Hindi Music Emotion Recognition

Divyanshi Jagetiya (B19BB012)

Hiteshi Singh (B19EE039)

Kwanit Gupta (B19EE046)

Abstract

In this report, we propose an end-to-end pipeline for identifying the type of emotions induced by different genres of music by studying the spectrogram of audio snippets from Hindi Music obtained from the publicly available MER500 dataset on Kaggle. We have employed the use of multiple Deep Learning based models including VGG16, Resnet-50, Mobilenet-v3 small, Densenet-121, Alexnet, and Efficientnet-B0. The audio snippets from the MER500 dataset are pre-processed to extract the corresponding spectrogram and passed as input to these models from which we classify the induced emotion of the listener. This project has been implemented from scratch and very little relevant work has been done prior to classifying Hindi music.

1 Introduction

Music has always been a huge part of our quotidian routine. The kind of music we listen to has a strong hold on our emotional core. The signals delivered by the audio snippets from a class of music could be captured and are visually represented as a plot between frequencies of the signal with time as a spectrogram. On the heatmaps generated, we can distinguish the power of different pitches of the music based on the intensity of those regions.

In this project, we identify what class of music will induce which kind of emotions in the mind of the listener. Convolutional Neural Networks are preferable for detecting heatmap-like signatures because of pooling layers and convolutional kernels. We attempt to compare and contrast various models and approaches in order to determine which approach will work best for the task. This proposed pipeline could be helpful to neuroscientists, especially those based out of India, for studying the effects of music on different segments of the brain and thereby aid in alleviating various neurological disorders.

2 Dataset and Features

To train our classifier for music emotion recognition, we used the **MER500 dataset** publicly available on Kaggle.

Source. The MER500 data set that we have used here consists of songs in 5 popular emotional categories for Hindi film songs as Romantic, Happy, Sad, Devotional, and Party. It has approximately 100 audio files of about 10 seconds of song clips from the original song are available for music emotion recognition experimentation. This data set of Indian Hindi film music will be useful for music information retrieval.

Preprocessing. We pre-processed the data, as shown in Figure 1, and set the frequency of all audio samples to 16000 Hz. Further, we padded shorter samples, followed by normalization and standardization steps. To represent the data, we use mel-spectrograms, which are the input of our model.

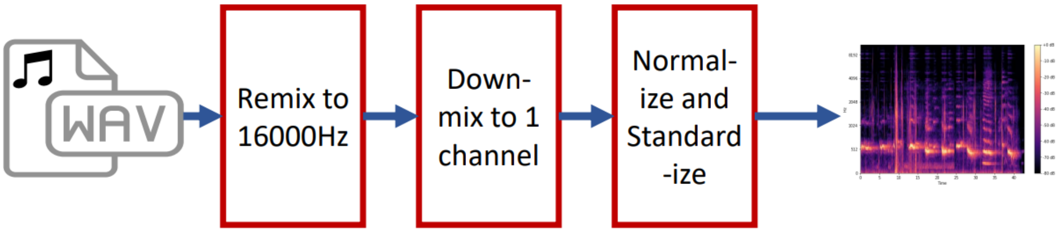


Figure 1. Audio Sample Pre-Processing Pipeline

Training-test split. We defined the data loaders and re-splited the already existing dataset in the ratio of 80:20.

Labeling. Since the machines cannot interpret strings so we converted the data into numeral digits.

Normalization. To ensure that our entire data feature has zero mean and unit variance, we normalized our data set before training.

4 Methods

This section discusses the various models used for extracting features from the spectrograms obtained from the audio dataset. For comparison purposes, we have applied multiple models to the dataset.

4.1 Resnet 18. ResNet-18 is a convolutional neural network that is 18 layers deep. ResNet includes several residual blocks that consist of convolutional layers, batch normalization layers, and ReLU activation functions. We used the pretrained Resnet 18 model to extract features from the spectrograms obtained from the audio dataset.

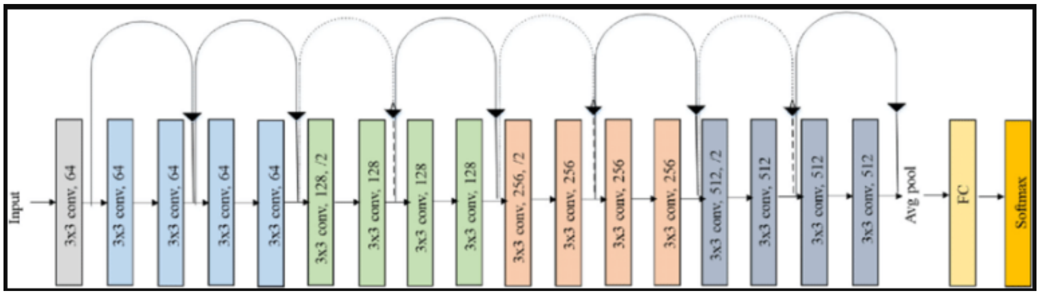


Figure 2.

