

**A FEATURE-BASED DUAL-LAYER ENSEMBLE
CLASSIFICATION METHOD FOR
EMOTIONAL STATE RECOGNITION**

MEHDI MALEKZADEH

**FACULTY OF COMPUTER SCIENCE AND
INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR**

2017

**A FEATURE-BASED DUAL-LAYER ENSEMBLE
CLASSIFICATION METHOD FOR
EMOTIONAL STATE RECOGNITION**

MEHDI MALEKZADEH

**THESIS SUBMITTED IN FULFILMENT
OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY**

**FACULTY OF COMPUTER SCIENCE AND
INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR**

2017

UNIVERSITY OF MALAYA
ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: **MEHDI MALEKZADEH**

Registration/Matric No: **WHA110029**

Name of Degree: **DOCTOR OF PHILOSOPHY**

Title of Project Paper/Research Report/Dissertation/Thesis (“this Work”):

**A FEATURE-BASED DUAL-LAYER ENSEMBLE CLASSIFICATION
METHOD FOR EMOTIONAL STATE RECOGNITION**

Field of Study: **Software Engineering - Human Computer Interaction**

I do solemnly and sincerely declare that:

- (1) I am the sole author/writer of this Work;
- (2) This Work is original;
- (3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
- (4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
- (5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya (“UM”), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
- (6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate’s Signature

Date:

Subscribed and solemnly declared before,

Witness’s Signature

Date:

Name:

Designation:

ABSTRACT

It is important to recognize an individual's emotional state as it can be used in many disciplines and research in areas such as medicine and education. Human emotions can be recognized through the analysis of several modalities, which include speech, facial appearance, gestures, and human physiology. Among the different modalities of human emotion expression, the physiological data that can be gathered from people, especially the speech impaired people is probably the most reliable for human emotion recognition. The physiological modality has the advantage of being more robust against possible artifacts of human interpersonal hiding since they will be instantaneously managed by the human autonomic nervous system.

The current automatic physiological-based emotion recognition systems call for improvement in two main respects which are applying a feature selection method for selecting an optimal feature subset and selecting a suitable classifier that maximizes the classification performance of the emotion recognition system. The main aim of this research is to improve the classification accuracy of physiological-based emotion recognition systems by proposing a feature-based dual-layer ensemble classification method. In addition, we analyse the accuracies of various classification methods with different physiological modalities and feature selection methods in order to understand the effect of each component on the overall performance of the emotion recognition system and recommend a system's design that can achieve the best classification accuracy for emotion recognition systems.

The results show that for single classifiers, Support Vector Machine (SVM) achieved the best classification method to be used for developing emotion recognition system and there is no single type of modality that is suitable for all the classifiers. In addition, feature selection methods have positively contributed to the improvement of multi-classifier methods compared to single classifiers. Compared to the best single classifiers, the

proposed feature-based dual-layer ensemble classification method has improved the accuracy around 5% to 17%. The proposed classification method can be used or tested on other emotion databases or even on other medical diagnosis problems that use physiological data.

University of Malaya

ABSTRAK

Adalah penting untuk mengenali keadaan emosi individu kerana ia boleh digunakan dalam pelbagai disiplin dan penyelidikan seperti bidang perubatan dan pendidikan. Emosi manusia boleh dikenali melalui analisis beberapa modaliti termasuk ucapan, penampilan wajah, gerak isyarat, dan fisiologi manusia. Di antara modaliti yang berbeza untuk ekspresi emosi manusia, data fisiologi yang boleh dikumpul daripada manusia, terutamanya individu dengan masalah pertuturan mungkin adalah yang paling berkesan untuk pengecaman emosi manusia. Modaliti fisiologi mempunyai kelebihan untuk menjadi lebih kuat terhadap artifak interpersonal manusia yang bersembunyi kerana mereka akan serta-merta diuruskan oleh sistem saraf autonomi manusia.

Sistem pengecaman emosi berasaskan fisiologi secara automatik masakini memerlukan penambahbaikan dalam dua perkara utama iaitu menggunakan kaedah pemilihan ciri untuk memilih subset ciri optimum dan memilih pengelas yang sesuai untuk memaksimumkan prestasi pengelasan sistem pengecaman emosi. Tujuan utama penyelidikan ini adalah untuk meningkatkan ketepatan pengelasan sistem pengecaman emosi berasaskan fisiologi dengan mencadangkan kaedah pengelasan ensemble lapisan dwi berasaskan ciri. Di samping itu, kami menganalisis ketepatan pelbagai kaedah pengelasan dengan kaedah modaliti fisiologi dan kaedah pemilihan ciri yang berbeza untuk memahami kesan setiap komponen terhadap prestasi keseluruhan sistem pengecaman emosi dan mencadangkan reka bentuk sistem yang boleh mencapai ketepatan pengelasan yang terbaik untuk sistem pengecaman emosi.

Keputusan menunjukkan bahawa untuk pengelas tunggal, Mesin Vektor Sokongan (SVM) mencapai kaedah pengelasan yang terbaik untuk digunakan bagi membangunkan sistem pengecaman emosi dan tidak ada modaliti tunggal yang sesuai untuk semua pengelas. Di samping itu, kaedah pemilihan ciri telah memberi sumbangan positif kepada peningkatan kaedah multi-pengelasan berbanding pengelasan tunggal. Berbanding

dengan pengelas tunggal yang terbaik, kaedah pengelasan ensemble dwi lapisan beasaskan ciri yang dicadangkan telah meningkatkan ketepatan sekitar 5% hingga 17%. Kaedah pengelasan yang dicadangkan boleh digunakan atau diuji pada pangkalan data emosi lain atau juga pada masalah diagnosis perubatan lain yang menggunakan data fisiologi.

University of Malaya

ACKNOWLEDGEMENTS

First and foremost, praises and thanks to God, the Almighty, for his blessings, protection, and guidance throughout my research work to complete this Ph.D. successfully.

This thesis would not have been successfully completed without the guidance and the help of several people who contributed their valuable assistance during the preparation and the development of the thesis work. I would like to express my deep and sincere gratitude to my Research Supervisors Prof Dr. Siti Salwah Salim and Dr. Mumtaz Begum for their full support, expert guidance, understanding and encouragement throughout my Ph.D. research journey. Without their incredible patience and timely wisdom and counsel, my research would have been an overwhelming pursuit.

I would also like to express my thanks to Dr. Adel Lahsasna for his motivation, guidance, support and patience during my study. Thanks also go to all my fellow friends in Human-Computer Interaction (HCI) lab and Multimodal Interaction (MMI) lab, University of Malaya and all people involved directly and indirectly upon completing this research.

Last but not least, I would like to express my deepest appreciation to my parents for their understanding, inspiration, love, and patience, and for putting up with me at all times of great distress.

TABLE OF CONTENTS

Abstract	iii
Abstrak	v
Acknowledgements	vii
Table of Contents	viii
List of Figures	xiii
List of Tables	xvii
List of Symbols and Abbreviations	xx
List of Appendices	xxi
CHAPTER 1: INTRODUCTION	1
1.1 Introduction	1
1.2 Research Background	2
1.2.1 Physiological Signals	2
1.2.2 Human Emotions	5
1.2.3 Physiological Signal based Emotion Recognition	7
1.3 Research Motivation	10
1.4 Problem Statement	11
1.5 Research Objectives	13
1.6 Research Questions	14
1.7 Research Scope	15
1.8 Research Methodology	16
1.9 Thesis Outline	19
CHAPTER 2: LITERATURE REVIEW	21
2.1 Introduction	21

2.2	Theories of Emotion	21
2.2.1	James-Lange Theory	22
2.2.2	Cognitive Appraisal Theory	22
2.2.3	Universality of Basic Emotions	23
2.3	Discrete and dimensional emotion models	23
2.4	Emotion Detection Modalities.....	25
2.5	Methods in Multimodal Emotion Recognition.....	28
2.6	Psychophysiology of Emotion	29
2.6.1	The Electroencephalogram (EEG).....	30
2.6.2	Autonomic Nervous System (ANS) Measures and Its applicability for Emotion Recognition	32
2.6.2.1	Respiration (RSP).....	34
2.6.2.2	Electrocardiogram (ECG).....	34
2.6.2.3	Electromyogram (EMG).....	35
2.6.2.4	Skin Conductivity (SC)	36
2.7	Review on Physiological-based Emotion Recognition System	36
2.7.1	Physiological data preparation.....	38
2.7.1.1	Emotion model selection	38
2.7.1.2	Physiological data collection	39
2.7.1.3	Pre-processing.....	48
2.7.1.4	Feature Extraction	48
2.7.1.5	Normalization.....	49
2.7.1.6	Feature Dimension Reduction.....	52
2.7.2	Classification	58
2.7.2.1	Classification performance evaluation	59
2.7.2.2	Nearest Neighbors	59

2.7.2.3	Naïve Bayes classifier	63
2.7.2.4	Discriminant Analysis	66
2.7.2.5	Support vector machines.....	66
2.7.2.6	Classification Trees	73
2.7.2.7	Artificial neural networks	73
2.7.2.8	Ensemble Classification.....	78
2.7.3	Discussion on classification techniques	89
2.8	Concluding Marks	90
 CHAPTER 3: THE PROPOSED CLASSIFICATION METHOD.....		93
3.1	The main finding of literature review	93
3.2	Steps involved in the design and development of a physiological-based emotion recognition system.....	95
3.3	Data set selection and preparation	97
3.3.1	General description	97
3.3.2	Features extracted from physiological Signals.....	99
3.3.2.1	Peripheral features	100
3.3.2.2	Electroencephalogram (EEG) features	103
3.4	Design emotion recognition system using benchmark classifiers.....	106
3.4.1	LDA	106
3.4.2	CART	106
3.4.3	ANN.....	107
3.4.4	SVM.....	108
3.5	Design the Feature-based Multi-Classifer methods	109
3.5.1	Feature selection methods	111
3.5.2	Majority vote method.....	114

3.6	Design the proposed Feature-based Dual-Layer Ensemble Classification Method	114
3.7	Evaluation	122
3.8	Experimental setups.....	123
3.8.1	Parameter specifications of the classification methods and feature selection methods	124
3.8.2	Classification Performance Evaluation	125
3.8.2.1	Testing classification accuracy rate calculation.....	125
3.8.2.2	Average Rank.....	126
3.8.2.3	Statistical test	127
3.9	Experiment 1: Comparative analysis of CARs of Benchmark classifiers	129
3.10	Experiment 2: Comparative analysis of CARs of Feature-based Multi-classifier methods	129
3.11	Experiment 3: Comparative analysis of CARs of Feature-based Dual-layer ensemble classifiers	130
3.12	Summary	132
CHAPTER 4: RESULTS AND DISCUSSIONS.....		133
4.1	The results of benchmark classification methods.....	133
4.1.1	Classification accuracy rates of the four benchmark classifiers based on the modality used.....	138
4.1.2	Results obtained by the classifiers based on given modality	141
4.2	Results obtained after applying feature-based multi-classifier methods	143
4.2.1	Comparison between single classifiers and feature-based multi-classifier using different modalities.....	146
4.2.2	Comparison between different feature subset sizes in terms of accuracy	

4.3	The results of proposed classification method: Feature-Based Dual-layer Ensemble Classification Method	151
4.4	Comparison between feature-based dual-layer ensemble classification methods and other classifiers	157
4.4.1	Physiological modalities performance comparison for emotional state recognition	163
4.5	Comparison between classifiers using Statistical method	166
4.6	Comparison between existing works	167
4.7	Conclusion.....	172
CHAPTER 5: CONCLUSIONS		173
5.1	Research objectives revisited	173
5.1.1	Research objective 1	173
5.1.2	Research objective 2	174
5.1.3	Research objective 3	175
5.2	Research Contribution	178
5.3	Research Limitation.....	179
5.4	Suggestions for future works.....	180
	References	181
	List of Publications and Papers Presented	200
	Appendix A: The samples of code used for feature extraction	201
	Appendix B: Classification accuracy results of valence, arousal, and linking recognition using different feature subset sizes	207

LIST OF FIGURES

Figure 1.1: Biosensors locations on the body and the associated waveforms: (a) ECG. (b) RSP. (c) SC. (d) EMG (Kim & André, 2008)	5
Figure 1.2: Two-dimensional model of emotion based on valence and arousal (Kim & André, 2008).....	7
Figure 1.3: The main components of an automatic physiological-based emotion recognition system (Novak, Mihelj, & Munih, 2012)	7
Figure 1.4: The DSRM research method adopted for this research	18
Figure 2.1: Emotional experiences described in two dimensions valence and arousal, accompanied by distribution of some discrete prototypes of emotions along the two dimensions (adopted from (Russell & Barrett, 1999, p.808))	24
Figure 2.2: The framework of a multimodal emotion recognition system (Tao & Tan, 2005)	27
Figure 2.3: Divisions of human nervous system (Andreassi, 2007).....	30
Figure 2.4: The four major types of EEG waves and associated human state (Adopted from (Webster & Clark, 2010))	32
Figure 2.5: The general process to create an automated physiological-based emotion recognition system (Novak, Mihelj, & Munih, 2012)	37
Figure 2.6: Images used by Self-Assessment Manikin (SAM) method for the individual self-report emotional state. Self-evaluation scales for the dimensions of valence (top), arousal (bottom) (Adopted from Hettich et al., 2016).	42
Figure 2.7: The concept of an ensemble classification (Witten, Frank, & Hall, 2011)..	78
Figure 2.8: The concept of stacking ensemble classification (Witten et al., 2011).....	86
Figure 3.1: The steps followed to design and develop the proposed feature-based dual-layer ensemble classification method	96
Figure 3.2: The SAM used to rank the emotion dimension of valence (top), arousal (second) and dominance (third) of subjects. Thumbs up/down (last) to scale liking (Koelstra et al., 2012).....	98
Figure 3.3: Distribution of emotion ranking for 40 videos by subject1 based on two-dimensional (valence–arousal) emotion model	99
Figure 3.4: A graphical representation of a typical feedforward neural network	108

Figure 3.5: The method for creating feature-based multi-classifier method.....	110
Figure 3.6: Concept of the proposed feature-based dual-layer ensemble classification method.....	118
Figure 3.7: The pseudocode related to the first layer of the proposed dual-layer classification method	119
Figure 3.8: The pseudocode related to the second layer of the proposed dual-layer classification method	120
Figure 3.9: A sequence of activities for proposed dual-layer ensemble classification methods.	121
Figure 3.10: The steps followed to evaluate the proposed feature-based dual-layer ensemble classification method	123
Figure 4.1: Average ranking score of four benchmark classification methods.....	135
Figure 4.2: Average accuracy rates of four benchmark classification methods for valence, arousal and liking targets using peripheral modality	135
Figure 4.3: Average accuracy rates of four benchmark classification methods for valence, arousal and liking targets using EEG modality	136
Figure 4.4: Average accuracy rates of four benchmark classification methods for valence, arousal and liking targets using (Peripheral+EEG) modalities	137
Figure 4.5: Average ranking score of each modality using four benchmark classifiers	139
Figure 4.6: Average ranking score of each modality using four benchmark classifiers	139
Figure 4.7: Average testing rates of the best feature-based multi-classifier methods for Peripheral modality.....	144
Figure 4.8: Average testing rates of the best feature-based multi-classifier methods for EEG modality	145
Figure 4.9: Average testing rates of the best feature-based multi-classifier methods for (Peripheral+EEG) modalities	145
Figure 4.10: Average classification accuracies for each classifier on valence, arousal, and liking using Peripheral modality	147
Figure 4.11: Average classification accuracies for each classifier on valence, arousal, and liking using EEG modality.....	148

Figure 4.12: Average classification accuracies for each classifier on valence, arousal, and liking using (Peripheral+EEG) modalities	148
Figure 4.13: Average accuracy ranking scores of Liking, Arousal and Valence using different subset sizes	150
Figure 4.14: The best accuracy rates among the feature-based dual-layer ensemble classification methods	152
Figure 4.15: Average ranking scores of the feature-based dual-layer ensemble classification methods over the three modalities	153
Figure 4.16: Testing accuracy rates of feature-based dual-layer ensemble classification methods on Valence, Arousal and liking with peripheral modality	154
Figure 4.17: Testing accuracy rates of feature-based dual-layer ensemble classification methods on Valence, Arousal and liking with EEG modality	155
Figure 4.18: Testing accuracy rates of feature-based dual-layer ensemble classification methods on Valence, Arousal and liking with (Peripheral+EEG) modalities	156
Figure 4.19: Average ranking scores of the best classifiers taken from three different categories	160
Figure 4.20: The best testing accuracy rates of each of the three categories for peripheral modality	161
Figure 4.21: The best testing accuracy rates of each of the three categories for EEG modality	161
Figure 4.22: The best testing accuracy rates of each of the three categories for (Peripheral +EEG) modalities	162
Figure 4.23: The comparison of testing accuracy rates of the present study and three other similar studies for Arousal recognition using EEG modality	170
Figure 4.24: The comparison of testing accuracy rates of the present study and four other similar studies for Valence recognition using EEG modality	170
Figure 4.25: The comparison of testing accuracy rates of the present study and four other similar studies for Liking recognition using EEG modality	171
Figure 4.26: The comparison of testing accuracy rates of the present study and a benchmark study for Arousal, Valence and Liking recognition using Peripheral modality	171

Figure A. 1: The MATLAB code developed for EEG signal feature extraction	201
Figure A. 2: The MATLAB code developed for GSR signal feature extraction	203
Figure A. 3: The MATLAB code developed for GSR signal feature extraction (contd)	204
Figure A. 4: The MATLAB code developed for GSR signal feature extraction (contd)	205
Figure A. 5: The MATLAB code developed for GSR signal feature extraction (contd)	206

University of Malaya

LIST OF TABLES

Table 2.1: An overview of reviewed physiological-based emotional datasets and their characteristics	47
Table 2.2: Physiological-based emotion classification studies that used k-nearest classification algorithm	61
Table 2.3: Physiological-based emotion classification studies that used naïve Bayes classification algorithms	64
Table 2.4: Physiological-based emotion classification studies that used discriminant analysis classification algorithms	68
Table 2.5: Physiological-based emotion classification studies that used support vector machine classification algorithms	70
Table 2.6: Physiological-based emotion classification studies that used classification tree algorithms	75
Table 2.7: Physiological-based emotion classification studies that used neural network classification algorithms	76
Table 2.8: Summary of studies utilized different ensemble techniques for physiological emotion recognition	81
Table 2.9: Summary of some research studies that employed stacking ensemble classification in other domains	88
Table 3.1: Features extracted from EEG and peripheral physiological signals	104
Table 3.2: Ranking and average rank calculated for three different classifiers on two data sets based on their classification accuracies	127
Table 4.1: Average ranking score of four benchmark classification methods for valence, arousal and liking targets	134
Table 4.2: Average accuracy rates of four benchmark classification methods for valence, arousal and liking targets using peripheral modality	136
Table 4.3: Average accuracy rates of four benchmark classification methods for valence, arousal and liking targets using EEG modality	137
Table 4.4: Average accuracy rates of four benchmark classification methods for valence, arousal and liking targets using (Peripheral+EEG) modalities	137
Table 4.5: Average ranking score of each modality using four benchmark classifiers	138

Table 4.6: Average ranking score of each modality using SVM	140
Table 4.7: Average ranking score of each modality using ANN	140
Table 4.8: Average ranking score of each modality using LDA	140
Table 4.9: Average ranking score of each modality using CART	140
Table 4.10: Average ranking score of the four classifiers for each modality	141
Table 4.11: Average ranking score of the three modalities for Valence recognition ...	142
Table 4.12: Average ranking score of the three modalities for Arousal recognition ...	142
Table 4.13: Average ranking score of the three modalities for Liking recognition	142
Table 4.14: Average testing rates of the best feature-based multi-classifier methods for Peripheral modality.....	144
Table 4.15: Average testing rates of the best feature-based multi-classifier methods for EEG modality	144
Table 4.16: Average testing rates of the best feature-based multi-classifier methods for (Peripheral+EEG) modalities	145
Table 4.17: Average accuracy rates of Liking, Arousal, and Valence using different subset sizes.....	150
Table 4.18: Average accuracy rates of the three best feature-based dual-layer ensemble classification method	152
Table 4.19: Average ranking scores of the feature-based dual-layer ensemble classification methods over the three modalities.....	153
Table 4.20: Testing accuracy rates of feature-based dual-layer ensemble classification methods on Valence, Arousal and liking with peripheral modality	154
Table 4.21: Testing accuracy rates of feature-based dual-layer ensemble classification methods on Valence, Arousal and liking with EEG modality	155
Table 4.22: Testing accuracy rates of feature-based dual-layer ensemble classification methods on Valence, Arousal and liking with (Peripheral+EEG) modalities	156
Table 4.23: Average ranking scores of the best classifiers taken from three different categories.....	159
Table 4.24: Average ranking score of the three modalities for valence recognition	164

Table 4.25: Average ranking score of the three modalities for arousal recognition	164
Table 4.26: Average ranking score of the three modalities for liking recognition.....	165
Table 4.27: Results obtained by Wilcoxon test for feature-based dual-layer (SVM+CART) algorithm.....	167
Table 4.28: Accuracy rates comparison of five similar studies	169
Table B. 1: Valence classification accuracies using different feature subset sizes and CART classifiers on peripheral modality.....	207
Table B. 2: Average testing accuracy rates of the feature-based multi-classifier methods for valence recognition using different classifiers on peripheral modality.....	207
Table B. 3: Average testing accuracy rates of the feature-based multi-classifier methods for arousal recognition using different classifiers on peripheral modality	208
Table B. 4: Average testing accuracy rates of the feature-based multi-classifier methods for liking recognition using different classifiers on peripheral modality	208
Table B. 5: Valence classification accuracies using different feature subset sizes and CART classifiers on EEG modality.....	209
Table B. 6: Average testing accuracy rates of the feature-based multi-classifier methods for valence recognition using different classifiers on EEG modality.....	209
Table B. 7: Average testing accuracy rates of the feature-based multi-classifier methods for arousal recognition using different classifiers on EEG modality	210
Table B. 8: Average testing accuracy rates of the feature-based multi-classifier methods for liking recognition using different classifiers on EEG modality.....	210
Table B. 9: Valence classification accuracies using different feature subset sizes and CART classifiers on (Peripheral+EEG) modalities.....	211
Table B. 10: Average testing accuracy rates of the feature-based multi-classifier methods for valence recognition using different classifiers on (Peripheral+EEG) modalities ...	211
Table B. 11: Average testing accuracy rates of the feature-based multi-classifier methods for arousal recognition using different classifiers on (Peripheral+EEG) modalities....	212
Table B. 12: Average testing accuracy rates of the feature-based multi-classifier methods for liking recognition using different classifiers on (Peripheral+EEG) modalities.....	212

LIST OF SYMBOLS AND ABBREVIATIONS

Abbreviation	Meaning
Adaboost	Adaptive Boosting
ANNs	Artificial Neural Networks
ANS	Autonomic Nervous System
Bagging	Bootstrap Aggregating
BVP	Blood Volume Pulse
CARs	Classification Accuracy Rates
CART	Classification And Regression Tree
CIFE	Conditional Informative Feature Extraction
CMIM	Conditional Mutual Info Maximization
CNS	Central Nervous System
CONDRED	Conditional Redundancy
DSIR	Double Input Symmetrical Relevance
ECG	Electrocardiogram
EEG	Electroencephalography
EMG	Electromyography
EOG	Electrooculography
FACS	Facial Action Coding System
FDEC	Feature-Based Dual-Layer Ensemble Classification Method
GSR	Galvanic Skin Response
HR	Heart Rate
HRV	Heart Rate Variability
ICAP	Interaction Capping
JMI	Joint Mutual Information
KNNs	K-Nearest Neighbors
LDA	Linear Discriminant Analysis
MEG	Magnetoencephalogram
MIM	Mutual Information Maximization
PCA	Principal Component Analysis
PNS	Parasympathetic Nervous System
RF	Random Forest
RSP	Respiration
SAM	Self-Assessment Manikin
SBS	Sequential Backward Selection
SC	Skin Conductance
SFFS	Sequential Floating Forward Selection
SFS	Sequential Forward Selection
SNS	Sympathetic Nervous System
ST	Skin Temperature
SVMs	Support Vector Machines

LIST OF APPENDICES

Appendix A: The samples of code used for feature extraction	201
Appendix B: Classification accuracy results of valence, arousal, and linking recognition using different feature subset sizes.....	207

University of Malaya

CHAPTER 1: INTRODUCTION

1.1 Introduction

As our environment is rapidly being influenced by technology, there has also been a correspondingly rapid growth of computer-based devices over the past decades. For this reason, effective ways of improving the interaction between human and computer have become the topics of interest among researchers. Being able to know and analyse an individual's emotional experiences would certainly lead to the development of more effortless assistive technological devices, which known as digital companions (Wagner & Andre, 2005). Affective computing technology, which aims to improve human-machine interaction, employs various algorithms to build a robust and reliable classification models for recognizing human emotions. Emotion recognition is one of the key steps towards emotional intelligence in advanced human-machine interaction (Hrabal et al., 2012). Researchers have used different modalities, such as speech, facial expression, gesture, and physiological responses to recognize human emotional states. The physiological modality has the advantage of being more robust against possible artifacts of human interpersonal hiding, since they will be instantaneously managed by the human autonomous nervous system (ANS). In addition, physiological responses can be measured continuously and in some cases, especially in speech impaired or autistic people, are probably the only reliable way to recognize the human emotional state (Picard, Vyzas, & Healey, 2001). Physiological responses or physiological signals reflect the immediate change (increase or decrease) in one or more of the body systems in response to a stimulus, for examples, change in blood pressure, heart rate or skin and body temperature. These responses are commonly measured through the physiological signals using methods like ECG (to record heart rate response), SC sensors (to record skin

conductance response), etc. The information collected via these signal sensors go through feature extraction processes, where important features of each signal are extracted. For example, in ECG signals, heart rate variability is one of the important features to be extracted. Eventually, all the extracted features are fed into a machine-learning algorithm, called classifier, to assign or map the recorded data signals to their corresponding emotional states.

1.2 Research Background

This section describes some fundamental concepts related to this research including human physiological signals, human emotions and the main components of an automatic physiological-based emotion recognition system.

1.2.1 Physiological Signals

There are a number of physiological signals, which are often collected to provide details about an individual's well-being, emotions, and so on. The following sub-sections discuss some of the physiological signals and sensors that are normally used for emotion recognition.

Heart signal: The human heart is situated more towards the left side of the chest. Examination of its function through the electrocardiogram (ECG) provides a large amount of information about human feelings. Heart function is mainly measured based on the heart rate (HR), which is the number of electrical impulses caused by the depolarization and repolarization of the heart muscle (Khalili & Moradi, 2008). The electrical activities of the heart, shown as a form of waveforms, are generally produced by utilizing electrodes attached at various locations on the chest. Levenson et al. (1990) found that HR acceleration - when compared to a baseline HR - is higher for anger, sadness, and fear in contrast to happiness, disgust, and surprise. Heart rate (HR) and heart rate variability (HRV) are common measures that are extracted by the ECG electrical

sensors. On the other hand, the blood volume pulse (BVP) sensor, called photoplethysmography sensor, measures circulation of blood by applying infrared light to the head of a finger and measure the amount of light that is reflected. Therefore, physiological changes related to the heart's activities can be detected by both the BVP and ECG sensors (Haag, Goronzy, Schaich, & Williams, 2004)

Body temperature: It is a valuable physiological signal, which is easy to measure using the skin temperature sensor (SKT). The temperature changes can reveal the differences in mood and emotions. The body temperature is measured by fixing a sensor on the fingers to detect the temperature signal and its changes. The sensor can also be used to detect the excitement level of a person (Khalili & Moradi, 2008).

Muscle electrical activity: Muscle electrical activity signals are generated during muscle contraction and relaxation. Facial electromyogram (EMG) is used to recognize human emotional experiences by attaching electrodes to the skin of the face. It measures muscle response or electrical activity produced by a nerve stimulus on the muscle. For example, the frowns and smiles can be an important source of information for facial emotion recognition, can be detected by electrodes attached to muscles on the skin of the face (Bradley, Lang, Cacioppo, Tassinari, & Berntson, 2007).

Human respiration function: This function enables the transfer of air into the lungs to help the diffusion of oxygen into the blood stream and let the waste gasses out. At the time of inhaling and exhaling (i.e., breathing), the lung inflates and deflates, respectively, while the diaphragm pushes up and falls down. The deepness and quickness of the respiration can indicate the status of an individual's well-being and emotions. Respiration (RSP) sensors monitor accurately how deeply and quickly a person is breathing (Khalili & Moradi, 2008). A person's respiratory rate is the mean number of breaths he or she takes per minute. The breathing system is extremely complex and responsive to various

psychological events (Lorig, Cacioppo, Tassinary, & Berntson, 2007). For example, if respiration rate will go up, it indicated quicker and shorter breaths - whenever an individual is in a fear state, whilst a rise in breathing rate with deeper breaths is observed when an individual is mad or angry (Khalili & Moradi, 2008).

Skin conductance (SC): This is an index of the sympathetic nervous system (SNS) activity and emotional arousal (Lang, 1995; Levenson, 1992). Each time an individual feels stress and tense, the palms get humid because of raised activity in the SNS, which can result in accelerated hydration in the sweat channels and on the outside of the skin where skin conductivity is increased and can be measured through skin resistance to a small electrical current (Andreassi, 2007; Dawson, Schell, & Fillion, 2007). Skin conductance (SC) sensors measure the electrical conductance of the skin and are usually put on the finger. Galvanic skin resistance (GSR) sensor is also utilized to measure the electrical conductance of the skin.

Brain electrical activity signals: These can be very useful to provide information regarding a person's behaviors and emotions. These signals can be recorded by electroencephalography (EEG) sensors, which are electrodes positioned on the scalp. Many researchers believe that by examining the difference in activity between both hemispheres of the brain, measured by EEG, different emotions can be identified (Cacioppo, 2004; Schiffer et al., 2007). In the past, EEG-based studies of emotional specificity have shown that asymmetric activity at the frontal site of both hemispheres of the brain, recorded through EEG (especially in the alpha (8–12 Hz) band), is associated with different emotions (Cacioppo, 2004; Lee & Hsieh, 2014; Schiffer et al., 2007). For instance, Ekman and Davidson (1993) discovered that voluntary facial expressions of smiles of satisfaction generate higher left frontal activity, while another study found a decrease in left frontal activity during the voluntary facial expressions of fear (Krumhansl,

1997). Figure 1.1 shows the location of the sensors on the human body as well as the associated physiological signals recorded by each type of sensor (Kim & André, 2008).

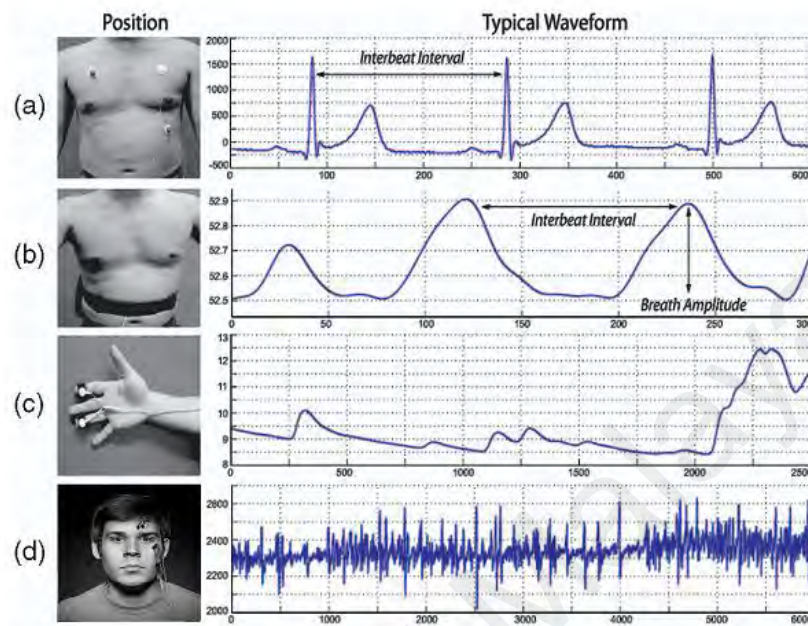


Figure 1.1: Biosensors locations on the body and the associated waveforms: (a) ECG. (b) RSP. (c) SC. (d) EMG (Kim & André, 2008)

1.2.2 Human Emotions

In any study of the human emotion, it is crucial to determine the accurate definition for emotion. There is still no precise definition of emotion and it differs from one research disciplines to another. For example, psychology, affective neuroscience, etc., have different perspectives of emotion. Basically, emotion is a state of mind or feeling that arises naturally and is usually associated with physiological transforms. Unlike mood that may remain for a long time, emotions persist for only a few seconds or minutes (Wioleta, 2013).

Models of emotions: In the literature, some of the well-known theories of emotions and emotion models for different fields have been proposed (Chanel, Kronegg, Grandjean, & Pun, 2006). Two of the well-known model will be discussed below. One model divides

emotions into discrete groups which can be described by specific labels such as sadness, happiness, joy, etc. (Kim & André, 2008). However, not every emotion could be characterized in this way since many types of emotions are combined, hence, it is difficult to categorize them into distinct groups. The famous American psychologist, Ekman (1957), was among the pioneer who methodically studied the human emotions, and proposed the discrete emotional model which consists of six common emotions - surprise, anger, disgust, happiness, sadness, and fear (Chanel et al., 2006). These basic emotions should have identical constructs across all human beings and cultures, and also distinctive universal features, for example, facial expressions and physiology. Emotions also have common characteristics - intuitive and momentary (Ekman & Cordaro, 2011). As opposed to the discrete emotion model, Russell (Russell & Barrett, 1999, 1980) stated that emotional experiences can be depicted in two-dimensional space of valence and arousal. The arousal dimension ranges from extremely activated (e.g., excited) to extremely deactivated (e.g., relaxed), and the valence dimension ranges from highly pleasant (positive) to highly unpleasant (negative). Therefore, an individual can easily indicate his/her feeling about generated emotion based on these scales (Figure 1.2). The emotions also can be generally evaluated as liking and disliking in a consistent manner by individuals (Hawknis, Mothersbaugh, & Mookerjee, 2011). For example, subjects' liking and disliking can be used for music and video tagging (Koelstra et al., 2012). The Russel emotion model also has been adopted for this research to identify the emotional state of the subject based on their arousal and valence level instead of distinct emotions.

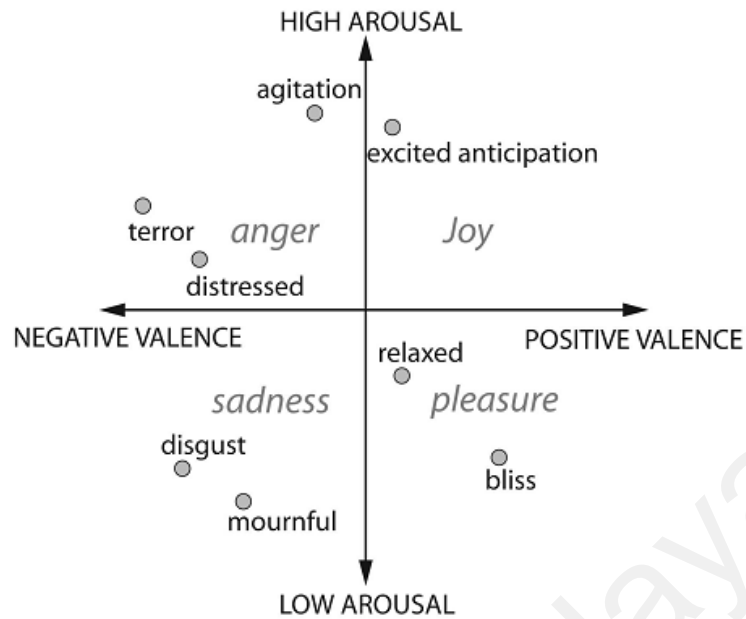


Figure 1.2: Two-dimensional model of emotion based on valence and arousal (Kim & André, 2008)

1.2.3 Physiological Signal based Emotion Recognition

To create the physiological-based human emotional state recognition system, several key components need to be designed and developed. The recorded physiological data signals need to pass through these components. Figure 1.3 shows the main components of a physiological-based emotion recognition system.

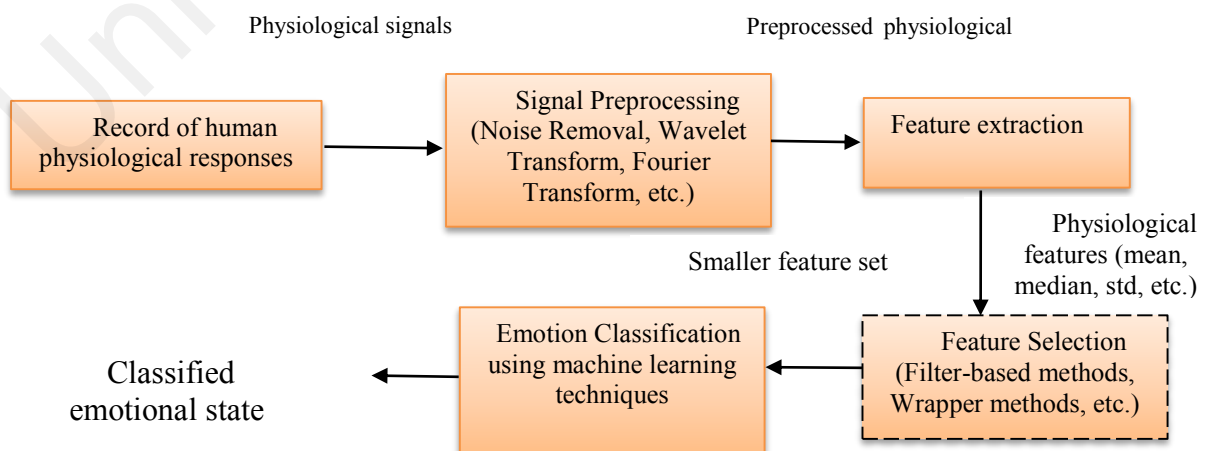


Figure 1.3: The main components of an automatic physiological-based emotion recognition system (Novak, Mihelj, & Munih, 2012)

- *Collection of human physiological signals:*

For the development of the emotion recognition system, accumulating a high-quality database of physiological signals is important (Kim, Bang, & Kim, 2004). Generally, whether a physiological data is of good quality, it will be determined by specialists (Picard, Vyzas, & Healey, 2001). The physiological signals originate from the human activity of ANS, and so they cannot be simply induced by any conscious or intentional control. The emotions need to be naturally elicited on the subjects in order to acquire good quality data. There are several emotion elicitation methods that can be used, such as pictures, movie/film clips (Abadi et al., 2015; Koelstra et al., 2012), and music/sound clips (Kim & André, 2008).

- *Signal preprocessing:*

The raw physiological signals are usually affected by noises and other external interventions. On top of the noises, artifacts such as electrostatic devices as well as muscular movements can have an adverse effect on the raw signals (Kim et al., 2004). These noise and artifacts need to be eliminated from the raw physiological signal before further processing. Commonly, different types of Low-pass filters including Adaptive filters, Elliptic filters, Butterworth filters etc., are employed to pre-process the raw physiological signals (Chang, Zheng, & Wang, 2010; Katsis, Katertsidis, Ganiatsas, & Fotiadis, 2008).

- *Feature Extraction:*

After the signals are pre-processed, the statistical information or features need to be extracted from the signal can be used to recognize the emotional information. Various statistical, time domain, frequency domain and time-frequency domain features could be extracted from the different physiological signals. As an illustration, a total of 110

features were extracted in Kim et al. (2008), using four physiological signals: ECG, SC, EMG and RSP.

- *Feature Selection:*

All physiological signals features that are extracted might not be associated with emotion. Therefore, it is vital to select only the features that have a correlation between the different emotional states. Although feature selection is optional and some researchers designed emotion recognition system without this component, uncorrelated features reduce the performance of the classifiers (Kim & André, 2008). To select the relevant features for effective emotion classification, a variety of feature selection algorithms like filter-based feature selection methods (e.g. Relief, T-test (Jenke, Peer, & Buss, 2014)), wrapper category of feature selection methods (e.g. Sequential Forward Selection (SFS) (Kolodyazhniy, Kreibig, Gross, Roth, & Wilhelm, 2011), Sequential Backward Selection (SBS) (Giakoumis, Tzovaras, & Hassapis, 2013), Sequential Forward Selection Search (SFFS) (Khezri, Firoozabadi, & Sharafat, 2015) are used.

- *Classification:*

The selected relevant features will be utilized to train a classifier, so that it can classify the different emotional states using the selected features (Maaoui & Pruski, 2010). There are several classifiers including, Decision Tree (Chen, Hu, Moore, Zhang, & Ma, 2015), K-Nearest Neighbour (KNN) (Verma & Tiwary, 2014), Support Vector Machines (SVM) (Khezri et al., 2015), Artificial Neural Network (ANN) (Singh, Conjeti, & Banerjee, 2013), and Linear Discriminant Analysis (LDA) (Jang, Park, Park, Kim, & Sohn, 2015), that are employed for emotion classification.

In fact, all the components of physiological-based emotion recognition systems are essential and improving them can have a positive effect on the accuracy of emotion

classification. However, the focus of this research is on the two major components which are the feature selection and classification.

1.3 Research Motivation

Emotional state recognition systems based on physiological signals has many potential applications. Remarkably, it is revealed from past studies that numerous fields are using physiological signals for system development in a variety of context. The most applied domains are for example:

a) Education:

An emotion-sensitive intelligent tutoring system can offer strategies and customized feedback to assist learners (Frasson & Chalfoun, 2010; Ghergulescu & Muntean, 2014; Malekzadeh, Mustafa, & Lahsasna, 2015).

b) Healthcare science:

Various physiological disorders exist and are directly correlated with the one of the different class of emotions. Several studies have been conducted to recognize the initial phase of stress to avoid the human's life going into the at-risk zone. As an example, in autistic spectrum disorder, since they are unable to use facial expressions and gestures to regulate social interactions, it is important to understand their emotions in order to teach them the social skills (Begeer, Koot, Rieffe, Meerum Terwogt, & Stegge, 2008; E. S. Kim et al., 2015; Uljarevic & Hamilton, 2013). The results of the researches (Bekele et al., 2016; Picard, 2009; Van Hecke et al., 2015) are some tools and algorithms that can help to identify the beginning of mental illness or aroused emotions. Therefore, emotion classification systems can play an important role in improving the health conditions of many people.

c) *Computer games:*

In this type of applications, the aim of using emotional state recognition systems is to track the emotional state of a game player and apply the system to adaptively modify the game in such a way to offer the player a lot more immersive experience, a greater gameplay (Kotsia, Patras, & Fotopoulos, 2012).

d) *Authentication system*

Different bio-signals (ECG, EEG and SC etc.) are combined and interpreted for generation of unique identification variables. These variables are unique and robust and sufficiently strong enough to be broken. This system could be considered in protecting very sensitive locations like defense and banking section etc. (Campisi & La Rocca, 2014; Pal, Gautam, & Singh, 2015).

1.4 Problem Statement

In physiological-based emotion recognition systems, training data sets are generally collected from multiple physiological modalities, which results in high dimensional data sets. This may cause some challenges including the computational complexity and the difficulty of convergence toward the optimal classification model. To overcome these difficulties, some studies apply various feature selection methods. Generally, these methods are not only aimed at reducing the computational cost but also at improving the overall performance of the recognition system. The literature shows that wrapper feature selection methods such as sequential forward selection (SFS) (Alpers, Wilhelm, & Roth, 2005; Kolodyazhniy et al., 2011; Kukulja, Popović, Horvat, Kovač, & Ćosić, 2014; Yannakakis, Martínez, & Jhala, 2010) and sequential backward selection (SBS) (Giakoumis et al., 2013; James Kim & André, 2008; Kolodyazhniy et al., 2011) are frequently used as feature selection methods. These methods were criticized for being overly dependent on the classifier used as the selected features cannot be used with other classifiers (Wang, Zhou, Yi, & Kong, 2014). Another popular feature reduction method

called Principal component analysis (PCA)(Jolliffe, 1986) was applied to reduce the high dimension of the physiological data set and projected it into a lower dimension by generating a new subset of features. This method's disadvantage is the lack of interpretability as the original set of features are replaced by a new set of features which cannot be used for interpretation and analysis (Singh, Conjeti, & Banerjee, 2013; Wei-Long Zheng & Bao-Liang Lu, 2015). Feature ranking methods which known as filter methods (Brown et al., 2012) are a different category of feature selection that also have been used in designing of physiological-based emotion recognition (Clerico, Gupta, & Falk, 2015). The feature ranking methods do not have the limitation of the other feature selection techniques like PCA and wrapper methods. However, it is common that the researchers used one feature ranking method in their proposed systems which may result in sub-optimal solution because two different feature ranking methods are likely to produce two different ranking sets and presenting only one set given by a particular method can be misleading (Kuncheva, 2007). One solution can be using more than one feature ranking method to increase the chance to choose the optimal feature set.

As to the emotion classification model, various single classification methods such as Support Vector Machine (SVM) (Cortes & Vapnik, 1995) and Artificial Neural Network (ANN) (Kohonen, 1982), have been used to develop emotion classification models for physiological emotion recognition systems. Beside single classifiers, ensemble classification methods are also known for their high classification ability compared to single classifiers, have been used in designing emotion recognition systems based on physiological signals (Bhatnagar, Bhardwaj, Sharma, & Haroon, 2014; Novak et al., 2012). Ensemble methods (Rokach, 2010) known as multiple classifier systems, combine a set of multiple classifiers' decisions (i.e. prediction results), usually by using majority vote method, to obtain the final classification output of a given testing pattern. Ensemble classification models may be used naturally in physiological based emotion recognition

systems because of the combination of different modalities such as peripheral data and brain data (i.e. EEG) to predict the emotional state (Novak et al., 2012). There are also some studies (AlZoubi, Fossati, D’Mello, & Calvo, 2014; Colomer Granero et al., 2016; Vaid, Singh, & Kaur, 2015) that have employed some other kind of ensemble classification methods like boosting and bagging (Rokach, 2010) in their proposed emotion recognition system. However, there are other types of ensemble classification methods like stacking ensemble (Wolpert, 1992), which is also called dual-layer ensemble method, that has not been thoroughly investigated in the emotion recognition systems based on physiological signals. In fact, stacking-based ensemble methods have proved its performance in other fields of research, such as network intrusion detection (Syarif, Zaluska, Prugel-Bennett, & Wills, 2012) and software fault prediction (Hussain, Keung, Khan, & Bennin, 2015). The reason for not using more advanced ensemble classification methods like stacking, probably because these methods have not been included in most popular software packages (e.g. SPSS), while benchmark single classifiers like ANN or Linear Discriminant Analysis (LDA) (Fisher, 1936) or other well-known ensemble classification methods like bagging are easily available in many software packages. In another word, more advanced ensemble methods are not widely accessible to be used by the researchers.

1.5 Research Objectives

The main aim of this research is to propose a feature-based dual-layer ensemble classification method to improve the accuracy of the physiological-based emotion recognition systems. This method is the result of combining different feature ranking methods along with more than one classifiers. In order to achieve this goal, the following intermediate objectives are identified:

1. To identify most used feature selection, classification methods and its related issues in the design of existing physiological-based emotion recognition systems.
2. To design and develop a feature-based dual-layer ensemble classification method to improve the accuracy rate of physiological-based emotion recognition system.
3. To evaluate the classification accuracy of the proposed feature-based dual-layer ensemble classification method by comparing with the benchmark classification methods using statistical analysis.

1.6 Research Questions

The following research questions are suggested as a guide for conducting this research at the different phases to accomplish the research objectives:

- 1- What are the most utilized feature selection and classification methods in the design of the current emotion recognition systems based on the physiological signals? (Objective #1)
- 2- What are the most prominent limitation and challenges of the current emotion recognition systems based on the physiological signals? (Objective #1)
- 3- How can we design an improved classification method to enhance the classification accuracy rate of the existing emotion recognition systems? (Objective #2)
- 4- How the classification accuracy rate of the emotion recognition systems is affected by using different data modalities? (Objective #3)
- 5- How the classification accuracy rate of emotion recognition systems is affected by using the benchmark classification methods combined with feature selection methods as compared to the same classification methods without the feature selection methods? (Objective #3)
- 6- Will the proposed classification have better classification accuracy as compared to other classification methods? (Objective #3)

- 7- Can the proposed classification method achieve significant improvement over the other methods? How can we prove that statistically? (Objective #3)

1.7 Research Scope

This research is mainly focused on the development of a physiological-based emotion recognition system and specifically addressing the problem of improving the emotion classification accuracy by designing an ensemble-based classification method. In addition, the effect of some feature selection methods, as well as the impact of different modalities on the classification accuracy rate of some benchmark classification methods, are examined. For evaluation, we use a multimodal publicly available dataset. This data set has three modalities: (1) peripheral physiological signals including electro-cardiogram (ECG), electro-myogram (EMG), electro-oculogram (EOG), blood volume pulse (BVP), respiration amplitude (RSP), skin temperature and galvanic skin response (GSR), (2) electroencephalogram (EEG) and (3) the combination of peripheral physiological signals and EEG. A total of nine data sets were created using the following criteria: The first three data sets consist of peripheral physiological signal recordings with three different targets, namely, Valence, Arousal and Liking rated between 1 to 9. Arousal and valence are two key components of human emotion based on 2D Russell's emotional model. Liking is also considered as distinct emotion. The second three data sets were extracted from EEG signal recordings and, as the previous three data sets, have the same three targets rated between 1 to 9. The last three datasets are formed by combining the three peripheral physiological data sets with the three EEG data sets. The recorded signals in the nine datasets are divided, based on the assigned rating scores of Arousal, Valence and liking by each subject, into two classes (binary classification): "low" if the score is less than 5 and "high" if it is five or more. The proposed method is evaluated using

classification accuracy rate and it is calculated using one-leave-out cross-validation¹. To check the existence of any significance difference between the classification accuracy rate of the proposed classification method and other benchmark classifiers, we apply a statistical test².

1.8 Research Methodology

In order to achieve the main objectives of this research, we have designed our research method based on the design science research (DSR) paradigm. Particularly, the Design Science Research Process Model (DSRM) proposed by (Peffer, Tuunanen, Rothenberger, & Chatterjee, 2007) is adopted to structure this research. We found DSRM process appropriate for our research, mainly because the key goal of a DSRM would be to deliver an artifact which has the characteristics of the research outcomes. In this research, the artifact is feature-based dual-layer ensemble classification method for physiological-based emotion recognition system.

Figure 1.4 depicts the DSRM adopted for this research where it comprises of five activities as briefly described below. The full details of each activity are explained throughout the thesis.

1. *Problem identification and motivation:* Having a defined problem that needs to be solved, is an early essential step in every research. A literature review (i.e. Narrative literature review), which describes and discusses the state of the science related to physiological-based emotion recognition systems from a theoretical and contextual point of view, is conducted to identify possible

¹ Please refer to section 3.8.2.1 for more details.

² Please refer to section 3.8.2.3 for more details.

problem/s in this field of research. The carried out literature review also assist us to propose our solution for the identified problem(s).

2. *Define the objectives for a solution:* As soon as the problem and the significance of a solution have been stated, the requirements that a solution should meet are derived from the stated problem(s). The solution proposed for this research is to design an improved classification algorithm for physiological-based emotion recognition system, particularly, the proposed algorithm has to the ability to enhance the recognition accuracy of emotion recognition system by addressing the problems identified in the existing systems.
3. *Design and Development:* This activity mainly focuses on creating the artifact. In this research, in order to create the proposed feature-based dual-layer ensemble classification method, some steps need to be considered, like database preparation, design emotion recognition system using benchmark classifiers and design feature-based multi-classifier methods.
4. *Evaluation:* Measure how well the artifact supports a solution to the problem. In our research, a series of comparative analysis based on the classification accuracies are conducted to compare classification accuracy rate of different benchmark classification methods. At the last stage of evaluations, the Wilcoxon signed-rank test is applied to check if there is any significance difference between the classification accuracy rate of the proposed classification method and other benchmark classifiers.
5. *Communication:* Once the research is completed, the process involved need to be documented and made available for knowledge sharing and discussion. In our research, all the research activities and processes are documented in form of a thesis. In addition, some major findings are submitted to related journals and conferences for possible publication.

