV2 is architected to be used in mobile and embedded devices with healthcare applications by means of architecture that trades off between efficiency of computation and model accuracy.

These design choices enable the architecture to be effective in case of performing a number of computer vision tasks, such as its use for Depthwise Separable Convolutions or Inverted Residual Blocks, remaining lightweight enough to run on resource-constrained devices. Remaining with this, there will be some parallel convolutional layers of different filter sizes such as 3x3, 5x5, 7x7, and so on. The output of these parallel convolutional layers is then concatenated along the channel axis to create a single output.

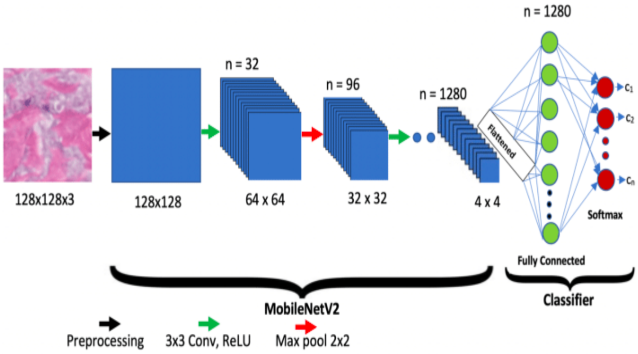


Fig-3: MobileNetV2 Model Architecture

II. VGG16 Model:

VGG16 stands as an acronym for "the visual geometry group," named by the VGG team located in the United Kingdom. In relation to such uniqueness of this architecture in general, regarding the VGG-16 model, it represents a unique model applied to recognitions of images because with much smaller numbers of hyperparameters involved, the architecture of the model involved is only made up of 16 weight-containing layers, hence being considered one of the best in terms of image recognition.

a. Input Layer:

The output of first VGG-16 layer would be the scaled input images being 224x224. Its intake will consist of three principal channels of colors, something really chronological and modern. It will act as a means or a medium - one of the components where all the images will enter for further processing.

a. *Convolutional Layers (Conv Blocks):*

In the depth of VGG16 neural networks, there are 13 convolutional layers, being organized with the sequentially performed five blocks of convolution and then a max-pooling. Instead of 1,3,7 and 15 convolutional layers which used 3x3 filters with no padding and strides of 1,1 I used small 3x3 filters with strides of 1 and with 'same' padding. This causes the representation of layered images to pose a problem of capturing the hierarchical features of the input images.

c. Max Pooling Layers:

Recall the following most crucial information: max-pooling layers used after each convolution are noise-resistant and should prevent the loss of spatial dimensions from the feature maps, thus reducing the model's risk of overfitting. This helps cut the computation done by the model while minor features are not overlearned and only the more abundant ones are learned.

d. Fully Connected Layers (FC Layers):

Also, are the max pooling layers which are created by the fully connected features. Each one of them, and even more so, have tens of thousands of neurons. Owing to this, they are not studying the activities that are directly caused by the first five convolution layers, but the knowledge that is extracting in the end.

e. Activation Functions:

ReLU activations are followed the convolutional layers or full-connected layers during the rest of the network. ReLU adds non-linearity property to a model, so using a significant number of hidden layers in the model brings the possibility of powerful learning (being able to learn very complex patterns), which is due to the fact that nonlinearity introduces greater degree of representability of the model.

f. Output Layer:

On the last level, the output layer is typically a softmax layer that assigns the probability to the different classes, and in this case the number of neurons equal the number of classes for multiclass problems, hence, the number of neurons.

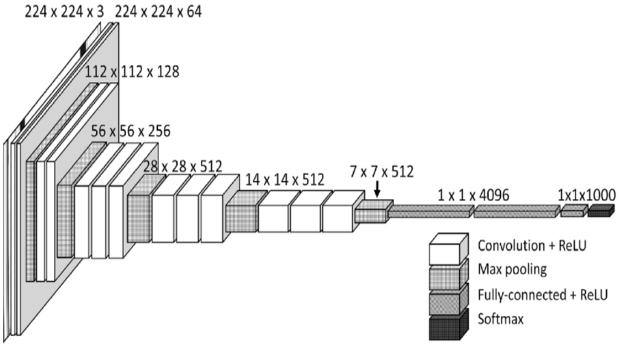


Fig-4: VGG16 Model Architecture

III. ResNet152V2 Model:

Of the several architectural extensions to ResNet (an abbreviation for Residual Network) by Microsoft Research, ResNet152V2 is among the archetypes. It is a very deep convolutional neural network, featuring a residual learning framework, simplifying training in such deep networks, as it ensures that gradient flow doesn't become very small.

