



**CLASSIFICATION OF MUSIC EMOTIONS WITH PRE-TRAINED
MODELS**

ZAINAB YASEEN TAHA TAQA

AUGUST 2024

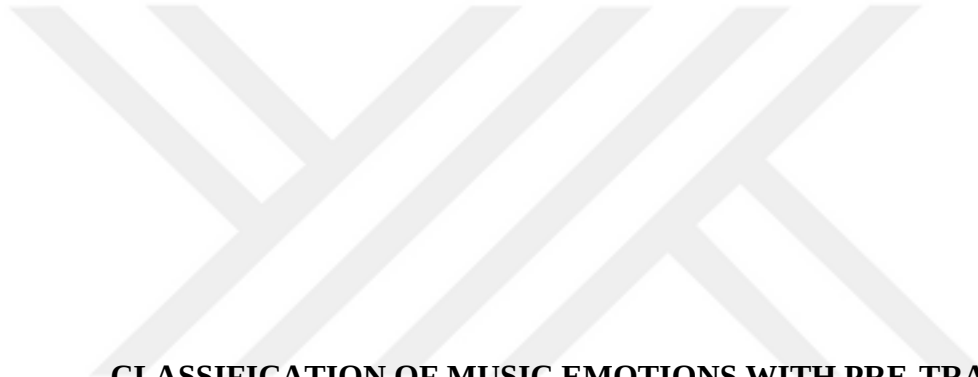
ÇANKAYA UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

DEPARTMENT OF COMPUTER ENGINEERING

M.Sc. Thesis in

INFORMATION TECHNOLOGIES



**CLASSIFICATION OF MUSIC EMOTIONS WITH PRE-TRAINED
MODELS**

ZAINAB YASEEN TAHA TAQA

AUGUST 2024

ABSTRACT

CLASSIFICATION OF MUSIC EMOTIONS WITH PRE-TRAINED MODELS

TAQA, ZAINAB

M.Sc. in Information Technologies

Supervisor: Assoc. Prof. Dr. ABDÜL KADIR GÖRÜR

August 2024, 135 pages

Music is the most emotional of all arts and is therefore the closest and dearest to people. Emotions are a central part of human life, influencing how we think, act, and make decisions. Many scientists have studied emotions, including Paul Ekman, who classified basic emotions into six kinds: happiness, sadness, fear, anger, surprise, and disgust. This classification helps in understanding how emotions influence our behavior and daily interactions.

Emotion analysis can benefit various fields such as healthcare, education, business, marketing, technology, and more. Various applications facilitate emotion extraction and analysis, including:

- **Text Analysis:** Natural Language Processing (NLP) analyzes written texts to extract emotions.
- **Voice Analysis:** Applications like Beyond Verbal and Affective analyze voice tone to extract emotions, which is useful in customer service centers.
- **Image and Video Analysis:** Computer Vision techniques analyze facial expressions to identify emotions from images and videos.

Music occupies a special place in stirring up and expressing human feelings, with its melodies, rhythms, and tones provoking a wide range of emotions.

The analysis of musical emotions is a growing research area where scholars seek to understand how different musical elements influence listeners' feelings.

My study focuses on classifying music sentiments using Deep Learning. To implement my proposed idea, I used some pre-trained models such as ResNet50, DenseNet121, and Inception V3 on two audio datasets after converting them to spectrogram images. The audio dataset was divided into ten-second segments, and two types of spectrogram transformations were used: classic spectrogram and Mel spectrogram.

The outcomes demonstrated that the ResNet50 model achieved the highest performance on both datasets, although DenseNet121 had a higher accuracy with the Mel spectrogram using one of the data balancing techniques on the second dataset, but the performance of the ResNet50 model is still higher in terms of inference and generalization task. It also achieved the highest accuracy on the first dataset. This study provides insight into how Deep Learning techniques can be applied to analyzing and classifying musical emotions, contributing to a better understanding of music and emotions.

Keywords: Music emotion recognition (MER), Spectrogram analysis, pre-trained models, Convolutional neural networks (CNN), Transfer learning, Deep learning.

ÖZET

ÖNCE DEN EĞİTİLMİŞ MODELLERLE MÜZİK DUYGULARININ SINIFLANDIRILMASI

TAQA, ZAINAB

Bilgi Teknolojileri Yüksek Lisans

Danışman: Prof. Dr. ABDÜL KADIR GÖRÜR

Ağustos 2024, 135 sayfa

Müzik tüm sanatların içinde en duygusal olanıdır ve bu nedenle de insanlara en yakın olan ve en sevilenidir. Duygular düşünme, davranış ve karar verme biçimlerimizi etkileyen, insan yaşamının merkezi bir parçasıdır. Temel duygularımızı mutluluk, üzüntü, korku, öfke, şaşkınlık ve iğrenme olmak üzere altı tür olarak sınıflandıran Paul Ekman da dahil olmak üzere pek çok bilim insanı duygular üzerinde araştırmalar yapmıştır. Bu sınıflandırma, duyguların davranışlarımızı ve günlük etkileşimlerimizi nasıl etkilediğini anlamamız açısından yardımcı olmaktadır.

Duygu analizi sağlık, eğitim, iş, pazarlama, teknoloji ve daha birçok alanda faydalı olabilir. Aşağıdakiler de dahil olmak üzere çeşitli uygulamalar duygu çıkarımını ve analizini kolaylaştırır:

- **Metin Analizi:** Doğal Dil İşleme (NLP) duygu çıkarımı yapmak için yazılı metinleri analiz eder.
- **Ses Analizi:** Beyond Verbal ve Affectiva gibi uygulamalar, duyguları çıkarmak için ses tonunu analiz eder, bu da müşteri hizmetleri merkezlerinde yararlıdır.
- **Görüntü ve Video Analizi:** Bilgisayarla Görüntüleme teknikleri, görüntü ve videolardan duyguları tespit etmek için yüz ifadelerini analiz eder.

Müzik, melodileri, ritimleri ve tonları ile çok çeşitli duyguları tetikleyerek insan duygularını harekete geçirme ve ifade etme konusunda özel bir yere sahiptir.

Müzikal duyguların analizi, araştırmacıların farklı müzikal unsurların dinleyicilerin duygularını nasıl etkilediğini anlamaya çalıştıkları ve giderek büyümekte olan bir araştırma alanıdır.,

Araştırmam, derin öğrenmeyi kullanarak müziksel duyguları sınıflandırmaya odaklanmaktadır. Öne sürdüğüm düşüncüyü gerçekleştirmek için iki ses veri kümesini spektrogram görüntülerine dönüştürdükten sonra ResNet-50, DenseNet-121 ve Inception-v3 gibi ön-eğitime sahip bazı modelleri bu görüntülerin üzerinde geliştirdim. Ses veri kümeleri on saniyelik bölümlere ayrıldı ve iki çeşit spektrogram dönüşümü kullanıldı: klasik spektrogram ve Mel spektrogramı.

Sonuçlar, her iki veri kümesinde de ResNet50 modelinin en yüksek performansı elde ettiğini, ancak DenseNet121'in ikinci veri kümesinde veri dengeleme tekniklerinden birini kullanarak Mel spektrogramı ile daha yüksek bir doğruluğa sahip olduğunu, ancak yine de ResNet-50 modelinin performansının çıkarım ve genelleme görevi açısından daha yüksek olduğunu göstermiştir. ResNet-50 aynı zamanda ilk veri kümesinde de en yüksek doğruluğu elde etmiştir. Bu araştırma, derin öğrenme tekniklerinin müzikal duyguları analiz etmek ve sınıflandırmak için nasıl kullanılabileceğine dair bir fikir sağlayarak müzik ve duyguların daha iyi anlaşılmasına katkıda bulunmaktadır.

Anahtar Kelimeler: Müzik duygu tanıma, Spektrogram analizi, Önceden eğitilmiş modeller, Konvolüsyonel sinir ağları, Transfer öğrenimi, Derin öğrenme.

ACKNOWLEDGEMENT

First and foremost, I want to express my deepest gratitude to all my teachers, especially my supervisor, Assist Prof. Abdül Kadir GÖRÜR for his helpful advice and guidance given throughout the program work. Special thanks to Mr. Hassan A. Taha/piano Center Institute - Mosul, who taught me the alphabet of music code. Special thanks to Mr. Kareem for his unwavering support, guidance, and encouragement. Thank you, my family, mom, sisters, and brothers especially my sister Alla for everything. I also want to extend a special thank you to my husband Ahmad who never stopped supporting and encouraging me throughout my study period. My heartfelt thanks go to my friends Balsam, Huda, Khadijah, Noor, Marwa, and Ohood for their backing.

TABLE OF CONTENTS

STATEMENT OF NONPLAGIARISM	III
ABSTRACT	IV
ÖZET.....	VI
ACKNOWLEDGEMENT.....	VIII
LIST OF TABLES	XIII
LIST OF FIGURES	XIV
LIST OF ABBREVIATIONS	XIX
CHAPTER I.....	1
INTRODUCTION.....	1
1.1 MUSIC WITH EMOTION	1
1.2 BACKGROUND	1
1.3 PSYCHOLOGY	2
1.4 MUSICOLOGY	2
1.5 ARTIFICIAL INTELLIGENCE (AI).....	3
1.6 PROBLEM STATEMENT	3
1.7 RESEARCH AIM.....	3
1.8 STRUCTURE OF THE THESIS.....	4
CHAPTER II	6
LITERATURE REVIEW.....	6
2.1 FEATURE EXTRACTION FOR MUSIC EMOTION RECOGNITION.....	6
2.2 APPROACHES TO MUSIC EMOTION CLASSIFICATION.....	7
2.3 RELATED STUDIES	8
CHAPTER III	17
SPECTROGRAMS AND PRE-TRAINED MODELS FOR MUSIC EMOTION RECOGNITION	17
3.1 SPECTROGRAMS.....	17

3.1.1	Introduction to Spectrograms	17
3.1.2	Types of Spectrograms	18
3.1.3	Standard spectrogram vs Mel spectrogram	18
3.2	PRE-TRAINED MODELS	19
3.2.1	Overview of Transfer Learning and its Benefits	19
3.2.1.1	What is Transfer Learning?	19
3.2.1.2	The Idea behind Pre-trained Models:	19
3.2.1.3	The Benefits of Pre-trained Models:	20
3.2.2	Overview of Deep Learning and Models Used in this Study	21
3.2.2.1	What is DL?	21
3.2.2.2	Layers existing in the pre-trained models and its importance....	22
3.2.2.3	Models.....	26
3.3	PRE-TRAINED MODELS ARCHITECTURE	27
3.3.1	ResNet50	27
3.3.2	DenseNet-121	30
3.3.3	Inception v3	32
3.3.3.1	Overview of Inception V3:	32
3.3.3.2	The architecture of the Inception V3:	34
3.4	Comparison of Model Architectures.....	35
CHAPTER IV.....		37
METHODOLOGY.....		37
4.1	AUDIO PROCESSING AND AUDIO FILES	37
4.1.1	Audio Files	37
4.1.1.1	Uncompressed Audio Files	37
4.1.1.2	Compressed Audio Files	38
4.1.2	Audio Technology	38
4.1.3	Audio Signals	38
4.1.4	Audio Applications.....	39
4.2	MUSIC AND EMOTIONS.....	39
4.2.1	Music and Their Relationship to Emotions	39
4.2.2	Characteristics of Musical Emotions.....	41

4.2.3	Categorical Classification of Emotions	41
4.3	DATASETS	42
4.3.1	Turkish Music Emotion Dataset	42
4.3.2	MER500 Dataset	43
4.4	DATA PREPROCESSING.....	44
4.4.1	First Stage Preprocessing	44
4.4.2	Second Stage Preprocessing: Conversion and Standardization.....	45
4.5	DATA BALANCING TECHNIQUES.....	47
4.5.1	Under-sampling	47
4.5.2	Oversampling	48
4.6	FEATURE EXTRACTION TECHNIQUES	49
4.7	MODEL ARCHITECTURE	55
4.8	MODEL TRAINING PROCESS.....	55
4.8.1	Training process on Turkish Music Emotion dataset	55
4.8.2	Training process on MER500 dataset.....	56
4.9	EVALUATION METRICS	57
4.9.1	Accuracy.....	57
4.9.2	Confusion matrix	57
4.9.3	Classification Report	58
4.9.4	Recall.....	58
4.9.5	Precision	58
4.9.6	F1 Score.....	58
CHAPTER V		59
EXPERIMENTS AND RESULTS		59
5.1	EXPERIMENTS WITH TURKISH MUSIC EMOTION DATASET	59
5.1.1	Results with Standard Spectrogram.....	59
5.1.2	Results with Mel-Spectrogram	64
5.1.3	Comparison of Models' Performance	66
5.2	EXPERIMENTS WITH MER500 DATASET	69
5.2.1	Results with Standard Spectrogram.....	69
5.2.1.1	Unbalanced Data	69
5.2.1.2	Undersampled Data	73

5.2.1.3	Oversampled Data.....	75
5.2.2	Results with Mel-Spectrogram.....	77
5.2.2.1	Unbalanced Data	77
5.2.2.2	Under-sampled Data	79
5.2.2.3	Oversampled Data.....	81
5.2.3	Comparison of Models' Performance	83
5.3	ANALYSIS OF RESULTS	88
5.3.1	Comparison between Datasets in first stage and in the second one: ..	88
5.3.2	Comparison with previous studies:	90
5.4	SUMMARY OF FINDINGS	92
CHAPTER VI.....		93
CONCLUSION AND FUTURE WORK		93
6.1	CONCLUDING REMARKS.....	93
6.2	LIMITATIONS OF THE STUDY.....	97
6.3	FUTURE RESEARCH DIRECTIONS	98
REFERENCES.....		99
APPENDICES		109
APPENDIX 1: TO MITIGATE OVERFITTING IN MY MODELS, I HAVE IMPLEMENTED SEVERAL STRATEGIES		109

LIST OF TABLES

Table 1: Models comparison	36
Table 2: Comparative Analysis of Emotional Categories in Music.....	41
Table 3: The MET audio files durations	43
Table 4: The MER-500 audio files durations.....	44
Table 5: MET Audio files after trimmed & standardizing their duration.	46
Table 6: MER500 Audio files after trimmed & standardizing their duration (unbalance).....	47
Table 7: the audio files after under-sampling.....	48
Table 8: the audio files after over sampling.....	48
Table 9: Test Accuracy & Mean Accuracy with Standard-Spectrogram.....	66
Table 10: Test Accuracy & Mean Accuracy with Mel-Spectrogram.....	66
Table 11: The total accuracy (Mel- spectrogram and standard spectrogram).....	83
Table 12: The highest accuracy of the first model	89
Table 13: The highest accuracy of the second model	90
Table 14: The comparing with previous studies MER500 dataset.....	91
Table 15: The comparing with previous studies for MET dataset.	92

LIST OF FIGURES

Figure 1: The elements of the music	2
Figure 2: Exploring music emotion classification using AI and neural networks	4
Figure 3: Emotions Behind the Melody	6
Figure 4: Russell’s circumplex model.	8
Figure 5: Working principles of the proposed.	8
Figure 6: Multifeatured combined network classifier based on.....	9
Figure 7: Block diagram of the proposed convolutional neural network.	10
Figure 8: System architecture of the experimental method	11
Figure 9: The flow chart of DBN and SVM combined algorithm to achieve classification.....	12
Figure 10: Flowchart of the Emotion Recognition Pipeline	13
Figure 11: ResNet50 Model Accuracy	13
Figure 12: ResNet50_trust structure	14
Figure 13: Experimental process.....	15
Figure 14: The proposed neural network	16
Figure 15: The overall structure of CNN–BiGRU–Attention.....	16
Figure 16: Standard Spectrogram	19
Figure 17: Mel Spectrogram	19
Figure 18: Comparison of different models.....	21
Figure 19: The layers of the convolution neural network in general (CNN).....	22
Figure 20: The operation of convolution layers.....	23
Figure 21: The operation of the max pooling layer.....	25
Figure 22: The operation of the fully connected layer.....	25
Figure 23: The architecture of ResNet50	28
Figure 24: Describe the method of residual block.	30
Figure 25: The architecture of DensNet-121.	32

Figure 26: Inception-v3.....	34
Figure 27: The architecture of the Inception_v3.....	35
Figure 28: MET Original Audio	45
Figure 29: MET Preprocessed Audio file	46
Figure 30: Standard Spectrogram vs. Mel Spectrogram (MET Happy class)	50
Figure 31: Standard Spectrogram vs. Mel Spectrogram (MET Sad class)	50
Figure 32: Standard Spectrogram vs. Mel Spectrogram (MET Angry class).....	51
Figure 33: Standard Spectrogram vs. Mel Spectrogram (MET Relax class).....	51
Figure 34: Original Spectrogram on left vs Resized Spectrogram (224*224*3) on right	52
Figure 35: Standard Spectrogram vs. Mel Spectrogram (MER500 Happy class)	52
Figure 36: Standard Spectrogram vs. Mel Spectrogram (MER500 Sad class).....	53
Figure 37: Standard Spectrogram vs. Mel Spectrogram (MER500 Romantic class)53	
Figure 38: Standard vs. Mel Spectrogram (MER500 Devotional class).....	53
Figure 39: Standard Spectrogram vs. Mel Spectrogram (MER500 Party class)	54
Figure 40: Resized Mel-Spectrogram images.....	54
Figure 41: Resized Standard-Spectrogram images.	54
Figure 42: CM formula	57
Figure 43: Confusion Matrix - MET ResNet50 Standard-Spectrogram.....	59
Figure 44: Classification report - MET ResNet50 on Standard-Spectrogram.....	60
Figure 45: Confusion Matrix - MET DenseNet121 Standard-Spectrogram.....	60
Figure 46: Classification report - MET DenseNet121 on Standard-Spectrogram....	61
Figure 47: Confusion Matrix - MET Inception V3 on Standard Spectrogram.....	61
Figure 48: Classification report - MET Inception V3 on Standard-Spectrogram.....	62
Figure 49: Confusion Matrix- MET ResNet50 on Standard Spectrogram.	62
Figure 50: Confusion Matrix - MET DenseNet121 on Standard Spectrogram.	63
Figure 51: Confusion Matrix- MET Inception V3 on Standard Spectrogram.....	63
Figure 52: Confusion Matrix - MET ResNet50 on Mel-Spectrogram.....	64
Figure 53: Confusion Matrix - MET DenseNet121 on Mel-Spectrogram.....	65
Figure 54: Confusion Matrix - MET Inception V3 on Mel-Spectrogram.....	65
Figure 55: Pre-trained Models on MET with Mel-Spectrogram bar plot.	66
Figure 56: Classification Report - MET ResNet50 on Mel-Spectrogram.	67

Figure 57: MET Best Model ResNet50 with Mel-Spectrogram fold-1.	67
Figure 58: MET Best Model ResNet50 with Mel-Spectrogram fold-2.	67
Figure 59: MET Best Model ResNet50 with Mel-Spectrogram fold-3.	68
Figure 60: MET Best Model ResNet50 with Mel-Spectrogram fold-4.	68
Figure 61: MET Best Model ResNet50 with Mel-Spectrogram fold-5.	68
Figure 62: Confusion Matrix - MER500 ResNet50 on Standard-Spectrogram.....	69
Figure 63: Classification Report - MER500 ResNet50 on Standard-Spectrogram. .	69
Figure 64: Confusion Matrix - MER500 DenseNet121 on Standard Spectrogram..	70
Figure 65: Classification Report - MER500 DenseNet121 on Standard Spectrogram.	70
Figure 66: Confusion Matrix -MER500 Inception V3 on Standard Spectrogram....	71
Figure 67: Classification report - MER500 Inception V3 on Standard Spectrogram	71
Figure 68: Confusion Matrix - MER500 ResNet50 (unbalance - Standard Spectrogram).....	72
Figure 69: Confusion Matrix - MER500 DenseNet121 (unbalance - Standard Spectrogram).....	72
Figure 70: Confusion Matrix - MER500 Inception V3.....	73
Figure 71: Confusion Matrix - MER500 ResNet50.....	74
Figure 72: MER500 DenseNet121 (undersampling - Standard Spectrogram)	74
Figure 73: Confusion Matrix - MER500 Inception V3.....	75
Figure 74: Confusion Matrix - MER500 ResNet50.....	76
Figure 75: Confusion Matrix - MER500 DenseNet121(oversampling - Standard Spectrogram).....	76
Figure 76: Confusion Matrix -MER500 Inception V3.....	77
Figure 77: Confusion Matrix - MER500 ResNet50 (unbalance - Mel Spectrogram)	78
Figure 78: Confusion Matrix -MER500 DenseNet121 (unbalance - Mel Spectrogram)	78
Figure 79: Confusion Matrix - MER500 Inception V3.....	79
Figure 80: Confusion Matrix -MER500 ResNet50.....	80
Figure 81: Confusion Matrix -MER500 DenseNet121 (undersampling - Mel Spectrogram).....	80

Figure 82: Confusion Matrix -MER500 Inception V3.....	81
Figure 83: Confusion Matrix -MER500 ResNet50 (oversampling - Mel Spectrogram)	82
Figure 84: Confusion Matrix -MER500 DenseNet121 (oversampling - Mel Spectrogram).....	82
Figure 85: Confusion Matrix - MER500 Inception V3 (oversampling - Mel Spectrogram).....	83
Figure 86: MER500 Best Model DenseNet121 (oversampling-standard spectrogram) fold-1.....	84
Figure 87: MER500 Best Model DenseNet121 (oversampling - standard spectrogram) fold-2.....	84
Figure 88: MER500 Best Model DenseNet121 (oversampling - standard spectrogram) fold-3.....	84
Figure 89: Pre-trained Models on MER500 (oversampling - standard spectrogram) bar plot.	85
Figure 90: Classification report - MER500 DenseNet121.....	85
Figure 91: Confusion Matrix - MER500 DenseNet121 cm.....	86
Figure 92: MER500 Best Model DenseNet121 (oversampling - Mel spectrogram)	86
Figure 93: MER500 Best Model DenseNet121 (oversampling - Mel spectrogram) fold-2.....	87
Figure 94: MER500 Best Model DenseNet121 (oversampling - Mel spectrogram) fold-3.....	87
Figure 95: Pre-trained Models on MER500 (oversampling - Mel spectrogram) bar plot.	87
Figure 96: Classification report - MER500 DenseNet121.....	88
Figure 97: Confusion Matrix - MER500 DenseNet121 cm.....	88
Figure 98: Inference Samples for ResNet50 with unbalance & Mel-spectrogram..	94
Figure 99: Inference Samples for ResNet50 with unbalance & Mel-spectrogram continued.....	94
Figure 100: Inference Samples for ResNet50 with unbalance & Mel-spectrogram continued.....	95

Figure 101: Inference Samples for ResNet50 with unbalance & Mel-spectrogram continued.....	95
Figure 102: Inference Samples for DenseNet121 with oversample standard-spectrogram.....	96
Figure 103: Inference Samples for DenseNet121 with oversample standard-spectrogram continued.	96
Figure 104: Inference Samples for DenseNet121 with oversample standard-spectrogram continued.	96
Figure 105: Inference Samples for DenseNet121 with oversample standard-spectrogram continued.	96
Figure 106: Model architecture.....	109
Figure 107: five folds for the Turkish music emotion dataset (First Stage)	110
Figure 108: five folds for the Turkish music emotion dataset (Second Stage)	110
Figure 109: Five folds for the MER500 dataset (First Stage).....	110
Figure 110: Three folds for the MER500 dataset (Second Stage)	110
Figure 111: Early Stopping	111
Figure 112: A sample of overfitting.....	111
Figure 113: ResNet50+Mel-Spectrogram on MET fold-1.....	112
Figure 114: ResNet50+Mel-Spectrogram on MET fold-2.....	112
Figure 115: ResNet50+Mel-Spectrogram on MET fold-3.....	112
Figure 116: ResNet50+Mel-Spectrogram on MET fold-4.....	113
Figure 117: ResNet50+Mel-Spectrogram on MET fold-5.....	113
Figure 118: MER500 Best Model DenseNet121 (Oversampling - Mel spectrogram) fold-1	113
Figure 119: MER500 Best Model DenseNet121 (Oversampling - Mel spectrogram) fold-2	114
Figure 120: MER500 Best Model DenseNet121 (Oversampling - Mel spectrogram) fold-3.....	114
Figure 121: Inference Function.....	115
Figure 122: Inference Sample on best model.....	115
Figure 123: Inference Sample on best model.....	115

LIST OF ABBREVIATIONS

MER	: Music Emotion Recognition
MIR	: Music Information Retrieval
AI	: Artificial Intelligence
ML	: Machine Learning
DL	: Deep Learning
DNN	: Deep Neural Network
CNNs	: Convolutional Neural Networks
CNN-LSTM	: Convolutional Neural Networks and Long Short-Term Memory
STFT	: Short Term Fourier Transform
SVM	: Support Vector Machine
RF	: Random Forests
DBN	: Deep Belief Network
MFDA	: Multifractal Detrended Fluctuation Analysis
DTL	: Deep Transfer Learning
ANN	: Artificial Neural Network
RNN	: Recurrent Neural Network
ReLU	: Rectified Linear Unit
TanH	: Hyperbolic Tangent
WAV	: Waveform
MP3	: MPEG-1 III
AIFF	: Audio Interchange File Format
FLAC	: Free Lossless Audio Codec
ALAC	: Apple Lossless Audio Codec
AAC	: Advanced Audio Codec
PCM	: Pulse Code Modulation

CHAPTER I

INTRODUCTION

1.1 MUSIC WITH EMOTION

Since ancient times, music has been one of the most ultimate sources of happiness for many individuals in their free time. Music is a vital aspect of our daily lives. The importance of music rises as we enter the digital age. In truth, music is more than just a source of entertainment; it is also a simple way for people to communicate ideas, a mood-measurement tool, and a repository for emotions and memories. Music has long been intertwined with human emotions. Despite the centuries-long debate about the actual relationship between music and emotion, it is undoubtedly regarded as a way for a human being to express his or her happiness, sadness, enthusiasm, joy, fear, surprise, or anger. Music has the extraordinary power to stir our emotions, feed our souls, and give rise to inspiration in ways that words often cannot. The melodies and the rhythms of music resound within us, accommodating emotions, memories, and connections that reach as far as others and even our customized experiences [1].

1.2 BACKGROUND

In recent years, music has witnessed the interest of researchers to understand and analyze how music affects the psyche. This is where Music Emotion Recognition, hereinafter referred to as MER, begins. It represents a subfield of Music Information Retrieval, MIR [2]. It is at the intersection of three disciplines: Psychology, Musicology, Artificial Intelligence, AI; and it seeks to understand the relationship between music and human emotions.

1.3 PSYCHOLOGY

Our emotional response to music is one of the core parts of MER. Researchers and psychologists confirm that music provokes many complex emotions, while emotions themselves are always fuzzy concepts [3].

Over the past few decades, music and emotion have been an area of sustained interest in music psychology literature [4]. Emotional responses to music are different from person to person, influenced by several factors, including cultural background [5], the musical experience of the individual [6], and the feelings that his memories brought up [7].

1.4 MUSICOLOGY

Musicology is another core part of MER. Music has eight basic elements. These are Tonality, Dynamics, Timbre, Form, Texture, Harmony, Rhythm, and Melody. Shown below in Figure 1. [8]. These elements form the foundation of musical expression and play a critical role in understanding how music evokes emotions.

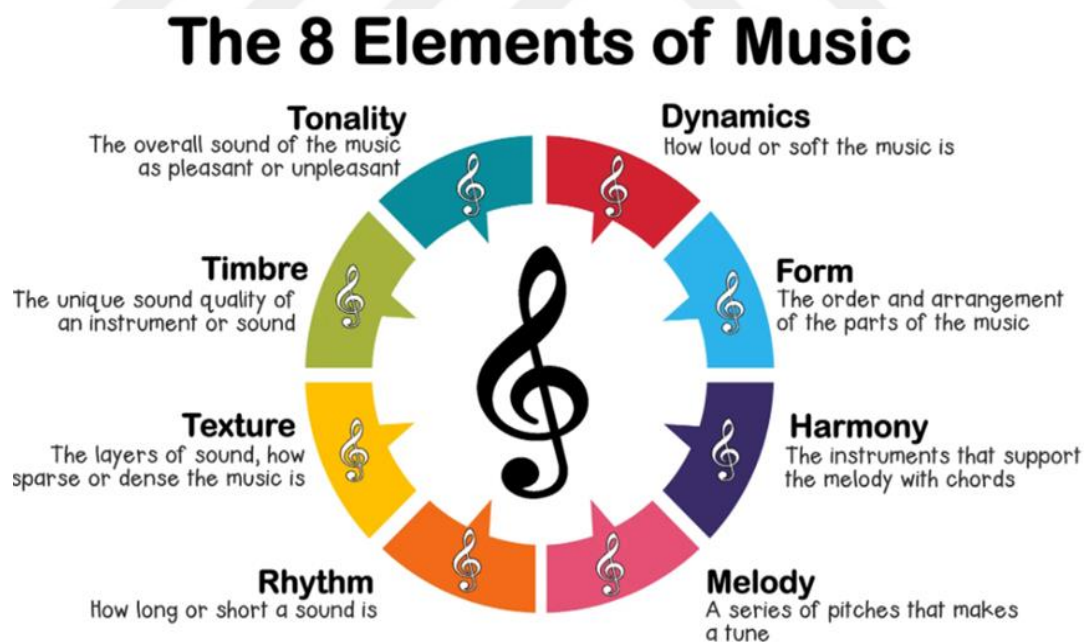


Figure 1: The elements of the music

1.5 ARTIFICIAL INTELLIGENCE (AI)

AI is the third core part of MER and is a sub-branch of computer science that focuses on developing systems capable of performing tasks similar to those done by the human brain, such as expert systems and robotics [9]. AI includes Machine Learning, ML, which deals with learning algorithms for prediction, and Deep Learning, DL which is a subset of ML that uses Deep Neural Networks, DNN to mimic the human brain's analytical abilities [10].

1.6 PROBLEM STATEMENT

Even with the development of DL techniques, and the incredibly high performance of DL algorithms in many applications that people use in daily life nowadays, music emotions classification and analysis remain challenging tasks due to:

- The complexity of the music, due to the overlapping tones, rhythms, and other sounds (including noise), makes the extraction of the right features a significant challenge.
- The variety of opinions and emotional responses expressed by diverse listeners result in significant diversity, making categorization confusing and difficult.
- A lack of properly tagged data for research purposes forces researchers to exert extra effort in collecting accurate and well-labeled data for their studies.

All these reasons make classifying music emotions a challenging task. Therefore, studying various methods was necessary to obtain high classification accuracy and improve performance.

1.7 RESEARCH AIM

The main goal of this research is to enhance the accuracy of music emotion classification to support a wide range of applications. First and foremost, I aim to contribute to the improvement of recommendation systems, that identify emotional aspects of music and suggest tracks matching the listener's mood. By achieving higher classification accuracy through the use of DL and pre-trained Convolutional Neural

Networks, CNNs, I aim to extend the benefits of my work to several other domains, including:

- Educational Tools: Music education tools for students to select songs and understand their emotional impact, and interactive educational content whose emotional tone changes based on user interaction.
- Personal Well-being and Lifestyle Applications: Development of apps that recommend music for mood regulation and generate playlists based on changing emotions throughout one's day.
- Research and Development: Contributing to cognitive and affective computing, as well as behavioral studies, by providing insights into how music influences emotions and behavior.



Figure 2: Exploring music emotion classification using AI and neural networks

1.8 STRUCTURE OF THE THESIS

This thesis contains six chapters:

- Chapter II discusses the literature review of recent studies in classification approaches.
- Chapter III discusses the types of spectrograms and pre-trained model details with their architectures (ResNet50, DenseNet121, Inception V3).

- Chapter IV discusses various methods of different types of spectrograms, Deep Learning, and CNN models for music emotion recognition.
- Chapter V includes the experimental setup, analysis, and results of all attempts using the MER500 and Turkish Music Emotion datasets.
- Chapter VI presents the study's conclusions and our suggestions for future work.



CHAPTER II

LITERATURE REVIEW

DL technology has made significant contributions to improving today's MER taxonomy, especially through pre-trained models, which are DNNs trained on large and diverse datasets. This technology has been highly successful across a wide range of applications. In this chapter, we will detail three main themes: first, the fundamental features on which MER is based; second, various music emotion classification techniques; and third, previous research related to the classification of music emotions, focusing on studies that have employed DL, especially pre-trained models. These state-of-the-art technologies have propelled us into an entirely new dimension of music emotion recognition systems.

2.1 FEATURE EXTRACTION FOR MUSIC EMOTION RECOGNITION

To analyze music emotions, I need to identify the features to extract and later use as input to my network. Music features are extracted from the audio signal, lyrics, and symbolism of the music scores.

These features can be divided into three levels [11]:

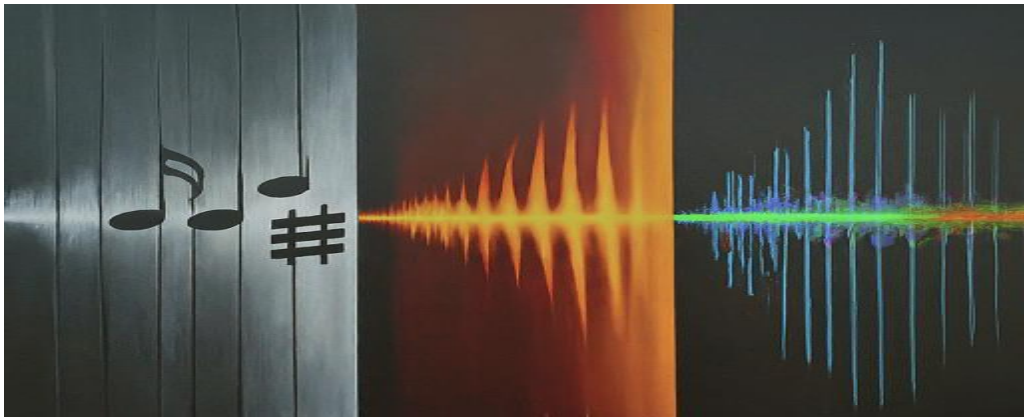


Figure 3: Emotions Behind the Melody

- High-level features: Related to the meaning that the listener can recognize [12], such as rhythm (It is a sequential arrangement of sounds and beats in music so that it follows a certain pattern of repetition and distinction between different tones) [13], timbre (the color of the sounds) [14].
- Middle-level features: Features that link high-level features to low-level ones, reflecting advanced musical concepts without access to specific details of musical structure [11], such as Chord (the simultaneous sounding of three or more musical notes) [12].
- Low-level features: Refers to the basic foundation of audio classification systems, such as spectrograms (which represent the distribution of sound power across frequencies) [11].

2.2 APPROACHES TO MUSIC EMOTION CLASSIFICATION

In this study, we classify music based on AI and DL. Let's take an in-depth look at the common ways of categorizing.. There are three primary approaches to classifying music emotion classification:

- The first type is categorical, which depends on distinct emotions such as happiness, sadness, relaxation, etc. [15]
- The second is dimensional [16], representing emotions in two dimensions, Arousal and Valence, as shown in the Thayer-Russell model in Figure 4 (The circumplex model is developed by James Russell. Adopted Russell's multidimensional complex model of scaling mood, X: Valence, Y: Arousal).
- The third type is appraisal, which depends on the mood of the listeners and the situation [15].

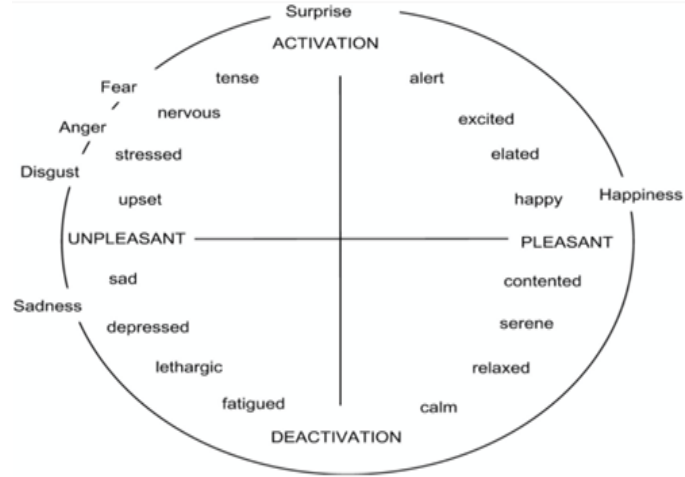


Figure 4: Russell's circumplex model.

2.3 RELATED STUDIES

The article by Er, M. B., & Aydılek, I. B. [18] extracted chroma spectrum features from a Turkish Music Emotion dataset to use as input for pre-trained models. They aimed to use two algorithms: AlexNet and VGG-16. The extracted deep visual features from different layers of AlexNet and VGG-16 were then used to train Support Vector Machine (SVM) and softmax classifiers. The achieved accuracies were 76% for AlexNet and 89.2% for VGG-16.

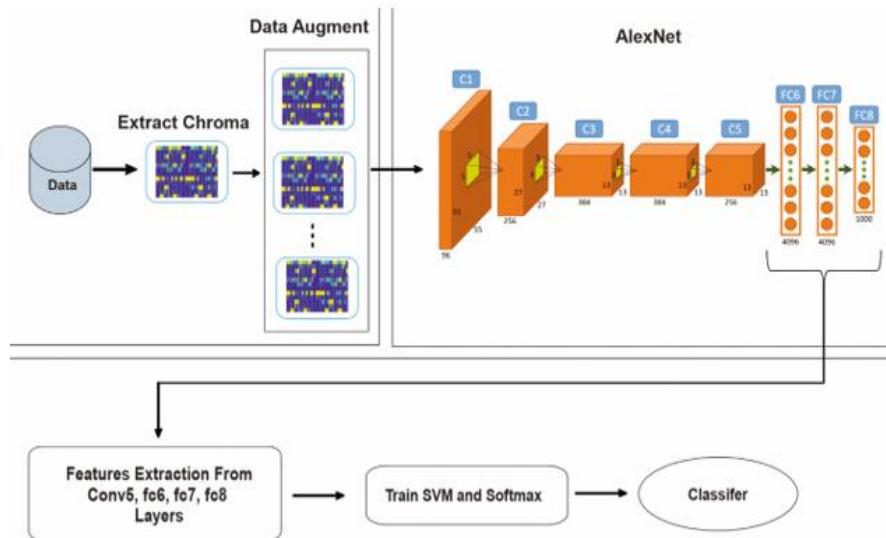


Figure 5: Working principles of the proposed.

In the article by Chen, C., & Li, Q. [19], the authors presented a study that proposed Convolutional Neural Networks with Long Short-Term Memory and DNN. This was done with two kinds of feature input: two-dimensional and one-dimensional for music emotion classification. Particularly, two-dimensional features, representing the spectrogram input through CNN-LSTM, were extracted using STFT (Short Term Fourier Transform), and one-dimensional features, representing the lyrics input, were processed through DNN. The accuracy reached 78% after the combination of those two types of features.

