Emotion Classification System for Digital Music with a Cascaded Technique

Kanawat Sorussa[1], **Anant Choksuriwong**[2], and **Montri Karnjanadecha**[3]

ABSTRACT

Music selection is difficult without efficient organization based on metadata or tags, and one effective tag scheme is based on the emotion expressed by the music. However, manual annotation is labor intensive and unstable because the perception of music emotion varies from person to person. This paper presents an emotion classification system for digital music with a resolution of eight emotional classes. Russell's emotion model was adopted as common ground for emotional annotation. The music information retrieval (MIR) toolbox was employed to extract acoustic features from audio files. The classification system utilized a supervised machine learning technique to recognize acoustic features and create predictive models. Four predictive models were proposed and compared. The models were composed by crossmatching two types of neural networks, the Levenberg-Marquardt (LM) and resilient backpropagation (Rprop), with two types of structures: a traditional multiclass model and the cascaded structure of a binary-class model. The performance of each model was evaluated via the MediaEval Database for Emotional Analysis (DEAM) benchmark. The best result was achieved by the model trained with the 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 in music classification because only a few music providers currently tag music with emotional terms.

Keywords: Artificial Neural Networks, Classification Algorithms, Emotion Recognition, Music Information Retrieval

1. INTRODUCTION

The appearance of digital music providers has changed the way people listen to music by offering 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 who search for music use emotional keywords [1].

Unfortunately, most music providers tag music by genre, artist 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. Yang and Chen proposed the conceptual framework for an MER system, as shown in Fig. 1 [1].

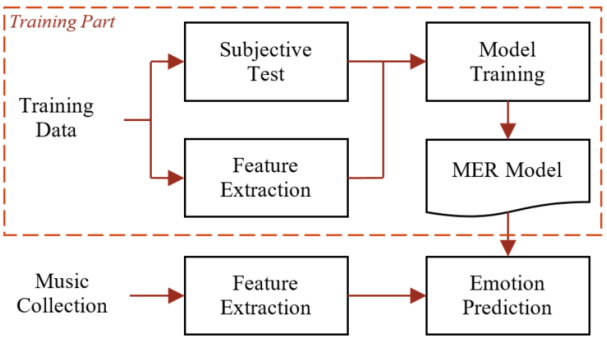

Fig.1: *A Music Emotion Classification System Framework [1].*

First, music was collected and annotated with one of the emotion models mentioned in section 2, and then acoustic features were extracted from the audio file. Finally, a supervised machine learning technique was applied to reveal the relationship between music emotions and acoustic features.

In section 3, we briefly review 14 studies published since 2008. Some studies adopted a dimensional emotional model as a quadrant of emotion or support only four emotional classes [2–4]. Some studies adopted a categorical emotional model of four to six classes [5–9]. Even though most of these works obtained

Manuscript received on July 24, 2019 ; revised on January 28, 2020.

Final manuscript received on February 5, 2020.

[1,2,3] The authors are with the Department of Computer Engineering Faculty of Engineering Prince of Songkla University, Songkhla, Thailand (e-mail: profkanawat@gmail.com, anant.c@psu.ac.th and montri.k@psu.ac.th

[1] The scholarship of the first author and funding for research material are supported by the Faculty of Engineering and Graduate School of Prince of Songkla University, Songkhla, Thailand.

DOI: 10.37936/ecti-cit.2020141.205317

over 80% accuracy, four to six classes are too limited to describe emotion in music. Resolution at eight emotions is sufficient in many applications, such as background music, for emotional scenarios in video games and commercial purposes [10–14]. Moreover, several studies used small datasets with limited variety [2][6][7][15], causing potential problems 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 single model [2][7][8]. We hypothesized that using multiple models with cascaded structures could reduce the number of false predictions if each model were specifically trained to discriminate only two classes at a time.

Therefore, this study makes a major contribution to the classification of 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 of various kinds.

We can approach the problem in two ways: regression and classification. To train regression models, the sample is labeled with the continuous values of the fundamental factors of emotion. Because regression models estimate the closest values of the factors, emotional classes can be predicted based on these values. In contrast, because the samples for training classification models are labeled with a discrete number of emotional classes, classification models can directly predict the most probable class.

These two approaches used two different algorithms. Both were specifically designed for the corresponding approach, and neither has previously been tested with the MediaEval Database for Emotional Analysis (DEAM) dataset. Levenberg-Marquardt (LM) backpropagation was chosen for the regression approach, and Resilient backpropagation (Rprop) was chosen for the classification approach. Additionally, we investigated each algorithm in two ways. A traditional multiclass model was employed and compared to seven cascading units of the binary-class model.

We found that the Rprop algorithm with a cascaded structure achieved the best accuracy when compared to the other three methods and previously proposed methods.

2. EMOTION REPRESENTATION

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

2.1 Categorical Psychometrics

Categorical psychometrics represents 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 [16].

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

2.2 Dimensional Psychometrics

Dimensional psychometrics represents 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 and arousal, to form a valence-arousal (VA) plane on a scale of -1 to 1, as shown in Fig. 2 [17].

Various valence and arousal coordinates define 28 emotional adjectives. This approach is flexible, measurable, and comparable, but the relationships between valence and arousal can be difficult to explain. Culture and language have an effect on the VA rating. For example, in English, “happy” is located at 0.11 for arousal and 0.83 for valence, but the coordinates can be different in other languages [18].

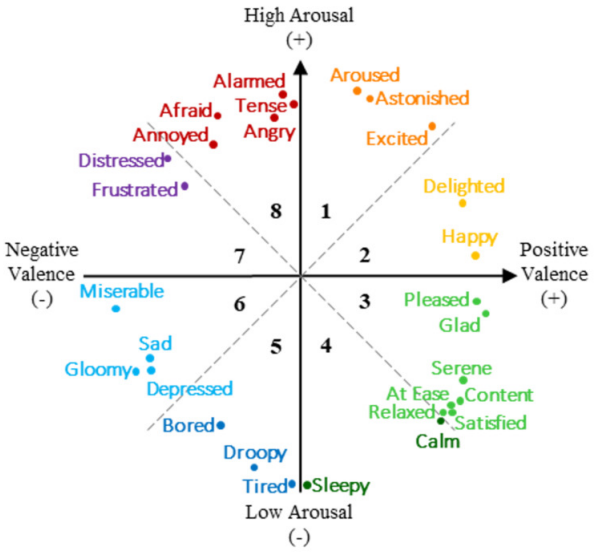

Fig.2: *Emotional Adjectives on the VA Plane [17].*

A systematic comparison of categorical and dimensional psychometrics by employing linear mapping techniques revealed a high correspondence between the two psychometrics formulations [19]. The study also involved three dimensions (valence, energy

Table 1: *Emotional Octant and the Associated 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
∑	**8**	**28**

arousal and tension arousal) and showed that only two dimensions were enough to represent perceived emotions in music. The major difference between the categorical and dimensional psychometrics was the resolution of emotional representation.