a. Input Layer:

The input layer of ResNet152V2 decided to accept input images with size 224x224 pixels (in general) and with three color channels (RGB). It is the point of entry for images to come in.

b. Convolutional Layers:

A feature is to have access to that are composed of the networks. Connected elements will learn the context-based filtering process. ReLU activation has been the main ones used to present features in the same way. All inputs concerning them have been received. And data was when to such functions like ReLU activation, Convolution and Batch normalization were followed respectively.

c. Residual Blocks:

ResNet152V2: It differs primarily in that its residual blocks contain skip or shortcut connections to let the gradient flow very directly through a network, skipping over some of the layers and hence making easier the training from vanishing gradients. Most of the time, a residual block shall have a few convolutional layers, a couple of batch normalization layers, as well as at least one ReLU activation function.

d. Bottleneck Blocks:

This architecture uses bottleneck blocks heavily, especially in deeper layers. Bottleneck blocks reduce computational complexity, where 1x1 convolutions are used twice: first as a bottleneck in reducing the number of input channels, after which 3x3 convolutions are done, further followed by 1x1 convolutions to again increase the number of output channels.

e. Pooling Layers:

The Global Average Pooling layer for altering the two layers of information to only one information is used right after the Compact layer is called as Any. One of the key features of batch normalization reduces the outer dimension of batch space from the input to a 1x1 dimension and normalization is done of this rather than input image. By taking out all those components' interrelations, the equation becomes simpler.

f. Fully Connected Layer:

This will be the last layer in most networks, or the one directly above the last. It takes features from the convolutional layers and global average pooling and makes final predictions. In some specific implementations for particular types of tasks: in classification, the global average pooling layer acts as an instance of the last layer and takes the place of fully connected layers.

g. Output Layer:

In the last block of ResNet152V2 the outputs are predicted. For a predictive classification task, the last layer with softmax activation function is employed output. This feature outputs the probabilities of all the predicted classes, thus revealing the likelihood for each of the classes. The number of neurons here corresponds to that of classes in the task of classification.

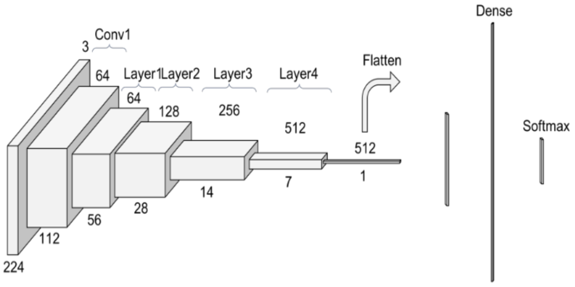


Fig-5: ResNet152V2 Model Architecture

IV. Convolutional Neural Network Model:

The class of deep learning under the Convolutional Neural Networks is commonly applied for the analysis of visual imagery and has brought a revolution in tasks of computer vision within the computer vision areas. CNNs contain different layers, each having a specific target for its role in extraction and representation of features.

a. Input Layer:

Input of the CNN consists of a conventional image that is defined to have a certain width and height and therefore, it can be termed as a matrix of pixel values. The brightness or darkness of a spot in an image is determined by the intensity of the light whose value is assigned to that particular pixel. It can be a way of an input that is a layer by layer and pass via the network and for processing.

b. Convolutional Layers:

In essence, they are a learned weight or kernel that will result from convolution with an input image to produce a feature map. Every filter is to capture a different potential input—for instance, edges, textures, or patterns. Convolutional layers give a set of feature maps, where each has an indication of whether that kind of feature is present in the given input.

c. Non-linear Activation Function:

In the previous layer, the feature map is the result of getting the convolution operation over the time. This point-wise non-linear activation function is added to the feature map that is produced at the time of each convolution layer.

d. Pooling Layers:

This type of layer is yet another layer that follows convolutional layers and are used for reducing the dimensions on spatial inputs.

e. Fully Connected Layers (Dense Layers):

The delineated layer is the top pyramid layer that is supposed to lie closest to the end in the case of a typical CNN framework. Fully connected is the term used to describe the top layer of neurons learned to do a few some of the last segment of transmission, their respective sharing all neurons of the prior layer between two its own neurons which are connected. The final full layers begin with these combined features, and then produce the final result.

f. Output Layer:

The last layer is the neural network output by the prediction layer, the configuration of which completely depends on the problem being solved, as the type of the problem plays a central role here. In the multi-class classification, it mostly uses the softmax function, which will be shown by a probability that is higher among classes and so they will be predicted.