4.2 VGG 16. VGG 16 was proposed by Karen Simonyan and Andrew Zisserman in 2014 in the paper "Very Deep Convolutional Networks for Large Scale Image Recognition".

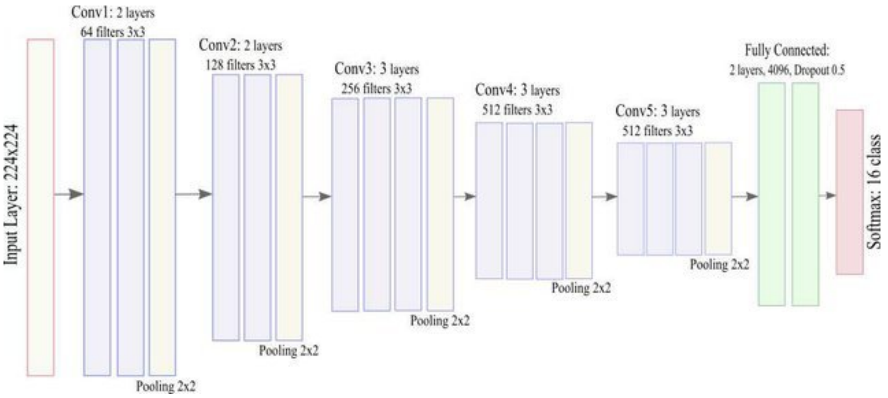


Figure 3.

4.3 Mobilenet -v3- small. MobileNet is a simple but efficient and not very computationally intensive convolutional neural network for mobile vision applications.

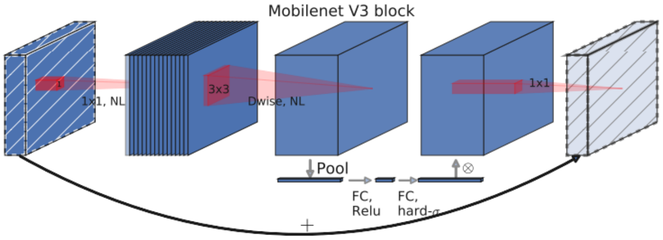


Figure 4.

4.4 Alexnet. Alexnet having eight layers with learnable parameters. The model consists of five layers with a combination of max pooling followed by 3 fully connected layers and they use Relu activation in each of these layers except the output layer.

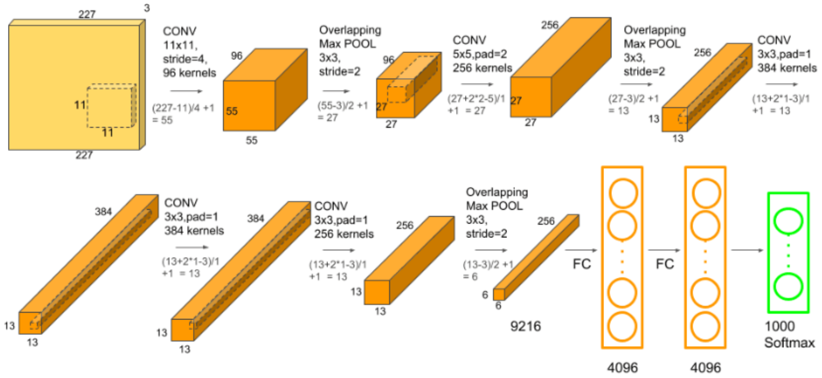


Figure 5.

4.5 Densenet 121

DenseNet-121 has 120 Convolutions and 4 AvgPool. DenseNets require fewer parameters and allow feature reuse. Therefore they result in more compact models and achieve state-of-the-art performances and better results across competitive datasets, as compared to their standard CNN or ResNet counterparts.

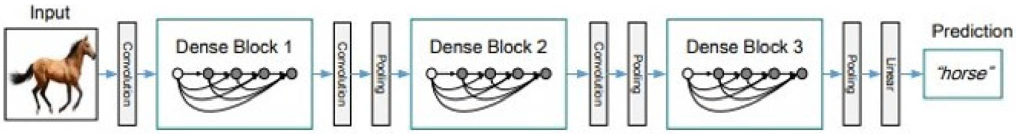


Figure 1: Various blocks and layers in DenseNet (Source: Original DenseNet paper)

Figure 6.