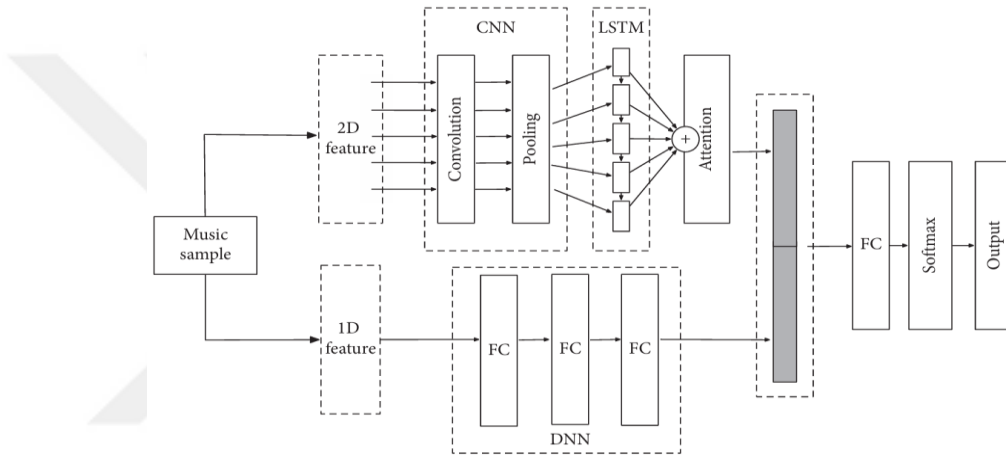


Figure 6: Multifeatured combined network classifier based on CNN-LSTM.

In the article by Sarkar, R., Choudhury, S., Dutta, S., Roy, A., & Saha, S. K. [20], the authors modified the VGGNet's architecture with reduced numbers of layers to prevent overfitting on small training datasets. Round benchmark datasets such as Soundtracks, Bi-Modal, and MER_taffc were used with a CNN to extract spectrograms, literally capturing time-frequency characteristics of music and obviating the need for handcrafted features. A novel post-processing method combining voting with run-length encoding was proposed to sustain an accurate description of the overall emotional category from a music clip based on its fragmented segments. This DL approach outperformed the traditional methods that relied on neural networks, SVM, and Random Forests, RF with handcrafted features. The results showed that the

methodology proposed was robust and could help improve music emotion recognition systems.

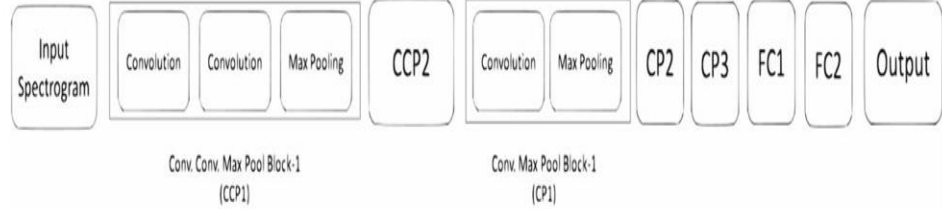


Figure 7: Block diagram of the proposed convolutional neural network.

In another study by Valentina Franzoni¹ et al. [21], the authors of this paper offer a strategy for generating sound spectrograms using fixed-length segments of original audio samples derived from real audio clips recorded at high-attendance events, where the crowd functions as a collective individual. A convolutional neural network that has been pre-trained on low-level features using the well-known ImageNet large dataset of visual information is employed to test transfer learning approaches. To generate an accurate spectrogram and refine the domain-specific features, the original sound clips are filtered and normalized in amplitude. The suggested model for categorizing the crowd's emotions performs promisingly in experiments conducted on the Convolutional Neural Network that has ultimately been trained.

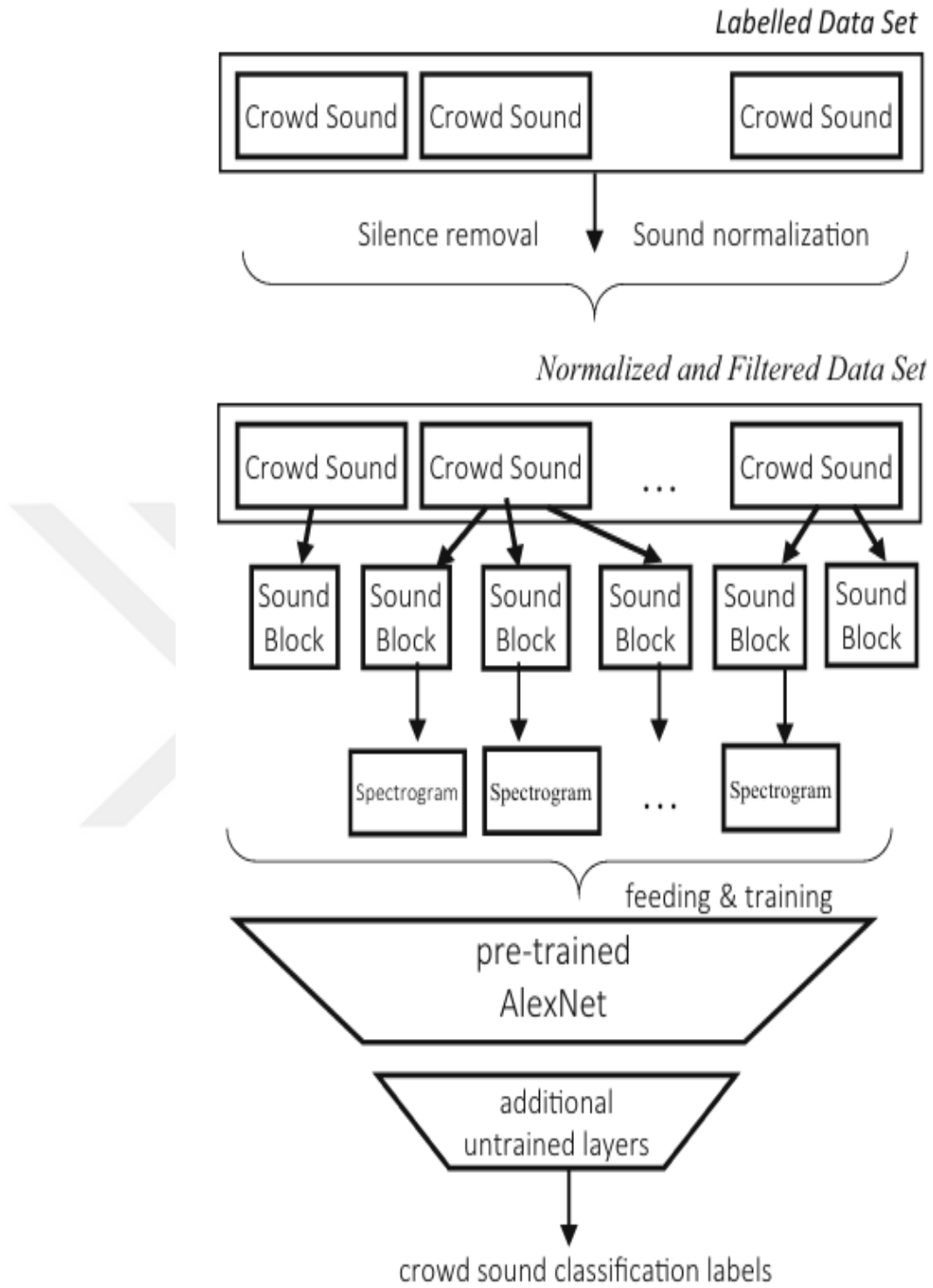


Figure 8: System architecture of the experimental method

The attempt made by the author Wu, Z. [22] in this paper was to recognize and classify the emotions in folk music. Taking features from the last layer in the Deep Belief Network, DBN acted as input to a classifier model—SVM. The dataset had four

classes for emotion: Happy, Woeful, Eager, and Peaceful. The attained accuracy for classification was 79.4%.

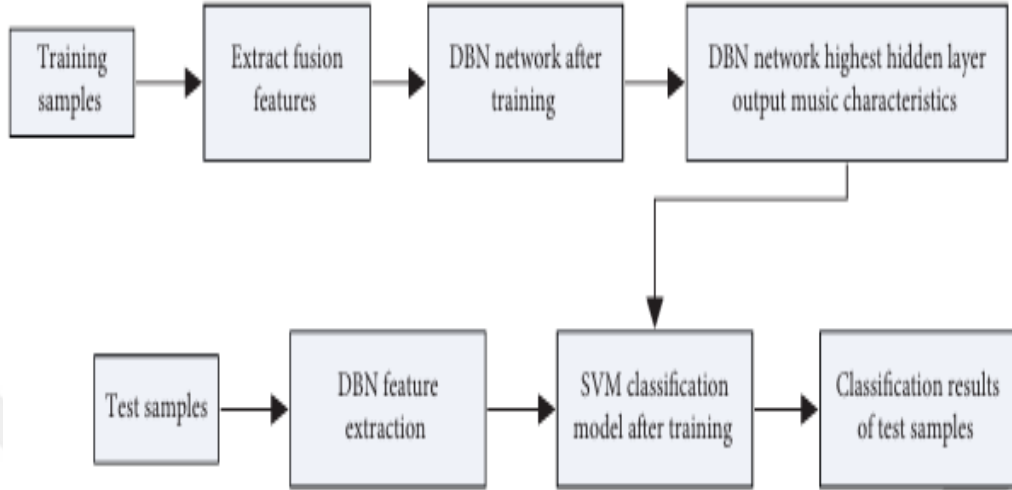


Figure 9: The flow chart of DBN and SVM combined algorithm to achieve classification.

In this study by Sayan Nag et al. [23], the authors introduced a novel dataset called JUMusEmoDB dataset, which contains 1600 audio clips of classical Indian music with four emotion categories: happy, sad, calm, and anxiety, each clip has a duration of 30 seconds. The study used three pre-trained models to analyze its data, namely "ResNet50", "MobileNet v2.0" and "SqueezeNet v1.0". They created a model named "ODE-Net" and trained it from scratch to enhance the accuracy of the classification and compare it with other pre-trained models. Additionally, emotion analysis and musical instrument extraction were performed using the Multifractal Detrended Fluctuation Analysis, MFDFA technique. They extracted the spectrogram using a CNN and Transform Fast Fourier and then converted the images to use as input to pre-trained models and the new model. The accuracy was: SqueezeNetV1 = 87.969%, MobileNetV2= 85.078%, ResNet50 = 91.250%, Proposed ODE- Net =92.969%.

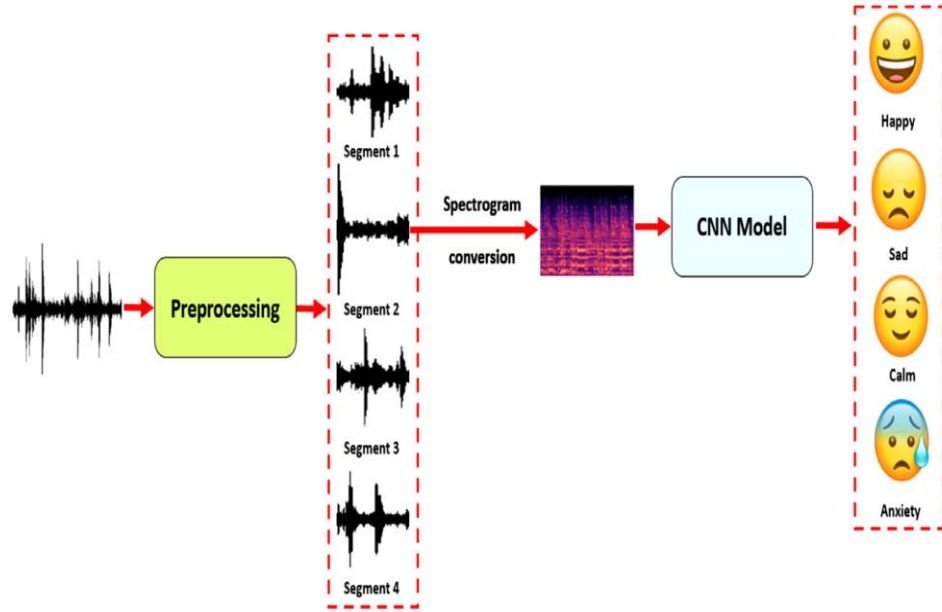


Figure 10: Flowchart of the Emotion Recognition Pipeline

The authors of this paper [24], used three datasets: FMA, GTZAN, and EMA. The dataset was designed to capture four emotions: angry, happy, calm, and sad. They implemented five DL models: DNN, VGGNet16, ResNet50, DenseNet121, NASNetMobile, and MobileNetV2. They extracted spectrograms as input and claimed that ResNet50 performed better as compared to the other models. They reported results in accuracy as shown in Figure 11:

DNN model	FMA	GTZAN_4	EMA	GTZAN_10
ResNet50	80.14	81.09	77.03	66.37
ResNet50_trust	81.33	83.58	79.45	71.56

Figure 11: ResNet50 Model Accuracy

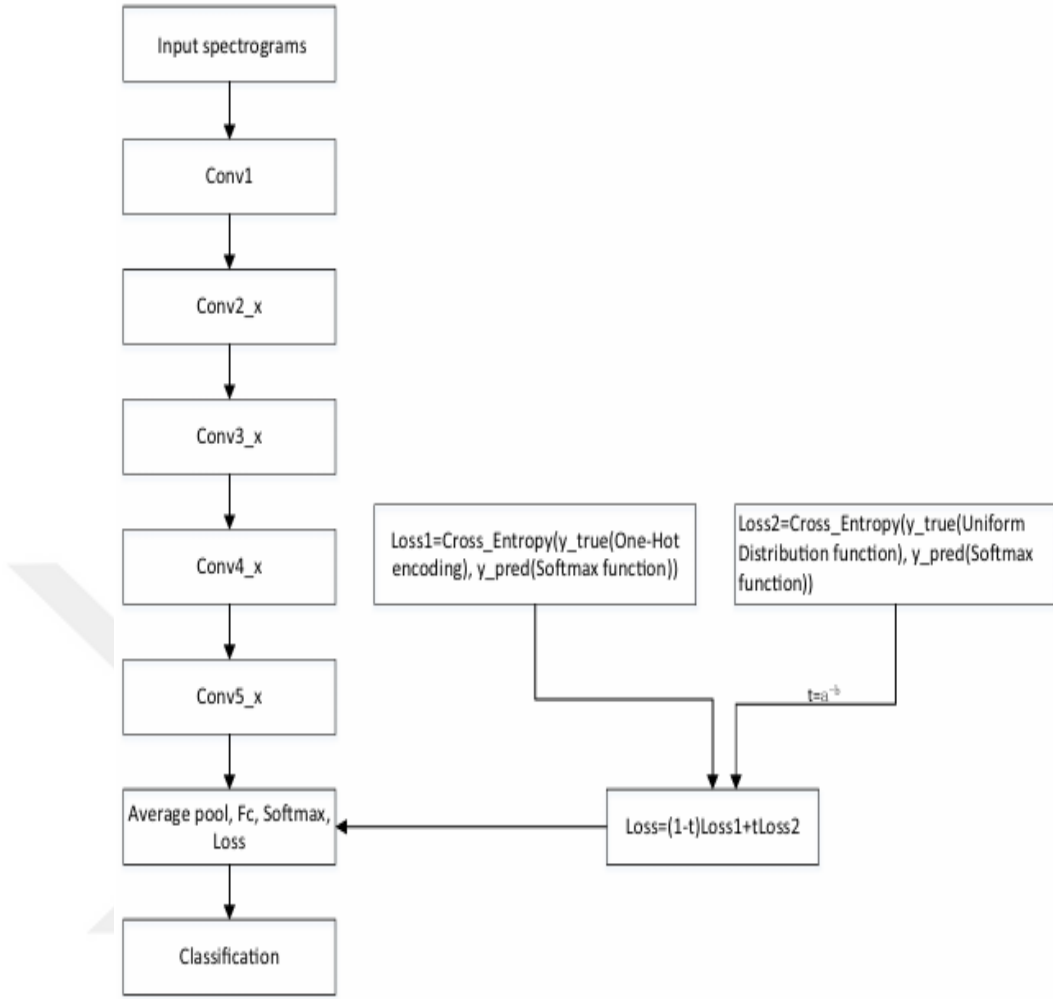


Figure 12: ResNet50_trust structure

The aim of this work done by Jingxian Li et al. [25], was to improve the accuracy of emotion classification from music. In this respect, the authors applied the spectrum with CNN and found that the loss function commonly used in this task—SoftMax loss—is inefficient in pulling apart the deep details of the music. With a view to attaining this objective, the authors constructed a new method where the angular margin and cosine margin were combined with weight and adjective normalization. They also added an angle parameter along with cosine parameters that could assist in allowing differentiation across attributes. They used four datasets: GTZAN_10 with an accuracy of 74.5%, FMA_4 with an accuracy of 85.48%, MER500_2 with two classes—happy and sad—with an accuracy of 83%, and MER500_400 with four classes of devotional, happy, sad, and romance tracks, with an accuracy of 73.98%.

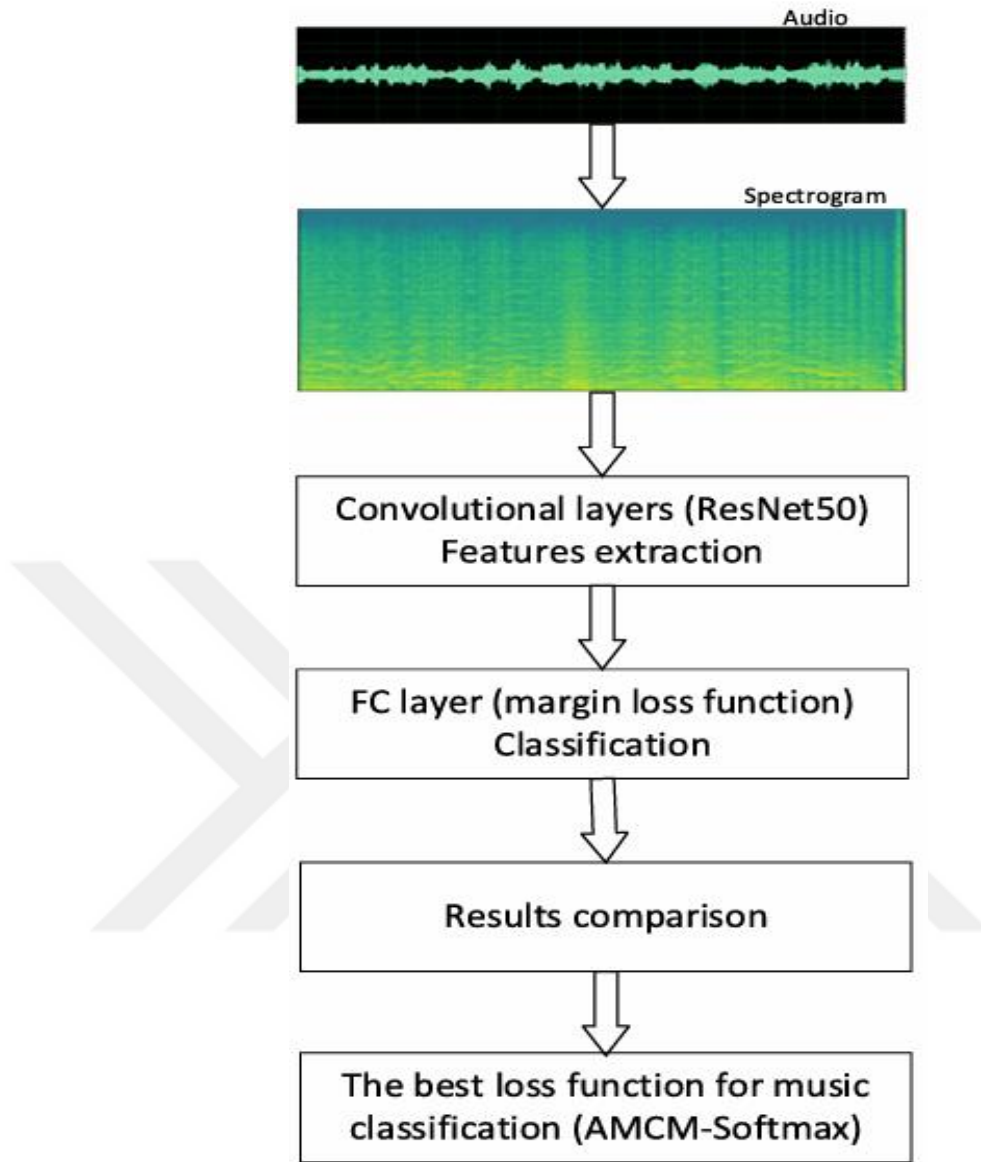


Figure 13: Experimental process

In this paper by Xiao Han et al. [26], he pointed out another way of music emotion classification. He used as input the logarithmic STFT spectral eigenvector matrix in the time domain and worked on a soundtrack dataset. An improved version of the model he developed has a 1-D CNN enacting an optimized inception-GRU residual structure. The accuracy reached 84.27%.

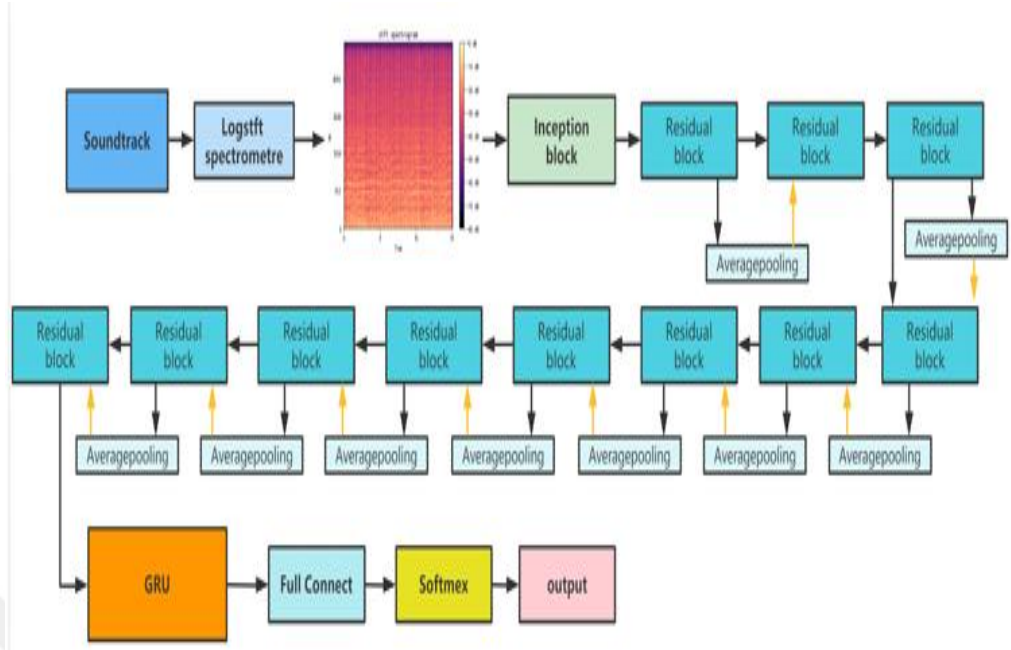


Figure 14: The proposed neural network

This study, by Sihan Ma and Ruohua Zhou [27], mainly focused on the classification of musical feelings only for a violin. They constructed the first single acoustic dataset for the violin, which was named VioMusic, and made use of the model based on the CNN-BIGRU-Attention model. On the part of the violin, rich emotional musical information can be obtained. It is equipped with a CNN to extract raw features from the audio signal itself at the upper layers of the audio spectrum, while a BIGRU attends to text contexts and temporal contexts: an attention mechanism is then added to focus on the important parts in terms of audio performance.

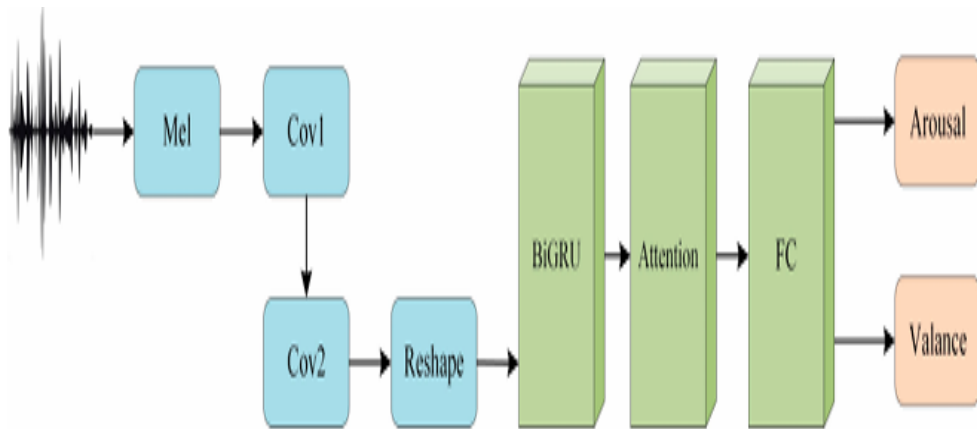


Figure 15: The overall structure of CNN-BiGRU-Attention

CHAPTER III

SPECTROGRAMS AND PRE-TRAINED MODELS FOR MUSIC EMOTION RECOGNITION

In this chapter, we explain in detail what spectrograms are and the types used in MER. We also briefly describe the architecture of some very important Pre-trained models, especially ResNet50, DenseNet121, and Inception V3 models, applied in our research. These models demonstrate their importance in exploiting DL techniques for effective musical emotion classification.

3.1 SPECTROGRAMS

3.1.1 Introduction to Spectrograms

Basically, the spectrogram (sound spectrum) is an optical representation of acoustics, showing the change of frequencies over time, it is generated by calculating the STFT of the audio signal, from which it produces a 2D matrix, time is its horizontal axis and frequency is its vertical axis, while the intensity of color or the level of illumination in the image represents energy (frequency strength) at each point in time [28].

The spectrum is considered one of the low-level features in signal processing [29], and this type of characteristic in music analysis refers to the basic information that can be extracted without any processors provided as I mentioned before, the spectrum means, the frequency, the time, and the intensity, and what distinguishes this information that: first, it's accurate because it's taken directly from the audio signal, second their flexibility as the spectrum is used as a basis for extracting advanced features such as chromatic features and temporal features.

Spectrum images are considered as an efficient tool in Music Analysis for the fact that spectrogram could analyze a wide range of signals. this technique can

1. Analyzing the frequency distribution in a music clip, which in turn helps to detect several things, such as identifying musical instruments and harmonic patterns [30].
2. Rhythm analysis, where it reveals rhythmic patterns and their repetition, as well as revealing distinctive segments [31].

3.1.2 Types of Spectrograms

There are several types of spectrograms [32], including:

- Linear Spectrogram: It is the traditional spectrogram, representing frequency over time.
- Mel spectrogram (nonlinear): It is essentially a branch of traditional spectrogram that uses mel scale, focus on the lower frequencies and less focus on the higher frequencies. because the human auditory system is more sensitive to changes in lower frequencies. It reflects how humans may perceive sound.

3.1.3 Standard spectrogram vs Mel spectrogram

The Standard spectrogram and Mel spectrogram are utilized as the main input in Deep Learning networks for a lot of applications, including music emotion recognition [33], speech recognition [34], and speech emotion classification [35]. They are able to extract subtle features of an audio signal with regard to rhythm and melody effectively for accurate emotion recognition. The most important features that can be realized in the spectrum relate to tempo, frequencies, harmonic patterns, intensity, temporal changes, and timbre, which have a relationship with emotions.

While both the standard spectrogram and mel spectrogram can be considered as a visual representation of the frequency content of a signal over time, they are different in the way these frequency axes are represented. Mel-spectrograms are commonly used in high application in music emotion analysis due to the fact that they do excellently in capturing time series differences in frequencies [36].

Images below belong to the same song, but with a different type of spectrograms.

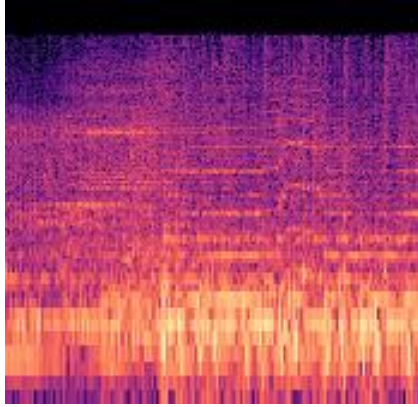


Figure 16: Standard Spectrogram

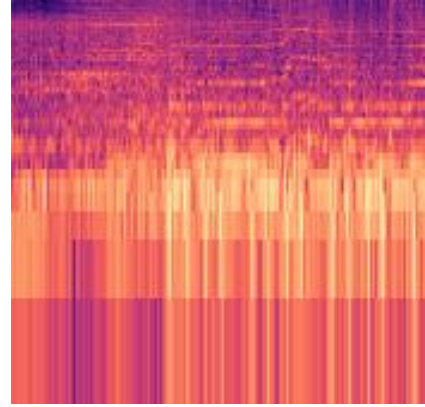


Figure 17: Mel Spectrogram

3.2 PRE-TRAINED MODELS

3.2.1 Overview of Transfer Learning and its Benefits

3.2.1.1 What is Transfer Learning?

Transfer learning is one of the recent active techniques of machine learning that uses acquired knowledge to improve performance in another, different but related problem. Owing to the transferable features, transfer learning becomes one of the most commonly used, gaining huge popularity in more fields such as speech recognition, image processing, and natural language processing because it can handle challenges with insufficient labeled data, incompatible computation power, and distribution mismatch [37].

In the context of DL, Deep Transfer Learning, DTL typically involves using a model that has been trained on a large dataset such as Resnet50, DenseNet121, Inception V3 models widely used in Transfer Learning applications, and already used in this research, then fine-tuning it for a specific task or dataset.

3.2.1.2 The Idea behind Pre-trained Models:

In 2010, the most crucial prominent competition in the world of computer vision and DL was established as the "ImageNet Large Scale Visual Recognition Challenge" (ILSVRC). The challenge was adopted an image dataset with over 14 million hand-annotated images in more than 20 thousand categories. This competition went with the requirements in image classification, object localization, and object

detection with segmentation by building a model and then applying it to this data. The competition's main objective was evaluating and developing ML models for the task of image recognition and classification; hence, it became one of the starting pivots in AI [38].

The competition boosted scientific research in the field through structural advances in neural network designs and improvements in performance, coupled with technological developments that brought new technologies as today's standards in DL. Some of the many important technologies and notable models include:

- AlexNet, 2012: This major model turned points towards the deeper structure to be applied further and focused on improvements needed in DL techniques [39]. Apparently, it won first place in 2012, proving its excellence in image recognition.
- VGG16, 2014: It was submitted by the Oxford University team. They came up with the idea of using deeper and simpler consecutive layers with 3x3 layers to achieve impressive results in 2014 [40].
- GoogLeNet, Inception 2014: This model was developed at Google using a structure called Inception that further uses many parallel paths. This model won first place in 2014 [41].
- ResNet50, 2015: It was developed at Microsoft in order to create a deeper network with the help of "Residual Connections". It took first place in 2015 [42].
- DenseNet: 2016, by Facebook AI Research, known as Dense Convolutional Network; it participated in the ImageNet competition and propelled image classification to a very high accuracy with its densely connected block design [43].

3.2.1.3 The Benefits of Pre-trained Models:

- A. Reduced training time: Pre-trained models have already learned general features, so less time is needed to train on a new task [44].
- B. Less data required: Since transfer learning learns useful features from pre-training on a large dataset, it can be quite effective even with less data.
- C. Better performance: Initial Training Using pre-trained weights most of the time leads to better performance compared to training from scratch, especially if this new task is similar to the original task.

D. Generalization: Usually, good generalization to new tasks is achieved with Pre-trained models due to the learning of robust features that turn out to have been acquired through a myriad of diverse data in many computer vision tasks [45].

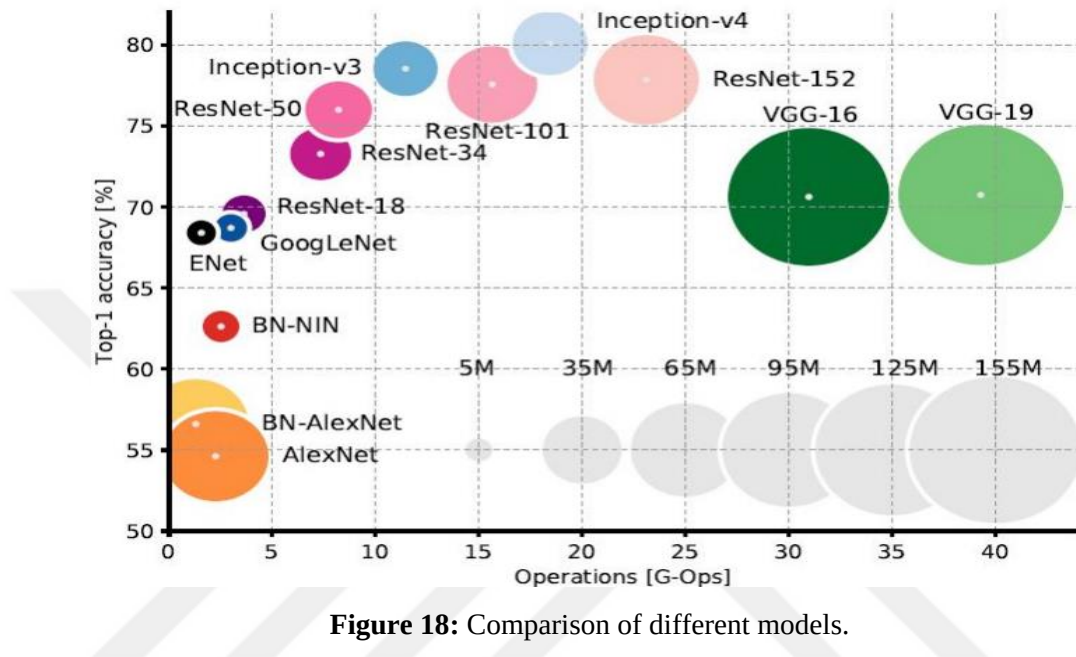


Figure 18: Comparison of different models.

3.2.2 Overview of Deep Learning and Models Used in this Study

3.2.2.1 What is DL?

Machine learning is the basis of artificial intelligence, and one of its most important parts is deep learning, which attempts to analyze data using a hierarchical architecture or what is known as multiple layers, where each layer extracts complex features such as edges and others and benefits from the information of the previous layers. Deep learning is applied in many fields, including transfer learning, computer vision, and more. The most important reasons that made deep learning prominent were expanded chip processing capacities, a decrease in the cost of the computer hardware, and the high development in machine learning techniques.

Among the most important networks:

- A. Artificial Neural Network, ANN: It consists of simple units, input layer, a neural hidden layer, and an output layer.

- B. Deep Neural Network, DNN: It consists of several hidden layers, input layers, and output layers.
- C. Convolutional Neural Network, CNN: This is a special kind of Neural Network used in the analysis and processing of images and videos. It contains convolutional layers, which are responsible for extracting spatial features from the input data.
- D. Recurrent Neural Network, RNN: One that deals with sequential data like text and speech; the possibility of its saving over time is allowed by units of frequency like LSTM and GRU.

3.2.2.2 Layers existing in the pre-trained models and its importance

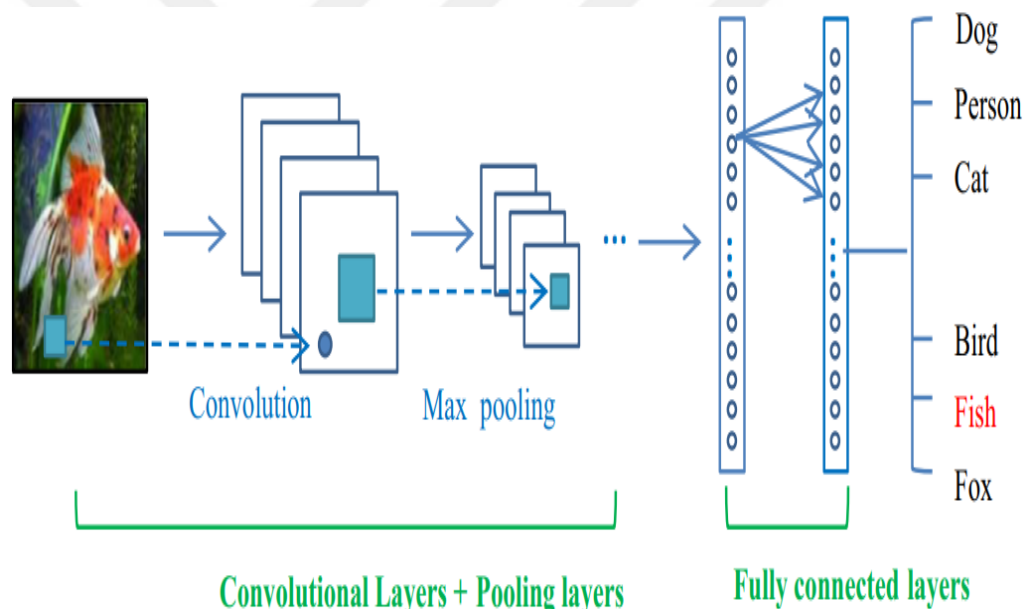


Figure 19: The layers of the convolution neural network in general (CNN).

1. Convolutional neural networks (CNNs) are a well-known deep learning method that uses multiple layers to learn in a strong way. They have proven to be very effective and are widely used in applications such as computer vision. Convolutional layers are typically the initial layer of a CNN. These layers provide the first operation with data containing spatial dimensions, more specifically images. These layers complete initial feature extraction and process data using filters. Indeed, filters can be considered small matrices of a

specific dimension (e.g., 1x1, 3x3) for the detection of edges or even textures. There might be many filters, and each filter has different features. In the application of the filters, each element of the sub-matrix is multiplied by the filter, after which the products are summed to give a value, which is then placed in the feature map. One of the convolutional layers processes is stride size, which defines how far the filter will move across the input. For instance, a stride of 1 would mean moving the filter one pixel at a time. Another process involved is padding. It is a process of adding zeros around the edge of the data. Two kinds of padding may be used:

- Valid padding: No zeros are added, so the data becomes small.
- Same padding: adding zeros keeps the original size of the data.

The number of feature maps equals the number of used filters, with each map representing another type of feature [46].

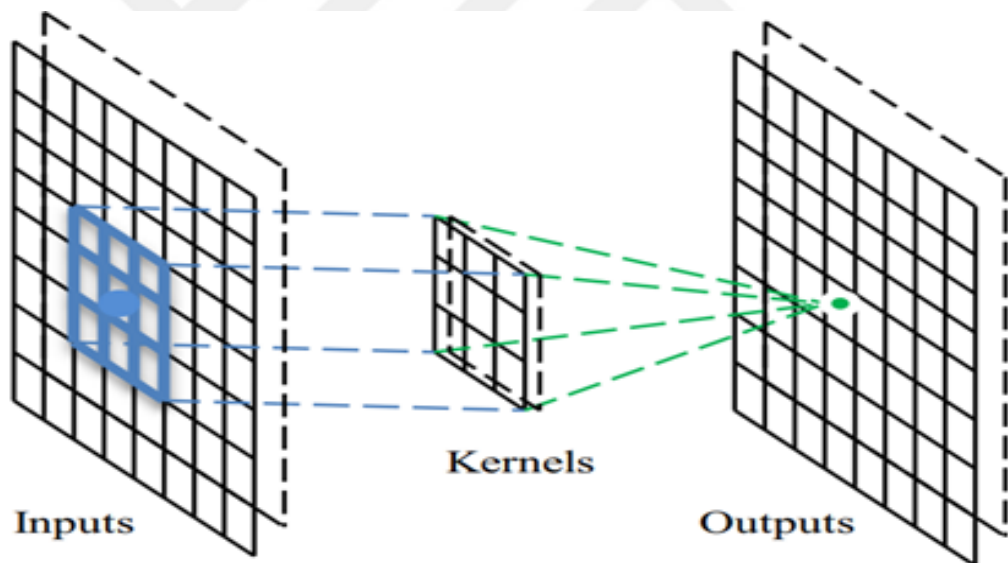


Figure 20:The operation of convolution layers.