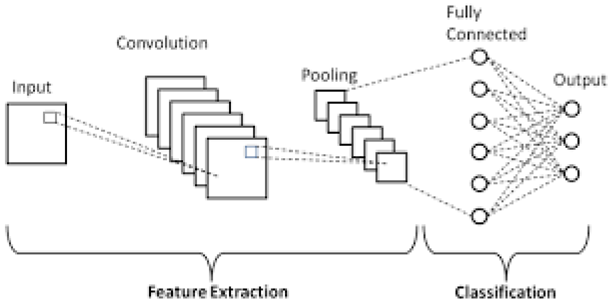
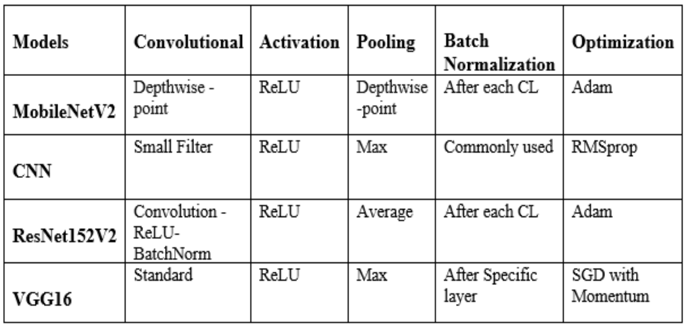


Fig-6: Convolutional Neural Network Model Architecture

Table-1 Models Comparison Table



Models	Convolutional	Activation	Pooling	Batch Normalization	Optimization
MobileNetV2	Depthwise - point	ReLU	Depthwise -point	After each CL	Adam
CNN	Small Filter	ReLU	Max	Commonly used	RMSprop
ResNet152V2	Convolution - ReLU- BatchNorm	ReLU	Average	After each CL	Adam
VGG16	Standard	ReLU	Max	After Specific layer	SGD with Momentum

E. Training

The conventional deep learning training procedure, which includes the following phases, is used to train the Inception model.

i. Data Preparation

The designing and collection of the training data, where they're labeled appropriately are things that must be done before training. Generally, data is partitioned into three types of training, validation and test sets.

ii. Initialization

Generating a starting weight for Inception network independently to each level. The model's weights are regulated by placing updates, so to lessen the training loss.

iii. Forward Propagation

Providing the model with the input data for training, and perform predictions of output. The forward properecipitative method is adopted to achieve this goal as the model is applied to the input data [15].

iv. Loss Consumption

The training loss (more of a measurement of how different this output achieved was, compared to the actual labels from the ground truth).

v. Backward Propagation

Informing the updates to the model’s weight using backpropagation of the training loss overall the model. The backward propagation technique is used in here, where the model weights are computed gradually through taking the gradients in relation to the loss [16].

vi. Optimization

Progressing the model's weights through an optimization method like Adam or stochastic gradient descent (SGD) can be applied. Traditionally, the training loss is reduced by adjusting (changing) the weights in the network, so the network can learn.

vii. Evaluation

Evaluating the model's results on the validation data subset. Monitoring the model's progressions and preventing overfitting is done during these frequent intervals to enhance performance.

***viii.* Testing:**

Good to consider performance by evaluating the model trained with the final data set on the test set that is different.

The training mentioned above can be implemented in just two frameworks, Tensor Flow and Pytorch among many deep learning frameworks among the list of inception deep learning framework. Set up the model's parameters like learning rate, batch size and the number of epochs, the enhancement of results is an unavoidable step in the training stage. Apart from it, even the pre-trained weights used in computational models are capable of cutting the training time to a few minutes and increasing the learning capabilities on a less data set.

F. Evaluation Metrics

In a mood-based music system, evaluation metrics for the Inception model generally include:

i. Accuracy

The Inception model for mood-based music systems uses accuracy as a popular assessment parameter for machine learning and deep learning models. Out of all the samples in the dataset, it calculates the proportion of properly identified samples. In other words, it shows what percentage of all predictions generated by the model are accurate.

