



**Eight Music Emotions Recognition System using Neural Network
with Cascaded Model**

Kanawat Sorussa

**A Thesis Submitted in Partial Fulfillment of the Requirements for the
Degree of Master of Computer Engineering
Prince of Songkla University**

2019

Copyright of Prince of Songkla University



**Eight Music Emotions Recognition System using Neural Network
with Cascaded Model**

Kanawat Sorussa

**A Thesis Submitted in Partial Fulfillment of the Requirements for the
Degree of Master of Computer Engineering
Prince of Songkla University
2019**

Copyright of Prince of Songkla University

Thesis Title Eight Music Emotions Recognition System using Neural Network
with Cascaded Model

Author Mr. Kanawat Sorussa

Major Program Computer Engineering

Major Advisor

.....
(Dr. Anant Choksuriwong)

Co-Advisor

.....
(Assoc. Prof. Dr. Montri Karnjanadecha)

Examining Committee:

.....Chairperson
(Asst. Prof. Dr. Nikom Suvonvorn)

.....Committee
(Dr. Anant Choksuriwong)

.....Committee
(Assoc. Prof. Dr. Montri Karnjanadecha)

.....Committee
(Assoc. Prof. Dr. Wattanapong Kurdthongmee)

The Graduate School, Prince of Songkla University, has approved this thesis
as Partial fulfillment of the requirements for the Master of Engineering Degree in
Computer Engineering

.....
(Prof. Dr. Damrongsak Faroongsarng)
Dean of Graduate School

This is to certify that the work here submitted is the result of the candidate's own investigations. Due acknowledgement has been made of any assistance received.

.....Signature

(Dr.Anant Choksuriwong)

Major Advisor

.....Signature

(Assoc. Prof. Dr.Montri Karnjanadecha)

Co-Advisor

.....Signature

(Mr. Kanawat Sorussa)

Candidate

I hereby certify that this work has not been accepted in substance for any degree, and is not being currently submitted in candidature for any degree.

.....Signature

(Mr. Kanawat Sorussa)

Candidate

Thesis Title	Eight Music Emotions Recognition System using Neural Network with Cascaded Model
Author	Mr. Kanawat Sorussa
Major Program	Computer Engineering
Academic Year	2018

ABSTRACT

Music selection is difficult without an efficient organization based on metadata or tags, and one effective tag scheme is based on the emotion expressed by the music. The main drawback of such a system is that manually tagging music files because tagging a large number of files is a tedious work and emotional perception of each person is different. Therefore, this thesis presents a music emotion classification system for eight emotional classes with cascaded model. Russell's emotion model was adopted as a common ground for emotional annotation. The system implements on MATLAB using MIR toolbox to extract acoustic features from audio files and employed a supervised machine learning technique to recognize acoustic features to create predictive models. Four predictive models were proposed and compared. The models were composed by crossmatching two types of neural networks, i.e., Levenberg-Marquardt (LM) and resilient backpropagation (Rprop) with two types of structures: a traditional multiclass unit and multiple units of binary-class with a cascaded structure. The performance of each model was evaluated via the DEAM benchmark. The best result was achieved by the model trained with a cascaded Rprop neural network (accuracy of 89.5%). In addition, correlation coefficient analysis showed that timbre features were the most impactful for prediction. Our work offers an opportunity for a competitive advantage because only a few music providers currently tag music with emotional terms.

ชื่อวิทยานิพนธ์	ระบบรู้จำแปดอารมณ์ดนตรี โดยวิธีการ โครงข่ายประสาทเทียม แบบ โครงสร้างน้ำตก
ผู้เขียน	นายกณวัฒน์ โสรัสสะ
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
ปีการศึกษา	2561

บทคัดย่อ

การเลือกเพลงให้ตรงกับความต้องการนั้นไม่ใช่เรื่องง่าย หากเพลงเหล่านั้นไม่ได้รับการจัดหมวดหมู่มาก่อนโดยใช้ข้อมูลอภิพันธ์ การจัดหมวดหมู่รูปแบบหนึ่งที่มีประสิทธิภาพคือ การจัดหมวดหมู่ตามอารมณ์ดนตรี แต่ทว่าการจัดหมวดหมู่รูปแบบดังกล่าวด้วยมนุษย์ สำหรับไฟล์เพลงจำนวนมากนั้น ไม่อาจจะกระทำได้อย่างมีประสิทธิภาพ เนื่องจากความเหนื่อยล้า และจินตคติที่แตกต่างกันไปในแต่ละบุคคล งานวิจัยชิ้นนี้จึงขอเสนอ ระบบจำแนก และตัวแบบ แบบน้ำตก สำหรับการจำแนกเพลงออกเป็นแปดกลุ่มของอารมณ์ดนตรี โดยได้อ้างอิงความหมายของอารมณ์ต่างๆ ตามแบบของ “รัสเซลล์” เพื่อใช้เป็นหลักเกณฑ์ในการจำแนกเพลงไปตามหมวดหมู่ต่างๆ ระบบต้นแบบพัฒนาบน “แมทแลป” โดยใช้ “เอ็มไออาร์ ทูลบ็อกซ์” เป็นเครื่องมือสกัดคุณลักษณะทางเสียง ตัวแบบสี่แบบถูกเปรียบเทียบ โดยใช้โครงข่ายประสาทเทียมสองชนิด (ประสาทเทียมตามแบบของ “ลีเวนเบิร์ก-มัลล์การ์ท” และประสาทเทียมแบบยัดหยุ่น) จับคู่กับวิธีจัดโครงสร้างสองวิธี (โครงข่ายประสาทเทียมแบบดั้งเดิมและ แบบหลายหน่วยโดยใช้โครงสร้างแบบน้ำตก) ประสิทธิภาพของตัวแบบถูกประเมินด้วย “ติเม เบนซ์มาร์ค” ผลลัพธ์แสดงให้เห็นว่า กรรมวิธีสังเคราะห์ตัวแบบด้วยโครงข่ายประสาทเทียมแบบยัดหยุ่นหลายหน่วยแบบน้ำตก ให้ความถูกต้องอยู่ที่ 89.5 เปอร์เซ็นต์สำหรับการจำแนกเป็นแปดกลุ่มอารมณ์ อีกทั้งจากการวิเคราะห์ค่าสหสัมพันธ์ของคุณลักษณะทางเสียงแสดงให้เห็นว่า คุณลักษณะประเภท อัตลักษณ์ของเสียงจากเครื่องดนตรี มีผลต่อการทำนายมากที่สุด งานวิจัยชิ้นนี้สามารถสร้างความได้เปรียบในการแข่งขันให้แก่ผู้ให้บริการดนตรีได้ เนื่องจากทุกวันนี้ยังคงมีผู้ให้บริการดนตรีจำนวนไม่มากนัก ที่จัดหมวดหมู่ดนตรีตามอารมณ์ดนตรี

ACKNOWLEDGEMENT

I would like to express my deepest gratitude toward my advisors, Assistant Professor Dr. Anant Choksuriwong and Associate Professor Dr. Montri Kajanadecha, for the helpful suggestions and I am very proud to be their advisee.

I would like to express my warmest gratitude toward Assistant Professor Dr. Nikom Suwanavorn and Assistant Professor Dr. Thanate Kkaorapapong, along with my seniors and friends in Intelligent Automation Research Center (IARC) for their useful comments.

I am very thankful to DR. Andrew Davison for proofreading my publications.

I really appreciate for the scholarship and financial support for research material, supported by faculty of engineering and graduate school of Prince of Songkla University.

Kanawat Sorussa

TABLE OF CONTENTS

	Page
ABSTRACT.....	v
บทคัดย่อ	vi
ACKNOWLEDGEMENT	vii
TABLE OF CONTENTS.....	viii
LIST OF TABLES	xi
LIST OF FIGURES	xiii
CHAPTER 1 INTRODUCTION.....	1
1.1 Motivation	1
1.2 Outline	2
1.3 Objectives	4
1.4 Scopes.....	4
1.5 Contributions	4
CHAPTER 2 LITERATURE REVIEW	5
2.1 Emotion Model.....	5
2.1.1 Categorical Psychometrics.....	5
2.1.2 Dimensional Psychometrics.....	6
2.2 Previously Methodologies	8
2.2.1 Regression Approach	8
2.2.2 Classification Approach.....	11
2.3 Datasets for Music Emotion Recognition.	13
2.4 Acoustic Features	14
2.4.1 Dynamic	14
2.4.2 Rhythm.....	15
2.4.3 Timbre.....	15
2.4.4 Pitch	16
2.4.5 Tonality	17
2.5 Music Information Retrieval Tools	17

2.5.1 MIR Toolbox	18
2.5.2 Timbre Toolbox	19
2.5.3 Tempogram Toolbox	19
2.5.4 Chromagram Toolbox	19
2.5.5 Essentia	20
2.6 Artificial Neural Network	21
2.6.1 Network Architecture.....	21
2.6.2 Backpropagation Algorithms	25
2.7 Summary	26
CHAPTER 3 METHODOLOGY	27
3.1 System Environment	27
3.2 Data Preparation	27
3.2.1 Audio files.....	27
3.2.2 Annotations	27
3.3 Feature Extraction	30
3.4 System Structure.....	31
3.4.1 Regression Approach	32
3.4.2 Classification Approach.....	33
3.5 Model Structure	33
3.5.1 Traditional Multiclass Model.....	33
3.5.2 Cascaded Model.....	34
3.6 Model Training.....	35
3.6.1 Preprocessing	36
3.6.2 Parameter Configuration	37
3.7 Summary	42
CHAPTER 4 RESULTS AND DISCUSSIONS	44
4.1 Feature Extraction	44
4.2 Feature Correlation.....	45
4.3 Modeling Results.....	47
4.3.1 Binary-Class LM Network for Valence Prediction	49
4.3.2 Binary-Class LM Network for Arousal Prediction.....	50

4.3.3 Traditional Multiclass LM Network for Four Emotions Prediction	50
4.3.4 Traditional Multiclass LM Network for Eight Emotions Prediction	51
4.3.5 Cascaded LM Network for Four Emotions Prediction	52
4.3.6 Cascaded LM Network for Eight Emotions Prediction	53
4.3.7 Binary-Class Rprop Network for Valence Prediction	55
4.3.8 Binary-Class Rprop Network for Arousal Prediction.....	55
4.3.9 Traditional Multiclass Rprop Network for Four Emotions Prediction	56
4.3.10 Traditional Multiclass Rprop Network for Eight Emotions Prediction	57
4.3.11 Cascaded Rprop Network for Four Emotions Prediction	58
4.3.12 Cascaded Rprop Network for Eight Emotions Prediction	59
4.4 Discussions	61
4.4.1 The Relationship of Submodels in Cascaded Models.....	61
4.4.2 Performance Measurement using F1-score.....	65
4.4.3 Results Comparison using Key Performance Indicators	68
4.5 Summary	73
CHAPTER 5 CONCLUSION	74
5.1 Conclusion.....	74
5.2 Future Work	75
REFERENCES	76
APPENDIX.....	82
VITAE.....	87

LIST OF TABLES

Table	Page
Table 2.1 Hevner's emotion adjectives [9]	5
Table 2.2 Emotional octant and their emotional adjectives	7
Table 2.3 List of input functions (f_1).....	23
Table 2.4 List of transfer functions (f_2).....	24
Table 3.1 Number of populations in each class	29
Table 3.2 Training dataset and purpose of each unit in Figure 3.9.....	35
Table 3.3 The parameters configuration	39
Table 4.1 List of MIR toolbox's function for acoustic feature extraction	44
Table 4.2 Correlation values between extracted features and VA-ratings.....	47
Table 4.3 The performance of binary-class LM networks for valence discrimination	49
Table 4.4 The performance of binary-class LM networks for arousal discrimination	50
Table 4.5 The performance of traditional multiclass LM networks for four emotions prediction	51
Table 4.6 The performance of traditional multiclass LM networks for eight emotions prediction	52
Table 4.7 The performance of cascaded LM networks for four emotions prediction .	53
Table 4.8 The performance of cascaded LM networks for eight emotions prediction	54
Table 4.9 The performance of binary-class Rprop networks for valence discrimination	55
Table 4.10 The performance of binary-class Rprop networks for arousal discrimination	56
Table 4.11 The performance of traditional multiclass Rprop networks for four emotions prediction	56

Table 4.12 The performance of traditional multiclass Rprop networks for eight emotions prediction.....	57
Table 4.13 The performance of cascaded Rprop networks for four emotions prediction	59
Table 4.14 The performance of cascaded Rprop networks for eight emotions prediction	60
Table 4.15 The F1-score of four emotions classification	65
Table 4.16 The F1-score of eight Emotion Classification	67
Table 4.17 The comparison of key performance indicators in each previous work	69

LIST OF FIGURES

Figure	Page
Figure 1.1 A music emotion classification system framework [1]	2
Figure 2.1 Russell’s emotion adjectives [10].....	6
Figure 2.2 Fixed aggregation (left), adaptive aggregation (right) [16]	9
Figure 2.3 Nguyen’s Segmented VA-plane [18]	10
Figure 2.4 Wang and Xin modeling methodology [2].....	11
Figure 2.5 Chiang <i>et al.</i> , modeling methodology [3].....	11
Figure 2.6 Number of music samples in the DEAM dataset associated with the eight emotions.....	14
Figure 2.7 The comparison of louder and lighter audio wave	15
Figure 2.8 The illustration of dynamic pattern	15
Figure 2.9 The close-up audio wave reveals a timbre	16
Figure 2.10 The comparison of consonant and dissonant audio waves.....	16
Figure 2.11 The comparison of C major and C minor and their audio wave	17
Figure 2.12 Acoustic feature extraction tools choosing path [37]	18
Figure 2.13 Comparation of musical note and chromagram [44].....	19
Figure 2.14 Schematic drawing of biological neurons [47].....	21
Figure 2.15 Schematic of artificial neurons.....	22
Figure 2.16 Function of hidden node (left) and output node (right).....	23
Figure 3.1 Dynamic annotation to static annotation transformation.....	28
Figure 3.2 The visualization of the dataset on VA plane.....	29
Figure 3.3 The acoustic features exacted by MIR toolbox	30
Figure 3.4 Feature extraction post process flow chart	31
Figure 3.5 Framework for regression approach system.....	32

Figure 3.6 Circular acceptable areas of error (left), triangular acceptable areas of error (right)	32
Figure 3.7 Framework for classification approach system	33
Figure 3.8 Traditional structure of predictive model, LM algorithm (left) & Rprop algorithm (right).....	34
Figure 3.9 Schematic cascaded structure diagram of multi-model neural network.....	35
Figure 3.10 Pre/Post processing of input/output data	36
Figure 3.11 Schematic of artificial neurons using in this study.....	37
Figure 3.12 Structure of traditional multiclass and each unit of cascade LM neural network	38
Figure 3.13 Structure of traditional multiclass Rprop neural network	38
Figure 3.14 Structure of each unit of cascaded Rprop neural network.....	38
Figure 3.15 Mean Squared Error (MSE) display	40
Figure 3.16 Stopping criteria display.....	40
Figure 3.17 The entire system flow chart	43
Figure 4.1 Scatter plot of feature No. 49, 80, and 29 against valence and arousal.....	46
Figure 4.2 The confusion matrix of binary-class LM networks for valence discrimination	49
Figure 4.3 The confusion matrix of binary-class LM networks for arousal discrimination	50
Figure 4.4 The confusion matrix of traditional multiclass LM networks for four emotions prediction.....	51
Figure 4.5 The confusion matrix of traditional multiclass LM networks for eight emotions prediction.....	52
Figure 4.6 The confusion matrix of cascaded LM networks for four emotions prediction	53

Figure 4.7 The confusion matrix of cascaded LM networks for eight emotions prediction	54
Figure 4.8 The confusion matrix of binary-class Rprop networks for valence discrimination	55
Figure 4.9 The confusion matrix of binary-class Rprop networks for arousal discrimination	56
Figure 4.10 The confusion matrix of traditional multiclass Rprop networks for four emotions prediction.....	57
Figure 4.11 The confusion matrix of traditional multiclass Rprop networks for eight emotions prediction.....	58
Figure 4.12 The confusion matrix of cascaded Rprop networks for four emotions prediction	59
Figure 4.13 The confusion matrix of cascaded Rprop networks for eight emotions prediction	60
Figure 4.14 Number of correctly predicted sample in each submodel in case of perfect accuracy (left), ratio of sample in each submodel (right)	63
Figure 4.15 Submodel accuracies of cascaded LM model (left), number of correctly predicted sample in each submodel in case of cascaded LM model (right)	63
Figure 4.16 Submodel accuracies of cascaded Rprop model (left), number of correctly predicted sample in each submodel in case of cascaded Rprops model (right).....	63
Figure 4.17 The comparison of F1-score of four emotion classification.....	65
Figure 4.18 The comparison of F1-score of eight emotion classification	67

CHAPTER 1

INTRODUCTION

1.1 Motivation

The appearance of digital music providers such as “iTunes” has changed the way people listen to music by offering more direct access to a vast collection of music. However, finding the right music is not easy without appropriate tags or metadata to help the search. Creating metadata manually is expensive and time consuming. Music information retrieval (MIR) attempts to address these problems. MIR is an interdisciplinary science that combines musicology, psychology, signal processing, and machine learning [1].

Emotional adjectives, such as search keywords, are particularly effective for nonvocal music, such as classical music and film soundtracks, and 28% of people use emotional keywords to search for music [1]. Unfortunately, most music providers tag music by genre, the artist’s name, year of production, and type of instrument, and rarely provide tags such as emotional terms. A branch of MIR known as music emotion recognition (MER) attempts to address this problem.

1.2 Outline

Yang and Chen proposed the conceptual framework for an MER system, as shown in Figure 1.1 [1].

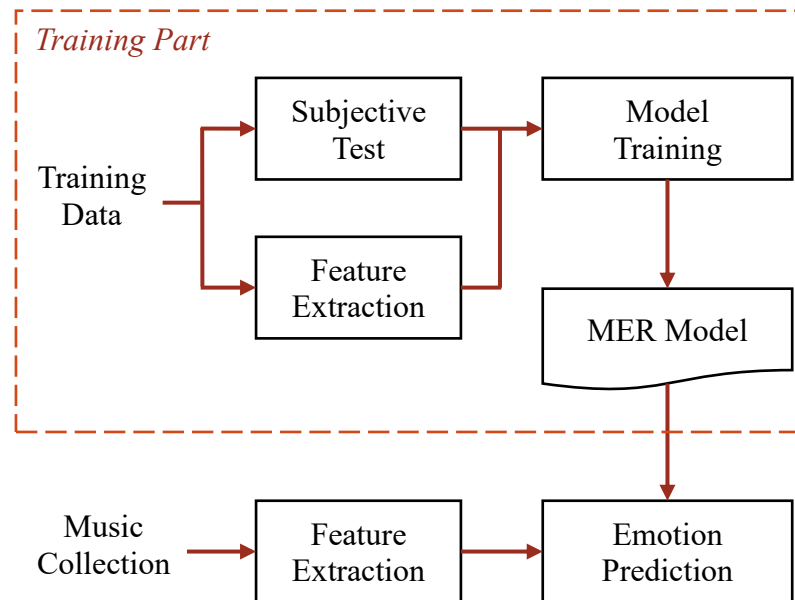


Figure 1.1 A music emotion classification system framework [1]

First, music was collected and annotated with one of the emotion models mentioned in Chapter 2.1, while acoustic features were extracted from the audio file. Then, a supervised machine learning technique was applied to reveal the relationship between music emotions and acoustic features.

In Chapter 2.2, We briefly review 14 publications since 2008. Several studies using four to six emotional classifications obtained over 80% accuracy [2–6], but six classes are insufficient in practice. Some studies used a small dataset with a limited variety [2][3][5][7], which could be a problem when the system tried to predict songs that were not in the dataset. According to the results of earlier studies using multiple models for prediction is more accurate than using a traditional multiclass model [2][3][8]. We hypothesized that using multiple models with cascaded structures could reduce the number of false predictions because each model is specifically trained to discriminate only two classes at a time. Therefore, this study makes a major contribution to classifying eight music emotions via a neural network with a cascaded structure while maintaining an accuracy greater than 80%. The models were trained

with a large dataset of 1,802 songs. Additionally, a large number of acoustic features were extracted.

We conducted four experiments using two types of neural networks and two methods of constructing the networks. The Levenberg-Marquardt backpropagation (LM) algorithm was investigated using a traditional structure and a cascaded structure. Additionally, the resilient backpropagation (Rprop) algorithm was examined with the same two types of structure.

We found that the cascaded Rprop algorithm achieves the best accuracy compared not only to the other three methods but also previously proposed methods. A comparison of the results is shown in Table 4.17 of Chapter 4.4.3.

1.3 Objectives

To design and develop music emotion recognition modeling method.

1.4 Scopes

1. Dataset: DEAM Benchmark
2. Number of emotional class: eight.
3. Implementation tool: MATLAB
4. Features extraction tool: Music Information Retrieval Toolbox.
5. Recognition algorithm: supervised machine learning technic using shallow neural network.

1.5 Contributions

1. The modeling method using traditional Levenberg-Marquardt backpropagation neural network algorithm.
2. The modeling method using cascaded structure of Levenberg-Marquardt backpropagation neural network algorithm.
3. The modeling method using traditional resilient backpropagation neural network algorithm.
4. The modeling method using cascaded structure of resilient backpropagation neural network algorithm.

CHAPTER 2

LITERATURE REVIEW

There are five subchapters in this chapter, firstly we will discuss about how psychologist define each motion. Then we will survey some recent studies on MER including available datasets. The last two subchapters are about acoustic features, the description acoustic features are given in Chapter 2.4 and we will introduce some tools to extract those features at the last subchapter.

2.1 Emotion Model

Emotions have been measured in two ways in psychological studies. Some psychologists believe that emotions are discrete perceptions and propose models based on categorical psychometrics. Others believe that emotion is a continuous level of perception and have proposed models employing dimensional psychometrics. We describe the most influential models from each psychometric perspective below.

2.1.1 Categorical Psychometrics

Categorical psychometrics represent emotional perception by a finite set of emotional descriptors. One of the earliest models, proposed by Hevner, consists of 66 emotional adjectives. Similar adjectives are arranged into related emotional groups, forming eight clusters of emotions [9]. Similar adjectives are arranged into related emotional groups, forming eight groups in total as shown on Table 2.1.

Table 2.1 Hevner's emotion adjectives [9]

Clusters	Emotional Adjectives
Cluster 1	spiritual, lofty, awe-inspiring, dignified, sacred, solemn, sober, serious
Cluster 2	pathetic, doleful, sad, mournful, tragic, melancholy, frustrated, depressing, gloomy, heavy, dark
Cluster 3	dreamy, yielding, tender, sentimental, longing, yearning, pleading, plaintive
Cluster 4	lyrical, leisurely, satisfying, serene, tranquil, quiet, soothing
Cluster 5	humorous, playful, whimsical, fanciful, quaint, sprightly, delicate, light, graceful

Table 2.1 (Continued) Hevner’s emotion adjectives [9]

Clusters	Emotional Adjectives
Cluster 6	merry, joyous, gay, happy, cheerful, bright
Cluster 7	exhilarated, soaring, triumphant, dramatic, passionate, sensational, agitated, exciting, impetuous, restless
Cluster 8	vigorous, robust, emphatic, martial, ponderous, majestic, exalting

This approach is easy to understand and more meaningful than dimensional psychometrics. However, some emotional adjectives do not exist in some languages, or have different meanings, and emotions are difficult to compare to each other.

