Emotion Classification Using Advanced Machine Learning Techniques Applied to Wearable Physiological Signals Data

Bahareh Nakisa

Master of Science in Computer Science and Information Technology

A thesis by Publication submitted in fulfilment of the required for the degree of Doctor of Philosophy (Ph.D)

School of Electrical Engineering and Computer Science
Science and Engineering Faculty
Queensland University of Technology
2019

Keywords

Emotion Recognition
Wearable Sensors
Machine Learning
Deep Learning
Feature Extraction
Feature Selection
Evolutionary Algorithms
Hyperparameter Optimization
Long Short Term Memory
Convolutional Neural Network
Temporal Multimodal Deep Learning
Early Fusion
Late Fusion

Abstract

Computers are becoming an inevitable part of our everyday life and thus, it will come to be crucial that we have the ability to have natural interactions with them, similar to the way that we interact with other humans. One of the key features that grants intuitive interaction to the process of human–computer interaction is emotional states. Affective computing incorporates human emotional states into computer applications, which subsequently, may improve communication between emotionally expressive humans and emotionally deficient computers. Human emotional states can be recognized using different modalities such as facial expression and physiological signals. Physiological signals offer several advantages over facial expression due to their sensitivity to inner feelings and insensitivity to social masking of emotions.

Recent advances in emotion recognition using ubiquitous miniaturized wearable physiological sensors have opened up a new era of human–computer interaction (HCI) applications like mental health care, intelligent tutoring and transportation safety. Many of these wearable sensors are easy to use and support real-world applications. The wearable technologies record different physiological signals in an unobtrusive and non-invasive manner. Physiological signals such as Electroencephalography (EEG), and Blood Volume Pulse (BVP) carry information about non-visible emotional changes, as these physiological signals originate from the autonomic and central nervous systems.

To build a reliable and accurate emotion recognition system using physiological signals (EEG and BVP signals), several challenges like feature extraction and selection, classification and fusion of physiological signals need to be addressed. Although several approaches have been proposed to address these issues, there is still a need for more efficient solutions to achieve an accurate and reliable emotion classification system.

A multitude of studies have focused on extracting features from physiological signals, particularly from EEG signals to improve the performance of emotion classification. However, combining different features may cause high dimensionality

and reduce the performance of emotion recognition. But, the development of a suitable feature selection method to overcome the problem of high dimensionality along with determining the salient set of EEG features to improve emotion recognition have not been thoroughly investigated in this domain. Since physiological signals consist of time-series data with variation over a long period of time and dependencies within shorter periods, there is a need to apply a classifier like the Long Short Term Memory (LSTM) classifier which considers temporal information. However, there is a lack of research on choosing the best possible LSTM network and optimizing its hyperparameters to accurately classify different emotions. Building automatic emotion recognition using multimodal data can substantially improve the performance of classification. The data from multiple sources are correlated and there are some emotional related information across them which can provide complementary information. In order to exchange such information, it is important to capture the correlation between modalities over time. Previous works failed to capture the non-linear correlation across data modalities, and only focused on the emotional related information within each modality. Therefore, to build an automatic emotion classification system based on multimodal physiological signals, it is essential to capture both emotional information within and across physiological signals over time.

This thesis by publication adopts different approaches to maximize the performance of emotion classification based on portable wearable physiological sensors. A total of three papers highlighting the proposed challenges associated with emotion classification comprise this thesis.

In Study 1, we proposed a new framework using evolutionary computation algorithms for feature selection to improve the EEG-based emotion recognition based on the selected salient set of EEG features. In this study, most of the state-of-the-art EEG features are reviewed and implemented. The performance of each evolutionary algorithm (particle swarm optimization (PSO), ant colony optimization (ACO), genetic algorithm (GA), differential evolution (DE) and simulated annealing (SA)) is evaluated and compared using two public datasets with a 32-channel EEG sensor and our new dataset collected from a wireless EEG sensor with only 5 channels.

Study 2 focused on proposing a new framework to optimize hyperparameters of LSTM network and to find the best possible LSTM network to maximize emotion

classification. The proposed framework is evaluated on our dataset collected from portable wearable physiological sensors (EEG and BVP signals). Optimizing LSTM hyperparameters and finding appropriate LSTM hyperparameters values in the proposed framework is done by using a DE algorithm. In this study, we evaluate and compare the performance of the proposed framework with other state-of-the-art hyperparameter optimization methods (PSO, SA, Random Search and Tree-of-Parzen-Estimators (TPE)) to classify dimensional emotions (four-quadrant dimensional emotion).