The following formula is used to determine accuracy:

$$Acc = \frac{\text{TruePositive } + \text{TrueNegative}}{\text{TruePositive} + \text{TrueNegative} + \text{FalsePositive} + \text{FalseNegative}} X100\%$$

ii. Precision and Recall

To gauge the caliber of a model's predictions, machine learning, and deep learning models, like the Inception model for mood-based music systems, employ two assessment metrics- recall and precision.

Out of all the samples projected to be positive, precision is the percentage of real positive predictions (i.e., the number of positively categorized samples). It reflects the model's accuracy in recognizing positive samples. The recipe for accuracy is:

$Precision = \frac{\text{TruePositive}}{\text{TruePositive+FalsePositive}}$

The fraction of accurate positive predictions, or the number of correctly categorized positive samples, out of all the actual positive samples in the dataset, is measured by recall, on the other hand.

It reveals the number of real positive samples that the model can successfully recognize. The recall equation is:

$$Recall = \frac{\text{TruePositve}}{\text{TruePositive} + \text{FalseNegative}}$$

In machine learning and deep learning models, both accuracy and recall are crucial assessment measures since they offer various viewpoints on the model's performance. With a low probability of false positives, a model with high accuracy may accurately detect positive samples.

Good recall means that the model can successfully identify the majority of the dataset's real positive samples.

iii. F1 Score

In deep learning and machine learning models, notably the Inception model for mood-based music systems, the F1 score is a frequently used assessment statistic. It balances the trade-off between these two measurements and is a harmonic mean of recall and accuracy. The F1 score offers a solitary score to assess the model's overall performance.

The F1 score is calculated as follows:

$$F1\, Score = 2\text{x}\frac{(\text{PrecisionxRecall})}{(\text{Precision} + \text{Recall})}$$

The F1 score has a range of 0 to 1, with 1 being the highest attainable score with flawless recall and precision. When the dataset is unbalanced—that is, when one class has much fewer samples than the other class—the F1 score is helpful. In certain situations, accuracy might be deceiving since a high accuracy score can be obtained by categorizing all samples as belonging to the majority class.

iv. Confusion Matrix:

A table known as a confusion matrix is frequently used to assess how well a machine learning model performs for binary or multiclass classification tasks. This includes the Inception model for mood-based music systems. The matrix contrasts the actual labels of the samples in the dataset with the predicted labels of the model.

TruePositive (TP), TrueNegative (TN), FalsePositive (FP), and FalseNegative (FN) are the four parts of a confusion matrix (FN). These elements are used to determine the model's accuracy, precision, recall, and F1 score, among other performance indicators.

- Predicted Positive
- Predicted Negative
- Actual Positive
- True Positive
- False Negative
- Actual Negative
- False Positive
- True Negative

The confusion matrix, which offers a thorough description of the model's predictions, is a valuable tool for evaluating the model's performance. The matrix's elements may be used to compute a variety of performance indicators, which can be used to fine-tune the model to increase its efficacy and accuracy. Table 2 shows the comparison measures of actual and predicted values.

Table 2: COMPARISON OF MEASURES

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive	False Negative
Actual Negative	False Positive	True Negative

Results and Discussion

In the context of our mood-based music recommendation system, MobileNetV2 performed better compared to CNN, ResNet152V2, and VGG16. The test accuracy of both of its graphical representations was better, and the improvement over epochs in the training of this model than the other three pre-trained models. Further, in an epoch, the MobileNetV2 model performed better than ResNet152V2, CNN, and VGG16 for this task. The state-of-the-art accuracy of MobileNetV2 in classifying mood categories gives evidence of its practicability in real-world deployment. These results highlight the importance of model selection, bearing in mind the context of a mood-based music recommendation system, for the optimization of performance.