2. Activation Layer: Basically, it introduces nonlinearity into the model. This is a nonlinear function, giving the model the ability to learn complex nonlinear patterns. They transform the output of convolution into a form suitable for the processing of the next layer. The types of activation functions are:

- Rectified Linear Unit, ReLU: Takes negative values to zero. This is yet another of the more common simple and effective functions used.
 - Leaky ReLU: In this activation function a small non-zero gradient for negative values is allowed to prevent dying neurons.
 - Sigmoid: This includes the output of values in the range (0, 1), typically used in the final layer for binary classification.
 - Hyperbolic Tangent, TanH: This will maintain values in -1 to 1 and can be used for normalized outputs.
 - Softmax: A module transforming a vector of values into a probability distribution. This is typically used in the last layer for multi-class classification.
3. The next layer after the convolution layer, known as the pooling layer, helps reduce the size of the feature maps and the number of network parameters. Pooling layers are used in a similar static and local manner to convolutional layers, by taking neighboring pixels into account [46]. Some types of pooling include:
- Max Pooling: It chooses the maximum value from every region of the feature map; it filters out only prominent features.
 - Average Pooling: This computes the mean values within each region, hence smoothening out features.
 - Global Pooling: Reduces each feature map to a single value by taking either the maximum (Global Max Pooling) or average (Global Average Pooling) of all values in the map.

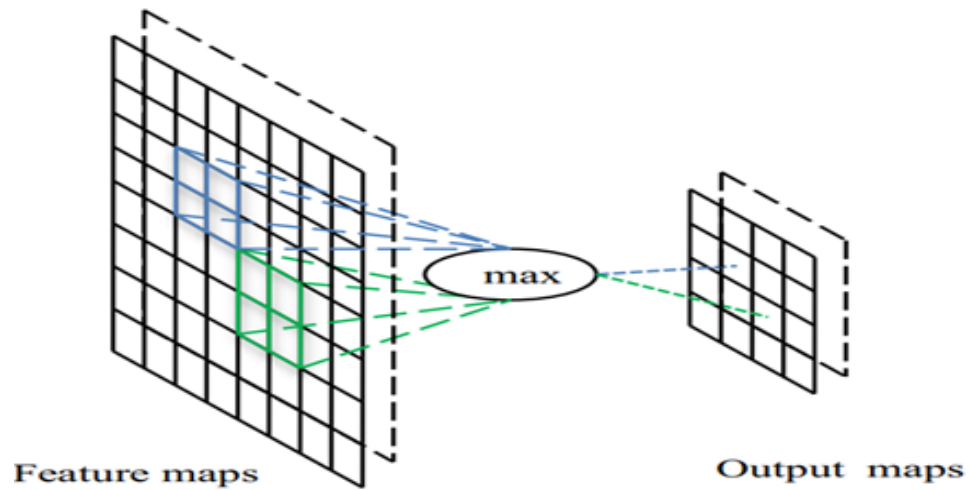


Figure 21: The operation of the max pooling layer.

4. Fully connect layers: eventually used to aggregate data. The task of this layer is to combine and synthesize data extracted from previous layers to produce important and useful outputs such as classification and prediction.

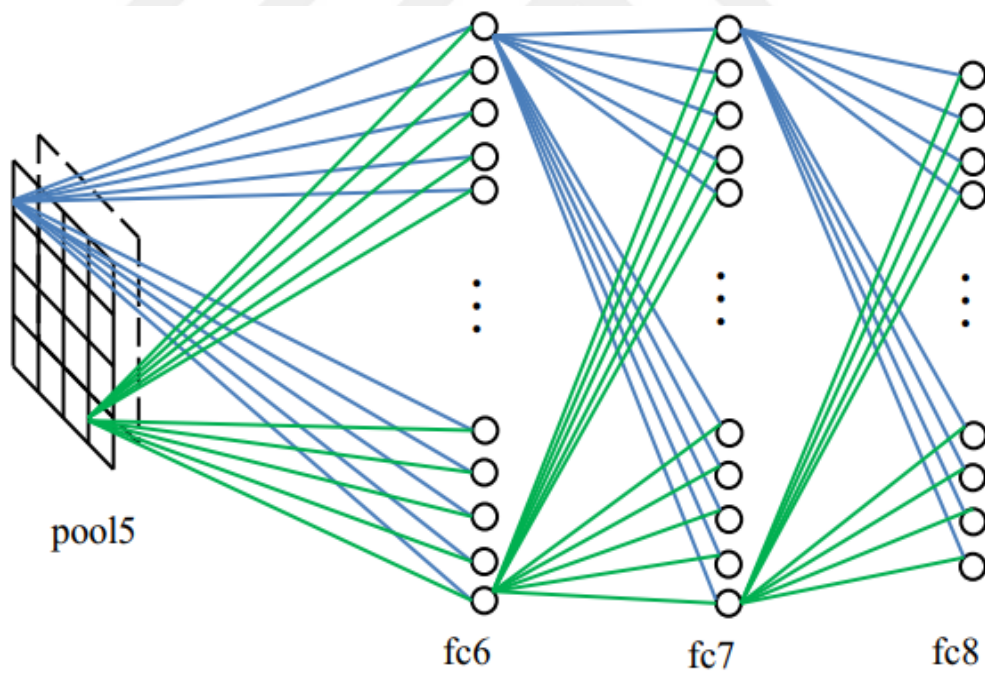


Figure 22: The operation of the fully connected layer.

3.2.2.3 Models

In this study, music emotion recognition is done using three very popular models that are pre-trained.

A. ResNet50:

This is a 50-layer Deep-CNN with roughly 25.6 million parameters. The residual learning framework that has made it famous enables training very deep networks. Originally, it was trained on the ImageNet dataset for image classification.

B. DenseNet121:

It is a 121-layer deep-CNN with about 8 million parameters. The network features a dense connectivity pattern whereby every layer is connected to all other layers in a feed-forward fashion. This means it efficiently reuses parameters and adds an improved flow of information and gradients from one layer to another.

C. Inception v3:

Inception-v3 is a 48-layer deep pre-trained convolutional neural network model, that contains approximately 23.9 million parameters. This model has inception modules, which are additionally equipped with multiple filter sizes and pooling operations in parallel. Known for their efficiency in computational cost and a number of parameters. One of the key features of Inception V3 is its ability to scale large datasets and to handle images of varying sizes and resolutions.

These models were selected because of their strong record in image classification tasks, and their potential to learn higher-level representations of complex patterns in spectrograms for music emotion recognition. We will illustrate how these models, pre-trained, can be exploited through transfer learning so that the features they have learned can, with adaptation, be used for the task of music emotion recognition. We detail their architecture and important features, explaining what makes them appropriate for music emotion recognition tasks.

3.3 PRE-TRAINED MODELS ARCHITECTURE

3.3.1 ResNet50

ResNet-50 is a convolutional neural network with 50 layers that focuses on learning residuals as opposed to features. This architecture introduces the Residual Network concept to address the problem of vanishing/exploding gradients [47].

The ResNet50 Stages consists of:

- The first stage (Conv1) Convolutional layer (1*1) with 64 filters, kernel size (7*7), stride (2), BatchNorm1: Batch Normalization, ReLU: Activation Function, Maxpool1: Max Pooling Layer with kernel size (3*3) and stride (2). This stage produces the signals required to drive the next stage of the network [48].
- The second stage (Conv2_x) is compounded by three series residual Units as well, the two upper layers have convolution operation with 64 filters plus Batch Normalization followed by ReLU activation function, and the, Shortcut connection, which is also known as Skip Connection, is a technique used in DNN to facilitate network training by by-passing or skipping one or more layers. In this approach, the original input is added directly to the output of a specific layer [42].

Conv2_1:

Conv [1*1,64], BatchNorm, ReLU.

Conv [3*3,64], BatchNorm, ReLU.

Conv [1*1,256], BatchNorm, Shortcut connection

(adding the original input to the final output)

Conv2_2 and Conv2_3 both are implementing the same Conv2_1 approach.

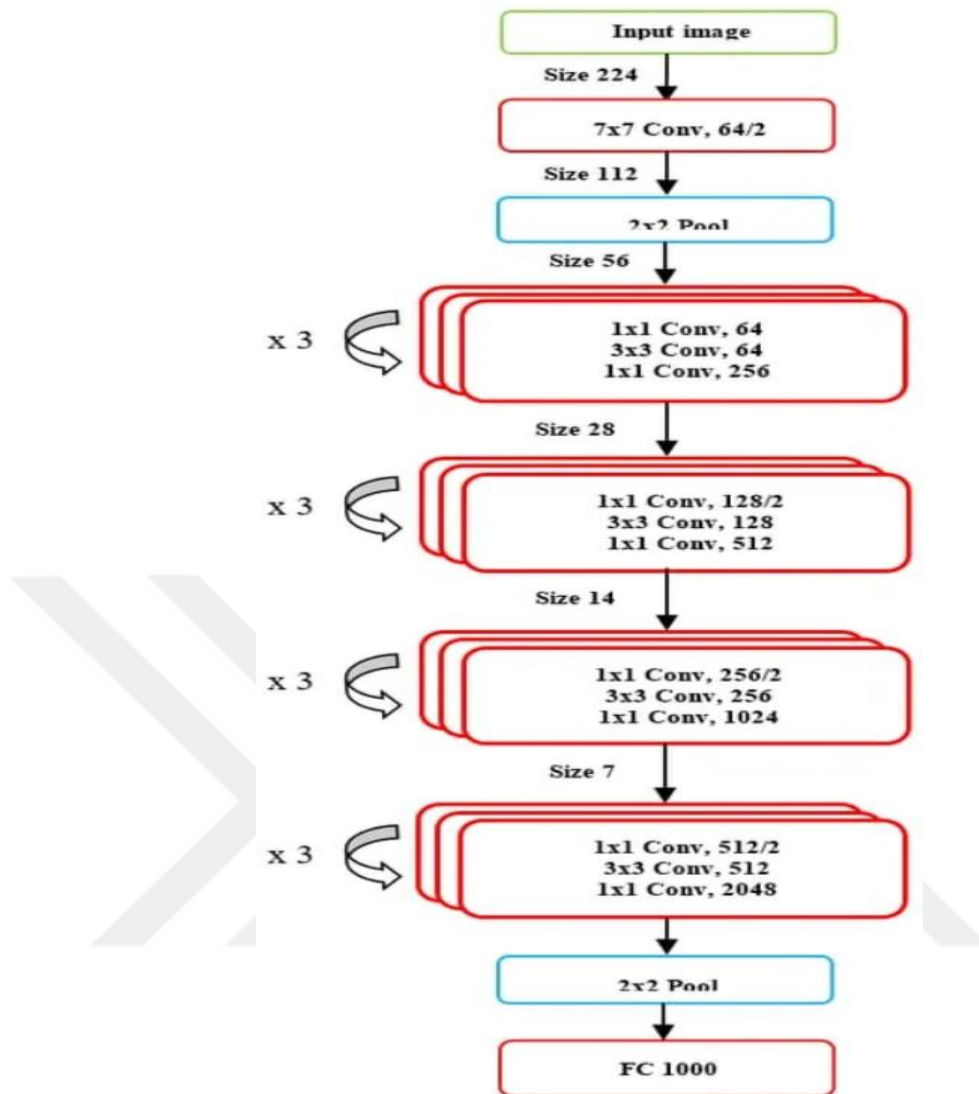


Figure 23: The architecture of ResNet50

- Now, the analogy of the third stage is dealt with through four series of residual units. The first layer is the convolution layer (1*1) with 128 filters, followed by batch normalization, then ReLU activation function.
Conv3_1 :
Conv [1*1,128], BatchNorm, ReLU.
Conv [3*3,128], BatchNorm, ReLU.
Conv [1*1,512], BatchNorm, Shortcut connection
(adding the original input to the final output)
Conv3_2, Conv3_3 and Conv3_4 are the same in their structure as Conv3_1.

- The fourth stage, with a stage ID equal to Conv4 of the ResNet50 model, contains six series ResUnits. The first convolution layer consists of (1*1), followed by batch normalization and the activation function ReLU. Then, the second convolution layer of the ResNet50 in Conv4 is (3x3) with 256 filters consecutively, followed by batch normalization and ReLU activation function. The batch normalization layer is attached finally with (1*1) layer from which the 1024 feature maps of the Residual units are formed. Here it is a direct connection and thus includes the original input in the final output.

Conv4_1:

Conv [1*1,256], BatchNorm, ReLU.

Conv [3*3,256], BatchNorm, ReLU

Conv [1*1,1024], BatchNorm, Shortcut connection

(adding the original input to the final output)

Conv4_2, Conv4_3, Conv4_4, Conv4_5 and Conv4_6, follow the same Conv4_1 strategy

- The fifth and very end stage contains three residual blocks, each consisting of two convolutional layers with 512 filters, followed by a down-sampling operation.

Conv5_1:

Conv [1*1,512], BatchNorm, ReLU.

Conv [3*3,512], BatchNorm, ReLU.

Conv [1*1,2048], BatchNorm, Shortcut connection

(adding the original input to the final output)

Conv5_2, Conv5_3 follow the same strategy as Conv5_1.

At the very end stage of them all, there is a Global Average Pooling layer, which calculates the average of the activations over the spatial dimensions of the output. The output will feed into a fully connected layer containing 1000 units, producing the final network output [49].

Note: BatchNorm layers are used for both improving the model's effectiveness and stabilizing it. the ReLU layers for nonlinear activation, i.e., for any negative value, the result is converted to zero [50].

The most significant problem that occurs while training DNN is vanishing gradients (gradients are very small) this problem makes the improvement and learning for layers very difficult. Skip connections (residual connections), in general, perform quite well in solving the problem of vanishing gradients by allowing the network to learn deeper architectures without vanishing gradients [51].

To address this problem, Skip connections allow the information to flow directly from the input to the output of the network, by passing one or more layers as shown in the Figure below.

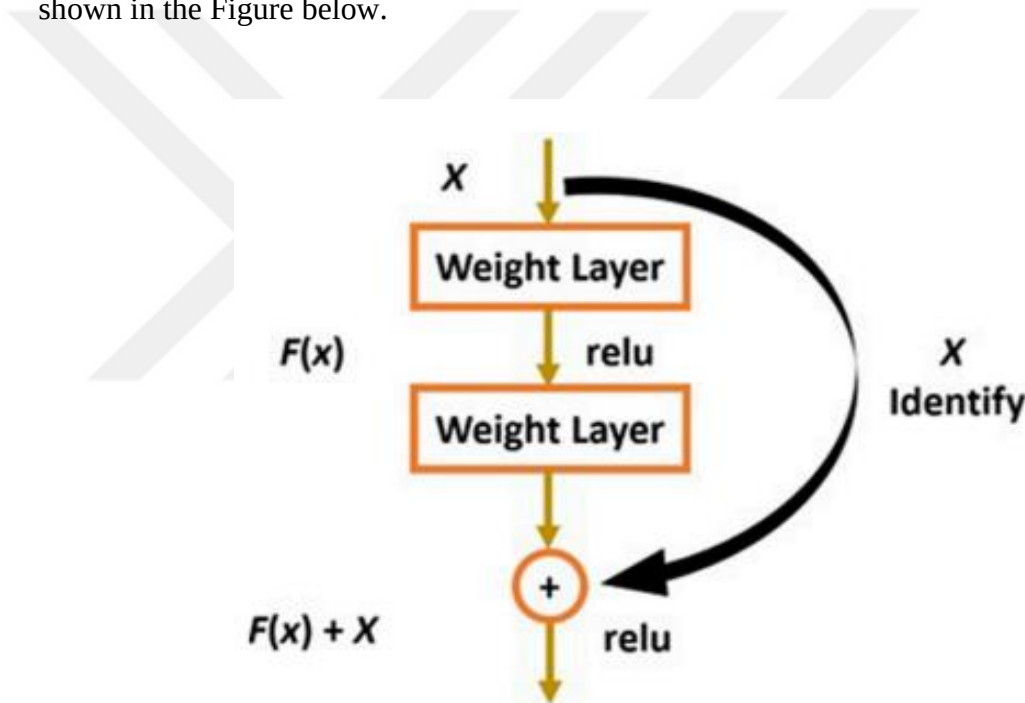


Figure 24: Describe the method of residual block.

This allows the network to learn residual functions that map the input to the desired output instead of having to learn the full mapping from scratch.

3.3.2 DenseNet-121

DenseNet-121 is one of the most innovative architectures designed for CNNs, which is considered one of the important tools in the field of DL, to be used in multiple

fields such as image classification, voice recognition and video applications. It was characterized by its design method based on dense links between its layers, which helped in improving the performance of the model through the transmission of information and gradations over the network [52].

Basic Structure:

- It consists of a number of dense layers and transition layers that contribute to reducing dimensions while preserving information [52].
 1. The inputs are images with size (224*224*3).
 2. The initial convolutions layers are Conv (7*7), stride =2, followed by a max-pooling layer, a filter size of 3 x 3, and stride 2.
 3. Dense Blocks: consist of 6, 12, 24, and 16 layers respectively.
 4. Transition Layers are layers introduced between dense blocks that reduce dimensions by using convolutional layers of size 1x1 and pooling layers of size 2x2.
 5. Final layers after passing through all previous layers, Global Average pooling layer, which is then followed by a fully connected layer containing 1000 units.

Network Components:

- Dense layers: In each layer the output from previous layers is concatenated to the output of the current layer itself. This gives rise to direct information flow and gradient propagation.
- Transition layers: These layers reduce the dimensions and size of data by convolution and pooling, thus maintaining network efficiency and not doing excessive computations.

Advantages of DensNet121:

1. Due to the presence of dense links, this improves the transition of the gradient and reduces the problem of fading due to the merging of the previous input with the suffix.
2. The number of parameters is less than other networks, so they are more memory efficient.
3. It is suitable for image classification, audio learning, video applications, and medical domains projects [52].

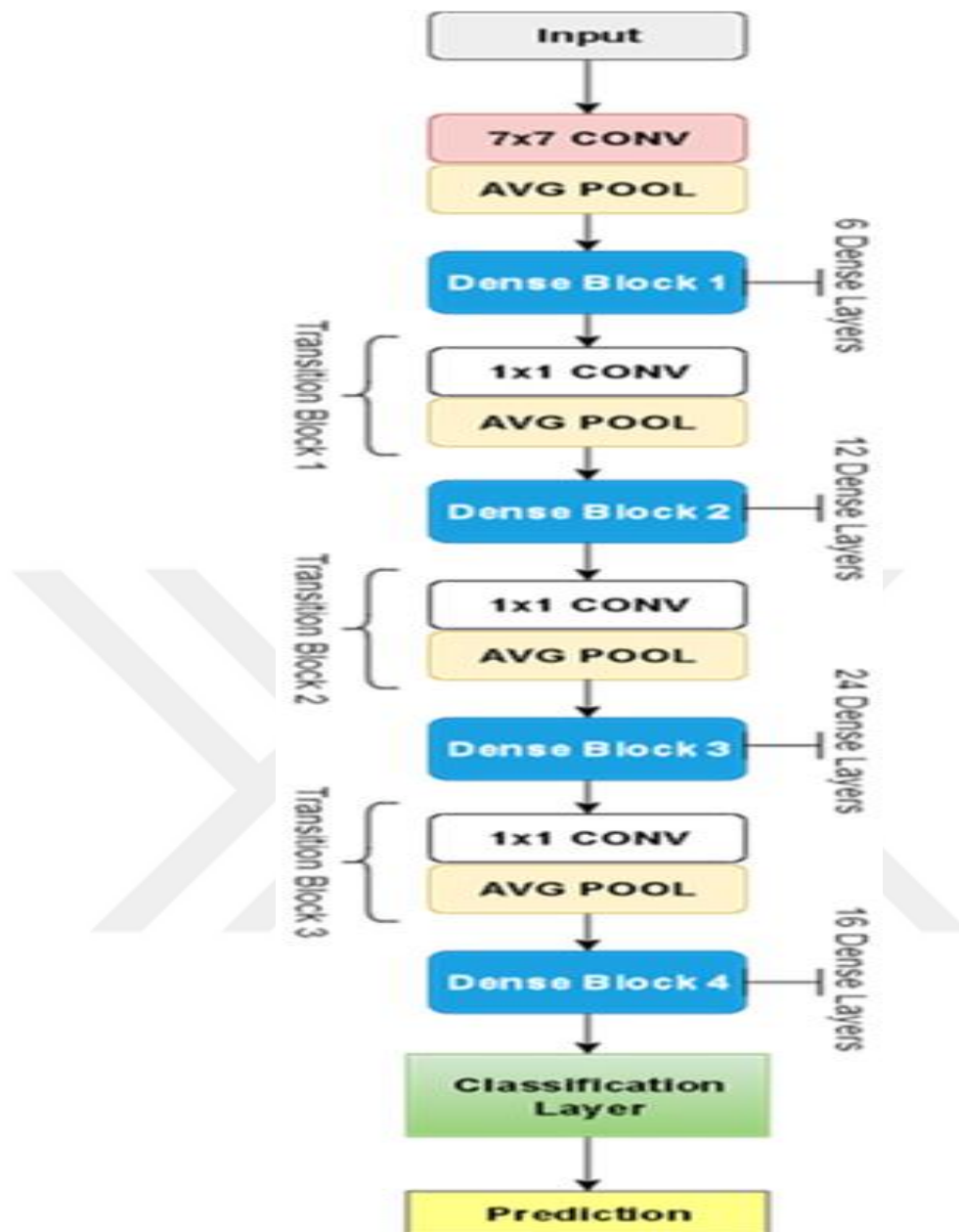


Figure 25: The architecture of DensNet-121.

3.3.3 Inception v3

Inception is a CNN image classification model, designed for the tasks of classifying Images in an efficient way. The network structure contained convolution, pooling, and inception modules make up this network.

3.3.3.1 Overview of Inception V3:

1. Inception Modules [43]:

- Convolution 1x1: Dimensionality reduction improves computational speed by reducing dimensions while capturing detailed information from the previous layers.
 - Convolution 3x3: Uses medium information from the image and detects different Patterns
 - Convolution 5x5: Reconstruction of large-scale details and identifying broader structures in an image.
 - Max Pooling 3x3: Reduce the spatial dimensions to focus on deeper and abstract information.
2. Factorization into Smaller Convolutions:
 - To improve quality and reduce the computational, large convolutions like 5*5 are broken down into series of smaller convolutions 3*3, then performance will improve, and the parameters will be reduced.
 3. Spatial Convolutions:
 - To reduce computational complexity while maintaining accuracy, large convolutions are broken down into smaller ones. For instance, from 7x7 into a convolution of 7x1 and another of 1x7.
 4. Regularization:
 - Dropout and label smoothing are embedded to act against overfitting within the Inception V3 model.
 5. Pre-trained Weights:
 1. Pre-trained weights speed up transfer learning on a new task.

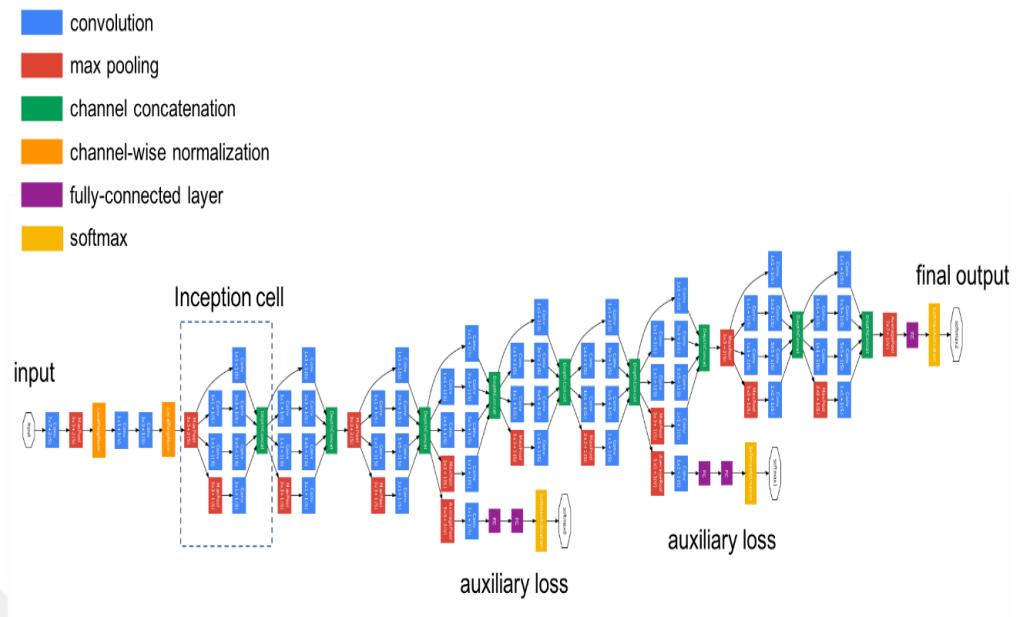


Figure 26: Inception-v3

3.3.3.2 The architecture of the Inception V3:

1. Initial Layers: [55]

Conv 3x3: The first convolutional layer with a filter size of 3x3 and 32 channels.

Conv 3x3: Followed by another convolutional layer with a 3x3 filter size and 64 channels.

2. Inception Modules: [55]

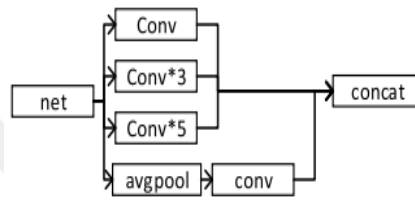
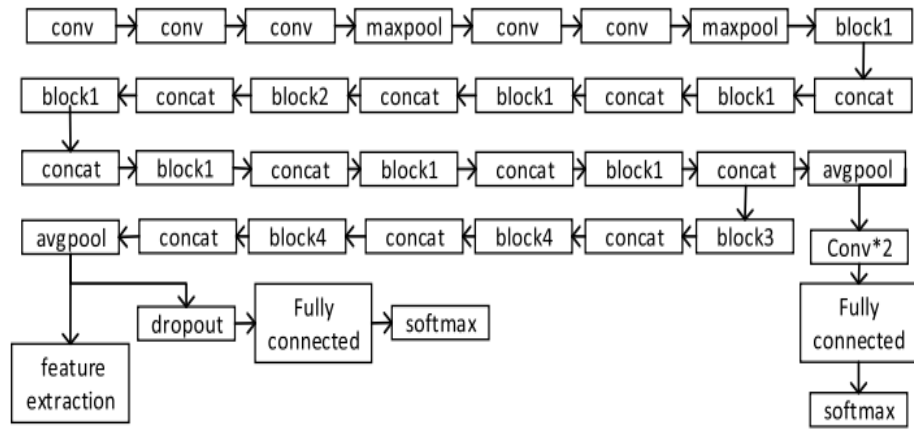
Inception-type A: 1x1, 3x3, and 5x5 convolutions followed by a pooling layer.

Inception-type B: Factorize convolutions to break down larger convolutions into smaller ones.

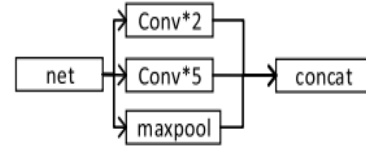
Inception-type C: 1x1, 3x3, and factorized convolutions like 7x1 and 1x7 to improve computational efficiency. Final Layers:

3. Global Average Pooling: It reduces each feature map to a single value by averaging. [55]

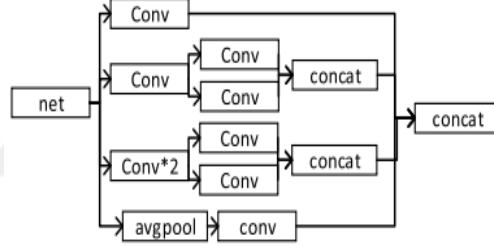
4. Fully Connected Layer: Contains often 1000 units for classification on the ImageNet dataset. [55]



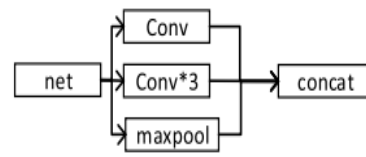
(a). The detail of block1



(b). The detail of block2



(c). The detail of block3



(d). The detail of block4

Figure 27: The architecture of the Inception_v3.

3.4 COMPARISON OF MODEL ARCHITECTURES

If we want to compare between the three networks, in terms of implementation speed and storage space, we will find:

- 1- The ResNet50 network has a large number of parameters, so it will need substantial memory to store these parameters in addition to additional memory to store intermediate outputs during execution, but because of its architecture in designing the residual blocks that allowed gradient flow more efficiently, the network improved its use of memory improved its speed.
- 2- DenseNet-121 Although the number of parameters is less than other networks, the presence of dense links in their architecture, which increase computational

complexity, made their speed of execution slower than the ResNet50 and increased memory consumption is very large. Because these dense links require storing intermediate outputs from each layer.

- 3- Inception-v3 is characterized by complex architecture due to Inception modules that combine convolution cores of different sizes, this can lead to increased execution time compared to ResNet50, it can be improved by using specific optimization techniques, but its space consumption is very large with many parameters in addition to Inception units.

Table 1: Models comparison

<i>Feature</i>	ResNet50	DenseNet121	Inception v3
<i>Parameters</i>	~25.6 million	~8 million	~23.9 million
<i>Memory Usage</i>	Moderate	High	Moderate to High
<i>Computational Efficiency</i>	High	Moderate	Moderate
<i>Key Architectural Feature</i>	Residual connections	Dense connections	Inception modules
<i>Depth</i>	50 layers	121 layers	48 layers
<i>Feature Extraction Capability</i>	Strong hierarchical features	Preserves fine-grained details	Multi-scale feature extraction
<i>Suitability for Capturing Spectral Features</i>	High	High	High
<i>Suitability for Temporal Features</i>	Good	Good	Very Good
<i>Adaptability to MER Tasks</i>	High	High	High
<i>Performance vs Resource Trade-off</i>	Well-balanced	Efficient parameter use, high memory	May require optimization

ResNet50 appears to offer a strong balance of features, efficiency, and adaptability for this task.

CHAPTER IV

METHODOLOGY

4.1 AUDIO PROCESSING AND AUDIO FILES

Sound waves are patterns of compression and expansion that travel through a medium like air at frequencies that are audible to humans. Thus, audio can be as simple as a person clapping their hands or singing outside, or it can be as complex as a symphony orchestra performing in a reverberant performance hall. The replication of sound as it passes through electrical circuits and wires is also referred to as audio [56].

4.1.1 Audio Files

Audio files are considered a basic part of our everyday life, used for music, podcasts, phone calls, and so on. There are two universally supported types of audio files [57]:

4.1.1.1 Uncompressed Audio Files

These files maintain their original size once they are stored in the memory. Common formats include:

1. Waveform Audio file format, WAV: Originally invented by Microsoft and IBM, it is one of the most widely used and oldest formats. It saves data in raw form using Pulse Code Modulation, PCM encoding without compression, preserving all audio information [58].
2. Audio Interchange File Format, AIFF: It was the next file developed by Apple as an alternate version for WAV files. It encodes the same PCM. It has very high sound quality and is used within the music industry; however, it takes further storage space and is not as well supported as the WAV file [59].

4.1.1.2 Compressed Audio Files

These are situated in two varieties:

A. Compressed Audio Files without quality loss:

- 1) Free Lossless Audio Codec, FLAC: Smaller than uncompressed files with high sound quality, but still larger than lossless compressed files like MP3 and AAC. Not universally supported [60, 61].
- 2) Apple Lossless Audio Codec, ALAC: Developed by Apple for iTunes, it assures premium sound experience on Apple products, but it is very bulky and less universally supported [62].

B. Compressed Audio Files with quality loss:

- 1) MPEG-1 Audio Layer III, MP3: A widely employed format that came to life in the early 1990s, MP3 is the format of choice for most handheld devices. It is an acceptable audio format, as it is developed through complex compression. This process takes away inaudible data, so the file is of lower quality than the uncompressed one, but it's a data saver as well [63][64].
- 2) Advanced Audio Codec, AAC: This format was developed as an improvement over MP3, used in iTunes and YouTube. It offers better audio quality than MP3 at similar bit rates but is less widespread [64].
- 3) OGG Vorbis: Is a full open source, non-proprietary, patent and royalty free compression audio format. [57] It's a format that is less known but still makes big ins. With this format you can have better results. The loss of audio quality is much smaller than MP3, however, the problem with this format is that it is not commonly used [65].

4.1.2 Audio Technology

The technology of audio refers to devices and programs used for audio modification and operation, such as microphones and speakers [66].

4.1.3 Audio Signals

Audio signals are based on electrical signals processed by PC digital processors, including compression and volume reduction [67].

4.1.4 Audio Applications

This involves various uses of audio in digital music and other digital applications [68].

4.2 MUSIC AND EMOTIONS

In this section, I will give an overview of the Music and its relationship with emotions: Music is those beautiful tones, played on musical instruments or those sung without musical instruments, which is called vocal music. [69] The basis of these melodies is the sound, which depends on the vibration of waves in media such as air and water. What we are hearing of the resonance of objects is known as a sound [70]. Melodies are sounds that have been collected in the form of notes that convey deep and complex sensations at the same time. Melodies have always evoked a great deal of emotions in our souls upon hearing them.

Scientists have tried to explain these feelings and emotions evoked by listening to music and consider them complex sensations. Understanding this music requires a deep and accurate analysis of its basic elements in order to comprehend the characteristics of the feelings evoked when hearing it.

Musical emotions depend on the music and on the listener as well. The emotional response of the listener includes a series of responses to the work of the brain in which psychological expressions and feelings are manifested.

4.2.1 Music and Their Relationship to Emotions

A. Melody [12]:

Defined as a horizontal succession of pitches, perceived by listeners as a single music line. The elements of melody are:

- 1) Pitch: It refers to the relative highness or lowness of musical sound. It depends on frequency. High pitch is related to the high frequency and the low pitch is related to the low frequency.
- 2) Pitch range: It is the difference between the frequency of the highest pitch and the lowest pitch. It is expressed as either Wide (divergent frequencies) or Narrow (close frequencies).
- 3) Pitch variation: It can be small or large.

4) melodic intervals: The distance between pitches in melody, it may be large interval, minor, perfect and major.

5) Melodic direction: The sound may be ascending or descending.

B. Harmony [12]:

Is the sound created by combining different pitches (notes or tones) in chords.

The elements of harmony are:

- 1) Harmonic perception (harmonic interval), maybe consonant (simple) associated with positive, dissonant (complex) associated with negative, high pitched, low pitched.
- 2) Tonality (A musical concept that refers to the frameworks in which tones and harmonies are organized): Major modes have commonly been associated with positive emotions, whilst minor modes are associated with negative ones.

C. Rhythm [12]:

It is a pattern of sounds and silence; it represents how long or short the sound is. Elements of Rhythm are:

- 1) Tempo: fast or slow.
- 2) Tempo and note value: Varies from slow to fast.
- 3) Rhythm type: regular, irregular (rough), regular (smooth), complex, firm, flowing/fluent and rests (after tonal closure the sequence start & end the same key, after no tonal closure)

D. Dynamics [12]:

The dynamic of a musical work is represented by the variation in the loudness or softness of its tones. Elements of Dynamics are:

- 1) Dynamic levels (high loud/ low soft)
- 2) Accent and change in the levels (Large, Small, Rapid variation, No changes)

E. Timbre [12]:

It is a tone color that is related to lower-level elements and sound characteristics such as amplitude and spectrum, which are important for distinguishing between instruments and voices. Elements of Timbre are:

- 1) Amplitude envelope: round, sharp.

- 2) Spectral envelope (no harmonic): low, high.
- 3) Spectral characteristics: positive correlation, negative correlation.

4.2.2 Characteristics of Musical Emotions

This study focuses on several emotional categories in music, each with distinct characteristics:

Table 2: Comparative Analysis of Emotional Categories in Music

Feature	Happy	Sad	Anger	Devotional	Calmness/Relax	Party	Tenderness/ Love
<i>Tempo</i>	Fast, small variability	Slow	Fast, small variability	Moderate to slow	Slow	Moderate to Fast	Slow
<i>Mode</i>	Major	Minor	Minor	N/A	N/A	N/A	Major
<i>Harmony</i>	Simple and consonant	Dissonant	Dissonant	N/A	N/A	N/A	Consonant
<i>Sound Level</i>	Medium-high, small variability	Low, moderate variability	High, small loudness variability	N/A	N/A	N/A	Medium-low, small variability
<i>Pitch</i>	High, much variability, wide range, ascending	Low, narrow range, descending	High, small variability, ascending	Varies, often uses mid-range pitches	Low	High	Low, fairly narrow range
<i>Rhythm</i>	Smooth and fluent	Slow, pauses present	Complex, sudden changes	Regular and steady	Regular	Regular	N/A
<i>Timbre</i>	Bright	Dull	Sharp	N/A	N/A	N/A	Soft
<i>Articulation</i>	Staccato, large variability	Legato, small variability	Staccato, moderate variability	N/A	N/A	N/A	N/A
<i>Dynamics</i>	N/A	N/A	N/A	Soft to moderate	Soft	Loud	N/A
<i>Melody</i>	N/A	N/A	N/A	Simple and repetitive	Consonant, harmonious, and pleasing	Harmonious and positive	N/A
<i>Vibrato</i>	Medium-fast rate, medium extent	Slow, small extent	Medium-fast rate, large extent	N/A	N/A	N/A	Medium fast, small extent
<i>Micro-structure</i>	Regular	Irregular	Irregular	N/A	N/A	N/A	Regular

4.2.3 Categorical Classification of Emotions

Categorical classification or the so-called discrete classification uses vocabulary or set of vocabulary to describe data based on certain characteristics, it is widely used in various fields such as the classification of images, texts, and musical

feelings. There are theories of psychology that deal with psychology and emotions and identify the categories of these emotions accurately. One of the most prominent of these theories is Eckman's theory ¹of basic emotions, which has been used as a basic tool for musical feelings [73].

Types of Categorical Classification:

- I. Binary Classification (simple type): It was adopted by classifying the data into two categories, for example, in the context of diseases, infected or uninfected, in the context of music, positive or negative, sedative or stimulating, etc. [76].
- II. Multi-classification: This was based on the classification of data into more than two categories, for example, in the musical field, songs can be classified into multiple emotions such as joy, sadness, calm, fear, fun, etc. [76].

The problem of categorical approach:

- Ambiguity: Happy = joyous = cheerful = delight. [77]
- Granularity [77]:
 - Few emotion classes, insufficient to capture the diversity of human perception.
 - Large number of emotion classes make determining truth value challenging.