In Study 3, a novel framework based on temporal multimodal deep learning models is proposed to automatically fuse EEG and BVP signals and maximize the performance of dimensional emotion classification (four-quadrant dimensional emotions). The proposed model is based on convolutional neural networks and LSTM (ConvNets LSTM) networks. The proposed model is evaluated based on early and late fusions in an end-to-end fashion. Based on an end-to-end learning approach, the network is trained from the raw data without any a priori feature extraction. The temporal multimodal deep learning models based on ConvNets LSTM networks help in fusing EEG and BVP signals temporally to capture the temporal emotional structure within and across the modalities. The performances of the proposed models are evaluated and compared with handcrafted feature extraction methods (Study 2) using our dataset collected from wearable sensors (EEG and BVP signals).

The results of Study 1 demonstrated that evolutionary algorithms can effectively support feature selection to identify the salient EEG features and channels to improve the performance of a dimensional emotion classification problem. The results also showed that a combination of time and frequency features is consistently more efficient than using only time or frequency features. This study also confirmed the validity of using wireless EEG sensors for emotion recognition.

In Study 2, we identified that optimizing the LSTM hyperparameters can result in a significant increase in emotion classification performance based on EEG and BVP signals. The results also showed that optimizing LSTM hyperparameters using DE algorithm was superior compared to the other-state-of-the-art hyperparameter optimization methods.

The finding regarding the impact of the temporal multimodal deep learning models on emotion classification, Study 3, demonstrated the ability of the proposed models applied to the fusion of EEG and BVP signals to capture the temporal emotional patterns within and across the signals, and improve the performance of emotion classification. The results also showed that the temporal multimodal deep learning models can be used to more accurately \recognize dimensional emotions on a dataset collected from wireless wearable sensors as compared to traditional handcrafted features (Study 2).

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AC	Affective Computing
ACO	Ant Colony
BVP	Blood Volume Pulse
ConvNet	Convolutional Neural Network
DE	Differential Evolution
HCI	Human Computer Interaction
EC	Evolutionary Computation Algorithms
ECG	Electrocardiogram
EMG	Electromyogram
EEG	Electroencephalography
ICA	Independent Component Analysis
GA	Genetic Algorithm
KNN	K-Nearest Neighbour Classification
LSTM	Long Short Term Memory Network
PSO	Particle Swarm Optimization
PPG	Photoplethysmography
PNN	probabilistic Neural Network
RNNs	Recurrent Neural Networks
RSP	Respiration
SA	Simulated Annealing
SC	Skin Conductivity
SVM	Support Vector Machine

List of Publications

List of Q1 Journal Papers

1) **B. Nakisa,** M. N. Rastgoo, D. Tjondronegoro, and V. Chandran, "Evolutionary Computation Algorithms for Feature Selection of EEG-based Emotion Recognition using Mobile Sensors," Expert Systems with Applications, 2017. **(Published)**

2) **B. Nakisa,** M. N. Rastgoo, A. Rakotonirainy, F. Maire and V. Chandran, "Long Short Term Memory Hyperparameter Optimization for a Neural Network Based Emotion Recognition Framework" IEEE Access, 2018. **(Published)**

3) B. Nakisa, M. N. Rastgoo, A. Rakotonirainy, F. Maire and V. Chandran," Automatic Emotion Recognition using Temporal Multimodal Deep Learning" Information Fusion, 2018. **(Submitted)**

Statement of Original Authorship

The work contained in this thesis has not been previously submitted to meet requirements for an award at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Signature: QUT Verified Signature

Date: May 2019

Acknowledgements

This PhD was completed due to the immense support I received from numerous QUT staff members. I would like to thank my supervision team including Professor Andry Rakotonirainy (Principal Supervisor), Dr Frederic Maire (Associate Supervisor), and Professor Vinod Chandran (Associate Supervisor). I am tremendously grateful to all of my supervisors for their endless support throughout my PhD study. I would like to express my gratitude to Professor Andry Rakotonirainy for his support during my PhD. He helped me with my technical knowledge, writing and presenting the results. I am so fortunate to have Professor Vinod Chandran as my supervisor. He actively helped me with designing the methods and writing the articles, and provided in-depth feedback to improve my work. I have learnt a lot from him about writing, research, and professional knowledge. In addition, he always suggested helpful connections with the right people during my candidature. Dr Frederic Maire provided on-time and valuable feedback on my work. He helped me with technical issues, methods and writing. I acknowledge the financial support received from QUT, via QUT's postgraduate research scholarships (Research Training Program and International Postgraduate Research Scholarship) and Excellence Top-Up Scholarships. I am thankful to the Information Systems School, the School of Electrical Engineering and Computer Science and the Centre for Accident Research and Road Safety-Queensland (CARRS-Q), for all their support throughout my journey.