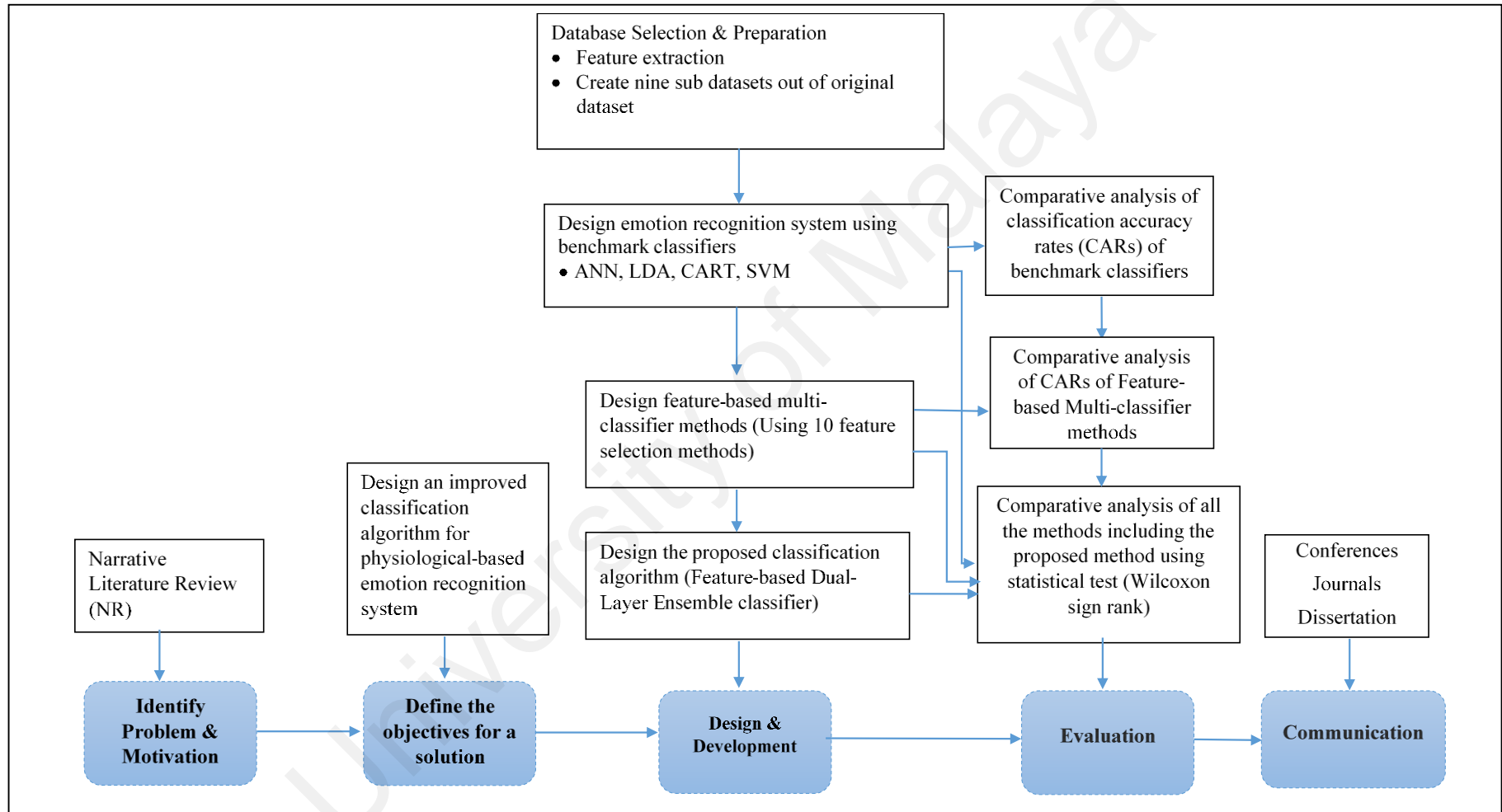


Figure 1.4: The DSRM research method adopted for this research

1.9 Thesis Outline

The remaining of the thesis is organized as follows:

Chapter 2 presents a set of basic concepts related to the human emotional state and the emotion recognition systems. That information is including Emotion Theories as well as different existing emotion models. An overview of different emotion detection modalities and the strategies for emotion elicitation are also presented. We also explain how information related to Human Nervous System (NS) can help to recognize human emotional states and how it can be collected.

Additionally, this chapter mainly describes physiological-based emotion recognition systems and related components. The processes for creating an emotion recognition system which starts from physiological data preparation to emotion classification are discussed in details. The related works are also presented and they are mainly related to the components of physiological-based emotion recognition system that include feature selection techniques as well as classification methods. A summary of benchmark physiological-based emotional datasets is also provided.

Chapter 3 describes the proposed method which consists of three main phases: feature selection that aims to select an optimal set of features using a different combination of feature ranking algorithms while the objective of the second phase is to construct single-layer classification technique. The last phase combines the single-layer classification technique with another classification layer to create proposed feature-based dual-layer ensemble classification technique.

Chapter 4 discusses the results obtained from applying the proposed method on the benchmark data sets. The results are compared with benchmark classification methods using a statistical method.

Chapter 5 concludes this thesis with highlights of contributions and main findings derived from this research. In addition, we propose some recommendations for future works.

University of Malaya

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This chapter first addresses the scientific outlook on emotion, which explains several theories of emotion. The major theories that will be reviewed in this chapter are: James-Lang theory and cognitive-appraisal theory; the notion of the universality of emotion and the discrete and dimensional models of emotions. This chapter also provides the relevant key point regarding the emotion recognition through different modalities like speech and facial modalities, as well as several including the psychophysiology of emotion as well as the rationale behind employing physiological measures for emotion identification.

This chapter will also describe the obtainable physiological-based emotional datasets that were collected for research needs and its features. Furthermore, the approach associated with the development of an emotion recognition system will be clearly explained. Ultimately, at the final section, the relevant works to our study is going to be reviewed.

2.2 Theories of Emotion

With regards to the study of emotion-related topics, the vital question is what is right definition for emotion? There is still no precise definition of emotion and every research disciplines (e.g. psychology, affective neuroscience) has adopted different perspectives of emotions. Based on the survey done by (Calvo & D'Mello, 2010), there are six important theoretical views which have been used to investigate human emotions and their usability to affective computing research. These six theories take into account emotions as expressions, embodiments, solutions of cognitive appraisal, social constructs, outcomes of neural circuitry, or cognitive and social interpretations of

adjustments in core emotion. A summary of these theories regarding the present research will be explained below.

2.2.1 James-Lange Theory

The James-Lange theory of emotion was suggested by psychologist William James and physiologist Carl Lange separately in the middle of the 1880s. The James-Lange theory of emotion suggests that emotions are created due to physiological responses to the events. Based on their theory, emotion is equivalent to the range of physiological arousal caused by external events. The two scientists suggested that for someone to feel emotion, he/she must first experience bodily responses such as increased respiration, increased heart rate, or sweaty hands. Once this physiological response is recognized, then the person can say that he/she feels the emotion. This means that every emotion possesses a specific physiological pattern. Since physiological responses are instantly managed by the autonomic nervous system (ANS), it will be practical to anticipate patterns of ANS activity to be linked to specific emotional states (Regan & Atkins, 2007). This idea has inspired most of the emerging research on physiological-based emotion identification research.

2.2.2 Cognitive Appraisal Theory

Cognitive Appraisal is a theory of emotion that implicates people's personal interpretations of an event in identifying their emotional reaction. An individual, in order to feel an emotion, must appraise an event (was the event a positive or negative occurrence) or a stimulus directly disturbing him. Throughout this process, an assessment of that event or stimulation takes place based on some factors such as relevance, ability to cope, consequences and opportunity (Lazarus, 1982; Scherer, 1999).

The scientific studies provide evidence that the cognitive theory of emotion supports the idea that the brain is the main organ that is responsible for processing and evaluating

emotional events (Farquharson, 1942). Thus brain signals provide a possible source for evaluating emotional experience (Khosrowabadi, Quek, Wahab, & Ang, 2010), and this forms the basis for our research study in which we utilized EEG data for detecting of emotional states.

2.2.3 Universality of Basic Emotions

The concept of universality of emotions was initiated at the time of Darwin and his research (Darwin, 1965), where he considered emotions as an evolutionary product. This means that emotions are directly connected to the brain and not a learned skill. Ekman and other researchers (Ekman, 1992; Izard, 1992) have supported this point of view remarkably and suggested the existence of six basic emotions (e.g. joy, anger, fear, etc.) that are the building blocks for all other emotional experiences. These basic emotions should have identical constructs across human beings and cultures, and also distinctive universal features (Ekman & Cordaro, 2011).

According to Picard (1995), from an affective computing research outlook, common patterns are anticipated for universal emotions and these might possibly differ for the derived emotions from the basic emotions. This links us to the next section where a discussion on different models of emotion is provided.

2.3 Discrete and dimensional emotion models

Research on emotion is widely known as the presence of a set of discrete emotional prototypes (Ekman, 1992; Izard, 1992). These discrete emotions possess a unique profile in experience physiology and behavior (Mauss & Robinson, 2009). According to Mauss and his fellow researcher (Mauss & Robinson, 2009), they recommended around 20 discrete emotional experiences which most frequently mentioned are fear, anger, sadness and joy. As an alternative to the idea of discrete emotions, Russell (Russell & Barrett, 1999, 1980) proposed that emotional experiences are most beneficial to be depicted in the two-dimensional space of valence and arousal.

The arousal dimension ranges from extremely activated (e.g. excited) to extremely deactivated (e.g. relaxed), and the valence dimension from highly pleasant (positive) to highly unpleasant (negative). Nevertheless, Barrett (Barrett, 1998) claimed that none of the mentioned models may well precisely present the subjective emotional state of the human.

As such, Russell and Barrett (Russell & Barrett, 1999) attempted to merge both ideas together and suggested that discrete emotional experiences are associated in some way to these two dimensions, so they can vertically be arranged as a fuzzy hierarchy and horizontally (Figure 2.1).

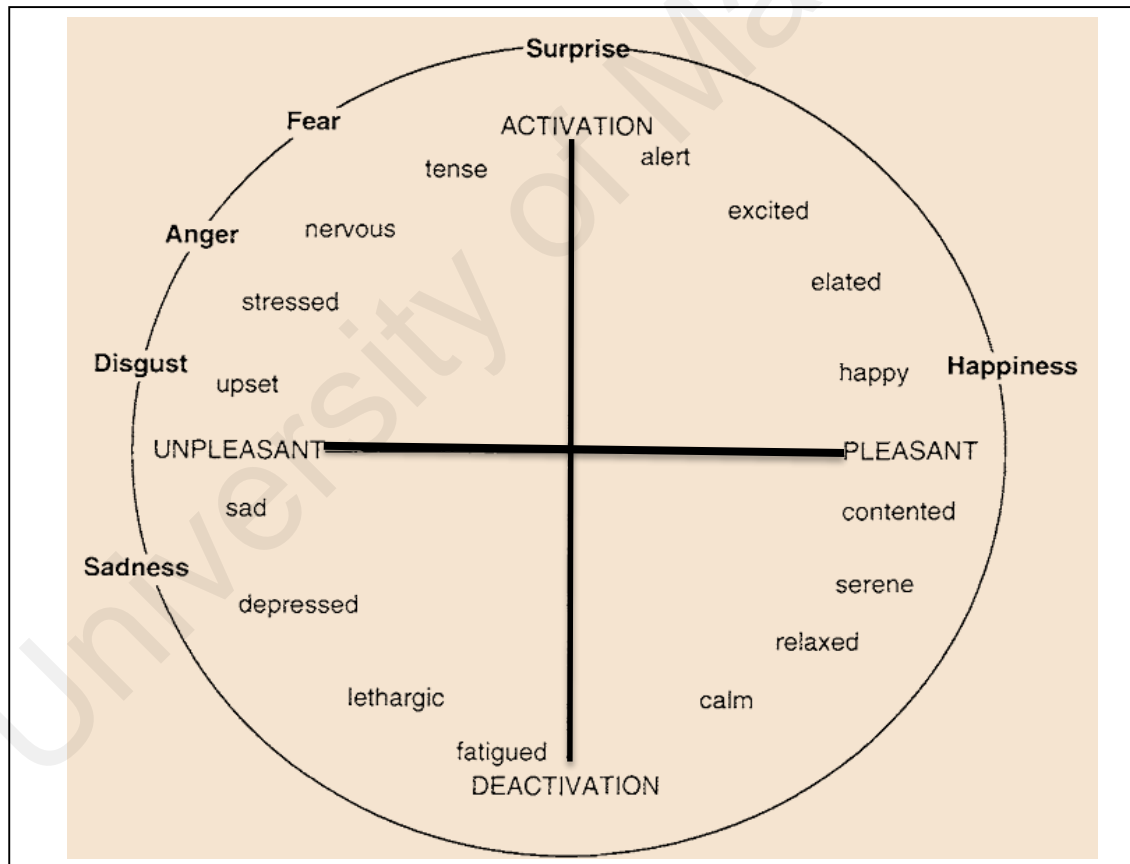


Figure 2.1: Emotional experiences described in two dimensions valence and arousal, accompanied by distribution of some discrete prototypes of emotions along the two dimensions (adopted from (Russell & Barrett, 1999, p.808))

As for physiological-based emotion recognition, relatively reliable recognition performance rate could be seen for both discrete (Kim et al., 2004; Wen et al., 2014) and dimensional (Horlings, Datcu, & Rothkrantz, 2008; Koelstra et al., 2012) models of

emotion. However, some researchers (Horlings et al., 2008; Kim, Bee, Wagner, & André, 2004) found that predicting emotion categories based on dimensional valence/arousal modeling of emotion is more successful compared to predicting the intensity of emotion or in another word modeling discrete emotion.

2.4 Emotion Detection Modalities

In order to recognize human emotional state by computer systems, some modalities have been widely utilized. The most well-known modalities that have been utilized for this purpose are facial expression and vocal patterns. Additionally, body gestures and movements also have been employed (Zeng, Pantic, Roisman, & Huang, 2009). At the same time, with the improving capability of computer systems, and also the advances in communication capability of signal acquisition systems, new modalities similar to emotion recognition from physiological sensors are achievable. Enhanced wearable electronic devices could provide a more natural, non-obstructive means of acquiring this sort of physiological measures from people today (Picard et al., 2001).

The capability of computers to receive a large amount of physiological data from individuals conveniently allows scientists the ability to analyze large amounts of physiological data to obtain and discover signal features that associate with the individuals' emotional states. The significance of physiological data signals for human emotion recognition, is that it provides another possibility of recognizing human emotions aside from the conventionally available modalities of the tone of voice and facial and body gestures (Picard et al., 2001).

In addition, using physiological data for emotion recognition is certainly essential for particular applications similar to deception identification or the applications that provide feedback to speech disable people or autistic people. Therefore, physiological-based emotion recognition will perhaps be the only reliable way to recognize their emotional state since those people are not fully capable of speaking or showing their facial emotion

respectively. Aside from that, physiological signals are continuously generated so it can be measured and monitored continuously, and since it is governed by the central and autonomic nervous systems (ANS), therefore it is difficult to pretend or manipulate them (Peter, Ebert, & Beikirch, 2009). This is definitely advantage compared with other modalities like voice or eye movement which their actions might be masked.

Progresses in emotion recognition performance of computer systems using these single modalities make it easy for the researcher to take into consideration designing multimodal emotion recognition systems where these systems combine several modalities in a whole package to recognize the individual emotion (Lisetti & Nasoz, 2002). Figure 2.2 depicts the framework of a multimodal emotion recognition system.

Many researches have been carried out to design multimodal emotion recognition systems along with physiological measures using facial and voice features. Based on the review conducted by Tao and Tan (2005), the multimodal emotion recognition systems are capable of recognizing human emotional states with higher accuracy compared to sole modalities.

For instance, Gunes, Piccardi, & Pantic (2008) developed a bimodal emotion identification system driven by facial expression features taken from 41 subjects' videos and their physiological responses (ECG and EDA). They measured subjects' sadness and amusement levels when subjects were watching emotionally reminiscent (evocative) movie clips. Their experimental results demonstrated that the integration of facial and physiological measures obtained better classification performance rate compared to using each modality separately.

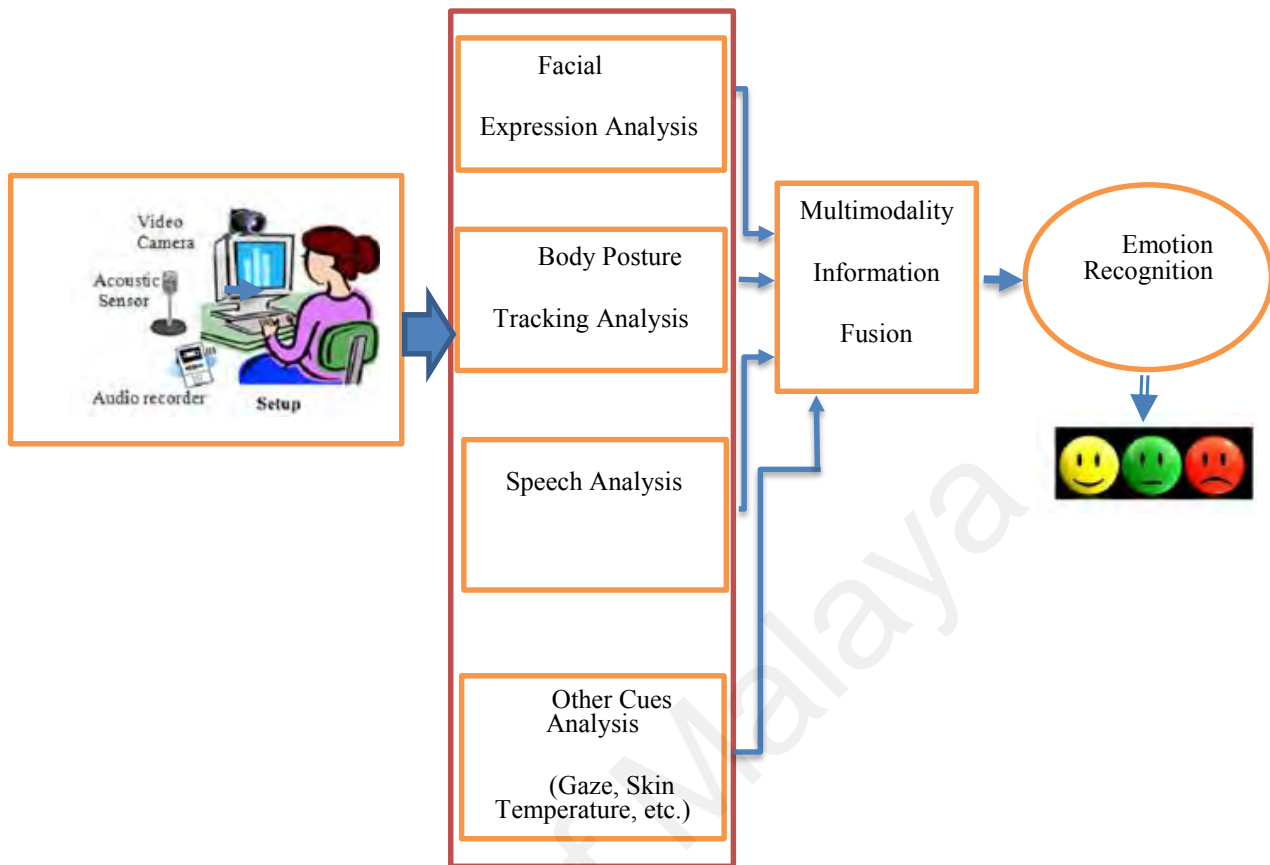


Figure 2.2: The framework of a multimodal emotion recognition system (Tao & Tan, 2005)

In Kim et al. (2004), they reported about a designed bi-modal emotion recognition system that utilizes speech and physiological measures. Subjects employ a computer gaming software which is capable of emotional interactions with players. According to their findings, they offered a two-channel emotion recognition system which they assumed would likely enhance the sole modality system as it might facilitate the handling of ambiguities. Their results imply that SC (Skin Conductance) as a physiological measure is an efficient signal of arousal and voice harmonics can help identify positive emotions with high arousal from the negative emotions with high arousal.

Soleymani et al. (2012) have also designed a multimodal database for emotion recognition and implicit tagging. The multimodal data set was recorded based on face video, speech signals, eye gaze data, and physiological measures along with central

nervous system data. The multimodal information has been collected from 27 subjects while watching emotional videos. The reported results showed that the combination of modalities of eye gaze data features and central nervous system data features could achieve higher emotion recognition accuracy rate in comparison with other single modalities.

2.5 Methods in Multimodal Emotion Recognition

As we stated in the previous section, emotion recognition by using multiple sources and sensors (i.e. different modalities) can provide better emotion recognition accuracy rate compared to single modality emotion recognition systems. Thus, there is certainly a need for methods that integrate and synthesize information from these multimodal resources. This procedure is known as information fusion. There are various of fusion methods related to multimodal emotion recognition programs like feature level fusion and decision level fusion (Zeng et al., 2009). Feature level fusion is involved in the integration of extracted features from each modality into one combined feature vector. The risks with this method are that the features from different signals have different time frames, which might need more attention to synchronization of the extracted features. Another issue is the high dimensionality of the derived feature vectors, which influence the performance of emotion recognition system. On the contrary, in decision level fusion each modality is utilized to classify emotions individually, and the ultimate decision is achieved by merging the decisions of all the modalities according to some criteria such as averaging or voting. Still, formulating a supreme strategy for decision level fusion remains to be an open research issue (Kim & Andre, 2006).

As stated in the research by Chanel et al. (2006), fusion offers better results in an emotion recognition related experiment that combined EEG and peripheral signals (peripheral signals are referred to physiological signals linked to ANS responses such as ECG, EMG, SC and RSP). Based on their results, some subjects had better scores with

peripheral signals as compared to with EEG and the opposite. In the same manner, Kim & Andre (2006) discovered that employed feature-level fusion technique in their research given the best results using a combination of physiological signals together with speech modalities, stating that feature-level fusion is more suitable when merging modalities with similar aspects.

2.6 Psychophysiology of Emotion

Study of psychological phenomena (e.g. emotions, and moods) and human physiology is very important, in which understanding the main links between an emotional behavior and a particular physiological response is necessary to properly monitor and knowing how to identify emotions from physiological measures. This section gives a summary on the links between the emotional behavior and human physiology which is managed by human nervous system. In addition, the way emotions are generated and experienced, and type of the physiological responses that are linked with various emotions are presented.

The human nervous system is split into two components the central nervous system (CNS) which comprises of the brain and spinal cord, as well as the peripheral nervous system. The peripheral nervous system then comes with the a) autonomic nervous system (ANS), and b) the somatic nervous system. The ANS is in charge of organizing the function of spontaneous bodily organs like the heart and glands, and smoothes out muscle systems of the human body and their reactions to the environment via increased heart rate and sweaty glands in the event of frightened circumstances or slow heart rate and respiration rate in the event of gloomy occasions. At the same time, the somatic nervous system manages skeletal voluntary muscles like facial muscles and biceps which are

under human control (Andreassi, 2007). The branches of the human nervous system are shown in Figure 2.3.

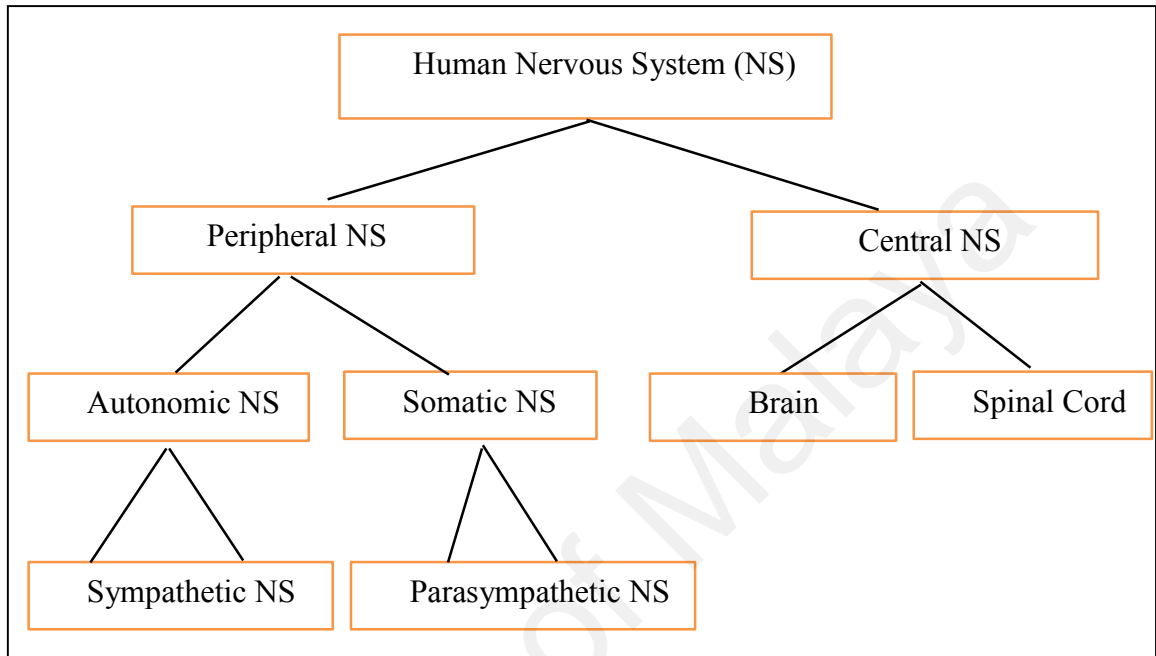


Figure 2.3: Divisions of human nervous system (Andreassi, 2007)

2.6.1 The Electroencephalogram (EEG)

Based on the works by many researchers (e.g. Cacioppo, Tassinary, & Berntson, 2007; Lang & Bradley, 2010), the activity of central nervous system straightly can be monitored by EEG, thereby this is valuable in discovering the hemispheric specialization and lateralization of emotion. The human brain neurons are successfully concerned with natural stimuli from receptors such as eyes and ears, or perhaps by neural prompt caused by some other nerves.

Hans Berger was a German psychiatrist who is known as the founder of electroencephalography (EEG) unveiled the initial result on human EEGs, where he recognized two wave patterns, a large regular wave (10-11 Hz) which he labeled alpha

waves, and a smaller irregular and faster wave which he labeled beta waves (20-30 Hz) (Andreassi, 2007; Pizzagalli, 2007).

According to the researches, specific human conditions have been linked to the activation of specific frequency bands in specific brain regions. For instance, a state of “high awareness or excitation” (i.e. high level of arousal) is linked to higher beta frequencies, at the same time a state of relaxed is linked to alpha activity; Delta waves are slow and brain waves which are linked to deep sleep in ordinary people, whereas theta brain waves happen more often during states of pleasure and displeasure (Pizzagalli, 2007). In addition, the state of engagement is linked with brain theta activity recorded from the frontal sites. Furthermore, the brain delta activity is linked to the state of sleepiness (Allanson & Fairclough, 2004; Fairclough, 2009). Figure 2.4 depicts the four main types of EEG brain waves, frequency level, and associated human state.

There are some theories about the relationship between brain signals and emotion. Several researchers believe that by checking at the difference in activity of both hemispheres of the brain recorded by EEG, different emotions can be identified (Cacioppo, 2004; Schiffer et al., 2007). In the research done by (Chanel et al., 2006), they found that EEG can be employed to recognize arousal level of human emotion while EEG performance was better than other peripheral signals for arousal level recognition.

There are more emotion recognition research studies that have proved the functionality and usefulness of EEG measurements to identify human emotional states including level of arousal, engagement and some basic emotions such as joy, anger, sadness, fear and relax (Chai et al., 2014; Heraz, Razaki, & Frasson, 2007; Horlings et al., 2008; Khalili & Moradi, 2008; Koelstra et al., 2012).

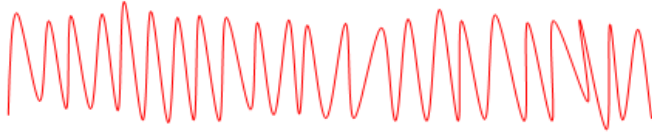



Nature of Sine Wave Activity in the Brain	Frequency Level & Description
	<p>BETA 14 – 30 Hertz</p> <p>Awake, normal levels of alertness. Also associated with overactive thinking patterns, stress, anxiety, frustration and other undesired states. People spend most of their daily life operating at this level.</p>
	<p>ALPHA 9 – 13 Hertz</p> <p>Relaxed, calm levels of mental activity occur at this level. A peaceful state associated with tranquillity and relaxation, which people can achieve through effective relaxation exercises and meditation.</p>
	<p>THETA 4 – 8 Hertz</p> <p>A deeper state of mindfulness associated with creative insight, cognitive & memory enhancement and feelings of deep connectedness. Also the level at which people naturally progress into sleep state.</p>
	<p>DELTA 1 – 3 Hertz</p> <p>The deepest brainwave level associated with dreamless (non-REM) sleep. Essential for proper restoration of health and immune system. Difficult to achieve this level if overactive at the Beta level.</p>

Figure 2.4: The four major types of EEG waves and associated human state (Adopted from (Webster & Clark, 2010))

2.6.2 Autonomic Nervous System (ANS) Measures and Its applicability for Emotion Recognition

The ANS is generally engaged in the regulation of essential activities of the human body such as cardiovascular system and respiratory activities, temperature, blood pressure and some other features of emotional attitude. The main functionality of the ANS would be to maintain the human body in a well-balanced inner condition in a deal with internal or external situations that may affect its balance (Andreassi, 2007; Carlson, 2012).

In accordance with the composition and functionality of the ANS, it includes two components, which are a sympathetic nervous system (SNS) together with the parasympathetic nervous system (PNS) (Carlson, 2012). As the sympathetic component prepares the body for high levels of somatic task, which may come from an interaction with a stimulus in the environment and prepares the body for a state of troubles while the

parasympathetic helps to bring a state of rest and relaxation to the human body (Andreassi, 2007; Lorig et al., 2007).

The SNS usually handles those tasks which are involved with urgent and stress conditions. These tasks may contain raised blood circulation to skeletal muscles, higher heart rate, outgoings of energy, expansion of eye pupils and increased sweating (Andreassi, 2007; Carlson, 2012). The reactions of the SNS to hazard situation are very adaptable since this improves survival. Alternatively, the tasks under PNS control consist of reduces in heart rate and blood pressure, activation of the digestive system, papillary constriction, resting and sleep (Andreassi, 2007). The tasks of the SNS and PNS probably have different functions. However, the tasks of the two systems are complementary in which allows an easy flow of body activities and behavior (Andreassi, 2007).

The change raises in the activities of these two components of ANS system while facing different environment conditions and their reflections in form of physiological measures as well as regulating human body physiology during the transition from one emotion to another, one could be measured through physiological sensing. The goal is then to model these emotion changes making use of computerized systems that would allow the automatic discovery of these alterations (Barreto, Zhai, & Adjouadi, 2007).

In many researches that have been carried out in the area of affective computing, to design automatic emotion identification systems, numerous physiological signals have been employed. The most essential ones are such as ECG, EMG, SC, RSP, ST, BVP and also EEG (AlZoubi et al., 2014; Barreto et al., 2007; Gunes et al., 2008; Haag et al., 2004; Hariharan & Adam, 2015; Hernandez, Paredes, Roseway, & Czerwinski, 2014; Khalili & Moradi, 2008; Koelstra et al., 2012; Picard et al., 2001; Soleymani, Lichtenauer, et al., 2012; Wagner, Kim, & Andre, 2005). In the following sections, a short presentation about each of the physiological signals which have been utilized for designing emotion recognition system in our research is introduced.

2.6.2.1 Respiration (RSP)

A person's respiratory rate is the mean number of breaths he or she take per minute. The normal respiration rate for an adult at rest is 12 to 20 breaths per minute. The breathing system is extremely complex and responsive to various psychological issues (Lorig et al., 2007). As an example, the respiration rate will rise whenever an individual is in a fear state which is also linked to quicker and shorter breaths, whilst becoming mad is linked to a rise in breathing rate with deeper breaths. Based on the research in (Allanson & Fairclough, 2004), using breathing patterns could indicate and differentiate between human emotional states such as calm against excitements states. As an example, the fast and deep breath represent excitement emotional states such as anger or fear and also joy, while quick short breathing can reveal anxious symptoms such as panic and fear (Philippot, Chappelle, & Blairy, 2002). Haag et al. (2004) also have mentioned that slow and deep breathing styles, which signify a relaxed resting state while slow and short breathing can reflect states of apathetic like depression or relaxed. Some other scientists like Lichtenstein et al. (2008), also shown that breathing styles that can be evaluated by RSP rate is noticeably dissimilar between happiness and fears, anger and sadness, satisfaction and happiness also sadness and happiness.

2.6.2.2 Electrocardiogram (ECG)

Electrocardiography (ECG or EKG) is the process of recording the electrical activity of the heart over a period of time using electrodes placed on an individual's body and it is usually measured based on beats per minute. Heart rate (HR) and heart rate variability (HRV) are common measures that would be extracted from ECG which are also an important physiological index for detecting different emotions. These two measures are comprised of details on the status of the ANS where both sympathetic and parasympathetic nervous system activities could be comprehended. Many researches

have been carried out by using HR and HRV related features in order to identify human emotional states. For instances, Levenson, Ekman, & Friesen (1990) found that HR acceleration rate compared to a baseline HR is higher for anger, sadness and fear in contrasted to happiness, disgust, and surprise. Furthermore, Cacioppo et al. (2007) discovered that anger, fear, and sadness were linked to more HR acceleration as compared with disgust, based on the meta-analysis that they performed on the impact of discrete emotions on physiological measures.

Sometimes, HR responses might be confusing to be analyzed due to the influence of sympathetic and parasympathetic activity of ANS (Kreibig, 2010). As a solution, some researchers proposed that add-on of an additional component of cardiovascular measures like HRV will help to reduce this issue (Hagemann, Waldstein, & Thayer, 2003). To illustrate this, (Rainville et al., 2006; Wagner et al., 2005) demonstrated that HR, HRV, together with RSP related features could possibly differentiate between four fundamental emotions which were fear, anger, happiness, and sadness.

2.6.2.3 Electromyogram (EMG)

Electromyography (EMG) measures muscle response or electrical activity due to a nerve's stimulus of the muscle. The muscles on the skin of the human face are prominent means of information for facial expression emotion recognition like frowns and smiles (Bradley et al., 2007). Therefore, facial muscle electrical activity can be obtained using EMG (Andreassi, 2007). The muscle above the jaw which is named masseter and the muscle above the eyebrow which is named corrugator are mostly analyzed facial muscles that are considered in emotion recognition researches. The corrugator face muscle is more connected with aroused emotional states like anger and surprise, at the same time the masseter muscle is connected with both arousal and valence elements of human emotion (Lichtenstein et al., 2008). Based on the research done by (Lee, Shackman, Jackson, &

Davidson, 2009), they illustrated steady changes in electrical activity of corrugator muscle measured by facial EMG, in reaction to pleasant or unpleasant stimulation.

2.6.2.4 Skin Conductivity (SC)

Skin conductance (SC) is an index of sympathetic nervous system (SNS) activity and emotional arousal (Lang, 1995; Levenson et al., 1990). Each time an individual feel stress and tension, the palms will get humid because of raised activity in the SNS, which can result in accelerated hydration in the sweat channels and on the outside of the skin in which skin conductivity is increased and then can be measured through skin resistance to a small electrical current (Andreassi, 2007; Cacioppo et al., 2007). Cacioppo et al. (2007) research shown that SC, unlike some other ANS measures, provides a direct reflection of sympathetic triggering once facing stressful circumstances. In addition, according to the experiment performed by (Lichtenstein et al., 2008), SC could distinguish between fear and sad, fear and anger, happy and sad, and additionally, helps to separate between struggle and no struggle situations. SC was typically considered as an index of arousal, however, Lichtenstein's research study revealed that the valence level may also be distinguished by this measure.

2.7 Review on Physiological-based Emotion Recognition System

In previous sections, the basic concepts related to the human emotional states, as well as the existing sensors for collecting the physiological data were discussed. This section reviews some of the prominent researches that have been performed in the area of physiological-based emotion recognition systems. This section discusses, the general process of creating automated physiological-based emotion recognition systems using human autonomic nervous system (ANS) responses and central nervous system responses (i.e. related to brain signals). We provide a review of the existing studies that worked on the development of physiological emotion recognition systems. The emphasis is on two

important phases, which are physiological data preparation for emotion classification (i.e. physiological data collection, feature extraction, normalization and dimension reduction) and different classification methods.

Figure 2.5 depicts the general process for developing a physiological-based emotion recognition system. Each block shows a specific task that must be performed and its resulting output which is mentioned on the left side of the related block. For example, the output of recording physiological signals from a given human subject is the physiological signals of that subject.

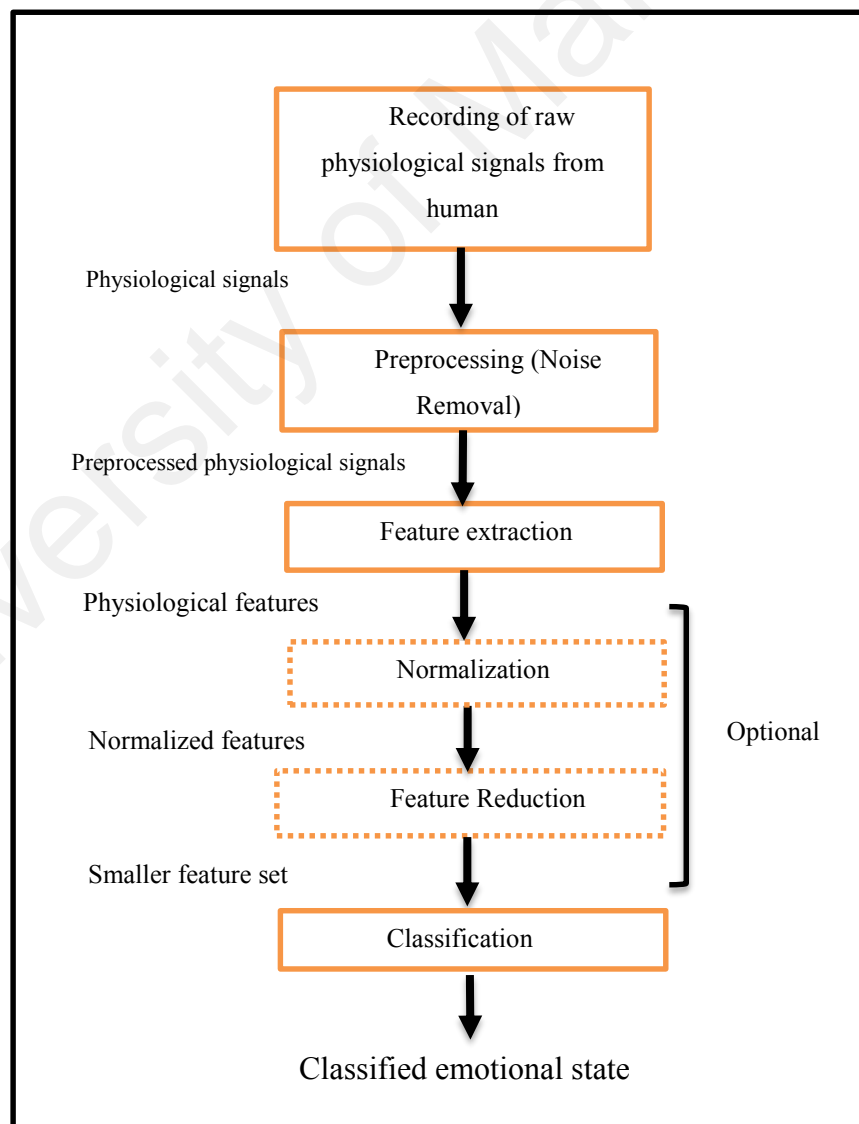


Figure 2.5: The general process to create an automated physiological-based emotion recognition system (Novak, Mihelj, & Munih, 2012)

2.7.1 Physiological data preparation

This section discusses the required steps that should be taken to acquire the final feature set that will be used for emotion classification in physiological-based emotion recognition system. The first step is to choose a suitable emotional model, in the second step, the recording of the raw physiological data from participants and labeling those raw data based on the selected emotional model. For the third step, the most useful features are extracted from the raw physiological signals, followed by normalization stage, which is optional; that is, it can be missed. The last stage, which also can be optional but especially important in building efficient emotion recognition systems, is the feature dimension reduction. Each of these steps is discussed in details in the following subsections.

2.7.1.1 Emotion model selection

The initial phase of developing a database for emotion recognition is to select a proper model that defines the participant's emotional state. Using this model, the emotional states that can be identified from physiological signals are defined. As we explained in section 2.3, the most common emotional models are: categorical model and two-dimensional arousal-valence model. The categorical model (i.e. discrete model) attempts to classify physiological data into one of the several basic emotions such as anger, sadness, surprise, happiness etc. (Ekman, 1992), while in the second model human emotional states are considered as multidimensional so that can be described with multiple variables. The most famous multidimensional emotion model is the arousal-valence model proposed by (Russell, 1980). The valence, which is also called pleasure is known as positive opposed to negative emotional states (e.g. disgrace, boredom, and anger at one side opposed to excitement, relaxation, and calmness at another side), whereas arousal is described in relation to the mental awareness and physical activity

(e.g. sleep, idleness, boredom, and relaxation at the lower end opposed to alertness, strain, exercise, and focus at the upper end) (Mehrabian, 1996). The arousal-valence space can be separated into quadrants as low arousal/positive valence, low arousal/negative valence, high arousal/positive valence and high arousal/negative valence. This split normally performed for classification purposes.

Selection of emotion model as the first step in emotion recognition research studies is very crucial because the emotion model influences every part of research study, from the experiment design to the data examination. It will be almost unfeasible to alter the models as soon as data signal collection has initiated. Though it may be possible, for instance, transform basic emotions to arousal-valence quadrants or the opposite way round (Christie & Friedman, 2004).

2.7.1.2 Physiological data collection

Collecting a high-quality physiological data signals as a data set is among the initial and very crucial steps for the development of emotion recognition systems (Novak, Mihelj, Zihlerl, Olenšek, & Munih, 2011). This physiological data, after passing through some preprocess and feature extraction steps, are employed as training dataset. The physiological data, which can be collected through physiological sensors (electrocardiogram, skin conductance, etc.) are linked with stimulated emotional states like anger, fear, happiness, low stress etc. Ultimately, a supervised data classification technique is commonly used on this training dataset to allow the computers to learn the associations between physiological data and related emotional states as both the inputs and outputs. The learned association forms a model that later it is normally utilized to identify the emotional state linked to the physiological data that is not still identified. The steps of collecting physiological data sets are described in the following subsections.

(a) Emotion Elicitation

To produce appropriate training dataset, the collected physiological data should properly represent the targeted emotional states that had been defined by the emotional model. Therefore, the targeted emotional states ought to be effectively provoked in the subjects to ensure that the training dataset consists of valuable physiological information related to the specific emotional states.

Physiological datasets that utilized in majority of studies were gained through the use of audio-visual emotion stimuli in lab settings in which individuals deliberately express preferred emotions while viewing picked pictures, watching movie clips or listening to music (Frantzidis et al., 2010; Koelstra et al., 2012; Rainville et al., 2006; Wei-Long Zheng, Bo-Nan Dong, & Bao-Liang Lu, 2014). Some studies (e.g. (G. Chanel, Rebetez, Bétrancourt, & Pun, 2011)) have also tried to create databases of natural emotional expressions, where the participants' emotions spontaneously arise as a consequence of significant circumstances in a location that resembles the real world. As an example, in the study by (Scheirer, Fernandez, Klein, & Picard, 2002), in an effort to stimulate frustration emotional experience in subjects for their related physiological data recording, they asked them to play a computer game, in which the target was to finish a variety of graphic puzzles quickly and precisely to win a cash prize. However, determined by random intervals, the computer mouse respond faulty (i.e. once the person clicked to proceed to the subsequent puzzle, nothing happened for some secs). Frustration assumed to take place throughout a multi-second window after every unsuccessful mouse-click. Wilson & Russell (2003) attempted to induce different levels of mental workload and at the same time recording of physiological data related to seven air traffic controllers during a simulated air traffic control task. The participants were required to handle a series of aircraft requesting air traffic control services. Subjects were responsible for controlling Aircraft arriving and departing and overflights. Task difficulty was manipulated in three

conditions: volume, complexity, and overload to stimulate the different level of mental workload in subjects. The just experienced mental workload levels of the subjects were collected through subjects self-report using the NASA Task Load Index (NASA-TLX; (Hart & Staveland, 1988)) at the same time with physiological data recording using biosensors. As another natural emotion induction example, the physiology and the emotional response of participants while interacting with a tutoring system (AutoTutor) are collected. The type of provided feedback by the system (i.e., positive, neutral, negative) has the main effect on the aroused emotional responses of participants (Aghaei Pour, Hussain, AlZoubi, D'Mello, & Calvo, 2010). As another example of natural emotion elicitation is the research done by (Healey & Picard, 2005) that the stress emotional state has been induced in seventeen drivers while they were driving a car and experiencing a different level of stress.

From the theoretical point of view and the researches performed by (Gross & Levenson, 1995), the first method of emotion induction is advantageous because it offers a well-structured database having all emotional states equivalently selected (supposing they have been effectively stimulated), whereas the 2nd method might be more conveniently attainable in utilized settings. However, several emotional states are special and hard to provoke. Consequently, the selection of the emotion induction will mainly depend on the objective of the research (Novak et al., 2012).

Another vital task that is performed at the same time with emotional state induction step is labeling (i.e. define output) of the physiological data recorded from subjects in each session of data collection. The typical approaches for this purpose are self-report approaches (e.g. Christie & Friedman, 2004; Haarmann, Boucsein, & Schaefer, 2009; Lisetti & Nasoz, 2004; Soleymani et al., 2012). The most desirable advantage of them is: being convenient to use and also being inexpensive. The self-report questionnaire better to be validated upfront, since certain strong points as well as weak points will be

identified. There are two widely used and well-validated samples of self-report questionnaires in psychophysiology that are the Self-Assessment Manikin (Bradley & Lang, 1949) and the NASA-TLX (Hart & Staveland, 1988). In Self-Assessment Manikin (Figure 2.6) individuals may put an X over a figure or on a point between any two figures which lead to a 9-point scale for the valence dimension (top panel) and arousal dimension (bottom panel). This was used successfully to measure emotional feedback in several situations including responses to pictures, sounds and other stimuli (Bradley & Lang, 1994).

On the other hand, in some studies, researchers have discovered that some subjects might be unmindful of their particular emotions, not able to state them, or are just reluctant to state them. In this specific circumstances, different techniques such as facial electromyography (Kreibig et al., 2007) or observation of the subject by expert (e.g. (Healey & Picard, 2005; Jones, Buhr, et al., 2014; Jones, Conture, & Walden, 2014; Katsis et al., 2008; Koenig et al., 2011; Liu, Conn, Sarkar, & Stone, 2008a; Schwerdtfeger, 2004) need to be considered to know what emotional state was stimulated in this type of subjects.

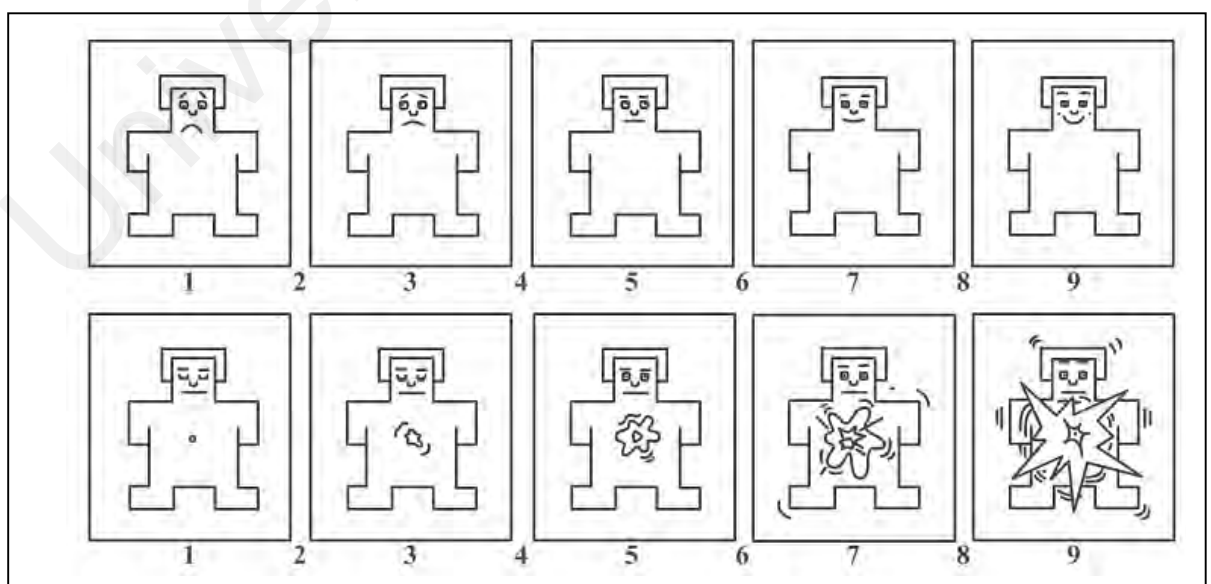


Figure 2.6: Images used by Self-Assessment Manikin (SAM) method for the individual self-report emotional state. Self-evaluation scales for the dimensions of valence (top), arousal (bottom) (Adopted from Hettich et al., 2016).

Another valuable tool that was used by researchers (e.g. (D. Kim, Frank, & Kim, 2014)) for measuring individual's emotional experience is Facial Action Coding System (FACS). This tool has been proposed by Ekman and Friesen (1987), and it is useful for measuring any facial expression a human being can make. Movements of individual facial muscles are encoded by FACS from little different quick changes in facial style. In this system, a total of 46 anatomically-based action units (AU) for facial motions were defined. Every AU gives information about a noticeable change in the face, as an example, AU-12 is related to raising the exterior lip corners, and AU-9 wrinkles the nose. The system represents all possible motions of the face obvious to the naked eye. This system involves comprehensive training and qualification. The FACS has been very beneficial in researches of emotion, for instance, it is able to differentiate between genuine and deceitful smiles (Larsen & Fredrickson, 1999).

(b) Selection of number of subjects (Sample size)

The number of subjects that should be included in a research study related to physiological recording for subject's emotional state recognition is based on the type of validation strategy that is employed which can be subject-dependent or subject-independent. In subject-dependent research studies, the quantity of subjects usually is immaterial considering that the emotional data classification is performed for every subject individually and it is just required to be sure that sufficient data is captured for every subject (e.g. Leon, Clarke, Callaghan, & Sepulveda, 2004; Picard et al., 2001; Wagner et al., 2005). That is why, subject-dependent researches commonly include either only one subject or a few subjects. This strategy is taken into account for a particular or small group of the people with a very short number of obtainable subjects like the study performed by (Liu, Conn, Sarkar, & Stone, 2008b) on 6 autistic children or physiological data recorded from only seven air traffic controllers in (Wilson & Russell, 2003). Having

small numbers of subjects also feasible in studies where the quantity of accessible subjects is somewhat limited. In particular, in the research carried out by (Novak et al., 2011), physiological data of 11 patients that had been having motor rehabilitation were reordered to create a model for offering feedback. Researchers should be aware in these research studies by which several recordings from every subject are mostly essential.