4.3 DATASETS

In this section, I will give an overview of the two datasets: (MER 500) and (MET), which were used in this study. Music was classified on these two datasets using the pre-trained models ResNet50, DenseNet121, and Inception V3. The aim of this study was to increase categorization accuracy.

4.3.1 Turkish Music Emotion Dataset

Base ingredients:

Origin: This dataset was a discrete music dataset have selected from different genres of Turkish music, prepared by Mehmet Bilal Er*, Ibrahim Berkan Aydilek,

¹ Ekman's theory of basic emotions: The theory of basic feelings presented by the well-known scientist Paul Ekman, stipulated the existence of six basic human feelings (happiness, sadness, anger, fear, surprise, and disgust) as he considered these six feelings the basic building block for the emergence of broad and complex feelings [74].

Department of Computer Engineering, Faculty of Engineering, Harran University, 63050 Sanliurfa, Turkey*Corresponding author. Email: bilal.er@harran.edu.tr.

Composition: Total number of audio files: 400

Format: MP3

Duration: Approximately 30 seconds each

Classes: Four emotional categories (Happy, Angry, Sad, Relax)

Distribution: 100 audio files per class

Characteristics: Includes both verbal and non-verbal music from different genres of Turkish music

Note: The files are not exactly equal to 30 secs. The details as shown in the table below:

Table 3: The MET audio files durations

SEQUENSE	CATEGORY	NO. OF FILES	DURATION TO EACH FILE
1	Happy	100	29-30 Secs.
2	Sad	100	30 Secs.
3	Angry	100	29-32 Secs.
4	Relax	100	28-30 Secs.
Total		400	

4.3.2 MER500 Dataset

Base ingredients:

Origin: This dataset was prepared by researchers Nikita Funde, Anuja Ahiwale, Sanjivani Bansode, Apoorva Sanglikar, and Pranita Pedukar from an Indian university for the task of analyzing MER to help researchers in developing models for classifying musical emotions.

Composition: Total number of audio files: 494 (original count before preprocessing)

Format: WAV

Duration: Varying lengths (in range, 8-59 seconds as shown in the table below)

Classes: Five emotional categories (Party, Happy, Sad, Devotional, and Romantic)

Distribution: Approximately 100 clips per category (there's an imbalance in the original data)

Characteristics: Popular Hindi film songs

Table 4: The MER-500 audio files durations

SEQUENSE	CATEGORY	NO. OF FILES	DURATION TO EACH FILE
1	Devotional	95	8-25 Secs.
2	Romantic	100	8-12 Secs.
3	Parties	100	8-9 Secs.
4	Happiness	100	9-11 Secs.
5	Sadness	99	8-59Secs.
Total		494	

4.4 DATA PREPROCESSING

In the context of data preprocessing, a two-step strategy was followed. The basic preprocessing technique across both datasets gave reasonable accuracies in the first step, Subsequently, a more advanced preprocessing method was applied. First of all, I wanted to make sure of a fair comparison between the two datasets used in the thesis by standardizing the preprocessing used by each of them. Secondly, I wanted to reach a more comprehensive level of generalization.

4.4.1 First Stage Preprocessing

➤ For Turkish Music Emotion Dataset:

1. Converted MP3 files to WAV format , for several reasons, the most important of which is that the WAV format is a standard format in research and audio works because it is compatible with most audio analysis programs and tools, it maintains the accuracy of the information, making it easier to extract audio features with high accuracy, and the spectrum extracted from it is more accurate and clearer than the data extracted from MP3.
2. Divided each audio file into 10-second segments.

3. Neglect during splitting, any period of less than ten seconds (without the use of a filter).
 4. Resulted in 1190 audio files of 10 seconds each.
- For MER500 Dataset:
1. No modification was made to the length of the duration.
 2. The duration of the clips varied between 8-59 seconds (the longest syllables were in the category of sadness).

4.4.2 Second Stage Preprocessing: Conversion and Standardization

- For Turkish Music Emotion Dataset:
1. Converted MP3 files to WAV format.
 2. Divided each audio file into 10-second segments.
 3. Discarded segments equal to or less than 5 seconds.
 4. Padded segments between 5-10 seconds to reach 10 seconds.
 5. Divided files longer than 10 seconds into multiple 10-second segments.
 6. Discarded any remainder less than 5 seconds.
 7. Resulted in 1200 audio files of 10 seconds each.

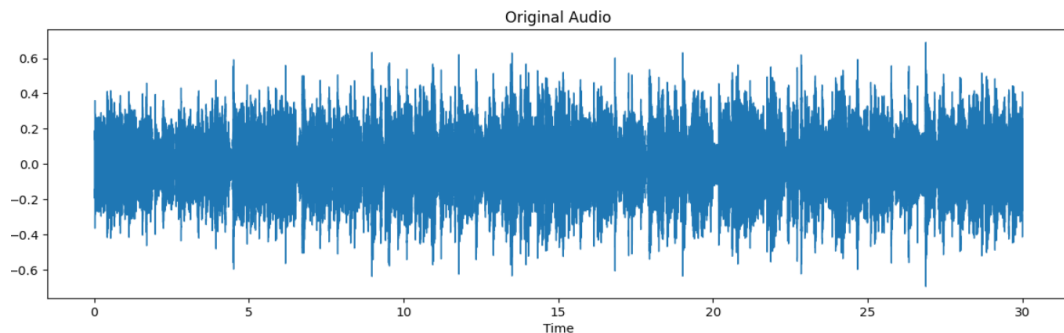


Figure 28: MET Original Audio

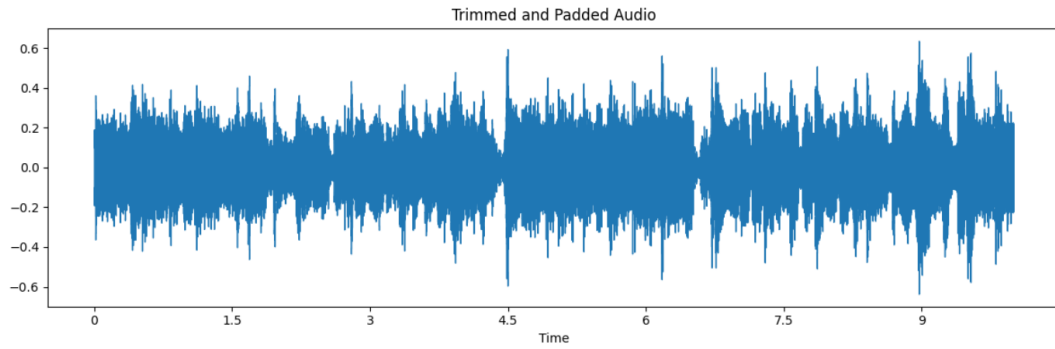


Figure 29: MET Preprocessed Audio file

Table 5: MET Audio files after trimmed & standardizing their duration.

SEQUENSE	CATEGORY	NO. OF FILES	DURATION TO EACH FILE
1	Happy	300	10 secs.
2	Sad	300	10 secs.
3	Angry	300	10 secs.
4	Relax	300	10 secs.
Total		1200	

➤ For MER500 Dataset:

1. Standardized audio files to 10 seconds.
2. Discarded files equal to or less than 5 seconds.
3. Padded files between 5-10 seconds to reach 10 seconds.
4. Divided files longer than 10 seconds into multiple 10-second segments.
5. Discarded any remainder less than 5 seconds.
6. Resulted in 603 audio files of 10 seconds each.

Table 6: MER500 Audio files after trimmed & standardizing their duration (unbalance).

SEQUENSE	CATEGORY	NO. OF FILES	DURATION TO EACH FILE
1	Devotional	97	10 Secs.
2	Romantic	100	10 Secs.
3	Parties	100	10 Secs.
4	Happiness	100	10 Secs.
5	Sadness	206	10 Secs.
Total		603	

4.5 DATA BALANCING TECHNIQUES

After standardizing the duration of audio files in the MER500 dataset, a new problem appeared, which is the weak distribution of data, as the sadness category has 206 files, while the rest of the files approximately have 100 audio files, in terms of solving this weakness, we have used some balancing techniques named:

4.5.1 Under-sampling

Under-sampling: It is a technique used to address the imbalance of data distribution when one category is more than the other, then the number of samples in the most possible category is reduced allowing the model to learn better.

There are many methods to apply Under-Sampling techniques, in this study, the Random Under-sampling technique was applied as follows.

- Random Under-sampling: A random sample is selected from the most representative category, the method is very simple but the possibility of losing important data is very likely, which may affect the accuracy of the model, one of the Under-sampling advantages is that it reduces the size of data and thus reduces training time, but the disadvantage is the probability to lose the important information and thus reduces the accuracy. The amount of data after applying under-sampling became: Devotional: 97, Happy: 97, Party: 97, Romantic: 97, Sad: 97

Table 7: the audio files after under-sampling.

SEQUENSE	CATEGORY	NO. OF FILES	DURATION TO EACH FILE
1	Devotional	97	10 Secs.
2	Romantic	97	10 Secs.
3	Parties	97	10 Secs.
4	Happiness	97	10 Secs.
5	Sadness	97	10 Secs.
Total		485	

4.5.2 Oversampling

Oversampling: It is a technique to augment data in the least numerous categories to make the distribution symmetrical and balanced, in this study, the random Oversampling technique was applied as follows.

- Random Oversampling: A random sample is selected from the least representative category and the technique is repeated until it reaches the required number. The amount of WAV data after Oversampling became: Devotional: 206, Happy: 206, Party: 206, Romantic: 206, Sad: 206

Table 8: the audio files after over sampling.

SEQUENSE	CATEGORY	NO. OF FILES	DURATION TO EACH FILE
1	Devotional	206	10 Secs.
2	Romantic	206	10 Secs.
3	Parties	206	10 Secs.
4	Happiness	206	10 Secs.
5	Sadness	206	10 Secs.
Total		1030	

4.6 FEATURE EXTRACTION TECHNIQUES

In the context of feature extraction, a two-step strategy was followed. The basic transformation to standard spectrogram across both datasets was applied, and gave reasonable accuracies in the first step. Subsequently, a more advanced feature extraction techniques was applied, in addition to implement a new feature extraction techniques called Mel Spectrograms. First of all, I aimed to leverage the strengths of each representation to enhance the performance and accuracy of the emotion recognition models. Secondly, I wanted to reach a more comprehensive analysis of the audio data, ultimately leading to generalization.

➤ Standard Spectrogram Generation

1. Frequency range: 20-8000 Hz
2. Hop length: 512
3. FFT number: 2048
4. Converted amplitude to decibel scale

➤ Mel-Spectrogram Generation

1. Frequency range: 20-8000 Hz
2. Hop length: 512
3. Number of mel bands: 128
4. Converted energy values to decibel scale

In this research, while obtaining a standard spectrogram in STFT values, I used a function that would convert amplitude into a decibel scale so that it is clearer to represent the variation between large and small amplitudes. The Mel-Spectrogram also used a function that renders energy values into a decibel scale for improving the handling of a wide range of values. Both methods support acoustic analysis in ways befitting the analyzed data. After extracting the two types of spectrograms, we convert the spectrogram to images in format PNG and later change the sizes of the spectrogram images to a size of 224*224*3.

- Here's a sample for the four classes: Happy, Sad, Angry, and Relax, of the Standard Spectrogram vs. Mel Spectrogram on the Turkish Music Emotion Dataset:

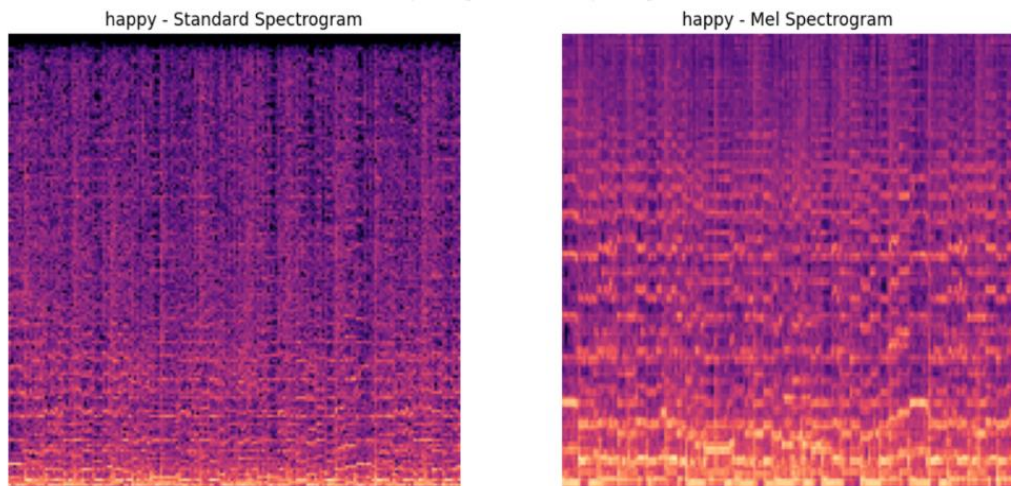


Figure 30: Standard Spectrogram vs. Mel Spectrogram (MET Happy class)

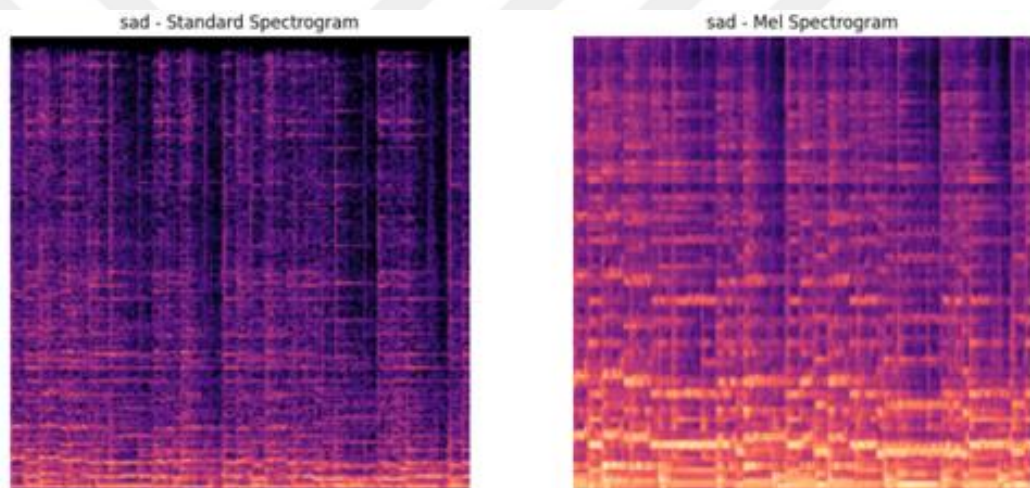


Figure 31: Standard Spectrogram vs. Mel Spectrogram (MET Sad class)

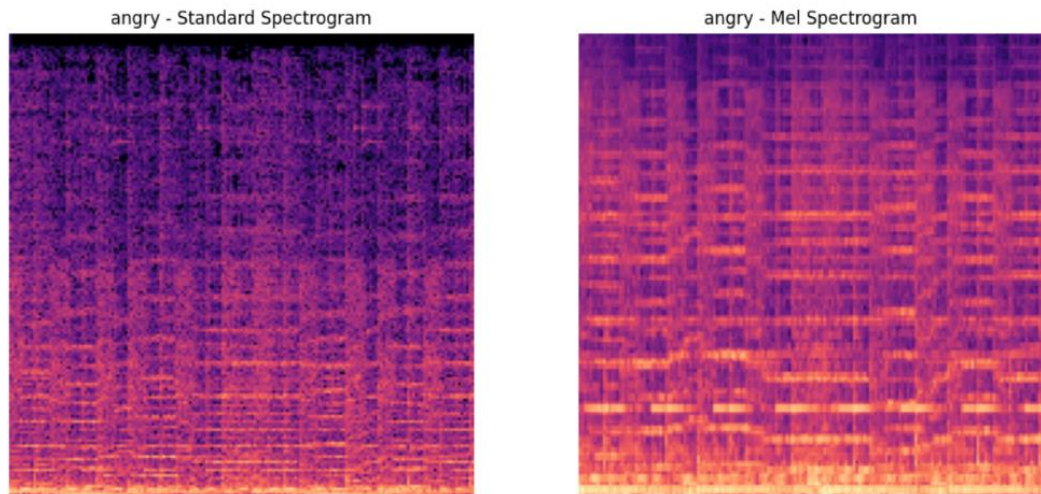


Figure 32: Standard Spectrogram vs. Mel Spectrogram (MET Angry class)

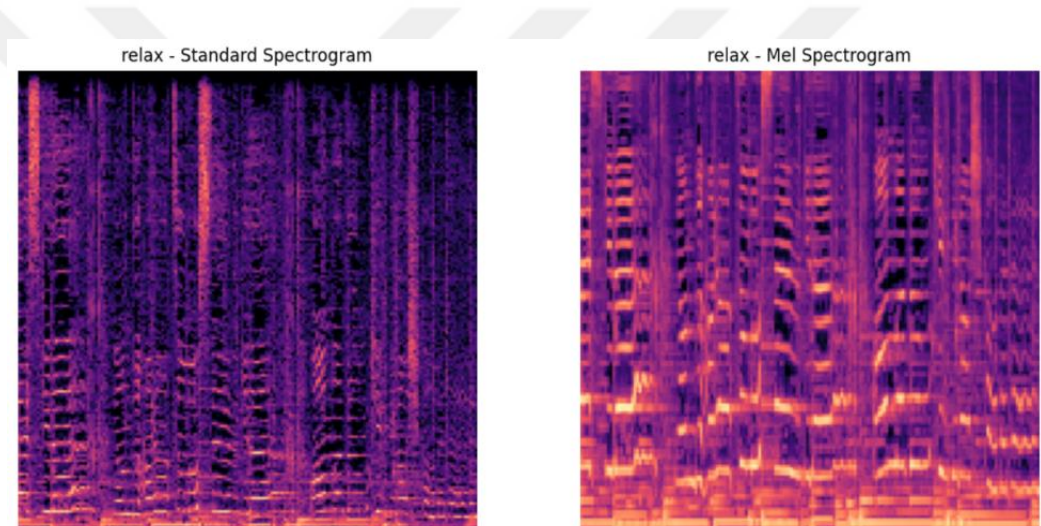


Figure 33: Standard Spectrogram vs. Mel Spectrogram (MET Relax class)

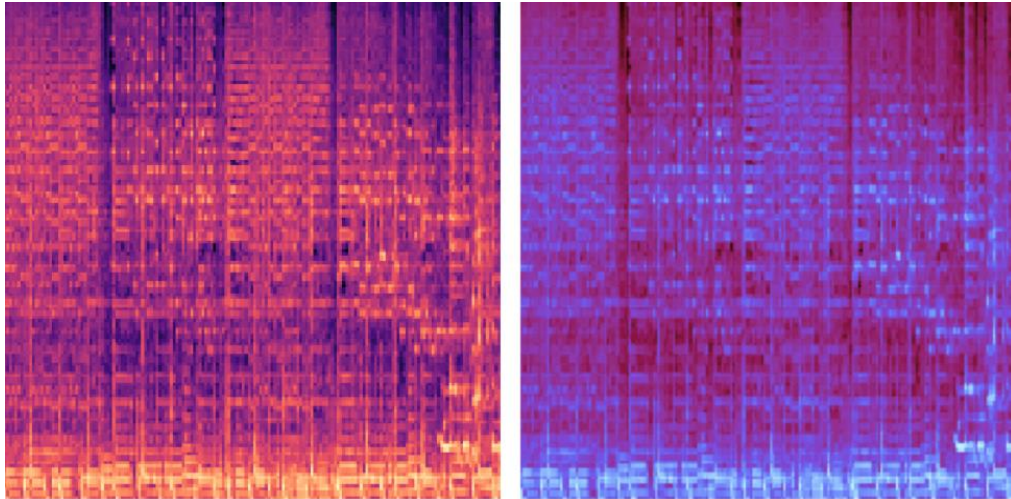


Figure 34: Original Spectrogram on left vs Resized Spectrogram (224*224*3) on right

- Here's a sample for the five classes: Happy, Sad, Romantic, Devotional and Party, of the Standard Spectrogram vs. Mel Spectrogram on the MER500 Dataset:

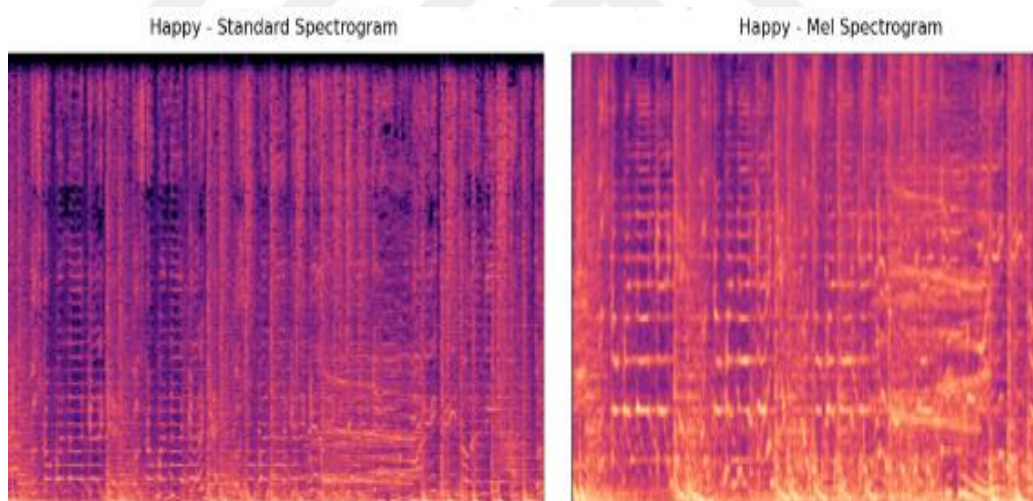


Figure 35: Standard Spectrogram vs. Mel Spectrogram (MER500 Happy class)

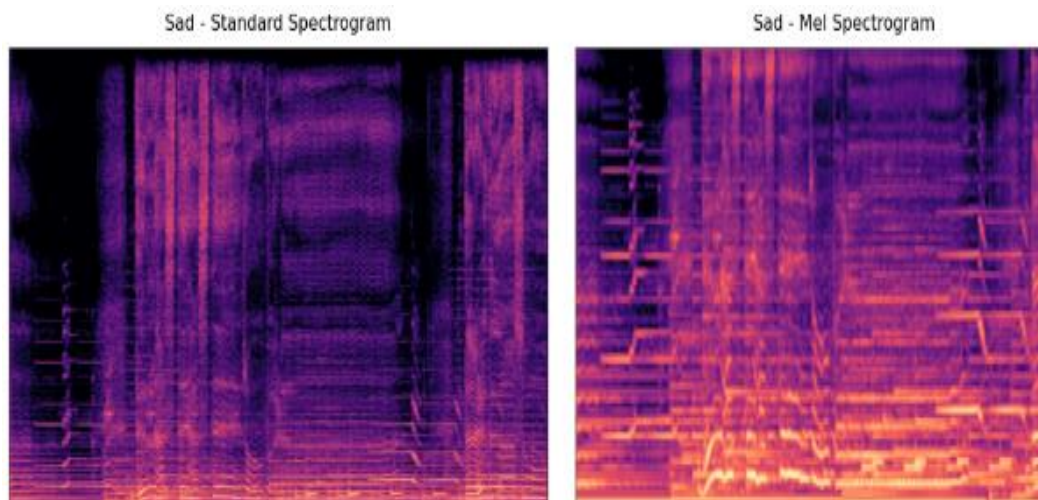


Figure 36: Standard Spectrogram vs. Mel Spectrogram (MER500 Sad class)

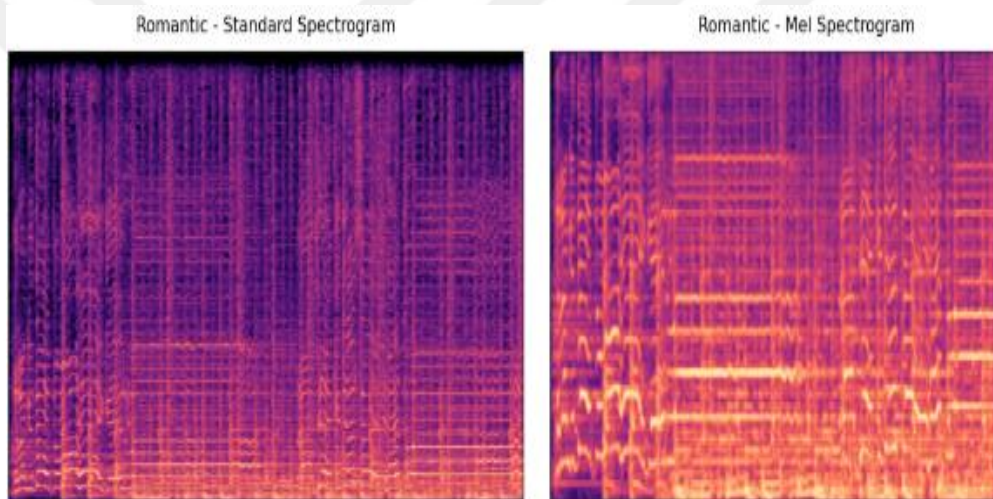


Figure 37: Standard Spectrogram vs. Mel Spectrogram (MER500 Romantic class)

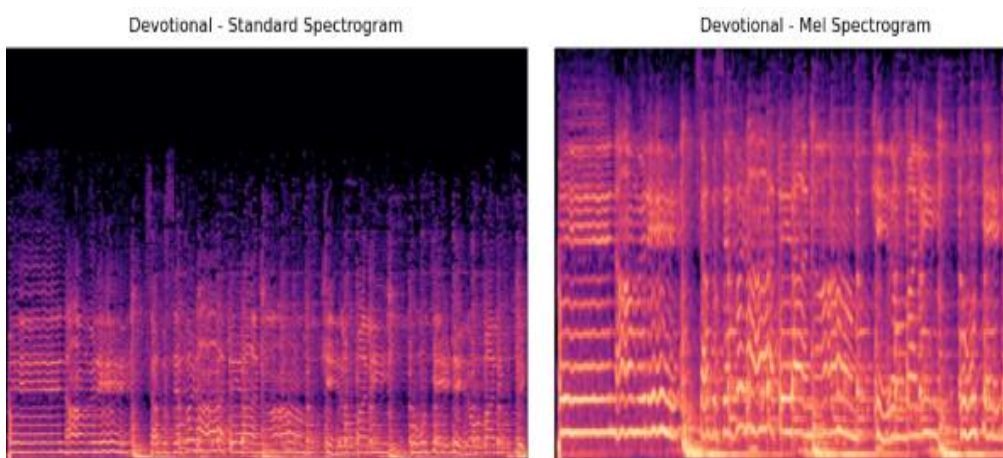


Figure 38: Standard vs. Mel Spectrogram (MER500 Devotional class)

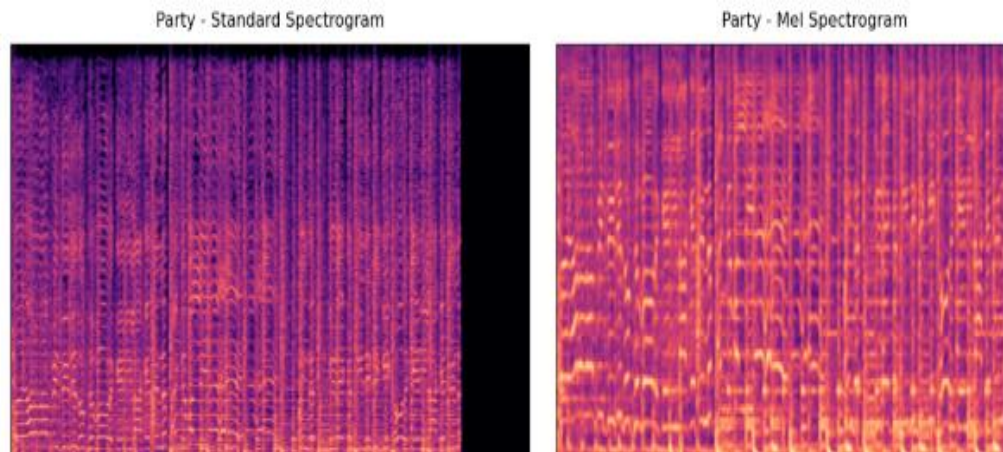


Figure 39: Standard Spectrogram vs. Mel Spectrogram (MER500 Party class)

- Samples of Mel-spectrogram images after resizing.

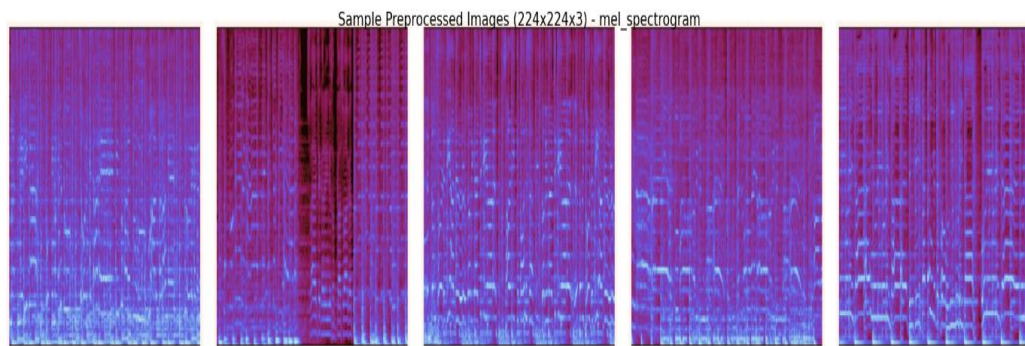


Figure 40: Resized Mel-Spectrogram images.

- Samples of Standard spectrogram images after resizing.

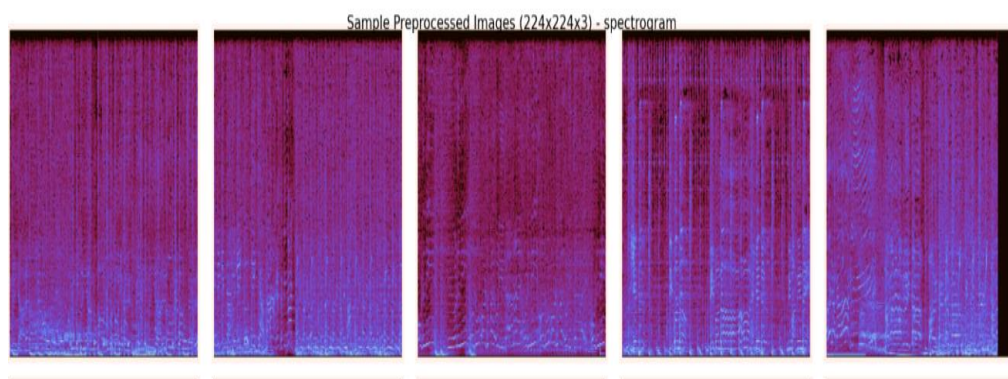


Figure 41: Resized Standard-Spectrogram images.

4.7 MODEL ARCHITECTURE

In the context of Modeling, I tried to freeze the Top layers of pre-trained models (ResNet50, DensNet121, Inception V3) in one attempt, and in another I made all the layers trainable, also added some additional layers in the last layers to improve my model for classifying music emotions.

➤ Additional Layers and Modifications:

- 1- GlobalAveragePooling2D: We reduced the dimensions of the spectrum in the early stages of our model using the GlobalAveragePooling2D layer. This layer reduces the spatial dimensions of the spectrum by calculating the mean of the data through the feature map to simplify the model, which in turn helps the model's ability to generalize and reduces excessive redundancy.
- 2- Dense layer (256 units, ReLU activation, L2 regularization with coefficient = 0.00001): We used a Dense layer containing 256 neural units that compute a linear combination of inputs to the previous layers, and we used the ReLU activation function, which helps to add nonlinear to the model in order to support the model to learn on more complex representations and to reduce overfitting, we used the generalization coefficient = 0.00001
- 3- BatchNormalization: It is the next layer we added to our model, which in turn normalizes and centralizes inputs from the previous layer to help stabilize and improve the model and reduce the need for large values.
- 4- Dropout (60%): It is the random dropping of 60% of the neural units during each step of training to prevent over-framing.
- 5- Final Dense layer with Softmax activation for multi-class classification with gradually changing number of filters.

4.8 MODEL TRAINING PROCESS

In the context of Training, I applied train-test split and Cross-Validation Technique with 5-fold cross-validation

4.8.1 Training process on Turkish Music Emotion dataset

In the first stage:

I have isolated 15% of the total data for testing and the remaining used for training & validation.

Number of epochs: 50

Batch size:32

Optimizer: sparse_categorical_crossentropy

Learning rate: 0.0001

In the second stage:

I have isolated 30% of the total data for testing and the remaining used for training & validation. That means I have 360 images for testing and 840 images for training & validation.

Number of epochs: 200

Batch size:32

Optimizer: categorical_crossentropy

Learning rate: 0.0001

4.8.2 Training process on MER500 dataset

In the first stage:

I have isolated 20% of the total data for testing and the remaining used for training & validation.

Number of epochs: 50

Batch size:32

Optimizer: sparse_categorical_crossentropy

Learning rate: 0.0001

In the second stage:

I have isolated 20% of the total data for testing and the remaining used for training & validation.

Number of epochs: 200

Batch size:32

Optimizer: categorical_crossentropy

Learning rate: 0.0001

4.9 EVALUATION METRICS

Performance metrics for the multi-class classification model:

4.9.1 Accuracy

The ratio of correct data classification to the total data [75].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4.1)$$

- Mean accuracy across cross-validation folds
- Test accuracy on the held-out test set

4.9.2 Confusion matrix

It is a tool to display the number of data that has been correctly or wrongly classified as shown below [76]:

True positive (TP): Data correctly classified as positive.

True negative (TN): Data correctly classified as negative.

False Positive (FP) Data classified false as positive.

False Negative (FN): Data classified false as negative.

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figure 42: CM formula

- Generated for each model to visualize performance across different emotion categories.

4.9.3 Classification Report

It is a tool used to evaluate the performance of a classification algorithm by providing a detailed breakdown of various metrics. It includes precision, recall, F1-score, and support for each class

- Generated for each model to visualize performance.

Note: Support refers to the number of actual occurrences of each class in the dataset, providing context to the precision, recall, and F1-score metrics.

4.9.4 Recall

The ratio of data that was correctly classified as positive to the total data that belonged to the positive category [75].

$$Recall = \frac{TP}{TP+FN} \quad (4.2)$$

4.9.5 Precision

The ratio of data that was correctly rated as positive to the total data that was actually rated positive [75].

$$precision = \frac{TP}{TP+FP} \quad (4.3)$$

4.9.6 F1 Score

It is a measure of model classification performance, as it is a mixture of precision and recall aimed at providing a balance between them [75].

$$2 \cdot \left(\frac{precision \cdot recall}{precision + recall} \right) \quad (4.4)$$

Precision, Recall, and F1-Score existed in the Classification Report and were calculated but specific formulas were not provided in our works.

CHAPTER V

EXPERIMENTS AND RESULTS

The projects undertaken during the previous studies on the MET as well as the datasets of MER500, the same datasets used in my thesis, showed the classification accuracies of 0.892 and 0.84 respectively. I will present the results of my study in detail in succession, and compare them with the previous values.

5.1 EXPERIMENTS WITH TURKISH MUSIC EMOTION DATASET

The result for each model shown below:

5.1.1 Results with Standard Spectrogram

First Stage:

1) ResNet50 Model:

ResNet50 - Test Accuracy: 0.97

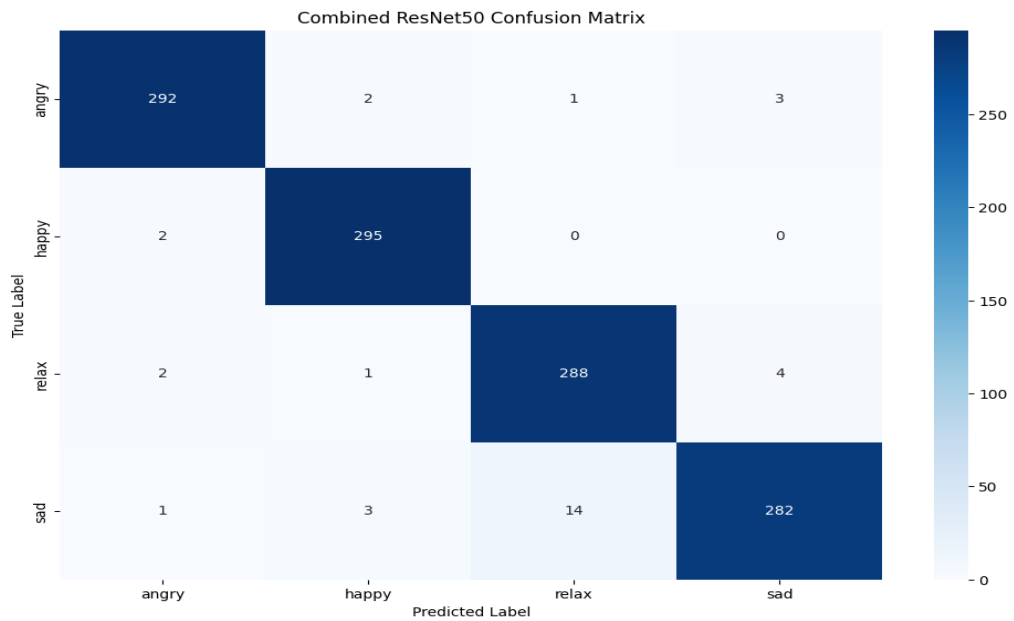


Figure 43: Confusion Matrix - MET ResNet50 Standard-Spectrogram.

	precision	recall	f1-score	support
0	0.98	0.98	0.98	298
1	0.98	0.99	0.99	297
2	0.95	0.98	0.96	295
3	0.98	0.94	0.96	300
accuracy			0.97	1190
macro avg	0.97	0.97	0.97	1190
weighted avg	0.97	0.97	0.97	1190

Figure 44: Classification report - MET ResNet50 on Standard-Spectrogram.