My heartfelt thanks go to my husband, Naim, for his steady mental support, consideration, patience, help, and encouragement. This work could not have been completed without his pure devotion, sacrifice and continued prayers towards my success. To my lovely Mum and Dad, words simply cannot thank you for all that you have done for me. Thank you for your encouragement, patience, positivity and loves from miles away. I would never have been able to achieve my goals without you, and for that I am grateful. To my sister, Banafsheh, for listening to all my phone calls and making me lough, I miss you every day.

Finally, I want to thank all my colleagues at QUT and to express my gratitude to all the anonymous students and staff of QUT who helped in my research via their participation in my studies.

Chapter 1: Introduction

1.1 BACKGROUND AND MOTIVATION

Computers are quickly becoming a ubiquitous part of human life. However, they are emotionally blind and cannot understand human emotional states. Reading and understanding human emotional states could maximize the performance of human–computer interaction (HCI). Therefore, it is essential to exchange this information and recognize the user's affective states to enhance HCI. Affective computing (AC) is a research domain which focuses on HCI through user affect detection and one of the main aims of the AC domain is to find ways for machines to recognize human emotion, which may enhance the capacity for communication between them (Calvo & D'Mello, 2010). There are many application areas like intelligent tutoring (Calvo & D'Mello, 2010; Du Boulay, 2011), computer games (Mandryk & Atkins, 2007), and e-Health applications (C. Liu, Conn, Sarkar, & Stone, 2008; Luneski, Bamidis, & Hitoglou-Antoniadou, 2008) that could benefit from an emotion recognition system.

An individual's emotional state can be recognized through physiological signals like electroencephalography (EEG), Blood Volume Pulse (BVP) and Galvanic Skin Response (GSR) (Koelstra et al., 2010), and physical indicators such as facial expression (Hossain & Muhammad, 2017). Physiological signals offer several advantages over physical indicators due to their sensitivity to inner feelings and insusceptibility to social masking of emotions (Kim, 2007). Gathering physiological signals can be done using two types of sensors: tethered-laboratory sensors and wireless physiological sensors. Although tethered-laboratory sensors are effective and can receive and record strong signals with higher resolution, they are more invasive and obtrusive and cannot be used for everyday life situations. Whereas, wireless physiological sensors can provide a non-invasive and unobtrusive way to collect physiological signals and can be utilized by individuals carrying out daily life activities. These miniaturized wearable sensors have rapidly been adopted by the general population due to the development of integrated sensor technologies, low

battery consumption and improved user experience design. Using these modern technologies, an automated system (learning model) could be developed that can accurately recognize different emotional states.

Among the physiological signals, EEG and BVP signals have been widely used to recognize different emotions, with evidence indicating a strong correlation between these signals and emotions such as sadness, anger, surprise (Haag, Goronzy, Schaich, & Williams, 2004; Kim, Bang, & Kim, 2004). An EEG sensor is usually placed on the scalp to record electrical activity in the brain. The strong correlation between EEG signals and different emotions is due to the fact that these signals come directly from the central nervous system (CNS), capturing features about internal emotional states. The use of everyday technology such as lightweight EEG headbands has been recently investigated in non-critical domains such as game experience (McMahan, Parberry, & Parsons, 2015), motor imagery (Kranczioch, Zich, Schierholz, & Sterr, 2014), and hand movement (Robinson & Vinod, 2015). Another physiological activity which correlates to different emotions is Blood Volume Pulse (BVP), a measure that determines the changes in blood volume in blood vessels and is regulated by the autonomic nervous system (ANS). In fact, the external stimuli which induce different emotions involuntarily modulate the activity of the ANS. BVP is measured by a photoplethysmography (PPG) sensor. PPG sensors can sense changes in light absorption density of skin and tissue when illuminated. PPG is a non-invasive and low-cost technique which has recently been embedded in smart wristbands. The usefulness of these wearable sensors has been proven in applications such as stress prediction (Ghosh, Danieli, & Riccardi, 2015) as well as emotion recognition (Haag et al., 2004).