Consequently, dimensional psychometrics can be use as categorical psychometrics by reducing the resolution or grouping similar emotions together.

Therefore, we adopted Russell's model as an octant of emotion for classification. The VA plane was divided into eight emotional classes by a range of fundamental factors. Emotional adjectives possessing a VA rating in a common range were grouped as one emotional class. The logic for developing the range of each class and the emotional adjectives is shown in Table 1, and the ranges are shown as dashed lines in Fig. 2.

3. PREVIOUS WORK

Music processing retrieves information in many forms, such as score notes, lyrics, audio signals, and chords [20–22]. Music emotion is often annotated based on verbal reports of emotional responses, although some studies have gathered data by monitoring biological or physical expressions [23]. However, we are interested only in the retrieval of information from audio signals and annotations from verbal reports.

Systems recognize music by referring to one of the psychometrics frameworks described in section 2. For dimensional psychometrics, a regression approach estimates the valence and arousal, whereas for categorical psychometrics, a classification approach is employed.

In the fallowing subsections, we reveal how other studies were addressed by employing key performance indicators (KPIs) to compare these studies in terms of five factors: methodology, number of samples, number of features, number of emotional classes, and accuracy claimed by the measurement method of each work. The exact numbers of samples and features, and the results of the studies mentioned in this section, are reported in Table 8, which includes the results of our work for comparison.

3.1 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 results of 28.1% valence and 58.3% arousal were achieved [15].

Weninger *et al.* captured time-varying emotion through music using recurrent neural networks (RNNs). The performance was evaluated by the R^2 statistic, and results of 50% valence and 70% arousal were achieved [24].

One of the common challenges with multiple-feature input data is ranking the most important features. Features that have a substantial effect on estimation should be weighted to improve the results of the calculation. For example, Fukayama and Goto utilized adaptive aggregation to obtain a feature ranking and estimated the VA-value via Gaussian process regressors The performance evaluated in terms of the root-mean-square error (RMSE) reached 77% for valence and 80% for arousal [25].

A recent study conducted 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 feature map was the input to both RNNs. The performance was evaluated in terms of the RMSE, with results of 73% for valence and 80% for arousal [26].

Most VA value estimation problems are solved by regression algorithms, but some researchers have used classification approaches by converting a continuous range to a finite range. Nguyen *et al.* divided valence and arousal levels into six segments and coordinated those segments to obtain a total of 36 segments. Then, the random forest algorithm was implemented in WEKA to classify valence and arousal as one of these six levels. The accuracies were 57.3% for valence and 70% for arousal [27].

3.2 Classification Approach

Hu and Yang created a dataset of Chinese-pop music (C-pop) for the MER task, which was atypical because most MER tasks have been conducted on Western music. The dataset was analysed by both regression and classification approaches. First, the dataset was investigated via SVM, and an accuracy of 85% was achieved for six-emotion classification. Then, the dataset was analyzed by means of a support vector re-

gressor (SVR), and the accuracy was 25% for valence and 79% for arousal [28].

A small-scale experiment demonstrated that music could be classified into four emotional classes with the help of a hierarchical SVM using only two features: tempo and mutation degree. Three SVM units were utilized, with each unit trained for a specific purpose. The first unit was trained to discriminate high and low tempos. High-tempo songs are happy or aggressive, while low-tempo songs are sad or soft. The second and third units were trained to discriminate between those emotions. The results were impressive, yielding 95% accuracy [7]. The analysis was repeated with a larger number of features and samples. The results were satisfactory, with an accuracy of 89.64% [2].

An investigation of six algorithms, which were SVM, k-nearest neighbors (KNN), neuro-fuzzy network classification (NFNC), fuzzy KNN (FKNN), a Bayesian classifier, and linear discriminant analysis (LDA), for classifying four emotional classes showed that the accuracies of the LDA, SVM, and FKNN algorithms were higher than 80% [3].

Nalini *et al.* investigated autoassociative neural networks (AANNs) and an SVM for classifying five music emotions; the accuracies were 94.4% and 85.0%, but the models were trained with a small dataset [6]. Another study applying the nearest multiprototype classifier to a very large dataset achieved only 56.43% accuracy [9].

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 [5].

Deng *et al.*, conducted a study 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 is impressive considering the number of samples, the number of emotional classes, and the proposed method [8].

Most MER studies focus on only acoustic features as inputs and ignore nonacoustic features, such as artist and genre. However, the impact of these nonacoustic 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 F1-scores were 46%, 40%, 37%, and 41%. The artist feature was not impactful, and the genre feature was only slightly beneficial for the decision tree method. The overall accuracy was not high because the experiments considered a limited number of acoustic features [4].

4. DATASET

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 [29–31]. Some datasets include extracted features [32], and some datasets consider the cultural background of the annotators [28][33]. Datasets that do not provide audio files can lead to problems because we cannot make any potentially required changes to the process.

Fortunately, the DEAM benchmark includes a dataset with audio files that can be redistributed under a Creative Commons (CC) license; thus, this dataset was utilized in this work. This DEAM benchmark includes 1,744 clips and 58 full-length songs. The audio files are in stereo MP3 format with a 44.1-kHz sampling rate and a 128-kbps bitrate. 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 included [34–36]. 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. Each song was annotated by five to ten people, and we used the average of the annotations [37–39].

In Fig. 3, the subfigure on the left 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 Fig. 2 and Table 1. The subfigure on the right shows where each sample was located on the VA plan.