4.6 EfficientNet-B0

EfficientNets, are very much efficient computationally and also achieve state of art results on the ImageNet dataset which is 84.4% top-1 accuracy. This model was developed using a multi-objective neural architecture search that optimizes both accuracy and floating-point operations.

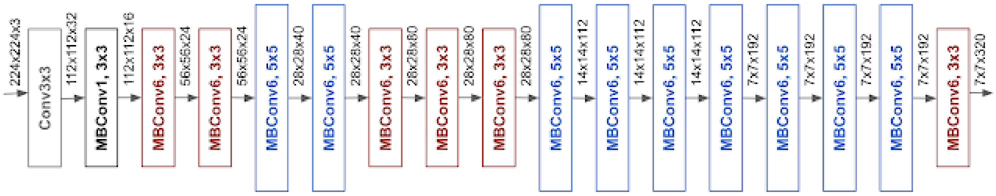


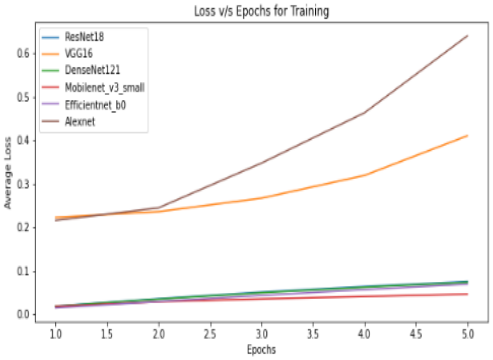
Figure 7.

5 Experiment/Results/Discussion

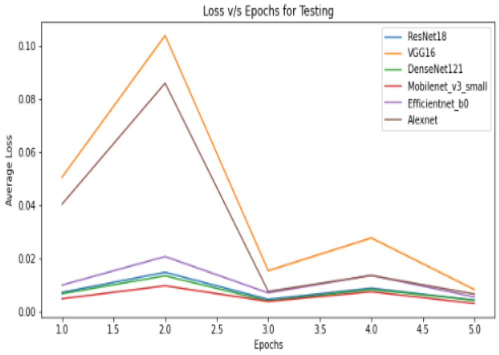
We tested a range of hyperparameters to maximize performance. We altered the batch size, trying values of 16, 32, and 64, and found the best results with a batch size of 64. The loss function used was Binary cross-entropy, Adam as optimizer, and accuracy with confusion matrix as an evaluation criterion. All the models were already trained on the Imagenet dataset.

Here's a comparative analysis of the plots between loss, epochs, and accuracy for both testing and training sets that we have obtained by comparing the different models that we have applied to the data set.

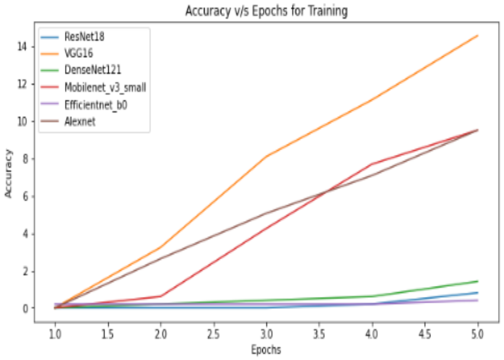
Loss v/s Epochs for training:



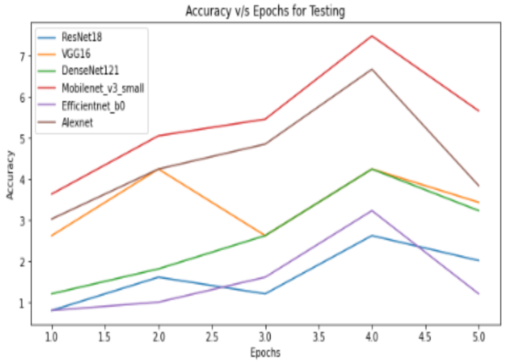
Loss v/s Epochs for testing:



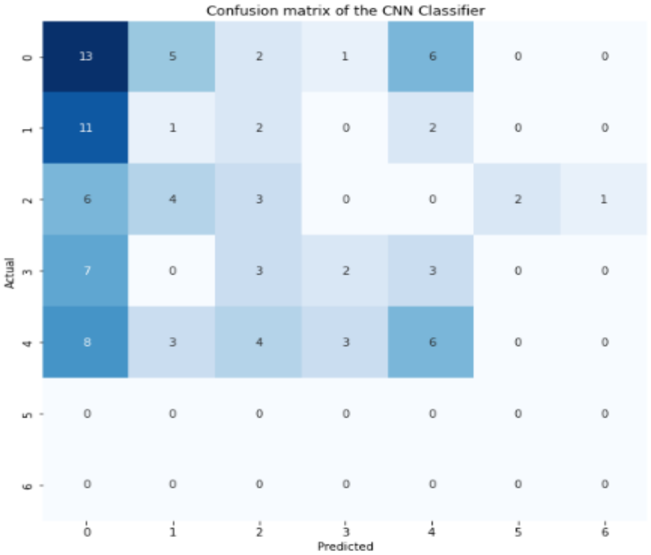
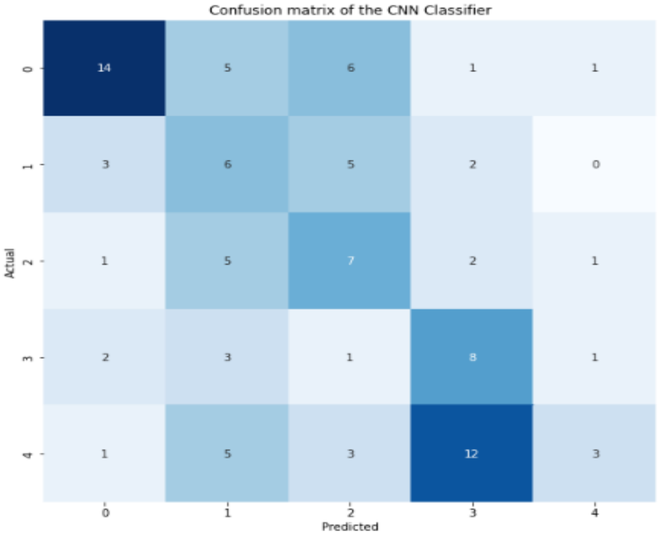
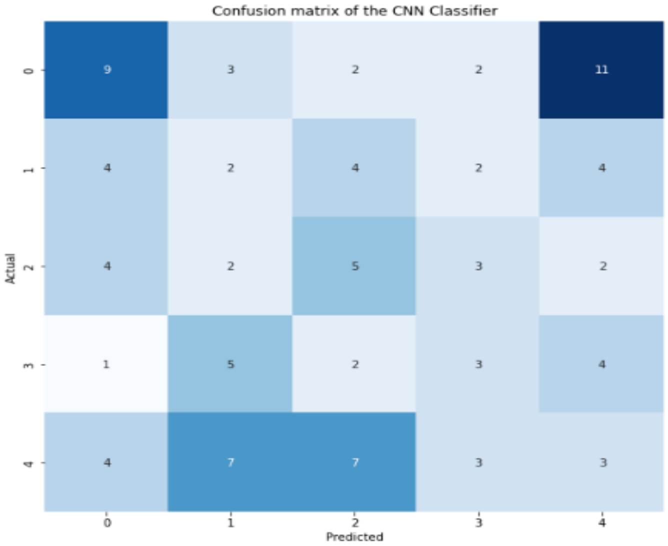
Accuracy v/s Epochs for training:



Accuracy v/s Epochs for testing:



For a perceptive understanding, we have also computed the confusion matrix for the test dataset for all the above mentioned pre-trained models, which are shown in Table 1.