Therefore, taking the advantage of these non-invasive technologies can enable us to develop an emotion recognition system which can be used in daily life situations. This thesis focuses on building accurate and reliable emotion classification models based on advance machine learning techniques using portable physiological sensors (capturing EEG and BVP signals).

1.2 RESEARCH PROBLEM

Building an emotion recognition system based on EEG and BVP signals to accurately classify different emotions is a challenging task. In order to build such a model, several methodological issues need to be addressed such as feature extraction and selection, choosing an appropriate classifier (time-series classifier) and the problem of hyperparameter optimization, and fusion of physiological signals. Although different approaches are proposed to address these issues in the domain of affective computing, the need exists for more efficient approaches to improve the performance of emotion classification. In the following sections, three main challenges: feature extraction and selection, classification and the problem of hyperparameter optimization, and finally the fusion of physiological signals are discussed.

1.2.1 Research Gap 1: Lack of Feature Selection Method to Find the Salient Set of EEG Features

Feature extraction, is one of the most important steps in emotion recognition. In essence, the process involves trying to determine the features which are able to differentiate different emotional states. To recognize emotions using physiological signals, there are a multitude of studies that focused on extracting new features.

Among the physiological signals, EEG signals are widely used for emotion recognition. In previous works many features from time, frequency and time-frequency domains have been extracted from EEG signals to recognize different emotions. However, there is no standardized set of EEG features that have been generally agreed to be the most suitable for emotion recognition. On the other hand, combining all the proposed EEG features from the literature can lead to a high-dimensionality issue, as not all of these features would carry significant information regarding emotions. Moreover, redundant features increase the feature space, making pattern detection more difficult and increasing the risk of overfitting. It is, therefore, important to identify salient features that have significant impact on the performance of the emotion classification model.

Feature selection methods have been shown to be effective in automatically decreasing high dimensionality by removing redundant and irrelevant features and maximizing the performance of classifiers. Feature selection is a difficult task because there can be complex interactions among features. An individually relevant (redundant or irrelevant) feature may become redundant (relevant) when working together with other features. Therefore, an optimal feature subset should be a group of complementary features that span the diverse properties of the classes to properly discriminate them. The feature selection task is also challenging because of the large search space. The size of the search space increases exponentially with respect to the number of available features in the data set (Guyon & Elisseeff, 2003). In order to better address feature selection problems, an efficient global search technique is needed which can find the salient subset of EEG features.

1.2.2 Research Gap 2: Lack of Optimized LSTM Classifier in emotion classification

Another challenging issue in the pipeline of building an emotion recognition system is choosing the best classifier which can accurately classify different emotions. Most of the studies in the AC domain have used simple learning techniques which failed to achieve the best possible performance using physiological signals. This is due to the fact that physiological signals are characterized by non-stationarities and non-linearities. In fact, physiological signals consist of time-series data with variation over a long period of time and dependencies within shorter periods. To capture the inherent temporal structure within the physiological data and recognize emotion signatures, we need to apply a classifier which is able to learn temporal patterns over time.

In recent years, the application of Recurrent Neural Networks (RNNs) to human emotion recognition has led to a significant improvement in recognition accuracy by modelling temporal data. RNN algorithms are able to elicit the context of observations within sequences and accurately classify sequences that have strong temporal correlation. However, RNNs have limitations in learning time-series data that stymie their training. Long Short Term Memory networks (LSTM), a special type of RNNs, have the capability of learning longer temporal sequences (Hochreiter & Schmidhuber, 1997). For this reason LSTM networks offer better emotion classification accuracy

over other methods when using time-series data (Kim, Lee, & Provost, 2013; Tsai, Weng, Ng, & Lee, 2017; Wöllmer et al., 2008; Wöllmer, Kaiser, Eyben, Schuller, & Rigoll, 2013).

Although the performance of LSTM networks in classifying different emotions is promising, training these networks depends heavily on a set of hyperparameters that determine many aspects of algorithm behaviour. To find a successful LSTM classifier which can accurately classify different emotions, we need to select optimal values for its hyperparameters to achieve a state-of-the-art result. These hyperparameters range from optimization hyperparameters such as number of hidden neurons and batch size, to regularization hyperparameters (weight). Given that these hyperparameters affect the quality of the model, it is therefore essential to find the best possible parameters values.