2.1.2 Dimensional Psychometrics

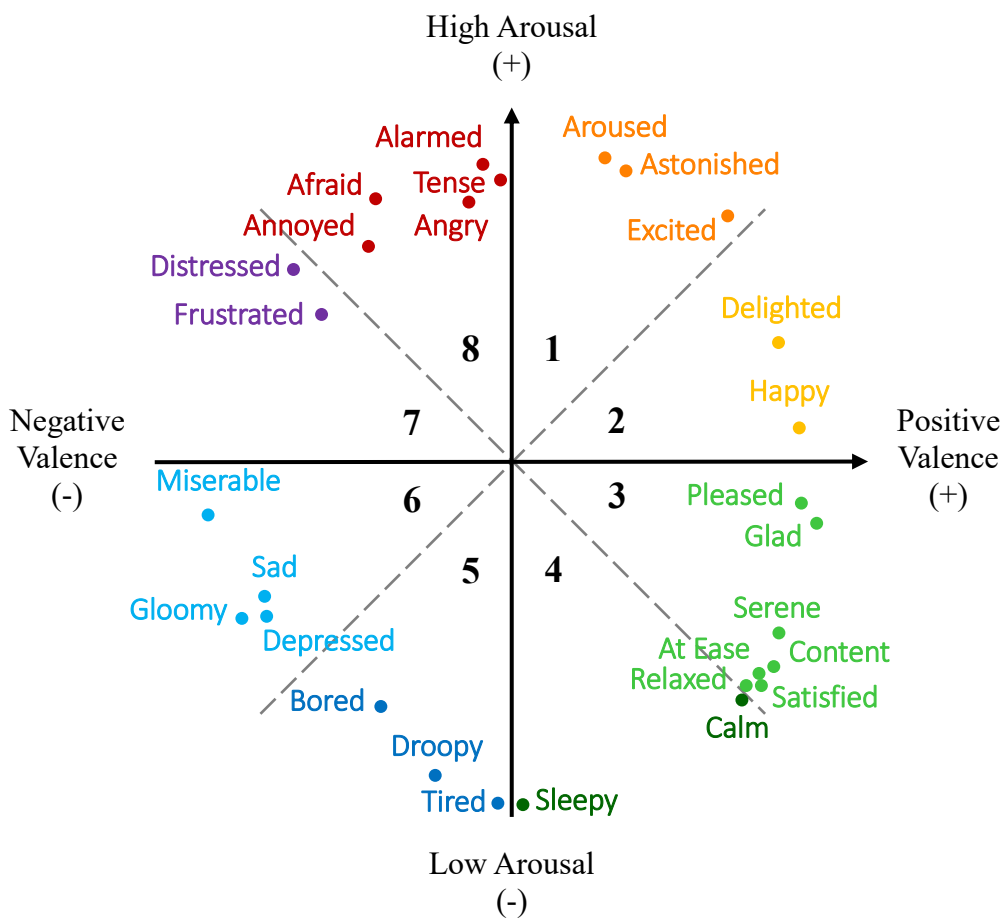


Figure 2.1 Russell’s emotion adjectives [10]

Dimensional psychometrics represent emotional perception by numeric values plotted along fundamental emotional axes. The most influential model, proposed by Russell, uses two dimensions of fundamental factors, i.e., valence (pleasantness, positive, and negative affective states) and arousal (danceability, activation, energy, and stimulation levels), to form the valence–arousal (VA) plane as shown in Figure 2.1 [10].

Various valence and arousal coordinates define 28 emotional adjectives. For example, “happy” is assigned arousal equal to 0.11 and valence equal to 0.83 with the axes scaled to be between -1 and +1. This approach is flexible, measurable, and comparable, but the relationships between valence and arousal can be difficult to explain.

We adopted Russell’s model, with only eight groups of emotion as our goal for classification. The emotional octant and their emotional adjectives are listed in Table 2.2.

Table 2.2 Emotional octant and their emotional adjectives

#	VA Logical Range	Emotions
1	High Arousal & Positive Valence (Valence < Arousal)	Aroused, Astonished, Excited
2	High Arousal & Positive Valence (Valence > Arousal)	Delighted, Happy
3	Low Arousal & Positive Valence (Valence > Arousal)	Pleased, Glad, Serene, Content, At Ease, Satisfied, Relaxed
4	Low Arousal & Positive Valence (Valence < Arousal)	Calm, Sleepy
5	Low Arousal & Negative Valence (Valence < Arousal)	Tired, Droopy, Bored
6	Low Arousal & Negative Valence (Valence > Arousal)	Depressed, Gloomy, Sad, Miserable
7	High Arousal & Negative Valence (Valence > Arousal)	Frustrated, Distressed
8	High Arousal & Negative Valence (Valence < Arousal)	Annoyed, Afraid, Angry, Tense, Alarmed
Σ	8	28

2.2 Previously Methodologies

Music processing retrieves information in many forms, such as score notes, lyrics, audio signals, and chords [11–13]. Music emotion is often annotated via people’s verbal reports of emotional responses, although some studies have gathered data by monitoring biological or physical expressions [14]. However, we are interested only in the retrieval of information from audio signals and annotations from verbal reports.

Systems recognize music by referring to the psychometrics described above by means of two approaches. For dimensional psychometrics, a regression approach estimates the valence and arousal, whereas for categorical psychometrics, a classification approach is employed.

2.2.1 Regression Approach

Yang *et al.* employed a support vector machine (SVM) as the regressor and ranked the importance of the predictors using the ReliefF algorithm for feature selection. The performance was evaluated with respect to the R^2 statistic, and 28.1% valence and 58.3% arousal was achieved [7].

Weninger *et al.* captured time-varying emotion through a music piece by recurrent neural networks (RNNs). The performance was evaluated by R^2 statistics, and 50% valence and 70% arousal was achieved [15].

One of the common problems with multiple-feature input data is the importance ranking of features. The features that have a substantial effect on estimation should be given some bias weight to improve the results of the calculation. For example, Fukayama and Goto utilized adaptive aggregation to obtain the feature ranking and estimated the VA-value via Gaussian process regressors. Figure 2.2 illustrate the concept of Fukayama’s method [16].

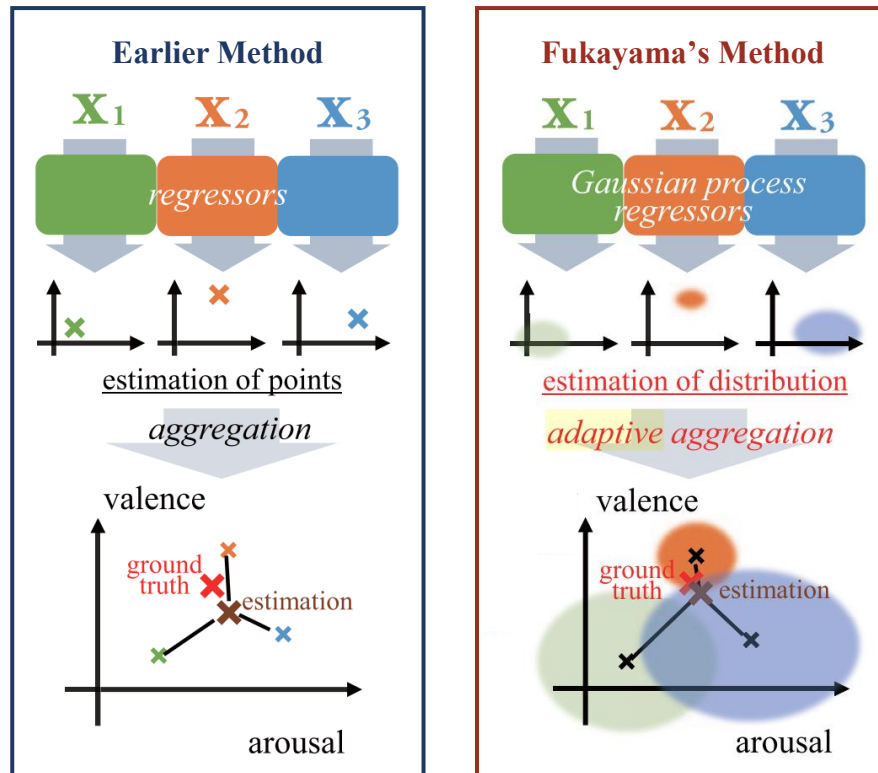


Figure 2.2 Fixed aggregation (left), adaptive aggregation (right) [16]

The inputs X_1, X_2, X_3 appear on Figure 2.2 refer to the feature elements. The fixed aggregation takes these features equally, but adaptive aggregation also considers variation range of those features too. The smaller circle area means the features have lower scatter and rarely to find these patterns of feature elsewhere on the VA-plane. The result show that adaptive aggregation overperform the fixed aggregation by 4% and 2.7% for valence and arousal estimation respectively, The performance evaluated in terms of the root-mean-square error (RMSE) reached 77% for valence and 80% for arousal [16].

A recent study proposed by Malik *et al.* used stacked neural networks. The authors employed a convolutional neural network (CNN) on the top layer, followed by two RNN branches, each trained separately, for valence and arousal. The RNNs were applied to time-varying features, while the CNN handled time-invariant features. The CNN's features map was the input to both RNNs. The performance was evaluated in terms of RMSE, which was 73% for valence and 80% for arousal [17].

Even though most of VA-value estimation problems can be solved by regression algorithm, but some researchers think of the problem as classification problems by convert a continuous range to a finite range. Nguyen *et al.*, divided valence and arousal level into six segments eventually, coordinate of those segment gave 36 segments in total [18]. Figure 2.3 show how Nguyen et al., divided valence and arousal level.

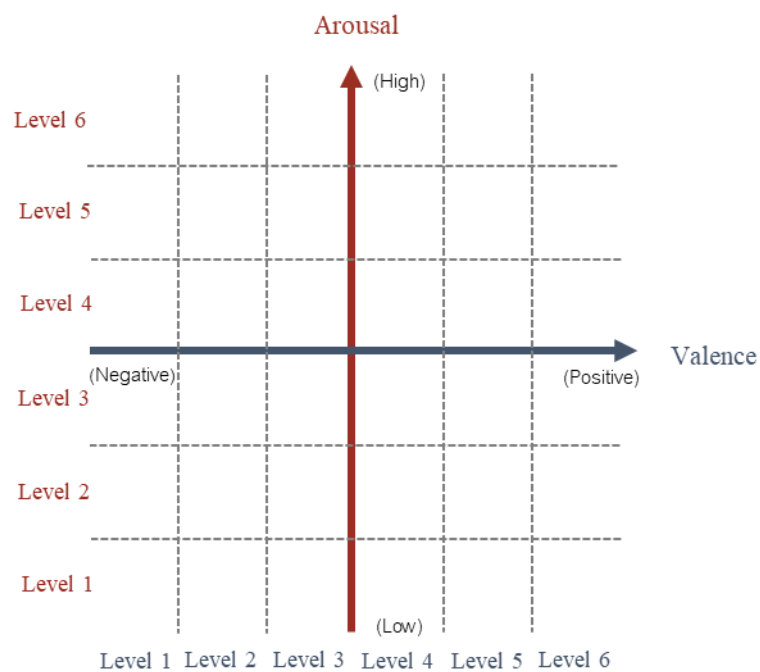


Figure 2.3 Nguyen's Segmented VA-plane [18]

Then, RandomForest algorithm is employed by implemented on WEKA, to classify valence and arousal to one of these six level. The accuracies are 57.3% for valence 70% for arousal.

Hu and Yang created dataset of Chinese-pop music (C-pop) for MER task, this is a rare work since most of MER task conducted on western music [6]. The dataset has tested by both regression and classification approach. First, the dataset is tested by Support Vector Machine (SVM) and got 85% of accuracy for six emotion classification. Then the dataset is tested again by Support Vector Regressor (SVR), the accuracies are 25% for valence 79% for arousal.

2.2.2 Classification Approach

A small-scale experiment demonstrated that music could be classified into four emotional classes with the help of a hierarchical SVM by using only two features, tempo and mutation degree. Three SVM nodes were utilized, with each node trained for a specific purpose. The first node was trained to discriminate high and low tempo: high-tempo songs are happy or aggressive, while low-tempo songs are sad or soft. The second and third nodes were trained to discriminate between those emotions. The results were impressive, with 95% accuracy [2].

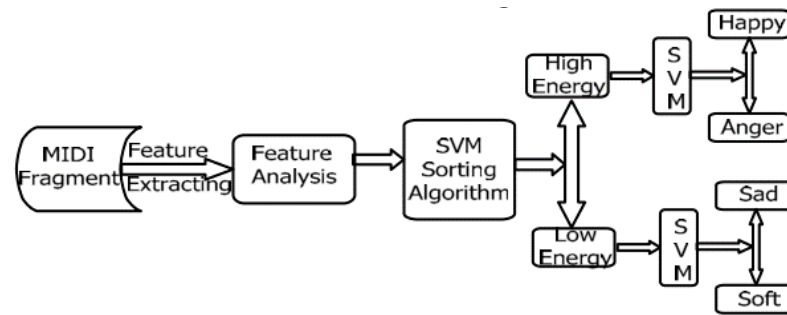


Figure 2.4 Wang and Xin modeling methodology [2]

A similar concept was investigated again but scaled up to larger number of feature and sample. The same SVM structure was employed, and the result were satisfactory, with 89.64% accuracy [3].

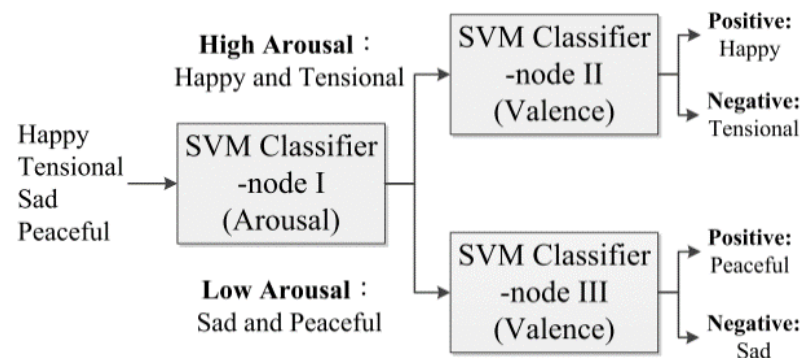


Figure 2.5 Chiang *et al.*, modeling methodology [3]

An investigation of six algorithms, i.e., SVM, k-nearest neighbors (KNN), neuro-fuzzy network classification (NFNC), fuzzy KNN (FKNN), a Bayes

classifier, and linear discriminant analysis (LDA), for classifying four emotional classes showed that the accuracy of the LDA, SVM, and FKNN algorithms was higher than 80% [4].

Nalini et al. investigated auto associative neural networks (AANN) and SVM for classifying five music emotions. The accuracies were 94.4% and 85.0%, but the models were trained with a small dataset [5]. The accuracy results were 94.4% and 85.0% respectively, but were conducted on a small dataset. Another study applying a nearest multiprototype classifier to a huge dataset achieved only 56.43% accuracy [19].

Trohidis et al. investigated how four algorithms handled six emotional classes. The four algorithms were binary relevance (BR), label powerset (LP), random k-labelsets (RAKEL), and multilabel k-nearest neighbor (MLkNN), and all achieved approximately 70% accuracy [20].

Deng et al., conducted a study of classifying eight music emotions by employing eight regressors to estimate the likelihood of each emotional class, with each regression model trained individually. This method did not classify each song separately but rated the likelihood of each emotion in each song. Therefore, more than one emotion could be assigned to each song. The accuracy was almost 60%, which was impressive considering the number of samples, number of emotional classes, and the proposed method [8].

Most MER studies focus on only acoustic features as inputs and ignore non-acoustic features, such as artist and genre. However, the impact of these non-acoustic features on the classification of four music emotions was studied by Vale et al. The experiment considered twenty-eight cases obtained by combining three groups of features (artist, genre, and acoustic features) and four types of classification algorithms (SVM, naïve Bayes, decision trees, and KNN). The models were evaluated with the DEAM benchmark, and their F-scores were 46%, 40%, 37%, and 41% respectively. The artist feature was not impactful, and the genre feature was only slightly beneficial for the decision trees method. The overall accuracy was not high because the experiments considered a limited number of acoustic features [21].

The exact numbers of samples and features and the results of the research mentioned in this chapter are reported in Table 4.17, which includes the results of our work for comparison.

In this work, we have employed the technique of using multiple models, training each model separately, and constructing a hierarchical classifier. We believe that multiple models with a cascaded structure would enhance the number of emotional classes.

2.3 Datasets for Music Emotion Recognition.

Most music datasets do not include audio files because of intellectual property concerns. Instead, the datasets provide emotional annotations, and lists of songs and where to find them [22–24]. Some datasets include extracted features [25], and some datasets consider the cultural background of the annotators [6][26]. Datasets that do not provide audio files can lead to problems because we cannot make any potentially required changes to the process.

Fortunately, the MediaEval Database for Emotional Analysis (DEAM) benchmark includes a dataset with audio files that can be redistributed under a Creative Commons (CC) license, so it was utilized in this work. The DEAM benchmark includes 1,802 songs. The audio files are in stereo MP3 format with a 44.1 kHz sampling rate. The music was collected from three sources (freemusicarchive.org, jamendo.com, and the medleyDB dataset) and includes a variety of genres (rock, pop, soul, blues, electronic, classical, hip-hop, experimental, folk, jazz, country, pop, rap, and reggae) in many languages. No more than five songs from the same artist are include [27–29]. The annotators were paid \$8 per hour to rate the valence and arousal separately via the crowdsourcing platform Amazon's Mechanical Turk (MTurk), and the annotators' background was not considered [30–32]. Each song was annotated by five to ten people, and we used the average of the annotations. Figure 2.6 shows the number of music samples in the DEAM dataset associated with each of the eight emotions, where numbers 1 to 8 refer to the emotional octant in Figure 2.1 and Table 2.2.

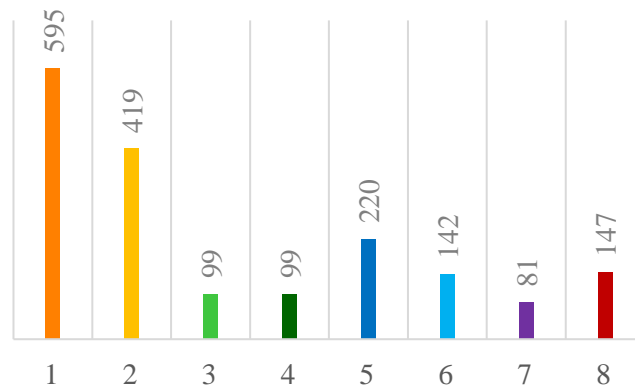


Figure 2.6 Number of music samples in the DEAM dataset associated with the eight emotions

As shown in Figure 2.6, the number of samples in each class is unequal. The inequality of training samples can bias the results, e.g., the prediction of classes 1 and 2, which have the largest populations, might achieve high accuracy, while the other classes might have low accuracy. The equalization of training samples by taking the number of samples in the smallest class as the ceiling and removing the excessive samples might solve this problem, but a previous study using the same dataset showed that even when the number of samples in each class was equalized, the prediction accuracy for classes 3, 4, 7, and 8 was not significantly improved [21]. The incorrect prediction of these classes must be caused by other characteristics; therefore, we did not equalize the number of samples.

2.4 Acoustic Features

The acoustic features were extracted using the MIR toolbox, which was run on MATLAB. The MIR toolbox relies on a built-in auditory toolbox and the Musical Instrument Digital Interface (MIDI) toolbox, which must be installed separately [33–35]. This tool was chosen because it can extract numerous features, including the five groups of features described below [22][36].

2.4.1 Dynamic

Dynamics is the physical intensity of a sound, and is often described as loudness, energy, volume, or audio power. This feature can be varied along the length of the song. Different dynamics audio waves are represented by amplitude height: the louder sound has a higher amplitude as shown in Figure 2.7.

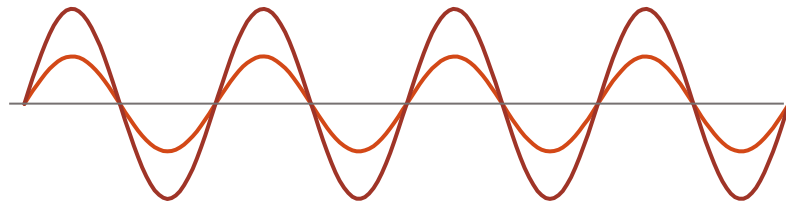


Figure 2.7 The comparison of louder and lighter audio wave

2.4.2 Rhythm

Rhythm is a periodic pattern of changes or events of pitch level, dynamics, or pulses. Pulse speed is known as meter, phrasing, tempo, or BPM (beat-per-minute). Figure 2.8 shows rhythm as part of the dynamics: at the beginning a low note is played with increasing loudness, then becomes steady, then suddenly changes to a higher note with the same pattern of dynamics.

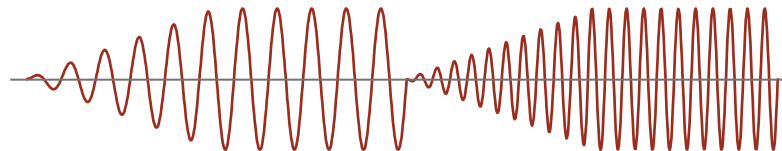


Figure 2.8 The illustration of dynamic pattern

2.4.3 Timbre

When a guitar or a violin plays the same note, the sound is similar, but we hear a difference which we call timbre. Each music instrument and sound-producing devices has its own timbre, which is particularly useful for musical instrument recognition. This appears in Figure 2.9 as imperfection in the sine wave. Difference musical instruments produce different characteristic wave forms. These characteristics are caused by the material and maintenance methods of the instruments.

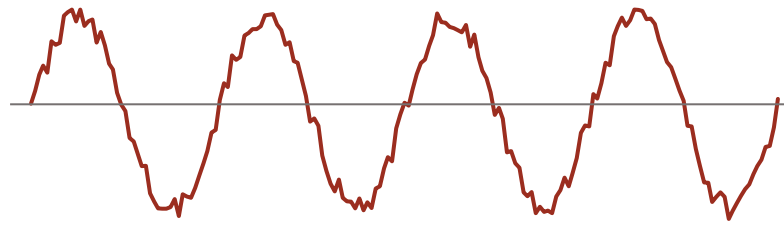


Figure 2.9 The close-up audio wave reveals a timbre

2.4.4 Pitch

Pitch is the different levels of sound, with Western music pitch encoded by the letters C, D, E, F, G, A, B. While a piano changes pitch level discretely, some instruments, such as a violin, can produce continuously change. In term of signal processing, the bass sound of a lower note is represented by less frequent oscillation, while the treble sound of a higher note is represented by more frequent oscillation, as shown in Figure 2.10.

Combinations of notes and their duration form a harmony, which is known as a chord or triad. There are two types of interval: those that sound harmonious or consonant as in Figure 2.10 (A), and those that sound dissonant as in Fig. 5 (B). Multiple consonant intervals sound naturally pleasant to our ears, but dissonant intervals conflict or clash with each other.