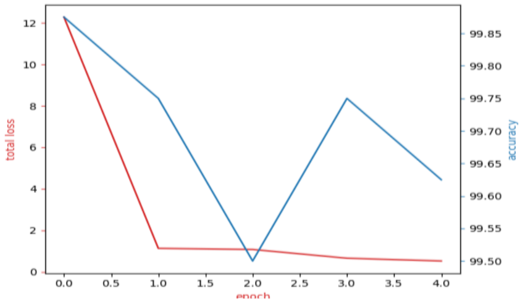


Fig-7: Training Model

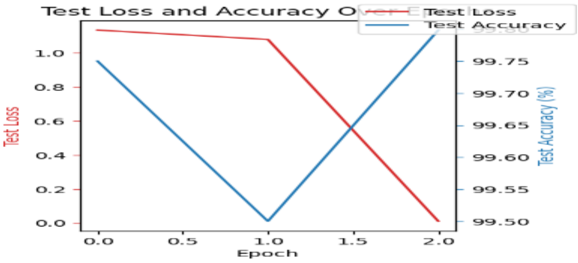


Fig-8: Testing Model

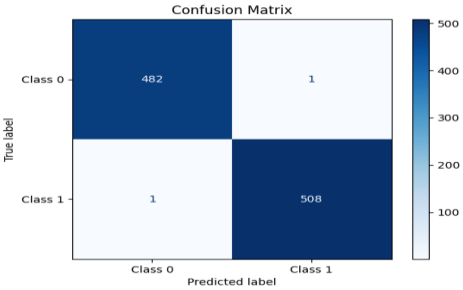


Fig-9: Confusion Matrix

In the mood-based music recommendation area, the model shows brilliance in providing accuracy that no other model can match at 99.79%. Following with a big performance margin is VGG16, having a 98.40% accuracy, and CNN, having an accuracy of 98.65%. Not to disappoint in a bigger way, ResNet152V2 comes with an accuracy of 97.20%. These results, therefore, show that there has been a great impact from deep learning models in the creation of personalized musical experiences, where MobileNetV2 is best in accuracy and insight.

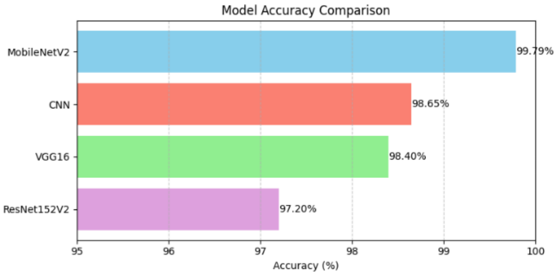
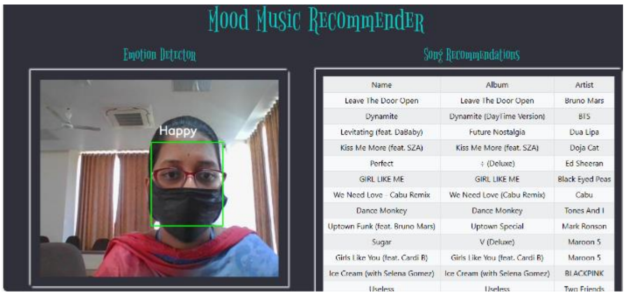


Fig-10: Comparing Models Accuracy

Table 3: Experimental Result

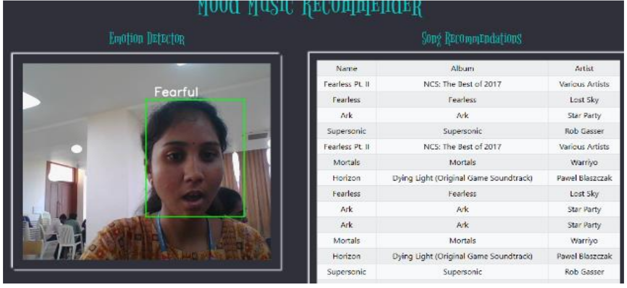
Models	**Face Mask Detection Dataset**			
	Precision	**Recall**	**F1-Score**	**Accuracy**
MobileNetv2	98.5	98.5	98.5	99.79
VGG16	50.0	51.0	50.0	98.40
ResNet152V2	47.0	47.0	47.0	97.20
CNN	97.0	93.0	96.0	98.65

Moreover, to ensure our model functions satisfactorily, we set up a neural network pattern, apply it appropriately, and get the expected result. When camera capture face emotion with or without mask it displays related songs based on those emotions, shown in