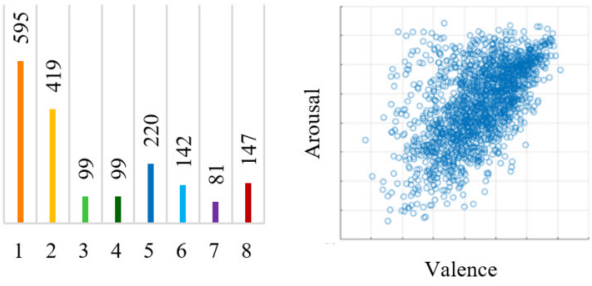

Fig.3: *Number of Music Samples in the DEAM Dataset for Each of the Eight Emotions (left); a Scatterplot of All Samples on the VA Plane (right).*

As shown in Fig. 3, each class has a different number of samples. The inequality of training samples

can bias the results. The prediction of classes 1 and 2, which have the largest populations, might achieve high accuracy, while the other classes might have lower 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 accuracy for classes 3, 4, 7, and 8 was not significantly improved [4]. The incorrect prediction of these classes must be caused by other characteristics. Therefore, we did not equalize the number of samples.

5. ACOUSTIC FEATURES

5.1 Feature Extraction

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 [40–42]. This tool was chosen because it can extract numerous features, including the five groups of features described below [29][43][44].

1) Dynamics is the physical intensity of a sound, and is often described as loudness, energy, volume, or audio power.

2) Rhythm is a periodic pattern of changes or events of pitch level, dynamics, or pulses. Pulse speed is known as meter, phrasing, tempo, or beats-per-minute.

3) Timbre can be explained as follows. When a guitar and a violin play the same note, the sound is similar, but we hear a difference; that difference is timbre. Each musical instrument has its own timbre, which is particularly useful for musical instrument recognition.

4) Pitch is the level of sound. In Western music, pitch is encoded by the letters C, D, E, F, G, A, B. While a piano changes pitch discretely, some instruments, such as a violin, allow continuous change.

5) Tonality is the arrangement of pitches and/or chords into major and minor scales and keys. Major and minor refer to the spaces between notes, with note separation measured in whole and half steps.

The functions used for feature extraction, their output and the running time for the entire dataset are presented in Table 2. A total of 122 features were extracted by the 37 functions, and some functions produced multiple feature elements, such as feature nos. 3 to 26, representing 12 notes in an octave.

The outputs of most extractors are a time series. Some provide a continuous numerical value, but feature no. 90, for example, is a discrete value taken from a set of twelve classes. To make the data compatible, we transformed the time series and discrete class data into individual numerical values by using the "mirmean" function to find the average of the time series, and the "mirgetdata" function to assign a numerical value to represent each of the discrete classes. For example, the classes of feature no. 90 were transformed into values from 1 to 12.

Table 2: *List of MIR Toolbox Functions for Acoustic Feature Extraction.*

Feature No.	Feature Type	Extractor Function	Output Type	Time (h)
1	Dynamics	mirrms	Time Series	0.89
2	Dynamics	mirlowenergy	Time Series	0.92
3-26	Rhythm	mirfluctuation	Time Series	0.93
27	Rhythm	mirbeatspectrum	Time Series	4.82
28	Rhythm	mirevents	Time Series	3.46
29	Rhythm	mireventdensity	Time Series	4.00
30	Rhythm	mirtempo	Numeric	3.62
31	Rhythm	mirmetroid	Time Series	5.85
32	Rhythm	mirpulseclarity	Time Series	3.74
33	Timbre	mirattacktime	Time Series	33.51
34	Timbre	mirattackslope	Time Series	35.44
35	Timbre	mirattackleap	Time Series	33.86
36	Timbre	mirdecaytime	Time Series	34.72
37	Timbre	mirdecayleap	Time Series	34.03
38	Timbre	mirdecayslope	Time Series	33.67
39	Timbre	mirduration	Time Series	33.73
40	Timbre	mirzerocross	Time Series	0.97
41	Timbre	mirrolloff	Time Series	1.33
42	Timbre	mirbrightness	Time Series	1.08
43	Timbre	mircentroid	Numeric	1.12
44	Timbre	mirspread	Numeric	1.39
45	Timbre	mirskewness	Numeric	1.52
46	Timbre	mirkurtosis	Numeric	1.51
47	Timbre	mirflatness	Time Series	1.22
48	Timbre	mirentropy	Time Series	1.01
49-61	Timbre	mirmfcc	Time Series	1.32
62	Timbre	mirroughness	Time Series	2.08
63	Timbre	mirregularity	Time Series	325.32
64	Pitch	mirpitch	Numeric	1.51
65	Pitch	mirmidi	Time Series	12.10
66-77	Tonality	mirchromagram	Time Series	1.02
78-89	Tonality	mirkeystrength	Time Series	1.02
90	Tonality	mirkey	Classes	1.40
91	Tonality	mirmode	Time Series	1.18
92-115	Tonality	mirkeysom	Time Series	1.01
116-121	Tonality	mirtonalcentroid	Time Series	1.01
122	Tonality	mirhcdf	Time Series	3.58

5.2 Feature Correlation

We measured the linear correlation between features and the fundamental factors of emotion using a linear correlation coefficient. The range of the correlation coefficient (r) is -1 to 1. An absolute r value of 1 indicates a perfect linear relationship, and an r value of 0 indicates the absence of a linear relationship. The closer the absolute value of r is to 1, the more ideal the linear relationship. Features with r values close to 0 are still useful (unless the r value is exactly 0) but have a less significant impact [45].

The correlations of each feature with valence and arousal were measured separately and can be visualized, as shown in Fig. 4.

To identify the impactful features, we ranked the feature correlations in ascending order. The partial correlation ranking is presented in Table 3. The correlations of features with valence and arousal were ranked separately, as shown in the "Rank of V" and "Rank of A" columns. The significance of each feature can be calculated by summing the absolute values of "Valence r" and "Arousal r" of each feature.

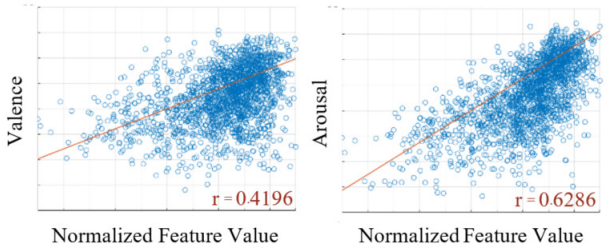

Fig.4: *Scatter Plot of Feature No. 29 Against Valence (left), and Arousal (right).*

Table 3: *Correlation Values Between Extracted Features and VA Ratings.*

Feature No.	Rank of V	Rank of A	Valence r	Arousal r	Correlation Status
49	1	1	-0.3339	-0.5241	Negative Correlation
45	2	2	-0.3262	-0.4497	
39	3	7	-0.2882	-0.2463	
36	5	3	-0.241	-0.3667	
62	60	108	0.0065	0.23	Poor Correlation
80	61	62	0.0081	0.0145	
90	62	65	0.0095	0.0294	
112	57	60	0.0015	0.0088	
99	45	61	-0.0206	0.0095	
48	120	122	0.3933	0.6232	Positive Correlation
29	121	121	0.4196	0.6286	
32	122	113	0.4266	0.3351	
42	116	120	0.3577	0.5828	

As shown by the correlations ranking, 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, an optimized model for use 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.