In contrast to subject-dependent validation strategy, subject-independent researches usually include more than 20 subjects in their experiments. As example, (Setz, Schumm, Lorenz, Arnrich, & Tröster, 2009), (G. Chanel et al., 2011), (Kapoor, Burlison, & Picard, 2007) employed between 20 to 24 subjects, (Abadi et al., 2015), (Setz et al., 2010), (Nasoz, Lisetti, & Vasilakos, 2010), (Bailenson et al., 2008), (Ringeval, Sonderegger, Sauer, & Lalanne, 2013) employed between 30 to 46 subjects and (Arroyo-Palacios & Romano, 2010), (Tognetti, Garbarino, Bonanno, Matteucci, & Bonarini, 2010) employed more than 55 subjects in their researches. Having data sets that include many subjects, doesn't mean that subject-dependent systems are not applicable to them. In such systems, subject dependent validation is also possible if enough recordings (i.e. multiple recording) could be collected from each subject.

(c) Physiological-based Emotional Datasets

A vital step toward developing an emotion recognition system is creating an emotional database. Latest advances in the area of automatic emotion identification have inspired the researchers for development of innovative data sets containing emotional expressions.

Most of these data sets consist of speech, visual, or audio-visual data (Fanelli, Gall, Romsdorfer, Weise, & Van Gool, 2010; Grimm, Kroschel, & Narayanan, 2008; McKeown, Valstar, Cowie, Pantic, & Schröder, 2012; Pantic, Valstar, Rademaker, & Maat, 2005). The emotional data sets that have visual modalities includes, face and/or body gestures, while the audio modality contains acted or genuine emotional speech,

which can be in different languages. Despite the availability of several audio-visual emotional datasets that can be used for research purposes, there are quite a few publicly available physiological signal emotional datasets to be employed freely by researchers to evaluate their proposed methods for the problems related to emotion recognition systems. The lack of data sets might be due to the cost and difficulty of collecting physiological data sets which usually involves the use of special equipment and presence of the participants. In this section, we review the publicly available physiological-based emotional datasets and describe their main characteristics.

One of the earliest data sets is MIT dataset, which has been created by (Healey & Picard, 2005). This dataset recorded responses of 17 drivers while they were experiencing various levels of stress by using an electrocardiogram (ECG), galvanic skin response (GSR) collected from drivers' hands and feet, electromyogram (EMG) from the right trapezius, and also the respiration pattern. This dataset is publicly available from the Physionet website (www.physionet.org).

Another large emotional dataset is HUMAINE (Douglas-Cowie et al., 2007) which is comprised of three natural and six persuaded reaction databases. The numbers of participants in each database are different and it ranges from 8 to 125 persons. The modalities that were recorded were various, from audio-visual to peripheral physiological signals. These databases had been created individually at separate sites and gathered under the HUMAINE project.

The AuBT physiological dataset is another dataset, which includes only one subject and has been collected by (Wagner et al., 2005) from Augsburg University in Germany. They produced four basic emotions: Anger, Joy, Pleasure and Sadness using four different music songs. Four peripheral physiological signal channels including electromyogram (EMG), electrocardiogram (ECG), skin conductivity (SC) and respiration change (RSP) are used to record the physiological data while a subject was

listening to the music songs. For each emotion, 25 recordings were collected during 25 days. Totally, they have recorded 100 signals.

The DEAP is a multimodal publicly available dataset for analysis of human emotional state recently created by (Koelstra et al., 2012). It consists of central nervous system signals (i.e. electroencephalogram (EEG)) and peripheral physiological signals which includes electro-cardiogram (ECG), electro-myogram (EMG), electrooculogram (EOG), blood volume pulse (BVP), respiration amplitude (RSP), skin temperature and galvanic skin response (GSR) of 32 subjects while watching 40 one minutes long music videos. Moreover, the face videos of 32 participants also were captured during experiments. A total of 1280 recorded signal were collected according to spontaneous responses of participants to music videos. Each subject rated each video in terms of the levels of arousal, valence, like/dislike, dominance using a 9-point Likert scale.

The MAHNOB-HCI dataset is also another advanced multimodal publicly available dataset for emotion recognition and implicit tagging developed by (Soleymani, Lichtenauer, et al., 2012). Like DEAP dataset, they have done synchronized recording of peripheral and central nervous system physiological signals in addition to face videos, audio signals, eye gaze data of 27 participants according to their response to emotional videos and images. They performed two separate experiments, the first experiment, which was called emotional reactions to videos, where they asked participants to watch 20 emotional videos and report their experienced emotion based on emotional keywords, the level of arousal, valence, and dominance using nine points' scales. SAM Mankins had been used to assist the self- assessment of valence, arousal, and dominance. In the second experiment, which was called implicit tagging, they presented 28 images and 14 short videos from flicker (www.flicker.com) and ask the participants if they agree with the assigned tags. All the signals, videos and bodily responses of participants were recorded and stored in a database.

Recently, (Abadi et al., 2015) presented DECAF as a multimodal data set for emotion recognition from human’s physiological responses. They have done synchronized recording of peripheral physiological signals including horizontal Electrooculogram (hEOG), Electrocardiogram (ECG), and trapezius-Electromyogram (tEMG) as well as central nervous system physiological signals (i.e. brain signal) using Magnetoencephalogram (MEG) in addition to near-infra-red (NIR) facial videos of 30 individuals while watching 36 movie clips and 40 one-minute music video segments (used in (Koelstra et al., 2012)). The creators of data sets claim that using Magnetoencephalogram (MEG) sensor instead of electroencephalogram (EEG) sensors for recording brain signals has the advantage of less physical contact with subject’s scalp and as a result allows for naturalistic emotional responses. In addition, as the advantage of DECAF to other data sets like DEAP or MAHNOB-HCI, it brings the possibility for performance comparison of emotion recognition system using MEG against EEG modalities as well as suitability comparison of using music-video versus movie clips for emotion induction and elicitation. Table 2.1 presents an overview of reviewed datasets and their characteristics.

Table 2.1: An overview of reviewed physiological-based emotional datasets and their characteristics

Dataset	#Participants	Natural / Stimulated	Audio	Visual	Peripheral physiological	EEG	Eye gaze
MIT	17	Natural	No	No	Yes	No	No
HUMAINE	8 to 125	Both	Yes	Yes	Yes	No	No
AuBT	1	Stimulated	No	No	Yes	No	No
DEAP	32	Stimulated	No	Yes	Yes	Yes	No
MAHNOB	27	Stimulated	Yes	Yes	Yes	Yes	Yes
DECAF	30	Stimulated	No	Yes	Yes	MEG	No

2.7.1.3 Pre-processing

In the time of physiological data collection, the raw physiological signals are usually polluted with noises and other external intrusions because of electrostatic devices and muscular movements that influence the raw signals (Kim et al., 2004). Therefore, to have clean data, noise and artifacts have to be eliminated from the raw physiological signal before sending to the next steps (i.e. feature extraction). The methods that are usually employed for preprocessing of raw ECG and EMG signals include different kinds of Low-pass filters such as Adaptive filters, Elliptic filters, Butterworth filters etc. In addition, to preprocess the raw GSR signals smoothing filters techniques also employed (Abadi et al., 2015; Chang et al., 2010; Katsis et al., 2008; Koelstra et al., 2012; G. Rigas, Katsis, Ganiatsas, & Fotiadis, 2007).

2.7.1.4 Feature Extraction

After the signals being pre-processed, the next important step is to extract statistical information that is called features from the signal which later can be employed to identify the emotional content of the signal. There are many types of features that can be categorized into statistical, time domain, frequency domain and time-frequency domain features that can be computed from the different physiological signals.

Some Features can be extracted quite easily like features calculated from skin temperature which just consists of the mean, standard deviation and mean absolute derivative over a specific time frame, whereas extraction of other features like heart rate variability from the electrocardiogram consists of conscientious filtering, peak identification, interpolation, and power spectral density computation (Task-force, 1996)

The outcome of feature extraction step is a vector comprised of a variety of physiological features computed from individual raw signals over a specific time frame. After that, this vector is going to be known as a 'feature vector'. A matrix comprises of numerous feature vectors from various subjects or different periods of time can be called

a 'data set'. Later, in data classification part, these feature vectors are employed to train and test the classification model.

Despite there is no agreement about which features need to be extracted from each physiological signal, there are certain features that are now relatively popular such as mean and standard values of a signal over a specific time frame, minimum and maximum values and mean absolute derivatives over a time period. Besides that, for some signals, specific features are needed to be extracted. For example, heart rate is mostly described by a variety of time and frequency- domain features of heart rate variability that were determined by European Society of Cardiology and the North American Society of Pacing and Electrophysiology (1996) (Abadi et al., 2015; Soleymani, Lichtenauer, et al., 2012). For EEG signals, power spectral features from theta, alpha, beta, and gammas bands are commonly extracted (Koelstra et al., 2012; Soleymani, Lichtenauer, et al., 2012). Since listing here all possible physiological features is not feasible, the readers can refer to the complete list of physiological features that have been collected by (Kreibig et al., 2007) and (Kreibig, 2010). Likewise, computational procedures for the calculation of the following three features: electrocardiography, skin conductance and skin temperature have been explained in details in the studies done by (J Kim et al., 2004). In another study, (Pramila Rani, Sarkar, Smith, & Kirby, 2004) explained the way of extracting electrocardiography, skin conductance, and electromyography features. Almost all features can be computed over a sliding window, however, because of theoretical restrictions some demand for wider windows, for example, certain features of heart rate variability may not be calculated over a window less than 2 mins (Task-force, 1996).

2.7.1.5 Normalization

Psychophysiological data signals and their related extracted features are extremely influenced by person's variability such as age, gender, time of the day along with matters. In features normalization, it is tried to decrease the impact of this variability before the

extracted data features be sent for data classification. For example, in a physiological research, different subjects may possibly express increased reactions than other ones or have diverse resting rates for psychophysiological features like resting heart rate of an individual (adult) can be in possibly be somewhere between 60 and 100 beats in minutes. This issue should be considered ahead of data classification. In addition, in feature extraction stage, various features are calculated in a variety of units, certain features receive bigger numerical values compared to others that can be troublesome for particular classifiers like a nearest-neighbor algorithm. To address these matters, using normalization process tries to lessen this consequence. There are three normalization strategies which are generally applied, nevertheless, it needs to be pointed out that not every psychophysiological research utilizes normalization process and sometimes the researchers do not state if it had been utilized (Novak et al., 2012).

In the first strategy subject's psychophysiological responses are recorded and the related features are extracted in which the subject is not put through stimuli or may be simply put through plain, calming stimuli (i.e. baseline or neutral situations). Then, the psychophysiological features from other situations once the subject is exposed to a serious task or different emotions stimuli are calculated and then these features can be normalized in different ways such as by subtracting from the baseline value (Jones, Buhr, et al., 2014; K. H. Kim et al., 2004; Stephens, Christie, & Friedman, 2010), dividing by the baseline value (Arroyo-Palacios & Romano, 2010; Zhai & Barreto, 2006b), subtracting the baseline value and dividing the result by the baseline value (Kukolja et al., 2014; Mohammad & Nishida, 2010; Nasoz, Alvarez, Lisetti, & Finkelstein, 2004), or an aggregate of them useful for different features (Novak et al., 2010; Setz et al., 2010). The purpose of subtraction scheme using the baseline values is to decrease intersubject variability because of different baseline values belong to different subjects, at the same

time division can be somewhat designed for decreasing variability because of various response sizes.

The second data normalization strategy starts like the first one, where the psychophysiological responses are recorded and related features are extracted in a baseline condition. Then, instead of subtracting or dividing the data features accumulated from exposing subjects to a serious task or different emotions stimuli, the baseline data features are added to the feature space as new features and create a feature vector. Therefore, the dimension of feature space becomes double. This strategy is also called the 'baseline matrix' which have been employed by some researchers like Picard et al. (2001) and Broek et al. (2010).

Another strategy for normalization simply converts the values of features to a certain range for example between 0 to 1 or 1 to -1. The specific range for each feature is calculated separately by, for example, subtracting each feature's value from the mean value of all feature vectors and dividing the result by the standard deviation of all feature vectors. This procedure can be done for each subject independently or across all subjects. When performed for every subject independently, the aim is usually to decrease intersubject variability by adjusting every subject's features values to a difference between their max and minimal rates. In case, it is performed across all subjects, the aim is to make sure that every psychophysiological feature includes the equivalent numerical scope. Applying this normalization approach should not have an impact on the outcomes of data classification only when it comes to employing classification algorithms just like k-nearest neighbors, which demand normalization. For example, Haag et al. (2004), Regan & Atkins (2007), Kulic & Croft (2007), Yannakakis & Hallam, (2008), Sakr et al. (2010), (Soleymani, Lichtenauer, et al., 2012) and (Verma & Tiwary, 2014) have utilized this normalization strategy in their studies.

There is no common agreement among researchers on the effectiveness of data normalization on the improvement of data classification. For example, the third strategy of normalization, which was converting the feature's values between certain ranges is a simple numerical resizing and should not have much impact on data classification, except for classification algorithms like k-nearest neighbors.

Many researchers found that classification results can be improved because of employing normalization approaches. However, some studies also reported that the best results were obtained without employing a normalization approach. Broek et al. (2010) utilized the baseline matrix normalization approach and they claimed minimal improvement on the classification results while at the same study they performed data normalization by subtracting and dividing the baseline (i.e. first normalization approach) and they identified significant enhancement in the classification outputs. In a different study performed by Setz et al. (2010), the assessment of data classification using normalized and non-normalized data features were presented. They noticed that using non-normalized features give better classification results.

2.7.1.6 Feature Dimension Reduction

Once several physiological channels are used to collect physiological responses from human body, normally, many features are extracted from them. Therefore, the size of feature set (dimension) grows easily and quickly where it will be challenging for the classifier to discover patterns and similarities in data. In another side, some extracted features from different physiological channels may possibly not be correlated with the emotion or redundant. Therefore, it is vital to find and eliminate the extracted features that are redundant or might not help the classifier to discriminate between the various emotional states. This uncorrelated features can easily lessen the performance of the classifiers (Kim & André, 2008).

For data classification purpose, a number of feature vectors, in another word, training data set plays an important role in data classification. If the size of the training dataset is too small or the number of feature vectors in training dataset less than number of features, overfitting issue may happen and the classification model obtained from a small dataset may not work well in the classification of the new data. Thus, reducing the number of features before classification task is advantageous (Liu & Motoda, 2008).

Feature dimension reduction methods can be used to identify and remove irrelevant and redundant attributes from data that do not contribute to the accuracy of a predictive model or may in fact decrease the accuracy of the model. Fewer attributes is desirable because it reduces the complexity of the model (Liu & Motoda, 2008). The majority of features dimension reduction methods that have been found in the field of psychophysiology are typically classified into three groups. The first group is known as feature ranking methods which incorporates the methods that select individual features while ignoring the correlation between these features. The second group includes the methods which project the feature space into a lower dimensional space. The methods in the last group select individual features while taking account the correlation between these features. All aforementioned groups are described in more details in the following sections.

(a) Feature Ranking Methods

In the feature ranking methods, also known as filter-based feature selection methods (Brown et al., 2012), the most suitable features for data classification purpose are chosen based on the power of each individual feature (i.e. based on a certain predefined threshold) in providing information for data classification (Jensen & Qiang Shen, 2009).

In field of physiological sciences, the most frequent solution to score particular features have been via analyses of variance, correlations and chi-square tests which are statistical approaches that are able to demonstrate statistically significant differences

between situations (e. g. between ‘sad’ and ‘angry’ emotions) or significant connections between individual parameters will be utilized in data classification. Exclusively those features that indicate statistically significant differences will be employed in data classification process (Novak et al., 2012).

For example, Wagner et al. (2005) used Analysis of variance (ANOVA) test to select most impactful features, in which physiological features were scored depending on their p-value. Broek et al. (2010) and Chanel et al., (2011) also employed ANOVA test in which the features with a p-value below 0.001 and below 0.1 were selected, respectively. Pour et al. (2010) utilized chi-square test for feature selection and they selected 10 most relevant features. Liu et al. (2009), Rani et al. (2007) and Bailenson et al. (2008), used analysis of correlations of features with self-reported psychological variables to select significant features, only they selected those physiological features that obtained an absolute correlation coefficient for a minimum of 0.3. Torres, Orozco, & Alvarez, (2013) used Recursive Feature Elimination (RFE) as feature ranking method for SVM classifier to reduce the dimension of the feature set that allows emotion classification. The minimum redundancy maximum relevance algorithm was used by Clerico, Gupta, & Falk (2015) as feature ranking method to rank features based on their importance and select subset of them (based on their rank) for EEG-based emotion recognition.

The common practice is to use only one feature ranking method but this approach may result in sub-optimal solution because two different feature ranking methods are likely to produce two different ranking sets and presenting only one set given by a particular method can be misleading (Kuncheva, 2007). One solution can be using more than one feature ranking method to increase the chance to choose the optimal feature set.

(b) Principal component analysis and Fisher’s projection

Principal component analysis (PCA) (Jolliffe, 1986) is also a technique for features dimension reduction that converts the original features into a lower space of uncorrelated

features which are known as principal components. Considering that the principal components are uncorrelated to each other, thus, PCA comes with an improvement over techniques from the earlier section which neglect correlations between features (J. Wang et al., 2014). As instances, in the psychophysiological research studies performed by Wagner et al. (2005), Rainville et al. (2006), G. Rigas et al. (2007), Broek et al. (2010), Wang, Nie, & Lu (2014) and (Guo et al., 2016) PCA technique has been employed for features dimension reduction.

Despite having certain advantages, this technique has got one drawback which is that the principal components (i.e. new features) do not guarantee to offer superior correlation with emotional states of the subjects than primary features (Novak et al., 2012).

Fisher's projection is another feature dimension reduction technique that aims to address the drawback of PCA technique which is also known as a supervised replacement for PCA. It projects the primary data into a lower-dimensional space at which distinct classes (e. g. anger, fear . . .) are much easier to linearly be distinguished.

Fisher's projection technique is basically known as a form of linear discriminant analysis (LDA) (Fisher, 1936) that is utilized for features dimension reduction purpose in place of the classification task. Researchers such as Picard et al. (2001), Healey and Picard (2005), Bonarini et al. (2008), Gu et al. (2010), and Kukulja et al. (2014) have utilized this method for feature reduction. Regarding a disadvantage of Fisher's projection technique, as it converts the primary features into less number of new features which have the functionality to identify different classes linearly, thus, it is substantially less recommended to be applied with nonlinear classifiers including support vector machines and neural networks. In addition, these methods namely, PCA and Fisher projection cannot preserve the original domain information such as channels and frequency bands that are very important for understanding the brain and physiological responses (Singh et al., 2013; Wei-Long Zheng & Bao-Liang Lu, 2015).

(c) Wrapper methods

Wrapper methods are dependent-classifier methods which search the space of feature subsets, using a specific measure like the training accuracy rate of a particular classifier to evaluate the utility for a candidate subset. Therefore, the performances of wrapper methods strongly depend on the given classifier. Sequential feature selection techniques (Kittler, 1978) are among famous wrapper methods. Wrapper methods have been applied for physiological-based emotion recognition systems to improve the system's accuracy and reduce its complexity.

(1) Sequential feature selection

Sequential feature selection techniques which are also known as stepwise techniques that select each feature sequentially from the feature space. This technique is different from PCA and Fisher's projection techniques that linearly convert the feature space. They are also different than feature ranking techniques since they will not disregard relations between features.

The most common sequential feature selection technique is a sequential forward selection (SFS) that operates as is stated in the following. At the beginning, there are no any features in the selection. The SFS examines every feature to decide the one that perfectly can distinguish between different classes in the training dataset by utilizing some measures like the F-value of every feature. That feature is picked and included in the selection. Next steps, the remaining features are examined one by one to figure out the one perfectly can distinguish between classes after the additions of all the formerly selected features that were already considered. This process goes on till there are no other features that may provide further information for better classification, to ensure their enclosure to the selection, for example, the F-value of rest of features is below specific rate. Many researchers including Alpers et al. (2005), Wagner et al. (2005), (Yannakakis & Hallam, 2008), Yannakakis & Hallam (2008), Tognetti et al. (2010), Kolodyazhniy et

al. 2011), Muaremi et al. (2013), and Martinez et al. (2013) have used SFS as a feature selection technique for their psychophysiological related studies.

Another alternative and similar technique is sequential backward selection (SBS) technique. While SFS is initiated with the empty selection and puts features in a sequence, SFS starts having all features in the selection and eliminates features in sequence based on which contributes the minimum to distinguish among classes. The procedure proceeds until the total score value of the remaining features reaches a certain threshold value. For example, the F-value of remaining features exceeds a specific rate. Research works like Kim & André (2008), Giakoumis et al. (2011), Kolodyazhniy et al. (2011), and Giakoumis et al. (2013) have used SBS as a feature selection technique for their psychophysiological related studies. Kim and Andre (2008) claimed that SBS outpaced SFS. However, they did not provide quantitative outputs for SFS.

It is possible to combine SFS and SBS techniques and the resulting method is known as sequential floating forward selection (SFFS) or sometimes sequential forward-backward selection. In the beginning, the selection is empty then features are added in consequent similar to SFS technique while in each step it also evaluates if any of existing features in the selection can be removed similar to SBS technique. The well-known criteria for including and excluding a feature to selection is based on F-value thresholds in which the feature with higher F-value will remain in the selection while the lower one is removed. Some researchers have employed SFFS technique to choose the most relevant features in Picard et al. (2001), Wilson & Russell (2003), Gu et al. (2010), Chanel et al. (2011), Singh et al. (2013), Kukolja et al. (2014) and Khezri et al. (2015).

There will be yet another potential for the composition of different feature selection techniques which can be using sequential feature selection techniques along with Fisher's projection where sequential technique is needed at first to select a subset of prominent features then apply Fisher's projection on the obtained subset. Picard et al. (2001) and

Kukolja et al. (2014) experimented the combination of the two techniques and revealed it can do better than using each technique on their own. Wagner et al. (2005) obtained better results by combining both approaches though not for all the classifiers. Lastly, a mixture of the two techniques was employed by Gu et al. (2010) but was not compared with the approaches separately.

The limitation of sequential feature selection techniques is that these methods do not examine all possible subsets, so no guarantee of finding the optimal subset (Jain & Lazzerini, 1999). In addition, the performances of wrapper methods strongly depend on the given classifier because the selected subset of features is used to train a specific classifier and evaluate that selected subset according to the performance of the classifier (J. Wang et al., 2014).

(2) Other wrapper techniques

Other wrapper techniques that were employed infrequently in psychophysiology, for example, are: Davies–Bouldin clustering used by Leon, Clarke, Callaghan, & Sepulveda (2007), the Simba algorithm used by Rigas et al. (2007) and genetic algorithms used by Tognetti et al. (2010). Consequently, using them in psychophysiology and related fields needed further researches before their appropriateness to be used in physiological computing and particularly affective computing.

2.7.2 Classification

After choosing the most relevant physiological features associated with the human emotional state, these features are used to train the classification model. Hence, later, the system will be able to classify different emotional states by using the provided features. There are several classifiers which have been utilized by many researchers for emotion classification including K-Nearest Neighbour (KNN), Support Vector Machines (SVM), Artificial Neural Network (ANN), Naïve Bayes classifier (NB), Regression Tree,

Bayesian Networks, Linear Discriminant and Analysis (LDA). In this section, we focus on the classification algorithms in the related research works that have been conducted in the area of human emotional state classification using human physiological data. Regarding the provided related works, in most cases, the comparison between different classification algorithms is not feasible because of the variation in experimental setups among these studies and the data sets for classification accuracy evaluation.

2.7.2.1 Classification performance evaluation

Measuring the performance of classification is an essential phase for evaluating any pattern recognition system. For physiological-based emotion recognition systems, the testing accuracy rate is the standard measure used for the assessment of these systems.

As can be seen from Table 2.2 to Table 2.7, all the reported studies used the testing classification accuracy (5th column) to evaluate the performance of their proposed systems. The testing accuracy rate is defined as the proportion of correctly classified testing instances to the total number of testing instances (He & Garcia, 2009).

2.7.2.2 Nearest Neighbors

The k-nearest neighbor (kNN) algorithm is among the least difficult classification algorithms. In the case of classification of an unknown feature vector, the kNN algorithm normally calculates the Euclidean or Mahalanobis distance to each feature vector in the training dataset. Through this, the training vectors are ranked based on their distance to the new sample, finally, majority class of the k (where $k \geq 1$) nearest training vectors (neighbors) is utilized to classify the new feature vector. In another word, the determination of assigning a class to the new sample will be based on the class which is most common among k nearest neighbors. It is normally advised to normalize features values between 0 and 1 to ensure all of the features devote similarly to the distance computation. This classification algorithm has been utilized in quite many research

studies related to psychophysiology and this may be due to its simplicity. Some well-known studies that have utilized this algorithm in their research are shown in Table 2.2.

University of Malaya

Table 2.2: Physiological-based emotion classification studies that used k-nearest classification algorithm

Study	Classification for	No of subjects/Name of Database	Feature selection / Feature reduction	Accuracy Rate	Signal Channels
(Picard et al. 2001).	8 Basic Emotions	1	Sequential Floating Forward Search & Fisher Projection	65%	Peripheral Signals
(C. Lisetti, Nasoz, LeRouge, Ozyer, & Alvarez, 2003)	5 Basic Emotions	10	Not Mentioned	70% to 90%	Peripheral Signals
(Nasoz et al., 2004)	6 Basic Emotions	29	Not Mentioned	67% to 87% Ave:72%	Peripheral Signals
(Wagner et al., 2005)	4 Basic Emotions	1	-SFS -Fisher -SFS+Fisher	79.55% to 90.91%	Peripheral Signals
(G. Rigas et al., 2007)	3 Basic Emotions	9	Principal Component Analysis (PCA)	62.5%	Peripheral Signals
(Nasoz et al., 2010)	4 Basic Emotions	34	Not Mentioned	65%	Peripheral Signals
(Kolodyazhniy et al., 2011)	3 Basic Emotions	34	SFS SBS	79.4%	Peripheral Signals
(Shen, Wang, & Shen, 2009)	High/Low Arousal-Valence (4 cases)	1	Not Mentioned	Peripheral:60.3 Peripheral+EEG signals:75.2	Peripheral Signals & EEG

Study	Classification for	No of subjects/Name of Database	Feature selection / Feature reduction	Accuracy Rate	Signal Channels
(Broek et al., 2010)	High/Low Arousal-Valence (4 cases)	21	- ANOVA - PCA	61.3%	Peripheral Signals
(Bonarini et al., 2008)	5 Stress levels	6	Fisher's projection	88.1%	Peripheral Signals
(C. Liu, Agrawal, Sarkar, & Chen, 2009)	3 Anxiety levels	15	Not Mentioned	80.4%	Peripheral Signals
(Levillain, Orero, Rifqi, & Bouchon-Meunier, 2010)	2 Amusement levels	25	Not Mentioned	77%	Peripheral Signals
(Verma & Tiwary, 2014)	13 Emotions	32 (DEAP Dataset)	Not Mentioned	57.74%	Peripheral Signals & EEG
(Chen et al., 2015)	2 levels of Arousal-Valence	32 (DEAP Dataset)	ANOVA	66.45%	EEG
(Khezri et al., 2015)	6 Basic emotions	25	SFFS	80%	Peripheral Signals & EEG

2.7.2.3 Naïve Bayes classifier

A Bayesian network is known as a probabilistic model of random parameters and their conditional dependencies. Nevertheless, the naïve Bayes classifier is considered as a basic type of Bayesian network, which considers all of that parameters are independent of another.

Throughout the training process, a probability model is generated that is utilized to evaluate the possibility that a feature vector is a member of a particular class. Then, a decision rule is employed to associate a class to the feature vector based on the probability model. The ‘maximum posteriori’ rule is the most popular rule employed to classify a feature vector based on the class with the maximum posterior probability.

Considering that naïve Bayes classifier assumes independence between features, it will take a smaller sized set of training data in comparison to complicated techniques. Consequently, this may be viewed as an advantage for this technique. On the other hand, in several researches related to physiological emotion recognition, more complicated classification Bayesian networks have been utilized, which tend not to assume that features are independent. Some well-known studies that have utilized Naïve Bayes classifier in their research related to the physiological emotion recognition are shown in Table 2.3.

Table 2.3: Physiological-based emotion classification studies that used naïve Bayes classification algorithms

Study	Classification for	No of subjects/Name of Database	Feature selection / Feature reduction	Accuracy Rate	Signal Channels
(Picard et al. 2001)	8 Basic Emotions	1	Sequential Floating Forward Search & Fisher Projection	81.3%	Peripheral Signals
(Zhai & Barreto, 2006a)	2 Stress Level	32	Not Mentioned	78.7%	Peripheral Signals & Eyes data
(Muller, 2006)	4 class of Arousal-Valence	1 (AUBT)	Not Mentioned	86%	Peripheral Signals
(Calvo, Brown, & Scheduling, 2009)	8 Basic Emotions	3	Not Mentioned	All sessions:43.6% One session:66.3	Peripheral Signals
(George Rigas, Goletsis, Bougia, & Fotiadis, 2011)	3 class of Fatigue and 2 class of Stress Level	1	Feature Ranking (based on metric of discrimination power of a feature)	Fatigue (74%) Stress (66%)	Peripheral Signals & Face video
(P. Rani & Sarkar, 2005)	3 levels of 5 basic emotions	15	Feature Selection based on high correlation with particular emotion	74.03%	Peripheral Signals

Study	Classification for	No of subjects/Name of Database	Feature selection / Feature reduction	Accuracy Rate	Signal Channels
(Kapoor et al., 2007)	2 levels of Frustration	24	Not Mentioned	79 %	Video camera, pressure-sensitive mouse, skin conductance sensor, and pressure sensitive chair.
(C. Liu et al., 2009)	3 Anxiety levels	15	Not Mentioned	80.6%	Peripheral Signals
(Calvo et al., 2009)	8 Basic emotions	3	Not Mentioned	-All sessions:64.3% -One session:81.3%	Peripheral Signals
(Koelstra et al., 2012)	2 levels of Arousal-Valence and liking	32 (DEAP)	Fisher's linear discriminant	Arousal-EEG (62%) Valence-EEG (57.6%) Arousal-Peripheral signals (57%) Valence-Peripheral signals (62.7%)	Peripheral Signals & EEG

2.7.2.4 Discriminant Analysis

Discriminant analysis techniques proposed by Fisher (1936) (e.g. linear discriminant analysis (LDA) or quadratic discriminant analysis (QDA)) are prominent classification techniques that discover linear composition of input features that are able to differentiate feature vectors into two or more classes perfectly. This input features composition is hyperplane in n-dimensional (i.e. n is the number of input features) area which can identify between feature vectors of diverse classes. This technique is primarily useful for two class situations. However, it may be expanded to several class scenarios. The major drawback of the technique is that it just permits linear or quadratic connections between input and outputs, Therefore when there is solid nonlinear connections exist between the data, using other classification techniques are preferred. Due to the fact that using discriminant analysis techniques is convenient and visibly reveals the involvement of every feature to discrimination among classes, it has become widely used classification technique in physiological-based emotion recognition systems. The summary of some well-known studies that have utilized discriminant analysis techniques in their research related to physiological emotion recognition is listed in Table 2.4.

2.7.2.5 Support vector machines

Support vector machines (SVMs) (Cortes & Vapnik, 1995) are much like discriminant analysis techniques which operate based on creating hyperplanes in the n-dimensional area to divide feature vectors to diverse classes. But different criteria are applied to estimate these hyperplanes of the two techniques. Even though LDA works to maximize a discriminative projection, SVM creates a hyperplane where in both sides the distance between the hyperlane and nearest feature vectors is maximized.

Due to characteristics of traditional SVMs and discriminant analysis techniques, the advantages and disadvantages are usually similar where both have been conveniently used in identifying the contribution of each input feature. Both are also linear classifiers

that could be a drawback of those techniques. To be able to prevent this constraint, SVMs can be extended utilizing kernels. The SVMs with nonlinear characteristics have brought on superior performance, therefore this has resulted in their common use in physiological computing. Table 2.5 presents some well-known examples of studies that have utilized this technique.

University of Malaya

Table 2.4: Physiological-based emotion classification studies that used discriminant analysis classification algorithms

Study	Classification for	No of subjects/Name of Database	Feature selection / Feature reduction	Accuracy Rate	Signal Channels
(C. Lisetti et al., 2003)	5 Basic Emotions	10	Not Mentioned	70%-90%	Peripheral Signals
(C. L. Lisetti & Nasoz, 2004)	6 Basic Emotions	29	Fisher Projection	75%	Peripheral Signals
(Christie & Friedman, 2004)	7 Basic Emotions	34	Not Mentioned	37.4%	Peripheral Signals
(Wagner et al., 2005)	4 Basic Emotions	1	SFS	92.05%	Peripheral Signals &EMG
(Rainville et al., 2006)	4 Basic Emotions	43	Not Mentioned	49%	Peripheral Signals
(Kreibig et al., 2007)	3 Basic Emotions	28	Not Mentioned	69%	Peripheral Signals
(Kolodyazhniy et al., 2011)	3 Basic Emotions	34	SFS and SBS	Dep:77% Ind:73.5%	Peripheral Signals &EMG
(Guillaume Chanel, Kierkels, Soleymani, & Pun, 2009)	3 areas of Arousal-Valence space	10	Not Mentioned	Peripheral Signals:51% EEG:70%	Peripheral Signals &EEG
(Healey & Picard, 2005)	3 levels of Stress	9	Not Mentioned	97.4%	Peripheral Signals &EMG

Study	Classification for	No of subjects/Name of Database	Feature selection / Feature reduction	Accuracy Rate	Signal Channels
(Giakoumis et al., 2011)	2 levels of Boredom	19	Fisher Projection	Dep:94.17% Ind:89.4%	Peripheral Signals
(Setz et al., 2010)	Differentiate of Stress from Cognitive load	33	Wrapper approach	82.8%	Peripheral Signals
(Alpers et al., 2005)	Phobic & non-phobic	38	Not Mentioned	95%	Peripheral Signals
(Setz et al., 2009)	4 or 5 Basic Emotions	20	Not Mentioned	4 Emotions:58.8% 5 Emotions:49%	Peripheral Signals & EMG & EOG
(James Kim & André, 2008)	4 areas of Arousal-Valence space	3	SBS	Dep:95% Ind:70%	Peripheral Signals & EMG
(Giakoumis et al., 2013)	Low/high stress levels	24	SBS	94.96%	SC and ECG
(Jang et al., 2015)	3 Basic Emotions	217	Not Mentioned	74.9%	Peripheral Signals
(Jenke et al., 2014)	5 Basic emotions	16	mRMR, Relief	25.0 to 47.5	EEG

Table 2.5: Physiological-based emotion classification studies that used support vector machine classification algorithms

Study	Classification for	No of subjects/Name of Database	Feature selection / Feature reduction	Accuracy Rate (%)	Signal Channels
(K. H. Kim et al., 2004)	3 or 4 Basic emotions	50	No	3 classes:78.4 4 classes:61.8	Peripheral Signals
(Katsis, Ganiatsas, & Fotiadis, 2006)	5 Basic emotions	4	Not mentioned	86	Peripheral Signals &EMG
(Katsis et al., 2008)	4 Basic emotions	10	Not mentioned	79.3	Peripheral Signals
(Calvo et al., 2009)	8 Basic emotions	3	Not mentioned	85.7	Peripheral Signals
(Pour, Hussain, AlZoubi, D'Mello, & Calvo, 2010)	2 Basic emotions	16	chi-square	42-84	Peripheral Signals
(Broek et al., 2010)	4 areas of Arousal-Valence space	21	-ANOVA+PCA	60.7	Peripheral Signals
(G. Chanel et al., 2011)	3 difficulty levels	20	-ANOVA+ fast Correlation-based filter(FCBF)+SFFS	56	Peripheral Signals &EEG
(George Rigas et al., 2011)	2 levels of stress and 3 levels fatigue	1	DAUC (area under curve) feature ranking	Stress:78 Fatigue:85	Peripheral Signals
(Wu et al., 2010)	3 levels of Arousal	18	SFS	96.5	Peripheral Signals &EEG
(Setz et al., 2010)	2 levels of stress or cognitive load	33	A wrapper method	81	Electrodermal activity (EDA)

Study	Classification for	No of subjects/Name of Database	Feature selection / Feature reduction	Accuracy Rate (%)	Signal Channels
(Soleymani, Lichtenauer, et al., 2012)	3 levels of Arousal and Valence	27	ANOVA	Arousal-EEG (52.4%) Valence-EEG (57%) Arousal-Peripheral signals (46.2%) Valence-Peripheral signals (45.5%)	Peripheral Signals & EEG
(X.-W. Wang et al., 2014)	2 levels (Positive and negative emotions)	6	-PCA -LDA -Correlation-based feature selector (CFS).	91.77	EEG
(Abadi et al., 2015)	2 levels of Arousal and Valence	30	Fisher	Arousal-MEG (60%) Valence-MEG (61%) Arousal-Peripheral signals (55%) Valence-Peripheral signals (60%)	Peripheral Signals & MEG

Study	Classification for	No of subjects/Name of Database	Feature selection / Feature reduction	Accuracy Rate (%)	Signal Channels
(Verma & Tiwary, 2014)	13 emotions	32 participants (DEAP dataset)	Not Mentioned	81.45%	Peripheral Signals & EEG
(Kukolja et al., 2014)	5 Basic emotions	14	-SFFS+Fisher projection	57.61%	Peripheral Signals
(Khezri et al., 2015)	6 Basic emotions	25	SFFS	84.7%	Peripheral Signals & EEG
(Wei-Long Zheng & Bao-Liang Lu, 2015)	4 profiles of EEG electrodes sets	15	Not Mentioned	83.99%	EEG

2.7.2.6 Classification Trees

In classification trees, as a result of many branching of IF–THEN logical rules, a class is allocated to each feature vector. Because of branching composition, they are simply known as trees. An illustration of one psychophysiological classification tree rule could possibly be “if skin conductance feedback frequency is below five per minute, the subject is bored”. These rules are not determined manually, there are some algorithms that can assist to extract the rules from training data. These particular algorithms at every fresh node of tree, choose the most relevant feature which will be capable of discriminating between classes.

Classification trees operate in the upfront path for classification of physiological feature vectors. Since the trees can be visualized graphically, these classifiers help the users to track the decision-making process quite easily. Furthermore, the process for constructing the tree may possibly work as kind of features dimension reduction technique because several trees constructing algorithms are involved with tree pruning. It avoids the tree turning into overly complicated and overfitting the data.

Table 2.6 presents some well-known studies that have utilized classification trees in physiological or affective computing.

2.7.2.7 Artificial neural networks

The artificial neural networks (ANNs) are comprised of numerous interrelated elements which are usually known as neurons that function in parallel. Every neuron gets a variety of inputs and makes use of those to compute the ‘activation’ of the neurons. The output of each neuron is next given to the subsequent layer of neurons and this process continues until calculating the final output. This layered network that comprises of weighted sums and threshold is usually named dual-layer perceptron. As long as sufficient layers and neurons are being employed, the dual-layer perceptrons will be able to model the operations with high complication. There are other forms of ANNs that

combine more complex components in their composition. For example, once the output of a single layer of neurons is considered as inputs of each previous and subsequent layers, the complex network is created which have been known as a feedback network.

ANNs are trained to accomplish a specific task by employing a training data set by changing the weights of the connections between various neurons. ANNs could be linear as well as non-linear methods and suitable for modeling highly complicated connections between features, that will be very efficient in affective /physiological computing.

Probably the main drawback of ANNs is a lack of transparency. ANNs will not give many details about the fundamental system to the users, once the network has been trained, it may not be obvious to see exactly how various variables (i.e. features) produce a specific output. Regardless of this drawback, ANNs have been commonly utilized for the classification purpose based on physiological data. Table 2.7 lists the summary of these studies.

Table 2.6: Physiological-based emotion classification studies that used classification tree algorithms

Study	Classification for	No of subjects/Name of Database	Feature selection / Feature reduction	Accuracy Rate (%)	Signal Channels
(Pramila Rani, Liu, Sarkar, & Vanman, 2006)	3 levels of 5 basic emotions	15	A person-specific correlated features (Statistical method)	83.5%	Peripheral Signals
(G. Rigas et al., 2007)	3 Basic emotions	9	Simba algorithm	62.4%	Peripheral Signals
(Calvo et al., 2009)	8 Basic emotions	3	Not Mentioned	89%	Peripheral Signals
(C. Liu et al., 2009)	3 Anxiety levels	15	Not Mentioned	88.5%	Peripheral Signals
(Levillain et al., 2010)	2 Amusement levels	25	Not Mentioned	75.9%	Peripheral Signals
(Mohammad & Nishida, 2010)	2 Classes of behavior naturalness	44	ANOVA	79%	Peripheral Signals
(Plarre et al., 2011)	2 Classes of stress	21	Correlation-based feature selection algorithm (CFS)	90.2%	Peripheral Signals
(Y.-H. Lee et al., 2014)	3 Classes of meditation experience	10	Not Mentioned	79%	EEG
(Chen et al., 2015)	2 levels of Arousal and Valence	32 (DEAP Data Set)	ANOVA	69.09%	EEG

Table 2.7: Physiological-based emotion classification studies that used neural network classification algorithms

Study	Classification for	No of subjects/Name of Database	Feature selection / Feature reduction	Accuracy Rate (%)	Signal Channels
(C. L. Lisetti & Nasoz, 2004)	6 Basic Emotions	29	Fisher Projection	84%	Peripheral Signals
(Wagner et al., 2005)	4 Basic Emotions	1	Hybrid SFS and Fisher	88.6%	Peripheral Signals
(Muller, 2006)	4 classes of Arousal-Valence	1 (AUBT)	Not Mentioned	81%-86%	Peripheral Signals
(Yannakakis & Hallam, 2008)	2 classes of entertainment preferences	72	nBest, SFS	79.8% (SFS) 70.26% (nBest)	Peripheral Signals
(Calvo et al., 2009)	8 Basic Emotions	3	Not Mentioned	All sessions:97.8% One session:97.1%	Peripheral Signals
(Broek et al., 2010)	4 areas of Arousal-Valence space	21	ANOVA+PCA	56.2%	Peripheral Signals
(Arroyo-Palacios & Romano, 2010)	4 areas of Arousal-Valence space	59	Not Mentioned	78.4%	Peripheral Signals
(Kolodyazhniy et al., 2011)	3 Basic Emotions	34	SFS	77.5%	Peripheral Signals
(Singh et al., 2013)	3 Levels of stress	20	Variance filter and combination of SFS&SBS(SFFS)	89.23%	Peripheral Signals
(Kukolja et al., 2014)	5 Basic emotions	14	SFFS	60.30%	Peripheral Signals

Study	Classification for	No of subjects/Name of Database	Feature selection / Feature reduction	Accuracy Rate (%)	Signal Channels
(Verma & Tiwary, 2014)	13 emotions	32 participants (DEAP dataset)	Not Mentioned	74.37%	Peripheral Signals & EEG
(Chen et al., 2015)	2 levels of Arousal and Valence	32 participants (DEAP dataset)	ANOVA	65%	EEG

2.7.2.8 Ensemble Classification

In ensemble classification, multiple classifiers' decisions (i.e. prediction results) are combined, usually by using majority vote method, to obtain the final classification output of a given testing pattern. The majority voting is a final decision rule that selects one of the several choices, based on the predicted classes with the highest votes (Lam & Suen, 1997). The ensemble method is generally more accurate compared to single classifiers (Novak et al., 2012). One possible interpretation of this superiority is that errors made by each of the classifiers are not identical and if we combine multiple classifier outputs in an efficient manner, we may be able to correct some of these errors (Leo Breiman, 1996). Figure 2.7 depicts the main concept of an ensemble classifiers.

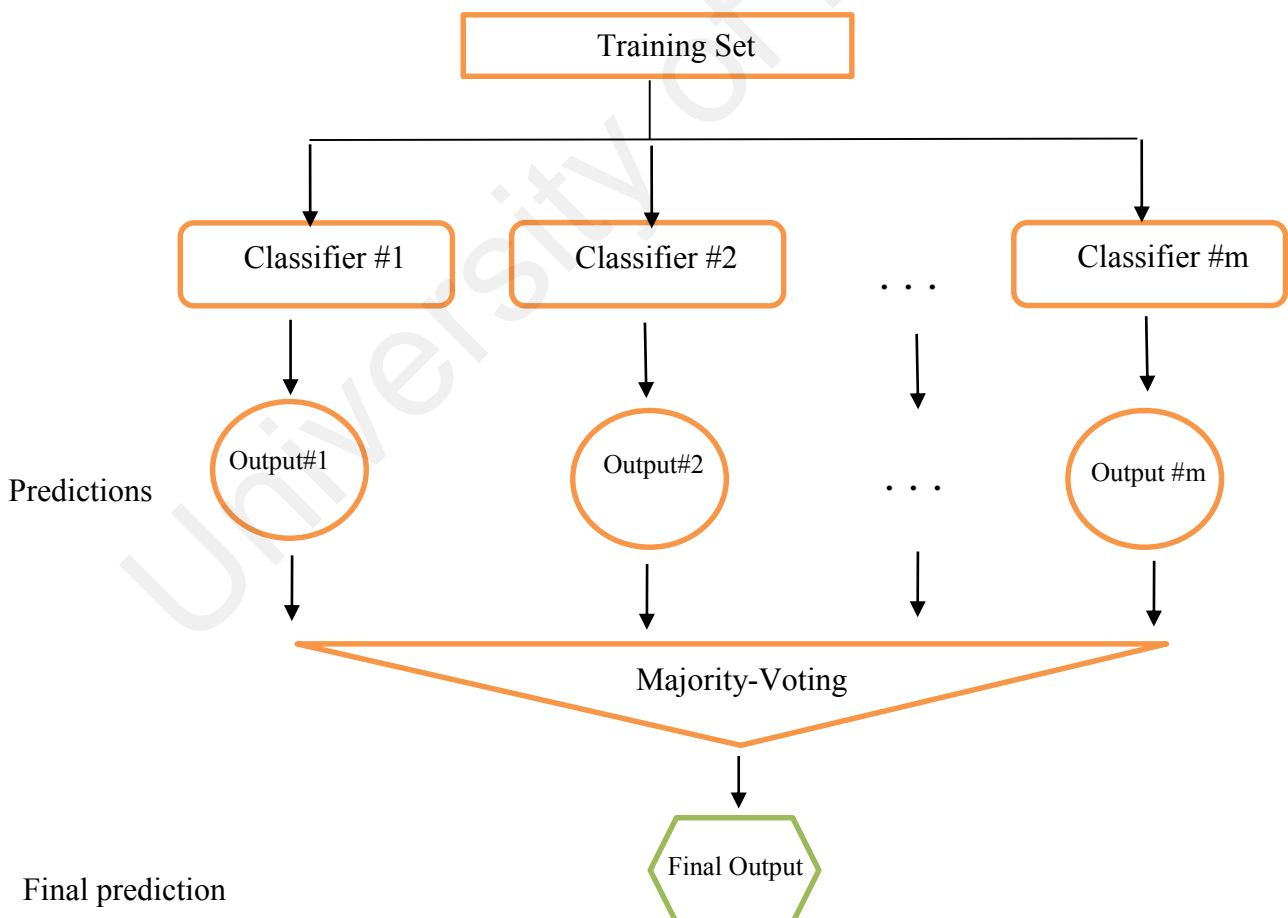


Figure 2.7: The concept of an ensemble classification (Witten, Frank, & Hall, 2011)

Ensemble classification techniques have not been utilized widely in affective computing and particularly in emotion recognition. However, since in most cases combination of different modalities such as peripheral data, brain data (i.e. EEG) and even speech signals are utilized for physiological based emotion recognition systems, ensemble classification may be used naturally. For example, classification outputs in emotion recognition systems can be obtained using data from different modalities separately and then these outputs are combined to get final classification output. This technique is also known as decision level fusion technique and has been employed for physiological-based emotion recognition by researchers like (Jonghwa Kim, 2007), Chanel et al. (2009, 2011) and Soleymani, Pantic, & Pun (2012). The fusion technique can also be applied on sensors level in which the classification can be trained on features extracted from each sensor like ECG, SC or EMG individually and then combines their outputs to get the final classification decision.

Another category of ensemble methods includes both Bagging and Boosting (Rokach, 2010). These two methods create a set of classifiers by manipulating the training data set in order to generate different training sets and then train the classifiers on the generated training sets. The final classification results will be obtained through applying majority voting on the results of all classifiers. The researches have been performed by (Bailenson et al., 2008; Colomer Granero et al., 2016; Plarre et al., 2011; G. Rigas et al., 2007; Takahashi, Namikawa, & Hashimoto, 2012) are examples of using such ensemble classification method on physiological data. Table 2.8 depicts a summary of some prominent studies that have utilized different ensemble techniques in their research related to physiological emotion recognition. Reviewing these studies show that overall, all ensemble classifiers outperformed the single classifiers. In addition, using feature

selection methods along with an ensemble classification method has a positive impact on the final classification accuracy (Diao, Chao, Peng, Snooke, & Shen, 2014). Diao, Chao, Peng, Snooke, & Shen (2014) have also proposed a method that uses feature selection technique to support classifier ensemble reduction (CER), by transforming ensemble predictions into training samples, and treating classifiers as features. The aim was to reduce the amount of redundancy in a reconstructed classifier ensemble, to form a much reduced subset of classifiers that can still deliver the same classification results. Obtaining a reduced amount of classifiers will prevent a portion of run-time overheads, making the ensemble processing a lot quicker; low memory and storage demands. Eliminating redundant ensemble members using feature selection may also lead to enhanced diversity within the group, and maximize the prediction accuracy of the ensemble.

Table 2.8: Summary of studies utilized different ensemble techniques for physiological emotion recognition

Study	Classification for	No of subjects/Name of Database	Feature selection / Feature reduction	Accuracy Rate	Signal Channels
(Plarre et al., 2011)	2 Classes of stress	21	Correlation-based feature selection algorithm (CFS)	SVM: 89.17% J48 Decision Tree: 87.67% J48 with Adaboost (ensemble classifier):90.17%	Peripheral Signals
(Kuncheva, Christy, Pierce, & Mansoor, 2011)	2 Emotions (positive, negative)	1	Not Mentioned	Average Accuracies: ANN:61.46% CART:63.27% SVM:60.80% Bagging: 66.69% RF ³ :65.6% Adaboost:65.96%	EEG, GSR, and Pulse Reader (kind of heart signal)

³ Random Forest (RF) is kind of ensemble classifier

Study	Classification for	No of subjects/Name of Database	Feature selection / Feature reduction	Accuracy Rate	Signal Channels
(Takahashi et al., 2012)	2 Emotions (positive, negative)	13	SFS	Without FS ⁴ : SVMs Bagging:52% LDAs Bagging:45% MNNS ⁵ Bagging:48.5% Decision Tree Bagging: 57% With FS: SVMs Bagging: 56% LDAs Bagging: 57% MNNS Bagging:50% Decision Tree Bagging:52%	Peripheral Signals

⁴ FS=Feature Selection

⁵ Multilayer neural networks (MNNs)

Study	Classification for	No of subjects/Name of Database	Feature selection / Feature reduction	Accuracy Rate	Signal Channels
(AlZoubi et al., 2014)	Low/high level of Arousal and Valence	4	feature ranking (Chi-square)	SVM with FS: 53.25% (Valence) 50.75% (Arousal) Winnow ensemble algorithm with FS: 71.75% (Valence) 72.25% (Arousal)	Peripheral Signals
(Vaid et al., 2015)	4 Emotions happy, sad exciting and hate	32 DEAP	Not Mentioned	Overall for classification of emotions (happy, sad, exciting, hate): RF: 98.1% ANN: 46.3% KNN:69.6% SVM:50.5%	EEG

Study	Classification for	No of subjects/Name of Database	Feature selection / Feature reduction	Accuracy Rate	Signal Channels
(Colomer Granero et al., 2016)	3 Emotions (positive, negative and neutral)	47	A wrapper method	<p>Naive Bayes, Logistic Regression, Multilayer Perceptron, Support Vector Machines, Random Forest and Bagging.</p> <p>The Best obtained Average Accuracies:</p> <p>EEG: 79.52% (RF, Bagging) RSP:69.84%(RF) HRV:79.95% (RF, Bagging) GSR:77.33% (RF) EEG+GSR+HRV:81.90% (RF, Bagging) (GSR+HRV)+FS:87.62% (RF)</p>	Peripheral Signals and EEG

(a) ***Dual-Layer Ensemble Classification (Stacking Ensemble Classification)***

In fact, there are two approaches for combining classification models (i.e. results of all classifiers). One of them uses voting in which the class predicted by a majority of the models is selected, another one is stacking where the predictions by each different model are given as input for a meta-layer classifier whose output is the final class (Wolpert, 1992). In another word, by using the first layer, a meta-dataset containing a tuple related to each feature vector in the original dataset is created. The second layer uses the predicted classifications by the classifiers in the first layer (meta-dataset) as the input features for second layer classification. The target feature remains as in the original training set. A test instance is first classified by each of the base (first layer) classifiers and second layer classifier combines the different predictions into a final one. Consequently, the meta-classifier predictions reflect the true performance of base-layer learning algorithms (Džeroski & Ženko, 2004; Rokach, 2010).