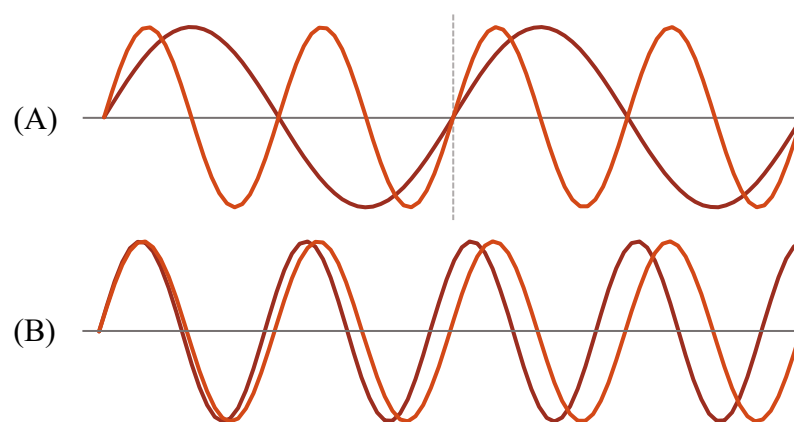


Figure 2.10 The comparison of consonant and dissonant audio waves

The brighter line in Figure 2.10 (A) is note A5 and the darker line is A4. When A4 completes one cycle, A5 has finished two, forming a consonance pattern. In Figure 2.10 (B), the brighter line is A5 and the darker line is B5, and are largely out of synchronization, and almost never complete their cycles together, thereby forming a dissonance pattern. Mostly consonance relates to positive valence, while dissonance corresponds to negative valence [6].

2.4.5 Tonality

Tonality is the arrangement of pitches and/or chords onto major and minor scales and keys. Major and minor scales refer to the spaces between notes, with note separation measured in whole and half steps. Figure 2.11 (A) illustrates a C major chord and their wave form is given in Figure 2.11 (a). Figure 2.11 (B) illustrates a C minor chord and their wave form is given in Figure 2.11 (b). On a piano keyboard, every half-step separates one-note and every whole-step is a two-note interval.

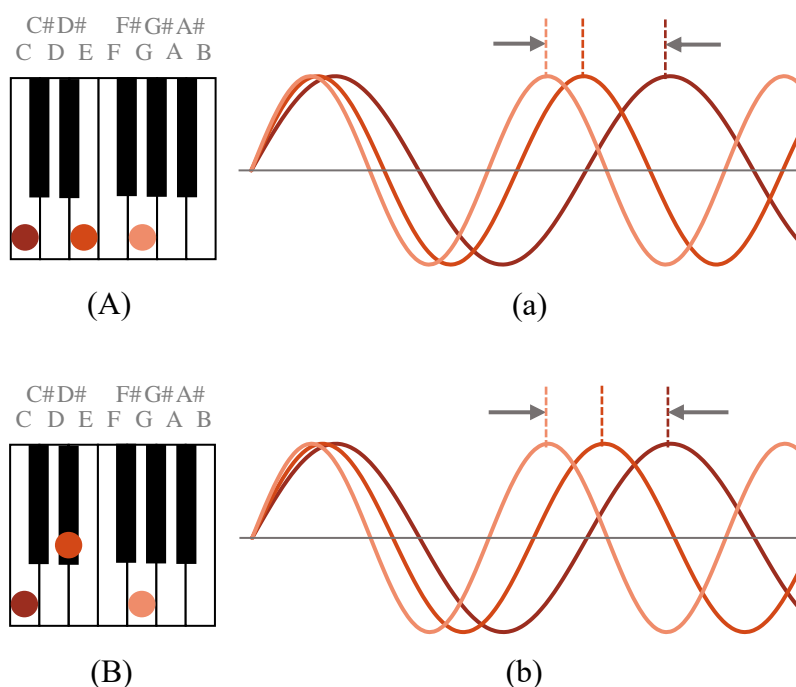


Figure 2.11 The comparison of C major and C minor and their audio wave

2.5 Music Information Retrieval Tools

Since the variety of tools for acoustic features extraction has been proposed to the community, choosing a suitable one for the project without any clue

could be a time-consuming process. Fortunately, D. Moffat *et al.*, have evaluated ten libraries and toolboxes for acoustic features extraction and propose a choosing method for a particular project [37]. Figure 2.12 shows the tool choosing path and the chosen path for this work is highlighted with the thicker arrows.

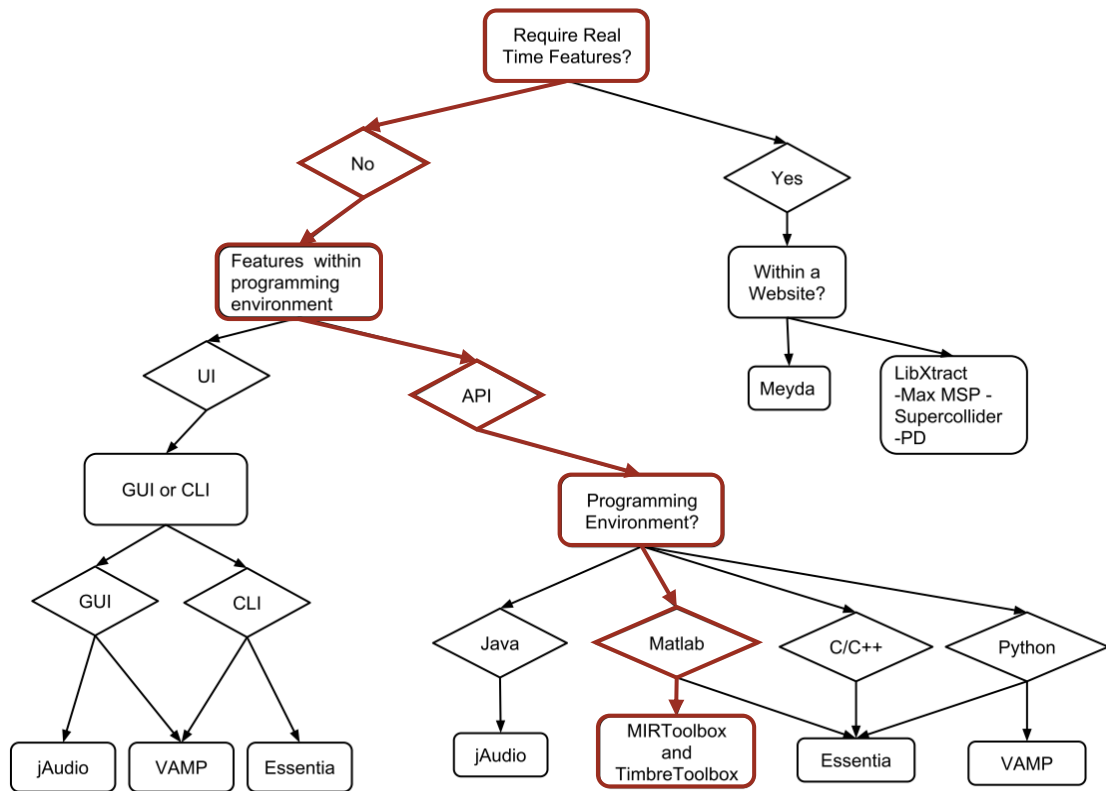


Figure 2.12 Acoustic feature extraction tools choosing path [37]

Since this work was conducted on MATLAB programming environment and it was not a real time processing, for those reason the suitable tools must be MIR toolbox, Timbre toolbox, and Essentia. Moreover, we have found other toolboxes which are compatible with MATLAB those are Tempogram and Chromagram toolbox. We will introduce these toolboxes in subchapters below.

2.5.1 MIR Toolbox

Music Information Retrieval (MIR) toolbox is a MATLAB library developed by O. Lartillot and P. Toivainen [33][35]. The MIR toolbox relies on a built-in auditory toolbox for basic audio processing. However, some additional function need Musical Instrument Digital Interface (MIDI) toolbox, which must be installed separately [34]. The advantage of the MIR toolbox its ability to extract all kinds of

acoustic feature mentioned in Chapter 2.4, number of features those can be extracted by this toolbox is more than one-hundred, but that ability sacrifices with heavy computational load. Nonetheless, this is the most powerful tool for music information retrieval task.

2.5.2 Timbre Toolbox

Timbre toolbox is a MATLAB library developed by G. Peeters *et al.*, [38][39], The timbre toolbox is especial made for musical instrument recognition problem, is able to distinguish the sound of the same note played by different kind of instrument or even the same kind of instrument but different model. The number of features that can be extracted by this toolbox is 32.

2.5.3 Tempogram Toolbox

Tempogram toolbox is a MATLAB library developed by P. Grosche and M. Müller for measure pulse speed of music [40][41]. The tempogram toolbox works base on Fourier and autocorrelation methods. This tool widely uses of beat or pulse tracking.

2.5.4 Chromagram Toolbox

Chromagram toolbox is a MATLAB library developed by M. Müller and S. Ewert design for extract Chroma-based (pitch or note) audio features [42][43]. Chromagram is representation of pitch level inform of spectrogram as shown in Figure 2.13.

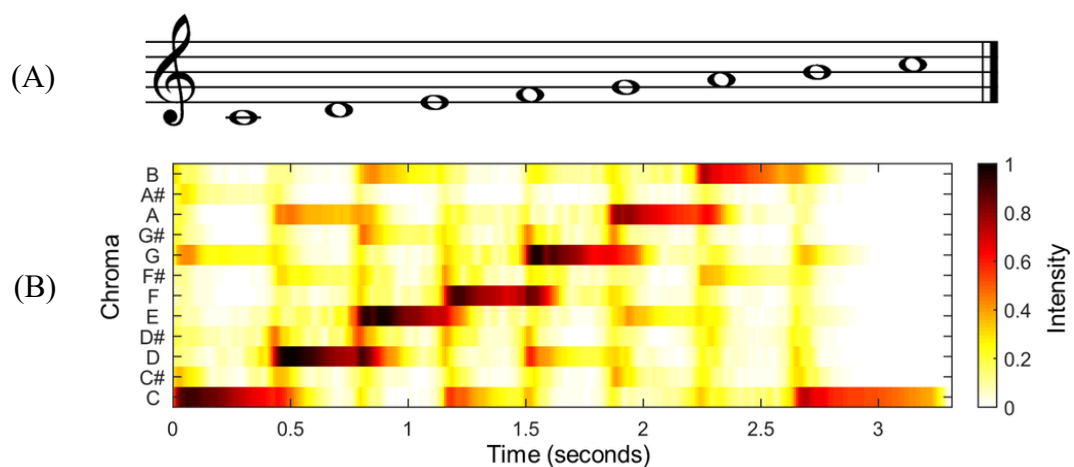


Figure 2.13 Comparison of musical note and chromagram [44]

Figure 2.13 shows the conversion of note Do middle C (C4) to higher Do (C5) to chromagram. Notes C to B represent by row of the chart, while the same note but difference tone such as C4 and C5 represented by tone of color, lower note has more intensity of sound or air pressure. Therefore, lower tone is represented by darker color. This feature useful for pitch tracking and these pitches also lead to harmony feature.

2.5.5 Essentia

Essentia is a standalone application developed by D. Bogdanov *et al.*, [45][46]. Even though it is a standalone application, it also provides application programming interface (API) so it is compatible with variety of programming environment (i.e. C, C++, Python, MATLAB). Essentia is the most flexible tool we have found.

2.6 Artificial Neural Network

The artificial neural network algorithm is an algorithm for recognized pattern of data to give a prediction or estimation. The algorithm inspired by biological neural network. A biological neural cell consists of three principle components: Dendrites, Cell Body, and Axon as illustrated in Figure 2.14. The dendrites are receptive nerves that collect electrical signals into the cell body. The cell body is computational unit for these signals then axon carries the output signal from the cell body out to other neurons or make a response to other organs [47].

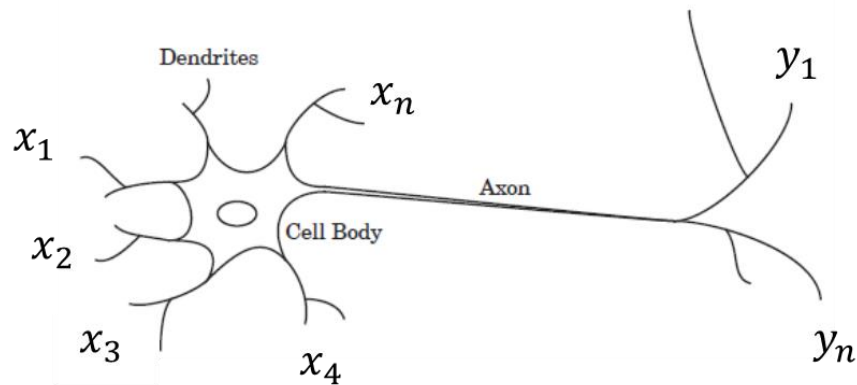


Figure 2.14 Schematic drawing of biological neurons [47]

2.6.1 Network Architecture

The artificial neural network inherits three principle components of biological neural network by these three layers: Input Layer, Hidden Layer, and Output Layer as illustrated in Figure 2.15. Each node in hidden layer similar to the dendrite which collects input signal (x_n) from many sources then nodes in hidden layer compute these input signals with mathematical function. The output layer compute output from hidden layer again to make the last result compatible with desire output (y_n) [47].

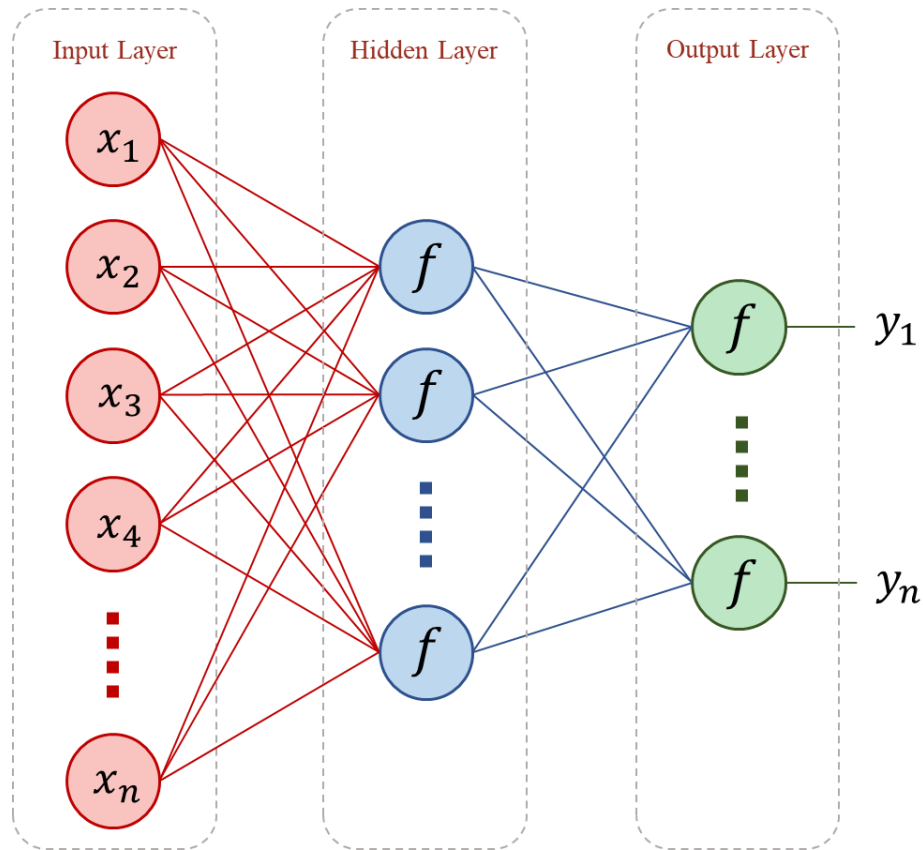


Figure 2.15 Schematic of artificial neurons

The training algorithms work in a similar way to the artillery discipline: “Ready-Fire-Aim”. There is no need to be precise initially: the firing trajectory is continuously adjusted until the artillery shells hit the target. The network parameter adjustment uses the same principle. The input nodes are passive node which contain static values while hidden and output nodes are active node which produce dynamic values with backpropagation algorithm. The backpropagation algorithm similar to trajectory adjustment of fire-control system in artillery.

Figure 2.16 reveal the inside of hidden and output node which consist of two mathematical functions: Input Functions (f_1) and Transfer Functions (f_2). The backpropagation algorithm adjusts the network’s output by feedbacks the differentiation between actual outputs and predicted or estimated outputs to weight (w_n) and bias (b) and recalculate the entire network and keep repeating this procedure until predicted outputs are closest to actual outputs (an iteration of this procedure call epoch). There are many kinds of function as list on Table 2.3 and Table 2.4. and also, many

backpropagation algorithms have been proposed. In order to choose the suitable function and algorithm goal of the task, data type of inputs and desired outputs are needed to be considered.

Even though MATLAB Neural Network toolbox™ provides many instant functions for constructing such a network but in this work we use only four functions: summation function was used as input function, hyperbolic tangent sigmoid function was used as transfer function of hidden layer because our input data have both negative and positive values, linear and softmax functions were used as transfer functions of output layer of regression and classification neural networks respectively. The Levenberg-Marquardt and resilient backpropagation algorithms have been chosen for regression and classification neural networks respectively, details of these two algorithms will be discussed in the next two subchapters and further implementation guides can be found in [48].

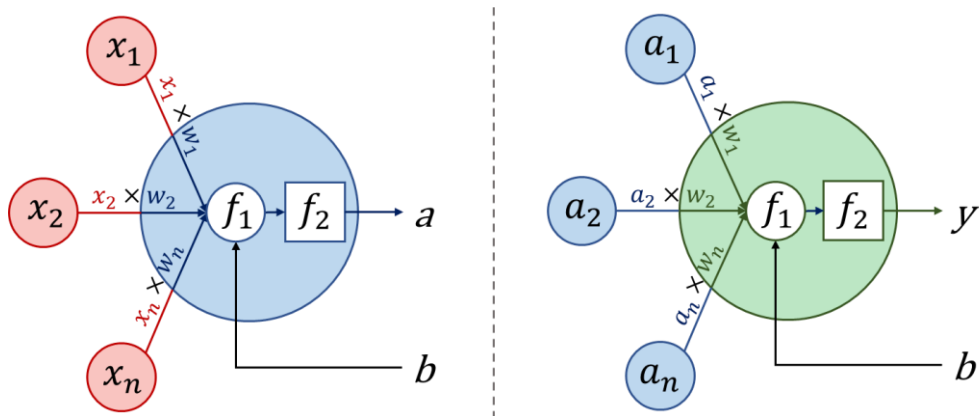








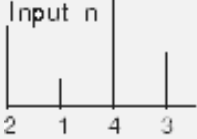
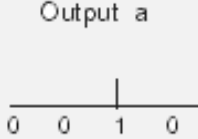

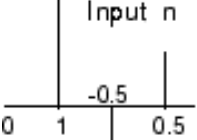
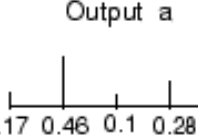



Figure 2.16 Function of hidden node (left) and output node (right)

Table 2.3 List of input functions (f_1)

No.	Name	Icon	MATLAB Function
1	Summation	\oplus	netsum
2	Production	\otimes	netprod

Table 2.4 List of transfer functions (f_2)

No.	Name	Input/Output Relation	Icon	MATLAB Function
1	Hard Limit	$a = 0 \quad n < 0$ $a = 1 \quad n \geq 0$		hardlim
2	Symmetrical Hard Limit	$a = -1 \quad n < 0$ $a = +1 \quad n \geq 0$		hardlims
3	Linear	$a = n$		purelin
4	Saturating Linear	$a = 0 \quad n < 0$ $a = n \quad 0 \leq n \leq 1$ $a = 1 \quad n > 1$		satlin
5	Symmetric Saturating Linear	$a = -1 \quad n < -1$ $a = n \quad -1 \leq n \leq 1$ $a = 1 \quad n > 1$		satlins
6	Log-Sigmoid	$a = \frac{1}{1 + e^{-n}}$		logsig
7	Hyperbolic Tangent Sigmoid	$a = \frac{e^n - e^{-n}}{e^n + e^{-n}}$		tansig
8	Positive Linear	$a = 0 \quad n < 0$ $a = n \quad 0 \leq n$		poslin
9	Competitive	<div style="display: flex; justify-content: space-around;"> <div style="text-align: center;"> <p>Input n</p>  </div> <div style="text-align: center;"> <p>Output a</p>  </div> </div>		compet
10	Softmax	<div style="display: flex; justify-content: space-around;"> <div style="text-align: center;"> <p>Input n</p>  </div> <div style="text-align: center;"> <p>Output a</p>  </div> </div>		softmax

2.6.2 Backpropagation Algorithms

Backpropagation stand for “the backward propagation of errors”. This algorithm used for network adjustment to update weight and bias value by errors. in There are many backpropagation algorithms for regression and classification in MATLAB. We took two algorithms which highly recommend by MATLAB, one for regression approach another one for classification approach.

Levenberg–Marquardt backpropagation algorithm (LM) was chosen for regression approach. The Levenberg–Marquardt algorithm also known as damped least-squares (DLS). This algorithm used for solving nonlinear equations systems or generic curve-fitting problems, first developed by Kenneth Levenberg in 1944 and improved by Donald W. Marquardt in 1963. Implemented in MATLAB by Martin T. Hagan in 1994.

Resilient backpropagation algorithm (Rprop) was chosen for classification approach. This algorithm was developed by Martin Riedmiller and Heinrich Braun in 1992 for solving classification problem. To overcome the inherent disadvantages of the pure gradient descent technique of the original backpropagation algorithm by performs an adaptive weight according to the behavior of the errorfunction. For further instruction and description of these algorithms please see [49–53].

2.7 Summary

Psychological studies have proposed two major approaches to give the definition on each emotion. Some psychologist proposed categorical psychometrics which completely separate each emotional definition, while some other psychologist think dimensional psychometrics is more flexible each emotion definition based on fundamental factors valence and arousal. We introduced Heavner' model and Russell's model as an example of each psychometrics. We adopted Russell's model, but with only eight group of emotion to be our goal for classification.

Commonly, goal of MER task is prediction of music emotion on one of psychometric mentioned above. In order to do so, classification algorithm has been employed to approach categorical psychometrics problem, while regression algorithm has been employed to approach dimensional psychometrics problem. We reviewed 14 work on both approach since 2008, the interested one is using multiple model and hierarchy structure to discriminate only two classes at the time.

We graphical demonstrated five group of acoustic features of i.e. 1.Dynamic (level of loudness) 2.Rhythm (speed and pattern of note or pulse, tempo, beat per minute (BPM)) 3.Timbre (auditory characteristic of each musical instrument) 4.Pitch (level of sound frequency) 5.Tonality (arrangement of pitches and/or chords onto major and minor scales).

We have introduced five acoustic features extraction tools which compatible with MATLAB including MIR toolbox, Timbre toolbox, Tempogram toolbox, Chromagram toolbox, and Essentia. All of these tools are MATLAB library except Essentia which is a standalone application. Each tool is made for different purpose, the interested one is MIR toolbox because this toolbox able to extract all kinds of acoustic features. The only drawback of MIR toolbox is heavy computational resources are need.

We have introduced two neural network algorithms one for regression approach and another one for classification approach. We employed these two algorithms for train the predictive models (details are in the Chapter 3.5).