6. SYSTEM BUILDING

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

6.1 System Structures

Two different system structures were built to determine which method is better for pursuing our goal. The first was a regression approach that classifies the VA response into classes at the output of the predictive model. The second was a classification approach that classifies the VA response into classes at the input of the predictive model.

Previous work on the DEAM dataset achieved low accuracy because a limited number of features were used and because the SVM, naïve Bayes, decision trees, and k-NN techniques do not work well on this dataset [4]. Therefore, we chose the LM algorithm for the regression approach and the Rprop algorithm for the classification approach. These two algorithms appear to be the best choices in the MATLAB neural network toolbox for each approach, as proven with a variety of datasets. For further detail information about these algorithms, please see [46–52].

6.1.1 Regression Approach

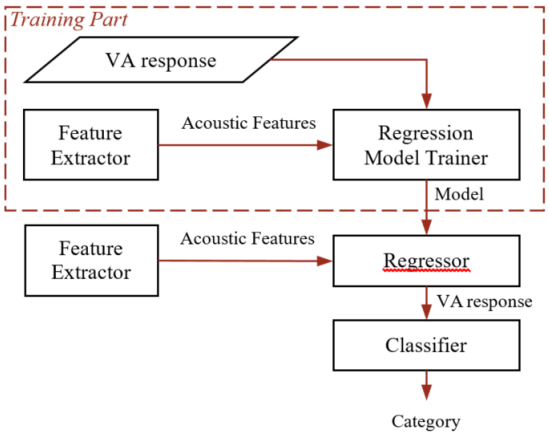

Fig.5: *Framework for the Regression Approach System.*

The DEAM dataset provides annotations in the form of valence and arousal responses. The regression approach can use these annotations directly. Fig. 5 shows the framework of the regression system. The LM algorithm was employed by the regression model trainer module, and the regressor module produced estimated VA responses. To determine whether these predictions were correct, acceptable areas of error were extended for the classifier module, as shown in Fig. 6.

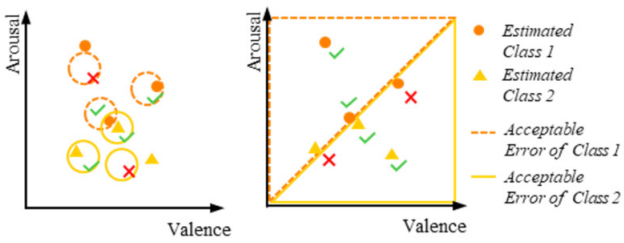

Fig.6: *General Acceptable Areas of Error (left) and Extended Acceptable Areas of Error in the Classifier Module (right).*

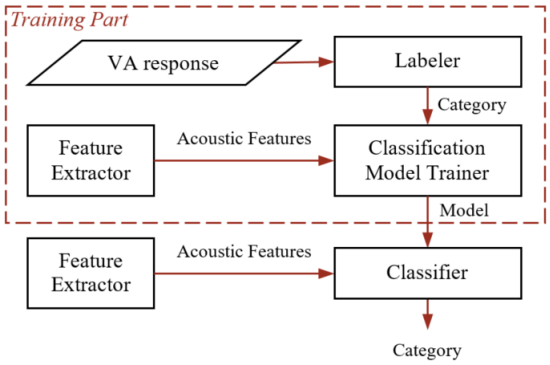

Fig.7: *Framework for the Classification Approach System.*

6.1.2 Classification Approach

Fig. 7 shows the framework of the classification system. The labeler module converts the VA responses to classes before their use as annotations in the model trainer module because the raw annotations cannot be used directly. The Rprop algorithm is then employed in the classification model trainer module.

6.2 Model Structures

The predictive model in each system was customized to have two different structures (traditional multiclass and cascading binary-class) to determine which is better for model training.

6.2.1 Traditional Multiclass Model

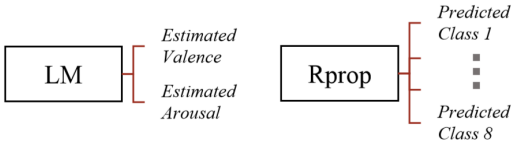

Fig.8: *Traditional Structure of the Predictive Model with the LM Algorithm (left) and the Rprop Algorithm (right).*

The LM and Rprop algorithms in the model trainer module were trained with 122 acoustic features of 1,802 songs to predict music emotion, but the outputs of the model were different, as illustrated in Fig. 8.

6.2.2 Cascaded Model

The cascaded model was obtained by connecting several units of binary-class submodels as a cascaded structure, as shown in Fig. 9. Each submodel was specifically trained to discriminate only two classes, as described in Table 4.

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][7][8][28]. Additionally, many regression approach studies have shown that arousal prediction is always more accurate than valence prediction [15][24–28]. Therefore, we initiated unit 1 to discriminate between high and lowarousal 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, etc. (Quadrants 1 to 4 and classes 1 to 8 refer to the quadrants and octants on the VA plane in Fig. 2).

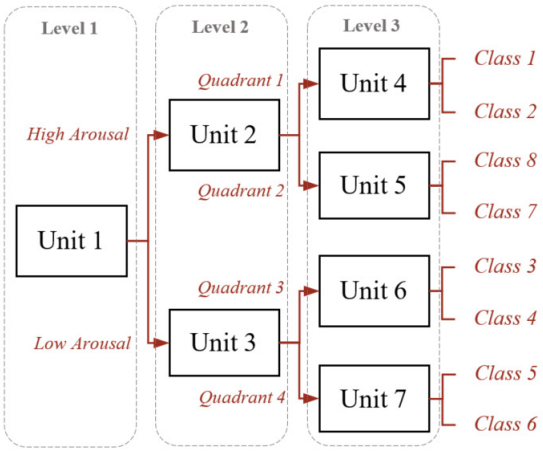

Fig.9: *Cascaded Structure of the Predictive Model.*

Table 4: *Training Dataset and Purpose of Each Unit.*

Unit	Trained With	To Discriminate	Number of Training Samples
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

6.3 Model Configuration

The dataset was divided into three parts to evaluate the performance of the models: 15% of the data was randomly selected for 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 100%. The result on the testing set was completely independent of the training process.