2) DenseNet121 Model:

DenseNet121 - Test Accuracy: 0.961

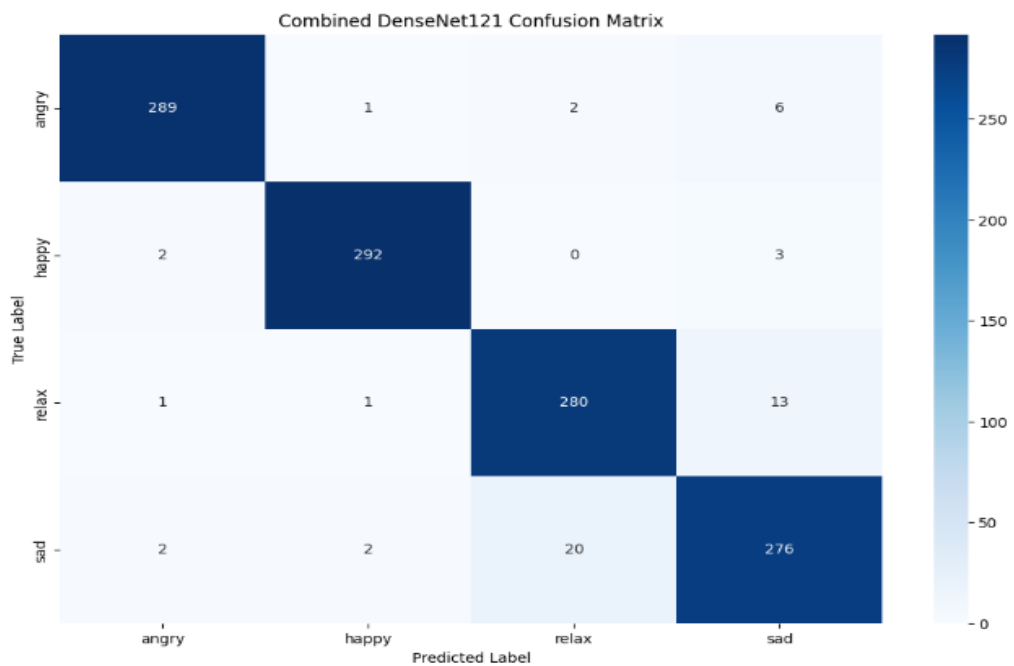


Figure 45: Confusion Matrix - MET DenseNet121 Standard-Spectrogram.

	precision	recall	f1-score	support
0	0.98	0.97	0.98	298
1	0.99	0.98	0.98	297
2	0.93	0.95	0.94	295
3	0.93	0.92	0.92	300
accuracy			0.96	1190
macro avg	0.96	0.96	0.96	1190
weighted avg	0.96	0.96	0.96	1190

Figure 46: Classification report - MET DenseNet121 on Standard-Spectrogram

3) Inception V3 Model:

Inception V3 - Test Accuracy: 0.93

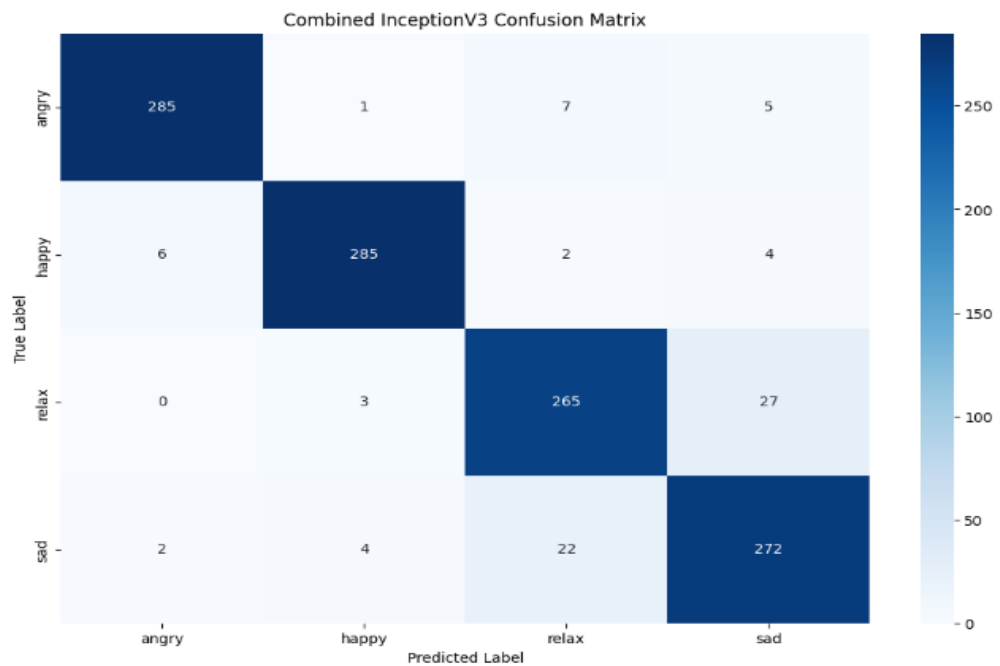


Figure 47: Confusion Matrix - MET Inception V3 on Standard Spectrogram.

	precision	recall	f1-score	support
0	0.97	0.96	0.96	298
1	0.97	0.96	0.97	297
2	0.90	0.90	0.90	295
3	0.88	0.91	0.89	300
accuracy	0.93 1190			
macro avg	0.93	0.93	0.93	1190
weighted avg	0.93	0.93	0.93	1190

Figure 48: Classification report - MET Inception V3 on Standard-Spectrogram.

Second Stage:

A. ResNet50 Model:

ResNet50 - Mean Accuracy: 0.9536 (± 0.0542)

ResNet50 - Test Accuracy: 0.8417

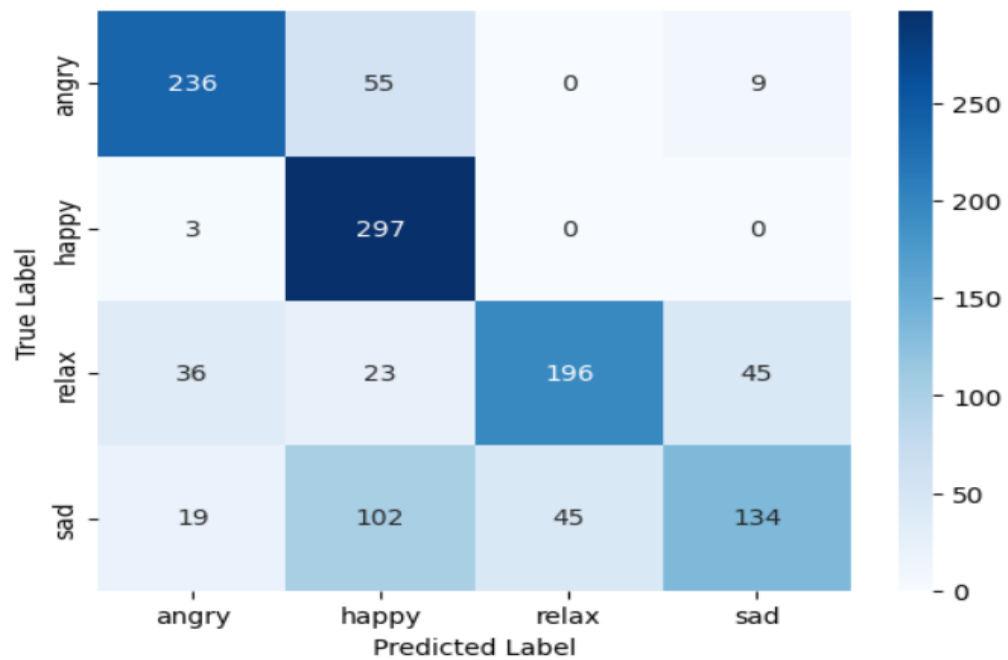


Figure 49: Confusion Matrix- MET ResNet50 on Standard Spectrogram.

B. DenseNet121 Model:

DenseNet121 - Mean Accuracy: 0.9214 (± 0.0666)

DenseNet121 - Test Accuracy: 0.8556

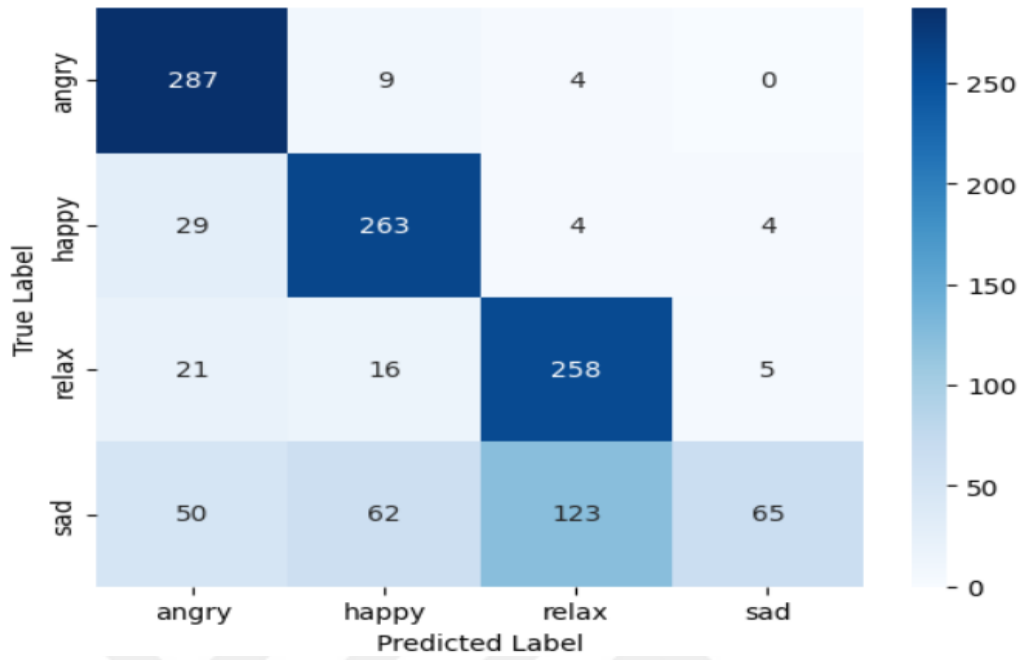


Figure 50: Confusion Matrix - MET DenseNet121 on Standard Spectrogram.

C. Inception V3 Model:

Inception V3 - Mean Accuracy: 0.8810 (± 0.1078)

Inception V3 - Test Accuracy: 0.8278

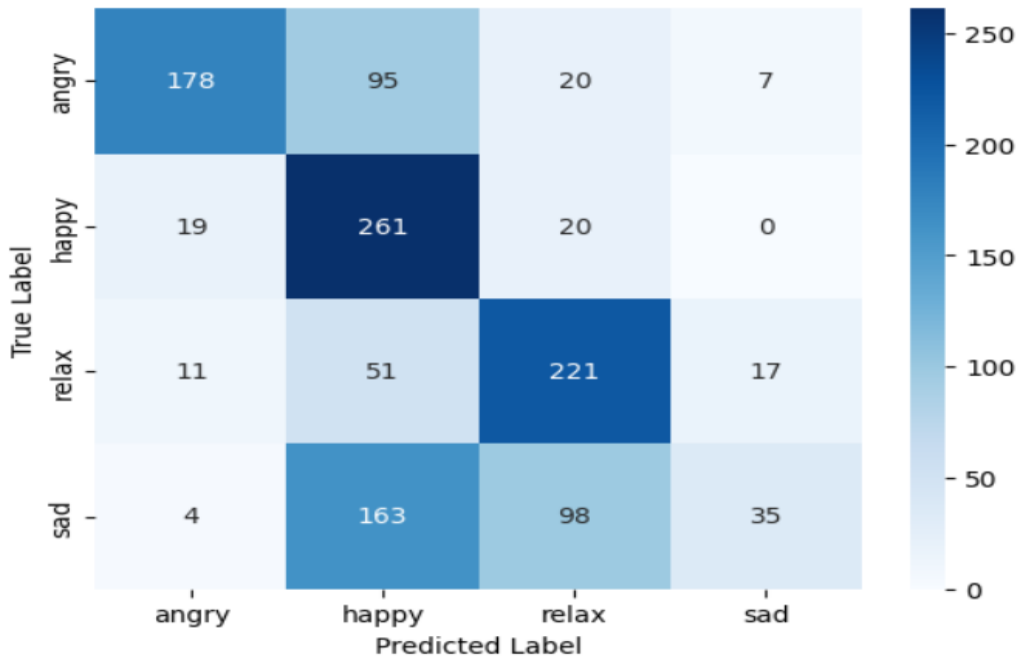


Figure 51: Confusion Matrix- MET Inception V3 on Standard Spectrogram.

5.1.2 Results with Mel-Spectrogram

1. ResNe50 Model:

ResNet50 - Mean Accuracy: 0.9655 (± 0.0443)

ResNet50 - Test Accuracy: 0.9194

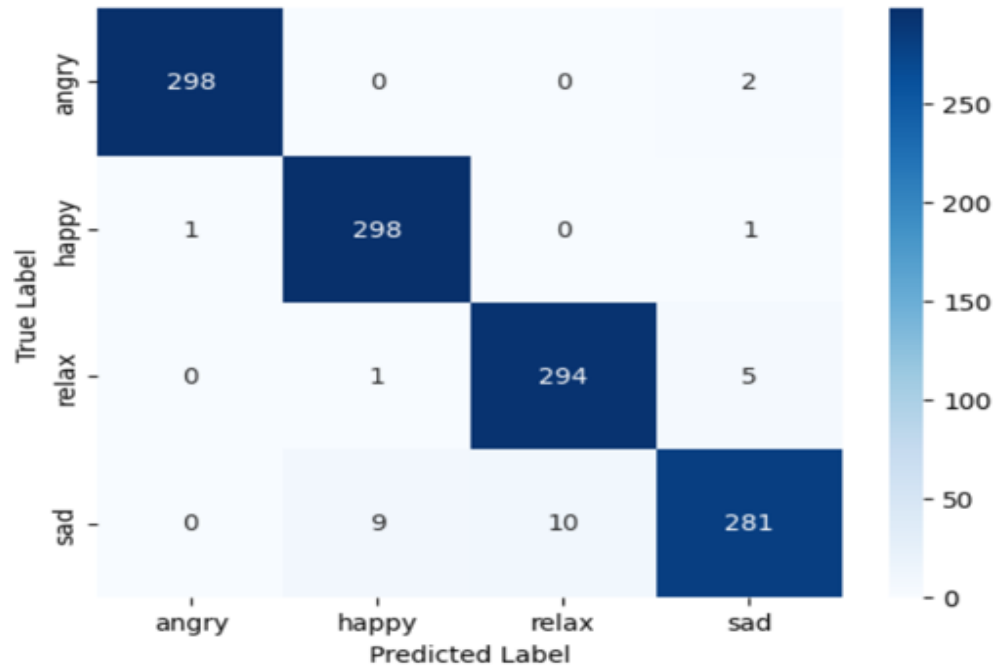


Figure 52: Confusion Matrix - MET ResNet50 on Mel-Spectrogram.

2. DenseNet121 Model:

DenseNet121 - Mean Accuracy: 0.9393 (± 0.0515)

DenseNet121 - Test Accuracy: 0.9167

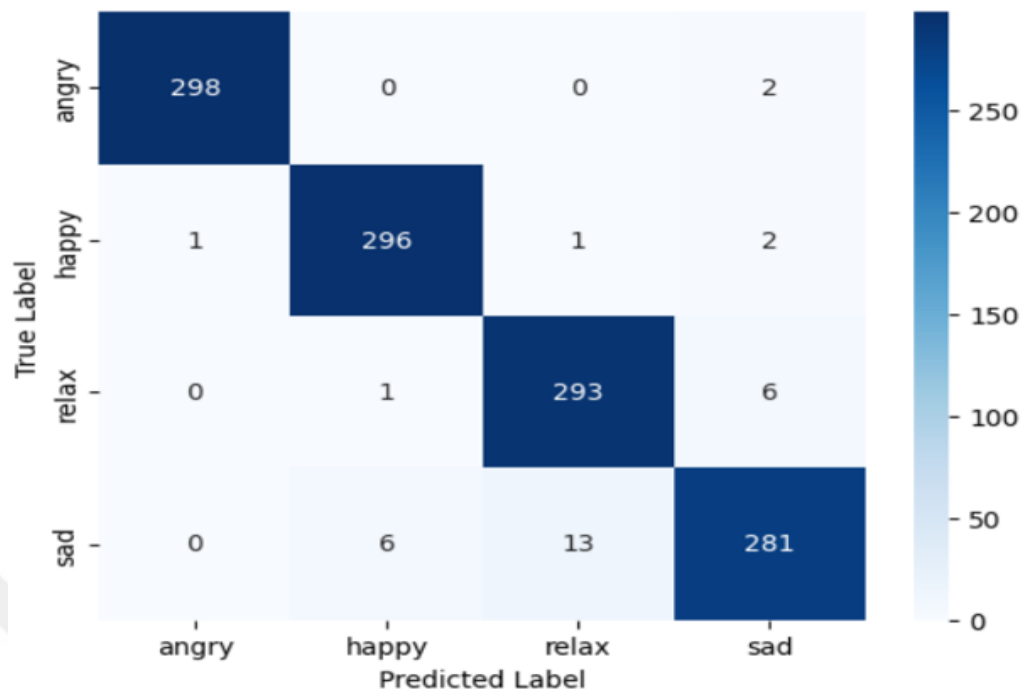


Figure 53: Confusion Matrix - MET DenseNet121 on Mel-Spectrogram.

3. Inception V3 Model:

Inception V3 - Mean Accuracy: 0.9119 (± 0.0829)

Inception V3 - Test Accuracy: 0.8250

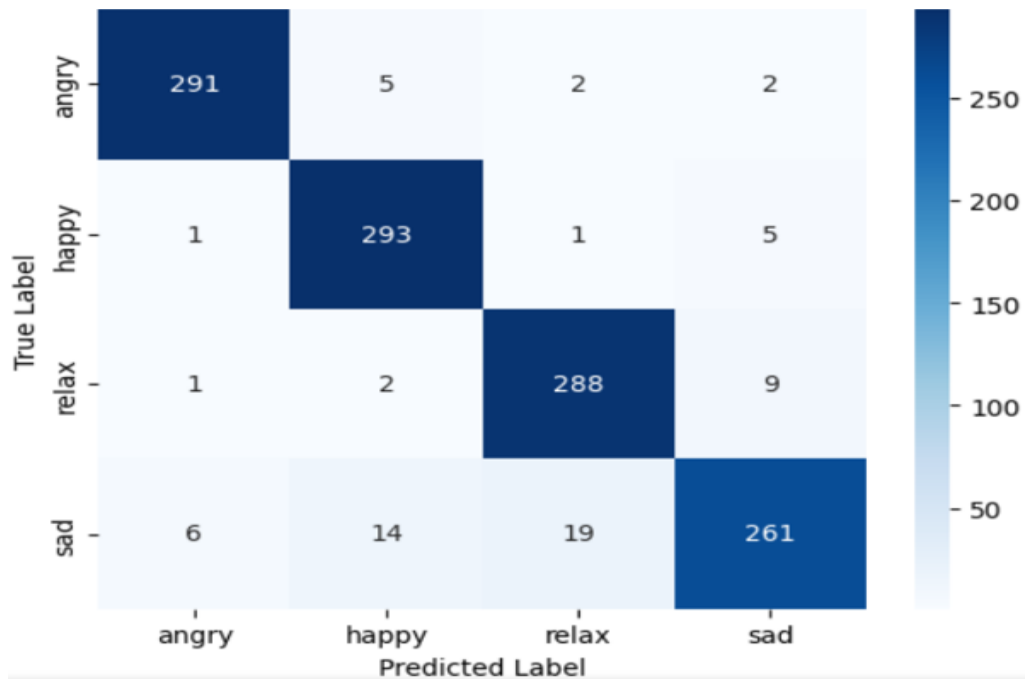


Figure 54: Confusion Matrix - MET Inception V3 on Mel-Spectrogram

5.1.3 Comparison of Models' Performance

I provide the final results as shown in the tables below, we can see the mean and test accuracies, The model that got the highest mean and test accuracy was ResNet50 with Mel-Spectrogram in terms of the second attempt, and in terms of the first one the ResNet50 with Standard Spectrogram achieved 0.97 which is reasonable too.

Table 9: Test Accuracy & Mean Accuracy with Standard-Spectrogram

Spectrogram	Resnet50	DenesNet121	Inception v3
Test Accuracy	0.8417	0.8556	0.8278
Mean Accuracy	0.9536	0.9214	0.881

Table 10: Test Accuracy & Mean Accuracy with Mel-Spectrogram

Mel_Spectrogram	Resnet50	DenesNet121	Inception v3
Test Accuracy	0.9194	0.9167	0.825
Mean Accuracy	0.9655	0.9393	0.9119

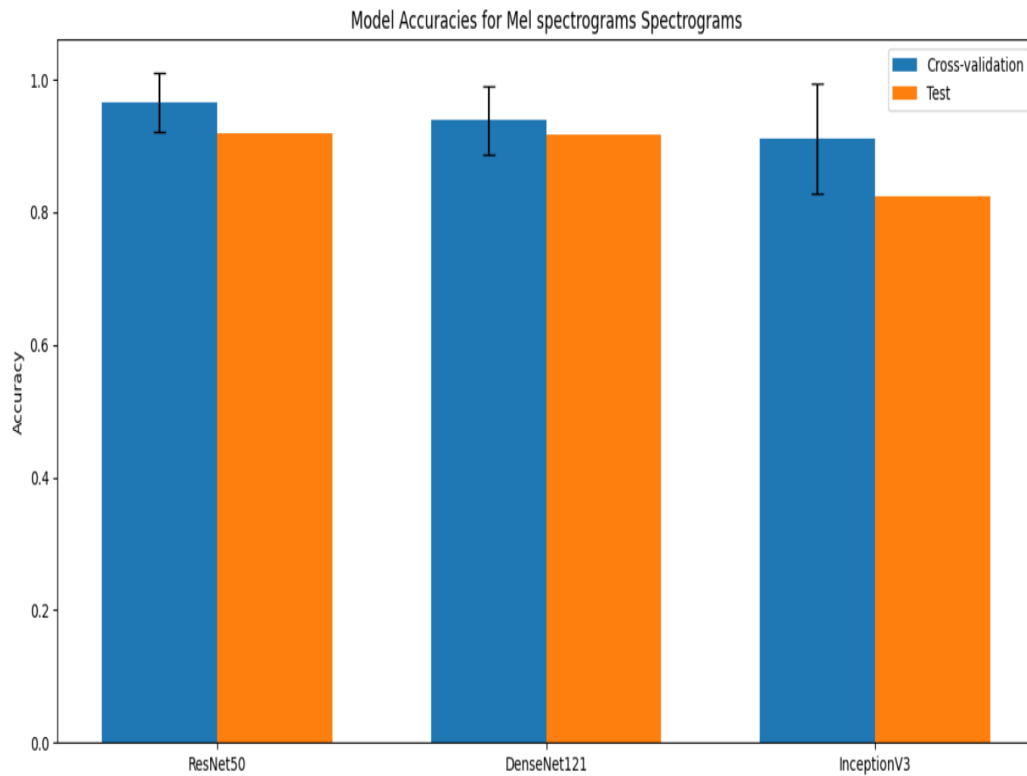


Figure 55: Pre-trained Models on MET with Mel-Spectrogram bar plot.

	precision	recall	f1-score	support
angry	1.00	0.99	0.99	300
happy	0.97	0.99	0.98	300
relax	0.97	0.98	0.97	300
sad	0.97	0.94	0.95	300
accuracy			0.98	1200
macro avg	0.98	0.98	0.98	1200
weighted avg	0.98	0.98	0.98	1200

Figure 56: Classification Report - MET ResNet50 on Mel-Spectrogram.

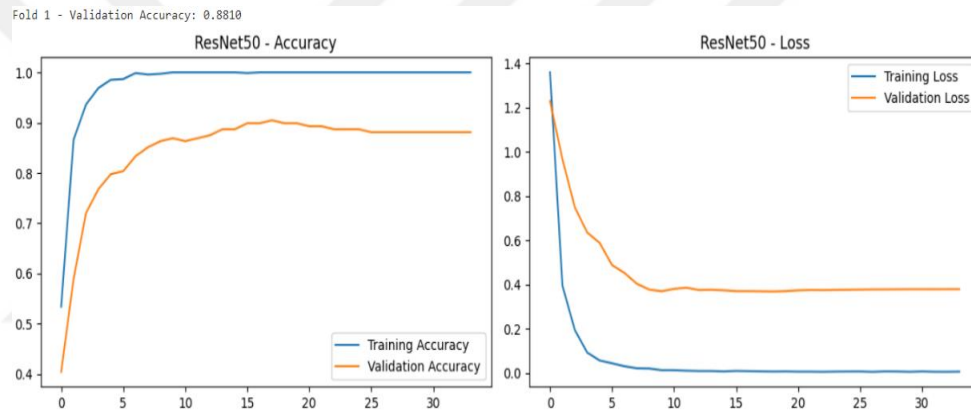


Figure 57: MET Best Model ResNet50 with Mel-Spectrogram fold-1.

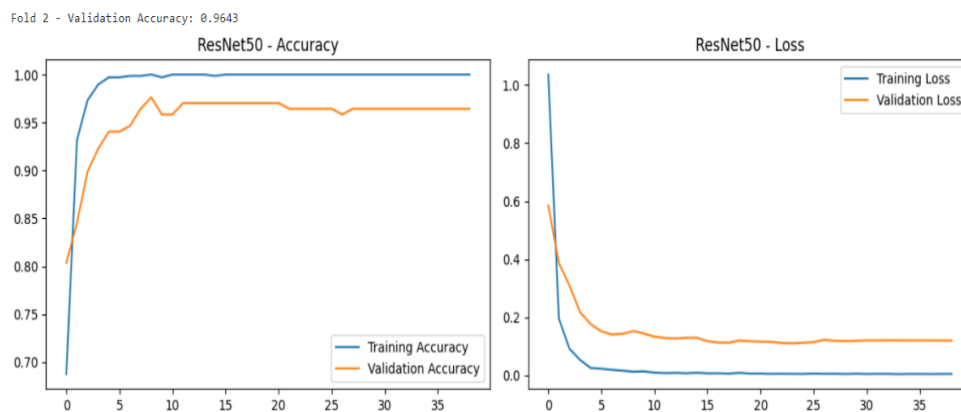


Figure 58: MET Best Model ResNet50 with Mel-Spectrogram fold-2.

Fold 3 - Validation Accuracy: 0.9821

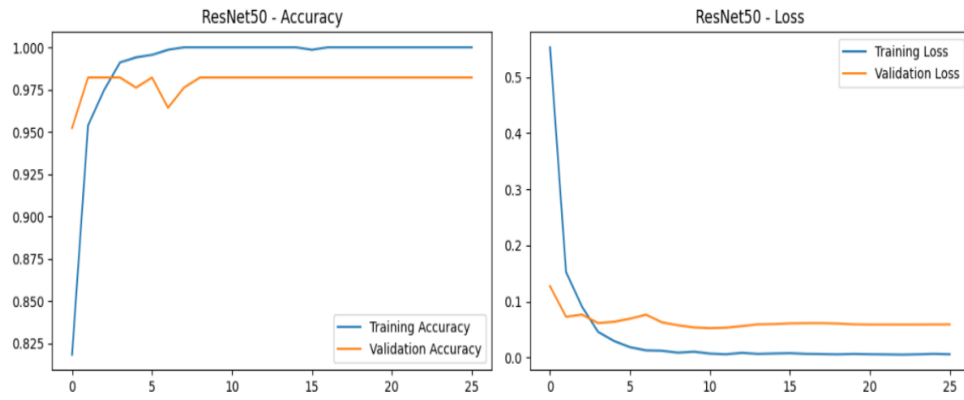


Figure 59: MET Best Model ResNet50 with Mel-Spectrogram fold-3.

Fold 4 - Validation Accuracy: 1.0000

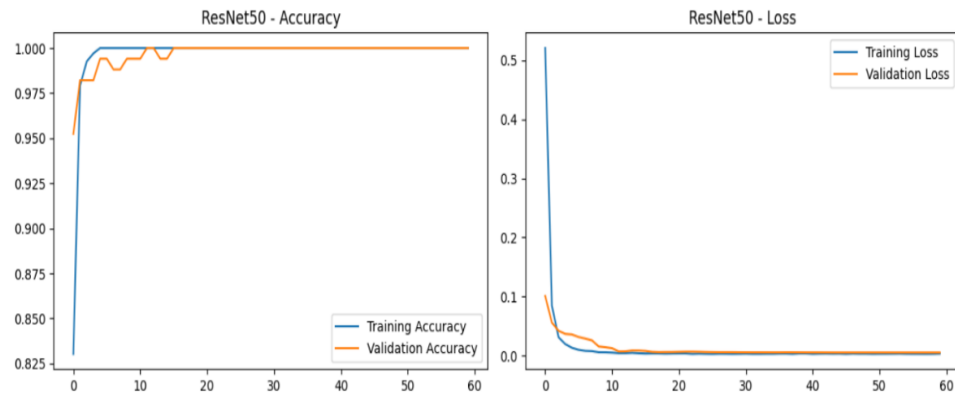


Figure 60: MET Best Model ResNet50 with Mel-Spectrogram fold-4.

Fold 5 - Validation Accuracy: 1.0000

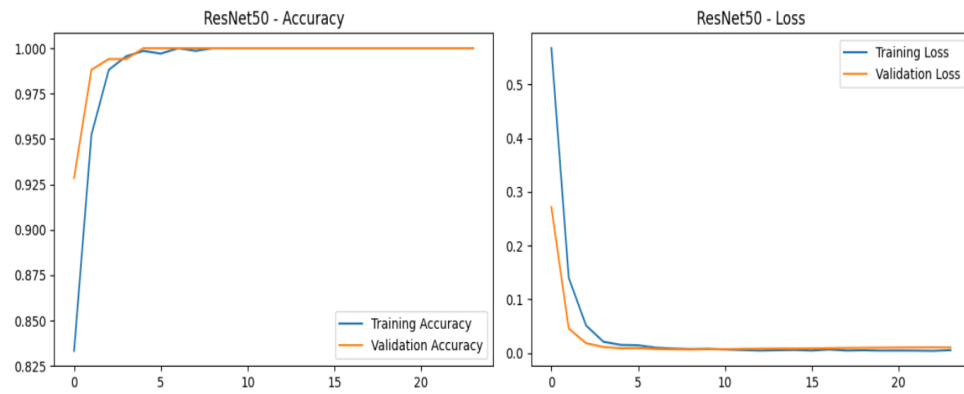


Figure 61: MET Best Model ResNet50 with Mel-Spectrogram fold-5.

5.2 EXPERIMENTS WITH MER500 DATASET

5.2.1 Results with Standard Spectrogram

5.2.1.1 Unbalanced Data

First Stage:

1) ResNet50 Model:

ResNet50 - Test Accuracy:0.82

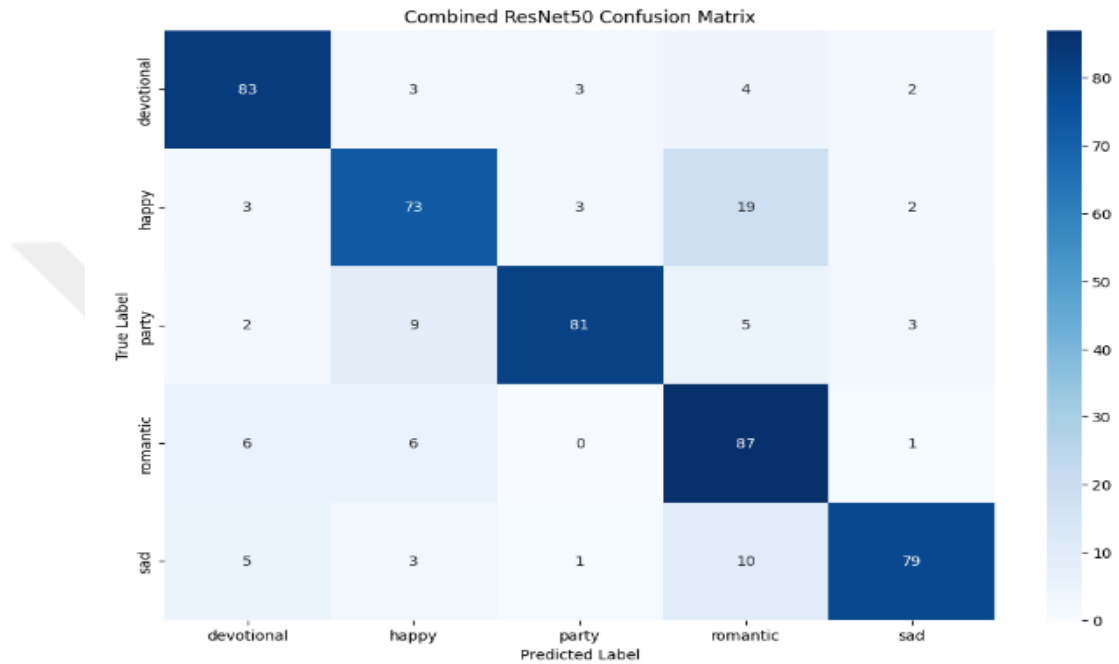


Figure 62: Confusion Matrix - MER500 ResNet50 on Standard-Spectrogram.

	precision	recall	f1-score	support
0	0.84	0.87	0.86	95
1	0.78	0.73	0.75	100
2	0.92	0.81	0.86	100
3	0.70	0.87	0.77	100
4	0.91	0.81	0.85	98
accuracy			0.82	493
macro avg	0.83	0.82	0.82	493
weighted avg	0.83	0.82	0.82	493

Figure 63: Classification Report - MER500 ResNet50 on Standard-Spectrogram.

2) DenseNet121 Model:

DenseNet121 - Test Accuracy: 0.77

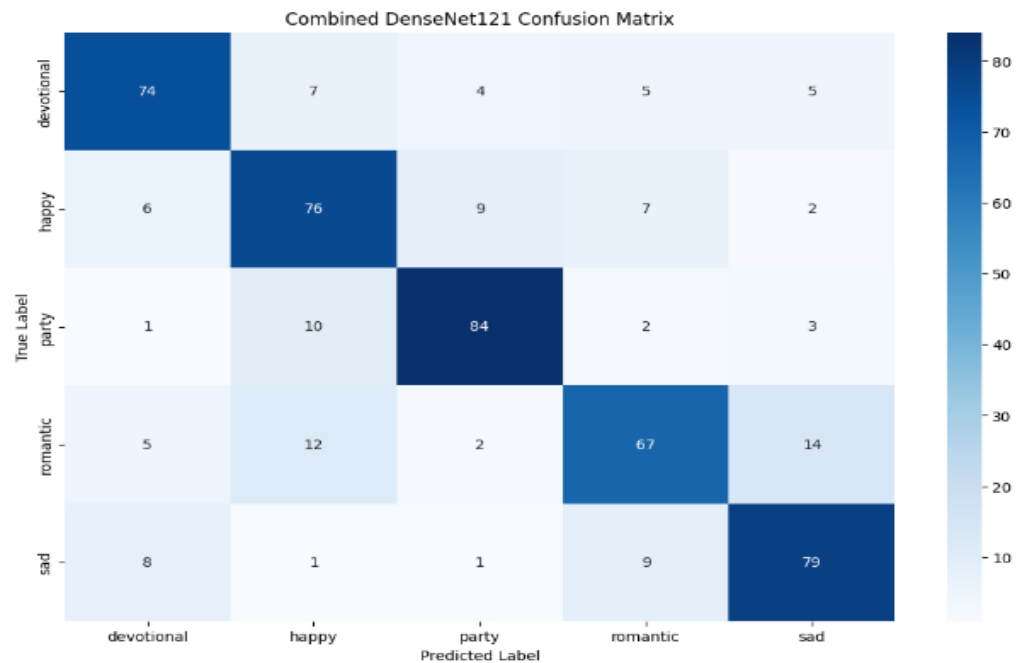


Figure 64: Confusion Matrix - MER500 DenseNet121 on Standard Spectrogram.

	precision	recall	f1-score	support
0	0.79	0.78	0.78	95
1	0.72	0.76	0.74	100
2	0.84	0.84	0.84	100
3	0.74	0.67	0.71	100
4	0.77	0.81	0.79	98
accuracy			0.77	493
macro avg	0.77	0.77	0.77	493
weighted avg	0.77	0.77	0.77	493

Figure 65: Classification Report - MER500 DenseNet121 on Standard Spectrogram.