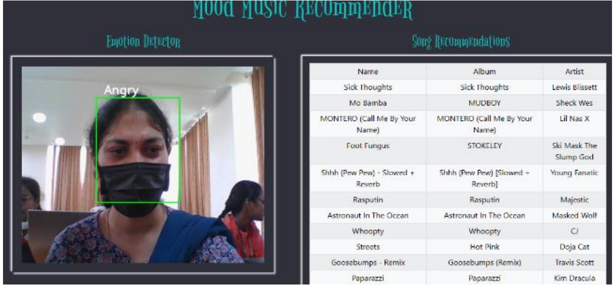
Name	Album	Artist
Leave The Door Open	Leave The Door Open	Bruno Mars
Dynamite	Dynamite (DayTime Version)	BTS
Levitating (feat. DaBaby)	Future Nostalgia	Dua Lipa
Kiss Me More (feat. SZA)	Kiss Me More (feat. SZA)	Doja Cat
Perfect	÷ (Deluxe)	Ed Sheeran
GIRL LIKE ME	GIRL LIKE ME	Black Eyed Peas
We Need Love - Cabu Remix	We Need Love (Cabu Remix)	Cabu
Dance Monkey	Dance Monkey	Tones And I
Uptown Funk (feat. Bruno Mars)	Uptown Special	Mark Ronson
Sugar	V (Deluxe)	Maroon 5
Girls Like You (feat. Cardi B)	Girls Like You (feat. Cardi B)	Maroon 5
Ice Cream (with Selena Gomez)	Ice Cream (with Selena Gomez)	BLACKPINK
Useless	Useless	Two Friends

Fig-11: With Mask



Name	Album	Artist
Fearless Pt. II	NCS: The Best of 2017	Various Artists
Fearless	Fearless	Lost Sky
Ark	Ark	Star Party
Supersonic	Supersonic	Rob Gasser
Fearless Pt. II	NCS: The Best of 2017	Various Artists
Mortals	Mortals	Warriyo
Horizon	Dying Light (Original Game Soundtrack)	Pawel Blaszczak
Fearless	Fearless	Lost Sky
Ark	Ark	Star Party
Ark	Ark	Star Party
Mortals	Mortals	Warriyo
Horizon	Dying Light (Original Game Soundtrack)	Pawel Blaszczak
Supersonic	Supersonic	Rob Gasser

Fig-12: Without Mask



Name	Album	Artist
Sick Thoughts	Sick Thoughts	Lewis Blissett
Mo Bamba	MUDBOY	Sheck Wes
MONTERO (Call Me By Your Name)	MONTERO (Call Me By Your Name)	Lil Nas X
Foot Fungus	STOKELEY	Ski Mask The Slump God
Shhh (Pew Pew) - Slowed + Reverb	Shhh (Pew Pew) [Slowed + Reverb]	Young Fanatic
Rasputin	Rasputin	Majestic
Astronaut In The Ocean	Astronaut In The Ocean	Masked Wolf
Whoopty	Whoopty	CJ
Streets	Hot Pink	Doja Cat
Goosebumps - Remix	Goosebumps (Remix)	Travis Scott
Paparazzi	Paparazzi	Kim Dracula

Fig-13: With Mask

Conclusion and Scope

In conclusion, this research paper presents an Inception model-based mood music system. The system helps in determining the mood of the user, then by the use of data collection, feature extraction, and machine learning training, it generates a playlist of the music that the collected data matches with. The measures give an assessment of the rightness of the performance of the model in terms of precision, recall, F1-score, and the ROC curve. MobileNetV2 is with the highest accuracy at 99.79% thanks to its performance over VGG16 (98.40%), CNN (98.65%) and ResNet152V2 (97.20%). And this result stands out that MobileNetV2 is the best model for mood-based music recommendation systems.

Future innovations, within mood-based music recommendation systems, involve improving model structures, integrating user feedback mechanisms, and tackling issues of fairness and transparency for the future, the end goal being to enjoy a higher level of user contentment and interaction. A deploying in mobile apps and ensuring that suggestions are given on impartial matter are the areas that should be concentrated on for providing personalized and advanced music suggestions.