Parameter adjustments 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. We defined the acceptable error rate as zero and allowed the maximum number of epochs, and the elapsed time to be infinite. The mean squared error (MSE) is set to be a loss function for both algorithms. Normally, cross entropy is a loss function for classification tasks, but our input data are continuous numbers. Therefore, the MSE is more suitable for our data. We concentrated on the minimum gradient and maximum failing adjustment as the stopping criteria for the training process and left the other parameters at their default values. The models in the same depth of cascaded structures were constructed with the parameter values reported in Table 5. 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. The list of parameters and their default values can be found in [51].

The architecture of neural networks is "Input node - Hidden node - Output node". We set the number of input nodes as the number of extracted features, 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. Thus, the number of desired outputs was 2. The desired output of the Rprop algorithm was the specific class, which differed between the multiclass model and the cascaded model. The desired number of output classes of the multiclass model was 8, while that of the cascaded model was 2, as previously noted.

Table 5: *Parameter Configuration.*

Model	Architecture	Min Gradient	Max Failing
Multiclass LM	122-10-2	1.0E-07	10
Lv. 1 of Cascaded LM	122-10-2	1.0E-07	10
Lv. 2 of Cascaded LM	122-10-2	1.0E-07	10
Lv. 3 of Cascaded LM	122-10-2	1.0E-21	10
Multiclass Rprop	122-10-8	1.0E-50	300
Lv. 1 of Cascaded Rprop	122-10-2	1.0E-50	300
Lv. 2 of Cascaded Rprop	122-10-2	1.0E-100	300
Lv. 3 of Cascaded Rprop	122-10-2	1.0E-100	300

There is uncertainty in the results when the network is retrained because of the variation in the randomly selected parameters such as weight, bias, and sample selection of subdatasets. Therefore, the network must be retrained several times to reduce the variation in the results. We performed retraining 11,000 times and chose the most accurate model.

7. RESULTS

The performance of model training is reported in Table 6. Accuracies of the training step, validation step, testing subdataset, and entire dataset are separately reported in the training dataset accuracy (Acc. Train), validation dataset accuracy (Acc. Val.), testing dataset accuracy (Acc. Test) and entire dataset accuracy (Acc. Entire) columns, respectively. The Min MSE column shows the minimum mean square error of each model unit. How many epochs, repetitions, and how long each unit takes to train the model are reported in the Best Epoch, Best Round, and Time columns, respectively. The confusion matrices in Fig. 10 show the prediction results of each model. Numbers 1 to 8 refer to the emotional classes in Table 1. Each cell shows the density of the population in terms of number and percentage. The sum of all cell numbers is equal to the sample size of the dataset, and the percentages sum to 100%. The sum of the numbers in the vertical cells is the total number of samples in each class. The sum of the numbers in the horizontal cells is the total number of predicted classes. The diagonal cells from the top-left to the bottom-right show the correct predictions for each class. The accuracy is shown at the bottom-right corner of the matrices. The bottom row of the matrices shows the recall. The right column of the matrices shows the precision.

Table 7 and Fig. 11 show the evaluation of prediction performance using the F1-score. The overall prediction performance based on the number of samples (with more training samples resulting a higher score) can be seen by comparing Fig. 11 with Fig. 3 (left). Prediction is problematic in classes. 3, 4, and 7 because these classes have fewer samples than the other classes.

With these limited samples, the feature patterns may not be concrete enough to form a strong prediction on these classes. In contrast, classes. 1, 2, and 5, which have an overwhelming number of training samples, obtained much higher scores than other classes.

The prediction accuracy for the cascaded LM model offers no significant improvement over that of the multiclass LM Thus, the LM algorithm is not suitable for a cascaded structure. However, the direct effect of the cascaded structure is obvious when using the Rprop algorithm, the F1-score of which is the highest in every class prediction.

8. DISCUSSION

We compared the performance of our work with that of other approaches. The progress of development and improvement can be observed in Table 8 by sorting the studies by time and the five employed KPIs. 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 predictions to make our re-

Table 6: *Performance Following Model Training.*

Model/Submodel	Acc. Train(%)	Acc. Val. (%)	Acc. Test(%)	Min MSE	Best Epoch	Best Round	Time (h)	Acc. Entire	F1-score
Multiclass LM	50.6	42.2	40.0	0.0327	7	1	6.1	47.8	See Table 7
Binary-Class LM Unit 1	85.9	86.7	84.8	0.0319	5	2	7.0	85.8	0.8980
Binary-Class LM Unit 2	83.9	81.7	84.9	0.0183	1	8	6.1	83.7	0.9055
Binary-Class LM Unit 3	76.8	81.0	81.0	0.0118	1	27	4.1	78.0	0.6886
Binary-Class LM Unit 4	69.9	67.1	64.5	0.0116	1	2,018	4.8	68.6	0.7531
Binary-Class LM Unit 5	78.8	79.4	64.7	0.0065	1	946	2.9	76.8	0.6667
Binary-Class LM Unit 6	79.0	63.3	70.0	0.0062	1	1,084	2.0	75.3	0.7656
Binary-Class LM Unit 7	71.7	59.3	74.1	0.0142	1	24	3.8	70.2	0.7882
Multiclass Rprop	59.6	62.2	52.2	0.2759	8	6	2.8	58.9	See Table 7
Binary-Class Rprop Unit 1	96.7	94.8	96.3	0.1361	18	3,455	2.6	96.3	0.9737
Binary-Class Rprop Unit 2	96.6	95.2	96.2	0.1141	4	3,074	2.2	96.3	0.9774
Binary-Class Rprop Unit 3	98.7	97.6	96.4	0.1369	1	3,484	1.6	98.2	0.9749
Binary-Class Rprop Unit 4	95.6	95.4	91.4	0.1349	2	10,603	1.6	95.0	0.9574
Binary-Class Rprop Unit 5	96.9	97.1	85.3	0.1397	1	122	0.3	95.2	0.9308
Binary-Class Rprop Unit 6	96.4	100.0	96.7	0.1250	19	9,950	0.1	97.0	0.9706
Binary-Class Rprop Unit 7	95.7	96.3	94.4	0.1435	1	357	0.1	95.6	0.9644