CHAPTER 3

METHODOLOGY

3.1 System Environment

The system runs on a workstation using a CPU Xeon E3-1270 with 48GB of 2133MHz ECC memory. A Samsung SSD 960 PRO is used for storage rather than an HDD to increase read/write speed to match the performance of the CPU and RAM.

The work was conducted in MATLAB version 2017b using its Music Information Retrieval (MIR) toolbox which was chosen to extract the acoustic features described in Chapter 2.4. The MIR toolbox relies on a built-in auditory toolbox and the Musical Instrument Digital Interface (MIDI) toolbox, which must be installed separately [33–35][37].

3.2 Data Preparation

MediaEval Database for Emotional Analysis in music (DEAM) is a benchmark for a dimensional MER system. This benchmark contains dataset with good quality control over the annotation process [30]. In order to utilize this dataset, first we need to understand how this dataset provides music and annotation data. Then converted into prefer format if it is necessary. The subchapters below described how audio files and annotation files are dealt.

3.2.1 Audio files

The dataset consists of 58 full-length songs, and 1,774 excerpts of 45 seconds length 1,802 songs in total, and the audio files are in stereo MP3 format at an audio CD sampling rate (44.1kHz). Audio files are named by serial number from 2.mp3 to 2058.mp3, some numbers are missing because during developing process the developer cancel some files. Files name 2.mp3 to 2000.mp3 are excerpts and files name 2001.mp3 to 2058.mp3 are full length songs.

3.2.2 Annotations

The emotional annotation files are in spreadsheet (.csv) format. There are two types of annotation static annotation and dynamic annotation. Static annotation

gives a couple value of valence and arousal while dynamic annotation gives VA values second-by-second over the duration of song. In this work, we use dynamic annotation only. The spreadsheet files are given individually (i.e. there is a file for valence of each song and there is a one another file for arousal of each song that mean there are 3,604 spreadsheet files in total to deal with) which each file contain 61 valence or arousal values in range -1 to 1 with fifth decimal depth annotate by 10 people or more. The annotation is in table form each column contains valence or arousal value by the time of song and each row is people who annotated it.

We converted dynamic annotation to static annotation for each song by using average VA-value at each point of time from many annotators, and then find average VA-value along the length of each song. How dynamic annotation is converted to static annotation, is illustrated in Figure 3.1.

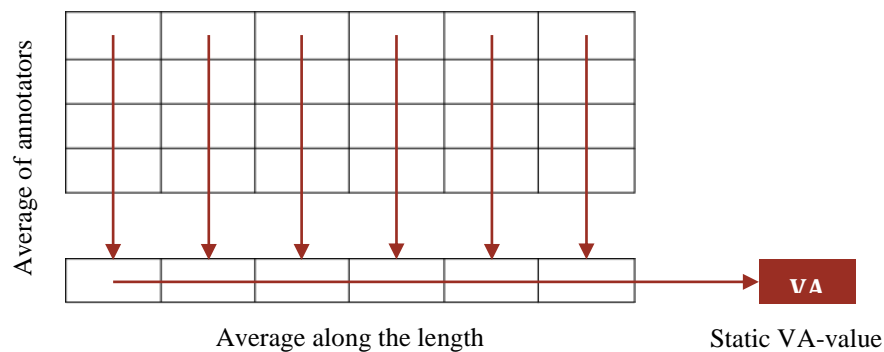


Figure 3.1 Dynamic annotation to static annotation transformation

We can visualize the entire dataset by implementing a scatterplot as shown in Figure 3.2, also we can classify the songs into 8 classes as shown in Table 3.1

Table 3.1 Number of populations in each class

Class No.	Excerpt	Full length	Total
1	588	7	595
2	405	14	419
3	93	6	99
4	95	4	99
5	213	7	220
6	138	4	142
7	70	11	81
8	142	5	147
Σ	1,744	58	1,802

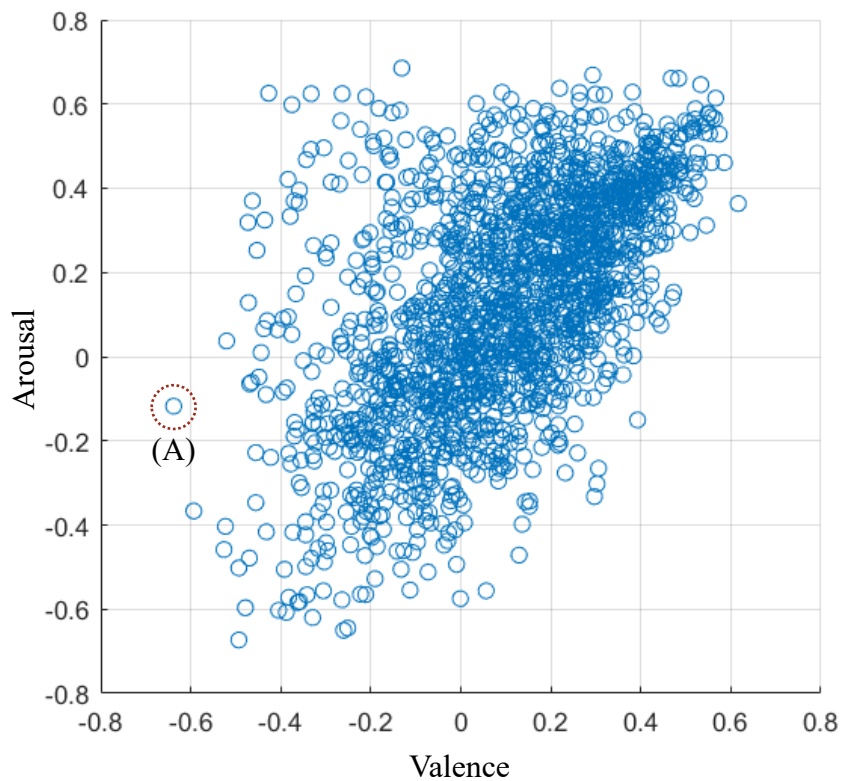


Figure 3.2 The visualization of the dataset on VA plane

Each circle on Figure 3.2 is a song and it is pinpointed by its VA-values. For example, point (A) is file No. 584 name 745.mp3 class No. 6 Sentimental, which

possess -0.64 of valence and -0.12 of arousal. Plotting in this way also shows the density of population of each emotional class.

3.3 Feature Extraction

Since, using multiple tools could be too complex and unnecessary if the final result is acceptable. Therefore, this work utilized only the MIR toolbox because this toolbox covers wide range of features including those features which Timbre, Tempogram, Chromagram toolbox able to extract.

The audio files were processed in terms of signal processing by MIR toolbox to retrieve acoustic features. There were five types of acoustic features and each type has different numbers of elements as shown in Figure 3.3, the number in the bracket is number of elements of each featured type.

In total 122 features were extracted by the 37 functions. Some feature extractors produced errors for some songs, which were ignored by the model training process.

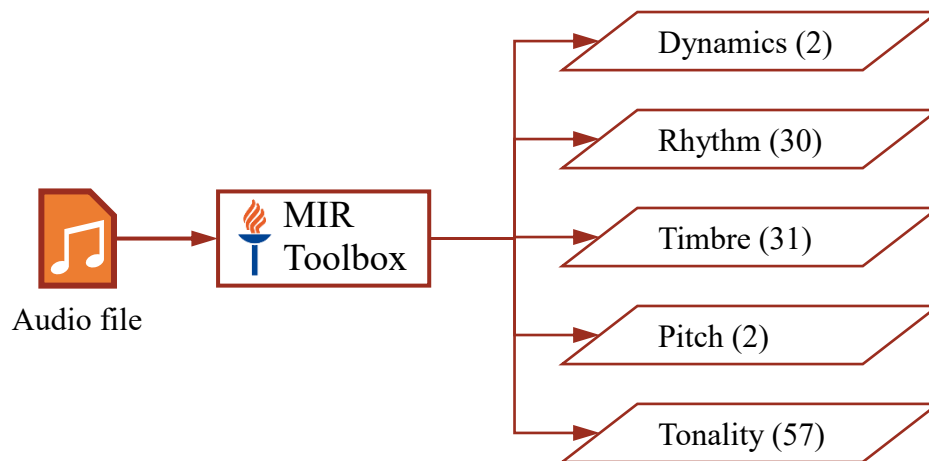


Figure 3.3 The acoustic features extracted by MIR toolbox

Some functions produced a continuous numerical value, some gave a time series, and some generated a discrete value from a finite set. To make the data compatible, we transformed time series and discrete class data into individual numerical values by using the “mirmean” and “mirgetdata” functions. Figure 3.4. shows how to transform raw extracted feature data to numerical data.

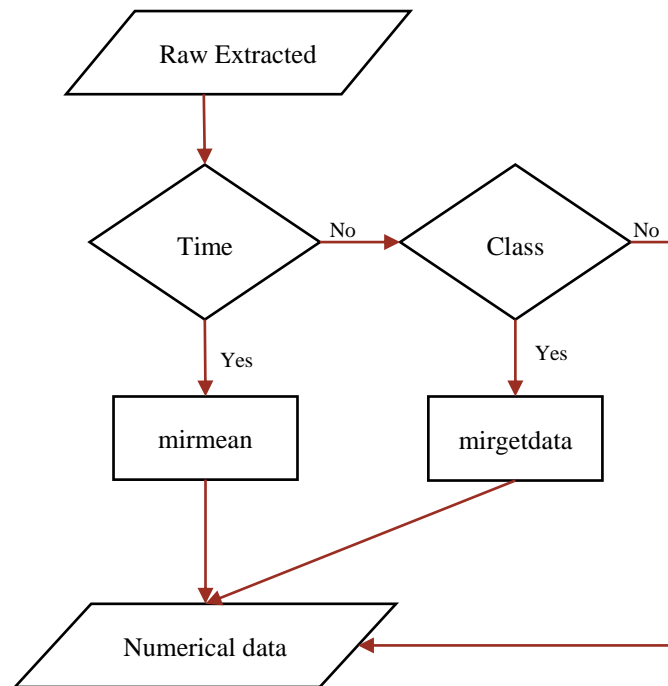


Figure 3.4 Feature extraction post process flow chart

3.4 System Structure

Two different structures of system were built-up and compared to find out which way was better to pursue our goal, between regression approach which classify VA-response to class at the output of predictive model and classification approach which classify VA-response to class at the input of predictive model.

The previous work on DEAM dataset got low accuracy because using limited number of features and SVM, Naïve Bayes, Decision Trees, and k-NN algorithm did not work well on this dataset though [21]. Therefore, we chose the LM algorithm for handling the regression approach and the Rprop algorithm for dealing with the classification approach.

3.4.1 Regression Approach

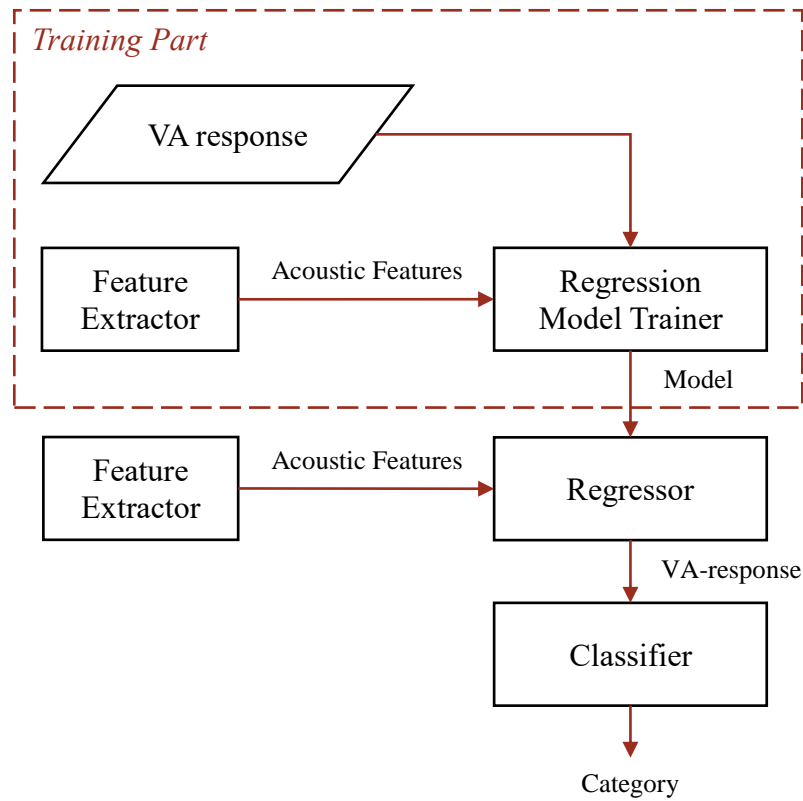


Figure 3.5 Framework for regression approach system

The dataset provided annotations in the form of valence and arousal responses. The regression approach can use these annotations directly, Figure 3.5 shows the framework of the system. The LM algorithm was employed at the regression model trainer module. The regressor module produced estimated VA-responses. In order to determine these predictions, correct or wrong. The acceptable areas of error were setting up at the classifier module, as can be seen in Figure 3.6.

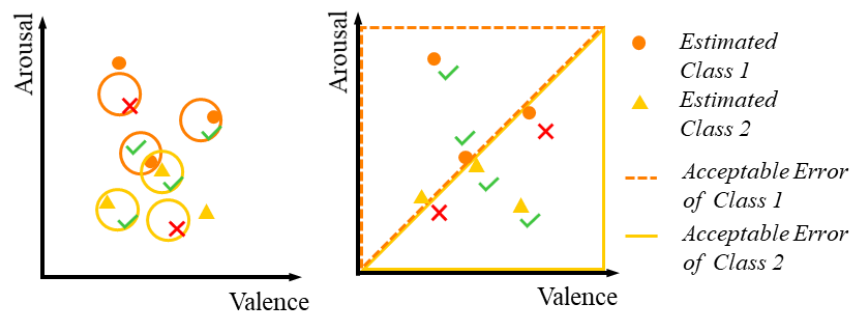


Figure 3.6 Circular acceptable areas of error (left), triangular acceptable areas of error (right)

3.4.2 Classification Approach

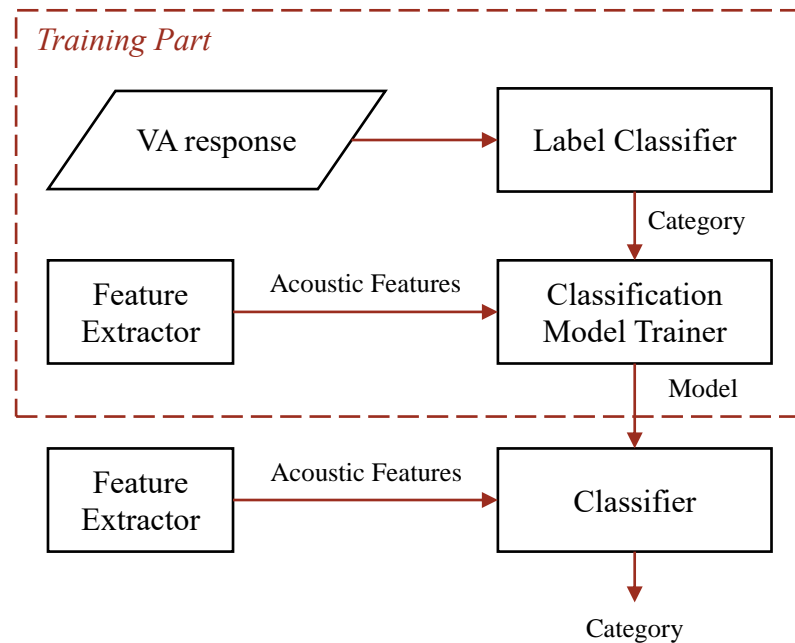


Figure 3.7 Framework for classification approach system

Figure 3.7 shows the framework of the system. The labeler module converts the VA response to classes before they were used as annotation in the model trainer module because raw annotations cannot be used directly, then Rprop algorithm was employed at the classification model trainer module.

3.5 Model Structure

The predictive model in each system was customized to have two different structures (traditional and cascaded) to determine which way is better to train the model.

3.5.1 Traditional Multiclass Model

The LM and Rprop algorithms in the model trainer module were trained with 122 acoustic features of 1,802 song to predict music emotion, but the outputs of the model were different as illustrated in Figure 3.8.

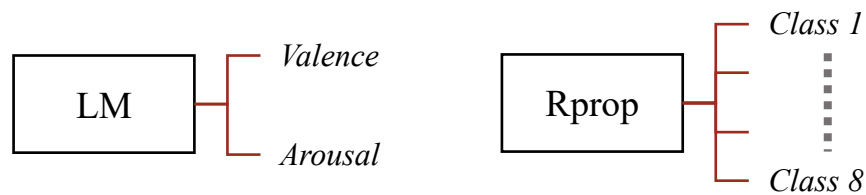


Figure 3.8 Traditional structure of predictive model, LM algorithm (left) & Rprop algorithm (right)

3.5.2 Cascaded Model

The cascaded model was obtained by connecting several traditional multiclass models as a cascaded structure, as shown in Figure 3.9. Each unit was specifically trained to discriminate only two classes as reported in Table 3.2.

Previous works that utilized a similar model structure demonstrated that the accuracy was better when discrimination started with arousal than when discrimination started with valence [2][3][6][8]. Additionally, many regression approach studies have shown that arousal prediction is always more accurate than valence prediction [6][7][15–18]. Therefore, we initiated unit 1 to discriminate between high- and low-arousal songs by training the model with the entire dataset. Unit 2 was trained with only high-arousal songs to discriminate positive valence songs from negative valence songs among those high-arousal songs that were predicted by unit 1, and so on. (Quadrants 1 to 4 and classes 1 to 8 refer to the quadrant of the VA plane and emotional octant in Figure 2.1; see Table 2.2 for the definition of each class.)

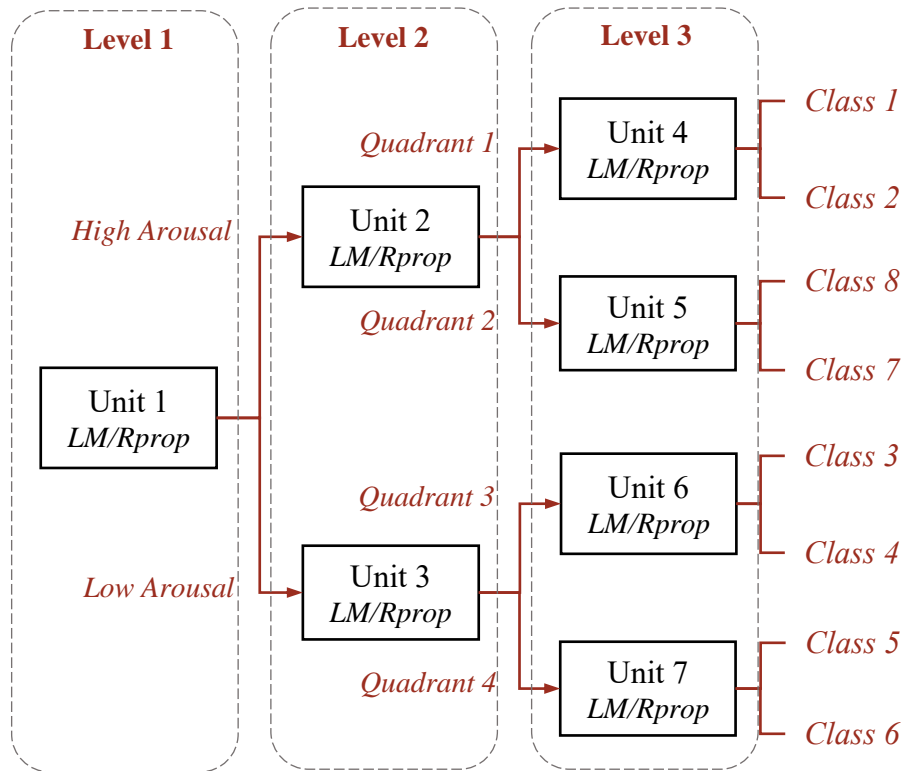


Figure 3.9 Schematic cascaded structure diagram of multi-model neural network

Table 3.2 Training dataset and purpose of each unit in Figure 3.9

Unit	Trained with	to Discriminate	Number of Training Sample
1	Entire Dataset	High & Low Arousal	1,802
2	High Arousal	Quadrant 1 & 2	1,242
3	Low Arousal	Quadrant 3 & 4	560
4	Quadrant 1	Class 1 & 2	1,014
5	Quadrant 2	Class 7 & 8	228
6	Quadrant 3	Class 3 & 4	198
7	Quadrant 4	Class 5 & 6	362

3.6 Model Training

Before feed in training algorithm some preprocessing such as standardization, cleansing defect or unrelated data, and dividing data are needed. The neural network toolbox provides these functions, we can instantly use these functions as describe in first subchapter below. During the training process several parameters of neural network must be adjusted for several time until by observed the result.

3.6.1 Preprocessing

All the neural networks use pre and post processing described in Figure 3.10, to standardize the data and reduce the computational power requirements. Constant rows are removed when there are unchanged elements in the data, because there is no need to count them in the calculations. The Min-Max function standardizes the data by equalizing the ranges of all the data to between -1 and 1. The “Fix Unknowns” function is applied to incomplete data by adding elements. Principle Component Analysis (PCA) finds out which input elements have a major effect or influence upon the output, and then applies some bias to those elements.

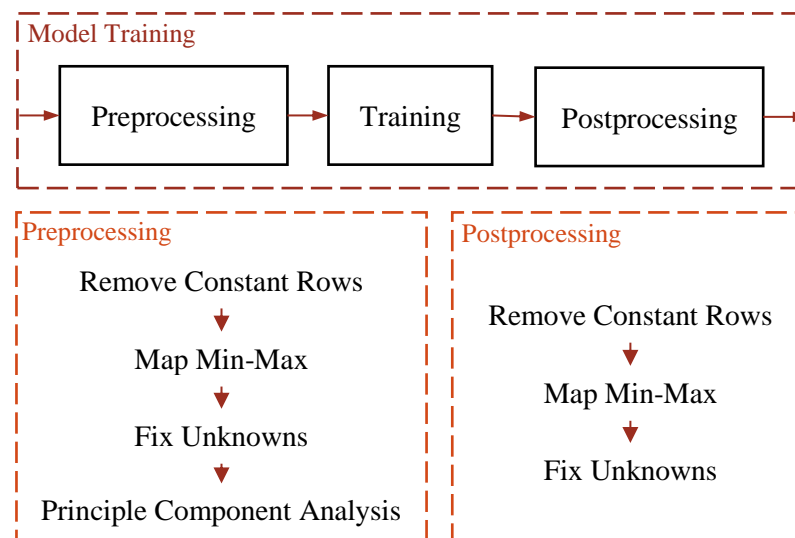


Figure 3.10 Pre/Post processing of input/output data

The dataset was divided into three parts to evaluate the performance of the models: 15% of the data was randomly selected as the validation, another 15% was assigned to the testing set, and the rest was used as the training set. The networks were trained with the training set and then tested with the validation and testing sets. The result on the validation set was used to update the weight parameter in the next epoch (a completed iteration of the training procedure) to shift the accuracy closer to perfect accuracy. The result on the testing set was completely independent from the training process. Accuracy on the training set that is much higher than that on the validation set and/or the testing set indicates overfitting, and we need to reconfigure the model.