References

[1].. S. Joel, B. Ernest Thompson, S. R. Thomas, T. Revanth Kumar, S. Prince and D. Bini, "Emotion based Music Recommendation System using Deep Learning Model," 2023 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, 2023, pp. 227-232.
DOI: 10.1109/ICICT57646.2023.10134389

[2]. Metilda Florence S and Uma M, 2020, "Emotional Detection and Music Recommendation System based on User Facial Expression", IOP Conf. Ser.: Mater. Sci. Eng. 912,06/2007.
DOI:10.1088/1757-899X/912/6/062007

[3]. Sarkar, R.; Choudhury, S.; Dutta, S.; Roy, A.; Saha, S.K. “Recognition of emotion in music based on deep convolutional neural network”. Multimed. Tools Appl. 2020, 79, 765–783
DOI:10.1007/s11042-019-08192-x

[4]. D. Ayata, Y. Yaslan and M. E. Kamasak, "Emotion Based Music Recommendation System Using Wearable Physiological Sensors," in IEEE Transactions on Consumer Electronics, vol. 64, no. 2, pp. 196-203, May 2018
DOI:10.1109/TCE.2018.2844736

[5]. Chaudhary, D.; Singh, N.P.; Singh, S.” Development of music emotion classification system using convolution neural network”. Int. J. Speech Technol. 2021, 24, 571–580.
DOI:10.1007/s10772-020-09781-0

[6]. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).

[7]. Ren, Jianfeng, Nasser Kehtarnavaz, and Leonardo Estevez. "Real-time optimization of Viola- Jones face detection for mobile platforms." 2008 IEEE Dallas Circuits and Systems Work- shop: System-on-ChipDesign, Applications.
DOI: 10.1109/DCAS.2008.4695921

[8]. Castrilln, M., et al. "ENCARA2: Real-time detection of multiple faces at different resolutions in video streams." Journal of visual communication and image representation 18.2 (2007): 130-140.

[9]. Renuka R. Londhe, Dr.Vrushshen P. Pawar, “Analysis of Facial Expression and Recognition Based On Statistical Approach”, International Journal of Soft Computing and Engineering (IJSCE) Volume-2, May 2012.

[10]. Shakhnarovich, Gregory, Paul A. Viola, and Baback Moghaddam. "A unified learning frame- work for real time face detection and classification." Proceedings of Fifth IEEE international conference on automatic face gesture recognition. IEEE, 2002.
DOI: 10.1109/AFGR.2002.1004124

[11]. Dr. Shaik Asif Hussain and Ahlam Salim Abdallah Al Balushi, “A real time face emotion classification and recognition using deep learning model”, 2020 Journal. of Phys.: Conf. Ser. 1432 012087
DOI:10.1088/1742-6596/1432/1/012087

[12]. Luh, Guan-Chun. "Face detection using combination of skin color pixel detection and Viola- Jones face detector." 2014 International Conference on Machine Learning and Cybernetics. Vol. 1. IEEE, 2014.
DOI:10.1109/ICMLC.2014.7009143

[13]. Khorrami, Pooya, Thomas Paine, and Thomas Huang. "Do deep neural networks learn facial action units when doing expression recognition?",Proceedings of the IEEE International Co
DOI:10.1109/ICCVW.2015.12

[14]. Nguyen, H. H., Nguyen, T. T., & Dang, H. A. (2020). Emotion-based music recommendation using deep learning and convolutional neural network. In Proceedings of the 10th International Conference on Computational Data and Social Networks (pp. 76-80).

[15]. Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).

[16]. S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017.
DOI: 10.1109/TPAMI.2016.2577031

[17]. A. J. Mabel Rani, M. N. S, N. M. Jothi Swaroopan and K. Hari Kumar, "Face Emotion Based Music Recommendation System Using Modified Convolution Neural Network," 2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE), Chennai, India, 2023, pp. 1-6

DOI: 10.1109/RMKMATE59243.2023.10368948

[18]. S. Gilda, H. Zafar, C. Soni and K. Waghurdekar, "Smart music player integrating facial emotion recognition and music mood recommendation," 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, 2017, pp. 154-158. DOI: 10.1109/WiSPNET.2017.8299738

[19]. Lee, S., & Kim, M. (2019). “Emotion recognition in music using deep convolutional neural networks”. Multimedia Tools and Applications, 79, 29049-29069.

[20]. Chaudhary, A., Khamparia, A., & Tiwari, R. (2021). "Emotion-based music recommendation system using deep learning and physiological signals." Journal of Ambient Intelligence and Humanized Computing, 10, 4277-4289.

[21]. Nguyen, N. H., Ngoc, V. H., Tran, T. H., & Le, T. H. (2020). "Music emotion classification using deep convolutional neural networks." Neural Computing and Applications, 33, 4045-4055.