VGG-16	
Mobile-Net-v3	
Densenet -121	

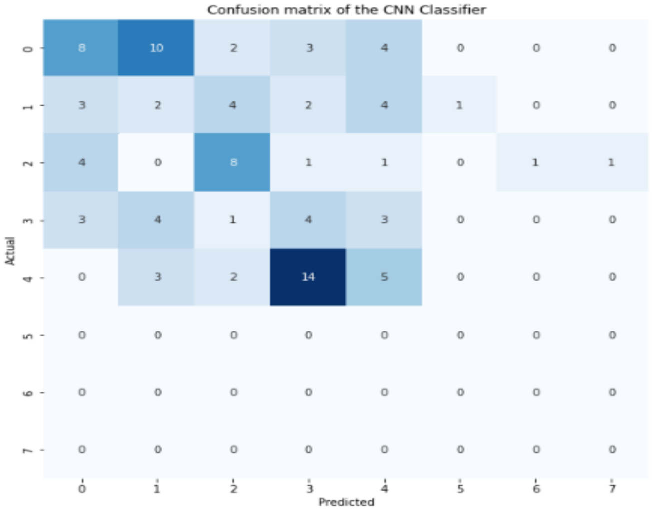
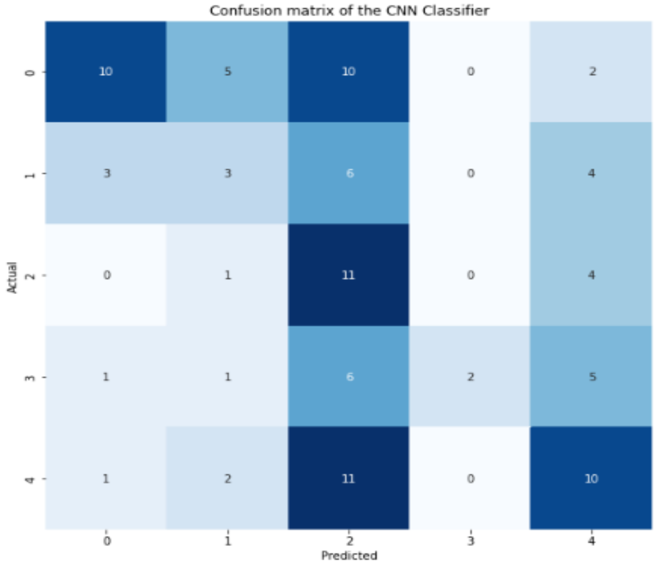
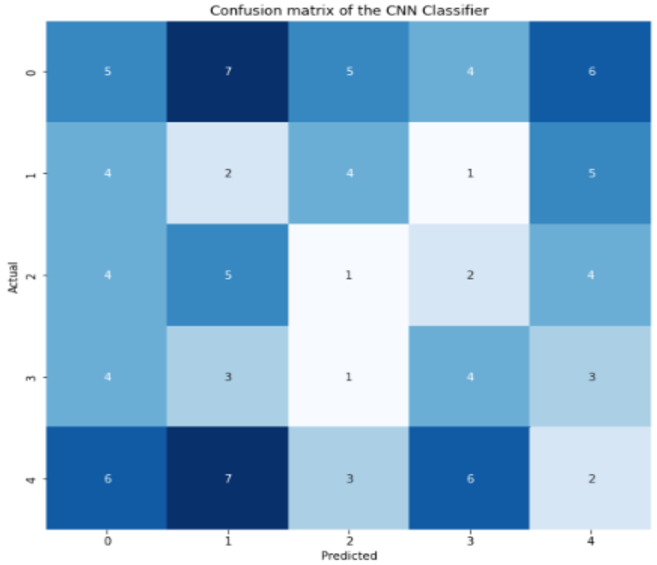
Efficientnet-B0	
Alexnet	
Resnet 18	

Table 1: Confusion Matrix of different models

6 Conclusion/Future Work

The spectrogram representation of music provided sufficient features and information for our convolutional neural network fit to and allowed our model to very accurately differentiate between musical instruments with very different timbres. Amongst the 6 variants of Pre-trained CNNs, Alexnet turned out to be the best, whereas ResNet performed the worst, despite possessing the crucial skip-connections. MobileNet-V3 also showed promising results since the dilated and point-wise convolutions helped to focus on minute changes in the spectrogram (specifically on the faint regions).yAs Convolutional Neural Networks lack the ability to extract Temporal relations from the spectrographs due to the audio and videos being sequential data, therefore, the accuracy of different models comes out to be comparatively less, as expected.

The following experiments could be carried out in the future to further develop the project:

- Further train these models on a larger dataset to get better results.
- By a combination of different architectures and ensembles we can identify some temporal relationships and hence improve the accuracy of models.
- With the help of transfer learning, we can extend our pipeline to multiple music genres and classifications based on different factors.

7 References

1. https://www.kaggle.com/datasets/makvel/mer500?resource=download
2. https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.researchgate.net%2Ffigure%2FVGG16-architecture-There-are-in-total-five-blocks-the-first-two-blocks-have-two-Conv_fig1_338572801&psig=AOvVaw0rejPJ2F2VL-CfLfTcXwSE&ust=1651612142168000&source=images&cd=vfe&ved=0CA4Q3YkBahcKEwig06Xr3MH3AhUAAAAAHQAAAAAQAw
3. http://cs230.stanford.edu/projects_fall_2019/reports/26225883.pdf
4. https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.researchgate.net%2Ffigure%2FOriginal-ResNet-18-Architecture_fig1_336642248&psig=AOvVaw0j-7ZknGc8kAoX0JummpcD&ust=1651612327640000&source=images&cd=vfe&ved=0CA4Q3YkBahcKEwjI3ajE3cH3AhUAAAAAHQAAAAAQAw