3.6.2 Parameter Configuration

Basically, neural networks consist of layers: Input layer, Hidden layer, and output layer. Input layer is passive node, while hidden and output layer are active node always update until one of the stopping criteria has been reach. Figure 3.12 shows the actual structure of LM neural network including require number node in each layer and type of transfer function in each active node.

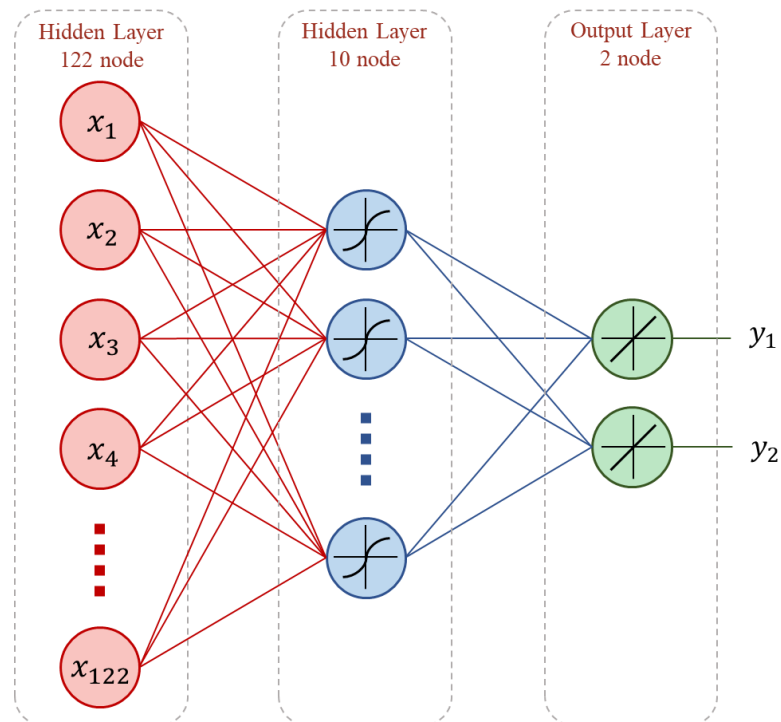


Figure 3.11 Schematic of artificial neurons using in this study

It is inconvenient to display the actual structure of neural network if there are large number of nodes so MATLAB represent this structures by simple block diagram as shown in Figure 3.12-14.

The traditional multiclass and each unit of cascaded LM models were constructed based on the diagram show in Figure 3.12. There are 122 feature inputs, with 10 hidden layers by default. Each hidden layer is a summation of weight (w) and bias (b), using a tangent sigmoid transfer function. Each output layer has the same structure as the hidden layer except that the transfer function is linear. The network estimates the values for valence and arousal as two outputs.

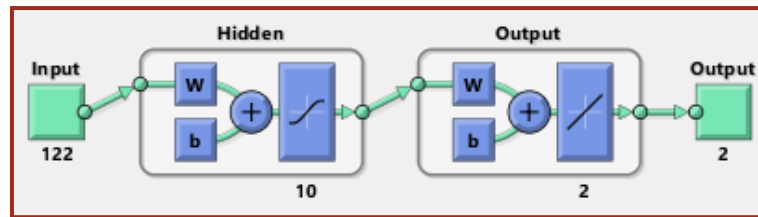


Figure 3.12 Structure of traditional multiclass and each unit of cascade LM neural network

The classification neural network employs a softmax transfer function in the output layer. This transfer function assigns probabilities to the eight outputs, and the output with the largest probability is taken as the predicted class. The structure is shown in Figure 3.13.

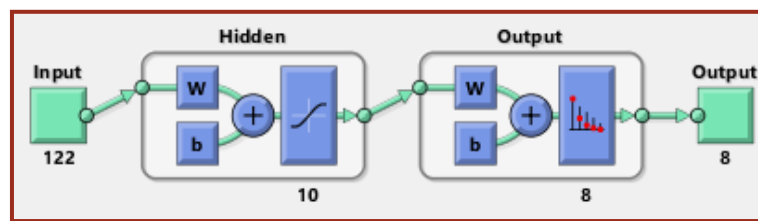


Figure 3.13 Structure of traditional multiclass Rprop neural network

The structure of the cascade classification network is similar to the traditional multiclass model, but the number of possible outputs is only two as shown in Figure 3.14. The networks are trained separately in order to discriminate only two classes, then the networks are composed using the same cascade structure as the cascaded LM (see Figure 3.9), but with the core algorithm changed to Rprop as shown in Figure 3.14.

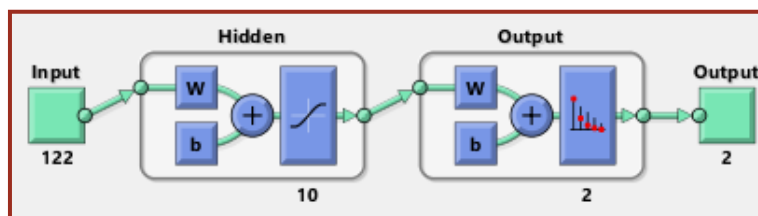


Figure 3.14 Structure of each unit of cascaded Rprop neural network

Various algorithms have different parameters that commonly include maximum number of epochs, elapsed time, acceptable error rate, minimum gradient

(convergent slope of training, validation, and testing subdatasets), and maximum failing (no improvement in accuracy). We attempted to maximize the accuracy, so we defined the acceptable error rate as zero and allowed the maximum number of epochs, and the elapsed time to be infinite. We concentrated on the minimum gradient and maximum failing adjustment as stopping criteria for the training process and left the other parameters at their default values (Maximum Failing or Validation Checks: number of epoch pass when performance no longer improve). The models in the same depth of cascaded structures were constructed with the parameter values reported in Table 3.3. The list of parameters and their default values can be found in [54].

Table 3.3 The parameters configuration

Model	Architecture	Min Gradient	Max Failing	Learning Rate
Traditional Multiclass LM	122-10-2	1.0E-07	10	1.0E-02
Level 1 of Cascaded LM	122-10-2	1.0E-07	10	1.0E-02
Level 2 of Cascaded LM	122-10-2	1.0E-07	10	1.0E-02
Level 3 of Cascaded LM	122-10-2	1.0E-21	10	1.0E-02
Traditional Multiclass Rprop	122-10-8	1.0E-50	300	1.0E-02
Level 1 of Cascaded Rprop	122-10-2	1.0E-50	300	1.0E-02
Level 2 of Cascaded Rprop	122-10-2	1.0E-100	300	1.0E-02
Level 3 of Cascaded Rprop	122-10-2	1.0E-100	300	1.0E-02

The number in the architecture column refers to the number of nodes in the neural network layer structure: “Input node - Hidden node – Output node”. The number of input nodes is the number of extracted features (122 features for all models). The number of hidden nodes remained 10 by default, and the number of output nodes of each model was set based on the number of desired outputs. The desired output of the LM algorithm was the estimated valence and arousal, so the number of desired outputs was 2. The desired output of the Rprop algorithm was the specific classes, differed between the traditional multiclass model and the cascaded model. The desired number of output classes of the traditional multiclass model was 8, while that of the cascaded model was 2, as previously noted. The reported values in the min gradient and max failing columns were the results of trial and error by observing the relationship

between changes in parameter values and accuracy: the values that gave the highest accuracy were selected.

We observe the mean squared error (MSE) of these three subdatasets, the lower the MSE the higher the accuracy. The value of the MSE on these three subdatasets should be almost equal. If the MSE of the training set is low, but the validation set and/or the testing set is high, then an overfitting problem has occurred, and we need to reconfigure the trial network. The network is trained with the training set, and then the network is tested with the validation and testing sets. The results for the validation set are used to update the weight in the next epoch to bring it closer to the target. This cycle repeats until one of the stopping criteria is reached. Figure 3.15 shows that the sixth epoch is the best-fitting; before sixth epoch there is underfitting and beyond sixth epoch there is overfitting. Figure 3.16 shows stopping criteria approaching during the training process.

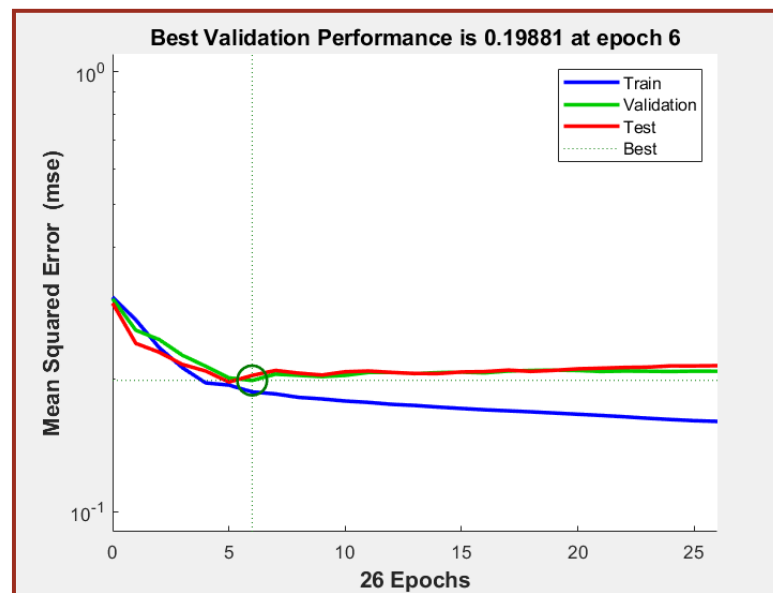


Figure 3.15 Mean Squared Error (MSE) display

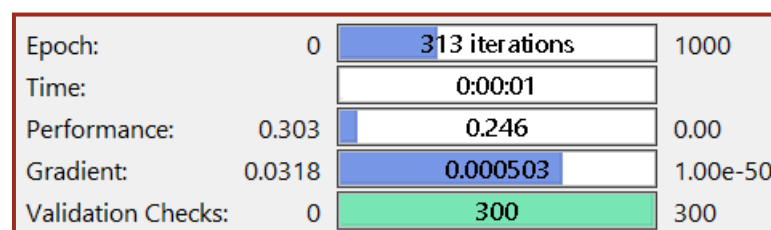


Figure 3.16 Stopping criteria display

The accuracies always vary slightly when the network is retrained because of the variation in the randomly selected parameters such as weight, bias, and sample selection of subdatasets, so the network must be retrained several times to reduce the variation in the results. The accuracy can be improved by approximately 5% by selecting the best result from 11,000 rounds of training.

3.7 Summary

We started this chapter with detailed information about dataset and how we used it. The most audio files in the dataset are 45 second excerpts from songs with each song annotated with valence and arousal levels along the song length by multiple annotators. We collapsed dynamic annotation to static annotation then converted these VA-value to eight emotional classes.

The acoustic features are extracted from audio files by 37 functions of MIR toolbox. We got 122 acoustic features as a result, but these features are divers. Some features data are time series data, some are discrete class, or continuous numerical data. Thus, we standardized these data by transforming them to numerical data in range of -1 to 1. We can see the relationship of VA-value and feature value by plot feature value of all sample against VA-plane.

We implemented model training by four methods i.e. 1) train with a traditional multiclass unit of Levenberg-Marquardt algorithm. 2) train with multiple units with cascaded structure of Levenberg-Marquardt algorithm. 3) train with a traditional multiclass unit of resilience algorithm. 4) train with multiple units with cascaded structure of resilience algorithm. The entire process is shown in Figure 3.17.

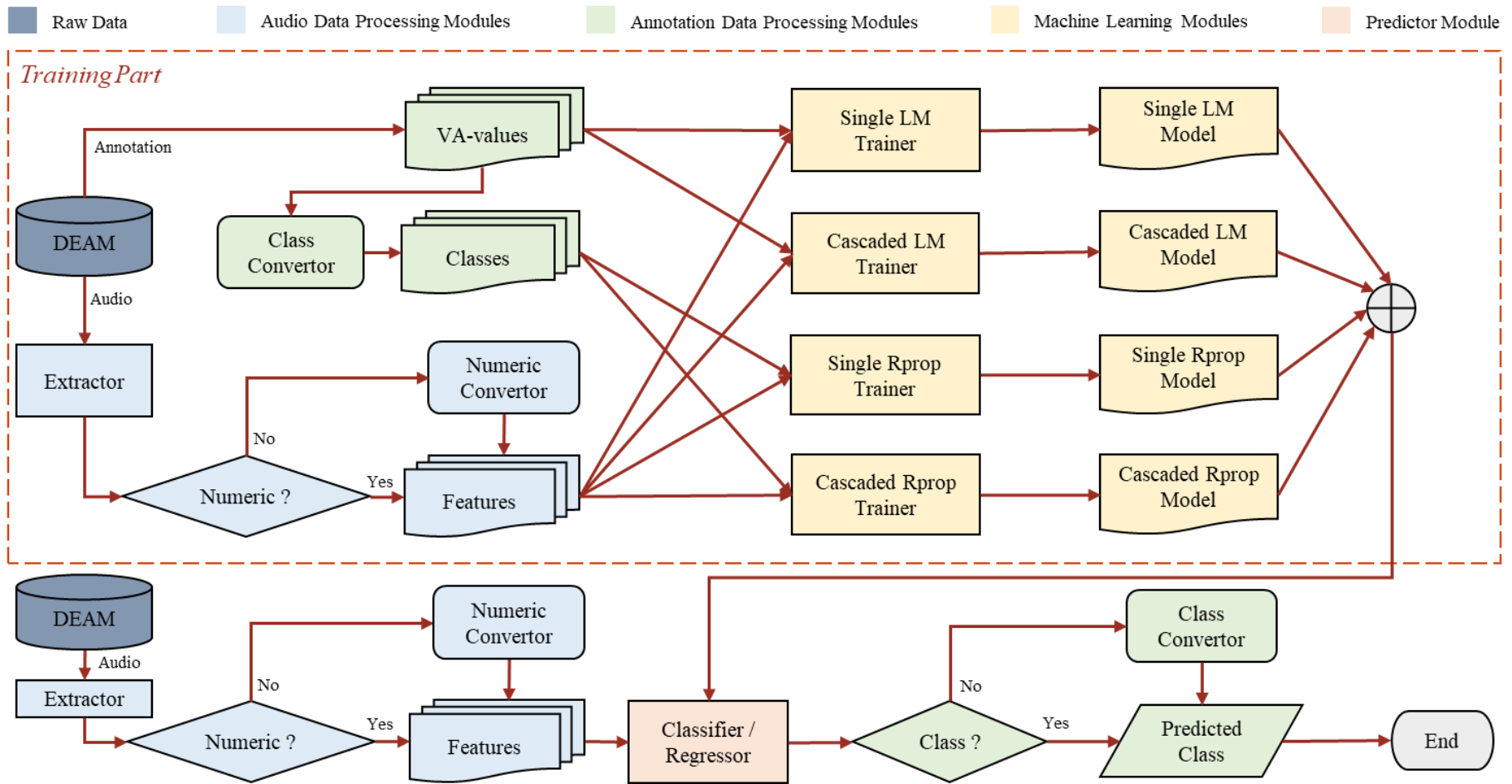


Figure 3.17 The entire system flow chart

CHAPTER 4

RESULTS AND DISCUSSIONS

4.1 Feature Extraction

The functions of MIR toolbox used for feature extraction, and their running time over the entire dataset are given in Table 4.1. Most of the feature extractions were completed in a few hours, but some took more than a day.

Table 4.1 List of MIR toolbox's function for acoustic feature extraction

#	Featured Type	Function Name	Output Data Type	Featured Element	Featured Number	Elapsed Time (h)
1	Dynamics	mirrms	Time Series	1	1	0.89
2	Dynamics	mirlowenergy	Time Series	1	2	0.92
3	Rhythm	mirfluctuation	Time Series	24	3-26	0.93
4	Rhythm	mirbeatspectrum	Time Series	1	27	4.82
5	Rhythm	mirerevents	Time Series	1	28	3.46
6	Rhythm	mirereventdensity	Time Series	1	29	4.00
7	Rhythm	mirtempo	Numeric	1	30	3.62
8	Rhythm	mirmetroid	Time Series	1	31	5.85
9	Rhythm	mirpulseclarity	Time Series	1	32	3.74
10	Timbre	mirattacktime	Time Series	1	33	33.51
11	Timbre	mirattackslope	Time Series	1	34	35.44
12	Timbre	mirattackleap	Time Series	1	35	33.86
13	Timbre	mirdecaytime	Time Series	1	36	34.72
14	Timbre	mirdecayleap	Time Series	1	37	34.03
15	Timbre	mirdecayslope	Time Series	1	38	33.67
16	Timbre	mirduration	Time Series	1	39	33.73
17	Timbre	mirzerocross	Time Series	1	40	0.97
18	Timbre	mirrolloff	Time Series	1	41	1.33
19	Timbre	mirbrightness	Time Series	1	42	1.08
20	Timbre	mircentroid	Numeric	1	43	1.12
21	Timbre	mirspread	Numeric	1	44	1.39
22	Timbre	mirskewness	Numeric	1	45	1.52
23	Timbre	mirkurtosis	Numeric	1	46	1.51

Table 4.1 (Continued) List of MIR toolbox's function for acoustic feature extraction

#	Featured Type	Function Name	Output Data Type	Featured Element	Featured Number	Elapsed Time (h)
24	Timbre	mirflatness	Time Series	1	47	1.22
25	Timbre	mirentropy	Time Series	1	48	1.01
26	Timbre	mirmfcc	Time Series	13	49-61	1.32
27	Timbre	mirroughness	Time Series	1	62	2.08
28	Timbre	mirregularity	Time Series	1	63	325.32
29	Pitch	mirpitch	Numeric	1	64	1.51
30	Pitch	mirmidi	Time Series	1	65	12.10
31	Tonality	mirchromagram	Time Series	12	66-77	1.02
32	Tonality	mirkeystrength	Time Series	12	78-89	1.02
33	Tonality	mirkey	Classes	1	90	1.40
34	Tonality	mirmode	Time Series	1	91	1.18
35	Tonality	mirkeysom	Time Series	24	92-115	1.01
36	Tonality	mirtonalcentroid	Time Series	6	116-121	1.01
37	Tonality	mirhcdf	Time Series	1	122	3.58

In total, 122 features were extracted by the 37 functions. All features extracted by MIR toolbox version 1.7 except function No. 30 extracted by the toolbox version 1.6 because in version 1.7 this function cause error. The tool is not perfect, for some extractor functions the error occurs for some song, so we employed error exception by give “NaN” as a result of the error process. In model training process, training algorithms will ignore “NaN” value by default.

4.2 Feature Correlation

We measured linear correlation between features and fundamental factors of emotion by Pearson correlation coefficient which is defined as (1).

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}} \quad (1)$$

We measured the linear correlation between features and the fundamental factors of emotion using a linear correlation coefficient. The correlations

of each feature with valence and arousal were measured separately, as shown in Figure 4.1.

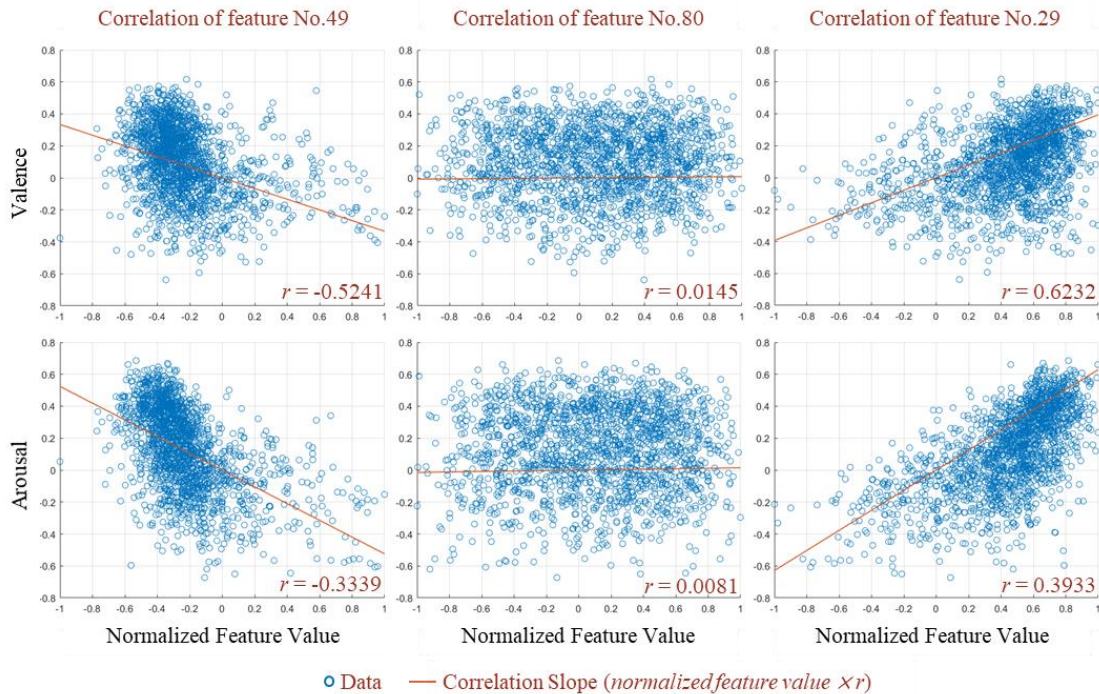


Figure 4.1 Scatter plot of feature No. 49, 80, and 29 against valence and arousal

A correlation coefficient (r) of 1 or -1 indicates, a perfect linear relationship, and r equal to 0 indicates the absence of a linear relationship. The most important features for prediction are those with r values close to 1 or -1. The features with r values close to 0 are still useful (unless r value is exactly 0) but have a lesser impact [55]. The correlation coefficient values of all features can be found in Table A in Appendix.

We ranked the feature correlations in ascending order to identify the impactful features; we present the partial correlation ranking in Table 4.2. The correlations of features with valence and arousal were ranked separately, as shown in the “Rank by V” and “Rank by A” columns.

Some features, such as feature No. 62, have a weak effect on valence but a strong effect on arousal. The most impactful feature for both valence and arousal is feature No. 29. Furthermore, most of the impactful features are timbre features.

On the basis of these correlation rankings, in scenarios with time or resource limitations we can select only impactful features to train the model rather than considering all the features, but this process may reduce the accuracy. However, the optimized model in the case of limited time and resources is not the focus of this work, so we included all 122 acoustic features as inputs for model training.

Table 4.2 Correlation values between extracted features and VA-ratings

Feature No.	Valence r	Arousal r	Rank by V	Rank by A	Correlation status
49	-0.3339	-0.5241	1	1	Negative Correlation
45	-0.3262	-0.4497	2	2	
39	-0.2882	-0.2463	3	7	
36	-0.2410	-0.3667	5	3	
62	0.0065	0.2300	60	108	Poor Correlation
80	0.0081	0.0145	61	62	
90	0.0095	0.0294	62	65	
112	0.0015	0.0088	57	60	
99	-0.0206	0.0095	45	61	Positive Correlation
48	0.3933	0.6232	120	122	
29	0.4196	0.6286	121	121	
32	0.4266	0.3351	122	113	
42	0.3577	0.5828	116	120	

4.3 Modeling Results

The model performances are present with tables, and the accuracies are explained with confusion matrices. The tables give the accuracy for training, validation, testing subdataset, and the entire dataset. They also include elapsed time for training each model, captured during the definite network training process, and show the round which the modeling performs the best.