The basic difference between stacking and voting is that in voting no learning takes place at the meta level, as the final classification is decided by the majority of votes casted by the first layer's classifiers whereas in stacking learning takes place at the meta level. By using a second layer classification, this method tries to induce which classifiers are reliable and which are not (i.e. specify which classifier should be used to obtain a prediction (Zenko, Todorovski, & Dzeroski, 2001)). For example, if a classifier steadily misclassified instances from one region because of incorrectly learning the feature space of that region, the Meta classifier may be able to discover this problem. Therefore, utilizing the learned behaviors of other classifiers, it may enhance this kind of training problems (Zhu, 2010).

Figure 2.8 depicts a general schematic of stacking ensemble classification method.

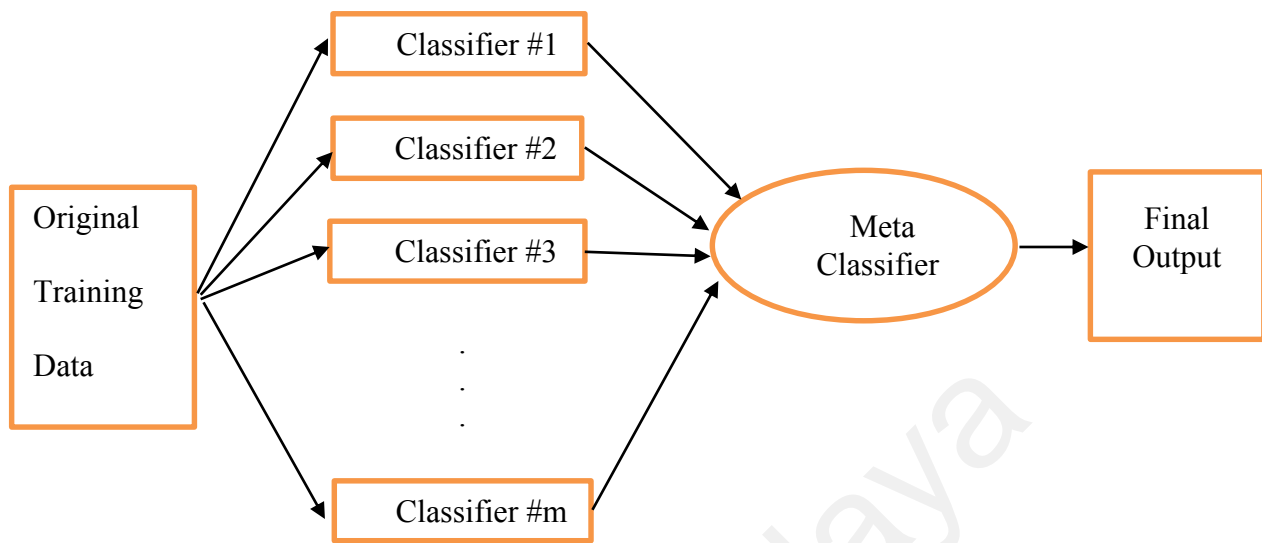


Figure 2.8: The concept of stacking ensemble classification (Witten et al., 2011)

It is claimed that stacking method is particularly more perfect for combining various types of classification models. The limitation of majority voting methods is that they are only able to capture linear relationships. While in stacking methods, the idea is that the meta-dataset provided to the Meta-learning algorithm (second layer classification) adjusts the errors in such a way that the classification of the combined model is optimized (Opitz & Maclin, 1999; Zhu, 2010).

To the best knowledge of author, the stacking technique of combination results of classifiers in physiological-based emotion recognition is a relatively new and untried area. Table 2.9 shows a summary of some research studies that used stacking ensemble classification in other domains. The obtained results in the reviewed studies show that ensemble method can improve classification performance as compared to a single classifier. In addition, among ensemble methods stacking in some studies received better performance compared to other ensemble methods. However, the results it highly depends on the characteristics of the data used. It seems that feature selection methods have not been used widely in the reviewed studies related to the application of stacking

ensemble method, one possible idea can be the number of features used in these studies were limited where feature selection won't bring much improvement to the results of ensemble method.

In the case of having high dimensional data, like physiological data, the using of multiple feature selection methods will increase the diversity but within certain boundary allowed by feature sets found by these methods. In this case, the diversity created here is directed and more meaningful as the features found are more relevant, even if they are not similar among different methods, compared to the ones (i.e. feature subsets) created by methods like random subspace ensemble method (Leo Breiman, 2001; Rokach, 2010). This will be significant since the performance of the ensemble systems depend on the quality (accuracy and diversity) achieved by the individual components which will increase disagreement among these components that eventually enhance classification performance of ensemble system (Oliveira, Morita, & Sabourin, 2006; Optiz, 1999; Santana & Canuto, 2014).

In ensemble systems, in the case of having high dimensional data, choosing the proper feature selection methods is crucial. The wrapper category of feature selection methods, as we mentioned earlier (Section 2.7.1.6), have the main limitation of being time-consuming and highly depend on the classifier used. It is stated that the feature selection methods used in feature-based ensemble structure should be fast, simple, and robust. Hence, the filter-based category of feature selection methods such as feature ranking methods (explained in Section 2.7.1.6) tends to be a remarkable (and efficient) preference for selecting features in ensembles where the computational cost is of a great importance (Santana & Canuto, 2014).

Table 2.9: Summary of some research studies that employed stacking ensemble classification in other domains

Author/year	Domain	Classification for	Feature selection method	Technique used & Results
(G. Wang, Hao, Ma, & Jiang, 2011)	Credit Scoring	2 Classes (Risk and Non-Risk) - (3 credit data sets)	Not Mentioned	Single classifiers: SVM (76.53%), ANN (75.28%), Decision Tree (DT) (78.11%), Logistic Regression Analysis (LRA) (78.26%) Ensemble methods: Boosting (DT):79.5%, Bagging (DT):80.76%, Stacking :80.38%
(Syarif et al., 2012)	Network Intrusion Detection	4 categories of intrusion (NSL-KDD intrusion data set)	Not Mentioned	Single classifiers: NB: 55.77% J48 (decision tree): 63.97% JRip (rule induction): 63.69 iBK (nearest neighbour):62.84% Ensemble methods: Bagging (64.51%), Boosting (37.60%) and Stacking (67.9%)
(Chanamaran, Tamee, & Sittidech, 2016)	Academic Achievement Prediction	2 Classes of graduate or Not graduate	Not Mentioned	Single classifiers: SVM (86.86%), ANN (81.02%), Decision tree (86.57%) Ensemble classifier: Stacking:87.1%
(Hussain et al., 2015)	Software Fault Prediction	Fault proneness classes (12 data sets)	Not Mentioned	Single Classifiers: NB, Logistic, J48, Votedperceptron and SMO Ensemble methods: Voting, Stacking, and Adaboost Results: Ensemble methods performed better than single classifiers, and among ensemble methods Stacking outperformed other selected ensemble methods.
(Cárdenas-Gallo et al., 2017)	Predict geometry degradation	3 different Geometric defect types	Not Mentioned	Single classifiers: Binary Logistic Regression(BLR):78.01% SVM:76.52% Ensemble classifiers: BLR Stacking: 78.88% SVM Stacking: 78.69%

This research work is also proposing an efficient combination of base and Meta level classifiers using a combination of feature selection methods for an emotion prediction from physiological data.

2.7.3 Discussion on classification techniques

As we described in previous sections, there are various classification algorithms which are commonly used in physiological-based emotion recognition systems. The task of choosing the most suitable classifier based on existing studies is difficult. Various results that have been obtained in different studies seemed to be confusing. For example, Nasoz et al. (2004) and Nasoz et al. (2010) showed that ANNs' performance is better than kNN while in a different study, van den Broek et al. (2010) claimed opposite results. Or in another research study, Zhai and Barreto (2006) found that SVMs performed much better than the naïve Bayes classifier although Müller (2006) claimed the two classifiers obtained quite the same accuracy rate. The mixed results obtained in the previously cited studies may be due mainly to the use of different physiological data sets which can be different because of a lot of reasons including emotion stimuli, subject physical exertion, environment temperature range and inter-subject variations in physiology (Novak et al., 2012).

In addition, comparing classification accuracies using the same physiological data set can be valid only if the comparison used the same experimental setups which include for example applying the same cross-validation ⁶technique for estimating the classification accuracy rate and the same feature selection methods for feature reduction.

⁶ Please refer to section 3.8.2.1 for more details.

It was also found that in most of the studies, application of the feature selection method with a single classifier shows that this combination has a positive influence on the performance of emotion recognition system compared to using a classifier without a feature selection technique. For selecting a suitable classifier, another potential solution is employing ensemble classification strategies which are still underutilized in developing emotion recognition systems despite having encouraging results in other fields.

In addition, for high dimensional data, like physiological data, the use of multiple feature selection methods in an ensemble classification system can enhance classification accuracy performance of ensemble system (Oliveira et al., 2006; Santana & Canuto, 2014). By using several feature selection methods, the diversity of the ensemble method will increase through the generation of different but relevant feature subsets. The diversity created in this case is not as the same as in the case of random subspace where the feature sets are randomly selected. It is rather the result of the selection of different but relevant feature sets. Compared to other feature selection method, filter-based feature selection methods are known to be simple and fast especially in high dimensional data problems where the computational cost is very important. (Santana & Canuto, 2014).

However, the feature-based ensemble classification strategy has not been thoroughly tested in physiological-based emotion recognition systems.

2.8 Concluding Marks

After discussing and reviewing several topics related to physiological-based emotion recognition systems in this chapter, we would like to highlight some important concluding remarks in the following points:

- 1- The main data modalities which have been used in physiological-based emotion recognition systems are: (1) peripheral physiological data mainly

includes ECG, EMG, SC, ST and RSP and (2) EEG physiological data (brain data).

- 2- Different feature dimension reduction techniques have been used in physiological-based emotion recognition systems. Commonly used methods are Wrapper methods which include SFS, SBS, SFFS. These methods, however, are time-consuming and overly specific to the classifier used. Another category includes PCA and Fisher's projection technique. The main limitations of these techniques are: they do not guarantee to offer superior correlation with emotional states of the subjects than primary features and (2) the new features do not have the physical meaning which results in a lack of the system's interpretation. Feature ranking techniques are fast but the use of only one feature ranking method, as it is commonly used, may result in a sub-optimal solution. One solution can be using more than one feature ranking method to increase the chance to choose the optimal feature set (Kuncheva, 2007).
- 3- Conventional comparisons of classification accuracies between existing studies is difficult because physiological data may be influenced by a lot of issues including emotion stimuli, subject physical exertion, etc. In addition, comparing classification accuracies using the same physiological data set can be valid only if the comparison used the same experimental setups which include for example applying the same cross-validation technique for estimating the classification accuracy rate and the same feature selection methods for feature reduction.

- 4- According to the review provided in previous sections, it is advisable to the researchers to apply a variety of classifiers along with feature reduction techniques to decide the one is the best suited solution for their data and condition (Novak et al., 2012).
- 5- The literature review showed that LDA, CART, ANN, SVM are widely used to develop physiological-based emotion recognition systems.
- 6- Ensemble classification strategies have been underutilized in physiological-based emotion recognition systems despite having encouraging results in other fields. The main advantage of these techniques is their ability to achieve better results than benchmark single classifiers. The main idea is that the errors made by each of the classifiers are not identical and combining several classifier outputs in an efficient manner may correct some of these errors (Leo Breiman, 1996).
- 7- Utilizing feature selection methods for ensembles has demonstrated to be a helpful strategy for ensemble methods development because of its ability to provide more robust and diverse feature subsets that make the classifiers of the ensemble disagree on challenging instances in which eventually increases classification accuracy of ensemble method (Oliveira et al., 2006; Santana & Canuto, 2014).
- 8- In physiological-based emotion recognition systems where a high dimensional data set is usually involved, filter-based feature selection methods are preferred because of being simple and fast (Santana & Canuto, 2014).

CHAPTER 3: THE PROPOSED CLASSIFICATION METHOD

This chapter describes the proposed classification method used to recognize human emotional states based on physiological signals. The chapter includes three main sections. The first section summarizes the findings of the literature review related to the limitations of existing methods that used to design the physiological-based emotion recognition systems. Then, section 3.2 explains the main steps involved in the design & development of the proposed physiological-based emotion recognition systems. These steps are described in section 3.3 to 3.6 and include data set selection and preparation, designing benchmark emotion recognition system using single classifiers, designing feature-based multi-classifier methods, which involve the use of multiple classifiers created by feature selection methods, and the proposed feature-based dual-layer ensemble classification methods, which involve the use of multiple classifiers created by feature selection methods embedded in dual-layer classification structure. The last section (3.7) of this chapter explains the evaluation activity adopted for this research which details in section 3.8 the experimental setups applied during the experiments and finally by describing experiment 1, 2 and 3 in sections 3.9, 3.10 and 3.11, respectively.

3.1 The main finding of literature review

As we discussed in conclusion section of chapter 2, the existing emotional state recognition systems based on physiological data have some limitations which affect their classification accuracy. The main limitations can be generally attributed to the feature selection method used for feature reduction and the classification algorithm applied in the recognition process. In the feature selection step, several studies have applied wrapper methods which are time-consuming especially for high dimensional data set and overly

specific to the classifier used. In addition, PCA and Fisher's projection methods, which are widely used to reduce the feature dimension, are lacking the interpretability as the new features produced by these methods do not have any physical meaning which prevents performing some analysis like identifying the most important features related to specific emotion. Feature ranking methods are fast but the common practice is to select one feature ranking method to rank the features according to their relevance to the emotion states. In fact, using only one method may result in sub-optimal solution because two different feature ranking methods are likely to produce two different ranking sets and presenting only one set given by a particular method can be misleading (Kuncheva, 2007). For classification methods, the common approach is to use one of the single classifiers but as it can be seen from the literature review, ensemble methods which are known for their classification ability have not been fully investigated in physiological-based emotion recognition systems probably because most of the ensemble methods, except for standard ones like Bagging and Adaboost, are not available in the software packages used by researchers.

To address the above-mentioned issues in the existing physiological-based emotion recognition systems, a feature-based dual-layer ensemble classification method was proposed. This method is designed based on stacking ensemble strategy where the first layer, which is also called base-layer, is where the classifiers are created from features generated by different feature ranking methods. These methods are known for being fast and do not suffer from some limitations such as classifier-dependency like in the case of wrapper methods or the lack for interpretability like in the case of feature projection methods (Saeys, Abeel, & Van de Peer, 2008; Santana & Canuto, 2014). In the second layer, the outputs of the first layer become the training data of the second layer, which is

also known as the meta-layer, are used to build the final classification model. Using several feature ranking methods together will prevent generating a sub-optimal subset of features. Furthermore, using multiple feature ranking methods will provide more accurate and diverse feature subsets in the ensemble classification method. This is important because the performance of the ensemble systems depend on accuracy and diversity achieved by the individual components of ensemble that will rise disagreement among these components on challenging instances which ultimately will increase the classification accuracy of the ensemble method (Guan, Yuan, Lee, Najeebullah, & Rasel, 2014; Santana & Canuto, 2014). As to the accuracy of the proposed method, the assumption is that proposed method will increase the classification accuracy rates because the features of the second layer, which are the outputs of the first layer (0 or 1), are relatively similar which makes the task of approximating the relation between the features and their respective classes easier for the classifier comparing to the features of the first layer. Our proposed classification method is described in more details in section 3.6.

3.2 Steps involved in the design and development of a physiological-based emotion recognition system

Design & Development of the physiological-based emotion recognition system is the third activity of research methodology presented in section 1.8. This activity, as summarized in Figure 3.1, involves four main steps and they are described in sections 3.3 to 3.6, namely, data base selection and preparation for producing data sets containing physiological data features, then designing the emotion recognition system using benchmark classifiers that produce a complete benchmark system of emotion recognition. This is followed with the designing of feature-based multi-classifier methods which produce an emotion recognition system augmented with feature selection ability. The last

step is designing of the proposed classification algorithm (feature-based dual-layer ensemble classifier) which produces an emotion recognition system augmented with the proposed feature-based dual-layer ensemble classification method.

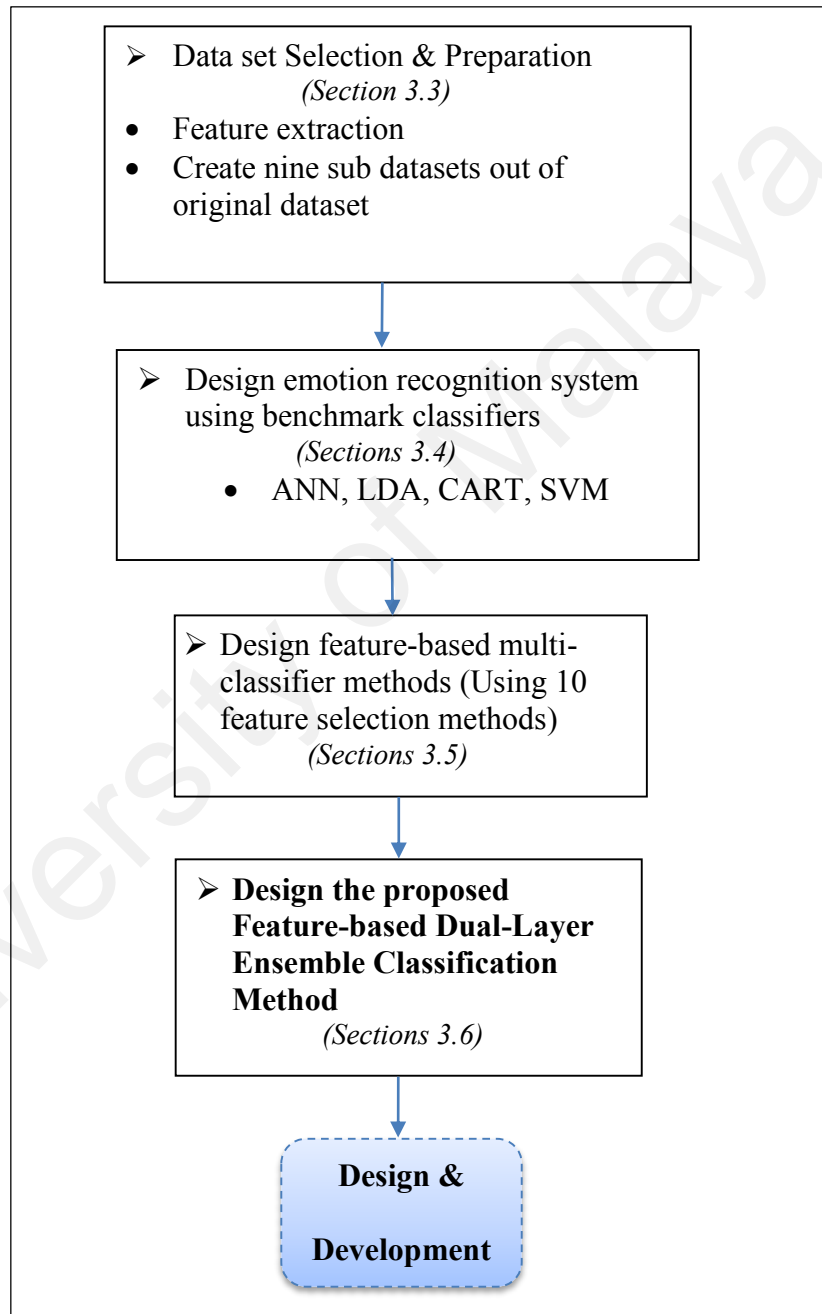


Figure 3.1: The steps followed to design and develop the proposed feature-based dual-layer ensemble classification method

3.3 Data set selection and preparation

Accumulating reliable physiological data signals as a data set is among the early and vital steps for the development of emotion recognition systems (Novak et al., 2011). In this research, the preference is on the usage of a comprehensive and standard data set that has been recorded in a proper manner and it was also used by other researchers to test their proposed emotion recognition systems. The detail information of this data set is explained in the next section.

3.3.1 General description

The DEAP physiological data set (Koelstra et al., 2012) which is a multimodal publicly available data set for analysis of human emotional state was used in this research. This data set was selected because of the following reasons: (1) it's publicly available, (2) it comprises of peripheral physiological data as well as brain physiological data, (3) it has enough number of participants for emotion recording and (4) highly cited by other researchers. This dataset consisted of electroencephalogram (EEG) and peripheral physiological signals which include electro-cardiogram (ECG), electro-myogram (EMG), electrooculogram (EOG), blood volume pulse (BVP), respiration amplitude (RSP), skin temperature and galvanic skin response (GSR) of 32 subjects while watching 40, one minutes long music videos. A total of 1280 recorded signal was collected and each subject rated each video in terms of the levels of arousal, valence, like/dislike, dominance using a 9-point Likert scale. Self-assessment manikins (SAM) (Bradley & Lang, 1994) were utilized to visualize the scales. For liking scale, thumbs up/down were used (see Figure 3.2).

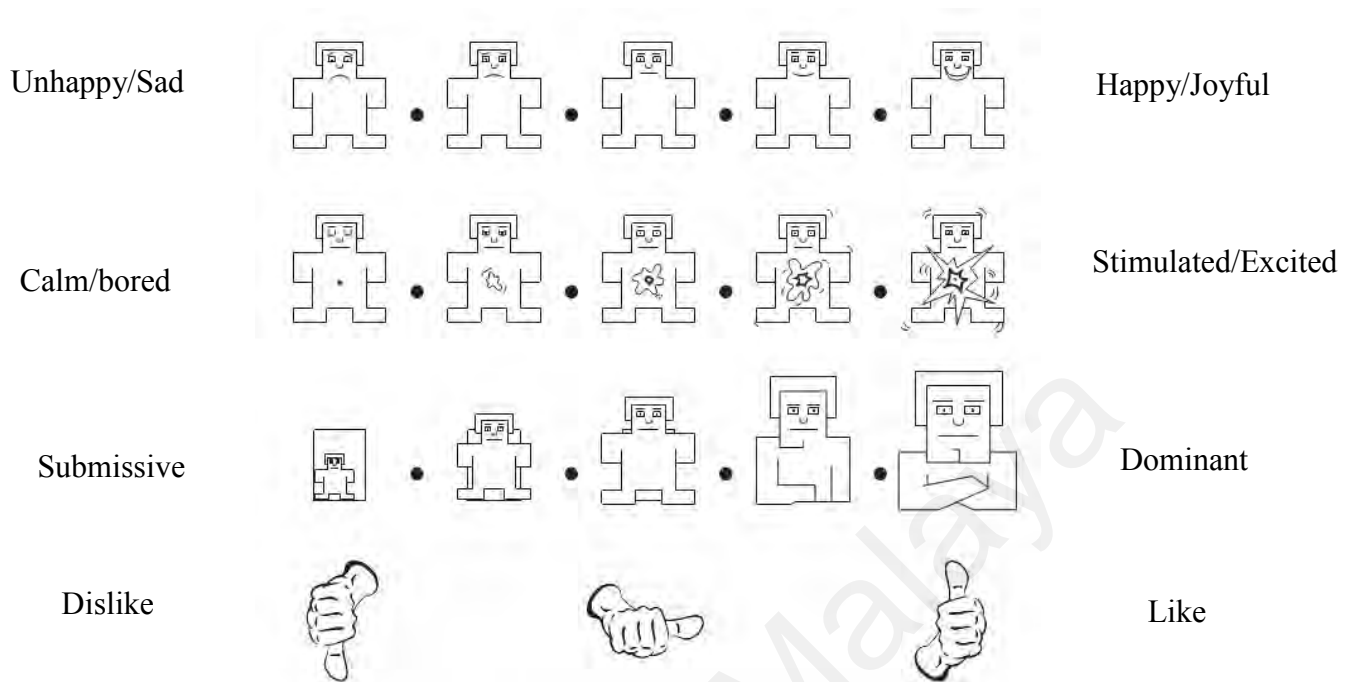


Figure 3.2: The SAM used to rank the emotion dimension of valence (top), arousal (second) and dominance (third) of subjects. Thumbs up/down (last) to scale liking (Koelstra et al., 2012)

The stimuli for making this dataset were selected to induce emotions in the four quadrants of the valence-arousal space. The subject emotions are classified based on Low-Arousal Low-Valence (LALV), Low-Arousal-High-Valence (LAHV), High Arousal-Low-Valence (HALV), and High-Arousal-High-Valence (HAHV). The rating value five and more is considered as high and less than five is low. A total of nine datasets has been created using Peripheral, EEG, and the combination of Peripheral and EEG data and used for binary classification tasks for arousal, valence, and liking based on “High” and “Low” ranking. Figure 3.3 shows the distribution of DEAP recorded signals for subject 1 in the Arousal-Valence space. Regarding scatter plot of Arousal-Valence space, the 40 selected music videos were quite successful to stimulate different emotions for subject1.

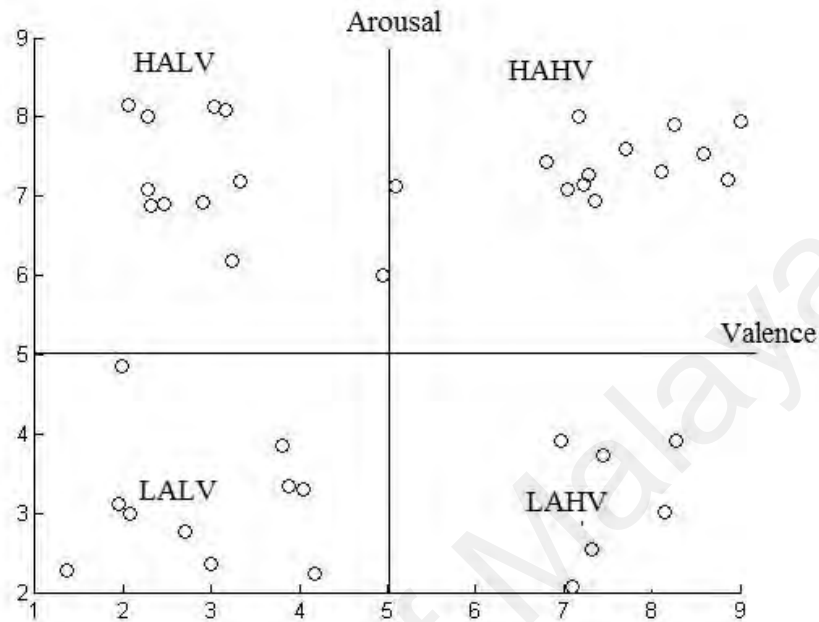


Figure 3.3: Distribution of emotion ranking for 40 videos by subject1 based on two-dimensional (valence–arousal) emotion model

In this research, we utilized the “Pre-processed version” of the data provided in MATLAB format for data analysis and feature extraction. The data set belongs to each subject was stored in a 3d array of size 40 (videos)×40(32 EEG channels and 8 peripheral physiological sensors) × 8046(63second long, sample rate 128Hz). In addition, we apply some pre-processing steps such as lowpass, highpass filtering and normalization on each signal and finally, the features such as mean, median, SD, maxima, minima and etc. can be calculated for each physiological signal.

3.3.2 Features extracted from physiological Signals

In a physiological-based emotion recognition system, signal processing techniques are commonly utilized to extract relevant features from physiological data signals. In this

study, we used two commonly types of features, namely, peripheral and EEG features. These two types of features were proposed by the creator of DEAP dataset (Koelstra et al., 2012) and the code used to extract them from the provided data signals was written in MATLAB. Some samples of code which used for feature extraction are provided in Appendix A (Figure A. 1 and Figure A. 2). In addition, the extracted features were normalized by converting the values of features to range between 0 to 1, by subtracting each feature's value from the mean value of all feature vectors and dividing the result by the standard deviation of all feature vectors. This process will help to decrease inter-subject variability by adjusting every subject's features values to a difference between their max and minimal rates.

3.3.2.1 Peripheral features

Peripheral features include the following set of features.

(a) *Electrocardiogram (ECG)*

The ECG signal commonly can be recorded from the left and right wrists or from other areas like the chest. The ECG signal measures the activity of heart contractions or beats. The measurement of heart rate (HR) or the speed at which the heart is beating is usually expressed in beats-per-minute (bpm) which is normally between 60-100 beats per minute.

ECG features extracted in this study are: Average and standard deviation of Heart Rate(HR), Energy Ratio between frequency bands [0.04-0.15] Hz, spectral power in the bands[0.1-0.2] Hz,[0.2-0.3]Hz,[0.3-0.4]Hz (Koelstra et al.,2012).

(b) *Electromyogram (EMG)*

The EMG signal is used to measure the frequency of muscle tension and contraction. The most common muscles that are involved in physiological emotion recognition

research belong to the human head and the facial part which are the masseter (the muscle above the jaw), the Trapezius (neck), the corrugator (the muscle above the eyebrow) and the zygomatic (cheek). As an example, the activity of Zygomaticus is monitored since this muscle is triggered when the person smiles or laughs. Most of the power in the spectrum of an EMG during muscle contraction is in the frequency range between 4 and 40Hz (koelstra et al., 2012).

EMG features extracted in this are: Average and variance of the signal, Median, Interquartile range, the energy of the signal (koelstra et al., 2012).

(c) *Electrooculography (EOG)*

This signal is used to measure rate of eye blinking which is related to anxiety in emotion recognition system. Eye blinking easily can be identified by peaks in the signal. The EOG features used in this study are: Average and variance of the signal, Median, Interquartile range, the energy of the signal (koelstra et al., 2012).

(d) *Skin Conductivity (SC) or Galvanic Skin Response (GSR)*

Skin Conductance is also known as Galvanic Skin Response (GSR). SC is a measure of the skin resistance (conductance) of a small electrical current. It is measured by placing two electrodes normally at the tips of index and middle fingers of the hand. SC also can be measured from palm, forearm and the soles of the feet. As soon as a person feels stress and nervous tension, the surface of the skin including the palms becomes moist. This cause increase in skin conductivity (decrease in skin resistance), and the skin can then be seen as a variable resistor.

The SC features used in this study are: Average of signal, Median of signal, Average of derivative, Average of derivative for negative values, Average rising time of the signal,

Proportion of negative derivatives values, Number of local minima, zero crossing rate of Skin conductance slow response (SCSR in [0,2.4]Hz), zero crossing rate of Skin conductance very slow response (SCVSR in [0,0.2]Hz), 10 spectral power in band [0-2.4]Hz, Mean of peaks magnitude for SCSR and SCVSR (koelstra et al., 2012).

(e) Respiration (RSP)

Respiration measures how fast and deep a person is breathing as he or she breathe in and breathe out air in their lungs. The rate and depth of this activity can be measured with a chest band sensor around the chest. The respiratory rate can be measured based on a number of breaths per minute (Lorig et al., 2007).

The RSP features used in this study are: Average and standard deviation of signal, Average of derivative, Band energy ratio, Total power of the signal, Average distance between local minima, range of greatest breath, breathing rate, breathing rhythm, 10 spectral power values in [0-2.4] Hz bands, Average and Median peak to peak time (koelstra et al., 2012).

(f) Skin temperature (ST)

It is a valuable physiological signal which is easy to measure using the skin temperature sensor (SKT). The temperature change can reveal differences in mood and emotions. The body temperature is measured by fixing the sensor on the fingers to detect the temperature signal and its change. The sensor can also be used to detect the excitement level of a person (Khalili & Moradi, 2008).

The ST features used in this study are: Average, Average of derivative, Median, Interquartile range, spectral power values in the bands [0-0.1] Hz and [0.1-0.2]Hz (koelstra et al., 2012).

3.3.2.2 Electroencephalogram (EEG) features

The EEG signal is measured by placing electrodes on the scalp (head surface) according to the 10-20 international system. It measures the electrical activity of the neurons of the brain in the form of oscillatory activity. The EEG has a high temporal resolution in milliseconds since it can measure brain electrical activity directly from the scalp (Andreassi, 2007; Lorig et al., 2007).

EEG features extracted in this study are: Theta, slow alpha, alpha, beta and gamma Spectral power for each electrode, the spectral power asymmetry between 14 pairs of electrodes in the four bands of alpha, beta, theta and gamma (koelstra et al., 2012).

Table 3.1 depicts a list of 216 features extracted from EEG signal and 78 extracted features for peripheral physiological signals including GSR, ECG, ST, EMG and EOG.

Table 3.1: Features extracted from EEG and peripheral physiological signals

Signal	#Channels	Extracted Features	#Extracted features
GSR	1	Average of signal, Median of signal, Average of derivative, Average of derivative for negative values, Average rising time of the signal, Proportion of negative derivatives values, Number of local minima, zero crossing rate of Skin conductance slow response (SCSR in [0,2.4] Hz), zero crossing rate of Skin conductance very slow response (SCVSR in [0,0.2] Hz, 10 spectral power in band [0-2.4]Hz, Mean of peaks magnitude for SCSR and SCVSR	21
ECG	1	Average and standard deviation of Heart Rate(HR), Energy Ratio between frequency bands [0.04-0.15] Hz, spectral power in the bands[0.1-0.2]Hz,[0.2-0.3]Hz,[0.3-0.4]Hz	6

Signal	#Channels	Extracted Features	#Extracted features
RSP	1	Average and standard deviation of signal, Average of derivative, Band energy ratio, Total power of the signal, Average distance between local minima, range of greatest breath, breathing rate, breathing rhythm, 10 spectral power values in [0-2.4] Hz bands, Average and Median peak to peak time	21
ST	1	Average, Average of derivative, Median, Interquartile range, spectral power values in the bands [0-0.1]Hz and [0.1-0.2]Hz	6
EMG and EOG	4	Average and variance of signal, Median, Interquartile range, energy of the signal	24 (6×4)
EEG	32	Theta, slow alpha, alpha, beta and gamma Spectral power for each electrode, The spectral power asymmetry between 14 pairs of electrodes in the four bands of alpha, beta, theta, and gamma.	216 (32×5+14×4)

3.4 Design emotion recognition system using benchmark classifiers

After physiological data features extracted as explained in section 3.3.2, then, we design our benchmark emotion recognition system. In designing this system four different single type classifiers are used. Thus, this section briefly presents the single classifiers that were utilized in this research as the benchmark classifiers. These methods will also be utilized for designing the feature-based multi-classifier methods as well as the proposed feature-based dual-layer ensemble classification methods. More details regarding each single classifier can be found in chapter 2.

3.4.1 LDA

Linear Discriminant Analysis (LDA) is a well-known classification method originally developed in 1936 by R. A. Fisher (1936) and has been used effectively in a wide variety of problems.

The main objective of LDA is to separate data samples to distinct groups which are called classes. LDA transforms the data to a different space, normally with lower dimension, which maximizes the between-class separability while minimising their within-class variability. (McLachlan, 2004).

LDA is commonly used in machine learning problems like pattern recognition, face recognition, feature extraction and data dimensionality reduction. It is a simple and mathematically robust method where usually generates models whose accuracy is similar to complicated methods (Miguel & Guerreiro, 2008).

3.4.2 CART

Classification and Regression Trees (CART) developed by Breiman, Freidman, Olshen, Stone (1984) is a classification method which uses past data to build decision tree

classification model and then use it to classify new data sample. CART algorithm will search for all possible features and all possible values in the data set to discover the best split question which later is named as splitting rule that splits the data into two parts with maximum homogeneity. This process is then repeated for each of the resulting data segments (Timofeev, 2004).

CART is a powerful and frequently-used classification algorithm, which can deal with incomplete data, multiple types of features (floats, unnumerated sets) for both input features, and predicted features, and the trees it produces often contain rules which are easily readable (Lewis et al., 2000).

3.4.3 ANN

Artificial neural networks (Kohonen, 1982) or neural networks are usually considered as a simulation of the information-processing in the nervous system. Early work in this field was inspired by studying systems of neurons and learning rules derived from biological models (Depenau, 1995).

There are different kinds of neural networks, but the commonly used neural network for classification is a feed-forward network with a simple perceptron and its extension the multi-layer perceptron. ANNs have been extensively used to model classification and regression problems in different fields. The input layer represents the features while the output layer is usually used for the classes. The hidden layer is used to approximate the input-output relation. In our study, we used Multilayer Perceptron, a frequently -used ANN model (Depenau, 1995). A classic multi-layer feed-forward network is shown in Figure 3.4.

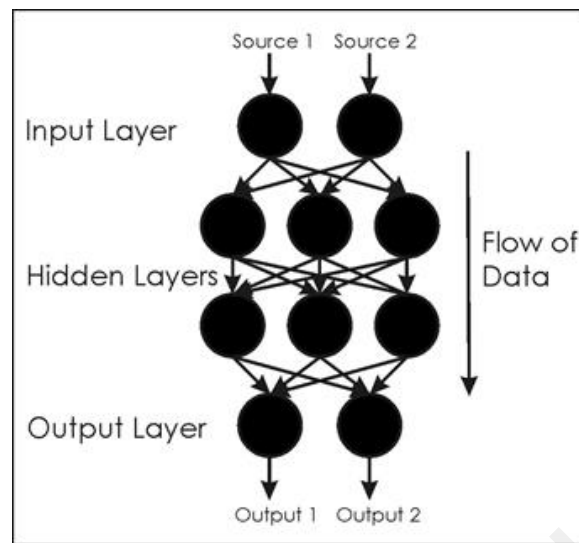


Figure 3.4: A graphical representation of a typical feedforward neural network

3.4.4 SVM

A support vector machine (SVM) is a powerful machine learning algorithm originated from statistical learning theory and first introduced by Vapnik (1998). It has been used successfully in a wide variety of problems. SVM map the input space into a very high dimensional feature space and then tries to find a linear separation between classes in the transformed version of the feature space (Webb, 2002). SVMs detect those samples of each class that determines the boundary of classes in the feature space. These samples are considered to be the most informative samples and are called support vectors.

SVM can employ a small training set for creating generalizable nonlinear classifiers which are main advantages of this classifier in high-dimensional feature space. In the case of having large training sets, SVM chooses a small set of support vectors that are required for designing the classifier. It can significantly decrease the computational cost of testing (A. K. Jain, Murty, & Flynn, 1999). Therefore, this method is popular because of its high level of generalizability and its capability to handle high dimensional input data relative to neural networks and decision trees (Theodoridis & Koutroumbas, 2006).

Because of above mentioned advantages of SVM, it is one of most popular classification technique in affective computing. SVM classifiers propose competitive performance results for emotion recognition compared to other classification techniques.

3.5 Design the Feature-based Multi-Classifier methods

It is useful to investigate the effect of using different feature selection techniques on the classification accuracy of the emotion recognition system, in which their results will be also used for comparison purpose in Chapter 4. The feature-based multi-classifier methods are a set of classifiers created by training sets whose features are selected using 10 different feature selection methods and whose decisions are taking by a majority vote method. Figure 3.5 displays the way in which this method is working. First, one of 10 feature selection methods, which is *feature selection1* for example, is applied to the original training data to rank the features from 1 to n where n is the number of features in the original training data set. The number of features n is divided by 9 to produce 9 feature subsets of equal number. This number is selected based on the total number of features. It could be another number but it should be moderate and around 10, which is good enough to prevent excessive computational cost, however still sufficiently good to produce all the possible essential features subsets.

For example, the peripheral physiological data set has 78 input features; which means we can get 9 parts with 8 features for each. The first training data is created by taking the top 8×1 or 8 features only while the second training data set is obtained by choosing the top 8×2 or 16 features and so on until the last, or the 9th training data set that includes the top 8×9 or 72 features. Each of the nine training sets is used to train one classifier algorithm, for example, we used ANN to create ANN1 using training data1, ANN2 using training data2 and so on until ANN9 using training data 9. The next step that follows

creating 9 ANN classifiers is to use them to produce their testing classification outputs which are used to get the final output using majority vote method⁷. To create another feature-based multi-classifier using the same ANN classifier, another feature selection method, for example, feature selection method2, is used. In this way, we can create 10 feature-based ANN classifiers using 10 different feature selection methods. The same thing is followed for the other 3 classifiers, namely, LDA, CART, and SVM.

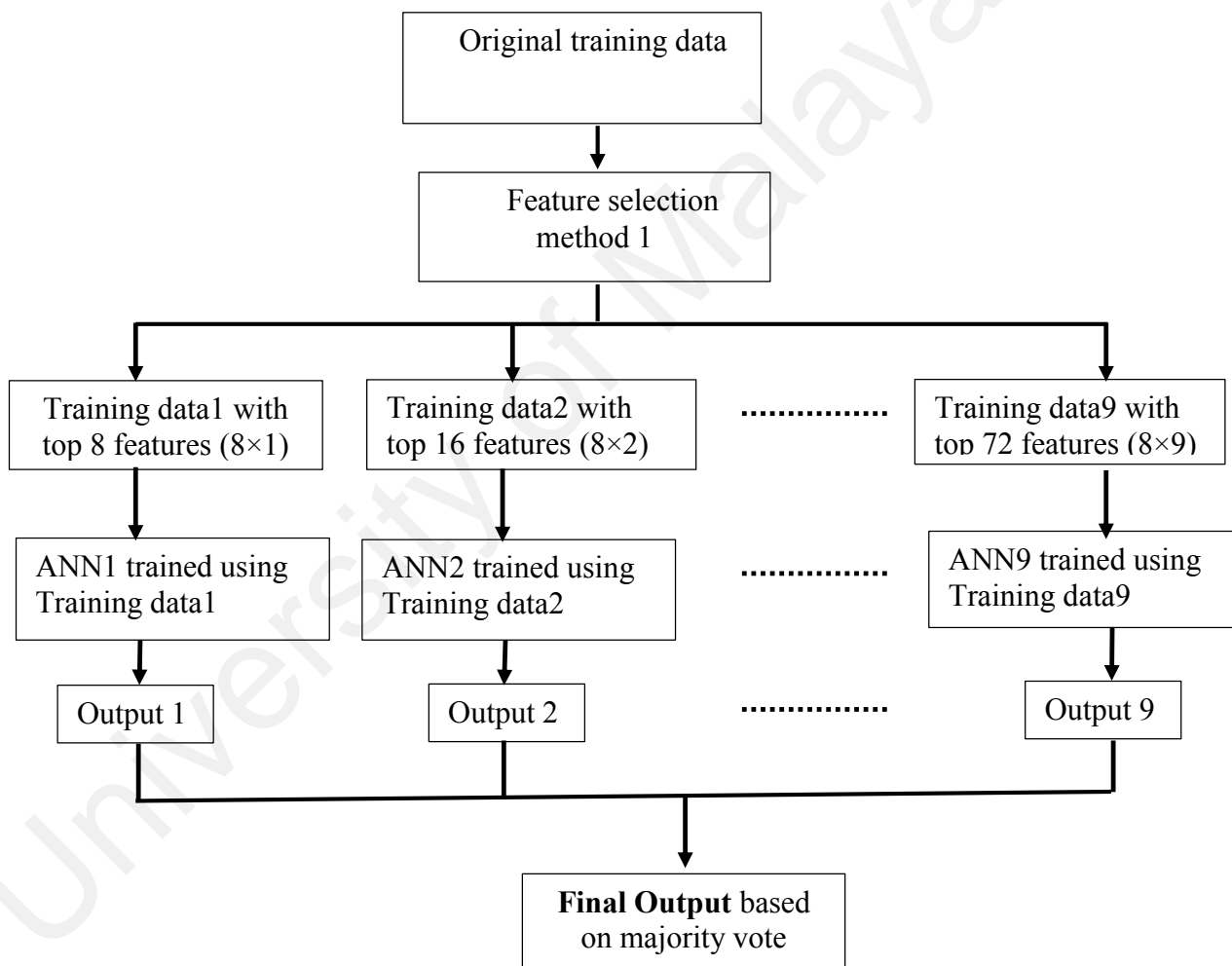


Figure 3.5: The method for creating feature-based multi-classifier method

⁷ majority vote method is explained in section 3.5.2

3.5.1 Feature selection methods

The 10 selected feature selection methods, which fall under the feature ranking category of feature selection methods (Section 2.7.1.6), are among the most commonly used methods in the literature. In, (Brown et al., 2012), the authors conducted a comprehensive study about their performances and characteristics. The objective of using different feature selection methods to create different training sets is to increase the diversity among the classifiers, which is a key feature in improving the performance of multi-classifiers systems. In addition, two different feature selection methods may give two different feature sets. Thus, presenting only one feature set can be misleading and may produce suboptimal results (Kuncheva, 2007). The 10 feature selection methods used in this study are briefly explained in the following subsection.

(a) *T-test*

The t-test is a filter-based feature ranking approach which is traditionally used to compare two normally distributed samples or populations. T-test (Student, 1908) method defines the score of an attribute as the ratio of the difference between its mean values for each of the two classes and the standard deviation, the latter considers the standard deviation values of the feature for every class and the cardinality of each. Finally, the weight of each feature is thus given by its computed absolute score (I.-H. Lee, Lushington, & Visvanathan, 2011).

(b) *Fisher score*

Fisher score is one of the most widely used supervised feature selection methods (i.e. the training data are labeled and determines feature relevance by evaluating feature's correlation with the class). The output of this algorithm is a list of ranked features based

on their computed feature weight/score under the Fisher criterion. Specified class labels $y = \{y_1, \dots, y_n\}$, Fisher Score (Duda, Hart, & Stork, 2001) selects features that assign similar values to the samples from the same class and assign different values to samples from different classes. The evaluation criterion used in Fisher Score can be formulated as below:

$$SC_F(f_i) = \frac{\sum_{j=1}^c n_j (\mu_{i,j} - \mu_i)^2}{\sum_{j=1}^c n_j \sigma_{i,j}^2} \quad (3.1)$$

where μ_i is the mean of the feature f_i , n_j is the number of samples in the j th class, and $\mu_{i,j}$ and $\sigma_{i,j}$ are the mean and the variance of f_i on class j , respectively.

Fisher Score is very effective feature selection algorithm, which has been widely utilized in many real applications (Zhao et al., 2010).

(c) **Relief**

It was introduced by Kira & Rendell (1992). This method is feature grading algorithm. The objective of this method is a quality estimation of features to differentiate samples that are near to each other in a dataset.

Original Relief only can handle boolean concept problems, but extensions have been developed to work in classification problems and in regression.

(d) **Mutual Information Maximisation (MIM)**

This method was proposed in (Lewis, 1992). It gives a score for each feature independently of others and finally, based on their mutual information, ranks the features. This method is commonly used in the literature to rank the features. Top features are usually selected for analysis or used as an input for the next algorithm (G. Brown et al.,

2012).

(e) *Conditional Mutual Info Maximisation (CMIM)*

This method searches for the most discriminative features by finding the optimal trade-off between relevancy and redundancy in the features (Fleuret, 2004). In this case, the feature is selected if it only maximizes the mutual information of the features while adds additional information to the already selected feature set.

(f) *Joint Mutual Information (JMI)*

This method was proposed by (Yang & Moody, 1999) to reduce the redundancy by increase the *complimentary* information between features (G. Brown et al., 2012).

(g) *Double Input Symmetrical Relevance (DISR)*

To reduce the redundancy, (Meyer & Bontempi, 2006) used *symmetric relevance* criterion which promotes the concept of complementary information between the features. This criterion measures the symmetrical relevance on all combination of two features (Meyer & Bontempi, 2006).

(h) *Interaction Capping (ICAP)*

This method which was proposed by (Jakulin, 2005) use interaction gain measure to detect the relevant features. In this method, any feature even if it is not relevant to the class by its own, it can be relevant when combined with another feature.

(i) *Conditional redundancy (Condred)*

It was proposed in (G. Brown et al., 2012) for a comparison purpose.

(j) ***Conditional Informative Feature Extraction (CIFE)***

This method which was proposed by (D. Lin & Tang, 2006) aims to maximize the class-relevant information by reducing the class-relevant redundancies among features (D. Lin & Tang, 2006).

3.5.2 Majority vote method

When classification outputs are produced by a set of classifiers, we can get the final output using *Majority vote method* by taking the output which receives the highest number of votes from these classifiers. In other words, the final output is the most frequently predicted class by the set of classifiers.

For examples, we have five classifiers used to predict classes 1 and 2 and which produced the following outputs: 1, 2, 1, 1, 2. By applying majority vote method, we can see that class 1 received three votes against two for class 2; which means 1 is final or the winner output (or class).

3.6 Design the proposed Feature-based Dual-Layer Ensemble Classification Method

This section explains the design of the proposed classification method to improve the classification accuracy performance of an emotion recognition system. As we mentioned in section 3.1, in the proposed feature-based dual-layer ensemble classification method (FDLEC), the concept of stacking ensemble (details in Section 2.7.2.8(a)) is adopted. In the first layer, the 10 feature selection methods (i.e. feature ranking methods) (detail in section 3.5.1) are used to generate several training data sets with different features subset sizes. Using different feature selection methods will help us to generate more diverse and higher quality subsets of features that lead to the creation of more accurate and diverse

set of classifiers in the first layer of our method which makes the classifiers of the ensemble disagree on challenging instances (Guan et al., 2014; Oliveira et al., 2006; Santana & Canuto, 2014). In addition, combining various feature ranking methods instead of using one method will prevent generating a sub-optimal subset of features (Kuncheva, 2007). Our selection of the 10 feature ranking methods is guided mainly by two criteria: the first one is to choose the frequently cited feature ranking methods in the literature and which proved their performance in other field of studies (Brown et al., 2012) and the second criterion is to select 10 methods or close to 10 which is modest enough to avoid any high computational cost but probably good enough to generate all the possible relevant features that can be missed if one or few methods are used. In addition, producing different feature sets improves the diversity of the ensemble method while ensures the quality of the selected feature sets.

The feature ranking methods that have been employed in our proposed methods fall under the filter-based category. The advantage of this type of methods compared to wrapper-based methods is being fast and independent from any particular classifier which makes them more suitable to be utilized in designing a feature-based ensemble classification method (Saeys et al., 2008; Santana & Canuto, 2014). In addition, the single classifiers chose to be used in our proposed method- LDA, ANN, SVM and CART- are among the well-known single classifiers that have been utilized widely in the area of physiological-based emotion recognition systems.

In the second layer (meta-layer), the prediction output of each classifier in the first layer is combined together to create a meta-dataset. The outputs of the first layer are composed of zeros and ones which indicate the predicted class for each instance. The second layer uses these outputs as the input features. The target feature of the second layer

remains the same as in the original training set. The second layer of our proposed classifier combines different predictions into a final one. By using a second layer classification, this method tries to induce which classifiers are reliable and which are not. For example, if a classifier steadily misclassified instances from one region because of incorrectly learning the feature space of that region, the Meta classifier may be able to discover this problem. Therefore, utilizing the learned behaviors of other classifiers may enhance this kind of training problems (Zhu, 2010).