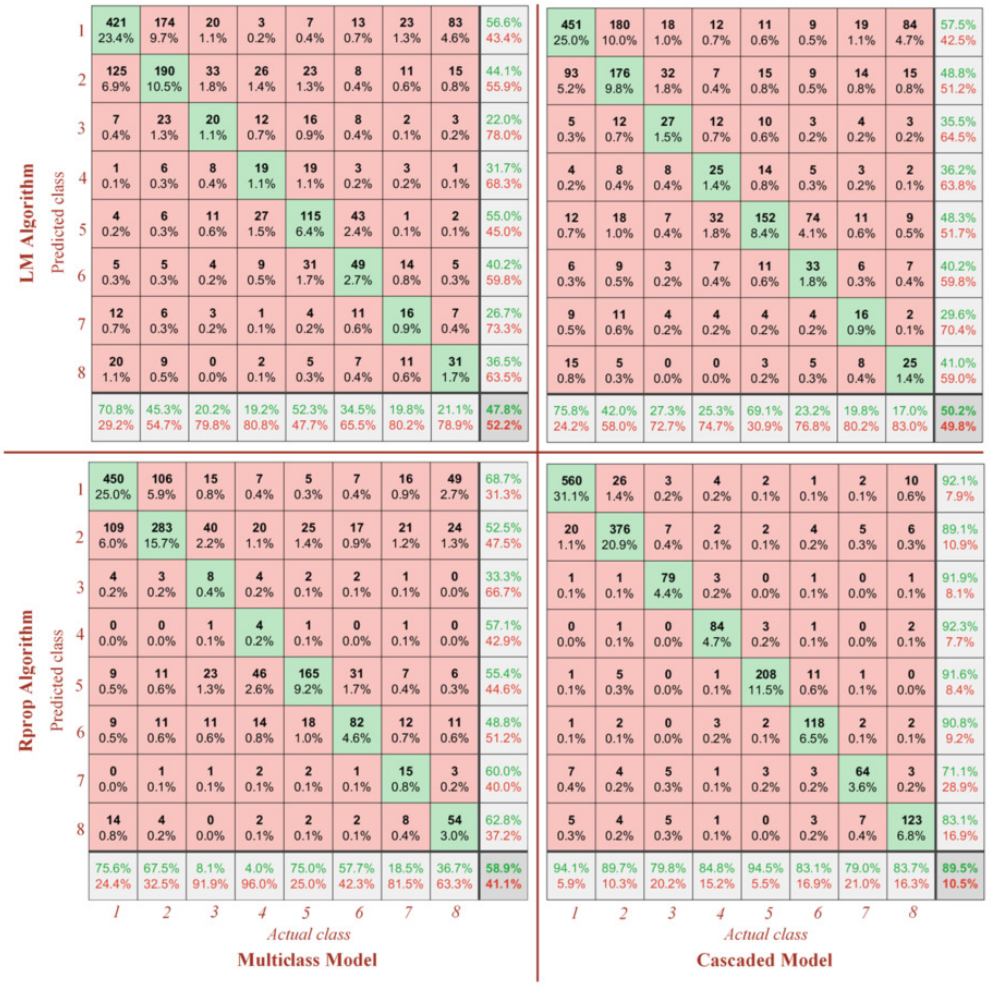

Fig.10: *The Comparison of Eight Emotion Prediction Results with the Multiclass LM Network (top left), Cascaded LM Network (top right), Multiclass Rprop Network (bottom left), and Cascaded Rprop Network (bottom right).*

Table 7: *The Evaluation of Prediction in Each Class using the F1-Score.*

Model	Class 1 F1-score	Class 2 F1-score	Class 3 F1-score	Class 4 F1-score	Class 5 F1-score5	Class 6 F1-score	Class 7 F1-score	Class 8 F1-score	Average F1-score	Acc.
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%
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%

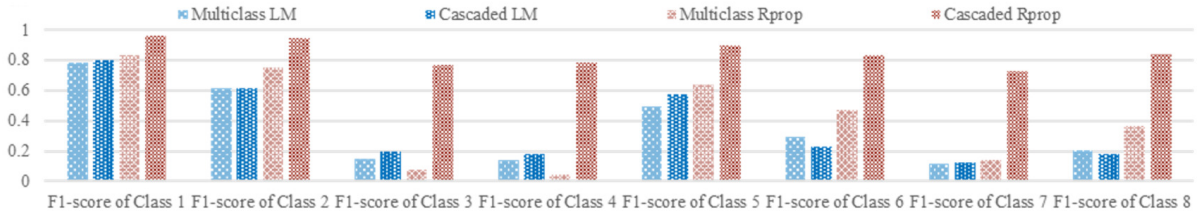

Fig.11: *The Evaluation of Prediction in Each Class using the F1-Score.*

sults comparable. When we measured our accuracy with only valence and arousal level (level 1), the results were similar to those of every study that measured their results in terms of valence/arousal level. Arousal prediction is always better than valence prediction, as observed in experiments 1, 9, 13 to 16, 28 and 29.

Among the four emotional classification works, in our work, when we measured our accuracy at level 2, nos. 30 to 32 we achieved average performance, but no. 33 outperforms the others. Experiments 8 and 10 are comparable to experiments 31 and 33. 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 24 to 27 were conducted based on the same dataset we used. However, the number of samples associated with each emotion were equalized, and the performance metric was the F1-score. Their achievements are comparable to our F1-score as shown in Table 7.

Among the eight emotional classification works, our methods (nos. 34 to 37) are comparable to those of experiment 11, which also employed multiple models. However, no. 11 simply employed eight models to regress each class simultaneously without a structure and their output format is "multilabel of multiclass", while nos. 35 and 37 employed a cascaded structure of seven models to discriminate two classes at a time and our output format is "single-label of multiclass".

The result of experiment no. 37 is outstanding because it took advantage of two aspects. First, is used the Rprop neural network which is designed for classes output. Second, we arranged of these Rprop units by using seven cascading units of the binary-class model in which each unit performs only one simple task. This reduces the workload of each unit and gains accuracy. When the accuracy increases unit by unit, level by level, the overall result can be substantially increased. This result confirms our hypothesis that using multiple model units can improve the overall performance.