While the confusion matrices show the detail of population in each class. Each cell shows the density of the population in terms of number and percentage. The summation of all the cell numbers, should equal the sample size of the dataset and the percentage summation should be 100%. The summation of the numbers in the vertical

cells is the total number of samples in each class. Summations of the horizontal cells are the total number of predicted classes. Diagonal cells from the top left to bottom right show the corrected prediction for each class (true positive). The accuracy (ACC) is shown on the bottom right corner of the matrices. The bottom of the matrices shows the True Positive Rate (TPR), sometime called sensitivity or recall. The right of the matrices shows the Positive Predictive Value (PPV), sometime called precision. The definition of ACC, TPR, and PPV are given by (2-4).

$$PPV = \frac{\sum True\ Positive}{\sum Predicted\ Condition\ Positive} \quad (2)$$

$$TPR = \frac{\sum True\ Positive}{\sum Actual\ Condition\ Positive} \quad (3)$$

$$ACC = \frac{\sum True\ Positive + \sum True\ Negative}{\sum Total\ Population} \quad (4)$$

Measuring performance by ACC alone may not give the details how the samples were incorrectly classified. Therefore, TPR and PPV were used to fulfill this weakness by showing how the model fail. The error of TPR (red number) indicates failure by omitting to count the samples from an actual class into a predicted class on the other hand the error of PPV indicates failure by incorrectly counting the samples in into a predicted class. The error of TPR and/or PPV could be high in some class if the models were train by imbalance class training data. To measure harmonic average of the TPR and PPV F1-score measurement was employed in every model and separately in each class. The definition of F1-score (F) are given by (5), number of class denoted by the symbol nC . The best value of F1-score is 1 (perfect TPR and PPV) and the worst is 0.

$$F = nC \cdot \frac{PPV \cdot TPR}{PPV + TPR} \quad (5)$$

Chapter 4.3.1 to 4.3.6 contain the results of the LM network, Chapter 4.3.1 and 4.3.2 carrying out binary classification of valence and arousal: class 1 of valence binary classification is labeled as positive valence and class 2 is negative valence on the other hand class 1 of arousal binary classification labeled as high arousal

and class 2 is low arousal. Chapter 4.3.3 and 4.3.4 carrying out four and eight emotional classification results with a traditional multiclass model. The class labels 1 to 4 of four emotional classification refer to emotional four quadrant as shown in Figure 2.1: i.e. class 1 is high arousal and positive valence, class 2 is high arousal but negative valence, class 3 is low arousal and negative valence, and class 4 is low arousal but positive valence. The class labels 1 to 8 of eight emotional classification refer to emotional octant as shown in Figure 2.1. Chapter 4.3.5 and 4.3.6 carrying out four and eight emotional classification results as well but with the cascaded structure of multiple model. The presentation order of Chapter 4.3.7 to 4.3.12 repeat the order of the first six subchapter but change the algorithm to Rprop network.

4.3.1 Binary-Class LM Network for Valence Prediction

The simplest task is binary classification of fundamental factors of emotion, firstly valence level is discriminated by traditional multiclass LM neural network to discriminate positive and negative valence. The modeling performance is shown in Table 4.3, and the classification result is shown in Figure 4.2.

Table 4.3 The performance of binary-class LM networks for valence discrimination

Model	Train (%)	Val. (%)	Test (%)	Entire (%)	F1-score	Best Epoch	Best Round	Time (h)
Traditional multiclass LM for Valence	77.5	80.7	83.7	78.9	0.8475	5	2	7.02

Predicted class	1	1056 58.6%	224 12.4%	82.5% 17.5%
	2	156 8.7%	366 20.3%	70.1% 29.9%
		87.1% 12.9%	62.0% 38.0%	78.9% 21.1%
		1	2	
		Actual class		

Figure 4.2 The confusion matrix of binary-class LM networks for valence discrimination

4.3.2 Binary-Class LM Network for Arousal Prediction

Another fundamental factor of emotion is arousal. These are result of arousal discrimination by LM neural network to discriminate high and low arousal. Table 4.4 shows that all subdataset gained a little bit accuracy when the evaluate in term of arousal, the detailed accuracy of entire dataset shows in Figure 4.3.

Table 4.4 The performance of binary-class LM networks for arousal discrimination

Model	Train (%)	Val. (%)	Test (%)	Entire (%)	F1-score	Best Epoch	Best Round	Time (h)
Traditional Multiclass LM for Arousal	85.9	86.7	84.8	85.8	0.8980	5	2	7.02

Predicted class	1	1122 62.3%	135 7.5%	89.3% 10.7%
	2	120 6.7%	425 23.6%	78.0% 22.0%
		90.3% 9.7%	75.9% 24.1%	85.8% 14.2%
		1	2	
		Actual class		

Figure 4.3 The confusion matrix of binary-class LM networks for arousal discrimination

4.3.3 Traditional Multiclass LM Network for Four Emotions Prediction

The four emotional classification is a combination of fundamental factors valence and arousal. The modeling performance is shown in Table 4.5, and the classification result is shown in Figure 4.4.

Table 4.5 The performance of traditional multiclass LM networks for four emotions prediction

Model	Train (%)	Val. (%)	Test (%)	Entire (%)	F1-score	Best Epoch	Best Round	Time (h)
Traditional Multiclass LM for 4 Emotion	70.6	70.7	70.4	70.6	See Table 4.15	5	2	7.02

Predicted class	1	910 50.5%	82 4.6%	51 2.8%	132 7.3%	77.4% 22.6%
	2	37 2.1%	59 3.3%	46 2.6%	9 0.5%	39.1% 60.9%
	3	20 1.1%	51 2.8%	238 13.2%	22 1.2%	71.9% 28.1%
	4	47 2.6%	6 0.3%	27 1.5%	65 3.6%	44.8% 55.2%
		89.7% 10.3%	29.8% 70.2%	65.7% 34.3%	28.5% 71.5%	70.6% 29.4%
		1	2	3	4	
		Actual class				

Figure 4.4 The confusion matrix of traditional multiclass LM networks for four emotions prediction

4.3.4 Traditional Multiclass LM Network for Eight Emotions Prediction

The eight emotional classification is a combination of fundamental factors valence and arousal with a more specific emotion. The modeling performance is shown in Table 4.6, and the classification result is shown in Figure 4.5. This result and the result in previous subchapter show that the more classes to predict, the lesser accuracy we might got.

Table 4.6 The performance of traditional multiclass LM networks for eight emotions prediction

Model	Train (%)	Val. (%)	Test (%)	Entire (%)	F1-score	Best Epoch	Best Round	Time (h)
Traditional Multiclass LM for 8 Emotion	50.6	42.2	40.0	47.8	See Table 4.16	5	2	7.02

Predicted class	1	421 23.4%	174 9.7%	20 1.1%	3 0.2%	7 0.4%	13 0.7%	23 1.3%	83 4.6%	56.6% 43.4%
	2	125 6.9%	190 10.5%	33 1.8%	26 1.4%	23 1.3%	8 0.4%	11 0.6%	15 0.8%	44.1% 55.9%
	3	7 0.4%	23 1.3%	20 1.1%	12 0.7%	16 0.9%	8 0.4%	2 0.1%	3 0.2%	22.0% 78.0%
	4	1 0.1%	6 0.3%	8 0.4%	19 1.1%	19 1.1%	3 0.2%	3 0.2%	1 0.1%	31.7% 68.3%
	5	4 0.2%	6 0.3%	11 0.6%	27 1.5%	115 6.4%	43 2.4%	1 0.1%	2 0.1%	55.0% 45.0%
	6	5 0.3%	5 0.3%	4 0.2%	9 0.5%	31 1.7%	49 2.7%	14 0.8%	5 0.3%	40.2% 59.8%
	7	12 0.7%	6 0.3%	3 0.2%	1 0.1%	4 0.2%	11 0.6%	16 0.9%	7 0.4%	26.7% 73.3%
	8	20 1.1%	9 0.5%	0 0.0%	2 0.1%	5 0.3%	7 0.4%	11 0.6%	31 1.7%	36.5% 63.5%
		70.8% 29.2%	45.3% 54.7%	20.2% 79.8%	19.2% 80.8%	52.3% 47.7%	34.5% 65.5%	19.8% 80.2%	21.1% 78.9%	47.8% 52.2%
		1	2	3	4	5	6	7	8	
		Actual class								

Figure 4.5 The confusion matrix of traditional multiclass LM networks for eight emotions prediction

4.3.5 Cascaded LM Network for Four Emotions Prediction

Instead of using only one network to predict all four emotional classes, the cascaded LM network for four emotions consist of three LM network to took advantage of binary classification of arousal (see Chapter 4.3.2). The modeling performance is shown in Table 4.7, and the classification result is shown in Figure 4.6.

Table 4.7 The performance of cascaded LM networks for four emotions prediction

Model	Train (%)	Val. (%)	Test (%)	Entire (%)	F1-score	Best Epoch	Best Round	Time (h)
Cascaded LM Unit 1	85.9	86.7	84.8	85.8	0.8980	5	2	7.0
Cascaded LM Unit 2	83.9	81.7	84.9	83.7	0.9055	1	8	6.1
Cascaded LM Unit 3	76.8	81.0	81.0	78.0	0.6886	1	27	4.1

Predicted class	1	904 50.2%	61 3.4%	43 2.4%	120 6.7%	80.1% 19.9%
	2	37 2.1%	88 4.9%	36 2.0%	10 0.6%	51.5% 48.5%
	3	42 2.3%	44 2.4%	268 14.9%	36 2.0%	68.7% 31.3%
	4	31 1.7%	5 0.3%	15 0.8%	62 3.4%	54.9% 45.1%
		89.2% 10.8%	44.4% 55.6%	74.0% 26.0%	27.2% 72.8%	73.4% 26.6%
	1	2	3	4	Actual class	

Figure 4.6 The confusion matrix of cascaded LM networks for four emotions prediction

4.3.6 Cascaded LM Network for Eight Emotions Prediction

Instead of using only one network to predict all eight emotional classes, the cascaded LM network for eight emotions which is developed from cascaded LM network for four emotions. The model consists of seven LM network to took advantage of binary classification of arousal and valence. Nonetheless this result and the result of previous subchapter show that the overall accuracy of both cascaded LM higher than traditional multiclass LM only a little as shown in Table 4.8 and Figure 4.7.

Table 4.8 The performance of cascaded LM networks for eight emotions prediction

Model	Train (%)	Val. (%)	Test (%)	Entire (%)	F1-score	Best Epoch	Best Round	Time (h)
Cascaded LM Unit 1	85.9	86.7	84.8	85.8	0.8980	5	2	7.0
Cascaded LM Unit 2	83.9	81.7	84.9	83.7	0.9055	1	8	6.1
Cascaded LM Unit 3	76.8	81.0	81.0	78.0	0.6886	1	27	4.1
Cascaded LM Unit 4	69.9	67.1	64.5	68.6	0.7531	1	2,018	4.8
Cascaded LM Unit 5	78.8	79.4	64.7	76.8	0.6667	1	946	2.9
Cascaded LM Unit 6	79.0	63.3	70.0	75.3	0.7656	1	1,084	2.0
Cascaded LM Unit 7	71.7	59.3	74.1	70.2	0.7882	1	24	3.8

Predicted class	1	451 25.0%	180 10.0%	18 1.0%	12 0.7%	11 0.6%	9 0.5%	19 1.1%	84 4.7%	57.5% 42.5%
	2	93 5.2%	176 9.8%	32 1.8%	7 0.4%	15 0.8%	9 0.5%	14 0.8%	15 0.8%	48.8% 51.2%
	3	5 0.3%	12 0.7%	27 1.5%	12 0.7%	10 0.6%	3 0.2%	4 0.2%	3 0.2%	35.5% 64.5%
	4	4 0.2%	8 0.4%	8 0.4%	25 1.4%	14 0.8%	5 0.3%	3 0.2%	2 0.1%	36.2% 63.8%
	5	12 0.7%	18 1.0%	7 0.4%	32 1.8%	152 8.4%	74 4.1%	11 0.6%	9 0.5%	48.3% 51.7%
	6	6 0.3%	9 0.5%	3 0.2%	7 0.4%	11 0.6%	33 1.8%	6 0.3%	7 0.4%	40.2% 59.8%
	7	9 0.5%	11 0.6%	4 0.2%	4 0.2%	4 0.2%	4 0.2%	16 0.9%	2 0.1%	29.6% 70.4%
	8	15 0.8%	5 0.3%	0 0.0%	0 0.0%	3 0.2%	5 0.3%	8 0.4%	25 1.4%	41.0% 59.0%
		75.8% 24.2%	42.0% 58.0%	27.3% 72.7%	25.3% 74.7%	69.1% 30.9%	23.2% 76.8%	19.8% 80.2%	17.0% 83.0%	50.2% 49.8%
		1	2	3	4	5	6	7	8	
		Actual class								

Figure 4.7 The confusion matrix of cascaded LM networks for eight emotions prediction

4.3.7 Binary-Class Rprop Network for Valence Prediction

Training the model with Rprop algorithm not only took lesser time than the LM algorithm but also better accuracy. The modeling performance is shown in Table 4.9, and the classification result is shown in Figure 4.8.

Table 4.9 The performance of binary-class Rprop networks for valence discrimination

Model	Train (%)	Val. (%)	Test (%)	Entire (%)	F1-score	Best Epoch	Best Round	Time (h)
Traditional Multiclass Rprop for Valence	76.6	76.3	77.4	76.7	0.8306	4	2,778	1.79

Predicted class	1	1030 57.2%	238 13.2%	81.2% 18.8%
	2	182 10.1%	352 19.5%	65.9% 34.1%
		85.0% 15.0%	59.7% 40.3%	76.7% 23.3%
		1	2	
		Actual class		

Figure 4.8 The confusion matrix of binary-class Rprop networks for valence discrimination

4.3.8 Binary-Class Rprop Network for Arousal Prediction

The binary classification of arousal with Rprop algorithm returned higher accuracy than valence which is similar to the LM algorithm but much better accuracy (see Chapter 4.3.2). The modeling performance is shown in Table 4.10, and the classification result is shown in Figure 4.9.

Table 4.10 The performance of binary-class Rprop networks for arousal discrimination

Model	Train (%)	Val. (%)	Test (%)	Entire (%)	F1-score	Best Epoch	Best Round	Time (h)
Traditional Multiclass Rprop for Arousal	96.7	94.8	96.3	96.3	0.9737	18	3,455	2.6

Predicted class	1	1222 67.8%	46 2.6%	96.4% 3.6%
	2	20 1.1%	514 28.5%	96.3% 3.7%
		98.4% 1.6%	91.8% 8.2%	96.3% 3.7%
		1	2	
		Actual class		

Figure 4.9 The confusion matrix of binary-class Rprop networks for arousal discrimination

4.3.9 Traditional Multiclass Rprop Network for Four Emotions Prediction

This result is comparable to four emotions prediction with traditional multiclass LM network (see the result in Chapter 4.3.3). The traditional multiclass Rprop for four classes returned higher accuracy about ten percent than the traditional multiclass LM for four classes as Table 4.11 and Figure 4.10 have shown.

Table 4.11 The performance of traditional multiclass Rprop networks for four emotions prediction

Model	Train (%)	Val. (%)	Test (%)	Entire (%)	F1-score	Best Epoch	Best Round	Time (h)
Traditional Multiclass Rprop for 4 Emotion	81.9	76.7	68.1	79	See Table 4.15	15	391	2.18

Predicted class	1	970 53.8%	67 3.7%	40 2.2%	133 7.4%	80.2% 19.8%
	2	7 0.4%	86 4.8%	16 0.9%	4 0.2%	76.1% 23.9%
	3	23 1.3%	41 2.3%	299 16.6%	22 1.2%	77.7% 22.3%
	4	14 0.8%	4 0.2%	7 0.4%	69 3.8%	73.4% 26.6%
		95.7% 4.3%	43.4% 56.6%	82.6% 17.4%	30.3% 69.7%	79.0% 21.0%
		<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	
		Actual class				

Figure 4.10 The confusion matrix of traditional multiclass Rprop networks for four emotions prediction

4.3.10 Traditional Multiclass Rprop Network for Eight Emotions Prediction

This result is comparable to eight emotions prediction with traditional multiclass LM network (see Chapter 4.3.4), similar to the result of four emotion classification in the previous subchapter, the accuracy of traditional multiclass Rprop for eight classes higher than traditional multiclass LM for eight classes about ten percent as well. The modeling performance is shown in Table 4.12, and the classification result is shown in Figure 4.11.

Table 4.12 The performance of traditional multiclass Rprop networks for eight emotions prediction

Model	Train (%)	Val. (%)	Test (%)	Entire (%)	F1-score	Best Epoch	Best Round	Time (h)
Traditional Multiclass Rprop for 8 Emotion	59.6	62.2	52.2	58.9	See Table 4.16	8	6	2.79

Predicted class	1	450 25.0%	106 5.9%	15 0.8%	7 0.4%	5 0.3%	7 0.4%	16 0.9%	49 2.7%	68.7% 31.3%
	2	109 6.0%	283 15.7%	40 2.2%	20 1.1%	25 1.4%	17 0.9%	21 1.2%	24 1.3%	52.5% 47.5%
	3	4 0.2%	3 0.2%	8 0.4%	4 0.2%	2 0.1%	2 0.1%	1 0.1%	0 0.0%	33.3% 66.7%
	4	0 0.0%	0 0.0%	1 0.1%	4 0.2%	1 0.1%	0 0.0%	1 0.1%	0 0.0%	57.1% 42.9%
	5	9 0.5%	11 0.6%	23 1.3%	46 2.6%	165 9.2%	31 1.7%	7 0.4%	6 0.3%	55.4% 44.6%
	6	9 0.5%	11 0.6%	11 0.6%	14 0.8%	18 1.0%	82 4.6%	12 0.7%	11 0.6%	48.8% 51.2%
	7	0 0.0%	1 0.1%	1 0.1%	2 0.1%	2 0.1%	1 0.1%	15 0.8%	3 0.2%	60.0% 40.0%
	8	14 0.8%	4 0.2%	0 0.0%	2 0.1%	2 0.1%	2 0.1%	8 0.4%	54 3.0%	62.8% 37.2%
			75.6% 24.4%	67.5% 32.5%	8.1% 91.9%	4.0% 96.0%	75.0% 25.0%	57.7% 42.3%	18.5% 81.5%	36.7% 63.3%
		1	2	3	4	5	6	7	8	
		Actual class								

Figure 4.11 The confusion matrix of traditional multiclass Rprop networks for eight emotions prediction

4.3.11 Cascaded Rprop Network for Four Emotions Prediction

Instead of using only one network to predict all four emotional classes, the cascaded Rprop network for four emotions consist of three Rprop network to took advantage of binary classification of arousal. This result is comparable to four emotions prediction with traditional multiclass Rprop network and both LM networks for four emotions (The results in Chapter 4.3.3, 4.3.5, and 4.3.9). The modeling performance is shown in Table 4.13, and the classification result is shown in Figure 4.12. These result show that the cascaded Rprop network is the winner among their competitors for four emotions prediction.

Table 4.13 The performance of cascaded Rprop networks for four emotions prediction

Model	Train (%)	Val. (%)	Test (%)	Entire (%)	F1-score	Best Epoch	Best Round	Time (h)
Cascaded Rprop Unit 1	96.7	94.8	96.3	96.3	0.9737	18	3,455	2.6
Cascaded Rprop Unit 2	96.6	95.2	96.2	96.3	0.9774	4	3,074	2.2
Cascaded Rprop Unit 3	98.7	97.6	96.4	98.2	0.9749	1	3,484	1.6

Predicted class	1	982 54.5%	16 0.9%	9 0.5%	23 1.3%	95.3% 4.7%
	2	3 0.2%	166 9.2%	5 0.3%	3 0.2%	93.8% 6.2%
	3	9 0.5%	4 0.2%	339 18.8%	5 0.3%	95.0% 5.0%
	4	20 1.1%	12 0.7%	9 0.5%	197 10.9%	82.8% 17.2%
		96.8% 3.2%	83.8% 16.2%	93.6% 6.4%	86.4% 13.6%	93.5% 6.5%
	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>		
	Actual class					

Figure 4.12 The confusion matrix of cascaded Rprop networks for four emotions prediction

4.3.12 Cascaded Rprop Network for Eight Emotions Prediction

Instead of using only one network to predict all eight emotional classes, the cascaded Rprop network for eight emotions consist of seven Rprop network to took advantage of binary classification of arousal and valence. This result is comparable to eight emotions prediction with traditional multiclass Rprop network and both LM networks for eight emotions (see Chapter 4.3.4, 4.3.6, and 4.3.10). The modeling performance is shown in Table 4.14, and the classification result is shown in Figure 4.13. These result show that the cascaded Rprop network is the winner among their competitors for eight emotions prediction.

Table 4.14 The performance of cascaded Rprop networks for eight emotions prediction

Model	Train (%)	Val. (%)	Test (%)	Entire (%)	F1-score	Best Epoch	Best Round	Time (h)
Cascaded Rprop Unit 1	96.7	94.8	96.3	96.3	0.9737	18	3,455	2.6
Cascaded Rprop Unit 2	96.6	95.2	96.2	96.3	0.9774	4	3,074	2.2
Cascaded Rprop Unit 3	98.7	97.6	96.4	98.2	0.9749	1	3,484	1.6
Cascaded Rprop Unit 4	95.6	95.4	91.4	95.0	0.9574	2	10,603	1.6
Cascaded Rprop Unit 5	96.9	97.1	85.3	95.2	0.9308	1	122	0.3
Cascaded Rprop Unit 6	96.4	100.0	96.7	97.0	0.9706	19	9,950	0.1
Cascaded Rprop Unit 7	95.7	96.3	94.4	95.6	0.9644	1	357	0.1

Predicted class	1	560 31.1%	26 1.4%	3 0.2%	4 0.2%	2 0.1%	1 0.1%	2 0.1%	10 0.6%	92.1% 7.9%
	2	20 1.1%	376 20.9%	7 0.4%	2 0.1%	2 0.1%	4 0.2%	5 0.3%	6 0.3%	89.1% 10.9%
	3	1 0.1%	1 0.1%	79 4.4%	3 0.2%	0 0.0%	1 0.1%	0 0.0%	1 0.1%	91.9% 8.1%
	4	0 0.0%	1 0.1%	0 0.0%	84 4.7%	3 0.2%	1 0.1%	0 0.0%	2 0.1%	92.3% 7.7%
	5	1 0.1%	5 0.3%	0 0.0%	1 0.1%	208 11.5%	11 0.6%	1 0.1%	0 0.0%	91.6% 8.4%
	6	1 0.1%	2 0.1%	0 0.0%	3 0.2%	2 0.1%	118 6.5%	2 0.1%	2 0.1%	90.8% 9.2%
	7	7 0.4%	4 0.2%	5 0.3%	1 0.1%	3 0.2%	3 0.2%	64 3.6%	3 0.2%	71.1% 28.9%
	8	5 0.3%	4 0.2%	5 0.3%	1 0.1%	0 0.0%	3 0.2%	7 0.4%	123 6.8%	83.1% 16.9%
		94.1% 5.9%	89.7% 10.3%	79.8% 20.2%	84.8% 15.2%	94.5% 5.5%	83.1% 16.9%	79.0% 21.0%	83.7% 16.3%	89.5% 10.5%
		1	2	3	4	5	6	7	8	
		Actual class								

Figure 4.13 The confusion matrix of cascaded Rprop networks for eight emotions prediction

4.4 Discussions

The modeling results will be discussed in three term in this subchapter. Firstly, relationship between accuracy of submodels and final accuracies of cascaded model will be explained in Chapter 4.4.1. The model performance measured by F1-score will be discussed in Chapter 0 and compared to the other proposed models in Chapter 4.4.3.