The proposed classification method composes -as depicted in Figure 3.6 - of double layers:

- **The first layer**

The first layer generates 90 classification outputs predicted by one type of classification method like SVM for each testing pattern. In fact, this 90-outputs vector is obtained by converting 10×9 matrix of outputs generated by the same classifier trained on 90 different training data sets where these data sets are obtained by taking 9 different feature subset sizes (represent the columns) from the original feature set using 10 different feature selection methods (represent the rows). We can say that we applied the same method as in feature-based multi-classifier method (details in section 3.5) except for one thing, in this method we have 90 outputs instead of 9 outputs because all the feature selection methods are combined. The first layer of this method is described in form of pseudocode in Figure 3.7

- **The second layer**

The second layer of training can be summarized as follows: since we apply leave-one-out cross-validation (details in Section 3.8.2.1) method to divide the training and testing sets and as we have 40 examples (explained in Section 3.3.1) for each subject, we will

get new data set of 40 examples with 90 features for each example. This data is again used to train another classifier using the same method that is leave-one-out cross-validation. Different combinations of classifiers are used in the first and second layer. Since we have 4 classifiers (Sections 3.4.1 to 3.4.4), this results in 16 different combinations. The best combination is chosen as the recommended method to be used for the emotion state recognition system. The pseudocode related to the second layer of this method is described in Figure 3.8.

University of Malaya

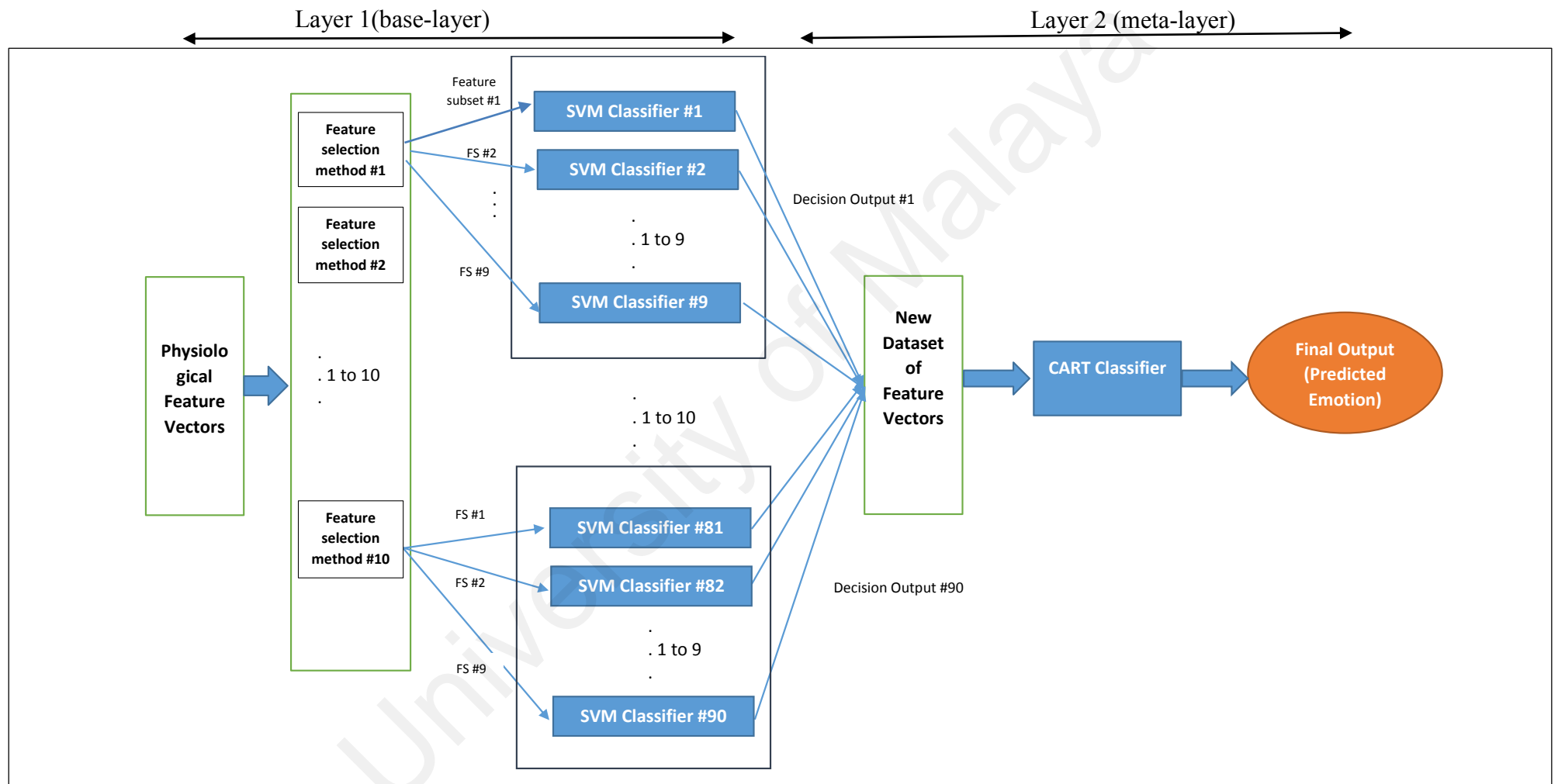


Figure 3.6: Concept of the proposed feature-based dual-layer ensemble classification method

```

% AllFeatures are the features of the original training data set

% OriginalTrainData is the original training data

% TestingData is the testing data set

For v=1 to 40 % 40 is the number of data sets per subject (or videos watched by each subject (leave-one out
method))

For SelectMethod =1 to 10 %

FeatureRanked= GetRanking(SelectMethod); % rank the n features of the original training data

% from top to bottom using SelectMethod which is

% one of 10 feature selection methods

For k =1 to 9 %

Subset= FeatureRanked (1:8×k) ; % Get the top 8×k features for each iteration

NuTrainingData= OriginalTrainData(Subset) ; % get the new training set using the selected Subset

SVMModel = TrainSVM(NuTrainingData); % Train SVM using the new training data set

TestOutput=Predict(SVMModel, TestingData);

AllOutput(SelectMethod,k)= TestOutput ; % get 10×9 outputs matrix of the first layer created by 9

% different subsets ranked using 10 different feature selection methods

End For k

End For SelectMethod

Outputvector=convert(AllOutput); % convert 10×9 matrix into 90 outputs vector

NewDataset(v, 10×9)= Outputvector;

End for v

```

Figure 3.7: The pseudocode related to the first layer of the proposed dual-layer classification method

```

%Second Layer%

MetaDataSet=NewDataset

For v=1 to 40 % 40 is the number of data sets per subject (or videos watched by each subject (leave-one out method))

    CARTModel = TrainCART ( TrainingData(MetaDataSet)); % Train CART using the new training meta-data set

    TestOutput=Predict (CARTModel,TestingDataSet(MetaDataset));

End for v

```

Figure 3.8: The pseudocode related to the second layer of the proposed dual-layer classification method

As an example, Figure 3.9 shows a sequence of activities that has been followed by an input data to generate final output using the proposed dual-layer ensemble classification method, as an example, in case of having peripheral data modality as input, there will be 40 feature vectors with total of 78 features for each subject (matrix of 40×78). Using the proposed method, the first layer will generate a new set of data with an equal number of rows and 90 features that its values will be 0 or 1 which are results of binary classification of 90 classifiers in the first layer (matrix of 40×90). The second layer will use the new set of data and train a single classifier. The output of the classification model will be a matrix of 40×1 . It shows the final classification output (i.e. 0 or 1) assigned to each feature vector. Each feature vector is related to physiological data of a person while he/she watching a particular movie (details in Section 3.3.1). Receiving 0 as output, means the proposed system could recognize the emotion of the person watching the video as a low level of arousal, valence or liking (more details in Section 2.3). Receiving 1 means the proposed system recognizes the high level of arousal, valence or liking recognized by the system.

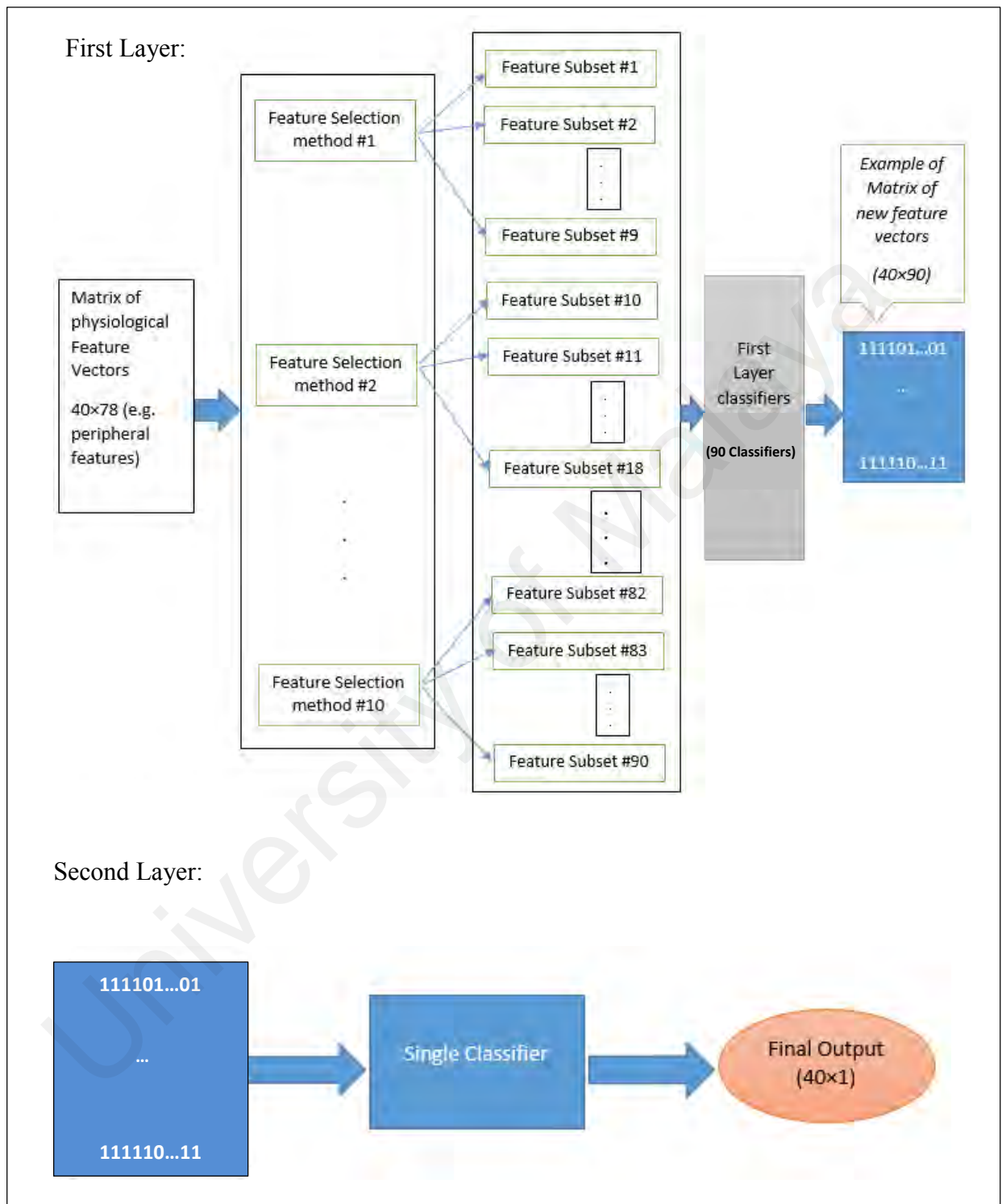


Figure 3.9: A sequence of activities for proposed dual-layer ensemble classification methods.

3.7 Evaluation

Evaluation is the fourth activity of the research methodology presented in section 1.8. This activity, as summarized in Figure 3.10, involves three main steps and they are described in sections 3.9 to 3.11, namely, comparative analysis of classification accuracy rates (CARs) of benchmark classifiers that describes experimental works related to the evaluation of each benchmark classifier on different modalities, followed by comparative analysis of classification CARs of feature-based multi-classifier methods that describes experimental works involved the evaluation of feature-based multi-classifier methods on different modalities as well as their comparison with benchmark classifiers. The last step is the comparative analysis of all the methods including the proposed feature-based dual-layer ensemble classification method using a statistical test. Before the details of each experiment are provided, the experimental setups (section 3.8) used in this study mainly related to the configuration of parameter specifications of classifiers and feature selection methods as well as the method for classification performance evaluation, are explained.

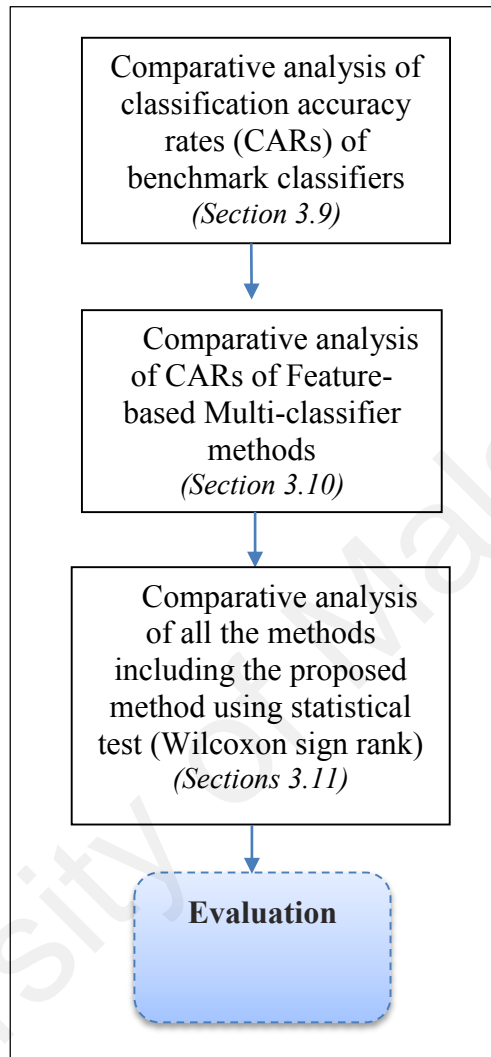


Figure 3.10: The steps followed to evaluate the proposed feature-based dual-layer ensemble classification method

3.8 Experimental setups

This section mainly describes the programming environment, software packages and its related configurations that have been used for developing the proposed methods.

3.8.1 Parameter specifications of the classification methods and feature selection methods

All the computer programs were written using MATLAB software version 14. This software is a well-known programming environment widely used by developers and researchers. This software is organized in specialized toolboxes that cover many domains ranging from engineering to finance. In our study, we used various toolboxes including: Statistics toolbox which used to build LDA, CART classification models and Neural network toolbox which employed to build ANN models. Another toolbox called libsvm and developed by Chih-Chung Chang and Chih-Jen Lin was used to create SVM classification model. This toolbox can be downloaded from the link below (<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>). In addition, MATLAB code for Random Forest, which is one the benchmark classifier used in this study for comparison purpose, was downloaded from <https://code.google.com/p/randomforest-matlab/>.

For feature selection methods, 7 out of 10 feature selection methods were obtained from FEAST toolbox which can be downloaded from <https://github.com/Craigacp/FEAST/>. In addition, the three other methods namely, Fisher, t-test, and Relief were developed by Arizona state university and are available for downloading from the following link (<http://featureselection.asu.edu/software.php>).

In this study, we used the default parameter values provided by MATLAB and other toolboxes developed by the third party. For ANN model, we used the following equation to calculate the number of neurons in the hidden layer:

The number of neurons in hidden layer = (the number of features +the number of classes)/2.

In addition, the data modalities used for the experiments are: single data modalities where we used peripheral and EEG features separately and multimodality which is a combination of both peripheral and EEG features.

3.8.2 Classification Performance Evaluation

This section provides the information related to the method used in this research for calculation of classification accuracy rates. It also explains the methods followed to compare different classifiers' performances on different data sets.

3.8.2.1 Testing classification accuracy rate calculation

In this study, following the creators of DEAP data sets Koelstra et al. (2012), we use leave-one-out cross-validation method to divide the data into training and testing data sets. In this method, only one example is used for testing while the remaining are used for training the classification model. We assume that our system is a dependent system which means that new users cannot use the system without prior training. In this case, the classification accuracy is calculated per subject and the final or average testing accuracy rate is averaged over the accuracies of all the subjects. To explain more, for each subject, we have 40 examples or videos, 39 of them are used for training while only one is used for testing. The process is repeated 40 times until all the 40 examples are used as a testing data set in of the 40 iterations. In this case, we have 40 testing accuracies for one subject and average testing for this subject is averaged over the 40 iterations. We used the same procedure for all the 32 subjects and the average accuracy is obtained by averaging the accuracies of the 32 subjects.

Classification accuracy rate is calculated as follows:

Classification accuracy rate (%) = $100 \times (\text{the number of correctly testing examples} / \text{the total number of testing examples})$

Since the total number of testing examples is 1, classification accuracy rate for each iteration is either 100% or 0%. Assuming that for one of the subjects, 30 out of 40 iterations, the classification rate is 100% while it is 0% in the 10 remaining iterations. In such case, the average classification rate for this subject is calculated as follows:

$$\text{Average classification accuracy rate (\%)} = (30 \times 100 + 10 \times 0) / 40 = 75\%.$$

Assuming also that 50% of the subjects or 16 of them got 75% and other half got 50%. Now, the average classification rate = $(16 \times 75 + 16 \times 50) / 32 = 62.5\%$

In this case, 62.5% represents the average classification accuracy for all the subjects and it is the final testing accuracy rate.

3.8.2.2 Average Rank

Since we have various classifiers applied to different data sets, each classifier produce different classification accuracy results on the different data set. In this case, to measure the overall performance of a classifier using different data sets, the ranking method proposed by Friedman's M statistic (Neave & Worthington, 1992) is used (Brazdil & Soares, 2000). In this method, each classifier receives a rank based on the measured accuracy rates on each data set where the classifier with highest accuracy rate on a data set is assigned rank 1 and the classifier with second highest accuracy rate assigned rank 2 and so on. If two classifiers achieved the equal accuracy rates, then the rank is divided between them. For example, if we have accuracies of 50%, 60%, 62%, 62%, 67% for five different classifiers on a data set, their ranking score would be 5, 4, 2.5, and 2.5, 1 respectively. To calculate the final ranking of a classifier on different data sets, the

different rank scores of that classifier on different data sets are averaged. Therefore, lowest average ranking score means the best classifier. For example, assuming that A, B, and C are three different classifiers that were tested on two data sets X and Y. Table 3.2 shows the measured accuracy rates of the three classifiers on the two data sets and associated ranks. In this example, classifier C is recognized as an overall best classifier with the lowest average rank of 1.5.

Table 3.2: Ranking and average rank calculated for three different classifiers on two data sets based on their classification accuracies.

Classifier/Data set	X	Rank	Y	Rank	Average Rank
A	56%	3	68%	1	$4/2=2$
B	58%	2	57%	3	$5/2=2.5$
C	68%	1	60%	2	$3/2=1.5$

3.8.2.3 Statistical test

To check whether there exists any significant difference between our proposed classification method and other methods we applied Wilcoxon ranks test (Wilcoxon, 1945). It is one of the safe and robust non-parametric tests for statistical comparisons of classifiers which works for comparison of two classifiers on multiple datasets. It ranks the differences in performances of two classifiers for each data set. The Wilcoxon ranks test will try to reject the null-hypothesis that both algorithms perform equally well (Demšar, 2006).

Assume that there are two classifiers that should be tested on N datasets, d_i is calculated as the difference between the accuracy performance rates of the two classifiers on i -th out of N data sets. The calculated differences are ranked according to their absolute values; In case of ties, the average ranks are assigned. Let R^+ be the sum of ranks for the

datasets on which the second algorithm outperformed the first while R^- is the sum of ranks where the first algorithm performs better than second one (opposite). Ranks of $d_i = 0$ are split equally among the sums; if there is an odd number of them, one is ignored:

$$R^+ = \sum_{d_i > 0} \text{rank}(d_i) + \frac{1}{2} \sum_{d_i = 0} \text{rank}(d_i) \quad (3.2)$$

$$R^- = \sum_{d_i < 0} \text{rank}(d_i) + \frac{1}{2} \sum_{d_i = 0} \text{rank}(d_i) \quad (3.3)$$

Later $T = \min(R^+, R^-)$ will be the minimum value among R^+ and R^- . Then the exact critical values for T for N (i.e. number of data sets) up to 25 can be found in general statistics books (i.e. Table of exact critical values for the Wilcoxon's test). As an example, with a confidence level of $\alpha = 0.05$ and $N = 14$ datasets, the difference between the classifiers is significant if T is equal or less than 21 (i.e. critical value). Therefore, the null-hypothesis is rejected. For a larger number of data sets, the statistics:

$$z = \frac{T - \frac{1}{4}N(N+1)}{\sqrt{\frac{1}{24}N(N+1)(2N+1)}} \quad (3.4)$$

is distributed approximately normally. For a confidence level of $\alpha = 0.05$, the null hypothesis can be rejected if z is smaller than -1.96 (Demšar, 2006).

3.9 Experiment 1: Comparative analysis of CARs⁸ of Benchmark classifiers

The objective of the first set of experiments is mainly to obtain the benchmark testing classification accuracy results (or simply classification accuracy rates) (details in Section 3.8.2.1) using four benchmark classifiers, namely, LDA, ANN, CART and SVM on nine different data sets (details in Section 3.4). Three data sets containing physiological data modalities, the other three containing EEG data modalities and the rest of data sets containing multi-modality data, which is a combination of both peripheral and EEG data modalities. All data sets are used for binary classification tasks for arousal, valence, and liking based on “High” and “Low” ranking (details in Section 3.3). The classifiers’ parameter specifications are explained in Section 3.8.1.

In addition, we are interested in investigating how the use of different data modalities- single modalities and multi-modality data- may affect the classification accuracy of the emotion recognition system. To compare the overall performance of different benchmark classifiers, the average rank score (details in Section 3.8.2.2) of each classifier over different modalities is calculated.

3.10 Experiment 2: Comparative analysis of CARs of Feature-based Multi-classifier methods

The objective of conducting the second set of experiments using feature-based Multi-classifier (details in Section 3.5) is to investigate the effect of using different feature selection techniques on the classification accuracy rate of the emotion recognition system. The testing classification accuracy results (details in Section 3.8.2.1) are used to evaluate the performance of Feature-based Multi-classifier methods on nine different data sets. All

⁸ CARs=Classification Accuracy Rates

the benchmark classifiers used in experiment 1 also used in this experiment. The classifiers' settings are as the same as experiment 1, explained in Section 3.8.1. More details about the 10 feature selection methods used can be found in sections 3.5.1 and 3.8.1.

The data modalities used for this set of experiments are as the same as experiment 1 which includes single modalities and multimodality data and all data sets used for binary classification tasks for arousal, valence, and liking based on "High" and "Low" ranking (details in Section 3.3). To compare the overall performance measurement of different feature-based multi-classifier, the average rank score of each classifier over different modalities (data modalities) is calculated (details in Section 3.8.2.2). In addition, to compare between results obtained by benchmark classifiers (experiment 1) and feature-based multi-classifier (experiment 2), the testing classification accuracy rates of these two methods are calculated based on valence, arousal, and liking and its averaged for each modality.

3.11 Experiment 3: Comparative analysis of CARs of Feature-based Dual-layer ensemble classifiers

The goal of the third set of experiments is mainly to evaluate accuracy performance of the proposed feature-based dual-layer ensemble classification method (Section 3.6). All different combination of LDA, ANN, CART, and SVM (details in Section 3.4) as first and second layer classification methods are implemented in such a way that all classification methods used for the first layer will be the same type. For example, one of proposal designs includes 90 SVM classifiers at the first layer. The classifiers' settings for the first and second layers are the same as experiment 1 and all the feature selection methods (details in Section 3.5.1&3.8.1) which were employed in experiment 2 are used in this experiment. In addition, Random Forest is considered as one of the benchmark

ensemble classifier used in this study for comparison purpose. Random forest (Leo Breiman, 2001) is a well-known ensemble learning method for classification. It works by creating several decision tree classifiers in the training phase and the final output is decided using majority vote method (details in Section 3.5.2). This method is included for comparison purpose as it is considered as one of the most efficient ensemble methods in the literature (Amaratunga, Cabrera, & Shkedy, 2014; Robnik-Šikonja, 2004).

The testing classification accuracy results (details in Section 3.8.2.1) are used to evaluate the performance of our proposed feature-based dual-layer ensemble classifiers on nine data sets. The data modalities used for this set of experiments are the same as experiment 1 and 2 which include single modalities and multi-modality data. All data sets used for binary classification tasks for arousal, valence, and liking based on “High” and “Low” ranking (details in Section 3.3). To compare overall performance measurement of different feature-based dual-layer ensemble classifiers, the average rank score of each classifier over different modalities is calculated. In addition, to compare between the results of benchmark classifiers (experiment 1), feature-based multi-classifier (experiment 2), and feature-based dual-layer ensemble classifiers (experiment 3) the classification accuracy rates of these three methods are calculated based on valence, arousal and liking and averaged for each modality.

Furthermore, in order to check whether the classification accuracy rate of the proposed feature-based dual-layer ensemble classifier is significantly better than the best classifiers of experiment 1, experiment 2, and experiment 3, the statistical Wilcoxon ranks test is used (more details in Section 3.8.2.3).

3.12 Summary

In this chapter, we described the steps conducted for third and fourth activities of the proposed research methodology (details in Section 1.8) which are Design & Development as well as Evaluation. In this chapter, we mainly described the main characteristics of the data sets, benchmark classification methods, feature-based multi-classifier methods as well as the proposed feature-based dual-layer ensemble classification method, used in this study. In addition, we provided the experimental settings used in developing our proposed system and conducting our experimental work. The leave-one-out cross-validation method used to divide the data set into training and testing data sets was also described. In addition, the Wilcoxon ranks test method used for the purpose of comparing the proposed classification method with other methods also was explained.

CHAPTER 4: RESULTS AND DISCUSSIONS

This chapter presents the emotion classification performance of the proposed method and the comparisons with benchmark methods. The chapter is divided into four parts. The first part reports the results of the four benchmark classifiers (details in Section 3.4) on the nine data sets related to DEAP. These classifiers, which have been frequently used for classification problems including in the field of emotion recognition, are: ANN, CART, LDA, and SVM. The second part of this chapter presents the effect of using feature selection methods on the performance of the four classifiers by testing of feature-based multi-classifier methods (details in Section 3.5). The objective of the first two parts (one and two) is to obtain benchmark results that can be compared with results of the proposed feature-based dual-layer ensemble classification method (details in Section 3.6) presented in the third part. In the fourth and final part, a statistical comparison using Wilcoxon rank test is conducted between the proposed method and the benchmark methods to evaluate the performance of our proposed method and identify the existence of any significant improvement made by our method compared to the benchmark methods.

- *Experiment 1:*

4.1 The results of benchmark classification methods

For overall performance measurement of different benchmark classifiers, the average rank score of each classifier over different modalities (See Table 4.1) is calculated. As stated in section 3.8.2.2 about the average rank method, the classifier with the lowest average ranking score is the best classifier. Therefore, we notice from Table 4.1 and Figure 4.1 that SVM achieved the best results followed narrowly by LDA and then ANN. CART received the lowest accuracy among the single classifiers. In addition, Table 4.2

to Table 4.4 and its corresponding figures (Figure 4.2 to Figure 4.4) depict the average accuracy rates of four benchmark classification methods for valence, arousal and liking targets, using different modalities, in which SVM performed well on liking and arousal, while it is the worst classifier that can be used for recognizing the valence for all modalities. Though SVM is a stable and powerful classifier, it could not be consistent in all the emotional states recognition. This may reveal one drawback of using single classifier for emotion recognition state which is the consistency. The alternatives to SVM can be LDA and ANN, as they achieved comparable results in most cases. CART is generally the worst among the four classifiers.

Table 4.1: Average ranking score of four benchmark classification methods for valence, arousal and liking targets

Modality Datasets	Target	SVM	ANN	LDA	CART
Peripheral data	Peri-Valence	4	2	1	3
	Peri-Arousal	1	4	3	2
	Peri-Liking	1	2	3	4
EEG Data	EEG-Valence	4	2	1	3
	EEG-Arousal	1	2	3	4
	EEG-Liking	1	2	4	3
peripheral & EEG data	(EEG+Peri)-Valence	4	2	1	3
	(EEG+Peri)-Arousal	1	4	2	3
	(EEG+Peri)-Liking	1	3	2	4
Average Ranking Score		2	2.555556	2.22222222	3.222222

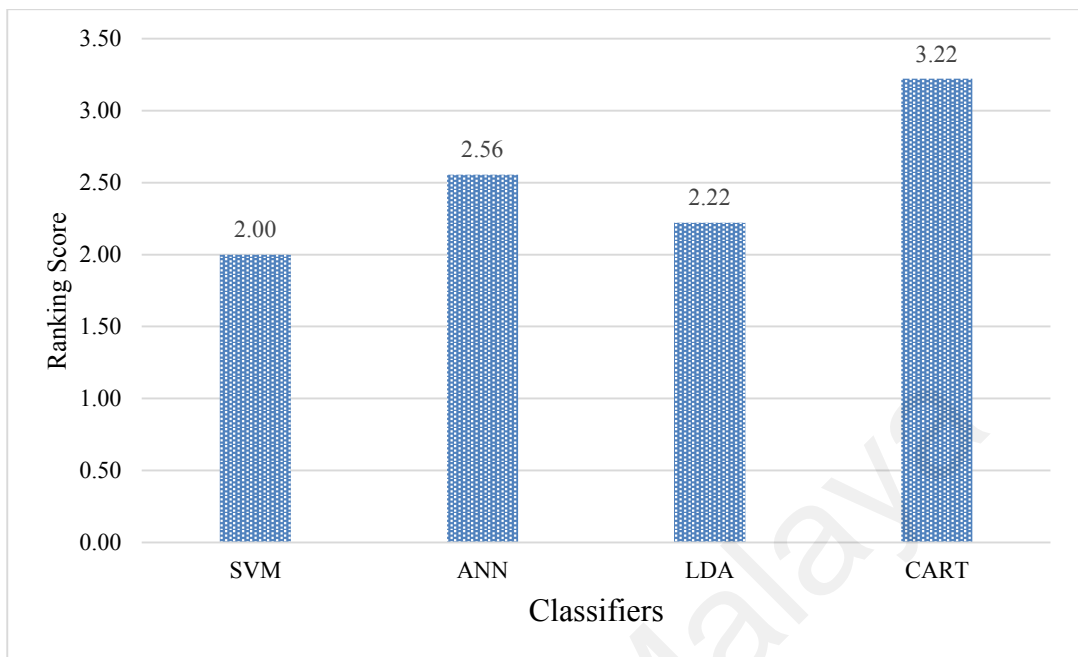


Figure 4.1: Average ranking score of four benchmark classification methods

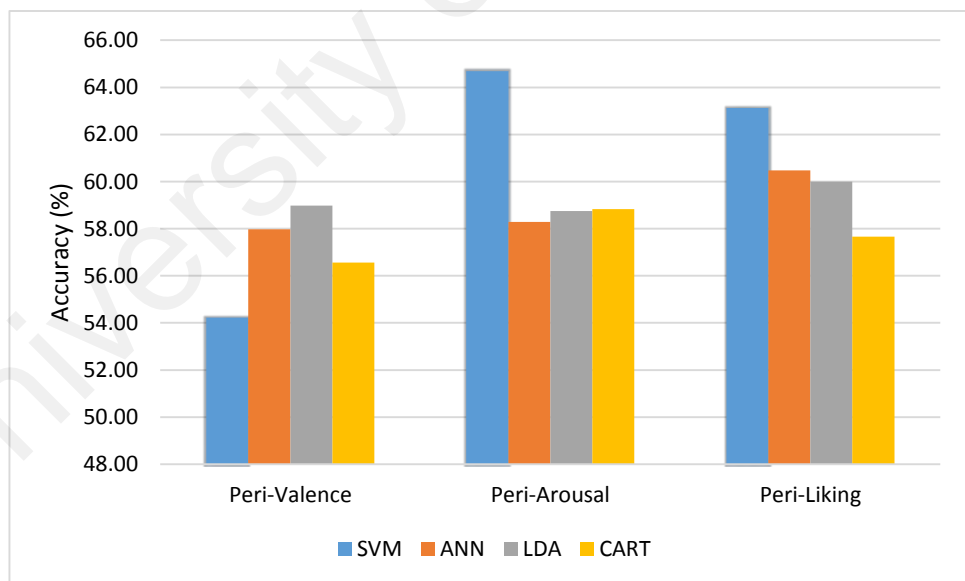


Figure 4.2: Average accuracy rates of four benchmark classification methods for valence, arousal and liking targets using peripheral modality

Table 4.2: Average accuracy rates of four benchmark classification methods for valence, arousal and liking targets using peripheral modality

Method	Peri-Valence	Peri-Arousal	Peri-Liking
SVM	54.2188	64.6875	63.125
ANN	57.96875	58.28125	60.46875
LDA	58.98438	58.75	60
CART	56.5625	58.82813	57.65625

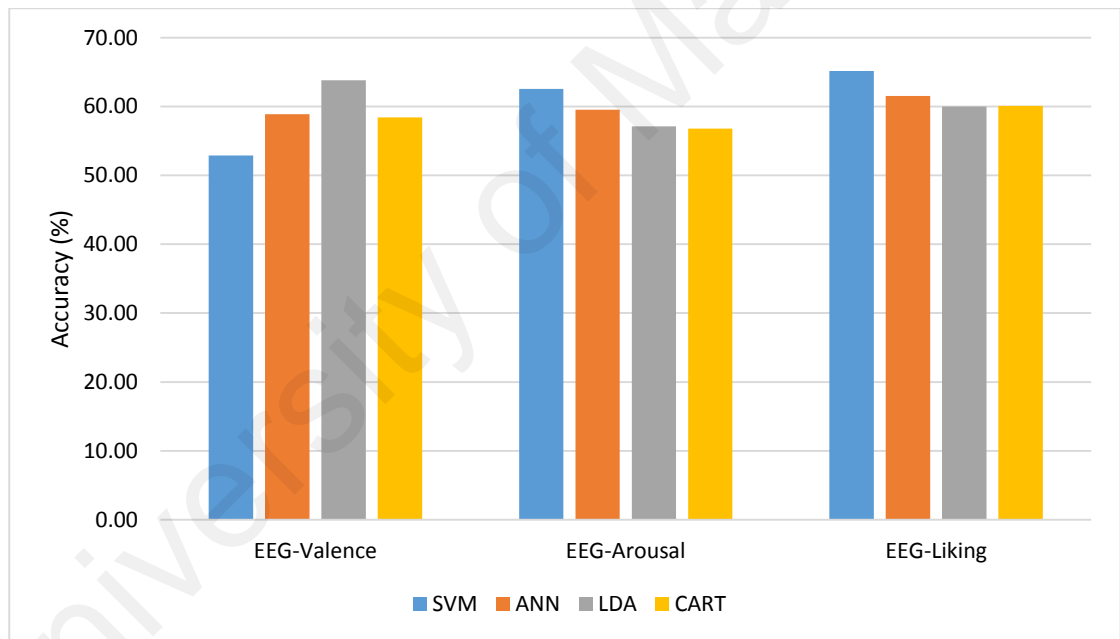


Figure 4.3: Average accuracy rates of four benchmark classification methods for valence, arousal and liking targets using EEG modality

Table 4.3: Average accuracy rates of four benchmark classification methods for valence, arousal and liking targets using EEG modality

Method	EEG-Valence	EEG-Arousal	EEG-Liking
SVM	52.8906	62.5781	65.1563
ANN	58.9063	59.5313	61.5625
LDA	63.8281	57.1094	60
CART	58.4375	56.7969	60.0781

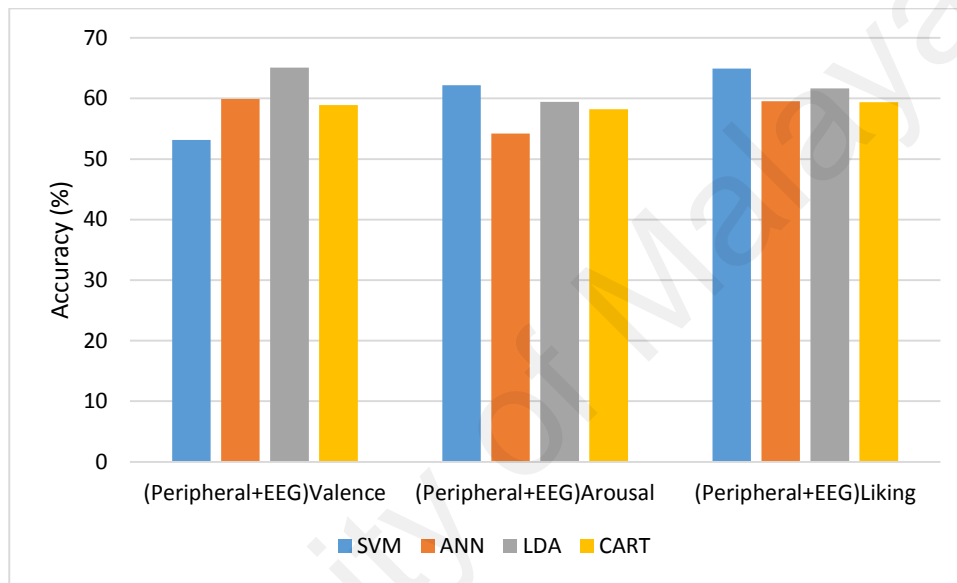


Figure 4.4: Average accuracy rates of four benchmark classification methods for valence, arousal and liking targets using (Peripheral+EEG) modalities

Table 4.4: Average accuracy rates of four benchmark classification methods for valence, arousal and liking targets using (Peripheral+EEG) modalities

Method	(Peri+EEG) Valence	(Peri+EEG) Arousal	(Peri+EEG) Liking
SVM	53.125	62.1875	64.92188
ANN	59.9219	54.2188	59.53125
LDA	65.0781	59.4531	61.64063
CART	58.9063	58.2031	59.375

4.1.1 Classification accuracy rates of the four benchmark classifiers based on the modality used

In order to study the effect of each modality on the performance of different benchmark classifiers, the classification accuracy rate of each classifier using different modalities for arousal, valence and liking recognition is ranked (See Table 4.6 to Table 4.9). Though Figure 4.5 shows that in overall using (Peripheral+EEG) data modality achieved the best results followed by EEG modality, the results in Table 4.5, which summarizes Table 4.6 to Table 4.9, and its corresponding Figure 4.6 indicate that there is no single type of modality that is suitable for all the classifiers. Specifically, SVM achieved the best results when it uses Peripheral modality while EEG is more suitable for ANN. The combination of the two modalities (Peripheral + EEG) enables LDA and CART to achieve their best accuracies. For a classification method, which has an internal mechanism for feature selection like CART, combining the two modalities can be better because it enables to select the relevant features from different modalities which result in more information about to the emotional state.

Table 4.5: Average ranking score of each modality using four benchmark classifiers

Modality/Classification technique	SVM	ANN	LDA	CART
Peripheral Modality	1.6667	2.3333	2.333333	2.333333
EEG Modality	2	1.3333	2.333333	2
(Peripheral+EEG) Modalities	2.3333	2.3333	1	1.666667

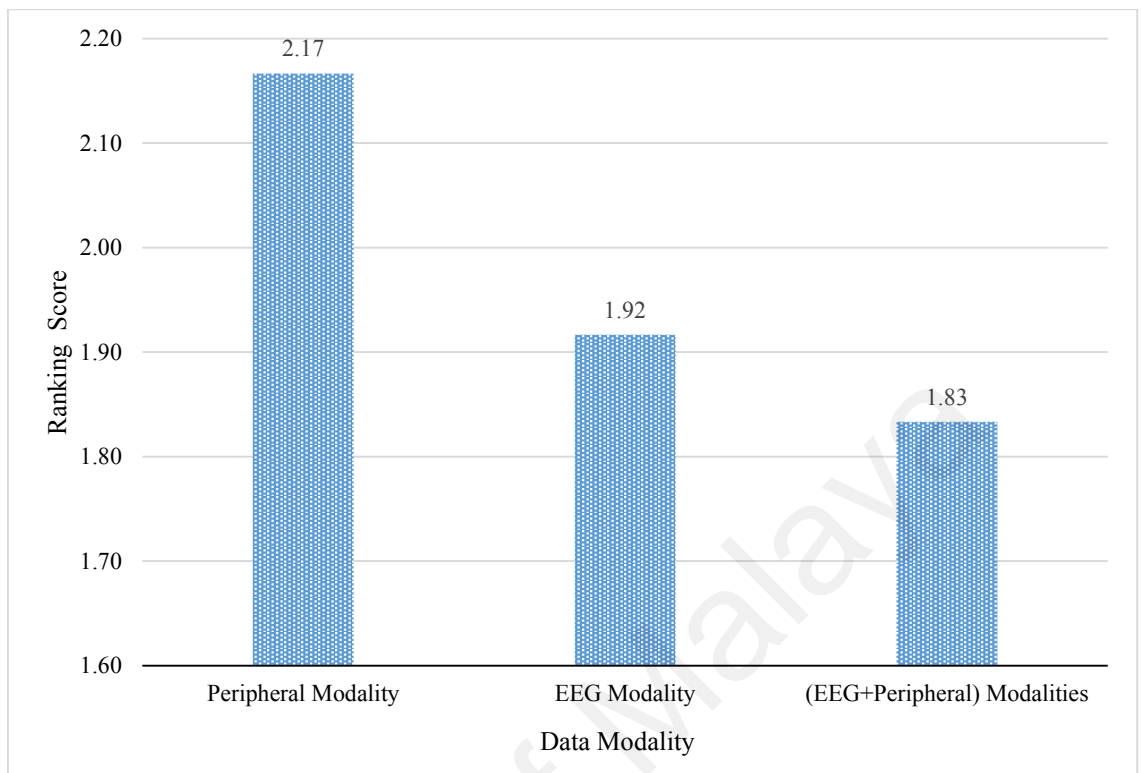


Figure 4.5: Average ranking score of each modality using four benchmark classifiers

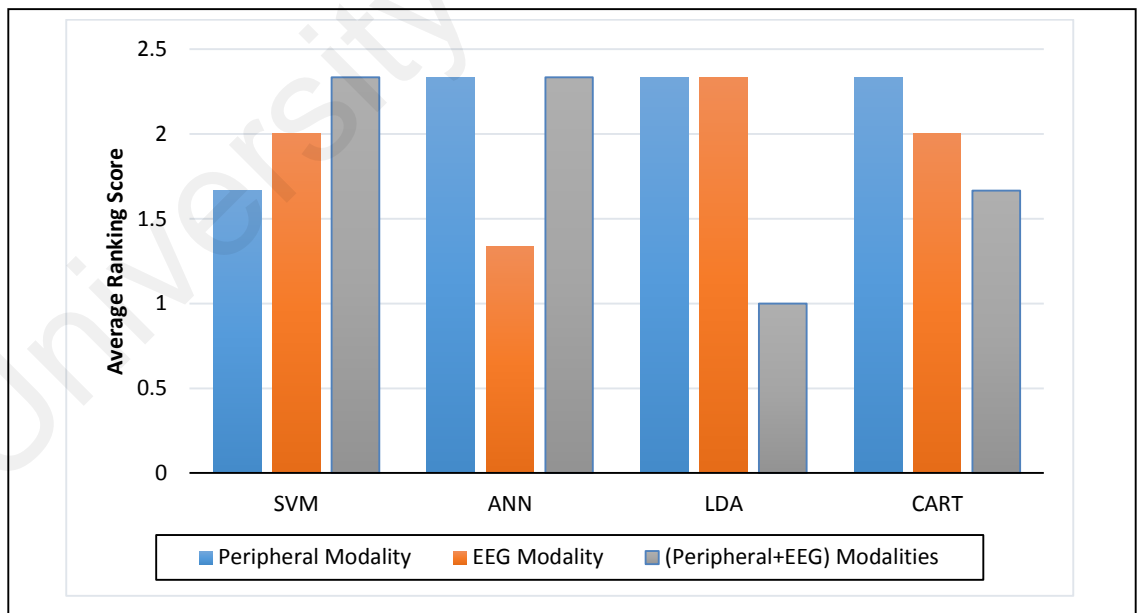


Figure 4.6: Average ranking score of each modality using four benchmark classifiers

Table 4.6: Average ranking score of each modality using SVM

Classification Method	SVM			Total Average Score
Modality Dataset	V ⁹	A	L	
Peripheral Modality	1	1	3	1.666666667
EEG Modality	3	2	1	2
(Peripheral+EEG) Modalities	2	3	2	2.333333333

Table 4.7: Average ranking score of each modality using ANN

Classification Method	ANN			Total Average Score
Modality Dataset	V	A	L	
Peripheral Modality	3	2	2	2.333333333
EEG Modality	2	1	1	1.333333333
(Peripheral+EEG) Modalities	1	3	3	2.333333333

Table 4.8: Average ranking score of each modality using LDA

Classification Method	LDA			Total Average Score
Modality Dataset	V	A	L	
Peripheral Modality	3	2	2	2.333333333
EEG Modality	2	3	2	2.333333333
(Peripheral+EEG) Modalities	1	1	1	1

Table 4.9: Average ranking score of each modality using CART

Classification Method	CART			Total Average Score
Modality Dataset	V	A	L	
Peripheral Modality	3	1	3	2.333333333
EEG Modality	2	3	1	2
(Peripheral+EEG) Modalities	1	2	2	1.666666667

⁹ V: Valence, A: Arousal, L: Liking

4.1.2 Results obtained by the classifiers based on given modality

“Suppose we have four classifiers and only one modality, and the question is which classifier is better to use with that modality?”

The average ranking score of the four benchmark classifiers for valence, arousal and liking recognition using three different modalities are calculated as shown in Table 4.10. According to this results, SVM seems the best classifier to use for Peripheral modality. For EEG, the best classifier is shared by SVM and ANN while LDA is the suitable classifier to use when EEG and Peripheral modalities are combined.

Table 4.10: Average ranking score of the four classifiers for each modality

Modality Datasets	Target	SVM	ANN	LDA	CART
Peripheral data	Peri-Valence	4	2	1	3
	Peri-Arousal	1	4	3	2
	Peri-Liking	1	2	3	4
		2	2.66	2.33	3
EEG Data	EEG-Valence	4	2	1	3
	EEG-Arousal	1	2	3	4
	EEG-Liking	1	2	4	3
		2	2	2.66	3.33
Peripheral & EEG data	(Peri+EEG)-Valence	4	2	1	3
	(Peri+EEG)-Arousal	1	4	2	3
	(Peri+EEG)-Liking	1	3	2	4
Average Ranking Score		2	3	1.66	3.33

“Which of the modalities is better to use when we intend to recognize a certain emotional state?”

The Table 4.11 to Table 4.13 show the average ranking scores of the three different modalities for valence, arousal, and liking recognition. It can be seen that combining the two modalities is better to employ if the objective is to recognize valence while arousal and liking are best recognized by using peripheral and EEG modalities, respectively.

Table 4.11: Average ranking score of the three modalities for Valence recognition

Modality/Classification technique	SVM	ANN	LDA	CART	Avg. Ranking Score
Peripheral Modality	1	3	3	3	2.5
EEG Modality	3	2	2	2	2.25
(Peripheral+EEG) Modalities	2	1	1	1	1.25

Table 4.12: Average ranking score of the three modalities for Arousal recognition

Modality/Classification technique	SVM	ANN	LDA	CART	Avg. Ranking Score
Peripheral Modality	1	2	2	1	1.5
EEG Modality	2	1	3	3	2.25
(Peripheral+EEG) Modalities	3	3	1	2	2.25

Table 4.13: Average ranking score of the three modalities for Liking recognition

Modality/Classification technique	SVM	ANN	LDA	CART	Avg. Ranking Score
Peripheral Modality	3	2	2	3	2.5
EEG Modality	1	1	2	1	1.25
(Peripheral+EEG) Modalities	2	3	1	2	2

- *Experiment 2:*

4.2 Results obtained after applying feature-based multi-classifier methods

This section presents and discusses the results of applying feature-based multi-classifier methods (see Section 3.5) using different data modalities. Table 4.14 to Table 4.16 and their corresponding Figures (Figure 4.7 to Figure 4.9) show average testing rates of the best feature-based multi-classifier methods for valence, arousal, and liking using different data modalities. From the results, it can be observed that there is no single feature selection method that is suitable for all the feature-based multi-classifier methods and all modalities. There are, however, some feature selection methods which frequently achieved the best results such as Fisher which obtained the highest accuracy in 11 out of 36 cases followed by relief and Condred where each has five cases in which they achieved the best accuracy. Specifically, Fisher worked relatively well with ANN (6 out of 9 cases) and somewhat with SVM (4 out 9 cases). In addition, Icap and Relief can be suitable feature selection methods to be used with CART. The detail results of using each feature selection method in feature-based multi-classifier methods for valence, arousal and liking recognition using different data modalities are presented in Appendix B.