However, there are two drawbacks associated with using cascaded structures. The first drawback is the complexity of the cascaded model. During training, we must separately configure each submodel (seven submodels in total), which might take at least seven times as long as the traditional multiclass model. For the three-level structure, 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 could be input while the previous sample was being 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, the programming effort to connect each submodel to form a cascaded structure can be complicated.

The second drawback is that the 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 seen by comparing the individual accuracies of the submodels (Table 6) and the final accuracies (Table 7). For example, in the cascaded Rprop model, even though the average accuracy of the submodels is 96%, the final accuracy is only 89.5%, as each percent loss in each submodel had a different effect.

In the simulation, 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, so a loss of 20% of unit 3 corresponds to 112 samples, or approximately 6.3% of the entire dataset (1,802). 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, so a loss of 20% of unit 6 corresponds to 40 samples, or 2.2% of the entire dataset.

There are two drawbacks associated with the MIR toolbox: 1) the time needed for feature extraction and 2) the large consumption of memory. However, with a powerful machine, we are able to run multiple model training and feature extraction tasks. Therefore, the total elapsed time is not the sum of the reported elapsed times in Tables 2 and 6.

9. CONCLUSION

This paper proposes a MER system using eight emotion classes. We evaluated the system with the DEAM benchmark. The dataset was divided in a 7/3 ratio (training set/testing set). The system was im-

Table 8: *Comparison of the KPIs in Each Work.*

♯	Ref. No.	Proposed Year	Algorithms and/or Methods	No. of Samples	No. of Features	No. of Classes	Achievement(%)	Measurement Method
1	[15]	2008	SVM Regressor with RReliefF feature selection	195	114	V/A	28.1/58.3	R^2
2	[5]	2008	Binary Relevance	593	74	6	73.8	Accuracy
3	[5]	2008	Label Powerset	593	74	6	76.7	Accuracy
4	[5]	2008	Random k-Labelsets	593	74	6	79.5	Accuracy
5	[5]	2008	Multilabel k-Nearest Neighbor	593	74	6	71	Accuracy
6	[6]	2013	Auto Associative Neural Networks	85	52	5	94.4	Accuracy
7	[6]	2013	Support Vector Machine	85	52	5	85	Accuracy
8	[7]	2013	Hierarchical SVM based on tempo & mutation degree	80	2	4	95	Accuracy
9	[24]	2014	Recurrent Neural Networks	1,000	70	V/A	50/70	R^2
10	[2]	2014	Hierarchical SVM	219	35	4	89.6	Accuracy
11	[8]	2015	Eight Regressors, individually trained	385	117	8	59.4	Accuracy
12	[9]	2015	Nearest multiprototype classifier	903	59	5	56.4	Accuracy
13	[25]	2016	Adaptive Aggregation of Gaussian Process Regressors	744	65	V/A	77/80	RMSE
14	[26]	2017	Stacked CNN & RNN	431	260	V/A	73/80	RMSE
15	[27]	2017	Random Forest	300	397	V/A	57.3/70	Accuracy
16	[28]	2017	Support Vector Regressor	818	539	V/A	25/79	Accuracy
17	[28]	2017	Support Vector Machine	818	539	6	85	Accuracy
18	[3]	2017	k-Nearest Neighbors	1,000	548	4	62	Accuracy
19	[3]	2017	Bayes classifier	1,000	548	4	69	Accuracy
20	[3]	2017	Linear Discriminant Analysis	1,000	548	4	80.4	Accuracy
21	[3]	2017	N euro-Fuzzy Network Classification	1,000	548	4	79.3	Accuracy
22	[3]	2017	Fuzzy KNN 1,000	548	4	83	Accuracy	
23	[3]	2017	Support Vector Machine	1,000	548	4	82.7	Accuracy
24	[4]	2017	Support Vector Machine	943	33	4	46	F1-score
25	[4]	2017	Naïve Bayes	943	33	4	40	F1-score
26	[4]	2017	Decision Trees	943	33	4	37	F1-score
27	[4]	2017	k-Nearest Neighbors	943	33	4	41	F1-score
28	|	2018	Binary-Class LM Neural Networks	1,802	122	V/A	78.9/85.8	Accuracy
29	|	2018	Binary-Class Rprop Neural Networks	1,802	122	V/A	76.7/96.3	Accuracy
30	|	2018	Multiclass LM Neural Networks	1,802	122	4	70.6	Accuracy
31	|	2018	Cascaded LM Neural Networks	1,802	122	4	73.4	Accuracy
32	This	2018	Multiclass Rprop Neural Networks	1,802	122	4	79	Accuracy
33	Work	2018	Cascaded Rprop Neural Networks	1,802	122	4	93.5	Accuracy
34	|	2018	Multiclass LM Neural Networks	1,802	122	8	47.8	Accuracy
35	|	2018	Cascaded LM Neural Networks	1,802	122	8	50.2	Accuracy
36	|	2018	Multiclass Rprop Neural Networks	1,802	122	8	58.9	Accuracy
37	|	2018	Cascaded Rprop Neural Networks	1,802	122	8	89.5	Accuracy

plemented in MATLAB. 122 acoustic features were extracted by the MIR toolbox, and four model training methods were investigated: multiclass LM, cascaded LM, multiclass Rprop, and cascaded Rprop. The corresponding accuracies were 47.8%, 50.2%, 58.9%, and 89.5%. The results for the cascaded Rprop model confirm the scalability of prediction with multimodel methods demonstrated in previous work [2][7][8]. We also found that features. 29, 48, 42, 43, and 49 (from Table 3) had higher impacts than the other features for music emotion prediction.

Future work should investigate how to index with the minimum cost in terms of computational time and hardware requirements, while maintaining acceptable accuracy. This might be achieved by, for example, considering only high-impact features. or applying feature selection or optimization methods before training the models.

Our work aims to encourage music providers to categorize music using emotional terms to improve search efficiency and access to music. This work may lead to additional applications, such as music playlist generation based on the listener's heart rate and automatic stage-lighting control based on music emotions [53][54].

References

[1] Y.-H. Yang and H. Chen, "01 Introduction," in *Music Emotion Recognition*, CRC Press, 2011, pp. 1–13.

[2] 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.