4.4.1 The Relationship of Submodels in Cascaded Models

As mentions in Chapter 3.5.2 that, cascaded structure (see Figure 3.9) composed of multiple submodels which each one has its own accuracy but these accuracies were not used to calculate the final accuracy. The final accuracies of cascaded models were evaluated as the same as traditional multiclass models i.e. output or predicted class of sample was compared to actual class of sample directly and calculated by a simple true or false logic. These accuracies are the number that reported at the bottom right of every confusion matrix in earlier subchapters. However, these number are able to calculate by the rule of three combines with the fractional percentage either. This calculation method also uses to verify the final accuracy of each cascaded model.

We have setup a simulation of accuracy calculation on spreadsheet “Excel” to verify the results of model. In this place we will give an example of finding number of correctly predicted sample in submodel unit two of cascaded model only. This calculation could be applied to every unit of cascade structure. The calculation begins with finding the percentage of actual sample in each unit as shown in Figure 4.14. The total sample number of each level is 1,802 which is equal to 100%, therefore the number of 1,242 in the unit two in level two is equal to 68.92%. When all of the ratio of sample was found out, the triangle of the rule of three is ready to apply. We will have four variables, first is number of Sample in Current Unit (nSCU) (red frame) which we want to find out. The second is percentage of Sample in Current Unit (pSCU) (yellow frame). The third is percentage of Accuracy in Parent Unit (pAPU) (green

frame). The fourth is number of Sample in Parent Unit (nSPU) (blue frame). These four variables have the relationship as defined by (6).

$$nSCU = \frac{pSCU}{100} \times \frac{pAPU \cdot nSPU}{100} \quad (6)$$

The results can be seen in Figure 4.15 and Figure 4.16, on the left is accuracy of cascaded submodels which were the same as reported in Table 4.8 and Table 4.14 and on the right is the calculated nSCU. The calculation shows that, the total correctly predicted sample of cascaded LM is 895 or 49.7% which is slightly difference about 0.5% from the actual accuracy of cascaded LM model. The total correctly predicted sample of cascaded Rprop is 1,604 or 89.0% which is only 0.5% difference from the actual accuracy of cascaded Rprop model as well. This error might be caused by the difference of dept of decimal between real accuracies on MATLAB and simulated accuracies on Excel.

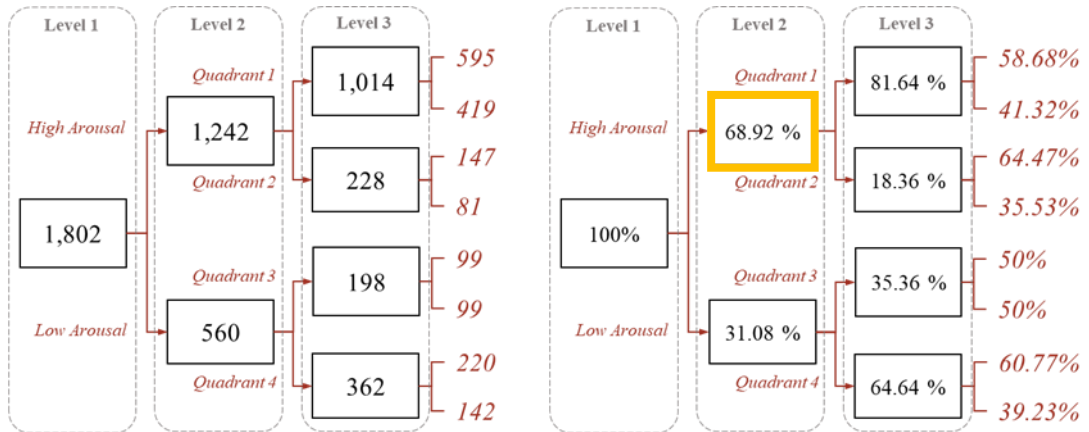


Figure 4.14 Number of correctly predicted sample in each submodel in case of perfect accuracy (left), ratio of sample in each submodel (right)

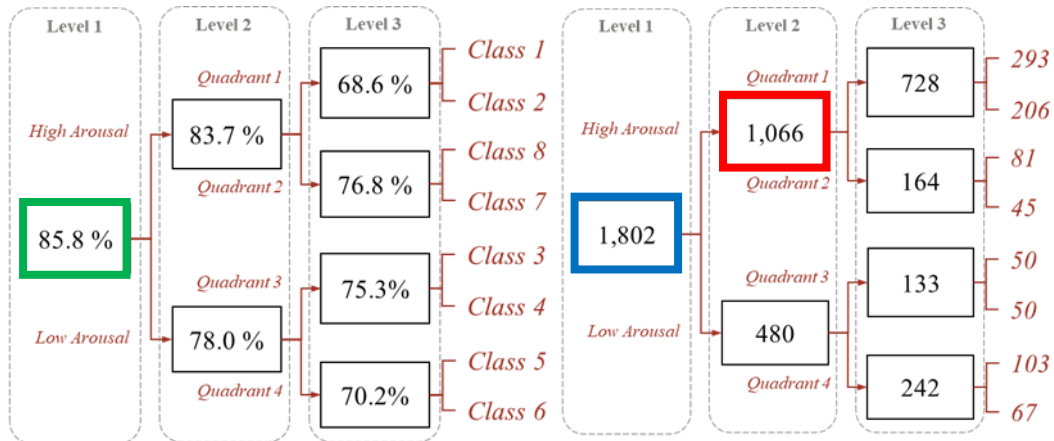


Figure 4.15 Submodel accuracies of cascaded LM model (left), number of correctly predicted sample in each submodel in case of cascaded LM model (right)

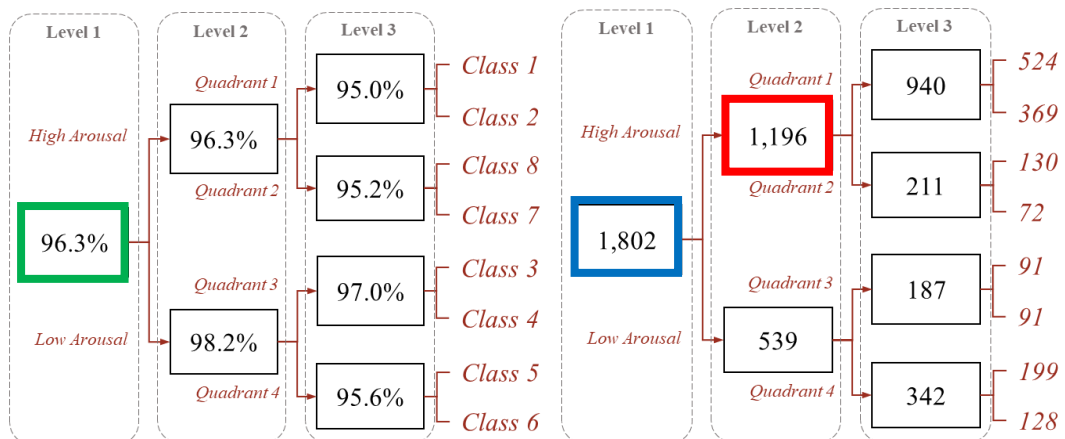


Figure 4.16 Submodel accuracies of cascaded Rprop model (left), number of correctly predicted sample in each submodel in case of cascaded Rprops model (right)

There are two drawbacks associated with using cascaded structures are. The first drawback is the complexity of the cascaded model. For the three-levels depth, a sample must pass through three models. The processing time can be slow if the next sample has to wait until the previous sample has passed the last level, but it would be more efficient if the next sample can be inputted while the previous sample is processed at level 2; then, when these two samples move to levels 3 and 2, the next sample can follow them, and so on. Additionally, programing for connecting each submodel to form a cascaded structure can be complicated.

The second is performance of the first level is crucial. If the submodel in the previous level fails to predict a sample, the levels after that are useless because the accuracy of the submodel in the first level greatly affects the final accuracy and the second level has more of an effect than the third, as these following three simulations shown.

In the simulation, if the accuracy of unit 1 (level 1) is 80% and that of the rest is 100%, then the final accuracy will be 80% because the total number of samples passing through unit 1 is 1,442, and so a loss 20% of unit 1 will be equal to 360 samples or also 20% of entire dataset (1,802 samples).

If the accuracy of unit 3 (level 2) is 80% and that of the rest is 100%, then the final accuracy will be 93.8% because the total number of samples passing through unit 3 is 560, and so a loss 20% of unit 3 will be equal to 112 samples or approximately 6.3% of entire dataset.

If the accuracy of unit 6 (level 3) is 80% and that of the rest is 100%, then the final accuracy will be 97.8% because the total number of samples passing through unit 6 is 198, and so a loss of 20% of unit 6 will be equal to 40 samples, or only 2.2% of the entire dataset.

4.4.2 Performance Measurement using F1-score

This subchapter present comparison tables of F1-score in each class for four and eight emotions classification separately in Table 4.15 and Table 4.16 respectively.

In four emotions classification, when classified the samples with cascaded Rprop model the F1-score of class two and four were much better than traditional multiclass Rprop as we can see in Figure 4.17. On the other hand, there was no significant different between traditional multiclass LM and cascaded LM.

Table 4.15 The F1-score of four emotions classification

Model	F1-score of Class 1	F1-score of Class 2	F1-score of Class 3	F1-score of Class 4	Average F1-score	Accuracy
Traditional Multiclass LM	0.8729	0.3081	0.6424	0.3291	0.5381	70.6%
Cascaded LM	0.8828	0.4231	0.6907	0.3407	0.5843	73.4%
Traditional Multiclass Rprop	0.9112	0.4765	0.7598	0.4220	0.6424	79.0%
Cascaded Rprop	0.9708	0.8491	0.9199	0.8698	0.9024	93.5%

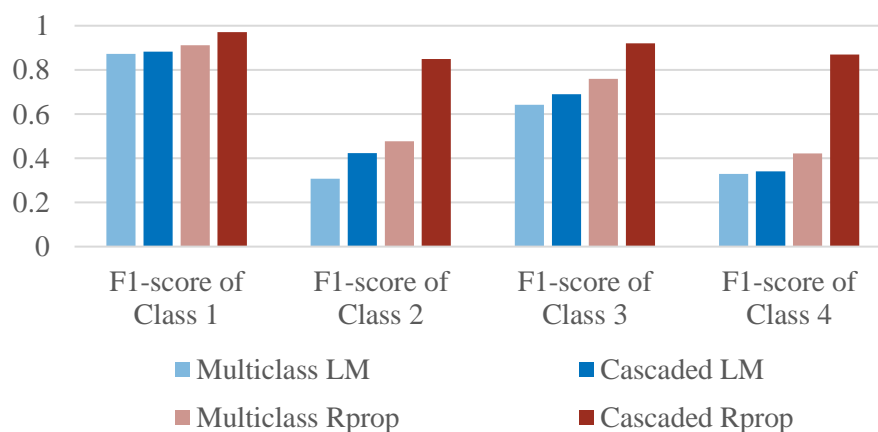


Figure 4.17 The comparison of F1-score of four emotion classification

The aim of the work is eight emotional prediction therefore we will pay more attention on Table 4.16 and Figure 4.18. The relationship of F1-score and number of training data can be seen by comparing bar chart of number of sample in Figure 2.6

with bar chart of F1-score in Figure 4.18. The height of each class and model of F1-score correspond to number of samples in each class. F1-score of the traditional multiclass LM, cascaded LM, and traditional multiclass Rprop, are similar to each other where the problematic predicted classes are one, two, five, and six. Since the samples are often wrongly predicted to belong to one of these classes because the population of these classes are larger and has a more uncertain pattern of data than the others. The prediction accuracy for the cascaded LM model is disappointing because it offers no significant improvement over the traditional multiclass LM. This means that, the LM algorithm is not suitable for a cascaded structure.

Table 4.16 The F1-score of eight Emotion Classification

Model	F1-score of Class 1	F1-score of Class 2	F1-score of Class 3	F1-score of Class 4	F1-score of Class 5	F1-score of Class 6	F1-score of Class 7	F1-score of Class 8	Average F1-score	Accuracy
Traditional Multiclass LM	0.7816	0.6176	0.1453	0.1390	0.4943	0.2940	0.1197	0.2086	0.3500	47.8%
Cascaded LM	0.8009	0.6108	0.1941	0.1823	0.5755	0.2274	0.1249	0.1823	0.3623	50.2%
Traditional Multiclass Rprop	0.8293	0.7534	0.0795	0.0414	0.6405	0.4696	0.1394	0.3683	0.4152	58.9%
Cascaded Rprop	0.9593	0.9406	0.7688	0.7796	0.8975	0.8324	0.7293	0.8382	0.8432	89.5%

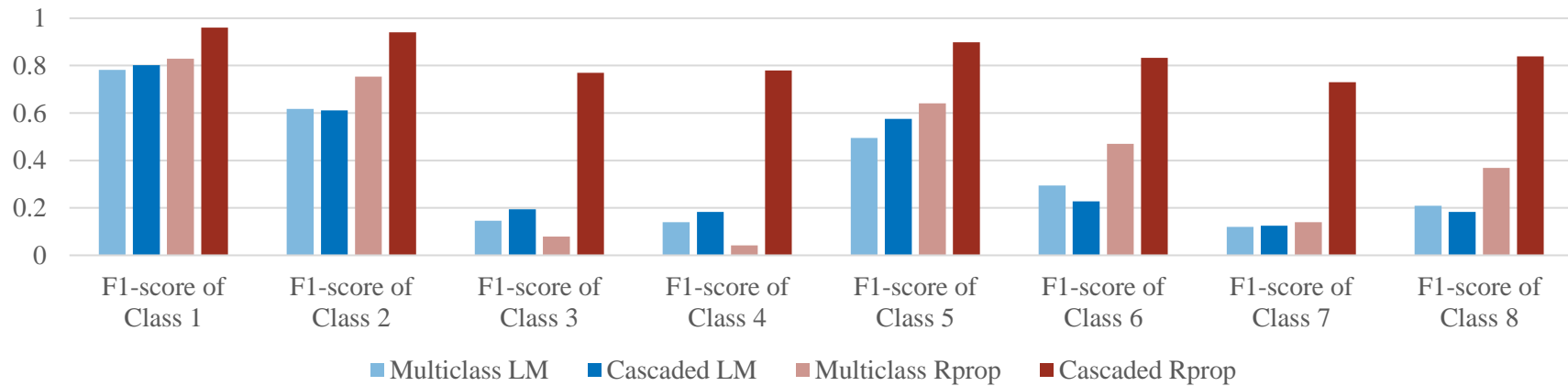


Figure 4.18 The comparison of F1-score of eight emotion classification

4.4.3 Results Comparison using Key Performance Indicators

We compared the performance of our work with that of other approaches, although direct comparison is impossible due to the different goals, resources, and achievement metrics. However, the progress of development and improvement can be observed in Table 4.17 by sorting the studies by time and capture mutual information or key performance indicators (KPIs) to compare these studies in terms of several factors such as methodology, number of samples, number of features, number of emotional classes, and percentage of achievement claimed by the measurement method of each work.

Normally a lower RMSE is better, but we invert the values such that a higher value is better to make the values comparable to those from other measurement methods. We also expanded our results to four emotions and valence/arousal level prediction to make our results comparable. When we measured our accuracy with only valence and arousal level, the results were similar to those of every work that measured their results in terms of valence/arousal level. Arousal prediction is always better than valence prediction, as observed in experiments No. 1, 9, 13, 14, 22, 27 and 28.

Among the four emotional classification works, our work (No. 32 to 34) performs at an average level, but No. 35 outperforms the others. Experiments No. 8 and 10 are comparable to No. 33 and 35. These works were conducted based on the same concept of using cascaded structures to distinguish only two classes at a time to classify a total of four emotional classes. Experiments No. 24 to 27 were conducted based on the same dataset that we used. However, the numbers of samples associated with each emotion were equalized, and the performance metric was the F1-score, while our work took the entire dataset and our performance metric was accuracy.

Among the eight-emotion classification works, our methods (No. 36 to 39) are comparable to experiment No. 11, which also employed multiple models. However, No. 11 simply employed eight models to regress each class simultaneously to form a structure, while No. 37 and 39 employed a cascaded structure of seven models to discriminate two classes at a time.

Table 4.17 The comparison of key performance indicators in each previous work

#	Key Algorithms and/or Methods	Year	Extractor	Sampling Rate (Hz)	File Format	Audio Channel	No. of Sample	No. of Feature	No. of Class	Achievement (%)	Measurement Method	Ref. No.
1	SVM Regressor with RReliefF feature selection	2008	Psysound, Marsyas, Spectral contrast, DWCH ¹	22.05k	WAV	Mono	195	114	V/A	28.1/58.3	R^2	[7]
2	Binary Relevance	2008	Marsyas	22.05k	WAV	Mono	593	74	6	73.78	Accuracy	[20]
3	Label Powerset									76.69	Accuracy	
4	Random k-Labelsets									79.54	Accuracy	
5	Multilabel k-Nearest Neighbor									71.04	Accuracy	
6	Auto Associative Neural Networks	2013	Praat (wav converter), MFCC ²	44.1k	MP3	Mono	85	52	5	94.4	Accuracy	[5]
7	Support Vector Machine									85	Accuracy	
8	Hierarchical SVM based on tempo & mutation degree	2013	N/A	N/A	MIDI	N/A	80	2	4	95	Accuracy	[2]
9	Recurrent Neural Networks	2014	ComParE ³	N/A	MP3	N/A	1,000	70	V/A	50/70	R^2	[15]
10	Hierarchical SVM	2014	NWFE, KBCS	22.05k	N/A	Mono	219	35	4	89.64	Accuracy	[3]

¹ Daubechies wavelets coefficient histogram

² Mel-frequency cepstral coefficient

³ Computational Paralinguistics Evaluation

Table 4.17 (Continued) The comparison of key performance indicators in each previous work

#	Key Algorithms and/or Methods	Year	Features Extractor	Sampling Rate (Hz)	File Format	Audio Channel	No. of Sample	No. of Feature	No. of Class	Achievement (%)	Measurement Method	Ref. No.
11	Eight Regressors, Individually trained	2015	Sound Description toolbox, MIR toolbox, PsySound	22.05k	WAV	Mono	385	117	8	59.35	Accuracy	[8]
12	Nearest multi-prototype classifier	2015	MIR toolbox	N/A	MP3	N/A	903	59	5	56.43	Accuracy	[19]
13	Adaptive Aggregation of Gaussian Process Regressors	2016	MediaEval 2014	N/A	N/A	N/A	744	65	V/A	77/80	RMSE	[16]
14	Stacked CNN & RNN	2017	openSMILE toolbox	N/A	N/A	N/A	431	260	V/A	73/80	RMSE	[17]
15	k-Nearest Neighbors	2017	Sound Description toolbox, MIR toolbox	44.1k	MP3	N/A	1,000	548	4	62	Accuracy	[4]
16	Bayes Classifier									69	Accuracy	
17	Linear Discriminant Analysis									80.4	Accuracy	
18	Neuro-Fuzzy Network Classification									79.3	Accuracy	
19	Fuzzy KNN									83	Accuracy	
20	Support Vector Machine									82.7	Accuracy	
21	RandomForest	2017	Sound Description toolbox, PsySound3, Marsyas	44.1k	N/A	N/A	300	397	V/A	57.3/70	Accuracy	[18]
22	Support Vector Machine	2017	MIR, Tempogram, Chroma toolbox, PsySound	N/A	N/A	N/A	818	539	6	85	Accuracy	[6]
23	Support Vector Regressor								V/A	25/79	Accuracy	

Table 4.17 (Continued) The comparison of key performance indicators in each previous work

#	Key Algorithms and/or Methods	Year	Features Extractor	Sampling Rate (Hz)	File Format	Audio Channel	No. of Sample	No. of Feature	No. of Class	Achievement (%)	Measurement Method	Ref. No.
24	Support Vector Machine	2017	MIR toolbox	44.1k	MP3	Stereo	943	33	4	46	F1-score	[21]
25	Naïve Bayes									40	F1-score	
26	Decision Trees									37	F1-score	
27	k-Nearest Neighbors									41	F1-score	
28	LM Networks for Valence	2018	MIR toolbox	44.1k	MP3	Stereo	1,802	122	2(V)	78.9	Accuracy	This Work
29	Rprop Networks for Valence									76.7	Accuracy	
30	LM Networks for Arousal								2(A)	85.8	Accuracy	
31	Rprop Networks for Arousal									96.3	Accuracy	
32	Traditional Multiclass LM Networks for 4 Emotions								4	70.6	Accuracy	
33	Cascaded LM Networks 4 Emotions									73.4	Accuracy	
34	Traditional Multiclass Rprop Networks 4 Emotions									79.0	Accuracy	
35	Cascaded Rprop Networks 4 Emotions	93.5	Accuracy									

Table 4.17 (Continued) The comparison of key performance indicators in each previous work

#	Key Algorithms and/or Methods	Year	Features Extractor	Sampling Rate (Hz)	File Format	Audio Channel	No. of Sample	No. of Feature	No. of Class	Achievement (%)	Measurement Method	Ref. No.
36	Traditional Multiclass LM Networks 8 Emotions	2018	MIR toolbox	44.1k	MP3	Stereo	1,802	122	8	47.8	Accuracy	This Work
37	Cascaded LM Networks 8 Emotions									50.2	Accuracy	
38	Traditional Multiclass Rprop Networks 8 Emotions									58.9	Accuracy	
39	Cascaded Rprop Networks 8 Emotions									89.5	Accuracy	

4.5 Summary

The 122 acoustic feature elements were extracted from audio files by MIR toolbox. The extracted features cover five types of acoustic features i.e. dynamics, rhythm, timbre, pitch, and Tonality. A drawback of using MIR toolbox is processing time. Elapsed time for each song is unequal up to fluctuation and complexity of each song, most of extraction task could be done in few hours for entire dataset.

The music emotion prediction results are presented step by step. Begin with such a simple task like valence or arousal discrimination to the most complicate task of eight music emotional classification at the end, 12 results in total. We present the model performance in table form, and accuracy is explained with a confusion matrix. The results show that cascaded Rprop gave the best accuracy at 89.5% for eight music emotion prediction. The result of traditional multiclass LM, cascaded LM and traditional multiclass Rprop model training methods have the similar problem, that is predicted confusion of class 1,2,5, and 6. The predicted class 1 confused with class 2 and predicted class 5 confused with class 6. The cascaded LM modeling method did not work well as were expected, perhaps LM neural network dose not fits for cascaded structure. However, the cascaded Rprop modeling method able to reduce confusion significantly.

The accuracies of cascaded models were verified with calculation using the rule of three and fractional percentage. These simulations have been shown to correspond well with actual accuracies. In order to explain how the models made incorrectly prediction, F1-score was employed to measure harmonic average of the precision and recall. The results were compared in various aspect with the other work. The comparation reveal that our work (cascaded Rprop) is the best in term of number of samples, number of emotional classes and accuracy.