Table 4.14: Average testing rates of the best feature-based multi-classifier methods for Peripheral modality

Recognition of/Classifier	CART	ANN	LDA	SVM
Valence	60.0781(Icap)	62.42(Mim)	60.9375(Mim)	55.9375(Fisher)
Arousal	59.60938 (Relief)	63.4375 (Fisher)	60.3125 (Icap)	64.76563 (Fisher)
Liking	63.59375 (Cife)	64.21875(Fisher)	62.42188 (Jmi)	63.82813(t-test)

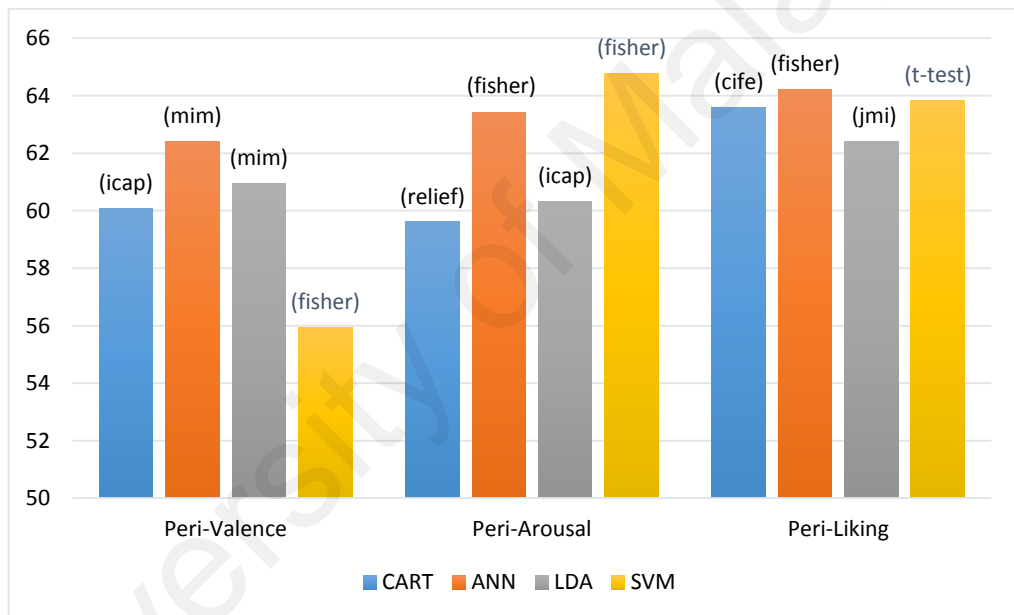


Figure 4.7: Average testing rates of the best feature-based multi-classifier methods for Peripheral modality

Table 4.15: Average testing rates of the best feature-based multi-classifier methods for EEG modality

Recognition of/Classifier	CART	ANN	LDA	SVM
Valence	60.78 (Condred)	66.3281(Fisher)	64.4531(Relief)	53.6719(Distr)
Arousal	60.625 (Condred)	63.9844(Relief)	59.375 (T-Test)	62.9688(Fisher)
Liking	58.5156 (Cmim)	66.1719 (Jmi)	62.4219 (Relief)	66.6406 (t-test)

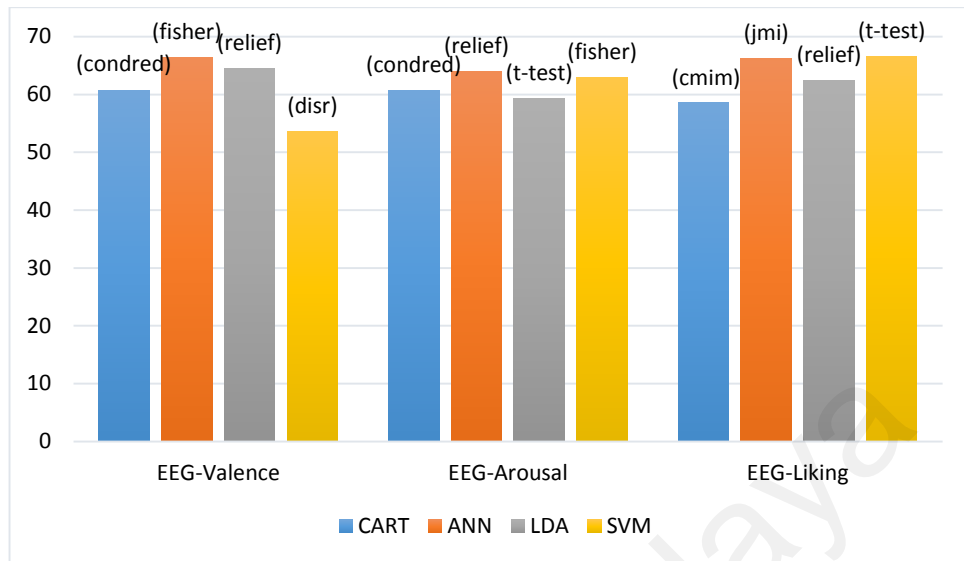


Figure 4.8: Average testing rates of the best feature-based multi-classifier methods for EEG modality

Table 4.16: Average testing rates of the best feature-based multi-classifier methods for (Peripheral+EEG) modalities

Recognition of/ Classifier	CART	ANN	LDA	SVM
Valence	61.02(Relief)	66.5625(Fisher)	65.7813(Cife&Fisher)	53.9844(Condred)
Arousal	59.5313(Icap)	65.0(Fisher)	61.1719(Condred)	63.75(Fisher)
Liking	61.9531(Icap)	65.4688(Fisher)	62.9688(Cife&Condred)	65.2344(Relief)

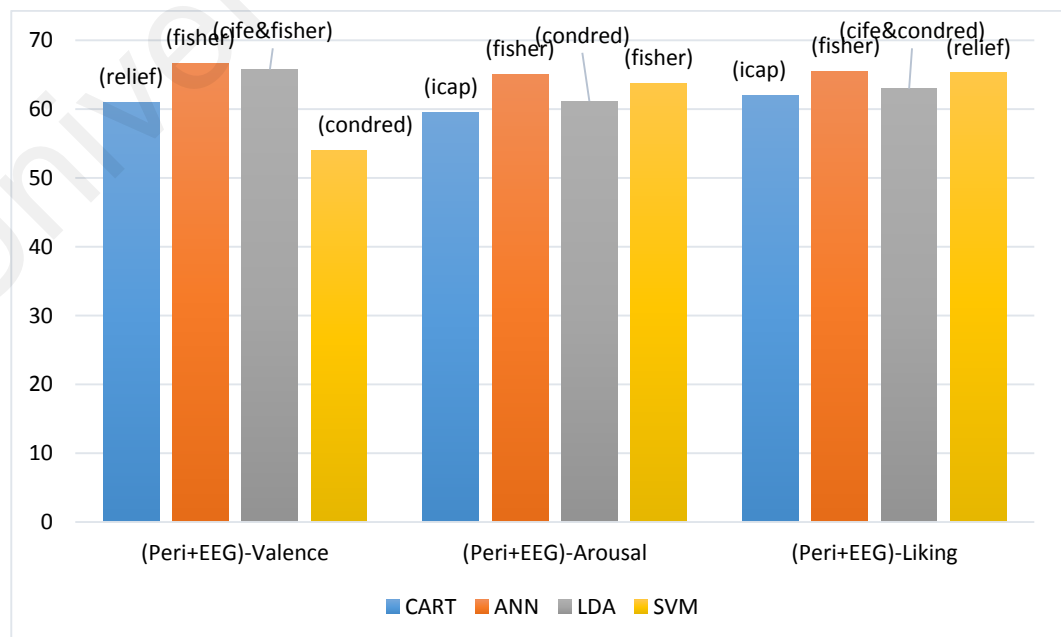


Figure 4.9: Average testing rates of the best feature-based multi-classifier methods for (Peripheral+EEG) modalities

4.2.1 Comparison between single classifiers and feature-based multi-classifier using different modalities

To compare between results of benchmark single classifiers and feature-based multi-classifier, the classification accuracy rates of these two methods are calculated for different emotions and averaged for each modality. In this case, recognition accuracies of valence, arousal, and liking are not separately calculated but averaged. For feature-based multi-classifier, the accuracy is averaged for all the feature selection methods used in this study.

The comparison results are depicted in Figure 4.10 to Figure 4.12. It can be seen that feature selection methods have positively contributed to the improvement of multi-classifier methods compared to single classifiers using single modalities and multimodality data. In addition, ANNs is the classifier that benefited the most after applying the feature selection method gaining between 7.56% and 13.45%, and followed by CART between 2.63% to 5.91%.

SVM gained the least among the classifiers between 1.37% to 1.52% behind LDA which gained between 2.01% and 3.34%. ANNs and CART could gain this much of accuracy because of their instability which enables them to generate more diverse classifiers and thus more accurate ensemble classifier. This result confirms previous findings which stated that CART and ANNs are unstable classifiers and thus are suitable for ensemble methods (Kuncheva, 2014).

The instability in CART is due to the fact that small changes in the training sets can produce very different trained classifiers (Kuncheva, 2014) while the source of instability for ANN is due to the randomness of the initial weight values in the training phase of ANNs. This behavior in addition to the fast training time has made CART the preferred base classifier for many ensemble methods including Bagging (Breiman, 1996), and Addaboost (Freund & Schapire, 1997).

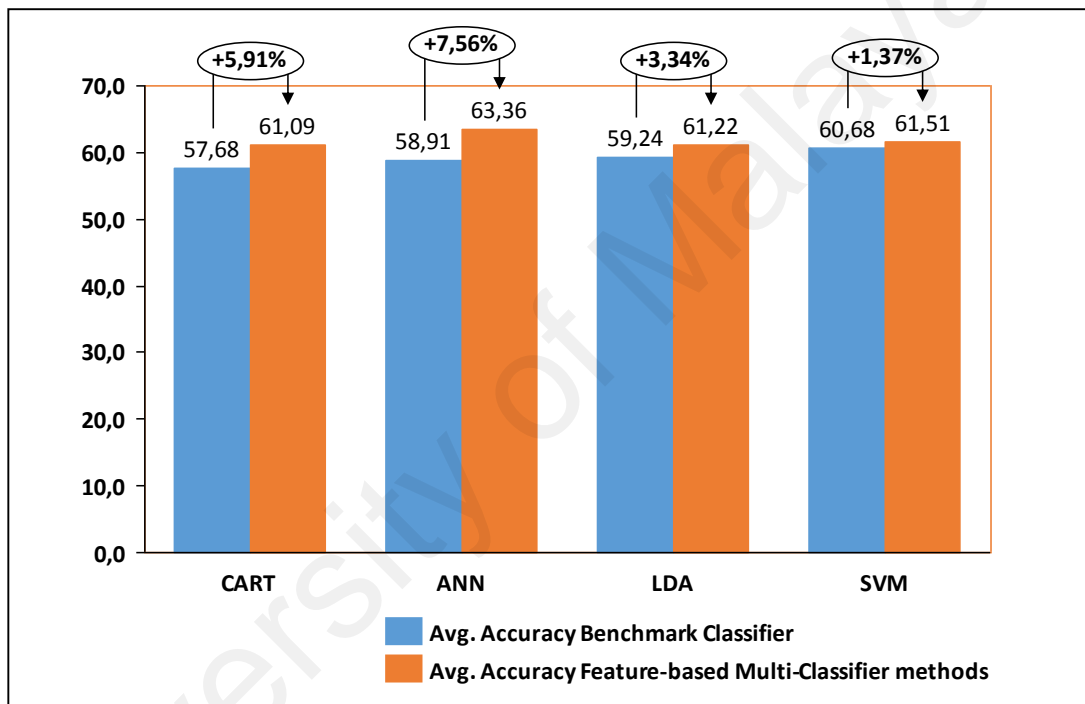


Figure 4.10: Average classification accuracies for each classifier on valence, arousal, and liking using Peripheral modality

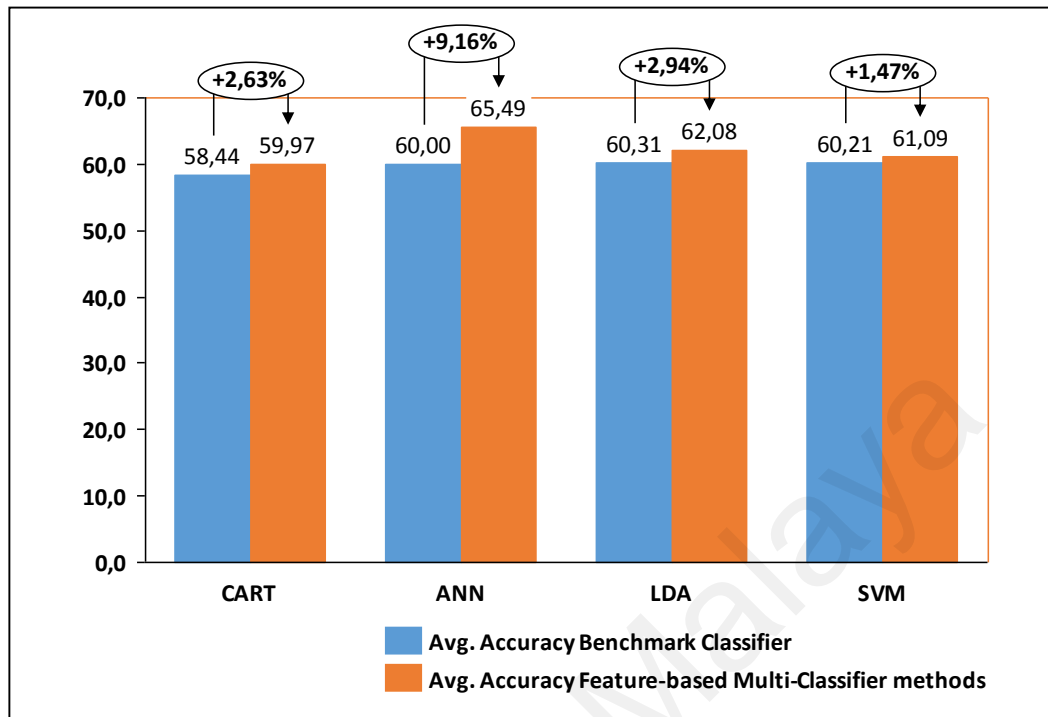


Figure 4.11: Average classification accuracies for each classifier on valence, arousal, and liking using EEG modality

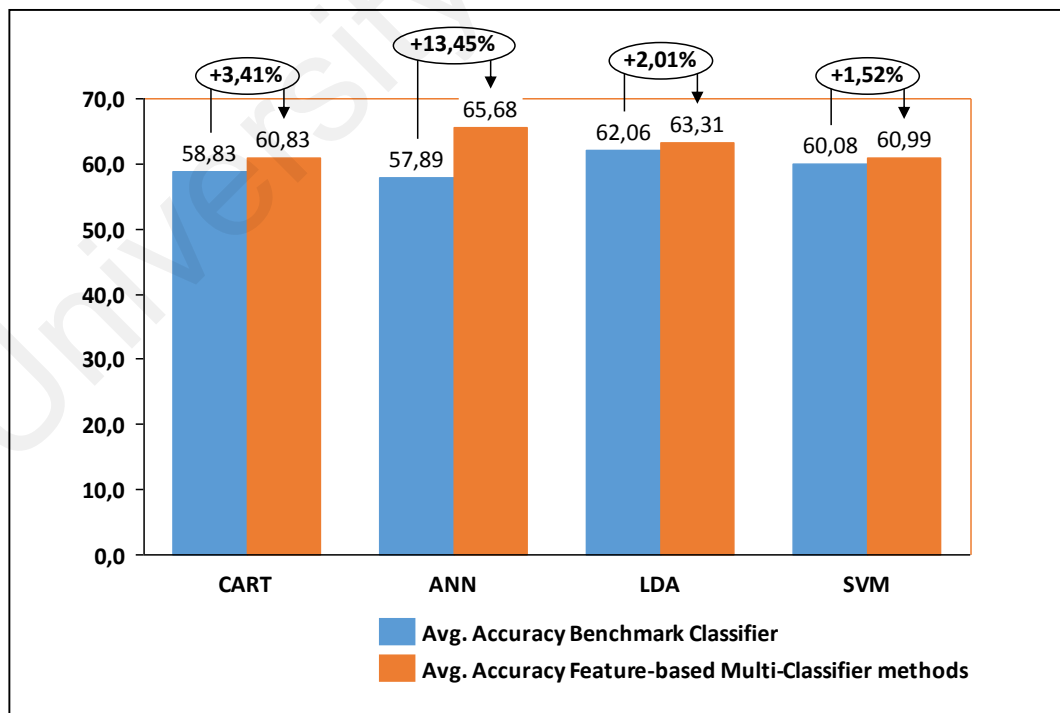


Figure 4.12: Average classification accuracies for each classifier on valence, arousal, and liking using (Peripheral+EEG) modalities

4.2.2 Comparison between different feature subset sizes in terms of accuracy

As it was explained in details in section 3.5, the feature set that belongs to each modality is divided into nine different feature subset sizes using different feature selection methods. The average accuracy rate for each subset in Table 4.17 is the result of averaging different accuracy rates achieved by the combination of different feature selection methods of the four proposed single classifiers for valence, arousal, and liking recognition using three different modalities.

The average ranking score of each feature subset is calculated based on the classification accuracy rates provided in Table 4.17 and it is visualized as in Figure 4.13. Based on the results shown in Figure 4.13 and its corresponding Table 4.17, it is observed that most of the best classification results were achieved by using either the first or second ranked feature subsets¹⁰. This is probably because the first and second feature subsets include the most discriminative and relevant features for valence, arousal and liking recognition. The results also indicate that using a subset of the features is more accurate than using all features. In addition, choosing a subset of features has also a positive effect on the computational cost because using 10% to 20% of the feature set for emotion classification is more computationally efficient than using all the features. Thus, the use of feature selection methods has a positive impact on both the accuracy and computational cost.

¹⁰ FST1 and FST2 represent the first and second subsets of the ranked features

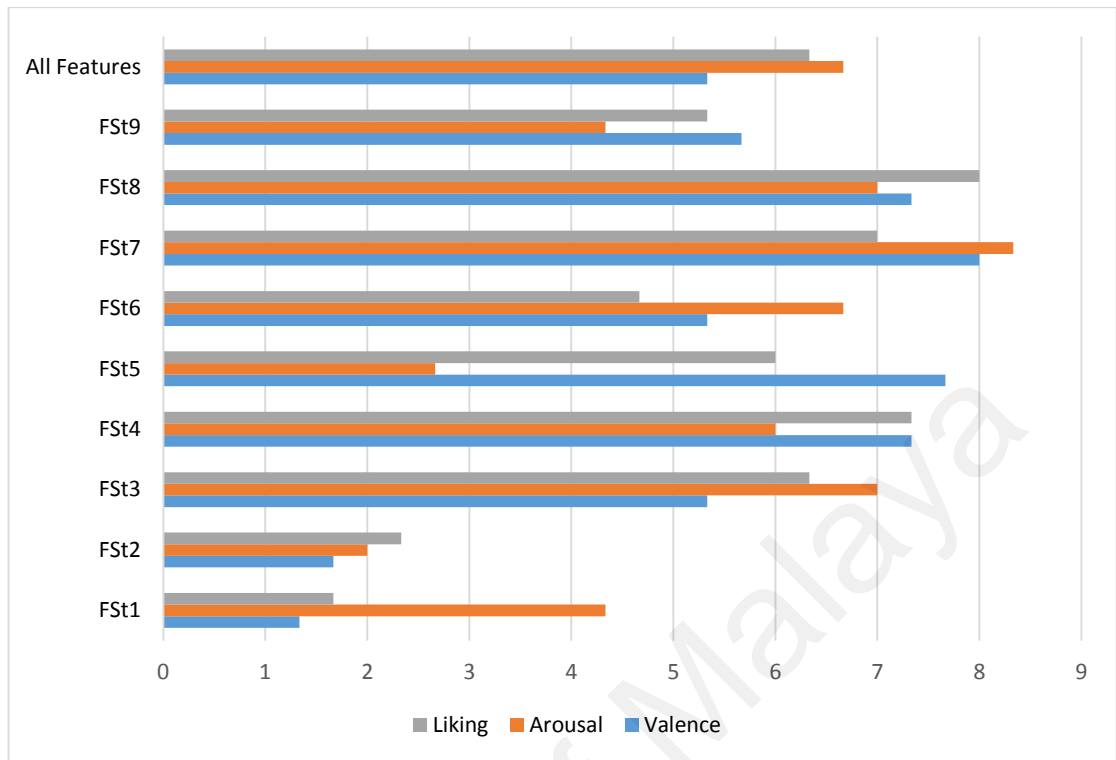


Figure 4.13: Average accuracy ranking scores of Liking, Arousal and Valence using different subset sizes

Table 4.17: Average accuracy rates of Liking, Arousal, and Valence using different subset sizes

Dataset	FSt1	FSt2	FSt3	FSt4	FSt5	FSt6	FSt7	FSt8	FSt9	All features
Peri-Valence	58.723	57.883	57.258	57.555	57.006	57.363	57.197	56.574	56.266	56.934
Peri-Arousal	59.305	59.693	59.100	59.537	59.980	59.646	59.219	59.635	59.736	60.137
Peripheral-Liking	61.842	61.285	61.111	60.549	60.277	61.836	60.555	59.869	59.342	60.313
EEG-Valence	59.793	60.191	58.645	58.160	58.395	58.398	57.580	58.305	58.854	58.516
EEG-Arousal	59.816	60.006	59.455	59.336	59.828	58.875	59.082	59.180	59.125	59.004
EEG-Liking	62.064	62.672	60.668	60.902	61.992	60.861	60.980	60.762	61.238	61.699
Peri+EEG-Valence	61.029	59.842	58.469	58.100	58.125	58.756	58.389	59.004	59.209	59.258
Peri+EEG-Arousal	59.984	60.316	59.443	59.590	59.744	59.719	58.797	58.664	59.838	58.516
Peri+EEG-Liking	62.154	62.152	61.807	61.336	61.434	61.875	60.859	61.674	62.207	61.367

- *Experiment 3:*

4.3 The results of proposed classification method: Feature-Based Dual-layer Ensemble Classification Method

The results of the three best feature-based dual-layer ensemble classification methods (FDLEC) for valence, arousal and liking recognition using different modalities are shown in Table 4.18 while Figure 4.14 depicts the highest accuracies that could be achieved employing these FDLEC methods. The results show that the best FDLEC method is SVM+CART followed by SVM+SVM and then CART+CART.

Table 4.19 and its corresponding Figure 4.15 also shown average ranking scores calculated for each tested FDLEC method according to the classification accuracies results for valence, arousal and liking recognition using different modalities provided in Table 4.20 to Table 4.22 and its related Figure 4.16 to Figure 4.18, respectively. Table 4.18 and Figure 4.14 show that FDLEC(SVM+CART) method dominates the results by achieving the best accuracy rate in 7 data sets out of 9. FDLEC(CART+CART) achieved the best results in 2 data sets out of 9. FDLEC(SVM+SVM) did not get the best results in any of the data sets despite it is ranked second in overall ranks (see Table 4.19) before FDLEC (CART+CART). One possible interpretation of this result is that SVM which achieved the best single classifier could maintain its dominant position while other classifiers especially ANN apparently could not gain much accuracy in the first layer to outperform SVM. In addition, in the second layer and since we have binary features (zeros and ones), CART seems more efficient than other classification algorithms in handling this type of feature variables (categorical variable).

Table 4.18: Average accuracy rates of the three best feature-based dual-layer ensemble classification method

Modality Datasets	Target	FDLEC CART+CART	FDLEC SVM+CART	FDLEC SVM+SVM
Peripheral data	Peri-Valence	62.0313	69.22	67.34
	Peri-Arousal	66.2675	70.78	67.58
	Peri-Liking	65	72.89	69.45
EEG Data	EEG-Valence	66.4844	68.20	66.56
	EEG-Arousal	65.7031	66.09	65.78
	EEG-Liking	70.7969	67.97	66.33
Peripheral & EEG data	(Peri+EEG)- Valence	69.375	66.56	66.72
	(Peri+EEG)- Arousal	67.7344	68.36	66.17
	(Peri+EEG)- Liking	71.0156	71.64	68.75

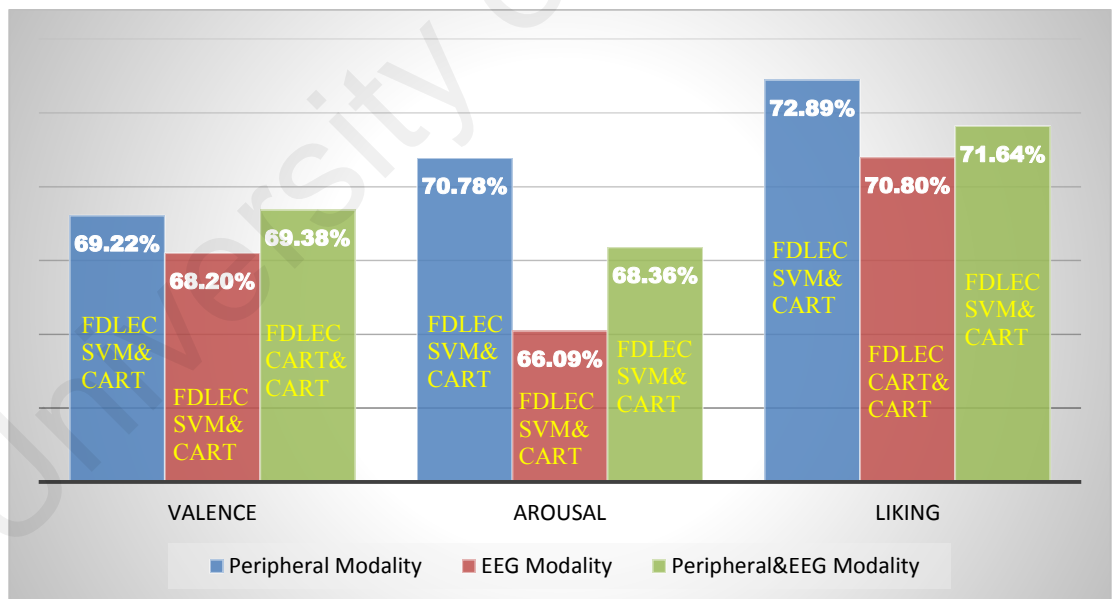


Figure 4.14: The best accuracy rates among the feature-based dual-layer ensemble classification methods

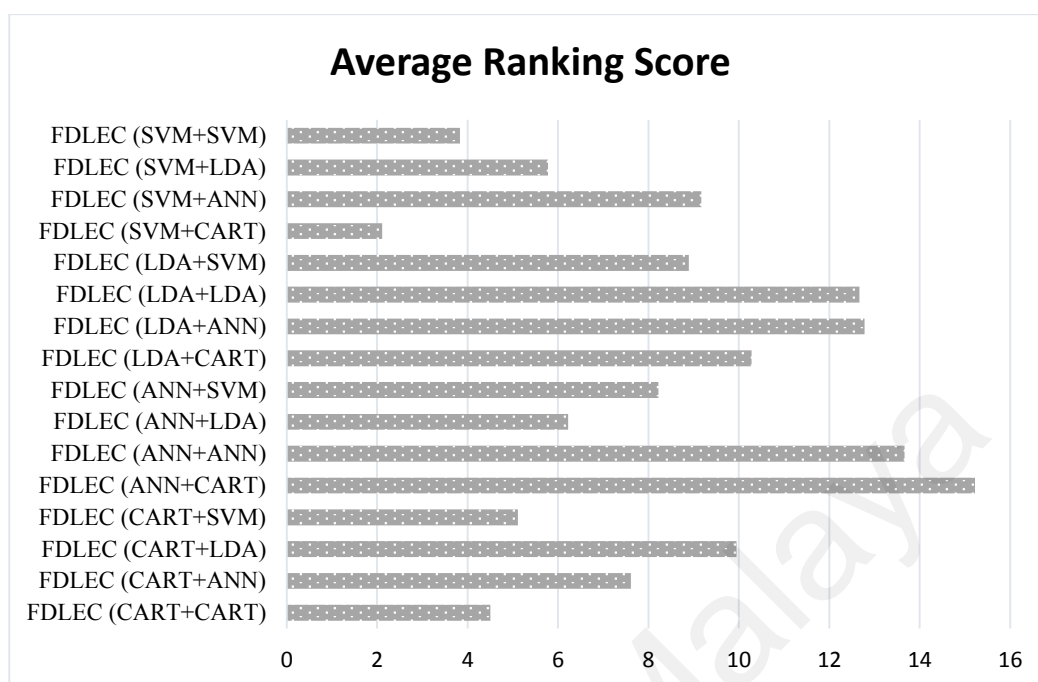


Figure 4.15: Average ranking scores of the feature-based dual-layer ensemble classification methods over the three modalities

Table 4.19: Average ranking scores of the feature-based dual-layer ensemble classification methods over the three modalities

Dataset	Peripheral data			EEG Data			Peripheral & EEG data			Method Ranking Score
	V	A	L	V	A	L	V	A	L	
FDLEC CART+CART	10.5	6	11	4	3	1	1	2	2	4.500
FDLEC CART+ANN	8	4	7	11.5	8	4	13	3	10	7.611
FDLEC CART+LDA	5	10	8	11.5	10	13	7	9	16	9.944
FDLEC CART+SVM	7	3	2	10	6	2	3	4	9	5.111
FDLEC ANN+CART	16	16	15	16	15	15	15	16	13	15.222
FDLEC ANN+ANN	15	13	9	15	13	14	16	13	15	13.667
FDLEC ANN+LDA	4	5	5	5.5	11	3	9	7	6.5	6.222
FDLEC ANN+SVM	10.5	7	13	8.5	7	8.5	8	5	6.5	8.222
FDLEC LDA+CART	14	12	10	8.5	4	12	10	11	11	10.278
FDLEC LDA+ANN	13	14	16	14	12	10	12	12	12	12.778
FDLEC LDA+LDA	12	15	14	7	16	16	6	14	14	12.667
FDLEC LDA+SVM	9	8	12	13	5	6	14	10	3	8.889
FDLEC SVM+CART	1	1	1	1	1	7	5	1	1	2.111
FDLEC SVM+ANN	3	11	4	5.5	14	11	11	15	8	9.167
FDLEC SVM+LDA	6	9	6	2	9	5	2	8	5	5.778
FDLEC SVM+SVM	2	2	3	3	2	8.5	4	6	4	3.833

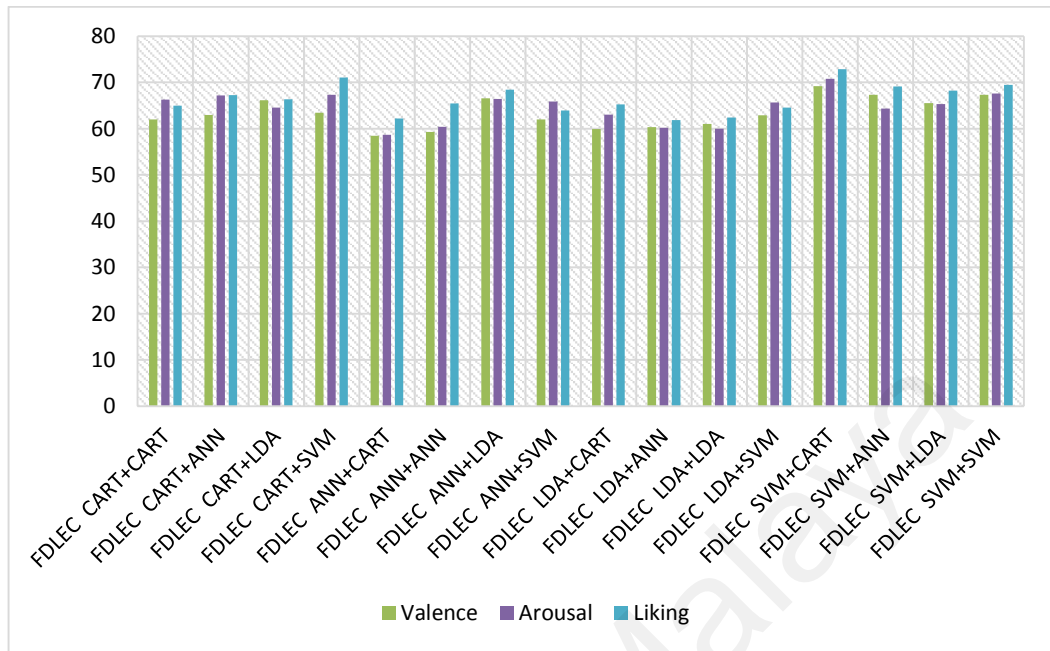


Figure 4.16: Testing accuracy rates of feature-based dual-layer ensemble classification methods on Valence, Arousal and liking with peripheral modality

Table 4.20: Testing accuracy rates of feature-based dual-layer ensemble classification methods on Valence, Arousal and liking with peripheral modality

Method/Target	Peri-Valence	Peri-Arousal	Peri-Liking
FDLEC CART+CART	62.0313	66.2675	65
FDLEC CART+ANN	62.9677	67.1875	67.253
FDLEC CART+LDA	66.1719	64.5313	66.3281
FDLEC CART+SVM	63.4375	67.3438	71.093
FDLEC ANN+CART	58.4375	58.6719	62.1875
FDLEC ANN+ANN	59.2969	60.3906	65.4688
FDLEC ANN+LDA	66.5625	66.4063	68.4375
FDLEC ANN+SVM	62.0313	65.8594	63.9063
FDLEC LDA+CART	59.9219	63.0469	65.2344
FDLEC LDA+ANN	60.3125	60.2344	61.875
FDLEC LDA+LDA	61.0156	60	62.4219
FDLEC LDA+SVM	62.8906	65.7031	64.5313
FDLEC SVM+CART	69.2188	70.7813	72.8906
FDLEC SVM+ANN	67.313	64.375	69.1406
FDLEC SVM+LDA	65.5469	65.3125	68.2031
FDLEC SVM+SVM	67.3438	67.5781	69.4531

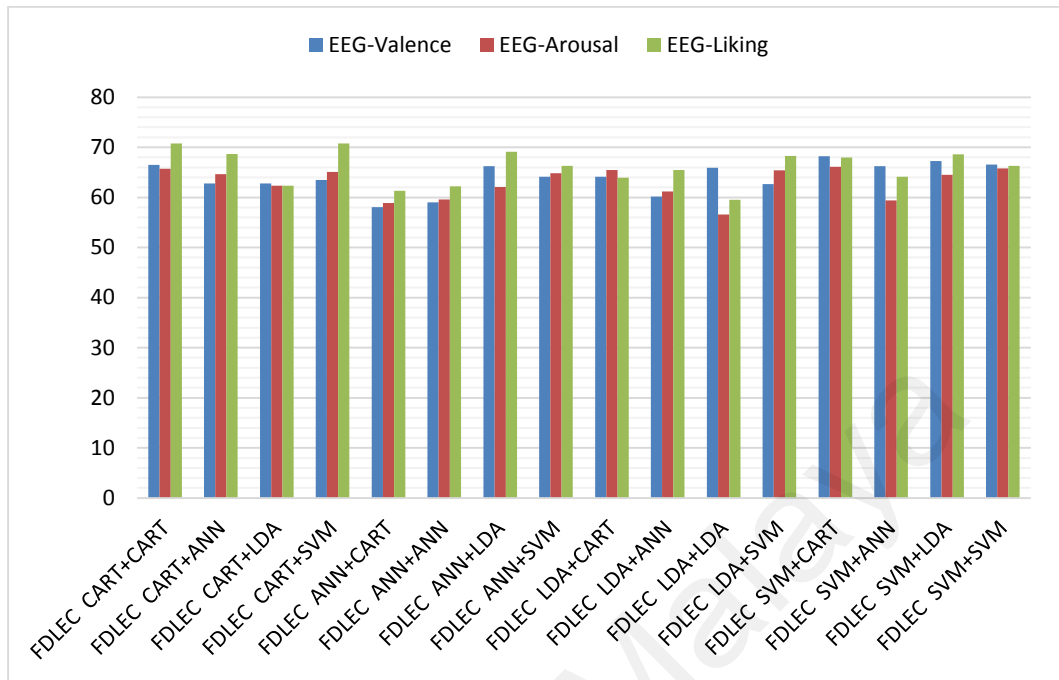


Figure 4.17: Testing accuracy rates of feature-based dual-layer ensemble classification methods on Valence, Arousal and liking with EEG modality

Table 4.21: Testing accuracy rates of feature-based dual-layer ensemble classification methods on Valence, Arousal and liking with EEG modality

Method/Target	EEG-Valence	EEG-Arousal	EEG-Liking
FDLEC CART+CART	66.4844	65.7031	70.7969
FDLEC CART+ANN	62.8115	64.6094	68.6719
FDLEC CART+LDA	62.8115	62.3438	62.3438
FDLEC CART+SVM	63.5156	65.0781	70.7813
FDLEC ANN+CART	58.0469	58.9063	61.3281
FDLEC ANN+ANN	58.9844	59.6094	62.1875
FDLEC ANN+LDA	66.25	62.1094	69.1406
FDLEC ANN+SVM	64.1406	64.8438	66.3281
FDLEC LDA+CART	64.1406	65.4688	63.9063
FDLEC LDA+ANN	60.1563	61.1719	65.4688
FDLEC LDA+LDA	65.9375	56.5625	59.5313
FDLEC LDA+SVM	62.6563	65.3906	68.2813
FDLEC SVM+CART	68.2031	66.0938	67.9688
FDLEC SVM+ANN	66.25	59.375	64.1406
FDLEC SVM+LDA	67.2656	64.5313	68.5938
FDLEC SVM+SVM	66.5625	65.7813	66.3281

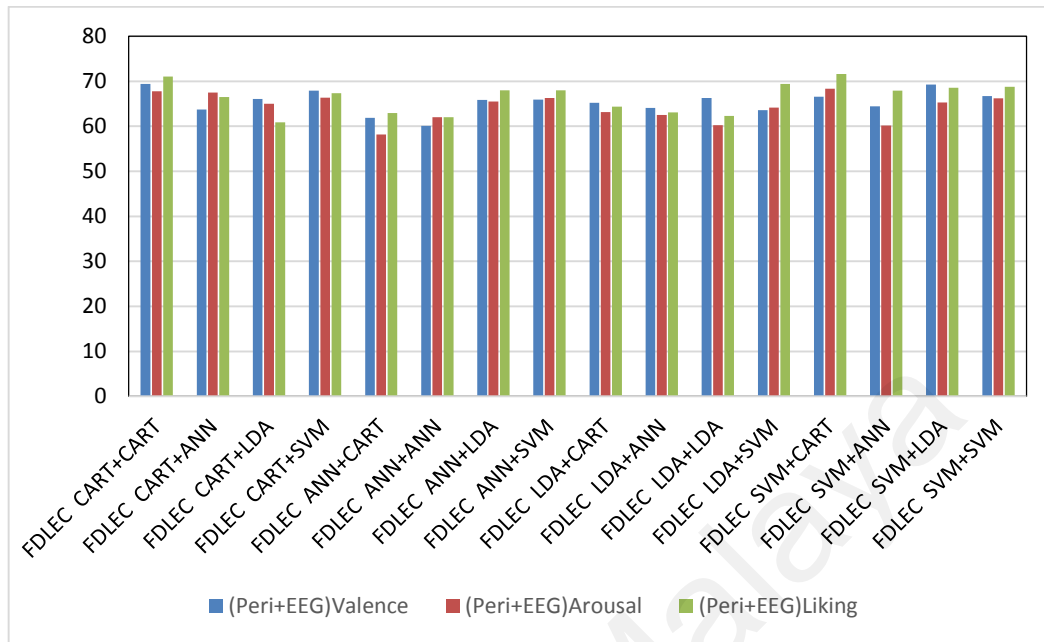


Figure 4.18: Testing accuracy rates of feature-based dual-layer ensemble classification methods on Valence, Arousal and liking with (Peripheral+EEG) modalities

Table 4.22: Testing accuracy rates of feature-based dual-layer ensemble classification methods on Valence, Arousal and liking with (Peripheral+EEG) modalities

Method/Target	(Peri+EEG) Valence	(Peri+EEG) Arousal	(Peri+EEG) Liking
FDLEC CART+CART	69.375	67.7344	71.0156
FDLEC CART+ANN	63.7	67.5	66.4844
FDLEC CART+LDA	66.0938	65	60.8594
FDLEC CART+SVM	67.8906	66.3281	67.3438
FDLEC ANN+CART	61.875	58.2031	62.9687
FDLEC ANN+ANN	60.0781	62.0313	62.0313
FDLEC ANN+LDA	65.8594	65.4688	67.9688
FDLEC ANN+SVM	65.9375	66.25	67.9688
FDLEC LDA+CART	65.2344	63.125	64.375
FDLEC LDA+ANN	64.0625	62.5	63.0469
FDLEC LDA+LDA	66.25	60.2344	62.2656
FDLEC LDA+SVM	63.5938	64.1406	69.375
FDLEC SVM+CART	66.5635	68.3594	71.6406
FDLEC SVM+ANN	64.4531	60.1563	67.8906
FDLEC SVM+LDA	69.2969	65.3125	68.5156
FDLEC SVM+SVM	66.7188	66.1719	68.75

- *Experiment 3(cont.):*

4.4 Comparison between feature-based dual-layer ensemble classification methods and other classifiers

Table 4.23 and Figure 4.19 show the average ranking results of the best classifiers taken from three different categories. The first category which represents single classifiers includes four classifiers, namely, CART, ANN, LDA, and SVM. The second category is represented by five feature-based multi-classifier methods that received the best results. These methods are: ANN+fisher, ANN+jmi, ANN+mim and SVM+fisher. The last category includes the following three best feature-based dual-layer ensemble classification (FDLEC) methods: CART+CART, SVM+CART, and SVM+SVM. Random Forest (RF) which is a well-known ensemble classifier is also included for comparison purpose.

The results show that the three FDLEC methods achieved the best results followed by Random Forest, and then feature-based multi-classifier methods. The single classifiers received the lowest accuracy rates compared to other methods. In addition, FDLEC (SVM+CART) is the best method among all the compared methods.

As we expected and based on the finding from literature review, the results also illustrated that using feature selection method generally has a positive effect on the classification accuracy rate of the recognition system compared to the approach of not using a feature selection method. In addition, the results proved that the proposed FDLEC method, which works based on a combination of first and second layer classifiers (stacking ensemble strategy) for an emotion prediction has obtained better classification accuracy rate compared to the use of a single feature selection in our tested feature-based multi-classifier methods. This result could be expected since the physiological data considered as high dimensional data and using multiple feature ranking methods in the

first layer (base level) of the proposed ensemble classification technique provided more diverse feature subsets and thus more accurate classification models in the first layer, and by using stacking ensemble strategy, the ensemble classifier can further learn, in the second layer, the behavior of first-layer classification methods and so it can successfully recognize which classifiers of the first layer are reliable to be used for final prediction result (Santana & Canuto, 2014). Eventually, the proposed ensemble method provided better classification accuracy compared to using feature-based multi-classifier methods that employed single feature selection method.

To compare the proposed FDLEC method with another kind of ensemble algorithms, the Random Forest (RF) method which is one of the most successful and powerful ensemble methods that works based on a bagging algorithm and exhibited performance comparable to the level of boosting (Robnik-Šikonja, 2004), was selected. The obtained results showed that the proposed FDLEC method could also outperform the RF.

Table 4.23: Average ranking scores of the best classifiers taken from three different categories

Dataset	Peripheral data			EEG Data			Peripheral & EEG data			Average Method Ranking Score
Method/Target	V	A	L	V	A	L	V	A	L	
CART	56.56	58.83	57.66	58.44	56.80	60.08	58.91	58.20	59.38	<i>11.556</i>
ANN	57.97	58.28	60.47	58.91	59.53	61.56	59.92	54.22	59.53	<i>10.889</i>
LDA	58.98	58.75	60.00	63.83	57.11	60.00	65.08	59.45	61.64	<i>10.111</i>
SVM	54.22	64.69	63.13	52.89	62.58	65.16	53.13	62.19	64.92	<i>8.944</i>
RF	62.34	63.59	65.00	65.00	62.19	65.39	67.03	63.60	66.09	<i>5.500</i>
ANN+Fisher	62.11	63.44	64.22	66.33	62.50	66.02	66.56	65.00	65.47	<i>5.667</i>
ANN+Jmi	60.47	63.28	63.75	63.59	63.59	66.17	64.61	61.95	64.14	<i>7.278</i>
ANN+mim	62.42	62.42	62.66	63.28	63.59	64.69	64.77	63.75	63.75	<i>7.500</i>
SVM+fisher	55.94	64.77	62.66	53.20	62.97	66.56	53.20	63.75	64.84	<i>7.778</i>
SVM+t-test	53.44	62.58	63.83	49.38	62.66	66.64	51.56	61.88	64.92	<i>8.944</i>
FDLEC CART+CART	62.03	66.268	65	66.48	65.7	70.8	69.38	67.73	71.02	<i>2.722</i>
FDLEC SVM+CART	69.22	70.78	72.89	68.20	66.09	67.97	66.56	68.36	71.64	<i>1.444</i>
FDLEC SVM+SVM	67.34	67.58	69.45	66.56	65.78	66.33	66.72	66.17	68.75	<i>2.667</i>

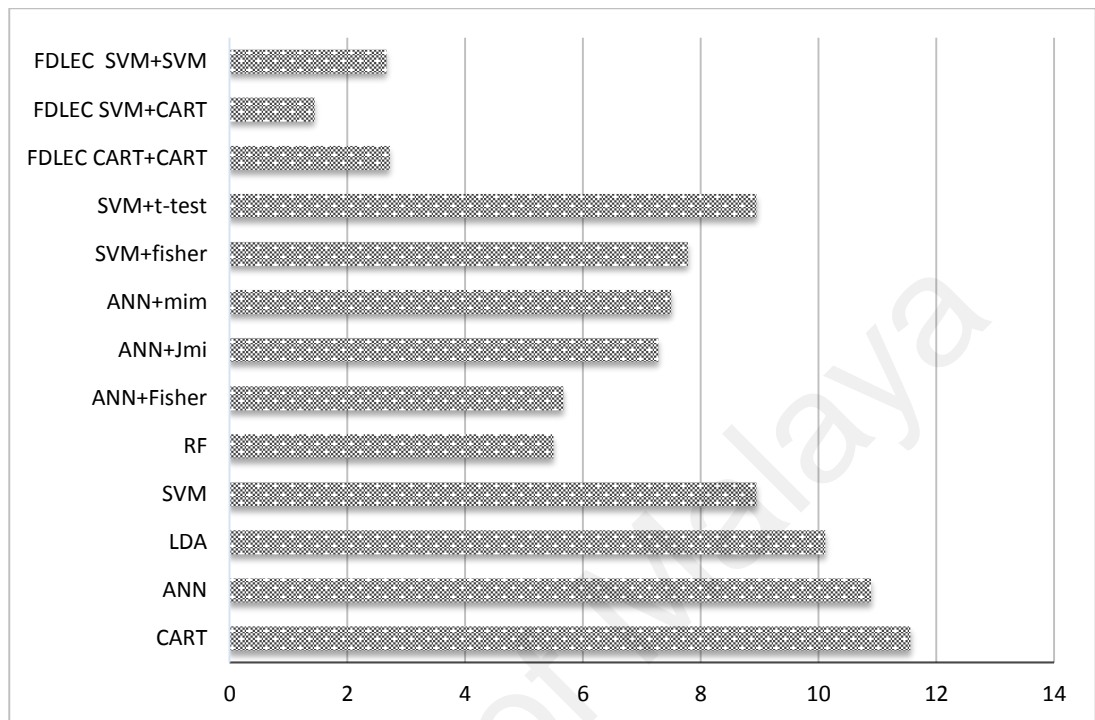


Figure 4.19: Average ranking scores of the best classifiers taken from three different categories

To evaluate the improvement made by using the proposed method, we plot Figure 4.20 to Figure 4.22 that represent the best accuracy recognition of valence, arousal, and liking achieved by different classification methods using peripheral, EEG and (Peripheral+EEG) modalities, respectively.

Compared to the best single classifiers, feature-based dual-layer ensemble classification methods have improved the accuracy between 5.62% and 17.36%.

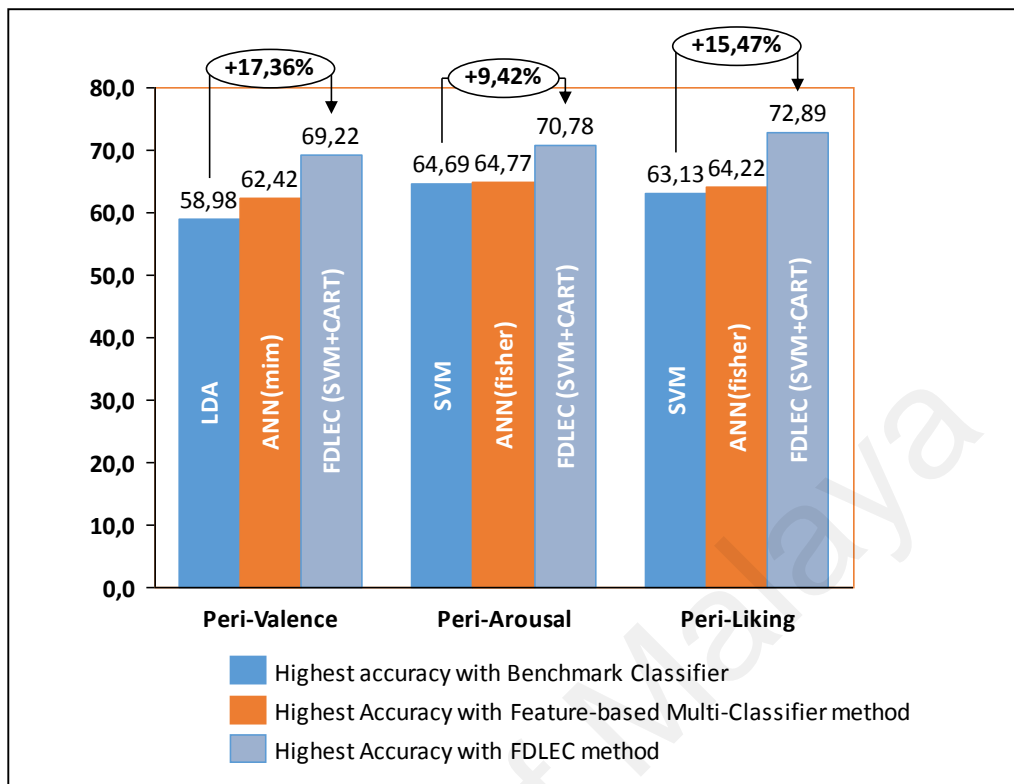


Figure 4.20: The best testing accuracy rates of each of the three categories for peripheral modality

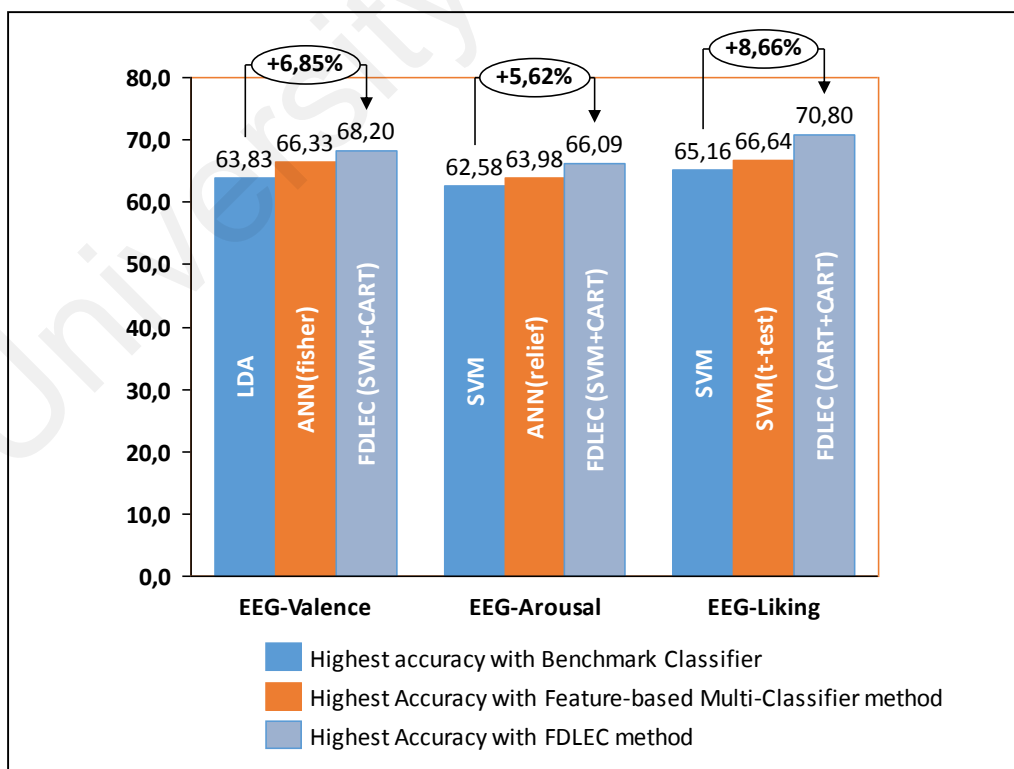


Figure 4.21: The best testing accuracy rates of each of the three categories for EEG modality

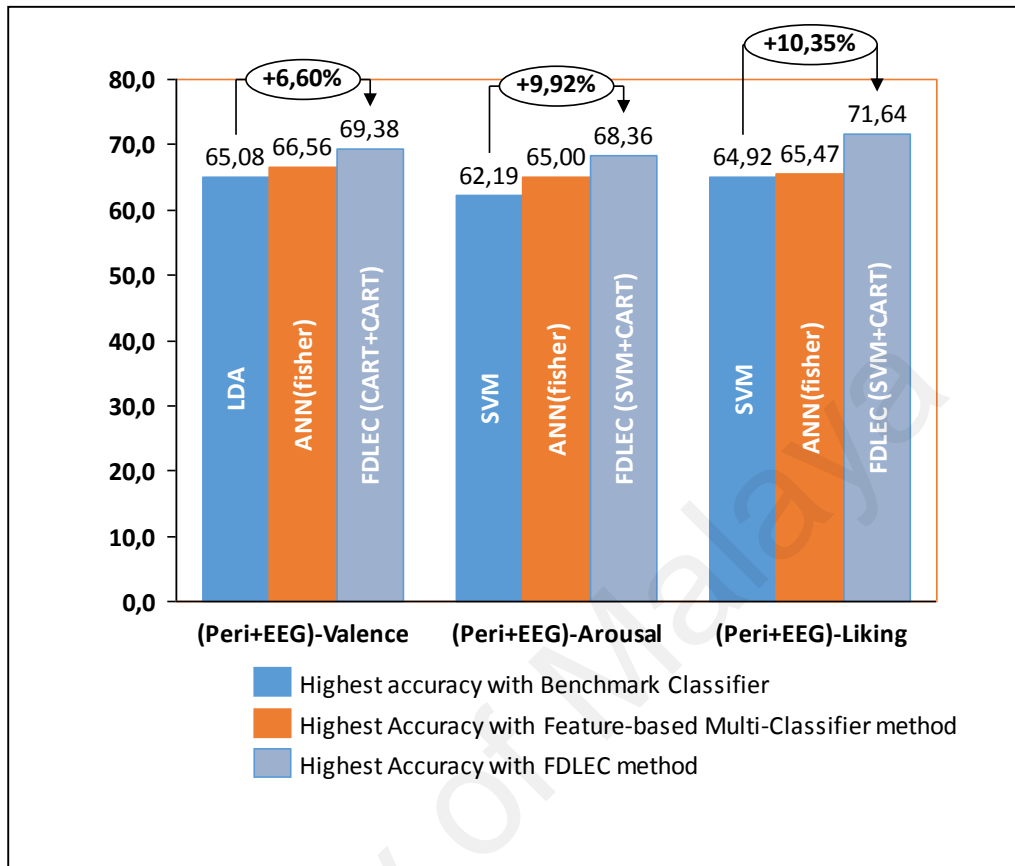


Figure 4.22: The best testing accuracy rates of each of the three categories for (Peripheral +EEG) modalities

4.4.1 Physiological modalities performance comparison for emotional state recognition

To check which of the modalities is probably better to use to detect a certain emotional state, an average ranking score of the three modalities for valence, arousal and liking recognition using classification accuracy rates of the best performed classifiers from previous experiments are calculated. The Table 4.24 to Table 4.26 shows the average ranking scores that were calculated. These results confirmed the early conclusion from benchmark experiments (see Section 4.1.2) that combining the two modalities is better to employ if the objective is to recognize valence while arousal and liking are best recognized by using peripheral and EEG modalities, respectively.