3) Inception V3 Model:

Inception V3 - Test Accuracy: 0.53

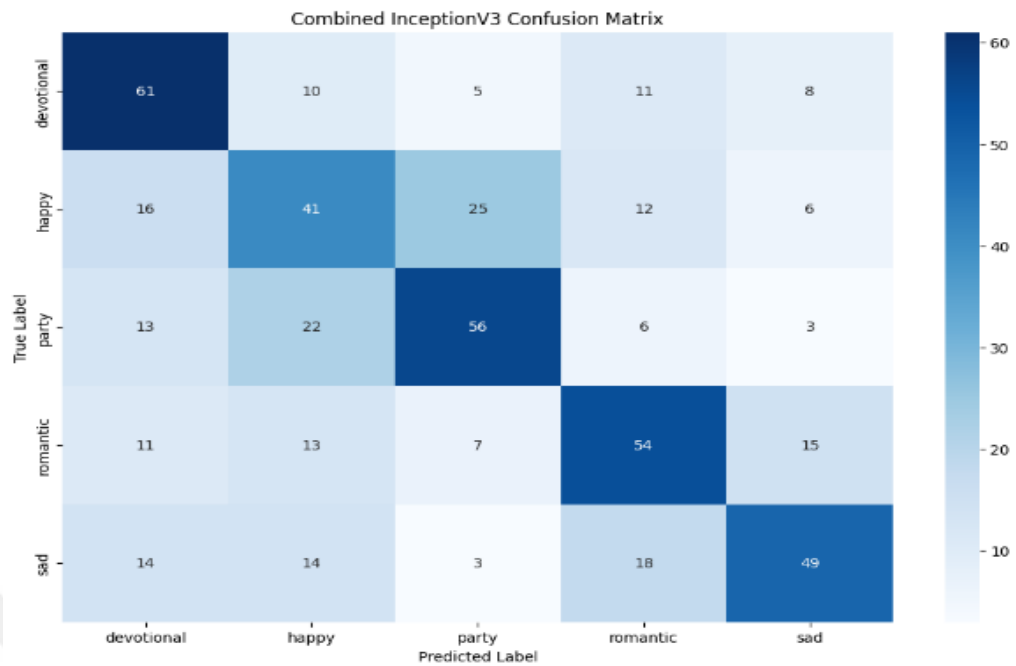


Figure 66: Confusion Matrix -MER500 Inception V3 on Standard Spectrogram.

	precision	recall	f1-score	support
0	0.53	0.64	0.58	95
1	0.41	0.41	0.41	100
2	0.58	0.56	0.57	100
3	0.53	0.54	0.54	100
4	0.60	0.50	0.55	98
accuracy			0.53	493
macro avg	0.53	0.53	0.53	493
weighted avg	0.53	0.53	0.53	493

Figure 67: Classification report - MER500 Inception V3 on Standard Spectrogram

Second Stage:

A. ResNet50 Model:

ResNet50 - Mean Accuracy: 0.9130

ResNet50 - Test Accuracy: 0.7521

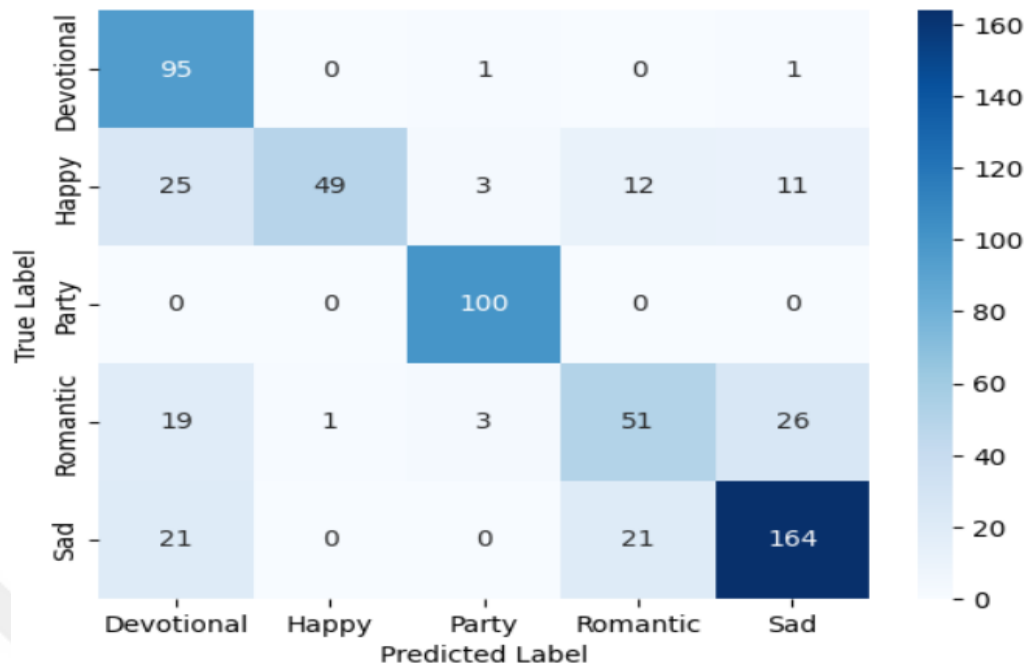


Figure 68: Confusion Matrix - MER500 ResNet50 (unbalance - Standard Spectrogram)

B. DenseNet121 Model:

DenseNet121 - Mean Accuracy: 0.8736

DenseNet121 - Test Accuracy: 0.7521

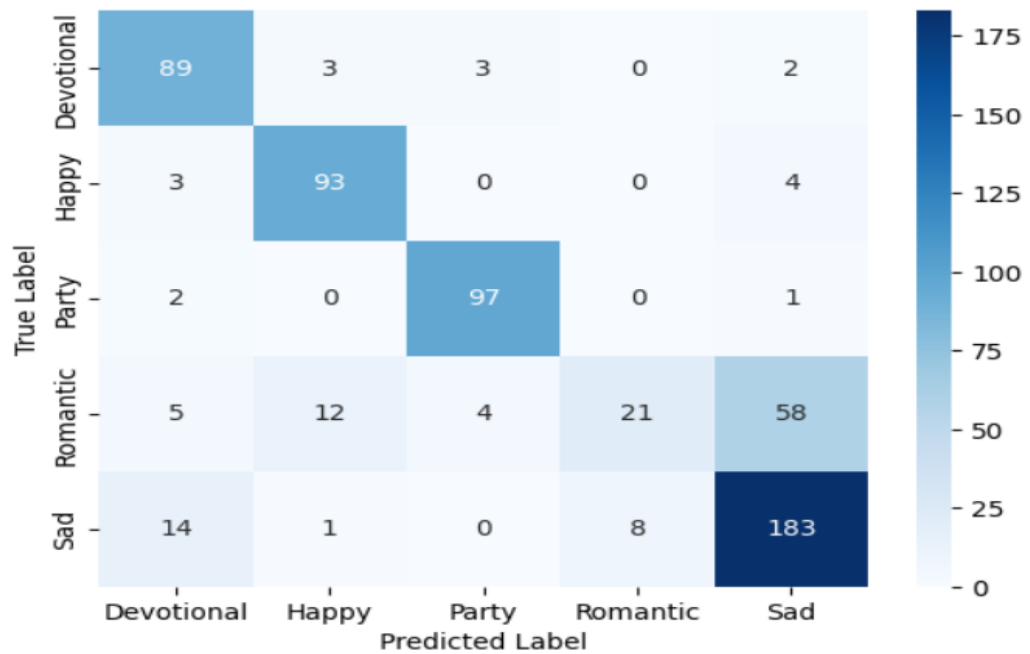
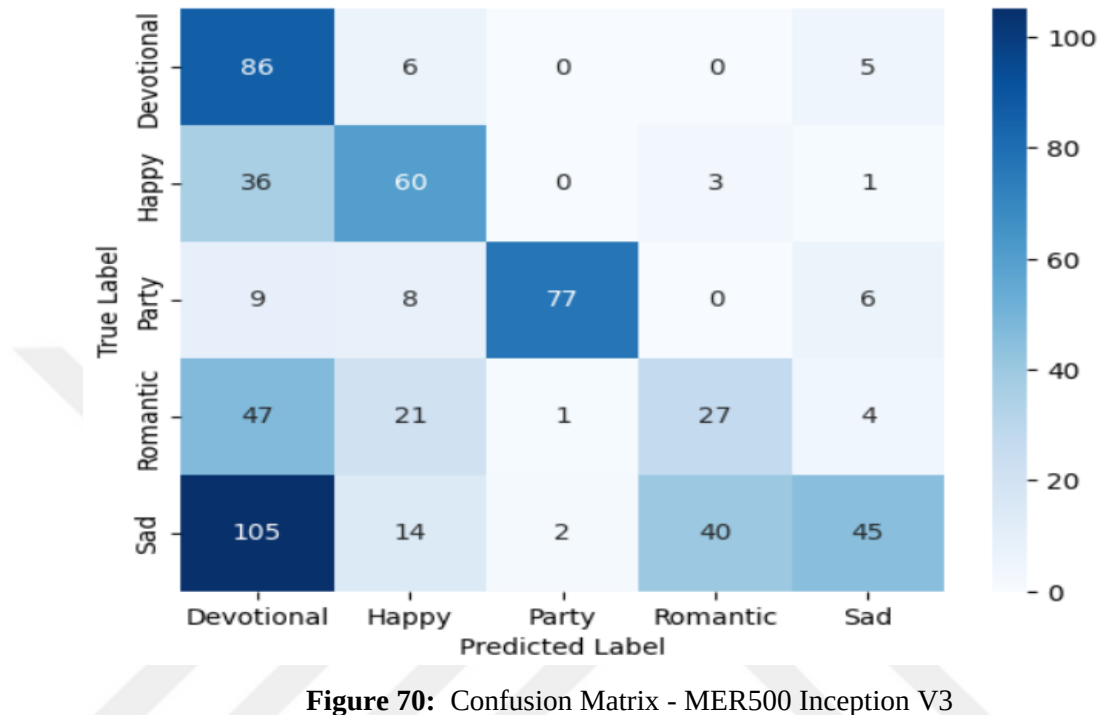


Figure 69: Confusion Matrix - MER500 DenseNet121 (unbalance - Standard Spectrogram)

C. Inception V3 Model:

Inception V3 - Mean Accuracy: 0.7970

Inception V3 - Test Accuracy: 0.7025



5.2.1.2 Undersampled Data

a. ResNet50 Model:

ResNet50 - Mean Accuracy: 0.9588

ResNet50 - Test Accuracy: 0.8866

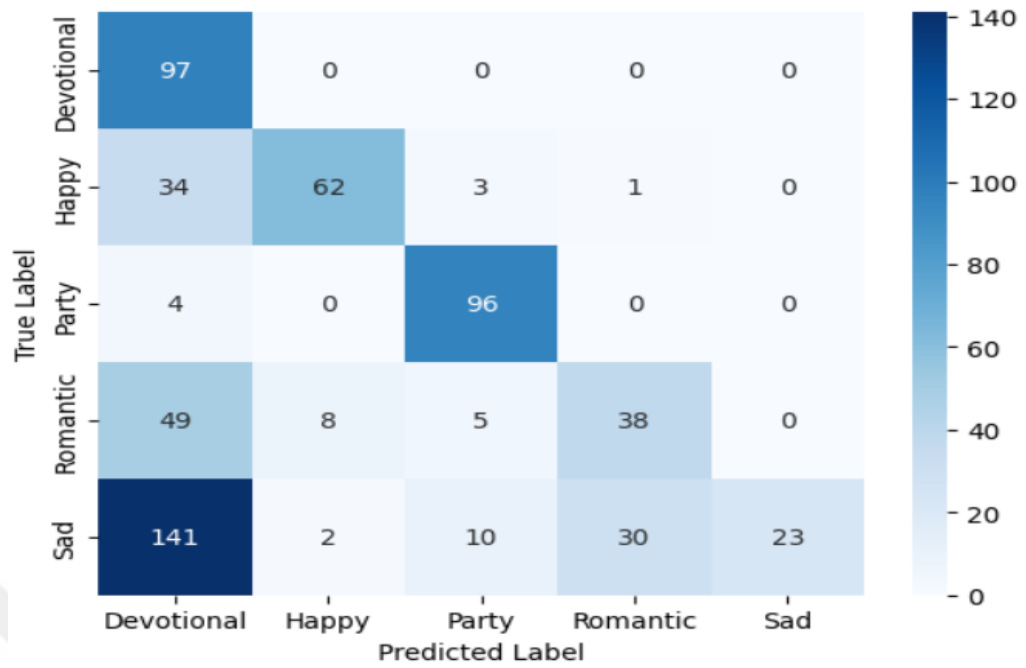


Figure 71: Confusion Matrix - MER500 ResNet50 (undersampling - Standard spectrogram)

b. DenseNet121 Model:

DenseNet121 - Mean Accuracy: 0.9846

DenseNet121 - Test Accuracy: 0.8969

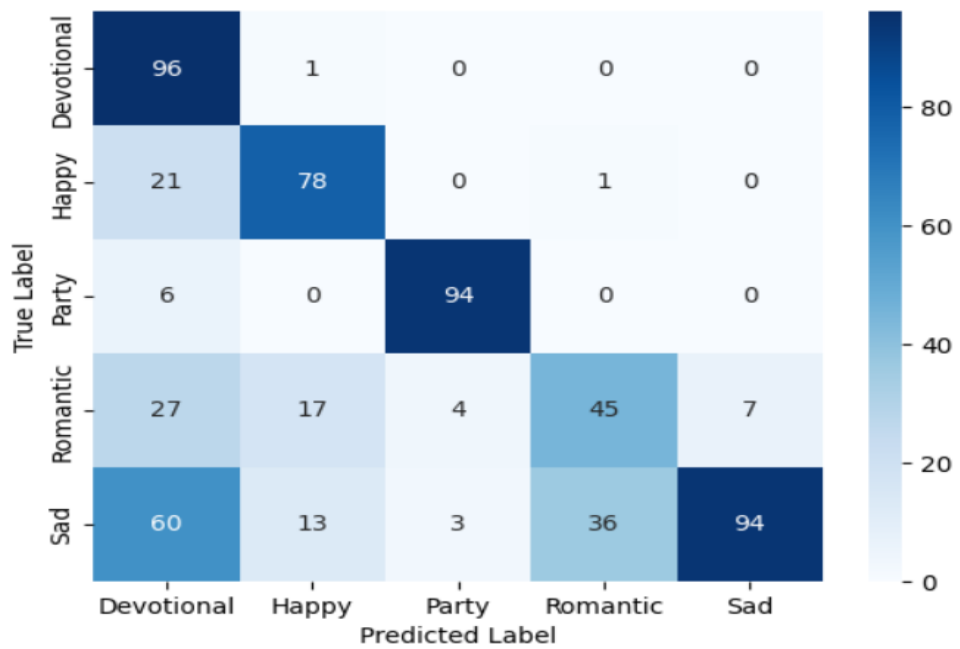


Figure 72: MER500 DenseNet121 (undersampling - Standard Spectrogram)

c. Inception V3 Model:

Inception V3 - Mean Accuracy: 0.9365

Inception V3 - Test Accuracy: 0.8454

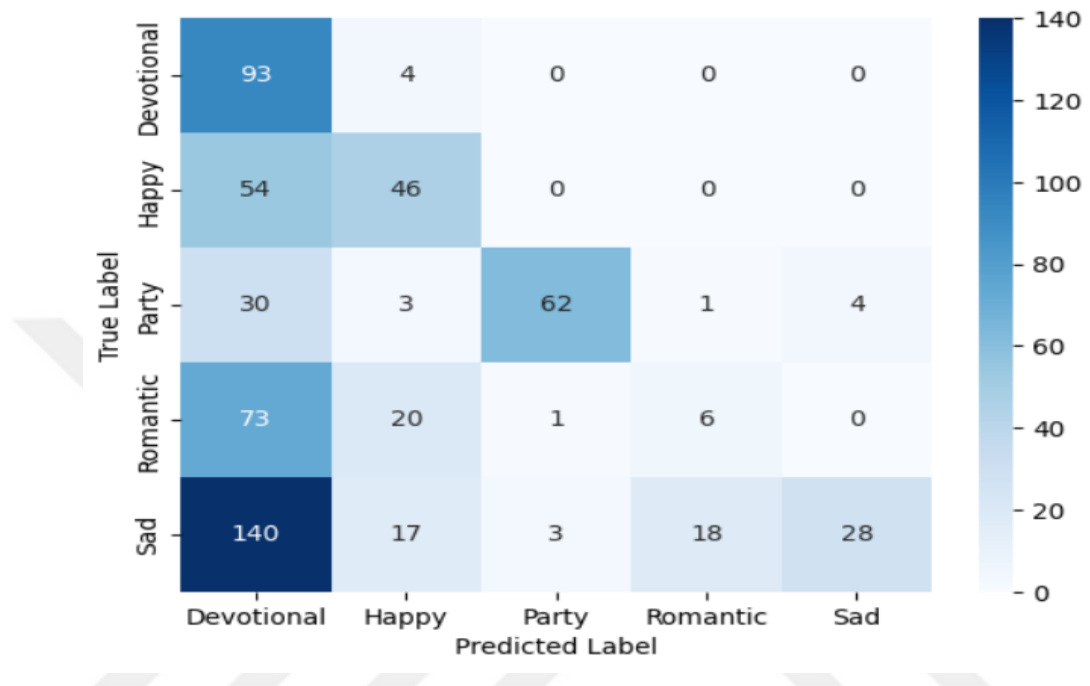


Figure 73: Confusion Matrix - MER500 Inception V3 (undersampling - Standard Spectrogram)

5.2.1.3 Oversampled Data

1. ResNet50 Model:

ResNet50 - Mean Accuracy: 0.9854

ResNet50 - Test Accuracy: 0.9563

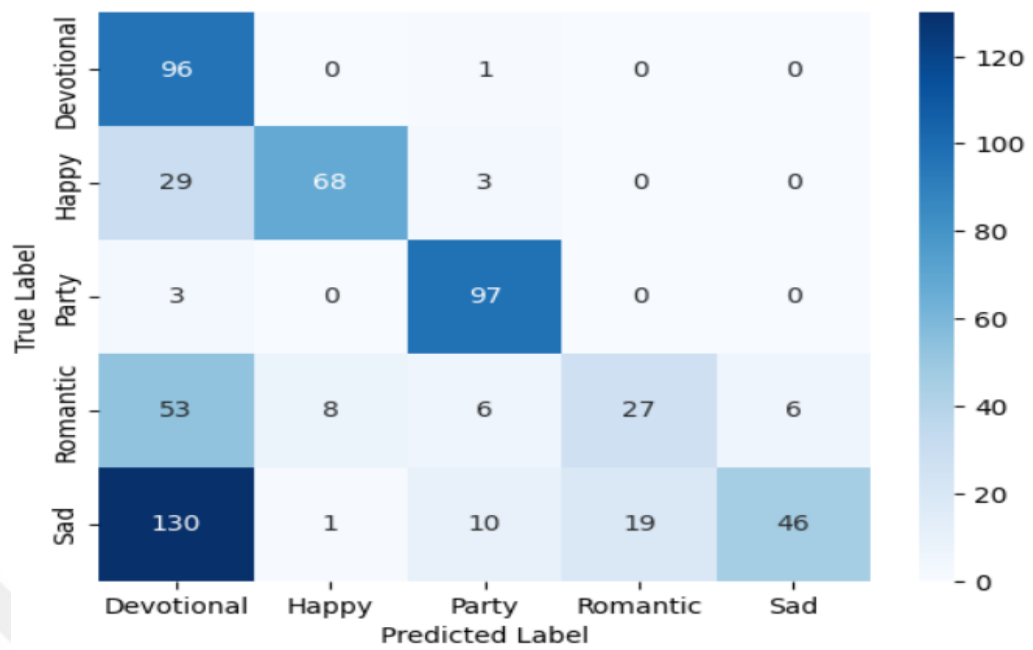


Figure 74: Confusion Matrix - MER500 ResNet50 (oversampling - Standard Spectrogram)

2. DenseNet121 Model:

DenseNet121 - Mean Accuracy: 0.9976

DenseNet121 - Test Accuracy: 0.9709

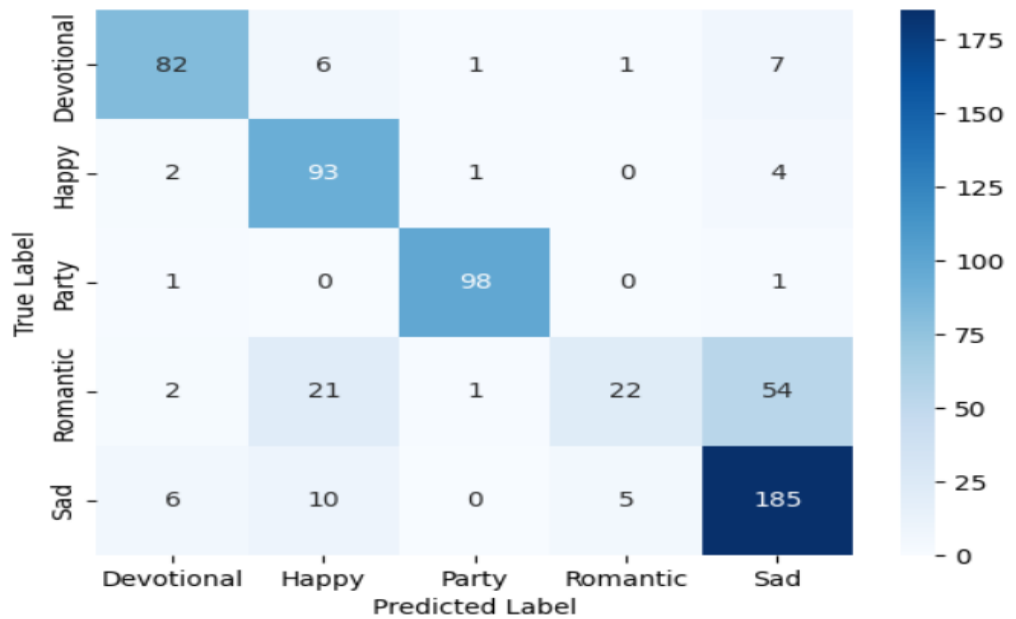


Figure 75: Confusion Matrix - MER500 DenseNet121(oversampling - Standard Spectrogram)

3. Inception V3 Model:

Inception V3 - Mean Accuracy: 0.9794

Inception V3 - Test Accuracy: 0.9563

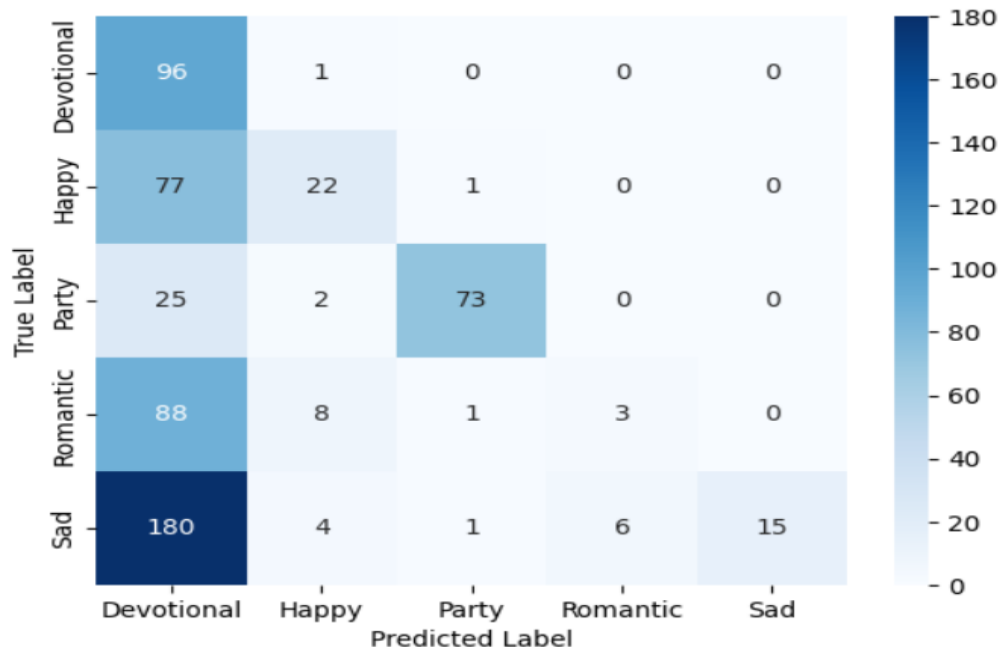


Figure 76: Confusion Matrix -MER500 Inception V3 (oversampling - Standard Spectrogram)

5.2.2 Results with Mel-Spectrogram

5.2.2.1 Unbalanced Data

A. ResNet50 Model:

ResNet50 - Mean Accuracy: 0.8175

ResNet50 - Test Accuracy: 0.7769

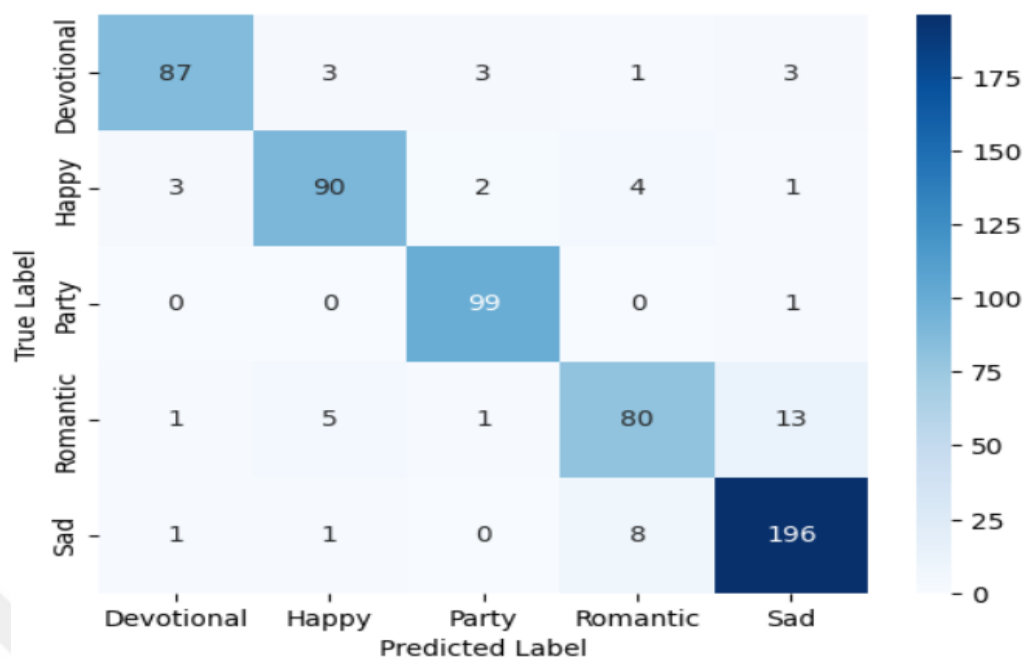


Figure 77: Confusion Matrix - MER500 ResNet50 (unbalance - Mel Spectrogram)

B. DenseNet121 Model:

DenseNet121 - Mean Accuracy: 0.7552

DenseNet121 - Test Accuracy: 0.6777

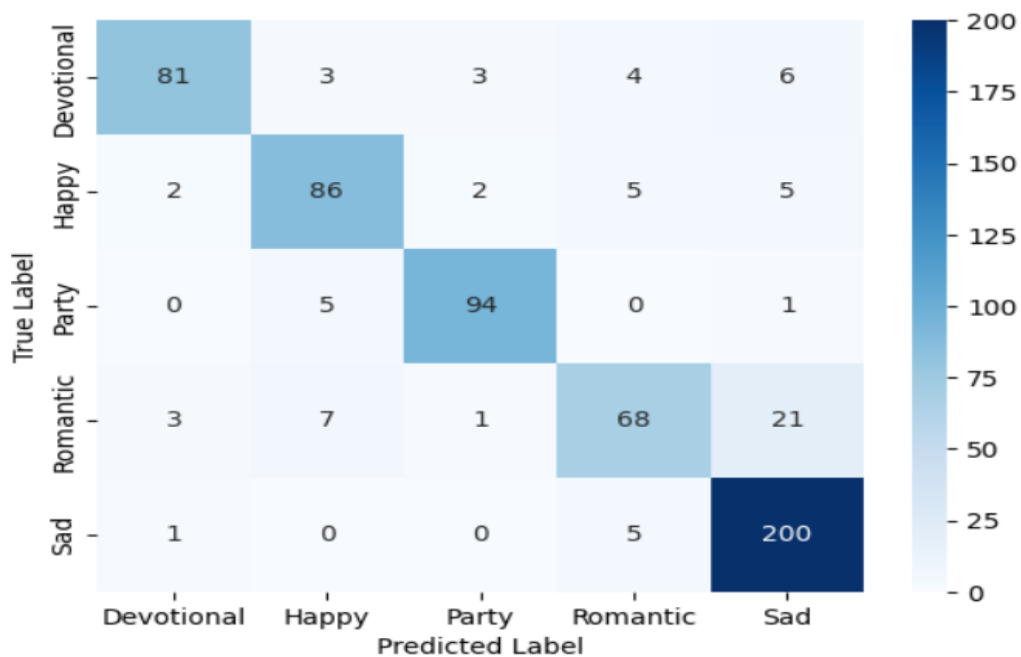
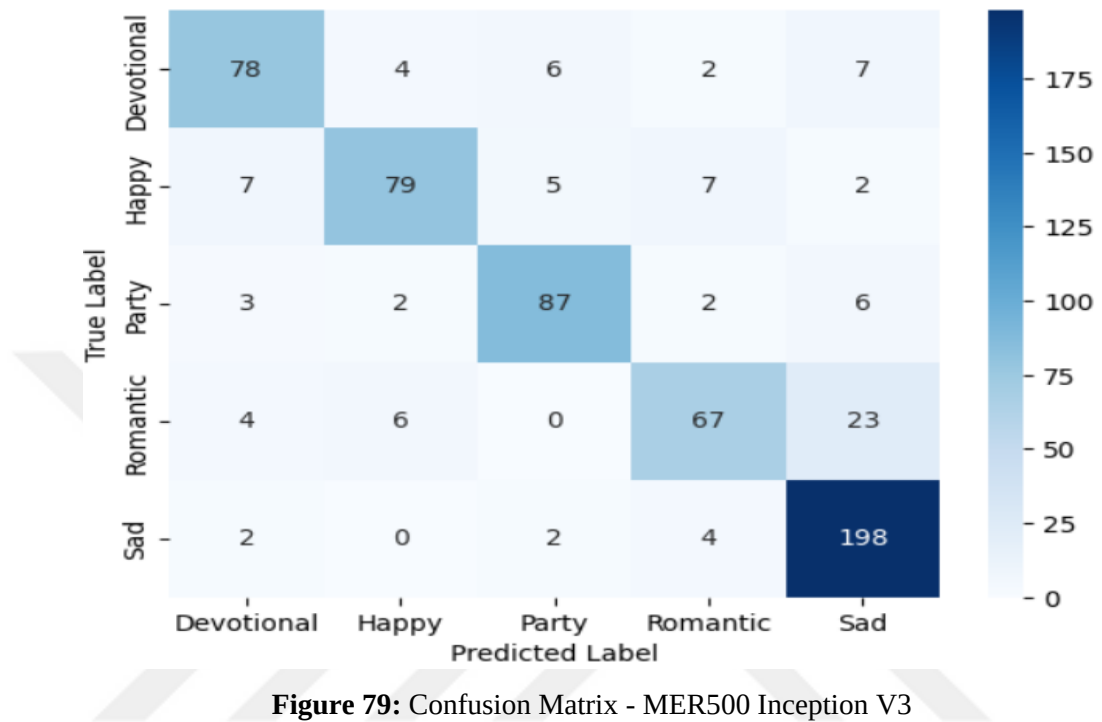


Figure 78: Confusion Matrix -MER500 DenseNet121 (unbalance - Mel Spectrogram)

C. Inception V3 Model:

Inception V3 - Mean Accuracy: 0.6661

Inception V3 - Test Accuracy: 0.5868



5.2.2.2 Under-sampled Data

a) ResNet50 Model:

ResNet50 - Mean Accuracy: 0.9357

ResNet50 - Test Accuracy: 0.7216

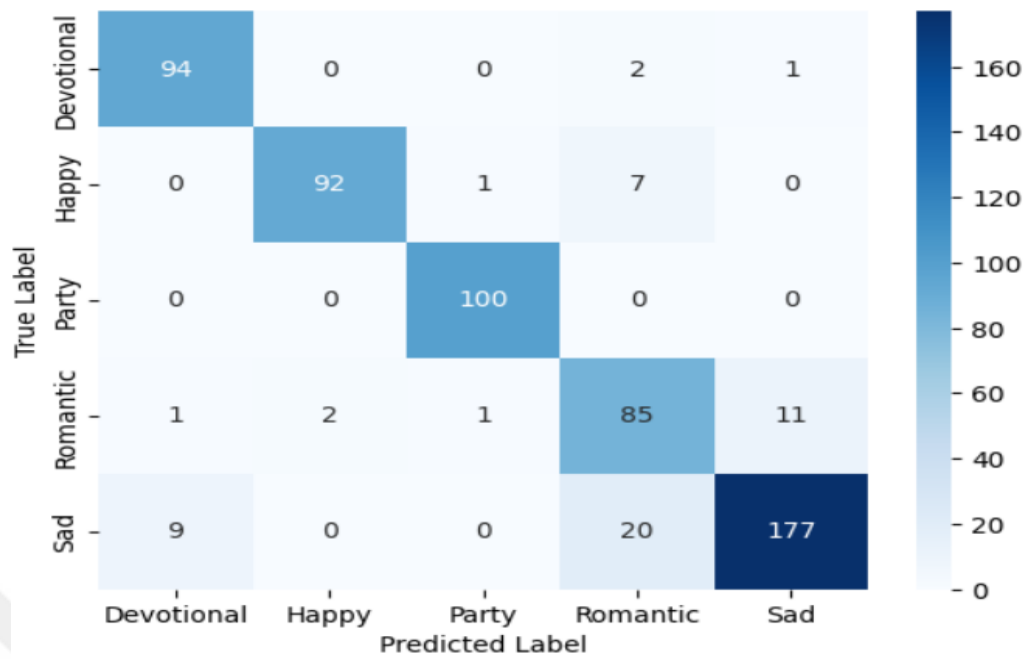


Figure 80: Confusion Matrix -MER500 ResNet50 (undersampling - Mel Spectrogram)

b) DenseNet121 Model:

DenseNet121 - Mean Accuracy: 0.8584

DenseNet121 - Test Accuracy: 0.7732

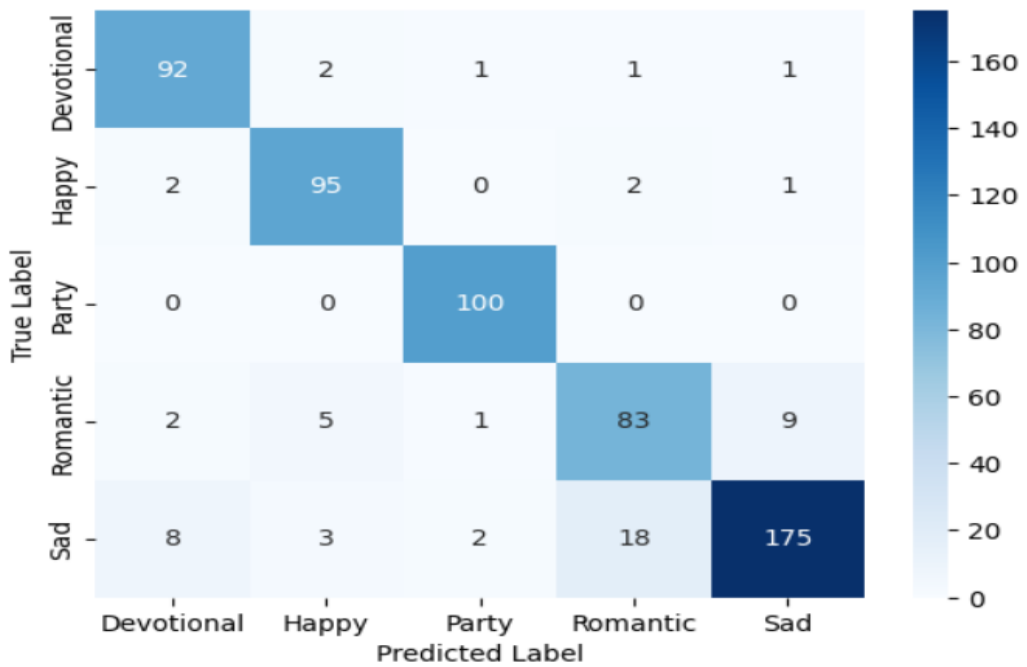


Figure 81: Confusion Matrix -MER500 DenseNet121 (undersampling - Mel Spectrogram)

c) Inception V3 Model:

Inception V3 - Mean Accuracy: 0.8762

Inception V3 - Test Accuracy: 0.6907

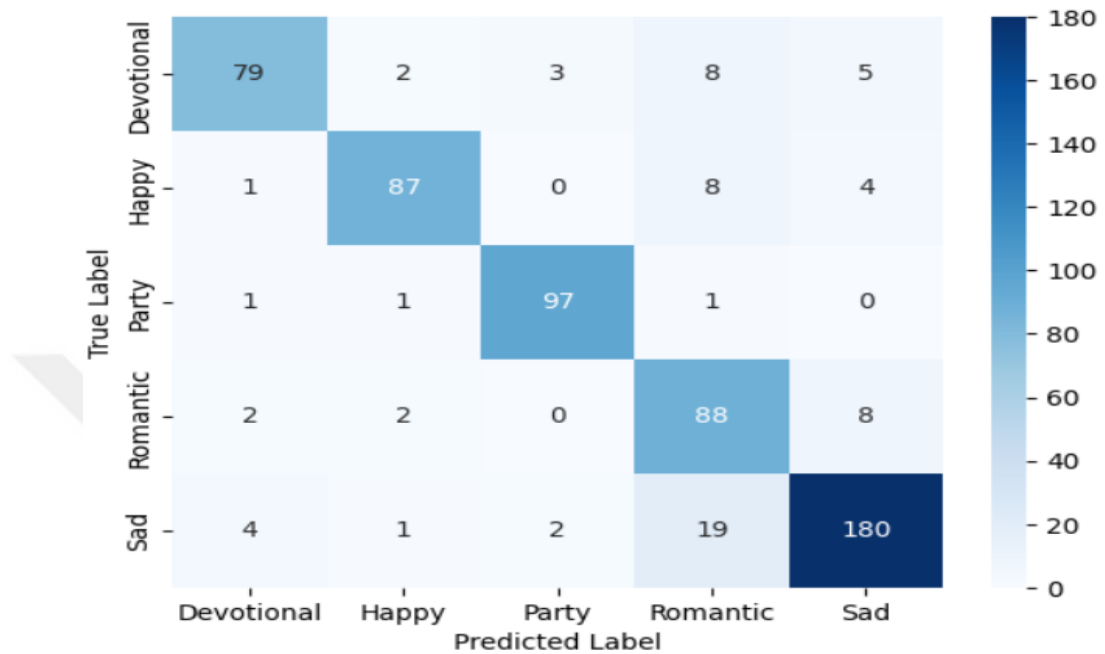


Figure 82: Confusion Matrix -MER500 Inception V3 (undersampling - Mel Spectrogram)

5.2.2.3 Oversampled Data

1. ResNet50 Model:

ResNet50 - Mean Accuracy: 0.9782

ResNet50 - Test Accuracy: 0.9223



Figure 83: Confusion Matrix -MER500 ResNet50 (oversampling - Mel Spectrogram)

2. DenseNet121 Model:

DenseNet121 - Mean Accuracy: 0.9733

DenseNet121 - Test Accuracy: 0.9563

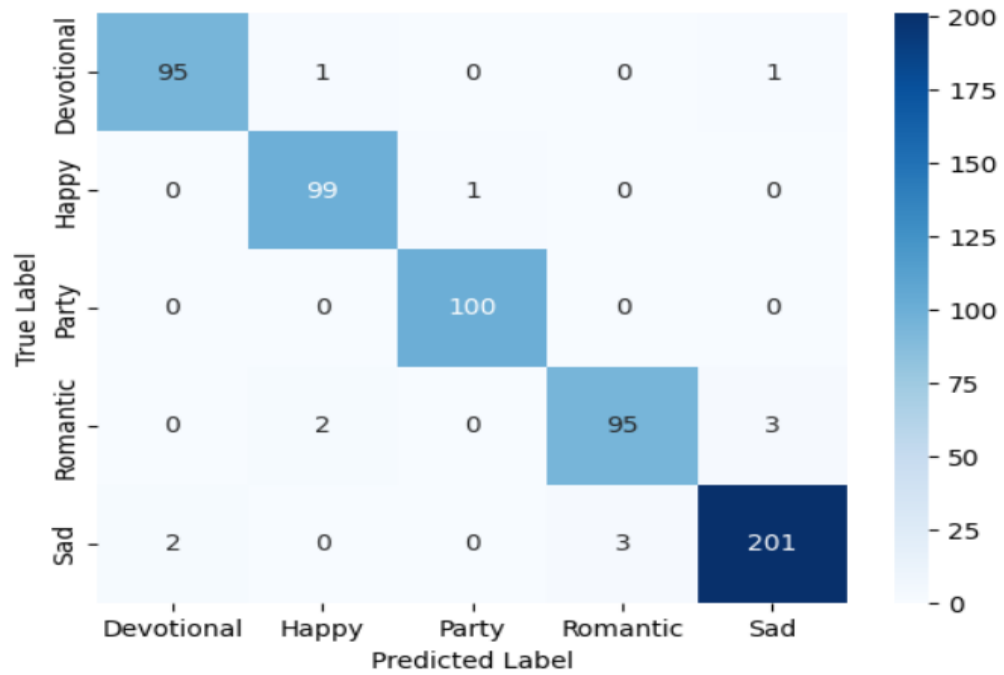


Figure 84: Confusion Matrix -MER500 DenseNet121 (oversampling - Mel Spectrogram)

3. Inception V3 Model:

Inception V3 - Mean Accuracy: 0.9539

Inception V3 - Test Accuracy: 0.9369



Figure 85: Confusion Matrix - MER500 Inception V3 (oversampling - Mel Spectrogram)

5.2.3 Comparison of Models' Performance

The best Model with standard spectrogram is DensNet121 with accuracy of 0.9709, (over-sample), The best Model with Mel spectrogram is DensNet121 as well with accuracy of 0.9563, (over-sample).