[3] J. Bai et al., "Music Emotion Recognition by Cognitive Classifcation Medthodologies," in *2017 IEEE 16th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC)*, 2017, pp. 121–129.

[4] P. M. F. Vale, "The Role of Artist and Genre on Music Emotion Recognition," Universidade Nova de Lisboa, 2017.

[5] 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.

[6] 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.

[7] 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.

[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] 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.

[10] S. R. Livingstone and A. R. Brown, "Dynamic Response: Real-Time Adaptation for Music Emotion," in *2nd Australasian Conference on Interactive Entertainment*, 2005, pp. 105–111.

[11] S. R. Livingstone, A. R. Brown, and R. Muhlberger, "Influencing the Perceived Emotions of Music with Intent," in *Proceedings of the Third International Conference on Generative Systems in the Electronic Arts*, 2005, pp. 161–170.

[12] S. R. Livingstone and W. F. Thompson, "Multimodal Affective Interaction A Comment on Musical Origins," *Music Percept.*, vol. 24, no. 1, pp. 89–94, 2006.

[13] S. R. Livingstone, R. Mühlberger, A. R. Brown, and A. Loch, "Controlling Musical Emotionality: An Affective Computational Architecture for Influencing Musical Emotions," *Digit. Creat.*, vol. 18, no. 1, pp. 43–53, 2007.

[14] "Audioblocks.com." [Online]. Available: `https://www.audioblocks.com/.` [Accessed: 08-Mar-2018].

[15] 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.

[16] K. Hevner, "Expression in Music: A Discussion of Experimental Studies and Theories," *Psychol. Rev.*, vol. 42, no. 2, pp. 186–204, 1935.

[17] J. A. Russell, "A Circumplex Model of Affect," *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980.

[18] J. A. Russell, "Culture and The Categorization of Emotions," *Psychol. Bull.*, vol. 110, no. 3, pp. 426–450, 1991.

[19] T. Eerola and J. K. Vuoskoski, "A Comparison of The Discrete and Dimensional Models of Emotion in Music," *Psychol. Music*, vol. 39, no. 1, pp. 18-49, 2011.

[20] 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.

[21] Y. Yang and H. H. Chen, "Machine Recognition of Music Emotion: A Review," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 40, 2012.

[22] 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)*, pp. 1505–1508, 2008.

[23] 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.

[24] F. Weninger, F. Eyben, and B. Schuller, "Online 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.

[25] S. Fukayama and M. Goto, "Music Emotion Recognition With Adaptive Aggregation of Gaussian Process Regressors," *Icassp 2016*, pp. 71–75, 2016.

[26] 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.`

[27] V. L. Nguyen, D. Kim, V. P. Ho, and Y. Lim, "A New Recognition Method for Visualizing Music Emotion," *Nguyen2017ANR*, vol. 7, no. 3, pp. 1246–1254, 2017.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] "Free Music Archive." [Online]. Available:

http://freemusicarchive.org/ [Accessed: 07-Nov-2018].

[35] "Jamendo Music." [Online]. Available: https://www.jamendo.com/ [Accessed: 07-Nov-2018].

[36] 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.

[37] 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.

[38] 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].

[39] "DEAM Dataset Release Page." [Online]. Available: http://cvml.unige.ch/databases/DEAM [Accessed: 01-Jun-2017].

[40] O. Lartillot and P. Toiviainen, "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.

[41] T. Eerola and P. Toiviainen, "MIR in MATLAB: The Midi Toolbox," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2004, pp. 22–27.

[42] 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].

[43] [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.

[44] H. Bowen et al., Eds., The Element of Classical Music," in *The Compleate Classical Music Guide*, 1st ed., New york: Dorling Kindersley Limited, 2012, pp. 10–17.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] M. Riedmiller, M. Riedmiller, and H. Braun, "RPROP - A Fast Adaptive Learning Algorithm," in *PROC. OF ISCIS VII*, UNIVERSITAT, 1992.

[50] M. Riedmiller and H. Braun, "A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm," in *IEEE International Conference on Neural Networks*, 1993, pp. 586–591.

[51] M. H. Beale, M. T. Hagan, and H. B. Demuth, *Neural Network Toolbox^{TM} Reference*. 3 Apple Hill Drive: Mathwork Inc., 2017.

[52] M. H. Beale, M. T. Hagan, and H. B. Demuth, *Neural Network ToolboxTM User's Guide*. 3 Apple Hill Drive: Mathwork Inc., 2017.

[53] 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)*, vol. 1, pp. 545–549, 2009.

[54] 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.

Kanawat Sorussa received his B.Eng. (Computer Engineering) degree from Rajamangala University of Technology Srivijaya, Songkhla, Thailand, in 2016. In December 2017, he participated in the 5th International Conference on Information Technology (ICIT) at Nanyang Technological University in Singapore; his paper became the basis of this work. He is studying for a master's degree in the Department of Computer Engineering, Faculty of Engineering, Prince of Songkla University, Songkhla, Thailand. His research interests are music retrieval, information retrieval, data mining, and machine learning.

Anant Choksuriwong received his bachelor's and two master's degrees in 2000 (PSU), 2003 (UJF) and, 2004 (INPG) and his Ph.D. degree from the School of Engineering ENSI de Bourges, France, in 2008. He is a researcher in the Laboratory of Computer Engineering PSU, Songkhla, Thailand. In November 2008, he became a lecturer in the Department of Computer Engineering, Prince of Songkla University (PSU), where he teaches courses in advanced image processing, machine learning, and robotics principles.

His research is on cognitive systems engineering. He is particularly interested in machine learning (Bayesian networks, probabilistic programming), computer vision (object detection and recognition), image processing (global and local invariant feature extraction), mobile robotics (autonomous position-sensing and navigation), and perception and multisensor data fusion with Bayesian inference.

Montri Karnjanadecha received his B.Eng. and M.Eng. degrees in Electrical Engineering from Prince of Songkla University, Thailand, in 1990 and 1995, respectively. He received his Ph.D. degree in Electrical Engineering from Old Dominion University, Virginia, USA, in 2000. He has been a faculty member of the Department of Computer Engineering, Faculty of Engineering, Prince of Songkla University, Thailand, since 1990. He is currently an associate professor.

His research interests include digital signal processing, speech modeling, speech recognition, speaker recognition, hardware implementation for speech processing and image processing, digital control systems, and robotics.