University of Malaysia

Table 4.24: Average ranking score of the three modalities for valence recognition

Modality/Classification technique	CART	ANN	LDA	SVM	RF	ANN+Fisher	ANN+Jmi	ANN+mim	SVM+fisher	SVM+t-test	FDLEC CART+CART	FDLEC SVM+CART	FDLEC SVM+SVM	Avg. Ranking Score
Peripheral Modality	3	3	3	1	3	3	3	3	1	1	3	1	1	2.23
EEG Modality	2	2	2	3	2	2	2	2	2.5	3	2	2	3	2.27
(Peripheral+EEG) Modalities	1	1	1	2	1	1	1	1	2.5	2	1	3	2	1.5

Table 4.25: Average ranking score of the three modalities for arousal recognition

Modality/Classification technique	CART	ANN	LDA	SVM	RF	ANN+Fisher	ANN+Jmi	ANN+mim	SVM+fisher	SVM+t-test	FDLEC CART+CART	FDLEC SVM+CART	FDLEC SVM+SVM	Avg. Ranking Score
Peripheral Modality	1	2	2	1	2	2	2	3	1	2	2	1	1	1.7
EEG Modality	3	1	3	2	3	3	1	2	3	1	3	3	3	2.39
(Peripheral+EEG) Modalities	2	3	1	3	1	1	3	1	2	3	1	2	2	1.93

Table 4.26: Average ranking score of the three modalities for liking recognition

Modality/Classification technique	CART	ANN	LDA	SVM	RF	ANN+Fisher	ANN+Jmi	ANN+mim	SVM+fisher	SVM+t-test	FDEC CART+CART	FDEC SVM+CART	FDEC SVM+SVM	Avg. Ranking Score
Peripheral Modality	3	2	2.5	3	3	3	3	3	3	3	3	1	1	2.58
EEG Modality	1	1	2.5	1	2	1	1	1	1	1	2	3	3	1.58
(EEG+Periphera) Modalities	2	3	1	2	1	2	2	2	2	2	1	2	2	1.85

- *Experiment 3(cont.):*

4.5 Comparison between classifiers using Statistical method

In this section, we check whether the difference in classification accuracy rates between the proposed method, FDLEC (SVM+CART), which achieved the best average ranks and some classifiers which received the best results, is statistically significant or not. These classifiers which we call them “selected classifiers” include the following methods:

- The four benchmark single classifiers (see Section 4.1).
- The four best feature-based multi-classifier methods (see Section 4.2)
- The three best feature-based dual-layer ensemble classification methods including the FDLEC (SVM+CART) (see Section 4.3).
- Random Forest as a benchmark ensemble classifier.

Based on the recommendation made by (Demsar, 2006), we apply Wilcoxon signed ranks test (see Section 3.8.2.3) separately between the proposed method, FDLEC (SVM+CART) and each of the selected classifiers.

Based on the table of exact critical values for the Wilcoxon’s test, for a confidence level of $\alpha = 0.05$ and $N = 9$ data sets, the difference between the classifiers is significant if $T = \min(R^+, R^-)$ is equal or less than 6. Table 4.27 depicts the calculated values for R^+ , R^- and T . R^+ is the sum of ranks for the datasets where the second algorithm outperformed the first and R^- is the sum of ranks where the first algorithm performs better than second one. To compare the performance of the selected algorithms with that of our proposed method, 9 data sets are used. The first algorithm is our proposed FDLEC (SVM+CART) and the second algorithms are each selected benchmark classifiers as shown in Table 4.27. It can be seen that, the value of T is either 0 or 1 and so less than 6, which means the null hypothesis is rejected for all the selected classifiers. We can say

also that the p-value of all comparison was less than 0.05 and thus FDLEC (SVM+CART) is significantly better than all the selected classifiers.

Table 4.27: Results obtained by Wilcoxon test for feature-based dual-layer (SVM+CART) algorithm

Versus Classification Method	$R+$	$R-$	Exact P-value	T
CART	45	0	0.003906	0
ANN	45	0	0.003906	0
LDA	45	0	0.003906	0
SVM	45	0	0.003906	0
RF	44	1	0.007812	1
ANN+Fisher	36	0	0.007812	0
ANN+Jmi	45	0	0.003906	0
ANN+mim	45	0	0.003906	0
SVM+fisher	45	0	0.003906	0
SVM+t-test	45	0	0.003906	0
FDLEC CART+CART	44	1	0.007812	1
FDLEC SVM+SVM	44	1	0.007812	1

4.6 Comparison between existing works

As it was mentioned in conclusion section of chapter 2, making a comparison between classification accuracy rates of different studies associated with physiological-based emotion recognition systems is not an easy task. One of the main reasons behind this difficulty is that different studies usually apply their own specific experimental setups like the selection of the type of the extracted features from the signals (e.g. frequency-based features, time-based features and etc.) and the calculation method of a particular feature. In this study, since the values of each feature was not provided by the developers of DEAP dataset, we develop our own code to calculate these features. Therefore, the comparison of accuracy rates obtained by the present study with similar studies may not be fair. But we prefer to include this comparison, which compares our results with the studies that use the same data set, as it can be useful for this area of research.

The comparison of various studies that used the same data set is given in Table 4.28 and its related Figure 4.23-26. Most of the studies have only used EEG data modality in their experiments. The developer of DEAP data set, (Koelstra et al., 2012), reported 62.00%, 57.60% and 55.4% (for arousal, valence and liking, respectively) accuracies using EEG modality and 57%, 62.7% and 59.1% (for arousal, valence and liking, respectively) using peripheral modality using Gaussian Bayes classifier. Since, our research study has tried to follow the same experimental setups and calculate the same physiological features as proposed by Koelstra et al. (2012), comparing results of their methods to that of our proposed method for arousal, valence and liking recognition (EEG: 66.09%, 68.20% and 68% for arousal, valence, and liking, respectively) and (Peripheral: 70.78%, 69.22% and 72.89% for arousal, valence and liking, respectively) proves the soundness and efficiency of our proposed system.

The achieved classification accuracy rates reported by (Naser & Saha, 2013) is 66.20% and 64.30%, (Clerico et al., 2015) is 66% and 61%, (Y. Liu & Sourina, 2012) is 76.51% and 50.80% for arousal and valence recognition, respectively. Liking recognition was tested by (Naser & Saha, 2013) and (Clerico et al., 2015) with accuracy rates of 70.2% and 62%, respectively. Regarding Liu and Sourina (2012), it was found that the reported high accuracy of 76.51% for arousal recognition, this is due to the combination of arousal and dominance dimensions together rather than considering arousal dimension only. Based on the reviewed similar studies, it is shown that the performance of our proposed feature-based dual-layer ensemble classification method is good and competitive with similar studies. Based on Table 4.28, Figure 4.23, 4.24 and 4.25 depict the comparison of accuracy rates obtained by present study and three similar studies that used the same data set for Arousal, Valence and Liking recognition using EEG modality respectively. The comparison of accuracy rates of present study and a benchmark study for Arousal, Valence and Liking recognition using peripheral modality is also shown in Figure 4.26.

Table 4.28: Accuracy rates comparison of five similar studies

Study	Modality/Type of extracted features	Feature Selection	Classifier	Accuracy Rates
(Koelstra et al., 2012)	Peripheral and EEG	Fisher's linear discriminant	Gaussian Naïve Bayes	EEG: Arousal (62%) Valence (57.6%) Liking (55.4%) Peripheral: Arousal (57%) Valence (62.7%) Liking (59.1%)
(Naser & Saha, 2013)	EEG	Singular value decomposition (SVD), QR factorization with column pivoting (QRcp) and F-ratio based method	SVM	EEG: Arousal (66.2%) Valence (64.3%) Liking (70.2%)
(Liu & Sourina, 2012)	EEG	-	SVM	EEG: Arousal (76.51%) Valence (50.8%)
(Clerico et al., 2015)	EEG	Minimum redundancy maximum relevance algorithm	SVM	EEG: Arousal (66%) Valence (61%) Liking (62%)
Present Study	Peripheral, EEG and combination of EEG& Peripheral	Combination of 10 feature ranking methods	Feature-based Dual-layer ensemble classification method FDLEC (SVM+CART)	EEG: Arousal (66.09%) Valence (68.20%) Liking (68%) Peripheral: Arousal (70.78%) Valence (69.22%) Liking (72.89%) Combination: Arousal (68.36%) Valence (66.56%) Liking (71.64%)

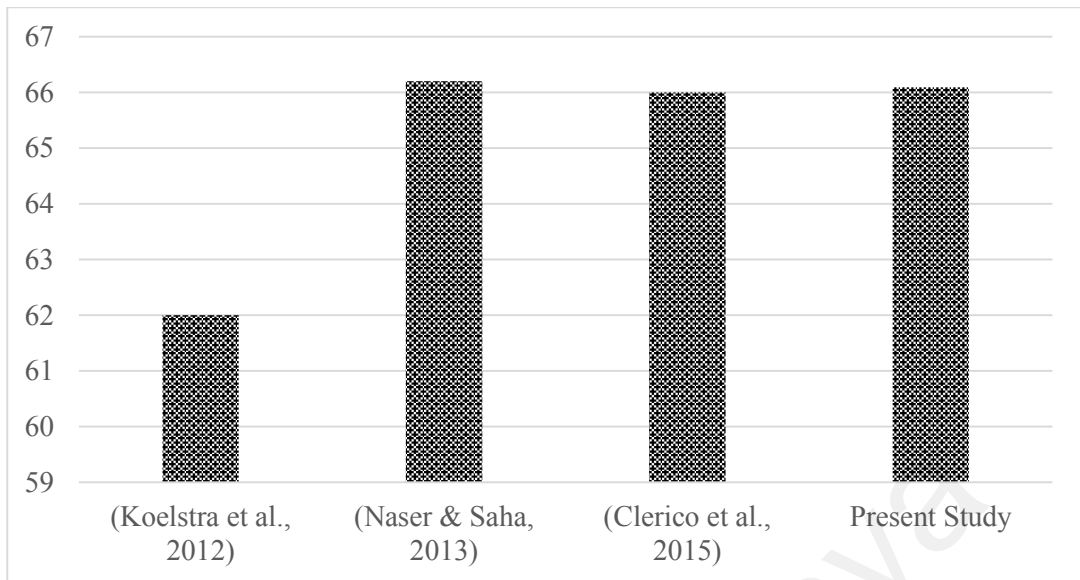


Figure 4.23: The comparison of testing accuracy rates of the present study and three other similar studies for Arousal recognition using EEG modality

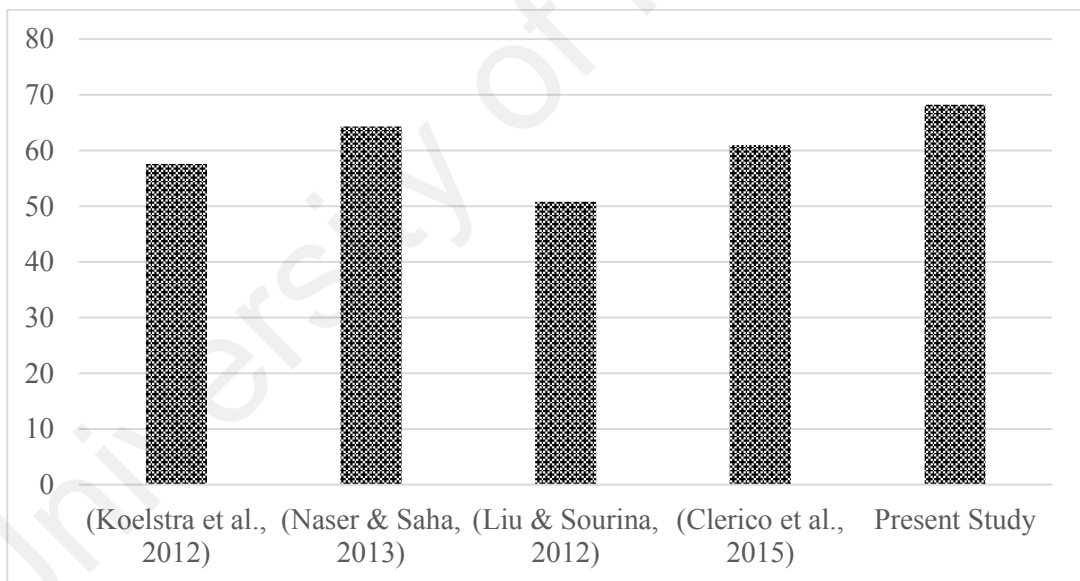


Figure 4.24: The comparison of testing accuracy rates of the present study and four other similar studies for Valence recognition using EEG modality

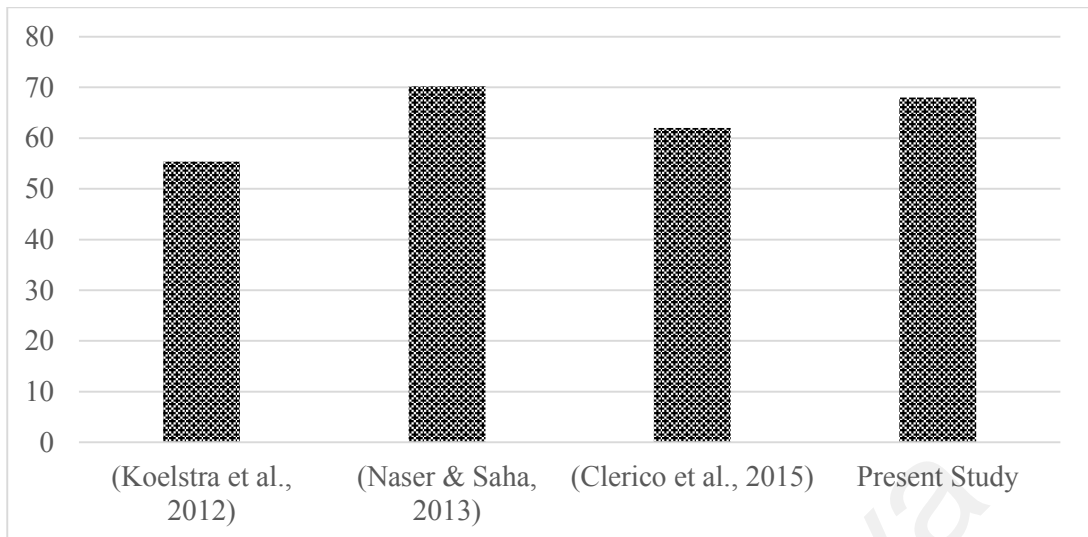


Figure 4.25: The comparison of testing accuracy rates of the present study and four other similar studies for Liking recognition using EEG modality

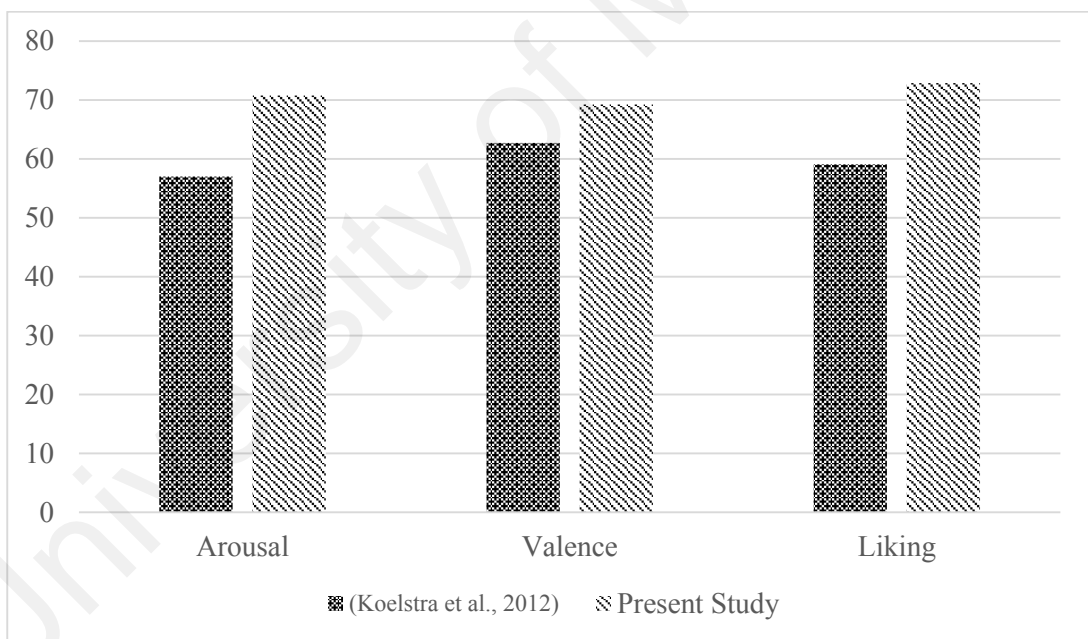


Figure 4.26: The comparison of testing accuracy rates of the present study and a benchmark study for Arousal, Valence and Liking recognition using Peripheral modality

4.7 Conclusion

This chapter reported the results of the proposed method and compare it with other benchmark classification methods. In addition, several analyses were conducted to study the effect of using a specific type of modality on the performance of the emotion recognition system. Furthermore, the benefits of using feature selection methods on the overall classification accuracy of the emotion recognition system were examined. Finally, the statistical test showed that the proposed method is significantly better than the benchmark classifiers considered in this study.

University of Malaya

CHAPTER 5: CONCLUSIONS

The main objective of this research is to propose a classification method to improve the performance accuracy of physiological-based emotion recognition system. This chapter summarizes the overall work carried out by this research. First, the research objectives defined in chapter 1 are revisited. Second, the research contributions are presented. Third, some limitations of this research are discussed and finally, the recommendations for future direction are provided.

5.1 Research objectives revisited

This section revisits the accomplishments of the research objectives defined in this research

5.1.1 Research objective 1

The first objective is to identify most used feature selection and classification methods and its related issues in designing the existing physiological-based emotion recognition systems. This objective is achieved with the analysis of literature in Chapter 2. There are two research questions that need to be answered for this objective. They are:

RQ1: What are the most utilized feature selection and classification methods in the design of the current emotion recognition systems based on the physiological signals?

RQ2: What are the most prominent limitation and challenges of the current emotion recognition systems based on the physiological signals?

The RQ1 is answered in Chapter 2 based on the analysis of findings from Table 2.2 to Table 2.8. There are varieties of feature dimension reduction techniques that have been used in physiological-based emotion recognition systems. Some of the commonly used

methods are wrapper feature selection methods such as SFS, SBS, SFFS. PCA, and Fisher's projection techniques, as well as filter-based feature section methods like feature ranking methods. The most utilized single classifiers used by researchers to develop physiological-based emotion recognition systems are: LDA, CART, ANN, and SVM. In some studies, also the well-known types of ensemble classification methods like bagging and boosting are applied.

For RQ2, as we explained in details in Section 2.7.1.6, the feature dimension reduction methods used in designing emotion recognition system has the following issues. First of all, the limitation of wrapper feature selection methods is being time-consuming and overly specific to the classifier used. The main limitations of PCA and Fisher's projection techniques are, including, (1) they do not guarantee to offer superior correlation with emotional states of the subjects than primary features and (2) the new set of features do not have physical meaning, which results in the lack of the system's interpretation. Feature ranking techniques as feature selection methods, are fast, but the use of only one feature ranking method, as it is commonly used, may result in a sub-optimal solution. One suggested solution is to use more than one feature ranking method to increase the chance to choose the optimal feature set. Ensemble classification methods which have encouraging results in other fields, have been underutilized in physiological-based emotion recognition systems despite. The main advantage of these techniques is their ability to achieve better results than benchmark single classifiers.

5.1.2 Research objective 2

The second objective is to design and develop a feature-based dual-layer ensemble classification method to improve the accuracy rate of physiological-based emotion recognition system. The related research question designed for this objective is:

RQ3: How can we design an improved classification method to enhance the classification accuracy rate of the existing emotion recognition systems?

Based on the limitation identified through the literature review of the existing emotion recognition systems based on physiological signals (objective #1), we found that proposing a classification technique based on a combination of feature ranking methods and using them in stacking ensemble classification strategies can enhance classification accuracy of physiological-based emotion classification system. Utilizing feature selection methods for ensembles has demonstrated to be a helpful strategy for ensemble development, because of its ability to provide more robust and diverse feature subsets that eventually increases the classification accuracy of the ensemble method. The feature ranking methods used in our proposed methods all are filter-based feature selection methods. The advantage of this types of feature selection methods is fast and classifier-independent. Thus, they impose less computational cost to the system, which makes them more suitable to be utilized in the proposed method. Additionally, the feature ranking selected are among the best performing feature ranking methods that have proved their performance in another field of studies (Brown et al., 2012). The single classifiers chosen to be used in our proposed method are LDA, ANN, SVM, and CART, which are among the well-known single classifiers that have been employed in most of the physiological-based emotion recognition systems.

5.1.3 Research objective 3

The third objective is to evaluate the classification accuracy of the proposed dual-layer ensemble classification method by comparing with the benchmark classification methods using statistical analysis. To evaluate achievement of this objective, four research questions are defined:

RQ4: How the classification accuracy rate of emotion recognition systems is affected by using different data modalities?

RQ5: How the classification accuracy rate of emotion recognition systems is affected by using benchmark classification methods combined with feature selection methods as compared to the same classification methods without the feature selection methods?

RQ6: Will the proposed classification have better classification accuracy as compared to other classification methods?

RQ7: Can the proposed classification method achieve significant improvement over the other methods? How can we prove that statistically?

The answers for RQ4 are summarized in the following points:

- There is no single type of modality that is suitable for all the classifiers.
- SVM achieved the best results when it uses Peripheral modality while EEG is more suitable for ANN. The combination of the two modalities (Peripheral + EEG) enables LDA and CART to achieve their best accuracies.
- Combining the two modalities, namely, peripheral and EEG, is better to employ if the objective is to recognize valence, while arousal and liking are best recognized by using peripheral and EEG modalities, respectively.
- SVM seems to be the best classifier to use for Peripheral modality. For EEG, the best classifier is shared by SVM and ANN while LDA is the suitable classifier to be used when EEG and Peripheral modalities are combined.

The answers for RQ5 are summarized in the following points:

- For single classifiers, SVM achieved the best classification method to be used for developing emotion recognition system. Another alternative to SVM can

be LDA and ANN, which achieved comparable results in most cases. CART is generally the worst among the four classifiers.

- While SVM performed well on liking and arousal, it is the worst classifier for recognizing the valence for all modalities. Despite being a stable and powerful classifier, SVM could not be consistent in all the emotional states recognition. This may reveal one drawback of using single classifier for automatic human emotion recognition system, which is the consistency
- There is no single feature selection method which is suitable for all the feature-based multi-classifier methods and all modalities.
- There are some feature selection methods, which frequently achieved the best results such as Fisher, Relief, and Condred.
- Fisher method worked relatively well with ANN (6 out of 9 cases) and somewhat with SVM (4 out 9 cases). In addition, Icap and Relief can be suitable feature selection candidates to be used with CART.
- Feature selection methods have positively contributed to the improvement of multi-classifier methods compared to single classifiers.
- ANNs which is benefited the most after applying feature selection method gaining between 4.45% and 7.79% followed by CART between 1.54% to 3.41%. SVM gained the least among the classifiers between 0.83% to 0.91% behind LDA which gained between 1.25% and 1.98%.
- The result confirms previous findings, which stated that CART and ANNs are unstable classifiers, and thus are suitable for ensemble methods.
- Feature selection methods have a positive impact on both the accuracy and computational cost.

The answers for RQ6 are summarized in the following points:

- The best feature-based dual-layer ensemble classification (FDLEC) methods are based on SVM+CART, followed by SVM+SVM, and then CART +CART.
- FDLEC method based on SVM+CART method dominates the results by achieving the best accuracy rate in 7 data sets out of 9 while FDLEC (CART +CART) achieved the best results in 2 data sets out of 9.
- FDLEC (SVM+CART) method is the best method among all the compared methods.
- Compared to the best single classifiers, the proposed feature-based dual-layer ensemble classification methods have improved the accuracy between 5.62% and 17.36%.

The answer for RQ7:

- Based on the Wilcoxon signed ranks test, FDLEC (SVM+CART) is significantly better than all the selected classifiers.

5.2 Research Contribution

The current study contributes to the fields of emotion recognition by proposing an improved classification method that can be used for enhancing the quality of emotion recognition system based on physiological signals. The main contributions of this study can be summarized in the following points:

- Analyse the accuracies of various classification methods with different modalities and feature selection methods in order to understand the effect of each component on the overall performance of the emotion recognition system.
- Analyse the performance of the emotion recognition system using different single and feature-based multi-classifier methods.

- Propose a design for a classification method called feature-based dual-layer ensemble classification (FDLEC) method that combines 10 different feature ranking techniques and used dual-layer ensemble classification model to improve the classification prediction.

5.3 Research Limitation

- *Difficulties in comparison of classification accuracies between existing studies and this research:*

The comparison between accuracy results obtained by our proposed classification method and the results achieved by other researchers in the literature is somewhat challenging. Physiological data can be influenced by many variables such as emotion stimuli, subject physical exertion, environment temperature range and inter-subject variations in physiology. In addition, comparing classification accuracies using the same physiological data set can be valid only if the comparison used the same experimental setups, which includes, for example applying the same cross-validation technique for estimating the classification accuracy rate, and the same feature selection methods for feature reduction as well as having the same extracted features' values are expected.

- *Human error in feature extraction process:*

Normally, in the existing physiological data sets for emotion recognition, the creators of data sets either do not provide exact name of extracted features or, do not provide its feature extraction code or calculated values. Thus, this requires each researcher to develop their own code for feature extraction from physiological data. Although this research has considered all the necessary

steps to calculate accurate values for different features, it is not guaranteed that human errors have been eradicated.

5.4 Suggestions for future works

Here are some possible ways to extend and improve this work:

- The proposed classification method can be used or tested on other emotion databases or even on other medical diagnosis problems that use physiological data. In fact, there are some other physiological data sets for emotion recognition like MAHNOB or DECAF that the proposed feature-based dual-layer ensemble classification method can be tested on them to study and compare the emotion recognition accuracy performance. There are also some medical applications that employ subject physiological data to detect level of physiological impairment, so the proposed method also can be applied to detect different level of impairment.
- Other feature ranking methods reviewed in the literature can be used in addition to the 10 feature ranking methods selected in this study. The new method of feature ranking can be added to the existing set of feature ranking methods, so it will be 11 feature ranking methods in the proposed feature-based dual layer ensemble classification method or a new method can be replaced with one or more existing feature ranking methods in our proposed method to test their efficiency in maximizing the accuracy of physiological-based emotion recognition system.
- The proposed classification method can be tested on other pattern recognition problems to evaluate its efficiency and possible use. For this purpose, standard data sets from other field of research can be selected and test the proposed classification method on those data sets to study the efficiency of our proposed method using other kind of data except physiological data.

REFERENCES

- Abadi, M. K., Subramanian, R., Kia, S. M., Avesani, P., Patras, I., & Sebe, N. (2015). DECAF: MEG-Based multimodal database for decoding affective physiological responses. *IEEE Transactions on Affective Computing*, 6(3), 209–222.
- Aghaei Pour, P., Hussain, M. S., AlZoubi, O., D’Mello, S., & Calvo, R. (2010). The impact of system feedback on learners’ affective and physiological states. In *Intelligent Tutoring Systems* (Vol. 6094, pp. 264–273). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-13388-6_31
- Allanson, J., & Fairclough, S. H. (2004). A research agenda for physiological computing. *Interacting with Computers*, 16(5), 857–878.
- Alpers, G. W., Wilhelm, F. H., & Roth, W. T. (2005). Psychophysiological assessment during exposure in driving phobic patients. *Journal of Abnormal Psychology*, 114(1), 126–39. <https://doi.org/10.1037/0021-843X.114.1.126>
- AlZoubi, O., Fossati, D., D’Mello, S., & Calvo, R. A. (2014). Affect detection from non-stationary physiological data using ensemble classifiers. *Evolving Systems*, 6(2), 79–92. <https://doi.org/10.1007/s12530-014-9123-z>
- Amaratunga, D., Cabrera, J., & Shkedy, Z. (2014). *Exploration and analysis of DNA microarray and other high-dimensional data*.
- Andreassi, J. L. (2007). *Psychophysiology: human behavior and physiological response*. Lawrence Erlbaum.
- Arroyo-palacios, J., & Romano, D. M. (2008). DM: towards a standardization in the use of physiological signals for affective recognition systems. In *Proceedings of Measuring Behavior*. Maastricht, The Netherlands.
- Arroyo-Palacios, J., & Romano, D. M. (2010). Bio-Affective computer interface for game interaction. *International Journal of Gaming and Computer-Mediated Simulations*, 2(4), 16–32. <https://doi.org/10.4018/jgcms.2010100102>
- Bailenson, J. N., Pontikakis, E. D., Mauss, I. B., Gross, J. J., Jabon, M. E., Hutcherson, C., ... John, O. (2008). Real-time classification of evoked emotions using facial feature tracking and physiological responses. *International Journal of Human-Computer Studies*, 66(5), 303–317. <https://doi.org/10.1016/j.ijhcs.2007.10.011>
- Barreto, A., Zhai, J., & Adjouadi, M. (2007). Non-intrusive physiological monitoring for automated stress detection in human-computer interaction. *Human-Computer Interaction*, 4796, 29–38. Retrieved from http://dx.doi.org/10.1007/978-3-540-75773-3_4
- Barrett, L. F. (1998). Discrete emotions or dimensions? The role of valence focus and arousal focus. *Cognition & Emotion*, 12(4), 579–599.
- Begeer, S., Koot, H. M., Rieffe, C., Meerum Terwogt, M., & Stegge, H. (2008, September). Emotional competence in children with autism: Diagnostic criteria and

empirical evidence. *Developmental Review*.

- Bekele, E., Wade, J., Bian, D., Fan, J., Swanson, A., Warren, Z., & Sarkar, N. (2016). Multimodal adaptive social interaction in virtual environment (MASI-VR) for children with Autism spectrum disorders (ASD). In *IEEE Virtual Reality (VR)* (pp. 121–130). IEEE. <https://doi.org/10.1109/VR.2016.7504695>
- Bhatnagar, V., Bhardwaj, M., Sharma, S., & Haroon, S. (2014). Accuracy–diversity based pruning of classifier ensembles. *Progress in Artificial Intelligence*, 2(2–3), 97–111. <https://doi.org/10.1007/s13748-014-0042-9>
- Bonarini, A., Mainardi, L., Matteucci, M., Tognetti, S., & Colombo, R. (2008). Stress recognition in a robotic rehabilitation task. In *Proceedings of the ACM/IEEE Human-Robot Interaction Conference (HRI08)* (pp. 41–48).
- Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1), 49–59. [https://doi.org/10.1016/0005-7916\(94\)90063-9](https://doi.org/10.1016/0005-7916(94)90063-9)
- Bradley, M. M., Lang, P. J., Cacioppo, J. T., Tassinary, L. G., & Berntson, G. G. (2007). *Emotion and motivation. Handbook of psychophysiology (3rd ed.)*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511546396>
- Brazdil, P. B., & Soares, C. (2000). A comparison of ranking methods for classification algorithm selection (pp. 63–75). Springer Berlin Heidelberg.
- Breiman, L. (1996). Bagging Predictors. *Machine Learning*, 24(421), 123–140.
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees. *The Wadsworth Statisticsprobability Series*, 19.
- Broek, E. L. van den, Lisy, V., Janssen, J. H., Westerink, J. H., Schut, M. H., & Tuinenbreijer, K. (2010). Affective Man-Machine Interface: Unveiling human emotions through biosignals. In *Biomedical Engineering Systems and Technologies* (pp. 21–47). Springer Verlag.
- Brown, G., Pocock, A., Zhao, M.-J., Luján, M., Brown, G., Pocock, A., ... Luján, M. (2012). Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *The Journal of Machine Learning Research*, 13(1), 27–66.
- Cacioppo, J. (2004). Feelings and emotions: roles for electrophysiological markers. *Biological Psychology*, 67(1–2), 235–43.
- Cacioppo, J., Tassinary, L., & Berntson, G. (2007). *The Handbook of Psychophysiology. Dreaming* (Vol. 44). <https://doi.org/10.1017/CBO9780511546396>
- Calvo, R. A., Brown, I., & Scheduling, S. (2009). Effect of experimental factors on the recognition of affective mental states through physiological measures. In *Lecture*

Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 5866, pp. 62–70). Berlin, Heidelberg: Springer Berlin Heidelberg.

- Calvo, R. A., & D’Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1), 18–37. <https://doi.org/10.1109/T-AFFC.2010.1>
- Campisi, P., & La Rocca, D. (2014). Brain waves for automatic biometric-based user recognition. *IEEE Transactions on Information Forensics and Security*, 9(5), 782–800. <https://doi.org/10.1109/TIFS.2014.2308640>
- Cárdenas-Gallo, I., Sarmiento, C. A., Morales, G. A., Bolivar, M. A., & Akhavan-Tabatabaei, R. (2017). An ensemble classifier to predict track geometry degradation. *Reliability Engineering & System Safety*, 161, 53–60.
- Carlson, N. R. (2012). *Physiology of behavior*. Prentice Hall.
- Chai, J., Ge, Y., Liu, Y., Li, W., Zhou, L., Yao, L., & Sun, X. (2014). Application of frontal EEG asymmetry to user experience research. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 8532 LNAI, pp. 234–243). Springer Verlag.
- Chanamarn, N., Tamee, K., & Sittidech, P. (2016). Stacking technique for academic achievement prediction. In *International Workshop on Smart Info-Media Systems in Asia (SISA 2016)* (pp. 14–17).
- Chanel, G., Kierkels, J. J. M., Soleymani, M., & Pun, T. (2009). Short-term emotion assessment in a recall paradigm. *International Journal of Human-Computer Studies*, 67(8), 607–627. <https://doi.org/10.1016/j.ijhcs.2009.03.005>
- Chanel, G., Rebetez, C., Bétrancourt, M., & Pun, T. (2011). Emotion assessment from physiological signals for adaptation of game difficulty. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 41(6), 1052–1063. <https://doi.org/10.1109/TSMCA.2011.2116000>
- Chanel, Kronegg, J., Grandjean, D., & Pun, T. (2006). *Emotion assessment: Arousal evaluation using EEG’s and peripheral physiological signals*. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 4105). Berlin, Heidelberg: Springer Berlin Heidelberg. <https://doi.org/10.1007/11848035>
- Chang, C.-Y., Zheng, J.-Y., & Wang, C.-J. (2010). Based on Support Vector Regression for emotion recognition using physiological signals. In *International Joint Conference on Neural Networks (IJCNN)* (pp. 1–7).
- Chen, J., Hu, B., Moore, P., Zhang, X., & Ma, X. (2015). Electroencephalogram-based emotion assessment system using ontology and data mining techniques. *Applied Soft Computing*, 30, 663–674. <https://doi.org/10.1016/j.asoc.2015.01.007>
- Christie, I. C., & Friedman, B. H. (2004). Autonomic specificity of discrete emotion and dimensions of affective space: A multivariate approach. *International Journal of*

Psychophysiology, 51(2), 143–153.

- Clerico, A., Gupta, R., & Falk, T. H. (2015). Mutual information between inter-hemispheric EEG spectro-temporal patterns : A new feature for automated affect recognition. In *7th International IEEE/EMBS Conference on Neural Engineering (NER)* (pp. 22–24). <https://doi.org/10.1109/NER.2015.7146774>
- Coan, J. a., & Allen, J. J. B. (2007). *Handbook of emotion elicitation and assessment. Physiology*.
- Colomer Granero, A., Fuentes-Hurtado, F., Naranjo Ornedo, V., Guixeres Provinciale, J., Ausín, J. M., & Alcañiz Raya, M. (2016). A comparison of physiological signal analysis techniques and classifiers for automatic emotional evaluation of audiovisual contents. *Frontiers in Computational Neuroscience*, 10, 74.
- Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3), 273–297. <https://doi.org/10.1023/A:1022627411411>
- Damasio, A. R. (1994). Descartes' error: Emotion, rationality and the human brain. *New York: Putnam*, 352.
- Darwin, C. (1965). *The Expression of the Emotions in Man and Animals. Penguin classics*. Oxford University Press Inc.
- Davidson, R. J. (1984). Affect, cognition, and hemispheric specialization. In *Emotion, Cognition, and Behavior* (pp. 320–365). Cambridge University Press.
- Davidson, R. J., Pizzagalli, D., Nitschke, J. B., & Kalin, N. H. (2003). Parsing the subcomponents of emotion and disorders of emotion: Perspectives from affective neuroscience. In *Handbook of the Affective Sciences* (pp. 8–24). Oxford University Press.
- Davidson, R., Schwartz, G., Saron, C., Bennett, J., & Goleman, D. (1979). Frontal versus parietal EEG asymmetry during positive and negative affect. *Psychophysiology*, 16, 202–203.
- Dawson, M. E., Schell, A. M., & Filion, D. L. (2007). *The electrodermal system. Handbook of Psychophysiology*.
- Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7, 1–30.
- Depenau, J. (1995, December 1). *Automated Design of Neural Network Architecture for Classification. DAIMI Report Series*. Aarhus University,. <https://doi.org/10.7146/dpb.v24i500.7029>
- Diao, R., Chao, F., Peng, T., Snooke, N., & Shen, Q. (2014). Feature Selection Inspired Classifier Ensemble Reduction. *IEEE Transactions on Cybernetics*, 44(8), 1259–1268. <https://doi.org/10.1109/TCYB.2013.2281820>
- Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., McRorie, M., ...

- Karpouzis, K. (2007). The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data. In *Proceedings of International Conference Affective Computing and Intelligent Interaction* (pp. 488–500).
- Duda, R. O., Hart, P. E. (Peter E.), & Stork, D. G. (2001). *Pattern classification*. Wiley.
- Džeroski, S., & Ženko, B. (2004). Is combining classifiers with stacking better than selecting the best one?. *Machine Learning*, 54(3), 255–273.
- Ekman, P. (1957). A methodological discussion of non-verbal behavior. *Journal of Psychology*, 43, 141–149.
- Ekman, P. (1992). Are there basic emotions?. *Psychological Review*, 99(3), 550–553.
- Ekman, P., & Cordaro, D. (2011). What is meant by calling emotions basic. *Emotion Review*, 3(4), 364–370. <https://doi.org/10.1177/1754073911410740>
- Ekman, P., & Davidson, R. J. (1993). Voluntary smiling changes regional brain activity. *Psychological Science*, 4(5), 342–345.
- Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., ... Ricci-Bitti, P. E. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, 53(4), 712–7.
- Ekman, Davidson, R. J., & Friesen, W. V. (1990). The Duchenne smile: emotional expression and brain physiology. II. *Journal of Personality and Social Psychology*, 58(2), 342–353.
- Eysenck, M. W., & Keane, M. T. (2005). *Cognitive Psychology: A Student's Handbook*. *Cognitive Psychology*.
- Fairclough, S. H. (2009). Fundamentals of physiological computing. *Interacting with Computers*, 21(1–2), 133–145. <https://doi.org/10.1016/j.intcom.2008.10.011>
- Fanelli, G., Gall, J., Romsdorfer, H., Weise, T., & Van Gool, L. (2010). A 3-D audio-visual corpus of affective communication. *IEEE Transactions on Multimedia*, 12(6), 591–598.
- Farquharson, R. F. (1942). The hypothalamus and central levels of autonomic function. *American Journal of Psychiatry*, 98(4), 625–625.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 179–188. <https://doi.org/10.1111/j.1469-1809.1936.tb02137.x>
- Frantzidis, C. A., Bratsas, C., Klados, M. A., Konstantinidis, E., Lithari, C. D., Vivas, A. B., ... Bamidis, P. D. (2010). On the classification of emotional biosignals evoked while viewing affective pictures: an integrated data-mining-based approach for healthcare applications. *IEEE Transactions on Information Technology in Biomedicine*, 14(2), 309–318. <https://doi.org/10.1109/TITB.2009.2038481>

- Frasson, C., & Chalfoun, P. (2010). Managing learner's affective states in intelligent tutoring systems (pp. 339–358). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-14363-2_17
- Ghergulescu, I., & Muntean, C. H. (2014). A novel sensor-based methodology for learner's motivation analysis in game-based learning. *Interacting with Computers*, 26(4), 305–320. <https://doi.org/10.1093/iwc/iwu013>
- Giakoumis, D., Tzovaras, D., & Hassapis, G. (2013). Subject-dependent biosignal features for increased accuracy in psychological stress detection. *International Journal of Human-Computer Studies*, 71(4), 425–439.
- Giakoumis, D., Tzovaras, D., Moustakas, K., & Hassapis, G. (2011). Automatic recognition of boredom in video games using novel biosignal moment-based features. *IEEE Transactions on Affective Computing*, 2(3), 119–133. <https://doi.org/10.1109/T-AFFC.2011.4>
- Grimm, M., Kroschel, K., & Narayanan, S. (2008). The Vera am Mittag German audio-visual emotional speech database. In *2008 IEEE International Conference on Multimedia and Expo* (pp. 865–868).
- Gross, J. J., & Levenson, R. W. (1995). Emotion elicitation using films. *Cognition & Emotion*, 9(1), 87–108. <https://doi.org/10.1080/02699939508408966>
- Gu, Y., Tan, S.-L., Wong, K.-J., Ho, M.-H. R., & Qu, L. (2010). A biometric signature based system for improved emotion recognition using physiological responses from multiple subjects. In *8th IEEE International Conference on Industrial Informatics* (pp. 61–66). IEEE. <https://doi.org/10.1109/INDIN.2010.5549464>
- Gualtieri, J., & Crompton, R. (1998). Support vector machines for hyperspectral remote sensing classification. In *The 27th AIPR Workshop: Advances in Computer Assisted Recognition* (pp. 221–232). Washington,.
- Guan, D., Yuan, W., Lee, Y.-K., Najeebullah, K., & Rasel, M. K. (2014). A review of ensemble learning based feature selection. *IETE Technical Review*, 31(3), 190–198. <https://doi.org/10.1080/02564602.2014.906859>
- Gunes, H., Piccardi, M., & Pantic, M. (2008). From the lab to the real world: Affect recognition using multiple cues and modalities. In *Affective Computing: Focus on Emotion Expression, Synthesis, and Recognition* (pp. 185–218). InTech Education and Publishing.
- Guo, H.-W., Huang, Y.-S., Lin, C.-H., Chien, J.-C., Haraikawa, K., & Shieh, J.-S. (2016). Heart rate variability signal features for emotion recognition by using principal component analysis and support vectors machine. In *IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE)* (pp. 274–277). IEEE. <https://doi.org/10.1109/BIBE.2016.40>
- Haag, A., Goronzy, S., Schaich, P., & Williams, J. (2004). Emotion recognition using bio-sensors: First steps towards an automatic system. In E. André, L. Dybkjær, W. Minker, & P. Heisterkamp (Eds.), *Affective dialogue systems* (Vol. 3068, pp. 36–48). Berlin, Heidelberg: Springer Berlin Heidelberg. <https://doi.org/10.1007/b98229>

- Haarmann, A., Boucsein, W., & Schaefer, F. (2009). Combining electrodermal responses and cardiovascular measures for probing adaptive automation during simulated flight. *Applied Ergonomics*, *40*(6), 1026–1040.
- Hagemann, D., Waldstein, S. R., & Thayer, J. F. (2003). Central and autonomic nervous system integration in emotion. *Brain and Cognition*, *52*(1), 79–87.
- Hariharan, A., & Adam, M. T. P. (2015). Blended emotion detection for decision support. *IEEE Transactions on Human-Machine Systems*, *45*(4), 510–517.
- Hart, S. G., & Staveland, L. E. (1988). *Development of NASA-TLX (Task Load Index): results of empirical and theoretical research*. *Advances in Psychology* (Vol. 52). Elsevier. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- Hawknis, D. I., Mothersbaugh, D. L., & Mookerjee, A. (2011). *Consumer behavior : building marketing strategy*. Tata McGraw Hill.
- He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284.
- Healey, J. A., & Picard, R. W. (2005). Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*, *6*(2), 156–166. <https://doi.org/10.1109/TITS.2005.848368>
- Hecke, V., Vaughan, A., Stevens, S., Carson, A. M., Karst, J. S., Dolan, B., ... Brockman, S. (2015). Measuring the plasticity of social approach: a randomized controlled trial of the effects of the peers intervention on EEG asymmetry in adolescents with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *45*(2), 316–335. <https://doi.org/10.1007/s10803-013-1883-y>
- Heraz, a, Razaki, R., & Frasson, C. (2007). Using machine learning to predict learner emotional state from brainwaves. *Seventh IEEE International Conference on Advanced Learning Technologies 2007*, *0*(Table 1), 853–857.
- Hernandez, J., Paredes, P., Roseway, A., & Czerwinski, M. (2014). Under pressure: Sensing Stress of Computer Users. *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems - CHI '14*, 51–60.
- Horlings, R., Datcu, D., & Rothkrantz, L. J. M. (2008). Emotion recognition using brain activity. *Proceedings of the 9th International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing*, (March), 6. <https://doi.org/10.1145/1500879.1500888>
- Hussain, S., Keung, J., Khan, A. A., & Bennin, K. E. (2015). Performance evaluation of ensemble methods for software fault prediction. In *24th Australasian Software Engineering Conference* (pp. 91–95). New York, New York, USA: ACM Press. <https://doi.org/10.1145/2811681.2811699>
- Izard, C. E. (1992). Basic emotions, relations among emotions, and emotion-cognition relations. *Psychological Review*, *99*(3), 561–5.

- Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, 31(3), 264–323.
- Jain, L. C., & Lazzerini, B. (1999). *Knowledge-based intelligent techniques in character recognition*. CRC Press.
- James, W. (1884). What is an Emotion?. *Mind*, 9(34), 188–205. <https://doi.org/10.1093/mind/LI.202.200>
- Jang, E.-H., Park, B.-J., Park, M.-S., Kim, S.-H., & Sohn, J.-H. (2015). Analysis of physiological signals for recognition of boredom, pain, and surprise emotions. *Journal of Physiological Anthropology*, 34, 25. <https://doi.org/10.1186/s40101-015-0063-5>
- Jenke, R., Peer, A., & Buss, M. (2014). Feature extraction and selection for emotion recognition from EEG. *IEEE Transactions on Affective Computing*, 5(3), 327–339. <https://doi.org/10.1109/TAFFC.2014.2339834>
- Jensen, R., & Qiang Shen. (2009). New Approaches to Fuzzy-Rough Feature Selection. *IEEE Transactions on Fuzzy Systems*, 17(4), 824–838. <https://doi.org/10.1109/TFUZZ.2008.924209>
- Jolliffe, I. T. (1986). *Principal component analysis*. Springer-Verlag, New York .
- Jones, R. M., Buhr, A. P., Conture, E. G., Tumanova, V., Walden, T. A., & Porges, S. W. (2014). Autonomic nervous system activity of preschool-age children who stutter. *Journal of Fluency Disorders*, 41(C), 12–31.
- Jones, R. M., Conture, E. G., & Walden, T. A. (2014). Emotional reactivity and regulation associated with fluent and stuttered utterances of preschool-age children who stutter. *Journal of Communication Disorders*, 48(1), 38–51.
- Kapoor, A., Burleson, W., & Picard, R. W. (2007). Automatic prediction of frustration. *International Journal of Human-Computer Studies*, 65(8), 724–736. <https://doi.org/10.1016/j.ijhcs.2007.02.003>
- Katsis, C. D., Ganiatsas, G., & Fotiadis, D. I. (2006). An integrated telemedicine platform for the assessment of affective physiological states. *Diagnostic Pathology*, 1, 16. <https://doi.org/10.1186/1746-1596-1-16>
- Katsis, Katertsidis, N., Ganiatsas, G., & Fotiadis, D. . (2008). Toward emotion recognition in car-racing drivers: A biosignal processing approach. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 38(3), 502–512. <https://doi.org/10.1109/TSMCA.2008.918624>
- Khalili, Z., & Moradi, M. H. (2008). Emotion detection using brain and peripheral signals. In *2008 Cairo International Biomedical Engineering Conference* (pp. 1–4). IEEE. <https://doi.org/10.1109/CIBEC.2008.4786096>
- Khezri, M., Firoozabadi, M., & Sharafat, A. R. (2015). Reliable emotion recognition system based on dynamic adaptive fusion of forehead biopotentials and

physiological signals. *Computer Methods and Programs in Biomedicine*, 122(2), 149–164.

- Khosrowabadi, R., Quek, H. C., Wahab, A., & Ang, K. K. (2010). EEG-based emotion recognition using self-organizing map for boundary detection. In *20th International Conference on Pattern Recognition* (pp. 4242–4245).
- Kim, D., Frank, M. G., & Kim, S. T. (2014). Emotional display behavior in different forms of Computer Mediated Communication. *Computers in Human Behavior*, 30, 222–229.
- Kim, E. S., Daniell, C. M., Makar, C., Elia, J., Scassellati, B., & Shic, F. (2015). Potential clinical impact of positive affect in robot interactions for autism intervention. In *International Conference on Affective Computing and Intelligent Interaction* (pp. 8–13). Institute of Electrical and Electronics Engineers Inc.
- Kim, J. (2007). Bimodal emotion recognition using speech and physiological changes. In *Robust Speech Recognition and Understanding* (pp. 265–280).
- Kim, J., & André, E. (2008). Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(12), 2067–83. <https://doi.org/10.1109/TPAMI.2008.26>
- Kim, J., Bee, N., Wagner, J., & André, E. (2004). Emote to win: Affective interactions with a computer game agent. *Lecture Notes in Informatics (LNI)*, P-50, 159–164.
- Kim, K. H., Bang, S. W., & Kim, S. R. (2004). Emotion recognition system using short-term monitoring of physiological signals. *Medical & Biological Engineering & Computing*, 42(3), 419–27.
- Kim, & Andre, E. (2006). Emotion recognition using physiological and speech signal in short-term observation. In *Perception and Interactive Technologies* (pp. 53–64).
- Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In *Proceedings of the ninth international workshop on Machine learning* (pp. 249–256). [https://doi.org/10.1016/S0031-3203\(01\)00046-2](https://doi.org/10.1016/S0031-3203(01)00046-2)
- Kittler, J. (1978). Feature set search algorithms. In C. H. Chen (Ed.), *Pattern recognition and signal processing* (pp. 41–60). Sijthof and Noordhoff, The Netherlands.
- Koelstra, S., Muhl, C., Soleymani, M., Yazdani, A., Ebrahimi, T., Pun, T., ... Patras, I. (2012). DEAP: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3(1), 18–31. <https://doi.org/10.1109/T-AFFC.2011.15>
- Koenig, A., Novak, D., Omlin, X., Pulfer, M., Perreault, E., Zimmerli, L., ... Riener, R. (2011). Real-time closed-loop control of cognitive load in neurological patients during robot-assisted gait training. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 19(4), 453–64.
- Kohonen, T. (1982). Self-organized formation of topologically correct feature maps.