Table 11: The total accuracy (Mel- spectrogram and standard spectrogram)

	DATA	Resnet50	DenesNet121	Inception v3
mel_spectrogram	unbalance data	0.7769	0.6777	0.5868
	oversample	0.9223	0.9563	0.9369
	undersample	0.7216	0.7732	0.6907
spectrogram	unbalance data	0.7521	0.7521	0.7025
	oversample	0.9563	0.9709	0.9563
	undersample	0.8866	0.8969	0.8454

- The Model DenseNet121 details (oversampling - standard spectrogram)

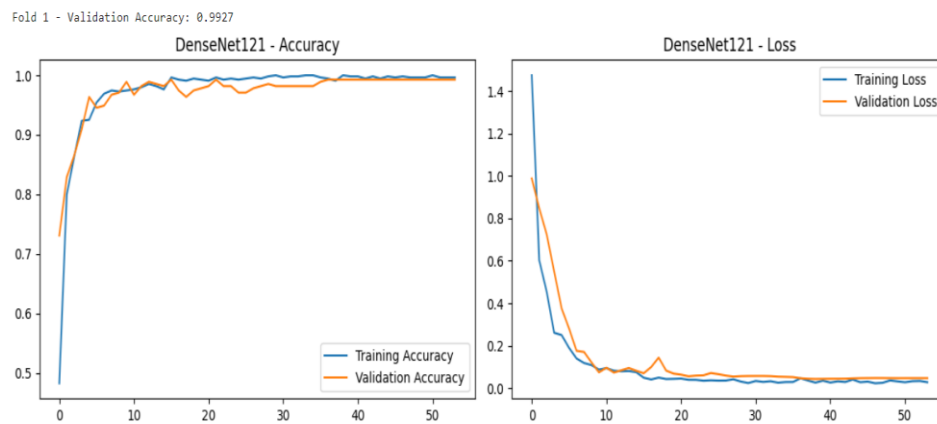


Figure 86: MER500 Best Model DenseNet121 (oversampling-standard spectrogram) fold-1.

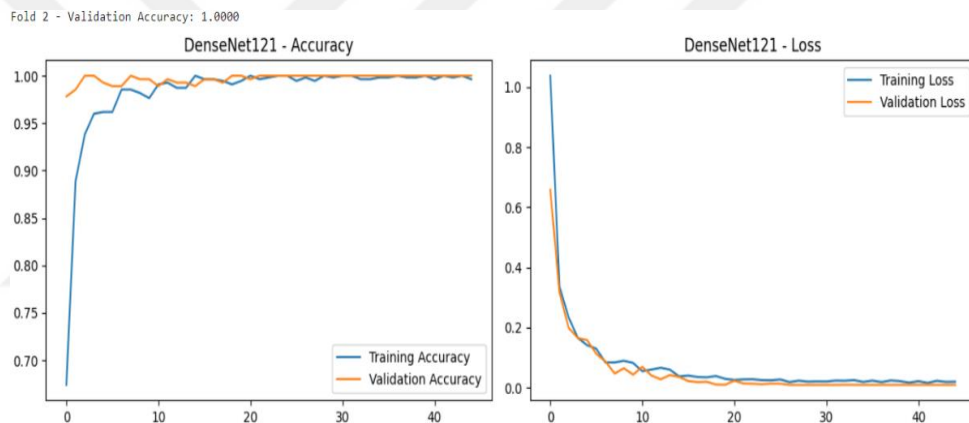


Figure 87: MER500 Best Model DenseNet121 (oversampling - standard spectrogram) fold-2.

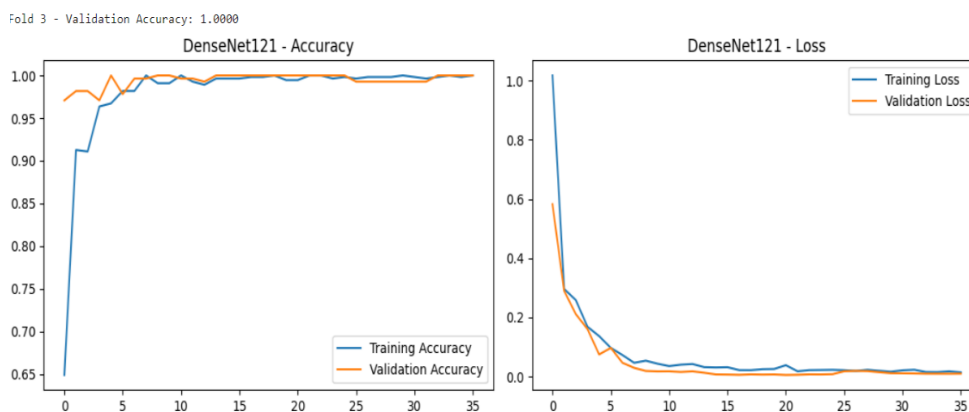


Figure 88: MER500 Best Model DenseNet121 (oversampling - standard spectrogram) fold-3.

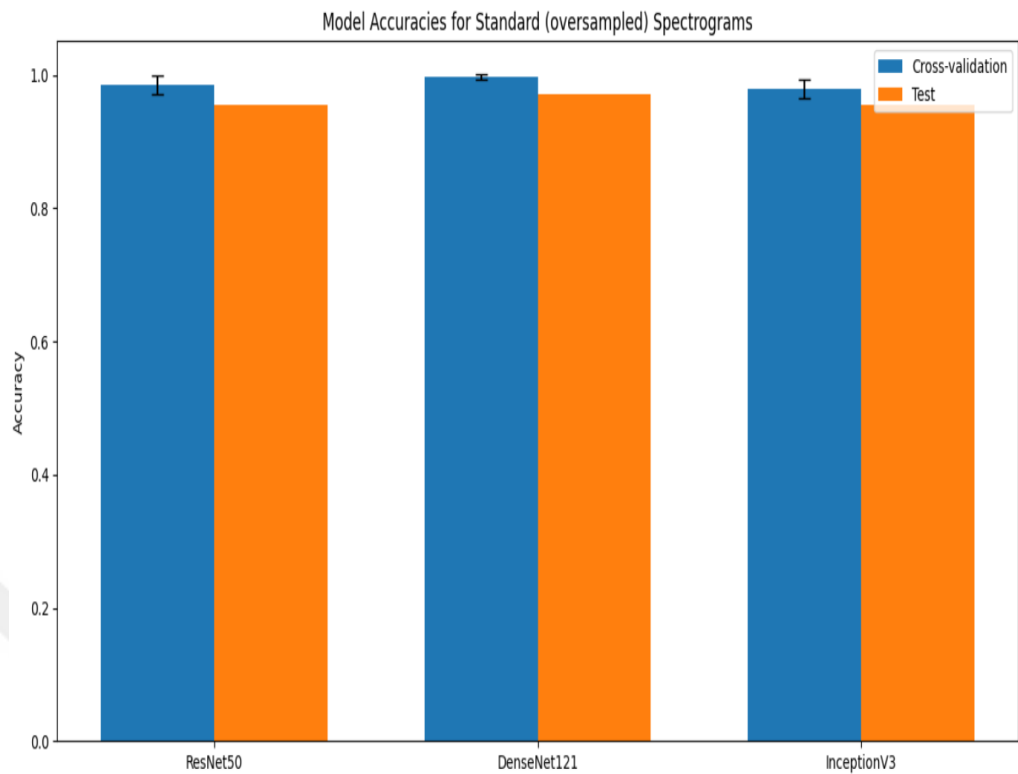


Figure 89: Pre-trained Models on MER500 (oversampling - standard spectrogram) bar plot.

	precision	recall	f1-score	support
Devotional	0.88	0.85	0.86	97
Happy	0.72	0.93	0.81	100
Party	0.97	0.98	0.98	100
Romantic	0.79	0.22	0.34	100
Sad	0.74	0.90	0.81	206
accuracy			0.80	603
macro avg	0.82	0.77	0.76	603
weighted avg	0.80	0.80	0.77	603

Figure 90: Classification report - MER500 DenseNet121 (oversampling - standard spectrogram)



Figure 91: Confusion Matrix - MER500 DenseNet121 cm (oversampling - standard spectrogram)

➤ The Model DenseNet121 details (oversampling - Mel spectrogram)

Fold 1 - Validation Accuracy: 0.9455

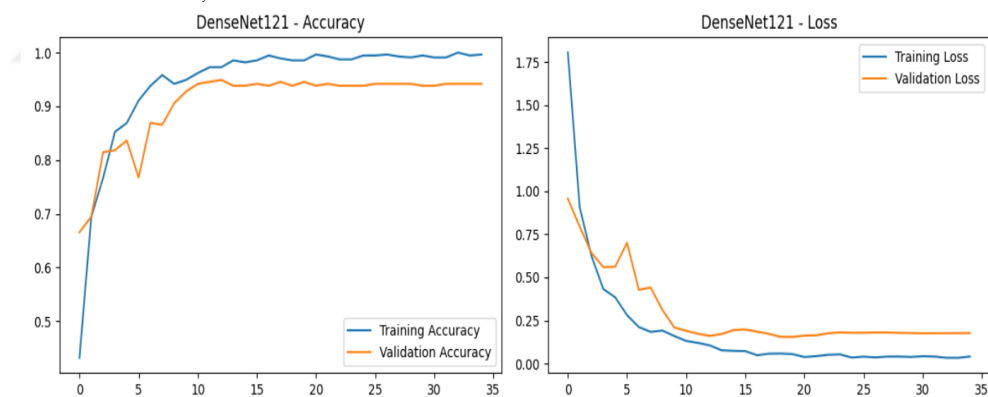


Figure 92: MER500 Best Model DenseNet121 (oversampling - Mel spectrogram) fold-1.

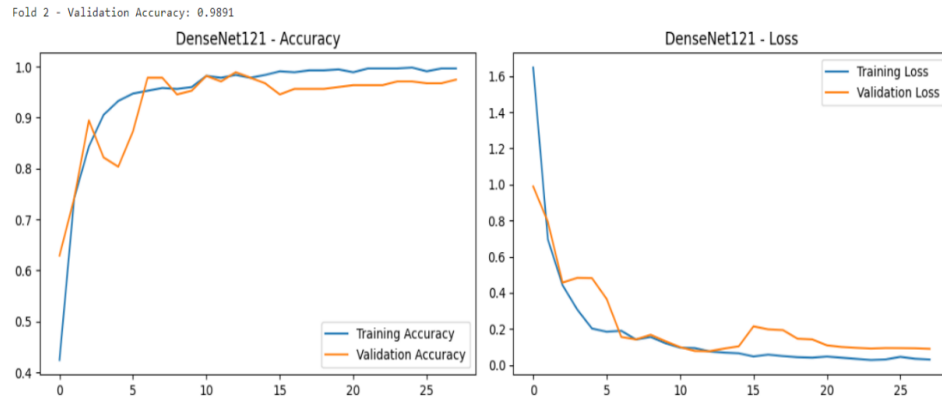


Figure 93: MER500 Best Model DenseNet121 (oversampling - Mel spectrogram) fold-2.

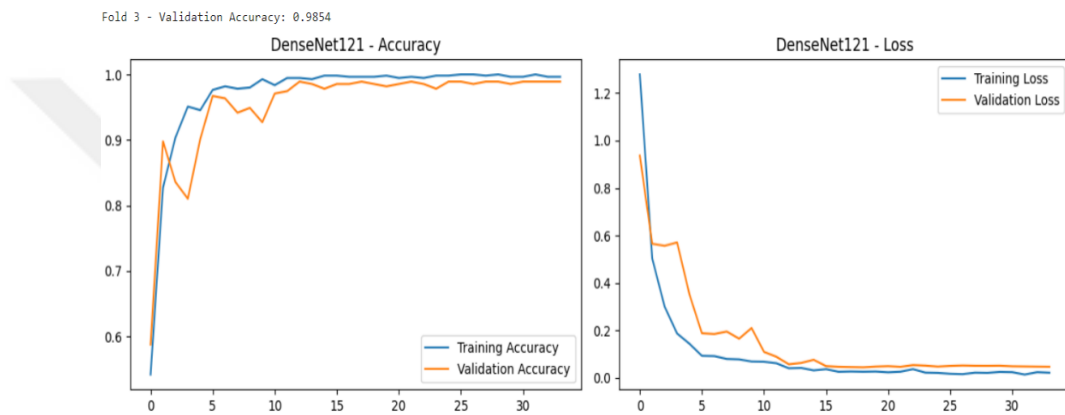


Figure 94: MER500 Best Model DenseNet121 (oversampling - Mel spectrogram) fold-3.

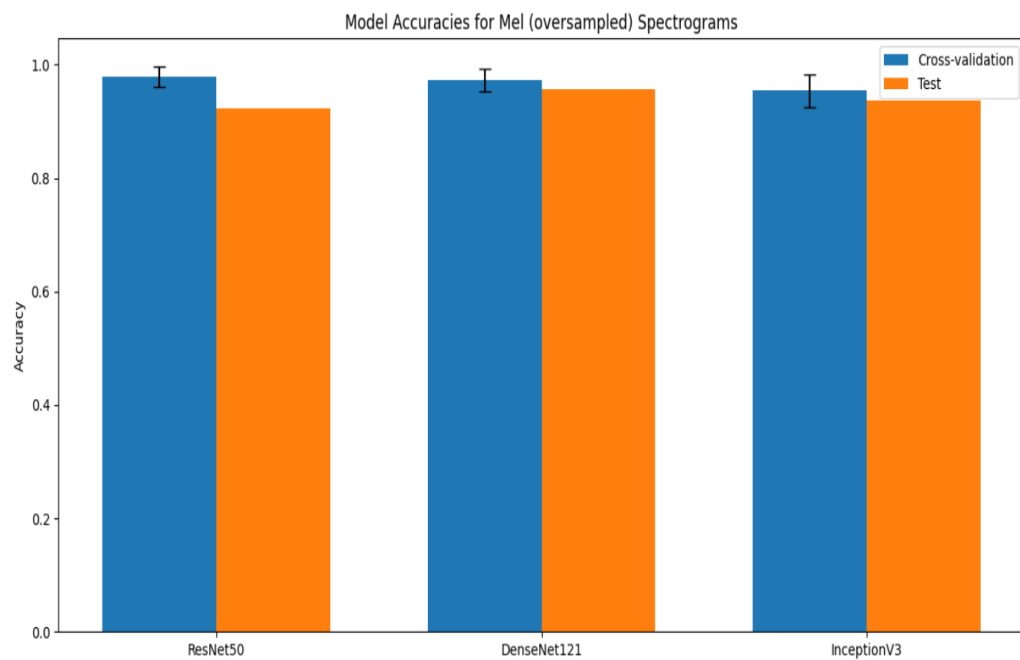


Figure 95: Pre-trained Models on MER500 (oversampling - Mel spectrogram) bar plot.

	precision	recall	f1-score	support
Devotional	0.98	0.98	0.98	97
Happy	0.97	0.99	0.98	100
Party	0.99	1.00	1.00	100
Romantic	0.97	0.95	0.96	100
Sad	0.98	0.98	0.98	206
accuracy			0.98	603
macro avg	0.98	0.98	0.98	603
weighted avg	0.98	0.98	0.98	603

Figure 96: Classification report - MER500 DenseNet121 (oversampling - Mel spectrogram)



Figure 97: Confusion Matrix - MER500 DenseNet121 cm (oversampling - Mel spectrogram)

5.3 ANALYSIS OF RESULTS

5.3.1 Comparison between Datasets in first stage and in the second one:

From the above results, we can find that the first model with standard spectrogram showed excellent results in terms of Turkish musical emotions where the ResNet50 model got the highest accuracy of 97%, the DenseNet121 model got an

accuracy of 96% and the Inception V3 model got the lowest accuracy of 93%. Regardless of the fact that all ratio of those models has a higher percentage in comparison to the MER500 dataset, where the ResNet50 model got 82% accuracy, the DensNet121 got 77% and the Inception V3 got a weak accuracy of 53%. as shown in the table below. While the second model with the Mel spectrogram and the standard spectrogram gave different results, after performing some advanced preprocessing on the dataset and distributing them evenly,

1- MER500 dataset with Mel spectrogram:

- DenseNet121 got the highest accuracy of 95.63% with over-sample techniques, whereas the Inception V3 model got an accuracy of 93.69% and the lowest accuracy was 92.23% for ResNet50.

2- MER500 dataset with standard spectrogram:

- DenseNet121 has the highest accuracy of 97.09% with over-sample techniques, while the ResNet50 and Inception V3 models got the same accuracy of 95.63% with over-sample as well.

3- Turkish music emotion with Mel spectrogram:

- ResNet50 got the highest accuracy of 91.94%, DenseNet121 got the accuracy of 91.67%, while the Inception V3 models got the lowest accuracy of 82.5%.

4- Turkish music emotion with standard spectrogram:

- The DenseNet121 model got the highest accuracy of 85.56%, while the ResNet50 model got an accuracy of 84.17%, and the lowest accuracy was 82.78% for the Inception V3 model.

Table 12: The highest accuracy of the first model

First Model					
Classic Spectrum					
Resnet50		DenseNet121		Inception V3	
MER500	TME	MER500	TME	MER500	TME
0.82	0.97	0.77	0.96	0.53	0.93

Table 13: The highest accuracy of the second model

Second Model											
Mel Spectrum						Classic Spectrum					
Resnet50		DenseNet121		Inception V3		Resnet50		DenseNet121		Inception V3	
MERS00	MET	MERS00	MET	MERS00	MET	MERS00	MET	MERS00	MET	MERS00	MET
0.9223	0.9194	0.9563	0.9167	0.9369	0.825	0.9563	0.8417	0.9709	0.8556	0.9563	0.8278
oversample		oversample		oversample		oversample		oversample		oversample	

5.3.2 Comparison with previous studies:

In a past study done by Mohamed Bilal Er* and Ibrahim Berkan Idelek, published in an article entitled " Music Emotion Recognition by Using Chroma Spectrogram and Deep Visual Features". International Journal of Computational Intelligence Systems", the article presented a method using pre-trained models with chroma spectrum classification of musical emotions for the Turkish musical emotion dataset, they were able to reach an accuracy rating of 89.2% which was a good ratio. Our current study raised this percentage of the previous study, where in the First model, accuracy increased by 7.8% using the ResNet50 model with standard spectrum. In the second model, accuracy also increased by 2.74%, another study by (Mehmet Bilal ER*, Emin Murat ESİN) published in an article entitled ("Music Emotion Recognition with Machine Learning Based on Audio Features"), [78], they used dataset (Turkish Music Emotion), the article presented a different approach to my study, they have used machine learning with audio features (energy, tempo, MFCC, attack time and chroma). Where SVM, K-NN and ANN classifiers have trained by using audio features, the best classification success has obtained from ANN as 79.30%, while our current study has got a higher accuracy, it increased by 12.64 % with the second model and 17.7% with the first model.

In another old study conducted by (Li, J., Han, L., Wang, Y., Yuan, B., Yuan, X., Yang, Y., & Yan, H.) and published in an article entitled ("Combined angular margin and cosine margin SoftMax loss for music classification based on spectrograms. Neural Computing and Applications"), They used MER500_2 with two

classes: happy and sad, with an accuracy of 83%, and also used MER500_400, which has four classes: devotional, happy, sad, and romance, with an accuracy of 73.98%. My second model increased the ratio classification to MER500 with 5 classes into 12.63 with the classical spectrum and 9.23 with the Mel spectrum, and another study by (Swati Goel) published in an article entitled (“Emotion Classification Using Nature Based Optimization with Transformers And Transfer Learning”), [79], he used the dataset MER500 with five classes (happy, sad, romantic, devotional, and party) and achieved an accuracy of 84% with lyrics, 81% with songs, and 82% with karaoke. My second stage model increased the ratio classification to MER500 with 5 classes into 13.09, 16.09, and 15.09 with the classical spectrum and with the Mel spectrum, 11.63, 14.63, and 13.63 respectively.

Table 14: The comparing with previous studies MER500 dataset.

Features & Models	Dataset	Methodology	Acc- uracy	Enhancement Ratio
ResNet50 with Spectro- gram	MER500-2: 2 classes (happy and sad) MER500-4: 4 classes (happy, sad, devotional, romantic)	Angular margin & cosine margin SoftMax loss using ResNet50 with spectrogram	0.83 0.74	Enhanced MER500-2 to 12.63 with classical spectrum 9.23 with Mel spectrum
BERT Model (lyrics) ResNet50 (songs)	MER500 with 5 classes	Lyrics Songs Karaoke	0.84 0.81 0.82	Enhanced songs to 16.09 Karaoke to 15.09 with Mel spectrogram

Table 15: The comparing with previous studies for MET dataset.

Chroma Spectrum with VGG16 and AlexNet	Turkish Music Emotion: 4 classes (happy, sad, relax, angry)	Chroma input used to train Support Vector Machine (SVM) and SoftMax classifiers	0.89	Enhanced Turkish Music Emotion dataset to 2.74 with Mel spectrogram
Machine Learning (SVM, ANN, K-NN)	Turkish Music Emotion: 4 classes (happy, sad, relax, angry)	Audio features (energy, tempo, MFCC, attack time, chroma). Best classification success achieved with ANN.	0.79	Enhanced Turkish Music Emotion dataset to 12.64

5.4 SUMMARY OF FINDINGS

Although the ResNet50 accuracy in the first model was higher compared to the second model, it got high accuracy with just Turkish music emotion; as for the MER500 dataset, the accuracy was not as well as in the attempt one. On the other hand, this second model would be more flexible and comprehensive, allowing for better generalization and with high classification accuracy for both, especially with the Mel spectrogram. It was able to handle two different types of datasets with respect to collection and storage, and in addition to the different levels of songs of the two different cultures; with all these challenges, the second model has been able to deal with it with complete flexibility and gave out percentages that were satisfactory for both datasets.

CHAPTER VI

CONCLUSION AND FUTURE WORK

6.1 CONCLUDING REMARKS

In this study, I analyzed the classification of musical emotions using deep neural network (DNN) models including ResNet50, DenseNet121, and Inception V3. The analysis was conducted on two datasets: the Turkish Music Emotion dataset, containing 400 audio files of duration (28-32 seconds) divided into four classes (sad, angry, relax, and happy), and the MER500 dataset, comprising 494 audio files of duration (8-59 seconds) with five classes (devotional, party, happy, sad, and romantic). From all previous work, I can conclude that when it comes to recognizing music emotions using pre-trained models, the ResNet50 model most often outperforms other models in terms of classification accuracy.

From the results of all attempts, the ResNet50 model demonstrated superior performance on both datasets. Specifically, for the Turkish Music Emotion dataset using standard spectrograms, ResNet50 achieved a test accuracy of 0.97, outperforming DenseNet121 and Inception V3 which achieved test accuracies of 0.961 and 0.93, respectively. When utilizing Mel-Spectrograms, ResNet50 again outperformed the other models with a mean accuracy of 0.9655 and a test accuracy of 0.9194.

For the MER500 dataset, the DenseNet121 model showed high performance during oversampling, it achieved an accuracy of 97% in the standard spectrograms. For the other models, ResNet50 and Inception V3, the accuracy of the test was the same at 95.63%. In the context of the inference, DenseNet121 performed well. It correctly predicted 6 out of 8 images in oversampled and undersampled conditions, with only one incorrect prediction in the no balance scenario using the standard spectrogram. With the Mel-Spectrogram, it predicted all 8 images correctly in undersampled and no balance situations, but made 2 errors in the oversample case.

The ResNet50 model also achieved a perfect 8 out of 8 prediction rate with the Mel-Spectrogram, it faced challenges with standard spectrogram predictions. The ResNet50 model made around 4 mistakes in undersampled and no balance scenarios and 6 errors in oversampling, indicating less accuracy compared to DenseNet121., as shown in the example below:

ResNet50 with unbalance Mel-spectrogram

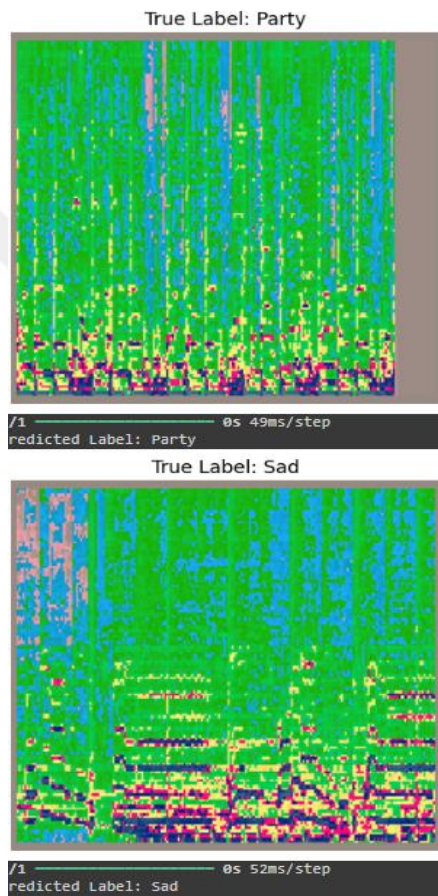


Figure 98: Inference Samples for ResNet50 with unbalance & Mel-spectrogram.

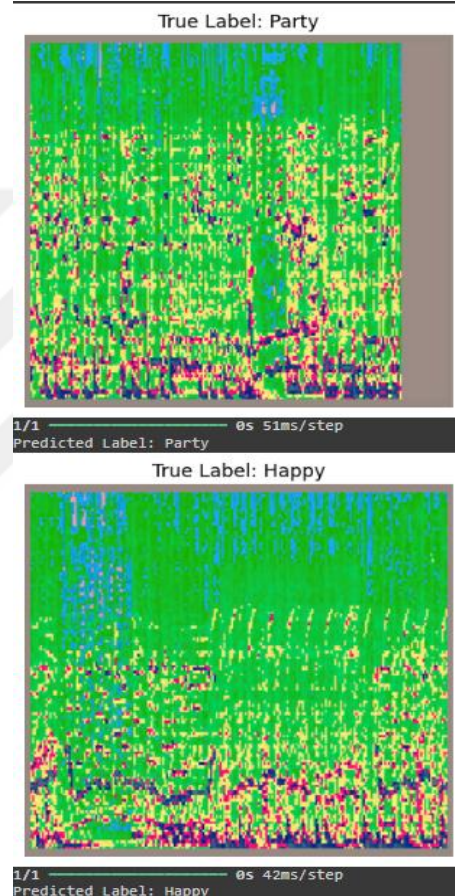


Figure 99: Inference Samples for ResNet50 with unbalance & Mel-spectrogram continued.

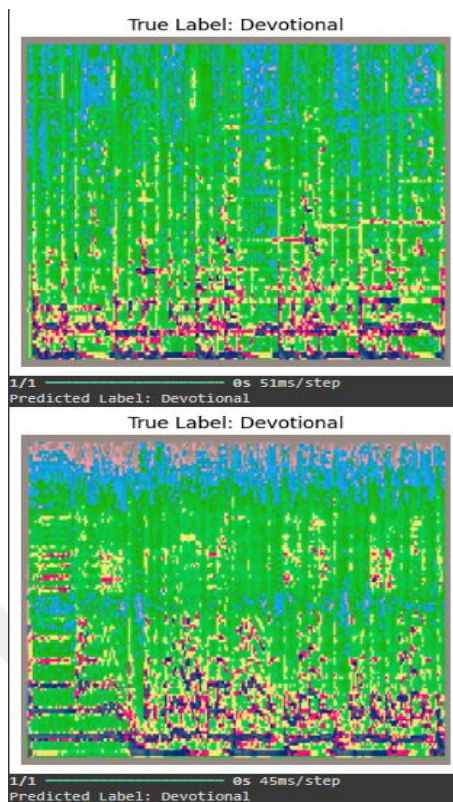


Figure 100: Inference Samples for ResNet50 with unbalance & Mel-spectrogram continued.

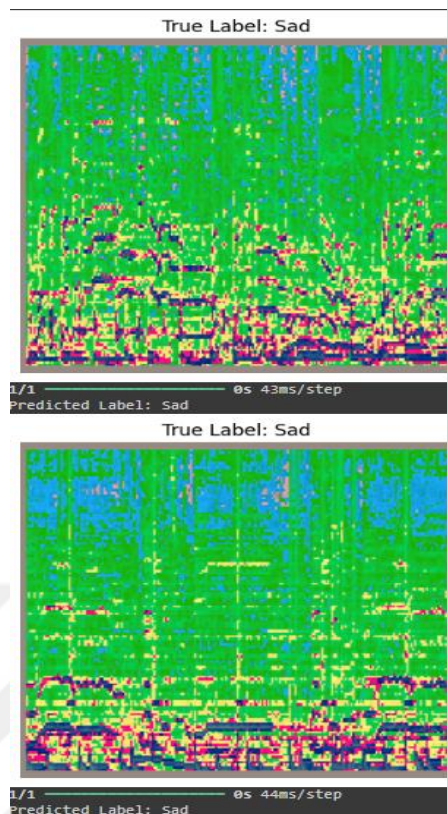


Figure 101: Inference Samples for ResNet50 with unbalance & Mel-spectrogram continued.

DenseNet121 with oversample standard-spectrogram

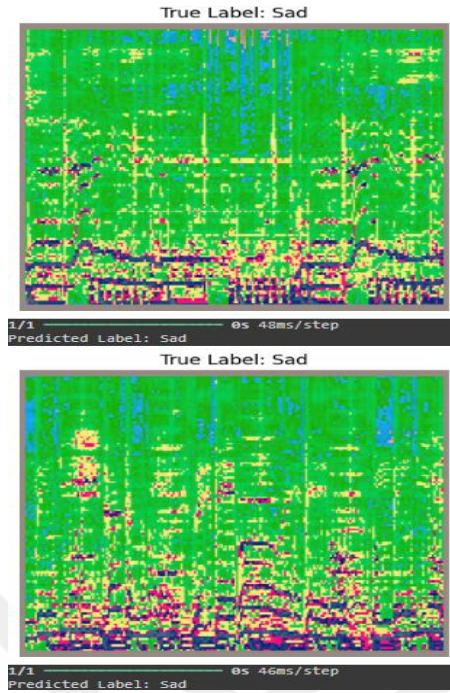


Figure 102: Inference Samples for DenseNet121 with oversample standard-spectrogram.

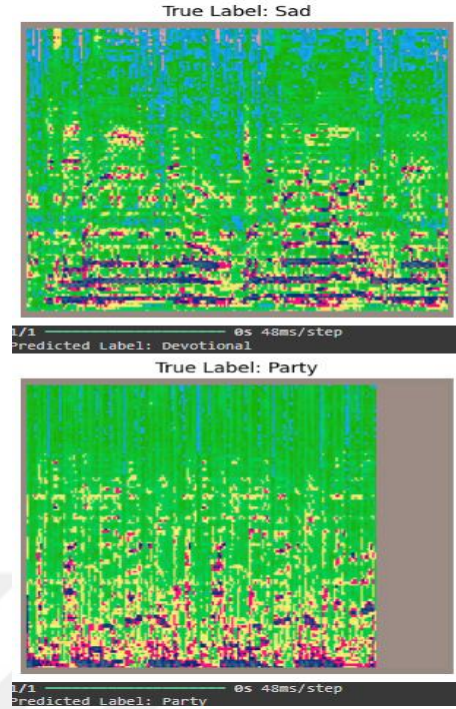


Figure 103: Inference Samples for DenseNet121 with oversample standard-spectrogram continued.

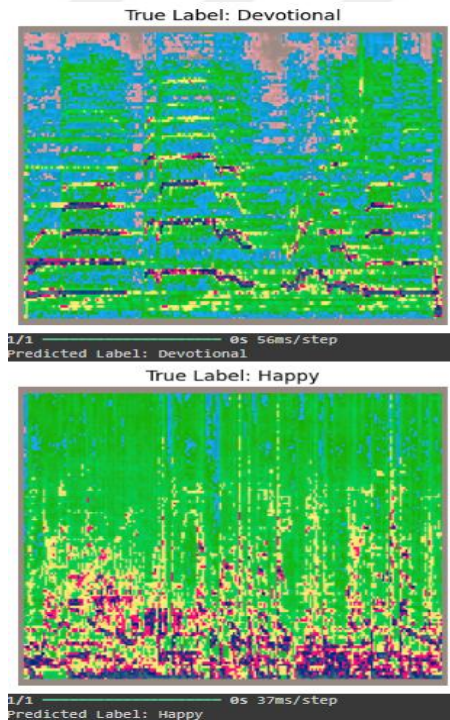


Figure 104: Inference Samples for DenseNet121 with oversample standard-spectrogram continued.

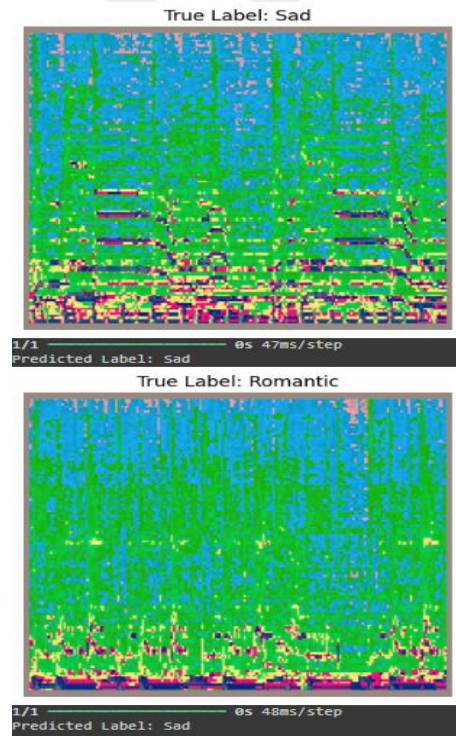


Figure 105: Inference Samples for DenseNet121 with oversample standard-spectrogram continued.

The study also adds to the existing literature by considerably enhancing the classification accuracies over the previous studies. For example, by using the Turkish Music Emotion dataset, the accuracy was raised by 7.8% with ResNet50 than the earlier research. Furthermore, it was found that for the MER500 dataset, the classification accuracy was higher by 12.63% and 9.23% for the standard and Mel spectrograms, respectively.

The study's key accomplishments lay in transfer learning, which worked very well, thereby giving huge gains in model performance without intense training from scratch, and the advanced spectrogram-based feature extraction methods that captured subtle characteristics of musical emotions.

Ultimately, it has been discovered that DenseNet121, and ResNet50 have shown to be the most efficient models for music emotion recognition analyzing the DNN models, considering their high accuracy and robustness under different conditions and with various spectrogram types.

6.2 LIMITATIONS OF THE STUDY

Despite the fact that our research excellently makes MER advance, still, the reality of its deficiencies cannot be neglected:

1. The limitation related to the size and diversity of the data sets is worth pointing out. Although efficient, the datasets involved in this research were of small size and limited diversity, a factor that could affect the generalization of the results to different genres and cultural contexts.
2. Deep Learning models used in this work are computationally intensive and consume large memory space. This in turn limits the real-world applicability of the proposed approach in scenarios with limited computational resources.
3. Investigating Different Audio Features.
4. Emotion classification in music itself is very subjective and may differ widely among listeners. Our trials have been generally binary like happy/sad, which, however, might not suffice in capturing all human responses to the music.

5. None of the models in this study are developed for real-time processing. In this respect, this constraint behaves as a direct hit on immediate applicability in life scenarios where humans require constant mood recognition.

6.3 FUTURE RESEARCH DIRECTIONS

Based on the findings and the limitations of this study, I would like to point out some future research directions as follows:

1. In my future study I should strive to collect and tailor bigger datasets that are more diversified and representative of a wide variety of musical genres, cultural influence, and emotional expressions so as to improve the generalizability of MER models.
2. I will build models for real-time MER, this research can be extended to be applied in live music recommendation systems, interactive music therapy, or real-time emotional feedback systems.
3. Further deepening the research and integration of more advanced DL architectures, such as transformer models (BERT) and attention mechanisms, should improve the accuracy and efficiency of MER systems.

REFERENCES

- [1] WORLU Chijioke (2017), "Predicting Listener's Mood Based on Music Genre: An Adapted Reproduced Model of Russell and Thayer", *Journal of Technology Management and Business*, Vol. 04, No. 01, pp. 1-15.
- [2] HIZLISOY Serhat, YILDIRIM Serdar, and TUFEKCI Zekeriya (2021), "Music Emotion Recognition Using Convolutional Long Short-Term Memory Deep Neural Networks", *Engineering Science and Technology, an International Journal*, Vol. 24, No. 3, pp. 760-767, DOI: 10.1016/j.jestch.2020.10.009.
- [3] YANG Yi-Hsuan, and CHEN Homer H. (2011), *Music Emotion Recognition*, CRC Press, Boca Raton, pp. 1-261, DOI: 10.1201/b10731.
- [4] WARRENBURG Lindsay Alison (2019), "Comparing Musical and Psychological Emotion Theories", *Psycho musicology: Music, Mind, and Brain (in press)*, Vol. 30, No. 1, pp. 1-19, DOI: 10.1037/pmu0000247.
- [5] XINYU Yang, YIZHUO Dong, and JUAN Li (2018), "Review of data features-based music emotion recognition methods", *Multimedia systems*, Vol. 24, pp. 365-389, DOI: 10.1007/s00530-017-0559-4.
- [6] GÓMEZ-CAÑÓN Juan Sebasti'an, CANO Estefan' ia, EEROLA Tuomas, HERRERA Perfecto, HU Xiao, YANG Yi-Hsuan, and GÓMEZ Emilia (2021), "Music Emotion Recognition Toward New, Robust Standards in Personalized and Context-Sensitive Applications", *IEEE Signal Processing Magazine*, Vol. 38, No. 06, pp. 106-114.
- [7] ZENTNER Marcel, GRANDJEAN Didier, and SCHERER Klaus R. (2008), "Emotions Evoked by the Sound of Music Characterization, Classification, and Measurement", *Emotion*, Vol. 8, No. 04, pp. 494-521, DOI: 10.1037/1528-3542.8.4.494.

- [8] TEIE David (2016), "A Comparative Analysis of the Universal Elements of Music and the Fetal Environment", *Frontiers in Psychology*, Vol. 7, p. 187715, DOI: 10.3389/fpsyg.2016.01158
- [9] ONGSULEE Pariwat (2017), "Artificial intelligence, machine learning and deep learning", In, *2017 15th International Conference on ICT and Knowledge Engineering (ICT&KE)*, pp. 1-6. IEEE.
- [10] SHARMA Neha, SHARMA Reecha, and JINDAL Neeru (2021), "Machine learning and deep learning applications-a vision", *Global Transitions Proceedings*, Vol. 02, No. 01, pp. 24-28.
- [11] FU Zhouyu, LU Guojun, TING Kai Ming, and ZHANG Dengsheng (2010), "A Survey of Audio-Based Music Classification and Annotation", *IEEE Transactions on Multimedia*, Vol. 13, No. 02, pp. 303-319.
- [12] PANDA Renato, MALHEIRO Ricardo, and PAIVA Rui Pedro (2020), "Audio Features for Music Emotion Recognition A Survey", *IEEE Transactions on Affective Computing*, Vol. 14, No. 01, pp. 68-88, DOI: 10.1109/TAFFC.2020.3032373.
- [13] Schlenker PHILIPPE (2017), "Outline of music semantics", *Music Perception: An Interdisciplinary Journal*, Vol. 35, No. 01, pp. 3-37, DOI: 10.1525/MP.2017.35.1.3.
- [14] FOLORUNSO Sakinat O., AFOLABI Sulaimon A., and OWODEYI Adeoye B. (2022), "Dissecting the Genre of Nigerian Music with Machine Learning Models", *Journal of King Saud University-Computer and Information Sciences*, Vol. 34, No. 08, pp. 6266-6279, DOI: 10.1016/j.jksuci.2021.07.009.
- [15] BARTHET Mathieu, FAZEKAS György, and SANDLER Mark (2013), "Music Emotion Recognition: From Content- to Context-Based Models", In, *From Sounds to Music and Emotions: 9th International Symposium, Revised Selected Papers*, Vol. 9, pp. 228-252, Springer, Berlin Heidelberg.
- [16] FELDMAN Lisa Feldman, and RUSSELL James A. (1998), "Independence and Bipolarity in the Structure of Current Affect", *Journal of Personality and Social Psychology*, Vol. 74, No. 04, pp. 967-984.