CHAPTER 5

CONCLUSION

5.1 Conclusion

We have designed a music emotion recognition system using eight emotion classes. The system was developed on MATLAB, 122 acoustic features were extracted by the MIR toolbox, and four modeling methods were investigated: traditional multiclass LM, cascaded LM, traditional multiclass Rprop, and cascaded Rprop. We evaluated the system with the DEAM benchmark. The dataset was divided into a 7/3 ratio (training set/testing set). The accuracies were 47.8%, 50.2%, 58.9%, and 89.5%, respectively. The results for the cascaded Rprop model confirm the scalability of prediction with multimodel methods demonstrated in previous work [2][3][8]. We also found that timbre features were the most important for music emotion prediction.

There are some difficulties in the prediction of class No. 3, 4, 7, and 8. This may be due to a lack of uniqueness of the features or because the annotation process contains ambiguity at first. Perhaps the music in these classes has a complex texture that is hard to distinguish in term of valence and arousal.

A major drawback of the proposed method is its heavy computational requirements, but this was handled by using high performance hardware. We have provided elapsed time data for features extraction, details on the model training process, and our hardware specification for reference, which should be helpful for reproducing or extending this work. We found that with our system environment mentioned in Chapter 3.1, the CPU was only fully loaded when eight extraction tasks were executing at the same time, which consumed approximately 30GB of RAM. During the modeling, we were able to run two LM algorithm tasks at the same time and five tasks for the Rprop algorithm. Therefore, the total elapsed time is not the summation of the reported elapsed times in Table 4.1 and 4.3 to 4.14.

The cascaded structure could be both advantage and drawback up to how its submodels were trained. The submodel in level 1 is the most significant and lesser and lesser in level 2 and level 3 as explained by simulations in Chapter 4.4.1.

5.2 Future Work

Future work should investigate how to index the minimum cost in terms of computational time and hardware requirements, while maintaining acceptable accuracy. One approach would be to consider only high-impact features, as demonstrated in Chapter 4.2. The overall accuracy might decrease slightly, but the elapsed time will decrease significantly. For instance, when we removed feature No. 63, which is the most time-consuming feature to extract, and trained the model with only the other 121 features, the accuracy of the cascaded Rprop model was still high (84.1%). Parallel processing and cloud base system could be beneficial to cascade model training because submodels able to be trained simultaneously.

In this work we use mean value of VA-values annotated by many people as described in Chapter 3.1. However, it possible to change annotation methods. Since, the DEAM benchmark provided raw annotation data. For instance, convert VA-values of each annotator then employ majority vote to pick up the most frequent annotated emotion.

Our work aims to encourage music providers to categorize music by using emotional terms. The benefits will be a more efficient search and better access to music. This work may also lead to additional applications, such as music playlist generation based on listener heart rate and automatic stage-lighting control based on music emotions [56][57].

REFERENCES

- [1] Y.-H. Yang and H. Chen, "01 Introduction," in *Music Emotion Recognition*, CRC Press, 2011, pp. 1–13.
- [2] J. Wang and S. Xin, "Emotional Classification Based on The Tempo and Mutation Degrees," in *2013 2nd International Symposium on Instrumentation and Measurement, Sensor Network and Automation (IMSNA)*, 2013, pp. 444–446.
- [3] W. C. Chiang, J. S. Wang, and Y. L. Hsu, "A Music Emotion Recognition Algorithm with Hierarchical SVM Based Classifiers," in *2014 International Symposium on Computer, Consumer and Control*, 2014, pp. 1249–1252.
- [4] J. Bai *et al.*, "Music Emotion Recognition by Cognitive Classification Methodologies," in *2017 IEEE 16th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC)*, 2017, pp. 121–129.
- [5] S. Nalini, N J and Palanivel, "Emotion Recognition in Music Signal using AANN and SVM," *Int. J. Comput. Appl.*, vol. 77, no. 2, pp. 7–14, 2013.
- [6] X. Hu and Y.-H. Yang, "The Mood of Chinese Pop Music: Representation and Recognition," *Int. Rev. Res. Open Distance Learn.*, pp. 90–103, 2017.
- [7] H. H. C. Yi-Hsuan Yang, Yu-Ching Lin, Ya-Fan Su, "A Regression Approach to Music Emotion Recognition," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 16, no. 2, pp. 448–457, 2008.
- [8] Y. Deng, Y. Lv, M. Liu, and Q. Lu, "A Regression Approach to Categorical Music Emotion Recognition," in *2015 IEEE International Conference on Progress in Informatics and Computing (PIC)*, 2015, pp. 257–261.
- [9] K. Hevner, "Expression in Music: A Discussion of Experimental Studies and Theories," *Psychol. Rev.*, vol. 42, no. 2, pp. 186–204, 1935.
- [10] Russell James A., "A Circumplex Model of Affect," *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980.
- [11] Y. E. Kim *et al.*, "Music Emotion Recognition : a State of The Art Review," in *11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, 2010, pp. 255–266.

- [12] Y. Yang and H. H. Chen, "Machine Recognition of Music Emotion : A Review," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 40, 2012.
- [13] H. Cheng, Y. Yang, Y. Lin, I. Liao, and H. H. Chen, "Automatic Chord Recognition for Music Classification and Retrieval," in *2008 IEEE International Conference on Multimedia and Expo (ICME) (2008)*, 2008, vol. 0, pp. 1505–1508.
- [14] J. Kim and E. André, "Emotion Recognition Based on Physiological Changes in Music Listening," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 12, pp. 2067–2083, 2008.
- [15] F. Weninger, F. Eyben, and B. Schuller, "On-line Continuous-Time Music Mood Regression with Deep Recurrent Neural Networks," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 338164, no. 338164, pp. 5412–5416, 2014.
- [16] S. Fukayama and M. Goto, "Music Emotion Recognition With Adaptive Aggregation of Gaussian Process Regressors," *Icassp 2016*, pp. 71–75, 2016.
- [17] M. Malik, S. Adavanne, K. Drossos, T. Virtanen, D. Ticha, and R. Jarina, "Stacked Convolutional and Recurrent Neural Networks for Music Emotion Recognition," 2017. [Online]. Available: <http://arxiv.org/abs/1706.02292>.
- [18] V. L. Nguyen, D. Kim, V. P. Ho, and Y. Lim, "A New Recognition Method for Visualizing Music Emotion," vol. 7, no. 3, pp. 1246–1254, 2017.
- [19] B. K. Baniya, C. S. Hong, and J. Lee, "Nearest Multi-Prototype Based Music Mood Classification," in *2015 IEEE/ACIS 14th International Conference on Computer and Information Science, ICIS 2015 - Proceedings*, 2015, pp. 303–306.
- [20] K. Trohidis and G. Kalliris, "Multi-Label Classification of Music Into Emotions," in *9th International Society for Music Information Retrieval Conference (ISMIR 2008)*, 2008, pp. 325–330.
- [21] P. M. F. Vale, "The Role of Artist and Genre on Music Emotion Recognition," Universidade Nova de Lisboa, 2017.
- [22] K. Sorussa, A. Choksuriwong, and M. Karnjanadecha, "Acoustic Features for Music Emotion Recognition and System Building," in *Proceedings of the 2017*

- International Conference on Information Technology - ICIT 2017*, 2017, pp. 413–417.
- [23] E. Çano and M. Morisio, “MoodyLyrics: A Sentiment Annotated Lyrics Dataset,” in *Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence - ISMSI '17*, 2017, pp. 118–124.
- [24] E. Çano and M. Morisio, “Music Mood Dataset Creation Based on Last.fm Tags,” in *4th International Conference on Artificial Intelligence and Applications (AIAP 2017)*, 2017, pp. 15–26.
- [25] Y. Chen, Y. Yang, J. Wang, and H. Chen, “The AMG1608 Dataset for Music Emotion Recognition,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 693–697.
- [26] X. Hu and Y.-H. Yang, “Cross-dataset and Cross-cultural Music Mood Prediction: A Case on Western and Chinese Pop Songs,” *IEEE Trans. Affect. Comput.*, vol. 8, no. 2, pp. 1–1, 2016.
- [27] “Free Music Archive.” [Online]. Available: <http://freemusicarchive.org/>. [Accessed: 07-Nov-2018].
- [28] “Jamendo Music.” [Online]. Available: <https://www.jamendo.com/>. [Accessed: 07-Nov-2018].
- [29] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. Bello, “MedleyDB: A Multitrack Dataset for Annotation - Intensive MIR Research,” in *International Society for Music Information Retrieval Conference*, 2014, no. Ismir, pp. 155–160.
- [30] A. Aljanaki, Y. H. Yang, and M. Soleymani, “Developing a Benchmark for Emotional Analysis of Music,” *PLoS One*, vol. 12, no. 3, pp. 1–22, 2017.
- [31] M. Soleymani, A. Aljanaki, and Y.-H. Yang, “DEAM: MediaEval Database for Emotional Analysis in Music,” 2016. [Online]. Available: <http://cvml.unige.ch/databases/DEAM/manual.pdf>. [Accessed: 08-Mar-2018].
- [32] “DEAM Dataset Release Page.” [Online]. Available: <http://cvml.unige.ch/databases/DEAM>. [Accessed: 01-Jun-2017].
- [33] O. Lartillot and P. Toivainen, “A MATLAB Toolbox for Musical Feature Extraction from Audio,” in *Proc. of the 10th Int. Conference on Digital Audio*

- Effects (DAFx-07)*, 2007, pp. 1–8.
- [34] T. Eerola and P. Toiviainen, “MIR in MATLAB: The Midi Toolbox,” in *International Society for Music Information Retrieval Conference (ISMIR)*, 2004, pp. 22–27.
- [35] M. Müller and S. Ewert, “MIR Toolbox Release Page.” [Online]. Available: <https://www.jyu.fi/hytk/fi/laitokset/mutku/en/research/materials/mirtoolbox>. [Accessed: 05-Jul-2017].
- [36] H. Bowen *et al.*, Eds., “The Element of Classical Music,” in *The Complete Classical Music Guide*, 1st ed., New York: Dorling Kindersley Limited, 2012, pp. 10–17.
- [37] D. Moffat, D. Ronan, and J. D. Reiss, “An Evaluation of Audio Feature Extraction Toolboxes,” in *Proc. of the 18th Int. Conference on Digital Audio Effects (DAFx-15)*, 2015, pp. 1–7.
- [38] G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams, “The Timbre Toolbox: Extracting audio descriptors from musical signals,” *J. Acoust. Soc. Am.*, vol. 130, no. 5, p. 2902, 2011.
- [39] G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams, “Timbre toolbox release page.” [Online]. Available: <http://www.cirmmt.org/l/research-tools/timbretoolbox/>. [Accessed: 05-Jul-2017].
- [40] P. Grosche and M. Müller, “Tempogram toolbox: Matlab implementations for tempo and pulse analysis of music recordings,” ... *Int. Conf. Music ...*, 2011.
- [41] P. Grosche and M. Müller, “Tempogram Toolbox Release page.” [Online]. Available: <https://www.audiolabs-erlangen.de/resources/MIR/tempogramtoolbox>. [Accessed: 05-Jul-2017].
- [42] M. Müller and S. Ewert, “Chroma Toolbox: Matlab Implementations for Extracting Variants of Chroma-Based Audio Features,” *12th Int. Soc. Music Inf. Retr. Conf. (ISMIR 2011)*, no. Ismir, pp. 215–220, 2011.
- [43] M. Müller and S. Ewert, “Chroma toolbox release page.” [Online]. Available: <http://resources.mpi-inf.mpg.de/MIR/chromatoolbox/>. [Accessed: 05-Jul-2017].
- [44] “Chroma feature.” [Online]. Available: <https://en.wikipedia.org/wiki/File:ChromaFeatureCmajorScaleScoreAudioColo>

r.png. [Accessed: 05-May-2017].

- [45] D. Bogdanov *et al.*, “ESSENTIA: An audio analysis library for music information retrieval,” *Proc. Int. Conf. Music Inf. Retr.*, pp. 493–498, 2013.
- [46] “Essentia official @ essentia.upf.edu.” [Online]. Available: <http://essentia.upf.edu/>. [Accessed: 08-Jan-2017].
- [47] H. B. Demuth, M. H. Beale, O. De Jess, and M. T. Hagan, *Neural Network Design*, 2nd ed. 2014.
- [48] M. H. Beale, M. T. Hagan, and H. B. Demuth, *Neural Network ToolboxTM User’s Guide*. 3 Apple Hill Drive: Mathwork inc, 2017.
- [49] K. Levenberg, “A Method for The Solution of Certain Non Linear Problems In Least Squares,” *Q. Appl. Math.*, vol. 2, pp. 164–168, Jan. 1944.
- [50] D. W. Marquardt, “An Algorithm for Least-Squares Estimation of Nonlinear Parameters,” *J. Soc. Ind. Appl. Math.*, vol. 11, no. 2, pp. 431–441, Jun. 1963.
- [51] M. T. Hagan and M. B. Menhaj, “Training Feedforward Networks with The Marquardt Algorithm,” *IEEE Trans. Neural Networks*, vol. 5, no. 6, pp. 989–993, 1994.
- [52] M. Riedmiller, M. Riedmiller, and H. Braun, “RPROP - A Fast Adaptive Learning Algorithm,” in *PROC. OF ISICIS VII, UNIVERSITAT*, 1992.
- [53] W. Saputra, T. Tulus, M. Zarlis, R. W. Sembiring, and D. Hartama, “Analysis Resilient Algorithm on Artificial Neural Network Backpropagation,” *J. Phys. Conf. Ser.*, vol. 930, no. 1, 2017.
- [54] M. H. Beale, M. T. Hagan, and H. B. Demuth, *Neural Network ToolboxTM Reference*. 3 Apple Hill Drive: Mathwork inc, 2017.
- [55] B. P. F. William H. Press, Saul A. Teukolsky, William T. Vetterling, “Chapter 14.5 Linear Correlation,” in *Numerical Recipes in C*, 2nd ed., Cambridge University Press, 1992, pp. 636–639.
- [56] H. Liu, J. Hu, and M. Rauterberg, “Music Playlist Recommendation Based on user Heartbeat and Music Preference,” in *2009 International Conference on Computer Technology and Development (ICCTD 2009)*, 2009, vol. 1, pp. 545–549.
- [57] S.-W. Hsiao, S.-K. Chen, and C.-H. Lee, “Methodology for Stage Lighting

Control Based on Music Emotions,” *Inf. Sci. (Ny)*, vol. 412–413, pp. 14–35, 2017.

APPENDIX

Table A. Correlation values between extracted features and VA-ratings

Feat. No.	Feat. Type	Raw Data Type	Valence Correlation	Valence Correlation Ranking	Arousal Correlation	Arousal Correlation Ranking
1	Dynamics	time series	0.0393	84	0.1883	104
2	Dynamics	time series	-0.1562	8	-0.2996	5
3	Rhythm	time series	-0.0029	50	-0.1480	30
4	Rhythm	time series	-0.0311	38	-0.1689	28
5	Rhythm	time series	-0.0628	16	-0.2000	26
6	Rhythm	time series	-0.0700	12	-0.2080	18
7	Rhythm	time series	-0.0730	11	-0.2123	16
8	Rhythm	time series	-0.0672	13	-0.2068	21
9	Rhythm	time series	-0.0652	14	-0.2061	24
10	Rhythm	time series	-0.0579	17	-0.2066	22
11	Rhythm	time series	-0.0574	18	-0.2079	19
12	Rhythm	time series	-0.0549	19	-0.2070	20
13	Rhythm	time series	-0.0488	23	-0.2041	25
14	Rhythm	time series	-0.0483	24	-0.2062	23
15	Rhythm	time series	-0.0482	25	-0.2106	17
16	Rhythm	time series	-0.0497	21	-0.2142	14
17	Rhythm	time series	-0.0489	22	-0.2171	13
18	Rhythm	time series	-0.0516	20	-0.2222	12
19	Rhythm	time series	-0.0480	26	-0.2225	11
20	Rhythm	time series	-0.0426	29	-0.2247	10
21	Rhythm	time series	-0.0392	32	-0.2278	9
22	Rhythm	time series	-0.0232	44	-0.2128	15
23	Rhythm	time series	-0.0005	54	-0.1856	27
24	Rhythm	time series	0.0192	68	-0.1565	29
25	Rhythm	time series	0.0318	75	-0.1376	31
26	Rhythm	time series	0.0433	86	-0.1256	33
27	Rhythm	time series	-0.2297	6	-0.2392	8
28	Rhythm	time series	0.1860	108	0.4642	117
29	Rhythm	time series	0.4196	121	0.6232	121
30	Rhythm	numeric	0.0998	100	0.1415	97
31	Rhythm	time series	0.3184	115	0.2857	111
32	Rhythm	time series	0.4266	122	0.3351	113

Table A. (Continued) Correlation values between extracted features and VA-ratings

Feat. No.	Feat. Type	Raw Data Type	Valence Correlation	Valence Correlation Ranking	Arousal Correlation	Arousal Correlation Ranking
33	Timbre	time series	-0.1403	9	-0.1307	32
34	Timbre	time series	0.2708	111	0.2552	109
35	Timbre	time series	0.2851	114	0.2049	106
36	Timbre	time series	-0.241	5	-0.3667	3
37	Timbre	time series	0.2089	109	0.2904	112
38	Timbre	time series	0.1776	107	0.0376	68
39	Timbre	time series	-0.2882	3	-0.2463	7
40	Timbre	time series	0.2804	113	0.5099	119
41	Timbre	time series	0.3847	119	0.4623	116
42	Timbre	time series	0.3577	116	0.5828	120
43	Timbre	numeric	0.3827	118	0.5068	118
44	Timbre	numeric	0.3748	117	0.4294	115
45	Timbre	numeric	-0.3262	2	-0.4497	2
46	Timbre	numeric	-0.2235	7	-0.3286	4
47	Timbre	time series	0.2788	112	0.3431	114
48	Timbre	time series	0.3933	120	0.6286	122
49	Timbre	time series	-0.3339	1	-0.5241	1
50	Timbre	time series	0.1281	104	0.0491	70
51	Timbre	time series	-0.1342	10	-0.1226	34
52	Timbre	time series	0.1157	103	0.033	67
53	Timbre	time series	-0.0009	53	-0.047	40
54	Timbre	time series	0.047	89	0.0627	75
55	Timbre	time series	0.0428	85	0.0617	74
56	Timbre	time series	0.0693	92	0.0896	88
57	Timbre	time series	0.0376	82	0.0867	86
58	Timbre	time series	0.0949	98	0.18	103
59	Timbre	time series	0.0293	74	0.1055	93
60	Timbre	time series	0.0799	95	0.087	87
61	Timbre	time series	0.0003	56	0.0898	89
62	Timbre	time series	0.0065	60	0.23	108
63	Timbre	time series	-0.2567	4	-0.2769	6
64	Pitch	numeric	0.1012	102	0.2823	110

Table A. (Continued) Correlation values between extracted features and VA-ratings

Feat. No.	Feat. Type	Raw Data Type	Valence Correlation	Valence Correlation Ranking	Arousal Correlation	Arousal Correlation Ranking
65	Pitch	time series	0.018	66	0.1794	101
66	Tonality	time series	0.0795	94	0.1148	95
67	Tonality	time series	0.0909	97	0.1685	100
68	Tonality	time series	0.0686	91	0.0938	92
69	Tonality	time series	0.0467	88	0.1136	94
70	Tonality	time series	0.0363	79	0.1211	96
71	Tonality	time series	0.0459	87	0.0744	79
72	Tonality	time series	0.0202	69	0.0554	73
73	Tonality	time series	0.0742	93	0.0932	91
74	Tonality	time series	0.1749	106	0.2297	107
75	Tonality	time series	0.1346	105	0.1796	102
76	Tonality	time series	0.09	96	0.1463	98
77	Tonality	time series	0.0996	99	0.203	105
78	Tonality	time series	-0.0004	55	-0.006	55
79	Tonality	time series	0.0186	67	0.0009	58
80	Tonality	time series	0.0081	61	0.0145	62
81	Tonality	time series	-0.013	47	-0.0442	44
82	Tonality	time series	0.0164	64	0.0824	84
83	Tonality	time series	-0.0013	52	-0.0335	47
84	Tonality	time series	-0.038	35	-0.0485	39
85	Tonality	time series	0.003	58	0.0054	59
86	Tonality	time series	0.0345	78	0.0189	63
87	Tonality	time series	0.0278	72	0.0722	78
88	Tonality	time series	-0.0251	43	-0.0705	35
89	Tonality	time series	-0.0387	33	-0.0065	54
90	Tonality	classes	0.0095	62	0.0294	65
91	Tonality	time series	0.0999	101	-0.03	50
92	Tonality	time series	-0.0373	36	-0.0445	43
93	Tonality	time series	-0.0407	31	-0.0458	41
94	Tonality	time series	-0.043	28	-0.0447	42
95	Tonality	time series	-0.0435	27	-0.0391	45
96	Tonality	time series	-0.0424	30	-0.0307	48

Table A. (Continued) Correlation values between extracted features and VA-ratings

Feat. No.	Feat. Type	Raw Data Type	Valence Correlation	Valence Correlation Ranking	Arousal Correlation	Arousal Correlation Ranking
97	Tonality	time series	-0.0382	34	-0.0194	51
98	Tonality	time series	-0.0304	39	-0.0049	56
99	Tonality	time series	-0.0206	45	0.0095	61
100	Tonality	time series	-0.0091	49	0.0253	64
101	Tonality	time series	0.0032	59	0.0405	69
102	Tonality	time series	0.0132	63	0.0537	72
103	Tonality	time series	0.0223	70	0.0668	77
104	Tonality	time series	0.0289	73	0.075	80
105	Tonality	time series	0.0338	77	0.0807	82
106	Tonality	time series	0.0375	81	0.0829	85
107	Tonality	time series	0.0383	83	0.0807	83
108	Tonality	time series	0.0373	80	0.0758	81
109	Tonality	time series	0.0334	76	0.0653	76
110	Tonality	time series	0.0276	71	0.0528	71
111	Tonality	time series	0.0165	65	0.0323	66
112	Tonality	time series	0.0015	57	0.0088	60
113	Tonality	time series	-0.0136	46	-0.0134	52
114	Tonality	time series	-0.0254	42	-0.0306	49
115	Tonality	time series	-0.0324	37	-0.0388	46
116	Tonality	time series	-0.0016	51	0.0005	57
117	Tonality	time series	-0.0096	48	-0.0567	38
118	Tonality	time series	-0.0278	40	-0.0089	53
119	Tonality	time series	0.0591	90	0.0898	90
120	Tonality	time series	-0.0637	15	-0.0575	36
121	Tonality	time series	-0.0277	41	-0.0569	37
122	Tonality	time series	0.2275	110	0.1684	99

VITAE

Name Mr. Kanawat Sorussa

Student ID 5910120073

Educational Attainment

Degree	Name of Institution	Year of Graduation
Bachelor of Engineering (Computer Engineering)	Rajamangala University of Technology Srivijaya	2016

Scholarship Awards during Enrolment

Scholarship supported by faculty of engineering, Prince of Songkla University.

Research material supported by graduate school, Prince of Songkla University.

List of Publication and Proceeding

K. Sorussa, A. Choksuriwong, and M. Karnjanadecha, "Acoustic Features for Music Emotion Recognition and System Building," in Proceedings of the 2017 International Conference on Information Technology - ICIT 2017, 2017, pp. 413–417.