- Kolodyazhnyi, V., Kreibig, S. D., Gross, J. J., Roth, W. T., & Wilhelm, F. H. (2011). An affective computing approach to physiological emotion specificity: toward subject-independent and stimulus-independent classification of film-induced emotions. *Psychophysiology*, 48(7), 908–22. <https://doi.org/10.1111/j.1469-8986.2010.01170.x>
- Kotsia, I., Patras, I., & Fotopoulos, S. (2012). Affective gaming: Beyond using sensors. In *5th International Symposium on Communications, Control and Signal Processing* (pp. 1–4). IEEE. <https://doi.org/10.1109/ISCCSP.2012.6217768>
- Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: a review. *Biological Psychology*, 84(3), 394–421.
- Kreibig, S. D., Wilhelm, F. H., Roth, W. T., & Gross, J. J. (2007). Cardiovascular, electrodermal, and respiratory response patterns to fear- and sadness-inducing films. *Psychophysiology*, 44(5), 787–806.
- Krumhansl, C. L. (1997). An exploratory study of musical emotions and psychophysiology. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, 51(4), 336–353.
- Kukolja, D., Popović, S., Horvat, M., Kovač, B., & Ćosić, K. (2014). Comparative analysis of emotion estimation methods based on physiological measurements for real-time applications. *International Journal of Human-Computer Studies*, 72(10), 717–727. <https://doi.org/10.1016/j.ijhcs.2014.05.006>
- Kulic, D., & Croft, E. A. (2007). Affective state estimation for human–robot interaction. *IEEE Transactions on Robotics*, 23(5), 991–1000.
- Kuncheva. (2007). A stability index for feature selection. In *International Multi-conference: artificial intelligence and applications* (pp. 390–395). IASTED.
- Kuncheva. (2014). *Combining pattern classifiers: methods and algorithms*. Wiley Blackwell.
- Kuncheva, L. I., Christy, T., Pierce, I., & Mansoor, S. P. (2011). Multi-modal biometric emotion recognition using classifier ensembles (pp. 317–326). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-21822-4_32
- Lam, L., & Suen, S. Y. (1997). Application of majority voting to pattern recognition: an analysis of its behavior and performance. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 27(5), 553–568.
- Lang, P. J. (1995). The emotion probe: Studies of motivation and attention. *American Psychologist*, 50(5), 372–385. <https://doi.org/10.1037/0003-066X.50.5.372>
- Lang, P. J., & Bradley, M. M. (2010). Emotion and the motivational brain. *Biological Psychology*, 84(3), 437–450.

- Larsen, R. J., & Fredrickson, B. L. (1999). Measurement Issues in Emotion Research. In *Response* (Vol. 14, pp. 40–60). Russell Sage Foundation.
- Lazarus, R. S. (1982). Thoughts on the relations between emotion and cognition. *American Psychologist*, *37*(9), 1019–1024.
- Lee, I.-H., Lushington, G. H., & Visvanathan, M. (2011). A filter-based feature selection approach for identifying potential biomarkers for lung cancer. *Journal of Clinical Bioinformatics*, *1*(1), 11. <https://doi.org/10.1186/2043-9113-1-11>
- Lee, Y.-H., Chen, S. C.-J., Shiah, Y.-J., Wang, S.-F., Young, M.-S., Hsu, C.-Y., ... Lin, C.-L. (2014). Support-vector-machine-based meditation experience evaluation using electroencephalography signals. *Medical and Biological Engineering*, *34*(6), 589–597. <https://doi.org/10.5405/JMBE.1776>
- Lee, Y., & Hsieh, S. (2014). Classifying different emotional states by means of EEG-based functional connectivity patterns. *PloS One*, *9*(4), e95415.
- Lee, Shackman, A. J., Jackson, D. C., & Davidson, R. J. (2009). Test-retest reliability of voluntary emotion regulation. *Psychophysiology*, *46*(4), 874–879.
- Leon, E., Clarke, G., Callaghan, V., & Sepulveda, F. (2004). Real-time detection of emotional changes for inhabited environments. *Computers & Graphics*, *28*(5), 635–642. <https://doi.org/10.1016/j.cag.2004.06.002>
- Levenson, R. (1992). Autonomic nervous system differences among emotions. *Psychological Science*, *3*(1), 23–27.
- Levenson, R., Ekman, P., & Friesen, W. (1990). Voluntary facial action generates emotion-specific autonomic nervous system activity. *Psychophysiology*, *27*(4), 363–84.
- Levillain, F., Orero, J. O., Rifqi, M., & Bouchon-Meunier, B. (2010). Characterizing player's experience from physiological signals using fuzzy decision trees. In *Proceedings of the 2010 IEEE Conference on Computational Intelligence and Games* (pp. 75–82). IEEE. <https://doi.org/10.1109/ITW.2010.5593370>
- Lewis, R. J., Ph, D., & Street, W. C. (2000). An Introduction to Classification and Regression Tree (CART) Analysis. In *Annual Meeting of the Society for Academic Emergency Medicine*, (310), 14. <https://doi.org/10.1.1.95.4103>
- Lichtenstein, A., Oehme, A., Kupschick, S., & Jürgensohn, T. (2008). Comparing two emotion models for deriving affective states from physiological data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 4868 LNCS, pp. 35–50).
- Lisetti, C. L., & Nasoz, F. (2004). Using noninvasive wearable computers to recognize human emotions from physiological signals. *EURASIP Journal on Advances in Signal Processing*, *2004*(11), 1672–1687.
- Lisetti, C., Nasoz, F., LeRouge, C., Ozyer, O., & Alvarez, K. (2003). Developing

- multimodal intelligent affective interfaces for tele-home health care. *International Journal of Human-Computer Studies*, 59(1–2), 245–255.
- Lisetti, & Nasoz, F. (2002). MAUI: a Multimodal Affective User Interface. *Proceedings of the Tenth ACM International Conference on Multimedia*, 161–170.
- Liu, C., Agrawal, P., Sarkar, N., & Chen, S. (2009). Dynamic difficulty adjustment in computer games through real-time anxiety-based affective feedback. *International Journal of Human-Computer Interaction*, 25(6), 506–529.
- Liu, C., Conn, K., Sarkar, N., & Stone, W. (2008a). Online affect detection and robot behavior adaptation for intervention of children with autism. *IEEE Transactions on Robotics*, 24(4), 883–896.
- Liu, C., Conn, K., Sarkar, N., & Stone, W. (2008b). Physiology-based affect recognition for computer-assisted intervention of children with Autism Spectrum Disorder. *International Journal of Human-Computer Studies*, 66(9), 662–677. <https://doi.org/10.1016/j.ijhsc.2008.04.003>
- Liu, H., & Motoda, H. (2008). Computational methods of feature selection. *Computer*, 198(1), 2–13. <https://doi.org/10.1016/j.cma.2008.05.004>
- Liu, Y., & Sourina, O. (2012). EEG-based valence level recognition for real-time applications. In *Proceedings of the 2012 International Conference on Cyberworlds, Cyberworlds 2012* (pp. 53–60). IEEE. <https://doi.org/10.1109/CW.2012.15>
- Lorig, T. S., Cacioppo, J. T., Tassinary, L. G., & Berntson, G. G. (2007). The respiratory system. *Handbook of Psychophysiology (3rd Ed.)*, 231–244. Retrieved from <http://www.mendeley.com/research/respiratory-system-306/>
- Maaoui, C., & Pruski, A. (2010). Emotion recognition through physiological signals for human-machine communication. In *Cutting Edge Robotics 2010* (pp. 317–333). InTech. <https://doi.org/10.5772/10312>
- Malekzadeh, M., Mustafa, M. B., & Lahsasna, A. (2015). A review of emotion regulation in intelligent tutoring systems. *Educational Technology & Society*, 18(4), 435–445.
- Martinez, H. P., Bengio, Y., & Yannakakis, G. N. (2013). Learning deep physiological models of affect. *IEEE Computational Intelligence Magazine*, 8(2), 20–33. <https://doi.org/10.1109/MCI.2013.2247823>
- Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition and Emotion*, 23(2), 209–237.
- McKeown, G., Valstar, M., Cowie, R., Pantic, M., & Schröder, M. (2012). The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1), 5–17.
- McLachlan, G. J. (2004). Discriminant analysis and statistical pattern recognition. *Wiley Series in Probability and Statistics*.

- Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament. *Current Psychology*, 14(4), 261–292.
- Melgani, F., & Bruzzone, L. (2004). Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, 42(8), 1778–1790. <https://doi.org/10.1109/TGRS.2004.831865>
- Miguel, P., & Guerreiro, C. (2008). *Linear discriminant analysis algorithms*. Lisbon.
- Mohammad, Y., & Nishida, T. (2010). Using physiological signals to detect natural interactive behavior. *Applied Intelligence*, 33(1), 79–92.
- Muaremi, A., Seiter, J., Gravenhorst, F., Bexheti, A., Arnrich, B., & Troester, G. (2013). Monitor Pilgrims: Prayer Activity Recognition using Wearable Sensors. In *Proceedings of the 8th International Conference on Body Area Networks* (pp. 161–164). ACM. <https://doi.org/10.4108/icst.bodynets.2013.253685>
- Muller, M. E. (2006). Why some emotional states are easier to be recognized than others: A thorough data analysis and a very accurate rough set classifier. In *IEEE International Conference on Systems, Man and Cybernetics* (Vol. 2, pp. 1624–1629). IEEE. <https://doi.org/10.1109/ICSMC.2006.384951>
- Naser, D. S., & Saha, G. (2013). Recognition of emotions induced by music videos using DT-CWPT. In *2013 Indian Conference on Medical Informatics and Telemedicine (ICMIT)* (pp. 53–57). IEEE. <https://doi.org/10.1109/IndianCMIT.2013.6529408>
- Nasoz, F., Alvarez, K., Lisetti, C. L., & Finkelstein, N. (2004). Emotion recognition from physiological signals using wireless sensors for presence technologies. *Cognition, Technology & Work*, 6(1), 4–14. <https://doi.org/10.1007/s10111-003-0143-x>
- Nasoz, F., Lisetti, C. L., & Vasilakos, A. V. (2010). Affectively intelligent and adaptive car interfaces. *Information Sciences*, 180(20), 3817–3836.
- Neave, H. R., & Worthington, P. L. (1992). *Distribution-Free Tests*. Routledge.
- Novak, D., Mihelj, M., & Munih, M. (2012). A survey of methods for data fusion and system adaptation using autonomic nervous system responses in physiological computing. *Interacting with Computers*, 24(3), 154–172.
- Novak, D., Mihelj, M., Zihlerl, J., Olenšek, A., & Munih, M. (2011). Psychophysiological measurements in a biocooperative feedback loop for upper extremity rehabilitation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 19(4), 400–410. <https://doi.org/10.1109/TNSRE.2011.2160357>
- Novak, D., Zihlerl, J., Olenšek, A., Milavec, M., Podobnik, J., Mihelj, M., & Munih, M. (2010). Psychophysiological responses to robotic rehabilitation tasks in stroke. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 18(4), 351–361.
- Oliveira, L. S., Morita, M., & Sabourin, R. (2006). Feature selection for ensembles applied to handwriting recognition. *International Journal of Document Analysis and*

Recognition (IJ DAR), 8(4), 262–279. <https://doi.org/10.1007/s10032-005-0013-6>

- Opitz, & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Artificial Intelligence Research*, 11, 169–198.
- Optiz, D. W. (1999). Feature selection for ensembles. In *Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence* (p. 998). Orlando, Florida, USA: AAAI Press.
- Pal, A., Gautam, A. K., & Singh, Y. N. (2015). Evaluation of bioelectric signals for human recognition. *Procedia Computer Science*, 48, 746–752.
- Pantic, M., Valstar, M., Rademaker, R., & Maat, L. (2005). Web-based database for facial expression analysis. In *IEEE International Conference on Multimedia and Expo, ICME 2005* (Vol. 2005, pp. 317–321).
- Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, 24(3), 45–77. <https://doi.org/10.2753/MIS0742-1222240302>
- Peter, C., Ebert, E., & Beikirch, H. (2009). Physiological sensing for affective computing. In *Affective Information Processing* (pp. 293–310). Springer London.
- Philippot, P., Chapelle, G., & Blairy, S. (2002). Respiratory feedback in the generation of emotion. *Cognition & Emotion*, 16(5), 605–627.
- Picard, R. (1995). *Affective computing*. Cambridge: MIT press.
- Picard, R. (2009). Future affective technology for autism and emotion communication. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 364(1535), 3575–84. <https://doi.org/10.1098/rstb.2009.0143>
- Picard, Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10), 1175–1191. <https://doi.org/10.1109/34.954607>
- Pizzagalli, D. a. (2007). Electroencephalography and High-Density Electrophysiological Source Localization. *Handbook of Psychophysiology*, 56–84.
- Plarre, K., Raij, A., Hossain, S. M., Ali, A. A., Nakajima, M., Al'absi, M., ... L. E. Wittmers. (2011). Continuous inference of psychological stress from sensory measurements collected in the natural environment. In *10th International Conference on Information Processing in Sensor Networks* (pp. 97–108).
- Polikar, R. (2006). Pattern Recognition. In *Wiley Encyclopedia of Biomedical Engineering*. Hoboken, NJ, USA: John Wiley & Sons, Inc. <https://doi.org/10.1002/9780471740360.ebs0904>
- Pour, P. A., Hussain, M. S., AlZoubi, O., D'Mello, S., & Calvo, R. A. (2010). *Intelligent*

Tutoring Systems. (V. Alevan, J. Kay, & J. Mostow, Eds.), *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 6094). Berlin, Heidelberg: Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-642-13388-6>

- Rainville, P., Bechara, A., Naqvi, N., & Damasio, A. R. (2006). Basic emotions are associated with distinct patterns of cardiorespiratory activity. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*, *61*(1), 5–18. <https://doi.org/10.1016/j.ijpsycho.2005.10.024>
- Rani, P., Liu, C., Sarkar, N., & Vanman, E. (2006). An empirical study of machine learning techniques for affect recognition in human--robot interaction. *Pattern Analysis and Applications*, *9*(1), 58–69. <https://doi.org/10.1007/s10044-006-0025-y>
- Rani, P., & Sarkar, N. (2005). An empirical study of machine learning techniques for affect recognition in human-robot interaction. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 2662–2667). IEEE. <https://doi.org/10.1109/IROS.2005.1545344>
- Rani, P., Sarkar, N., Smith, C. A., & Kirby, L. D. (2004). Anxiety detecting robotic system - towards implicit human-robot collaboration. *Robotica*, *22*(1), 85–95.
- Regan, L. M., & Atkins, M. S. (2007). A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. *International Journal of Human Computer Studies*, *65*, 329–347.
- Rigas, G., Goletsis, Y., Bougia, P., & Fotiadis, D. I. (2011). Towards driver's state recognition on real driving conditions. *International Journal of Vehicular Technology*, *2011*, 1–14. <https://doi.org/10.1155/2011/617210>
- Rigas, G., Katsis, C. D., Ganiatsas, G., & Fotiadis, D. I. (2007). A user independent, biosignal based, emotion recognition method. In C. Conati, K. McCoy, & G. Paliouras (Eds.), *UM '07 Proceedings of the 11th international conference on User Modeling* (Vol. 4511, pp. 314–318). Berlin, Heidelberg: Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-540-73078-1>
- Ringeval, F., Sonderegger, A., Sauer, J., & Lalanne, D. (2013). Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013*.
- Robnik-Šikonja, M. (2004). Improving random forests (pp. 359–370). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-30115-8_34
- Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*(1–2), 1–39. <https://doi.org/10.1007/s10462-009-9124-7>
- Russell & Barrett, 1999, P. 80. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178.
- Russell, & Barrett, L. (1999). Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of Personality and Social*

Psychology, 76(5), 805–819.

- Saeyns, Y., Abeel, T., & Van de Peer, Y. (2008). Robust feature selection using ensemble feature selection techniques. In *Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases - Part II* (pp. 313–325). Springer-Verlag. https://doi.org/10.1007/978-3-540-87481-2_21
- Sakr, G. E., Elhajj, I. H., & Huijjer, H. A.-S. (2010). Support vector machines to define and detect agitation transition. *IEEE Transactions on Affective Computing*, 1(2), 98–108. <https://doi.org/10.1109/T-AFFC.2010.2>
- Santana, L. E. A. dos S., & Canuto, A. M. de P. (2014). Filter-based optimization techniques for selection of feature subsets in ensemble systems. *Expert Systems with Applications*, 41(4), 1622–1631. <https://doi.org/10.1016/j.eswa.2013.08.059>
- Scheirer, J., Fernandez, R., Klein, J., & Picard, R. W. (2002). Frustrating the user on purpose: a step toward building an affective computer. *Interacting with Computers*, 14(2), 93–118. [https://doi.org/10.1016/S0953-5438\(01\)00059-5](https://doi.org/10.1016/S0953-5438(01)00059-5)
- Scherer, K. R. (1999). Appraisal theory. *Handbook of Cognition and Emotion*. <https://doi.org/10.1002/0470013494.ch30>
- Schiffer, F., Teicher, M. H., Anderson, C., Tomoda, A., Polcari, A., Navalta, C. P., & Andersen, S. L. (2007). Determination of hemispheric emotional valence in individual subjects: a new approach with research and therapeutic implications. *Behavioral and Brain Functions : BBF*, 3, 13. <https://doi.org/10.1186/1744-9081-3-13>
- Schwartz, G. E., Davidson, R. J., & Maer, F. (1975). Right hemisphere lateralization for emotion in the human brain: interactions with cognition. *Science (New York, N.Y.)*, 190(4211), 286–288.
- Schwerdtfeger, A. (2004). Predicting autonomic reactivity to public speaking: Don't get fixed on self-report data! *International Journal of Psychophysiology*, 52(3), 217–224.
- Setz, C., Arnrich, B., Schumm, J., La Marca, R., Tröster, G., & Ehlert, U. (2010). Discriminating stress from cognitive load using a wearable EDA device. *IEEE Transactions on Information Technology in Biomedicine*, 14(2), 410–7.
- Setz, C., Schumm, J., Lorenz, C., Arnrich, B., & Tröster, G. (2009). Combining worthless sensor data. In *Measuring Mobile Emotions Workshop at MobileHCI*.
- Shen, L., Wang, M., & Shen, R. (2009). Affective e-learning: using “emotional” data to improve learning in pervasive learning environment. *Educational Technology & Society*, 12(2), 176–189.
- Singh, R. R., Conjeti, S., & Banerjee, R. (2013). A comparative evaluation of neural network classifiers for stress level analysis of automotive drivers using physiological signals. *Biomedical Signal Processing and Control*, 8(6), 740–754. <https://doi.org/10.1016/j.bspc.2013.06.014>

- Soleymani, M., Lichtenauer, J., Pun, T., & Pantic, M. (2012). A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1), 42–55. <https://doi.org/10.1109/T-AFFC.2011.25>
- Soleymani, M., Pantic, M., & Pun, T. (2012). Multimodal Emotion Recognition in Response to Videos. *IEEE Transactions on Affective Computing*, 3(2), 211–223. <https://doi.org/10.1109/T-AFFC.2011.37>
- Stephens, C. L., Christie, I. C., & Friedman, B. H. (2010). Autonomic specificity of basic emotions: Evidence from pattern classification and cluster analysis. *Biological Psychology*, 84(3), 463–473.
- Student. (1908). The Probable Error of a Mean. *Biometrika*, 6(1), 1.
- Swenson, R. S. (2006). *Review of Clinical and Functional Neuroscience*. New York.: Castle Connolly Graduate Medical Publishing, Ltd.
- Syarif, I., Zaluska, E., Prugel-Bennett, A., & Wills, G. (2012). Application of bagging, boosting and stacking to intrusion detection (pp. 593–602). Springer Berlin Heidelberg.
- Takahashi, K., Namikawa, S., & Hashimoto, M. (2012). Computational emotion recognition using multimodal physiological signals: Elicited using Japanese kanji words. In *2012 35th International Conference on Telecommunications and Signal Processing (TSP)* (pp. 615–620). IEEE. <https://doi.org/10.1109/TSP.2012.6256370>
- Tao, J., & Tan, T. (2005). Affective computing: A review. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 3784 LNCS, pp. 981–995).
- Task-force. (1996). Heart rate variability. Standards of measurement, physiological interpretation, and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. *European Heart Journal*, 17(3), 354–381. <https://doi.org/0195-668X/96/030354> + 28 \$18.00/0
- Theodoridis, S., & Koutroumbas, K. (2006). *Pattern Recognition*, Third Edition.
- Timofeev, R. (2004). *Classification and Regression Trees –Theories and applications*. Humboldt University.
- Tognetti, S., Garbarino, M., Bonanno, A. T., Matteucci, M., & Bonarini, A. (2010). Enjoyment recognition from physiological data in a car racing game. In *Proceedings of the 3rd international workshop on Affective interaction in natural environments - AFFINE '10* (p. 3). New York, New York, USA: ACM Press.
- Torres, C. A., Orozco, A. A., & Alvarez, M. A. (2013). Feature selection for multimodal emotion recognition in the arousal-valence space. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* (pp. 4330–4333).
- Uljarevic, M., & Hamilton, A. (2013). Recognition of emotions in autism: A formal meta-

- analysis. *Journal of Autism and Developmental Disorders*, 43(7), 1517–1526.
- Vaid, S., Singh, P., & Kaur, C. (2015). Classification of human emotions using multiwavelet transform based features and random forest technique. *Indian Journal of Science and Technology*, 8(28).
- Vapnik, V. N. (1998). *Statistical Learning Theory*. John Wiley & Sons, Inc. <https://doi.org/10.2307/1271368>
- Verma, G. K., & Tiwary, U. S. (2014). Multimodal fusion framework: a multiresolution approach for emotion classification and recognition from physiological signals. *NeuroImage*, 102(P1), 162–72. <https://doi.org/10.1016/j.neuroimage.2013.11.007>
- Wagner, J., Kim, J., & Andre, E. (2005). From Physiological Signals to Emotions: Implementing and Comparing Selected Methods for Feature Extraction and Classification. In *IEEE International Conference on Multimedia and Expo* (pp. 940–943). IEEE. <https://doi.org/10.1109/ICME.2005.1521579>
- Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, 38(1), 223–230. <https://doi.org/10.1016/j.eswa.2010.06.048>
- Wang, J., Zhou, S., Yi, Y., & Kong, J. (2014). An improved feature selection based on effective range for classification. *TheScientificWorldJournal*, 2014, 972125.
- Wang, X.-W., Nie, D., & Lu, B.-L. (2014). Emotional state classification from EEG data using machine learning approach. *Neurocomputing*, 129, 94–106. <https://doi.org/10.1016/j.neucom.2013.06.046>
- Wassmann, C. (2010). Reflections on the “body loop”: Carl Georg Lange’s theory of emotion. *Cognition & Emotion*, 24(6), 974–990.
- Webb, A. R. (2002). *Statistical Pattern Recognition* (Vol. 9). Chichester, UK: John Wiley & Sons, Ltd.
- Wei-Long Zheng, W.-L., Bo-Nan Dong, B.-N., & Bao-Liang Lu, B.-L. (2014). Multimodal emotion recognition using EEG and eye tracking data. In *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 5040–5043). IEEE.
- Wei-Long Zheng, & Bao-Liang Lu. (2015). Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3), 162–175. <https://doi.org/10.1109/TAMD.2015.2431497>
- Wen, W., Liu, G., Cheng, N., Wei, J., Shanguan, P., & Huang, W. (2014). Emotion recognition based on multi-variant correlation of physiological signals. *IEEE Transactions on Affective Computing*, 5(2), 126–140.
- Wilcoxon, F. (1945). Individual comparisons by ranking method. *Biometrics Bulletin*, 1, 80–83.

- Wilson, G. F., & Russell, C. A. (2003). Operator functional state classification using multiple psychophysiological features in an air traffic control task. *Human Factors*, 45(3), 381–9.
- Wioleta, S. (2013). Using physiological signals for emotion recognition. In *2013 6th International Conference on Human System Interactions (HSI)* (pp. 556–561). IEEE. <https://doi.org/10.1109/HSI.2013.6577880>
- Witten, I. H., Frank, E., & Hall, M. A. (Mark A. (2011). *Data mining : Practical machine learning tools and techniques*. Morgan Kaufmann.
- Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2), 241–259. [https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1)
- Wu, D., Courtney, C. G., Lance, B. J., Narayanan, S. S., Dawson, M. E., Oie, K. S., & Parsons, T. D. (2010). Optimal arousal identification and classification for affective computing using physiological signals: virtual reality stroop task. *IEEE Transactions on Affective Computing*, 1(2), 109–118.
- Yannakakis, G. N., & Hallam, J. (2008). Entertainment modeling through physiology in physical play. *International Journal of Human-Computer Studies*, 66(10), 741–755. <https://doi.org/10.1016/j.ijhcs.2008.06.004>
- Yannakakis, G. N., Martínez, H. P., & Jhala, A. (2010). Towards affective camera control in games. *User Modeling and User-Adapted Interaction*, 20(4), 313–340. <https://doi.org/10.1007/s11257-010-9078-0>
- Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1), 39–58.
- Zenko, B., Todorovski, L., & Dzeroski, S. (2001). A comparison of stacking with meta decision trees to bagging, boosting, and stacking with other methods. *Proceedings 2001 IEEE International Conference on Data Mining*, 8, 669–670.
- Zhai, J., & Barreto, A. (2006a). Stress detection in computer users based on digital signal processing of noninvasive physiological variables. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 1, 1355–8. <https://doi.org/10.1109/IEMBS.2006.259421>
- Zhai, J., & Barreto, A. (2006b). Stress detection in computer users through non-invasive monitoring of physiological signals. *Biomedical Sciences Instrumentation*, 42, 495–500.
- Zhao, Z., Morstatter, F., Sharma, S., Alelyani, S., Anand, A., & Liu, H. (2010). Advancing feature selection research. *ASU Feature Selection Repository Arizona State University*, 1–28.
- Zhu, D. (2010). A hybrid approach for efficient ensembles. *Decision Support Systems*, 48(3), 480–487. <https://doi.org/10.1016/j.dss.2009.06.007>

LIST OF PUBLICATIONS AND PAPERS PRESENTED

Articles

- M. Malekzadeh, M.B. Mustafa, A.Lahsasna, 2015, “A Review of Emotion Regulation in Intelligent Tutoring Systems”, *Journal of Educational Technology & Society*, 18 (4), 435–445
- M. Malekzadeh, M.B. Mustafa, S.S. Salim, A.Lahsasna 2017, “Human Emotional State Recognition from Physiological Signals: Comparison of Benchmark Classification Methods”, *Journal of Applied Sciences* (Accepted for publication)

Book Chapter

- M. Malekzadeh, S.S. Salim and M.B. Mustafa, “Towards Integrating Emotion Management Strategies in Intelligent Tutoring System Used by Children”, *4th International Symposium on Pervasive Computing Paradigms for Mental Health*, Tokyo, Japan, May 8-9, 2014.

APPENDIX A: THE SAMPLES OF CODE USED FOR FEATURE EXTRACTION

```
clc
clear all
close all

for m=1:32 %Subjects

if m<10
loadfile=strcat('D:\emotion
databases\data_preprocessed_matlab_DEAP\s0',int2str(m),'.mat');
else
loadfile=strcat('D:\emotion
databases\data_preprocessed_matlab_DEAP\s',int2str(m),'.mat');
end

DEAP = load(loadfile);
for n=1:40 %#Videos
    for nn=1:32 %EEG channels

        R1(1,:)=DEAP.data(n,nn,:);

        % plot(R1)

        %%
        fs=128; % Sample rate
        L=8; % Order of Filter
        fl=1; % low cut_off frequency
        fh1=60; % high cut_off frequency
        L2=20;
        f2=48;
        fh2=52;
        d1 = fdesign.bandpass('N,F3dB1,F3dB2',L,fl,fh1,fs);
        Hd1 = design(d1,'butter');
        filtered_data = filter(Hd1,R1); %bandpass filter (EEG
frequency rang)
        d2 = fdesign.bandstop('N,F3dB1,F3dB2',L2,f2,fh2,fs);
        Hd2 = design(d2,'butter');
        EEG_data=filter(Hd2, filtered_data); %bandstop filter (notch)

        % hold on, plot(EEG_data,'r')

        wlen=ceil(fs*1.024);
        nfft=wlen;
        h=ceil(0.75*wlen);
        [stft1, f, t] = stft(R1, wlen, h, nfft, fs);
        a_stft1=abs(stft1);
        % figure, surf(t,f,a_stft1)
        dd_stft1(n,nn,:,:)=a_stft1(:,:);
```

Figure A. 1: The MATLAB code developed for EEG signal feature extraction

```

s_theta=5;
    e_theta=9;
    psd_thetal(n,nn)=sum(sum
(dd_stft1(n,nn,s_theta:e_theta,:))/(f(e_theta)-f(s_theta)));

    s_alpha=9;
    e_alpha=13;
    psd_alpha1(n,nn)=sum(sum
(dd_stft1(n,nn,s_alpha:e_alpha,:))/(f(e_alpha)-f(s_alpha)));

    s_alpha_slow=9;
    e_alpha_slow=11;
    psd_slwalphal(n,nn)=sum(sum
(dd_stft1(n,nn,s_alpha_slow:e_alpha_slow,:))/(f(e_alpha_slow)-
f(s_alpha_slow)));

    s_beta= 13;
    e_beta= 32;
    psd_beta1(n,nn)=sum(sum
(dd_stft1(n,nn,s_beta:e_beta,:))/(f(e_beta)-f(s_beta)));

    s_gamma= 32;
    e_gamma= 64;
    psd_gamma1(n,nn)=sum(sum
(dd_stft1(n,nn,s_gamma:e_gamma,:))/(f(e_gamma)-f(s_gamma)));

    end
end

%%
d=0;
e=[1:12 14 15 30:-1:19 18 17];
for i=1:40
    for b=1:14

% power assymetry (subtracting)
psd_theta2(i,b)=sum(sum (dd_stft1(i,e(b),s_theta:e_theta,:)-
dd_stft1(i,e(b+14),s_theta:e_theta,:))/(f(e_theta)-f(s_theta)));
psd_alpha2(i,b)=sum(sum (dd_stft1(i,e(b),s_alpha:e_alpha,:)-
dd_stft1(i,e(b+14),s_alpha:e_alpha,:))/(f(e_alpha)-f(s_alpha)));
psd_beta2(i,b)=sum(sum (dd_stft1(i,e(b),s_beta:e_beta,:)-
dd_stft1(i,e(b+14),s_beta:e_beta,:))/(f(e_beta)-f(s_beta)));
psd_gamma2(i,b)=sum(sum (dd_stft1(i,e(b),s_gamma:e_gamma,:)-
dd_stft1(i,e(b+14),s_gamma:e_gamma,:))/(f(e_gamma)-f(s_gamma)));

    end
end
temp_feature_table =
cat(2,psd_alpha1,psd_alpha2,psd_alpha3,psd_slwalphal,psd_beta1,psd_beta
2,psd_beta3,psd_gamma1,psd_gamma2,psd_gamma3,psd_thetal,psd_theta2,psd_
theta3);

```

Figure A. 1: The MATLAB code developed for EEG signal feature extraction (contd)


```

function [M1,
MD,Mean_neg_dev,pnsd,number_localmin,spectral_p,scsr,mean_peak_SCSR,scv
sr,mean_peak_SCVSR,GSR_power1,GSR_power2,GSR_power3,GSR_power4,GSR_powe
r5,GSR_power6,GSR_power7,GSR_power8,GSR_power9,GSR_power10,GSR_Totalpow
er] = GSR_feats(x)
DEAP = load(x);
y(:, :) = DEAP.data(:, 37, :);
x1=0;
x2=1;
for i=1:size(y,1)
Average_number=50; %number of data for averaging
% if your data is so noisy, you can increase Average_number
y1(i,:)=(smooth(y(i,:),Average_number))';
end
M=mean(y1,2); %Average-method1
for i=1:size(y1,1) %Moshtagh
D(i,:)=diff(y1(i,:));
end
for i=1:size(y1,1) %Average or mean
M1(i,1)=mean(y1(i,:));
GSR_Median(i,1)=median(y1(i,:)); %not compulsory
GSR_IntR(i,1) = iqr(y1(i,:)); %not compulsory
end
MD=mean(D,2); %Average Moshtagh
D1=diff(y1,1,2); %mohasebe moshtagh-ravesh jadid
p=D<0; % make flag for negative values in derivaive (D)
for i=1:size(y1,1) %count of of negative samples based on flag
(0/1)in p array
nsd(i,1)=sum(p(i,:));
end
for i=1:40 %calculation of proportion of negative samplesin
derivative vs all samples
pnsd(i,1)=nsd(i,1)/8063;
end
%calculation for average of derivative for negative samples
Sum_neg_dev=0
for i=1:size(D,1)
for j=1:size(D,2)
if D(i,j)<0
Sum_neg_dev=D(i,j)+Sum_neg_dev;
end
end
Mean_neg_dev(i,1)=Sum_neg_dev/nsd(i); %average of derivative for
negative samples
Sum_neg_dev=0;
end
% calculate localminimum
for i=1:40
for j=1:8062
if (y1(i,j)>y1(i,j+1)) && (y1(i,j+1)<=y1(i,j+2))
localmin(x2,1)=i; %save vector number 1...40
localmin(x2,2)=(j+1); %add position of localminimum
localmin(x2,3)=(y1(i,j+1)); %add value of each local minimum
x1=x1+1;
x2=x2+1;
end
end
end

```

Figure A. 2: The MATLAB code developed for GSR signal feature extraction

```

number_localmin(i,1)=x1; %save numbers of localminimum for each vector
x1=0;
end

%calculate spectral power in band [0-2.4]
fs=128;
wlen=ceil(fs*1.024/4);
nfft=wlen;
h=ceil(0.75*wlen);
for k=1:40
    all_A2=y1(k,:);
    [stft1, f1, t1] = stft(all_A2, wlen, h, nfft, fs); %all_A2=your
data
    a_stft1=abs(stft1);
    f = fs/2*linspace(0,1,nfft/2+1);
    t=0:30/length(t1):30-1/length(t1);
    ss=sum(a_stft1,2);
    ss1=0;
    for i=1:size(ss,1)-1
        if (ss(i+1)<= 10*ss(i))
            last=i+1;
            break
        end
    end
    for t=1:last
        ss1=ss(t)+ss1;
    end
    ss2=ss1+(ss(last)+ss(last+1))/2;
    ss2=ss2/2.5;
    spectral_p(k,1)=ss2;
    ss1=0;ss2=0;
end

%***Calculate power with fft
fs=128; % Sample rate
time=63;
T = 1/fs; % Sample time
Le = length(y1(1,:)); % Length of signal
tt = (0:Le-1)*T; % Time vector

for i=1:size(y1,1)
NFFT = 2^nextpow2(Le); % Next power of 2 from length of y
Y = fft(y1(i,:),NFFT)/2;
ff = fs/2*linspace(0,1,NFFT/2+1);
%Plot single-sided amplitude spectrum.
absY(i,:)=2*abs(Y(1:NFFT/2+1));
end

%***Calculate 10 spectral power in the bands[0-2.4]HZ
for i=1:size(y1,1)
GSR_power1(i,1)=sum(absY(i,1:16)); %where between f= 0 & 0.24...check
f vector
GSR_power2(i,1)=sum(absY(i,17:32));%where between f= 0.24&
0.48...check f vector
GSR_power3(i,1)=sum(absY(i,33:47));
GSR_power4(i,1)=sum(absY(i,48:63));
GSR_power5(i,1)=sum(absY(i,64:78));
GSR_power6(i,1)=sum(absY(i,79:93));

```

Figure A. 3: The MATLAB code developed for GSR signal feature extraction (contd)

```

GSR_power7(i,1)=sum (absY(i,94:109));
GSR_power8(i,1)=sum (absY(i,110:124));
GSR_power9(i,1)=sum (absY(i,125:139));
GSR_power10(i,1)=sum (absY(i,140:155));
GSR_Totalpower(i,1)=sum (absY(i,155)); %Calculate spectral power in the
bands [0-2.4]HZ
end
%**** calculation of zero crossing rate of skin conductance Slow
Response(SCSR)****
%make butter filter on time-based vector
fs=128; % Sample rate
L=8; % Order of Filter
fl=0.00001; % low cut_off frequency
fh1=0.2; % high cut_off frequency
d1 = fdesign.bandpass('N,F3dB1,F3dB2',L,fl,fh1,fs);
Hd1 = design(d1,'butter');
for i=1:40
    filtered_data(i,:) = filter(Hd1,y1(i,:)); %make butter filter
on the original time based vectors
    p1(i,:)=filtered_data(i,:)<0; %% make flag for negative values
in signalvectors (y)
    z=0;
    % calculation of zero crossing rate of skin conductance
    for j=1:size(p1,2)-1
        if (p1(i,j)~=p1(i,j+1))
            z=z+1;
            crs_scsr(z,1)=i;
            crs_scsr(z,2)=j+1;%save position of zero crossing
            crs_scsr(z,3)=filtered_data(j+1); %save related value of zero
crossing
        end
    end
    scsr(i,1)=z; %save zero crossing rate of skin conductance
slow response(SCSR 0-0.08)
end
%****calculate mean of peaks magnitude SCSR****

for i=1:40
    [pks1,locs1] = findpeaks(filtered_data(i,:));
    mean_peak_SCSR(i,1)=sum(pks1(1,:))/size(pks1,2);
    plot(filtered_data (i,:), '--r')
    hold on, plot(locs1,pks1,'b*')
end
%**** calculation of zero crossing rate of skin conductance Very Slow
Response(SCVSR)****
%make butter filter on time-based vector
fs=128; % Sample rate
L=8; % Order of Filter
fl=0.00001; % low cut_off frequency
fh1=0.08; % high cut_off frequency
d1 = fdesign.bandpass('N,F3dB1,F3dB2',L,fl,fh1,fs);
Hd1 = design(d1,'butter');
for i=1:40
    filtered_data1(i,:) = filter(Hd1,y1(i,:)); %make butter filter
on the original time based vectors
    p2(i,:)=filtered_data1(i,:)<0; %% make flag for negative values
in signalvectors (y)

```

Figure A. 4: The MATLAB code developed for GSR signal feature extraction (contd)

```

% calculation of zero crossing rate of skin conductance
for j=1:size(p2,2)-1
    if (p2(i,j)~=p2(i,j+1))
        z=z+1;
        crs_scvsr(z,1)=i;
        crs_scvsr(z,2)=j+1;%save position of zero crossing
        crs_scvsr(z,3)=filted_data1(j+1); %save related value of
zero crossing
    end
end
scvsr(i,1)=z; %save zero crossing rate of skin
conductance very slow response(SCSR 0-0.08)
end

%***calculate mean of peaks magnitude SCVSR***

x1=1;
w=1;

for i=1:40
    [pks,locs] = findpeaks(filted_data1(i,:));
    mean_peak_SCVSR(i,1)=sum(pks(1,:))/size(pks,2);
% plot(filted_data1 (i,:), '--r')
% hold on, plot(locs,pks,'b*')
end

Mean_neg_dev(isnan(Mean_neg_dev(:,:)))=0; %to remove Nan Values
mean_peak_SCSR(isnan(mean_peak_SCSR(:,:)))=0; %to remove Nan Values
mean_peak_SCVSR(isnan(mean_peak_SCVSR(:,:)))=0; %to remove Nan
Values
end

```

Figure A. 5: The MATLAB code developed for GSR signal feature extraction (contd)

**APPENDIX B: CLASSIFICATION ACCURACY RESULTS OF VALENCE,
AROUSAL, AND LINKING RECOGNITION USING DIFFERENT FEATURE
SUBSET SIZES**

Table B. 1: Valence classification accuracies using different feature subset sizes and CART classifiers on peripheral modality

FS Method ¹¹	Numbers of features in each feature subset									Majority Vote
	8	16	24	32	40	48	56	64	72	
1-Jmi	59.61	59.30	58.52	59.61	58.52	58.52	58.83	58.52	58.05	58.91
2-Cmim	58.91	60.16	58.20	58.05	58.05	58.05	58.05	58.05	58.05	58.05
3-Disr	56.95	58.91	58.75	58.75	58.13	58.67	58.36	58.59	58.44	58.05
4-Mim	57.50	57.42	59.84	58.83	58.59	58.52	58.98	58.75	58.52	59.45
5-Cife	57.19	57.50	57.42	56.88	56.64	55.78	56.48	56.25	56.33	57.03
6-Icap	58.20	59.06	59.61	58.52	59.84	59.45	57.89	56.48	56.41	60.08
7-Condred	55.78	56.80	59.30	58.52	59.69	59.06	59.06	58.52	57.81	59.06
8-Relief	57.89	58.28	58.98	58.05	57.34	57.81	57.66	57.89	57.11	58.28
9-Fisher	59.69	57.97	57.73	58.83	59.06	58.91	58.05	57.27	56.64	59.06
10-T-test	58.20	56.09	56.72	56.72	57.11	55.86	55.63	54.53	53.75	57.34

Table B. 2: Average testing accuracy rates of the feature-based multi-classifier methods for valence recognition using different classifiers on peripheral modality

FS Method	CART	ANN	LDA	SVM
1-Jmi	58.9063	60.47	60.3125	55.3125
2-Cmim	58.0469	58.59	59.2969	54.2969
3-Disr	58.0469	61.64	60.2344	55.7031
4-Mim	59.4531	62.42	60.9375	55.5469
5-Cife	57.0313	61.02	60.5469	54.2188
6-Icap	60.0781	59.38	59.6094	54.6094
7-Condred	59.0625	60.00	60.1563	55.1563
8-Relief	58.2812	60.39	59.375	54.375
9-Fisher	59.0625	62.11	59.2188	55.9375
10-T-test	57.3438	55.23	58.0469	53.4375

¹¹ FS: Feature Selection

Table B. 3: Average testing accuracy rates of the feature-based multi-classifier methods for arousal recognition using different classifiers on peripheral modality

FS Method	CART	ANN	LDA	SVM
1-Jmi	58.90625	63.28125	58.75	64.0625
2-Cmim	57.26563	60	56.64063	63.51563
3-Disr	57.73438	61.71875	59.21875	64.21875
4-Mim	58.125	62.42188	59.0625	64.375
5-Cife	59.45313	62.89063	58.59375	64.21875
6-Icap	57.5	62.1875	60.3125	63.98438
7-Condred	58.4375	61.5625	59.0625	64.0625
8-Relief	59.60938	62.8125	58.4375	64.21875
9-Fisher	58.4375	63.4375	59.84375	64.76563
10-T-test	54.45313	61.5625	57.5	62.57813

Table B. 4: Average testing accuracy rates of the feature-based multi-classifier methods for liking recognition using different classifiers on peripheral modality

FS Method	CART	ANN	LDA	SVM
1-Jmi	61.71875	63.75	62.42188	63.04688
2-Cmim	62.57813	62.57813	59.45313	63.04688
3-Disr	60.54688	62.42188	61.875	62.89063
4-Mim	60.85938	62.65625	61.79688	63.04688
5-Cife	63.59375	63.51563	61.64063	62.5
6-Icap	60.54688	62.5	61.25	62.96875
7-Condred	62.26563	63.51563	61.79688	62.96875
8-Relief	60.07813	61.25	61.01563	63.125
9-Fisher	58.35938	64.21875	60.70313	62.65625
10-T-test	57.03125	62.65625	60.625	63.82813

Table B. 5: Valence classification accuracies using different feature subset sizes and CART classifiers on EEG modality

FS Method	Numbers of features in each feature set									Majority Vote
	24	48	72	96	120	144	168	192	216	
1-Jmi	58.52	59.22	59.84	58.67	58.05	58.98	57.81	58.13	59.84	60.16
2-Cmim	59.38	59.77	59.77	59.77	59.77	59.77	59.77	59.77	59.77	59.77
3-Disr	58.05	58.59	60.00	58.36	58.20	59.38	57.73	58.20	59.69	59.84
4-Mim	57.19	59.53	59.84	58.52	57.50	58.91	57.97	58.36	59.69	60.31
5-Cife	57.73	57.11	60.70	59.69	60.39	60.63	57.97	58.67	60.16	60.63
6-Icap	59.06	58.52	60.78	59.22	58.67	59.53	58.75	58.83	59.38	59.61
7-Condred	60.23	59.69	59.84	59.22	58.05	59.53	58.05	58.44	60.08	60.78
8-Relief	59.14	60.78	60.31	59.22	58.75	58.98	58.52	59.45	58.98	59.53
9-Fisher	60.00	58.83	57.89	59.53	58.91	59.84	59.45	59.92	60.00	60.08
10-T-test	60.16	60.08	58.20	57.58	57.89	60.39	61.17	61.72	60.23	58.98

Table B. 6: Average testing accuracy rates of the feature-based multi-classifier methods for valence recognition using different classifiers on EEG modality

FS Method	CART	ANN	LDA	SVM
1-Jmi	60.16	63.5938	62.1875	53.4375
2-Cmim	59.77	63.125	58.4375	50.0781
3-Disr	59.84	63.6719	62.3438	53.6719
4-Mim	60.31	63.2813	62.1875	53.4375
5-Cife	60.63	65.4688	62.9688	52.8125
6-Icap	59.61	62.1875	63.5938	51.25
7-Condred	60.78	63.9063	61.5625	53.5156
8-Relief	59.53	63.5938	64.4531	52.1094
9-Fisher	60.08	66.3281	64.375	53.2031
10-T-test	58.98	65.1563	63.2813	49.375

Table B. 7: Average testing accuracy rates of the feature-based multi-classifier methods for arousal recognition using different classifiers on EEG modality

FS Method	CART	ANN	LDA	SVM
1-Jmi	59.9219	63.5938	57.4219	62.1875
2-Cmim	58.4375	61.0938	52.0313	62.3438
3-Disr	60.3125	62.0313	57.0313	62.4219
4-Mim	59.6094	63.5938	57.3438	62.1875
5-Cife	59.7656	62.3438	57.8125	62.5781
6-Icap	58.3594	62.9688	57.6563	62.6563
7-Condred	60.625	63.125	57.2656	62.2656
8-Relief	57.3438	63.9844	56.4063	62.5
9-Fisher	58.5156	62.5	57.6563	62.9688
10-T-test	57.8125	62.5	59.375	62.6563

Table B. 8: Average testing accuracy rates of the feature-based multi-classifier methods for liking recognition using different classifiers on EEG modality

FS Method	CART	ANN	LDA	SVM
1-Jmi	57.3438	66.1719	61.7969	66.0938
2-Cmim	58.5156	64.0625	57.5	65
3-Disr	57.8125	64.7656	61.7969	66.1719
4-Mim	57.5781	64.6875	61.7969	66.1719
5-Cife	58.2813	65.625	62.3438	66.0156
6-Icap	58.2031	65.0781	61.875	65.9375
7-Condred	56.3281	65.7813	61.9531	66.25
8-Relief	55.8594	63.9844	62.4219	65.0781
9-Fisher	57.2656	66.0156	61.875	66.5625
10-T-test	56.25	65.9375	61.9531	66.6406

Table B. 9: Valence classification accuracies using different feature subset sizes and CART classifiers on (Peripheral+EEG) modalities

FS Method	Numbers of features in each feature subset									Majority Vote
	32	64	96	128	160	192	224	256	288	
1-Jmi	57.97	57.27	55.94	54.30	57.34	57.89	58.75	58.13	58.13	58.67
2-Cmim	59.14	59.30	59.77	60.23	60.23	60.23	60.23	60.23	60.23	60.23
3-Disr	57.03	55.78	54.06	55.47	56.02	58.28	58.59	58.20	58.59	57.73
4-Mim	58.13	57.19	56.33	53.91	57.19	58.36	59.30	58.28	58.36	58.05
5-Cife	58.28	56.80	55.86	57.66	57.89	59.22	58.20	57.81	59.06	59.69
6-Icap	60.16	59.14	59.77	58.44	58.36	57.58	57.27	57.89	57.73	59.22
7-Condred	59.53	57.27	56.48	55.08	58.59	59.45	59.14	58.44	58.44	59.45
8-Relief	54.77	57.97	60.70	60.70	60.39	59.84	59.38	60.39	59.53	61.02
9-Fisher	62.27	58.28	56.72	57.81	57.73	58.67	58.52	59.22	59.77	58.44
10-T-test	59.22	61.02	61.09	59.77	60.00	59.30	58.59	59.22	59.53	60.39

Table B. 10: Average testing accuracy rates of the feature-based multi-classifier methods for valence recognition using different classifiers on (Peripheral+EEG) modalities

FS Method	CART	ANN	LDA	SVM
1-Jmi	58.67	64.6094	65.3125	53.8281
2-Cmim	60.23	64.9219	61.4844	53.5938
3-Disr	57.73	63.8281	65	53.8281
4-Mim	58.05	64.7656	65.0781	53.9063
5-Cife	59.69	62.6563	65.7813	53.2031
6-Icap	59.22	63.8281	64.9219	50.4688
7-Condred	59.45	64.0625	65.0781	53.9844
8-Relief	61.02	64.2969	65.1563	52.1094
9-Fisher	58.44	66.5625	65.7813	53.2031
10-T-test	60.39	63.3594	64.6094	51.5625

Table B. 11: Average testing accuracy rates of the feature-based multi-classifier methods for arousal recognition using different classifiers on (Peripheral+EEG) modalities

FS Method	CART	ANN	LDA	SVM
1-Jmi	58.6719	61.9531	60.7813	63.2031
2-Cmim	54.6094	60.7031	54.6875	63.0469
3-Disr	58.6719	62.1094	60.9375	63.4375
4-Mim	57.6563	63.75	61.0156	63.0469
5-Cife	58.75	62.3438	61.0938	62.3438
6-Icap	59.5313	61.5625	58.9063	62.5
7-Condred	58.125	61.875	61.1719	63.125
8-Relief	57.3438	62.5781	57.9688	61.9531
9-Fisher	58.75	65	60.7813	63.75
10-T-test	56.7188	63.0469	60.4688	61.875

Table B. 12: Average testing accuracy rates of the feature-based multi-classifier methods for liking recognition using different classifiers on (Peripheral+EEG) modalities

FS Method	CART	ANN	LDA	SVM
1-Jmi	60.8594	64.1406	62.4219	63.9844
2-Cmim	60.4688	62.8125	58.5938	63.8281
3-Disr	61.3281	64.0625	62.7344	63.9063
4-Mim	61.1719	63.75	62.8125	63.9063
5-Cife	61.875	62.9688	62.9688	64.0625
6-Icap	61.9531	62.8906	62.8125	64.9219
7-Condred	60	63.5156	62.9688	63.9063
8-Relief	61.5625	63.5156	61.6406	65.2344
9-Fisher	59.8438	65.4688	61.7188	64.8438
10-T-test	61.7969	63.9063	62.8906	64.9219