- [17] CHATURVEDI Vybhav, KAUR Arman Beer, VARSHNEY Vedansh, GARG Anupam, CHHABRA Gurpal Singh, and KUMAR Munish (2022), "Music Mood and Human Emotion Recognition Based on Physiological Signals a Systematic Review", *Multimedia Systems*, Vol. 28, No. 01, pp. 21-44, DOI: 10.1007/s00530-021-00786-6.
- [18] ER Mehmet Bilal, and AYDILEK Ibrahim Berkan (2019), "Music Emotion Recognition by Using Chroma Spectrogram and Deep Visual Features", *International Journal of Computational Intelligence Systems*, Vol. 12, No. 02, pp. 1622-1634, DOI: 10.2991/ijcis.d.191216.001.
- [19] CHEN Changfeng, and LI Qiang (2020), "A Multimodal Music Emotion Classification Method Based on Multi-Feature Combined Network Classifier", *Mathematical Problems in Engineering*, Vol. 2020, Article ID 4606027, pp. 1-11, DOI: 10.1155/2020/4606027.
- [20] SARKAR Rajib, CHOUDHURY Sombuddha, DUTTA Saikat, ROY Aneek, and SAHA Sanjoy Kumar (2020), "Recognition of Emotion in Music Based on Deep Convolutional Neural Network", *Multimedia Tools and Applications*, Vol. 79, No. 01, pp. 765-783, DOI: 10.1007/s11042-019-08192-x.
- [21] FRANZONI Valentina, BIONDI Giulio, and MILANI Alfredo (2020), "Emotional Sounds of Crowds Spectrogram-Based Analysis Using Deep Learning", *Multimedia Tools and Applications*, Vol. 79, No. 47, pp. 36063-36075, DOI: 10.1007/s11042-020-09428-x
- [22] WU Zijin (2022), "Research on Automatic Classification Method of Ethnic Music Emotion Based on Machine Learning", *Journal of Mathematics*, Vol. 2022, Article ID 7554404, pp. 1-11, DOI: 10.1155/2022/7554404.
- [23] NAG Sayan, BASU Medha, SANYAL Shankha, BANERJEE Archi, and GHOSH Dipak (2022), "On the Application of Deep Learning and Multifractal Techniques to Classify Emotions and Instruments Using Indian Classical Music", *Physica A: Statistical Mechanics and its Applications*, Vol. 597, p.127261, DOI: 10.1016/j.physa.2022.127261.

- [24] LI Jingxian, HAN Lixin, LI Xiaoshuang, ZHU Jun, YUAN Baohua, and GOU Zhinan (2022), "An Evaluation of Deep Neural Network Models for Music Classification Using Spectrograms", *Multimedia Tools and Applications*, Vol. 81, pp. 4621–4647, DOI: 10.1007/s11042-020-10465-9.
- [25] LI Jingxian, HAN Lixin, WANG Yang, YUAN Baohua, YUAN Xiaofeng, YANG Yi, and YAN Hong (2022), "Combined Angular Margin and Cosine Margin Softmax Loss for Music Classification Based on Spectrograms", *Neural Computing and Applications*, Vol. 34, No. 13, pp. 10337-10353, DOI: 10.1007/s00521-022-06896-0.
- [26] HAN Xiao, CHEN Fuyang, and BAN Junrong (2023), "Music Emotion Recognition Based on a Neural Network with an Inception-GRU Residual Structure", *Electronics*, Vol. 12, No. 04, p.978, DOI: 10.3390/electronics12040978.
- [27] MA Sihan, and ZHOU Ruohua (2024), "Violin Music Emotion Recognition with Fusion of CNN–BiGRU and Attention Mechanism", *Information*, Vol. 15, No. 04, p. 224, DOI:10.3390/info15040224.
- [28] LIU Xin, CHEN Qingcai, WU Xiangping, LIU Yan, and LIU Yang (2017), "CNN based music emotion classification", *arXiv preprint*, DOI: 10.48550/arXiv.1704.05665.
- [29] SEJDIC Ervin, DJUROVIC Igor, and STANKOVIC LJubiša (2008), "Quantitative Performance Analysis of Scalogram as Instantaneous Frequency Estimator", *IEEE Transactions on Signal Processing*, Vol. 56, No. 08, pp. 3837-3845.
- [30] MIRSAMADI Seyedmahdad, BARSOUM Emad, and ZHANG, Cha (2017), "Automatic speech emotion recognition using recurrent neural networks with local attention", In, *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2227-2231, IEEE, New Orleans, USA.
- [31] FULOP Sean A., and FITZ Kelly (2006), "A Spectrogram for the Twenty-First Century", *Acoustics Today*, Vol. 2, No. 03, pp. 26-33.

- [32] RUDD David Hason, HUO Huan, and XU Guandong (2022), "Leveraged mel spectrograms using harmonic and percussive components in speech emotion recognition", In, *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 392-404, Springer International Publishing , Cham.
- [33] ZENG Yuni, MAO Hua, PENG Dezhong, and YI Zhang (2019),"Spectrogram Based Multi-Task Audio Classification", *Multimedia Tools and Applications*, Vol. 78, pp. 3705-3722, DOI: 10.1007/s11042-017-5539-3.
- [34] ZUE Victor W., and COLE Ronald A. (1979), "Experiments on spectrogram reading", In, *ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing* Vol. 4, pp. 116-119, IEEE.
- [35] SATT Aharon, ROZENBERG Shai, and HOORY Ron (2017), "Efficient emotion recognition from speech using deep learning on spectrograms", *Interspeech*, pp. 1089-1093, DOI: 10.21437/Interspeech.2017-200.
- [36] KHURANA Aayush, MITTAL Sweta, KUMAR Deepika, GUPTA Sonali, and GUPTA Ayushi (2023), "Tri-Integrated Convolutional Neural Network for Audio Image Classification Using Mel-Frequency Spectrograms", *Multimedia Tools and Applications*, Vol. 82, No. 4, pp. 5521-5546, DOI: 10.1007/s11042-022-13358-1.
- [37] NIU Shuteng, LIU Yongxin, WANG Jian, and SONG Houbing (2020), "A Decade Survey of Transfer Learning (2010-2020)", *IEEE Transactions on Artificial Intelligence*, Vol. 1, No. 02, pp. 151-166.
- [38] RUSSAKOVSKY Olga, DENG Jia SU Hao, KRAUSE Jonathan, SATHEESH Sanjeev, MA Sean, HUANG Zhiheng, KARPATY Andrej, KHOSLA Aditya, BERNSTEIN Michael, BERG Alexander C., and FEI-FEI LI (2015), "ImageNet Large Scale Visual Recognition Challenge", *International Journal of Computer Vision*, Vol. 115, pp. 211-252, DOI: 10.1007/s11263-015-0816-y.
- [39] ALOM Md Zahangir, TAHA Tarek M., YAKOPCIC Christopher, WESTBERG Stefan, HASAN Mahmudul, ESEN Brian C Van, AWWAL Abdul A S., and ASARI Vijayan K. (2018), "The history began from AlexNet: A comprehensive survey on deep learning approaches", *arXiv preprint*, DOI: 10.48550/arXiv.1803.01164.

- [40] BHATT Dulari, PATEL Chirag, TALSANIA Hardik, PATEL Jigar, VAGHELA Rasmika, PANDYA Sharnil, MODI Kirit, and GHAYVAT dHemant (2021), "CNN Variants for Computer Vision History, Architecture, Application, Challenges and Future Scope", *Electronics*, Vol. 10, No. 20, p. 2470, DOI: 10.3390/electronics10202470.
- [41] SZEGEDY Christian, LIU Wei, JIA Yangqing, SERMANET Pierre, REED Scott, ANGUELOV Dragomir, ERHAN Dumitru, VANHOUCKE Vincent, and RABINOVICH Andrew (2015), "Going deeper with convolutions", *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1-9, Boston, USA, DOI: 10.48550/arXiv.1409.4842.
- [42] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, and SUN Jian (2016), "Deep residual learning for image recognition", *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778, Las Vegas, USA.
- [43] MILANI Pedro M. (2017), "The Power of Inception Tackling the Tiny ImageNet Challenge", *CS-231N Final Project Report*, Stanford University, California.
- [44] IMAN Mohammadreza, ARABNIA HamidReza, and RASHEED Khaled (2023), "A Review of Deep Transfer Learning and Recent Advancements", *Technologies*, Vol. 11, No. 2, p. 40, DOI: 10.3390/technologies11020040.
- [45] IKECHUKWU A. Victor, MURALI S., DEEPU R., and SHIVAMURTHY R. C. (2021), "ResNet-50 vs VGG-19 vs training from scratch A comparative analysis of the segmentation and classification of Pneumonia from chest X-ray images", *Global Transitions Proceedings*, Vol. 2, No. 02, pp. 375-381, DOI: 10.1016/j.gltp.2021.08.027.
- [46] GUOY anming, LIU Yu, OERLEMANS Ard, LAO Songyang, WU Song, and LEW Michael S. (2016), "Deep Learning for Visual Understanding A Review", *Neurocomputing*, Vol. 187, pp. 27-48, DOI: 10.1016/j.neucom.2015.09.116.

- [47] PAN Yuhang, LIU Junru, CAIL Yuting, YANG Xuemei, ZHANG Zhucheng, LONG Hong, ZHAO Ketong, YU Xia, ZENG Cui, DUAN Jueni, XIAO4 Ping, Jingbo LI, Feiyue CAI, YANG Xiaoyun, and TAN Zhen (2023), "Fundus Image Classification Using Inception V3 and ResNet-50 for the Early Diagnostics of Fundus Diseases", *Frontiers in Physiology*, Vol. 14, p. 1126780, DOI: 10.3389/fphys.2023.1126780.
- [48] PRAVEEN T.N.V.S., SIVATHMIKA Dasaradhi, JAHNAVI Goparaju, and BOLLEDU Jaswitha (2023), "An In-depth Exploration of ResNet-50 for Complex Emotion Recognition to Unraveling Emotional States", *In, 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT)*, pp. 1-5, IEEE, DOI: 10.1109/InCACCT57535.2023.10141774.
- [49] MANDAL Bishwas, OKEUKWU Adaeze, and THEIS Yihong (2021), "Masked face recognition using ResNet-50", *arXiv preprint*, DOI: 10.48550/arXiv.2104.08997.
- [50] SAKLI Nizar, GHABRI Haifa, SOUFIENE Ben Othman, ALMALKI Faris. A., SAKLI Hedi, ALI Obaid, and NAJJARI Mustapha (2022), "ResNet-50 for 12-Lead Electrocardiogram Automated Diagnosis", *Computational Intelligence and Neuroscience*, Vol. 2022, Article ID 7617551, 16 pages, DOI: 10.1155/2022/7617551.
- [51] SAHAAI Madona B., JOTHILAKSHMI G. R., RAVIKUMAR D., PRASATH Raghavendra, and SINGH Saurav (2022), "ResNet-50 based deep neural network using transfer learning for brain tumor classification", *AIP Conference Proceedings INTERNATIONAL CONFERENCE ON RECENT INNOVATIONS IN SCIENCE AND TECHNOLOGY (RIST 2021)*, Vol. 2463, Issue 1, pp. 020014-1 to 020014-8, AIP Publishing, Malappuram, India, DOI: 10.1063/5.0082328.
- [52] HUANG Gao, LIU Zhuang, VAN DER MAATEN Laurens, and WEINBERGER Kilian Q. (2017), "Densely connected convolutional networks", *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4700-4708, Honolulu, HI, USA, DOI: 10.1109/CVPR.2017.243.

- [53] WANG CHENG, CHEN DELEI, HAO LIN, LIU XUEBO, ZENG YU, CHEN JIANWEI, and ZHANG GUOKAI (2019), "Pulmonary Image Classification Based on Inception-V3 Transfer Learning Model", *IEEE Access*, Vol. 7, pp. 146533-146541, DOI: 10.1109/ACCESS.2019.2946000.
- [54] LIN Chunmian, LI Lin, LUO Wenting, WANG Kelvin C. P., and GUO Jiangang (2019), "Transfer Learning Based Traffic Sign Recognition Using Inception-V3 Model", *Periodica Polytechnica Transportation Engineering*, Vol. 47, No. 03, pp. 242-250, DOI: 10.3311/PPtr.11480.
- [55] SZEGEDY Christian, VANHOUCKE Vincent, IOFFE Sergey, SHLENS Jon, and WOJNA Zbigniew (2016), "Rethinking the inception architecture for computer vision", *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818-2826, Las Vegas, NV, USA, DOI: 10.1109/CVPR.2016.308.
- [56] WINER Ethan (2017), *The Audio Expert: Everything You Need to Know About Audio*, Second Edition, Routledge, New York, DOI: 10.4324/9781315223162.
- [57] GUNAWAN Teddy Surya, and KARTIWI Mira (2018), "Performance Evaluation of Multichannel Audio Compression", *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, Vol. 10, No. 1, pp. 146-153, pp146-153, DOI: 10.11591/ijeecs.v10.i1.
- [58] DOBSON Richard W. (2000), "Developments in audio file formats", *Proceedings of the International Computer Music Conference (ICMC)*, pp. 1-4, Berlin, Germany.
- [59] KUMAR Pramod, and GOYAL Dinesh (2014), "Performance Analysis for Audio Streaming in Cloud", *IOSR-JCE*, Vol. 16, pp. 98-104.
- [60] HUANG H., SHU H., and YU, R. (2014), "Lossless audio compression in the new IEEE standard for advanced audio coding", *In, 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6934-6938, IEEE.
- [61] RIVERO Cristobal, and MISHRA Prabhat (2008) "Lossless Audio Compression: A Case Study -Technical Report 08-415-" ,pp. 1-12, University of Florida, Department of Computer and Information Science and Engineering, Gainesville.

- [62] AHMED Rafid, ISLAM Md. Sazzadul, and UDDIN Jia (2018), "Optimizing Apple Lossless Audio Codec Algorithm Using NVIDIA CUDA Architecture", *International Journal of Electrical and Computer Engineering (IJECE)*, Vol. 8, No. 01, pp. 70-75, DOI: 10.11591/ijece.v8i1.pp70-75.
- [63] SMIRNOVA Tetiana (2020), "Comparative Analysis of Modern Formats of Lossy Audio Compression", *The proceedings of the CEUR Workshop*, Vol. 2654, pp. 233-240, DOI: 10.1007/978-3-030-50732-9_252.
- [64] BRANDENBURG Karlheinz (1999), "MP3 and AAC explained", *In, Audio Engineering Society Conference 17th International Conference High-Quality Audio Coding*, Florence, Italy.
- [65] CARLACCI Ayman Ammoura Franco (2002), *Ogg Vorbis and MP3 Audio Stream Characterization* (Unpublished Project Report), University of Alberta, Edmonton.
- [66] MOURJOPOULOS John N. (2005), *The Evolution of Digital Audio Technology*, *In, Communication Acoustics*, pp. 299-319, Springer Berlin Heidelberg.
- [67] SMITH Julius O. (2008), *Mathematics of the Discrete Fourier Transform (DFT): With Audio Applications*, Julius Smith, Second edition, BookSurge Publishing, Seattle, WA, USA.
- [68] GEMELLI Alessandro, TAMBUSI Marco, FUSSETTO Samuele, APRILE Antonio, MOISELLO Elisabetta, BONIZZONI Edoardo, and MALCOVATI Piero (2023), "Recent Trends in Structures and Interfaces of MEMS Transducers for Audio Applications A Review", *Micromachines*, Vol. 14, No. 04, p. 847, DOI: 10.3390/mi14040847.
- [69] JUSLIN Patrik N., and LAUKKA Petri (2003), "Communication of Emotions in Vocal Expression and Music Performance Different Channels, Same Code?", *Psychological Bulletin*, Vol. 129, No. 05, pp. 770-814, DOI: 10.1037/0033-2909.129.5.770.
- [70] BERNARDINI Nicola, and Giovanni DE POLI (2007), "The Sound and Music Computing Field Present and Future", *Journal of New Music Research*, Vol. 36, No. 03, pp. 143-148.

- [71] JUSLIN Patrik N., and LAUKKA Petri (2004), "Expression, Perception, and Induction of Musical Emotions A Review and a Questionnaire Study of Everyday Listening", *Journal of New Music Research*, Vol. 33, No. 03, pp. 217-238, DOI: 10.1080/0929821042000317813.
- [72] AL-SHOSHAN Abdullah I. (2006), "Speech and Music Classification and Separation A Review", *Journal of King Saud University-Engineering Sciences*, Vol. 19, No. 01, pp. 95-132.
- [73] STEIN Nancy L., and OATLEY Keith (1992), "Basic Emotions Theory and Measurement", *Cognition & Emotion*, Vol. 6, No. 3-4, pp. 161-168.
- [74] AN Sieun JI Li-Jun, MARKS Michael, and ZHANG Zhiyong (2017), "Two Sides of Emotion Exploring Positivity and Negativity in Six Basic Emotions Across Cultures", *Frontiers in Psychology*, Vol. 8, p. 610, DOI: 10.3389/fpsyg.2017.00610.
- [75] TEIMOORINIA Hossen, TOYONAGA Robert D., FABBRO Sebastien, and BOTTRELL Connor (2020), "Comparison of Multi-Class and Binary Classification Machine Learning Models in Identifying Strong Gravitational Lenses", *Publications of the Astronomical Society of the Pacific*, Vol. 132, No. 1010, p. 044501.
- [76] GRANDINI Margherita, BAGLI Enrico, and VISANI Giorgio (2020), "Metrics for Multi-Class Classification: An Overview", *arXiv preprint*, DOI: 10.48550/arXiv:2008.05756.
- [77] CALVO Rafael, and MAC KIM Sunghwan (2013), "Emotions in text: dimensional and categorical models", *Computational Intelligence*, Vol. 29, No. 03, pp. 527-543.
- [78] ER Mehmet Bilal, and ESIN Emin Murat (2021), "Music Emotion Recognition with Machine Learning Based on Audio Features", *Computer Science*, Vol. 06, No. 03, pp. 133-144, DOI: 10.53070/bbd.945894.
- [79] GOEL Swati (2022), "Emotion Classification Using Nature Based Optimization with Transformers and Transfer Learning", *Journal of Pharmaceutical Negative Results*, Vol. 13, No. 10, pp. 3052-3071, DOI: 10.47750/pnr.2022.13.S10.369.

APPENDICES

APPENDIX 1: TO MITIGATE OVERFITTING IN MY MODELS, I HAVE IMPLEMENTED SEVERAL STRATEGIES

- 1- L2 regularize and dropout technique have been used:
 - Applied L2 regularization to the dense layers to penalize large weights and prevent the model from becoming too complex.
 - Used dropout techniques to randomly drop units during training, which helps prevent the network from relying too heavily on specific nodes and promotes better generalization.

▼ Create the model architecture

```
def build_model(base_model, num_classes):  
  
    # Make the last few layers of the base model trainable  
    base_model.trainable = True  
  
    x = layers.GlobalAveragePooling2D()(base_model.output)  
    x = layers.Dense(256, activation='relu', kernel_regularizer=regularizers.l2(0.00001))(x)  
    x = layers.BatchNormalization(trainable=base_model.trainable)(x)  
    x = layers.Dropout(0.6)(x)  
  
    x = layers.Dense(64, activation='relu', kernel_regularizer=regularizers.l2(0.00001))(x)  
    x = layers.Dropout(0.4)(x)  
  
    outputs = Dense(num_classes, activation='softmax')(x)  
  
    model = Model(inputs=base_model.input, outputs=outputs)  
    return model
```

Figure 106: Model architecture

- 2- Cross- Validation techniques:
Employed cross-validation techniques, using five folds for the Turkish music emotion dataset and a different number of folds for the MER500 datasets, five folds in the first stage and three folds in the second stage. This helps ensure that the model generalizes well to unseen data.

```
def split_data(images, labels, n_splits=5, random_state=42):
    skf = StratifiedKFold(n_splits=n_splits, shuffle=True,
                          random_state=random_state)
    for train_index, val_index in skf.split(images, labels):
        train_images = [images[i] for i in train_index]
        train_labels = [labels[i] for i in train_index]
        val_images = [images[i] for i in val_index]
        val_labels = [labels[i] for i in val_index]
        yield (train_images, train_labels), (val_images, val_labels)
```

Figure 107: five folds for the Turkish music emotion dataset (First Stage)

```
def train_models_on_spectrograms(use_mel=True, save_preprocessed=True, n_splits=5, test_size=0.3):
    spectrogram_directory = mel_spectrogram_dir if use_mel else spectrogram_dir
    spectrogram_type = 'mel' if use_mel else 'spectrogram'

    # Prepare image data
    image_data, labels_encoded, label_encoder, file_paths = prepare_image_data(spectrogram_directory, save_preprocessed, use_mel)

    # Split data into train+val and test sets
    X_train_val, X_test, y_train_val, y_test = train_test_split(image_data, labels_encoded, test_size=test_size, stratify=labels_encoded, random_state=42)
```

Figure 108: five folds for the Turkish music emotion dataset (Second Stage)

```
def split_data(images, labels, n_splits=5, random_state=42):
    skf = StratifiedKFold(n_splits=n_splits, shuffle=True,
                          random_state=random_state)
    for train_index, val_index in skf.split(images, labels):
        train_images = [images[i] for i in train_index]
        train_labels = [labels[i] for i in train_index]
        val_images = [images[i] for i in val_index]
        val_labels = [labels[i] for i in val_index]
        yield (train_images, train_labels), (val_images, val_labels)
```

Figure 109: Five folds for the MER500 dataset (First Stage)

```
def train_and_evaluate_model(X, y, base_model, num_classes, model_name, n_splits=3):
    skf = StratifiedKFold(n_splits=n_splits, shuffle=True, random_state=42)
    fold accuracies = []
    best_model = None
    best_accuracy = 0
```

Figure 110: Three folds for the MER500 dataset (Second Stage)

- 3- Early stopping has been added to each epoch's circle.
Incorporated early stopping during training to halt the process once the model's performance on the validation set stops improving. This prevents the model from continuing to learn patterns specific to the training data.

```
callbacks = [
    EarlyStopping(monitor='val_loss', patience=15, restore_best_weights=True),
    ReduceLROnPlateau(monitor='val_loss', factor=0.5, patience=3, min_lr=1e-7)
]
```

Figure 111: Early Stopping

4- Performance Monitoring:

Observed the training vs validation loss curves in terms of insuring they track closely together, monitoring these curves helps in making decisions about when to stop training or adjust model parameters

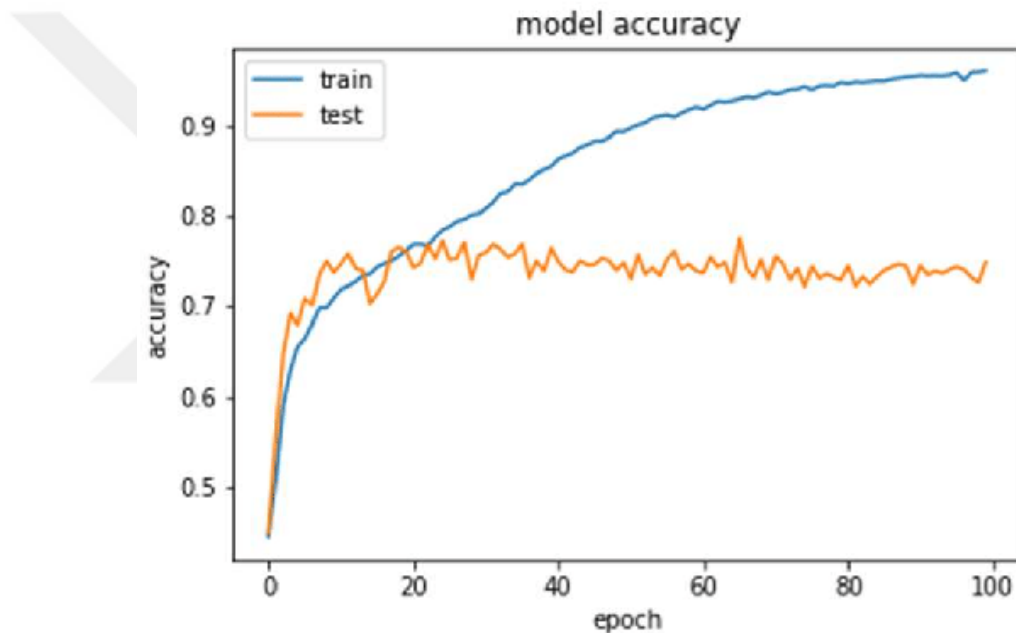


Figure 112: A sample of overfitting.

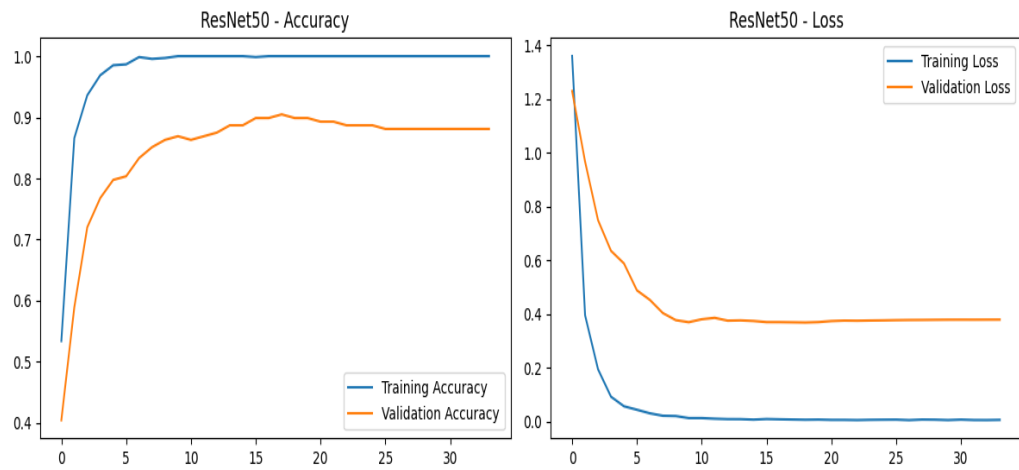


Figure 113: ResNet50+Mel-Spectrogram on MET fold-1

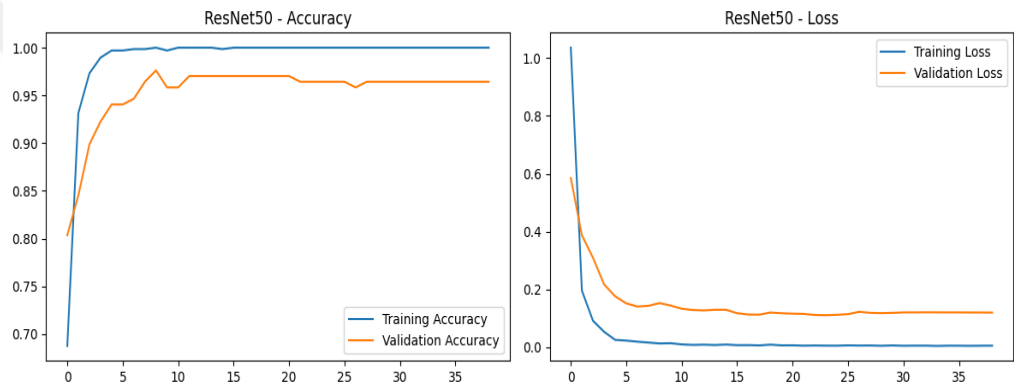


Figure 114: ResNet50+Mel-Spectrogram on MET fold-2

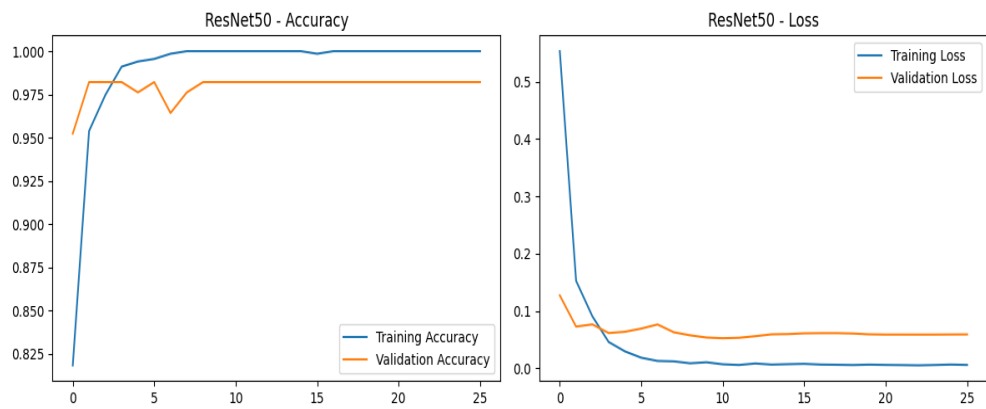


Figure 115: ResNet50+Mel-Spectrogram on MET fold-3

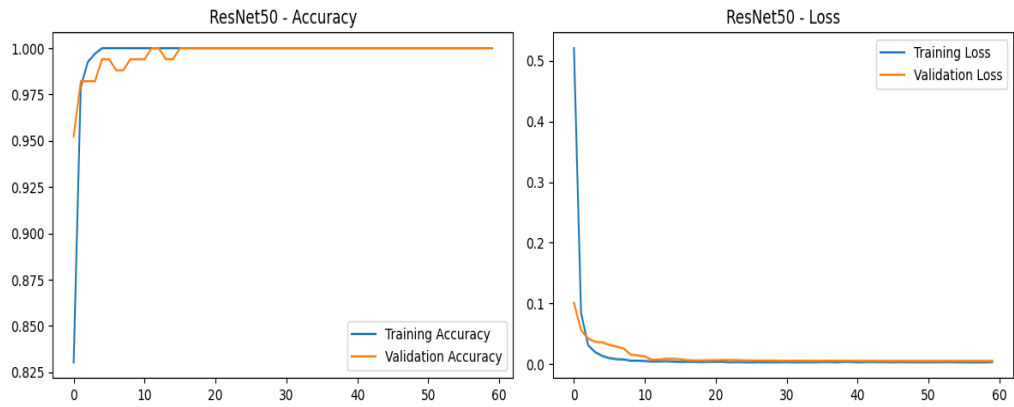


Figure 116: ResNet50+Mel-Spectrogram on MET fold-4

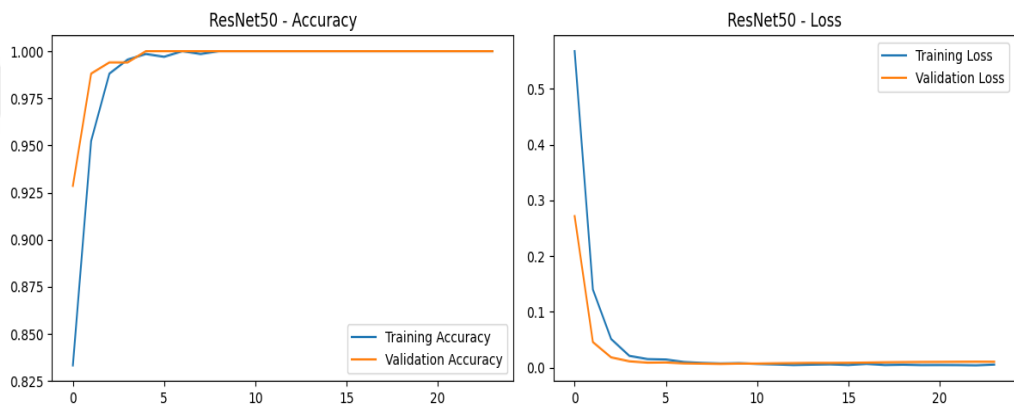


Figure 117: ResNet50+Mel-Spectrogram on MET fold-5

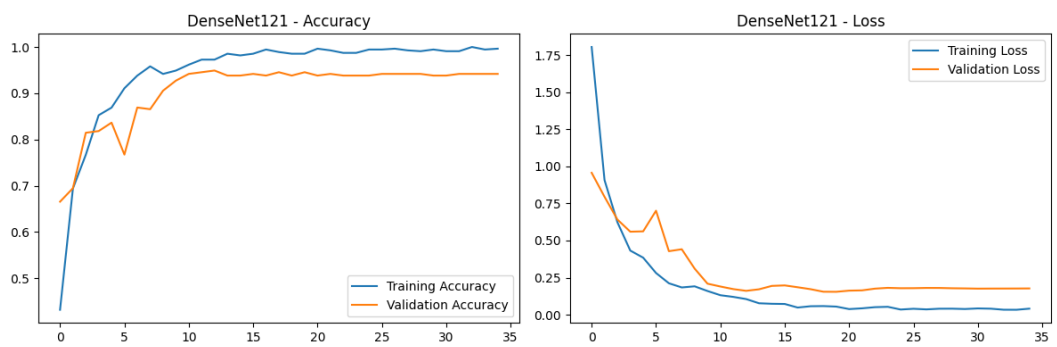


Figure 118: MER500 Best Model DenseNet121 (Oversampling - Mel spectrogram) fold-1

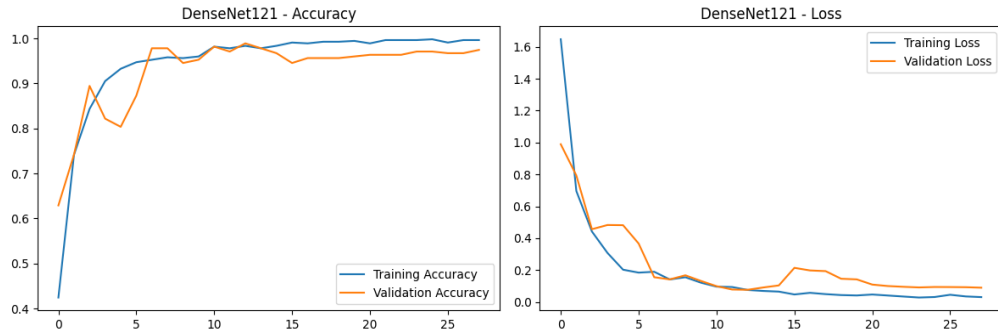


Figure 119: MER500 Best Model DenseNet121 (Oversampling - Mel spectrogram) fold-2

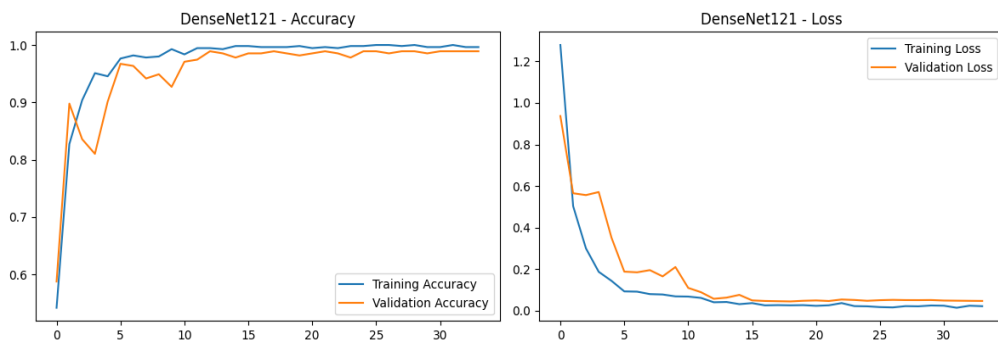


Figure 120: MER500 Best Model DenseNet121 (Oversampling - Mel spectrogram) fold-3

5- Inference Testing:

I have Developed a function to infer the spectrum images (both standard and Mel-spectrograms). By evaluating the model's performance on these images, I can determine if overfitting is occurring. Consistent and accurate predictions on the spectrum images indicate a well-generalized model.

▼ Inference on Random Sample

```
def inference_on_random_samples(model, X_test, y_test, label_encoder, num_samples=8):
    random_indices = np.random.randint(0, len(X_test), size=num_samples)

    for idx in random_indices:
        sample = X_test[idx]
        true_label = np.argmax(y_test[idx])

        plt.imshow(sample.astype("uint8"))
        plt.title(f"True Label: {label_encoder.classes_[true_label]}")
        plt.axis("off")
        plt.show()

        sample_expanded = np.expand_dims(sample, axis=0)
        predictions = model.predict(sample_expanded)
        predicted_label = np.argmax(predictions)

        print(f"Predicted Label: {label_encoder.classes_[predicted_label]}")
```

Figure 121: Inference Function

The best inference model will guess most of the spectrum images true, if the model fails to do so, it likely suffers from overfitting.

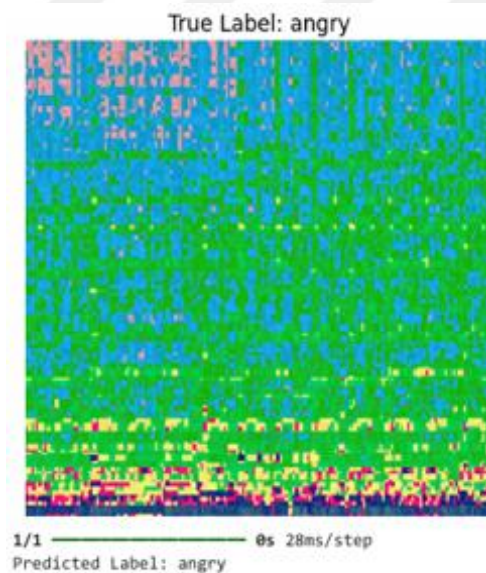


Figure 122: Inference Sample on best model

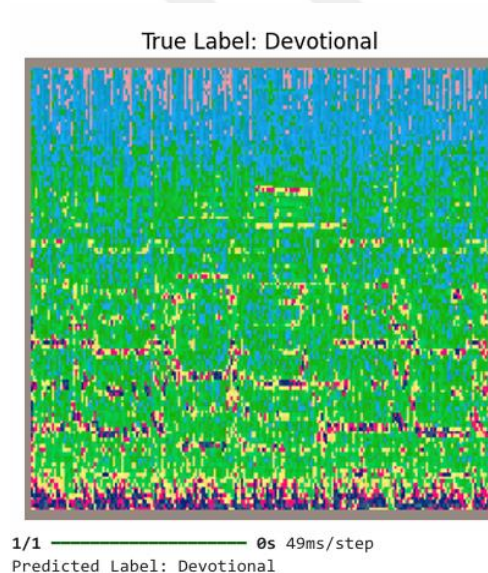


Figure 123: Inference Sample on best model.