1.2.3 Research Gap 3: Lack of Temporal Multimodal Fusion Model for Automatic Emotion Classification

Building an accurate emotion classification system often involves different data from different modalities. However, fusing information from different modalities is usually non-trivial due to the distinct statistical features (Srivastava & Salakhutdinov, 2012) and the non-linear relationships between modalities. Most of the current studies in the domain of affective computing focused on fusing multimodal physiological signals using concatenating features from each modality to create one single feature vector. Although this approach has been widely applied, it cannot capture the non-linear emotional states across modalities. This is due to the fact that this approach focuses more on the correlation between features within each modality rather than finding non-linear correlation across the modalities. The non-linear emotional information across modalities can provide complementary information for emotion classification.

Recently, deep architecture and learning techniques have been shown to be effective in capturing non-linear feature interaction in multimodal data. Multimodal learning methods have been proposed to jointly learn and explore the highly correlated representation across modalities after learning each channel data with single deep

network. But physiological signals are inherently sequential and temporal in nature, which means that the current pattern in a signal is influenced by the previous one. And the built multimodal networks like deep Autoencoder, Boltzmann Machine could not model the temporal multimodal data. In order to build an accurate emotion recognition system using the fusion of EEG and BVP, it is essential to propose a multimodal learning model that can temporally capture and learn the inherent emotional changes within and across modalities over time.

To the best of our knowledge, this is the first study that proposes a temporal multimodal learning method for emotion recognition based on physiological signals (EEG and BVP signals). The proposed method is based on convolutional neural networks (ConvNets) LSTM in an end-to-end manner. Using end-to-end learning, the constructed features using ConvNets are trained jointly with the classification step as a single network. Moreover, in an end-to-end learning approach, the network is trained from the raw data without any a priori feature extraction. In addition, the performance of temporal learning models based on early and late fusions is also investigated to measure which type of fusion level performs better.

1.3 RESEARCH AIM, QUESTIONS AND OBJECTIVES

This research aims to **maximize the performance of an emotion classification system using portable wearable sensors.** Specifically, the following research questions and objectives are addressed in this program of research:

***RQ1*. What impacts do feature selection algorithms have on emotion recognition system?**

Research objective1. A new framework for feature selection algorithms based on Evolutionary Computing (EC) algorithms are proposed to automatically find the most salient EEG features and channels and improve emotion classification. To find the most salient set, the state-of-the-art EEG features are reviewed and extracted in this study. Five well-known EC algorithms are adopted for feature selection to find the salient set of EEG features and improve the performance of emotion classification. The performance of the proposed framework is evaluated on two public datasets and

our collected dataset to demonstrate the benefit of using evolutionary algorithms to identify the salient EEG features and improve the performance of emotion classification.

RQ2. **How can we optimize the performance of emotion classification system based on physiological signals?**

Research objective 2. A new emotion classification framework for optimizing LSTM hyperparameters using differential evolution (DE) is presented to maximize emotion classification performance. The goal here is to choose appropriate values using the DE algorithm for the LSTM hyperparameters to enhance emotion classification, since there is no generic optimal values to use for different problems. The performance evaluation and comparison of the proposed model with state-of-the-art hyperparameter optimization methods (particle swarm optimization (PSO), simulated annealing (SA), Random search and Tree-of-Parzen-Estimators (TPE)) using a dataset collected from wireless wearable sensors (Emotiv and Empatica E4) show that the performance of our proposed method surpasses that of the other methods.

RQ3. **How can physiological signals be fused to capture temporal emotional changes and improve emotion recognition?**

Research Objective 3. To fuse physiological signals (EEG and BVP signals) temporally, a multimodal deep learning model is proposed to build an accurate automatic emotion recognition system in an end-to-end learning fashion. This approach is able to improve the performance of emotion recognition based on capturing the non-linear correlation within and across physiological signals over time. The proposed model is able to jointly learn the highly correlated emotional structure of EEG and BVP signals after learning each channel data with single deep network. The proposed models based on early and late fusion are evaluated and compared with handcrafted features using our dataset collected from portable wearable sensors to identify the efficacy of temporal multimodal learning on the improvement of multimodal emotion classification.

1.4 RESEARCH FRAMEWORK

The key components of this research are shown in Figure 1.1. The research involves data collection, modelling and validation of the models. In the data collection step, a dataset was collected using the wireless wearable physiological sensors (smart wristband and headset) in a controlled environment.

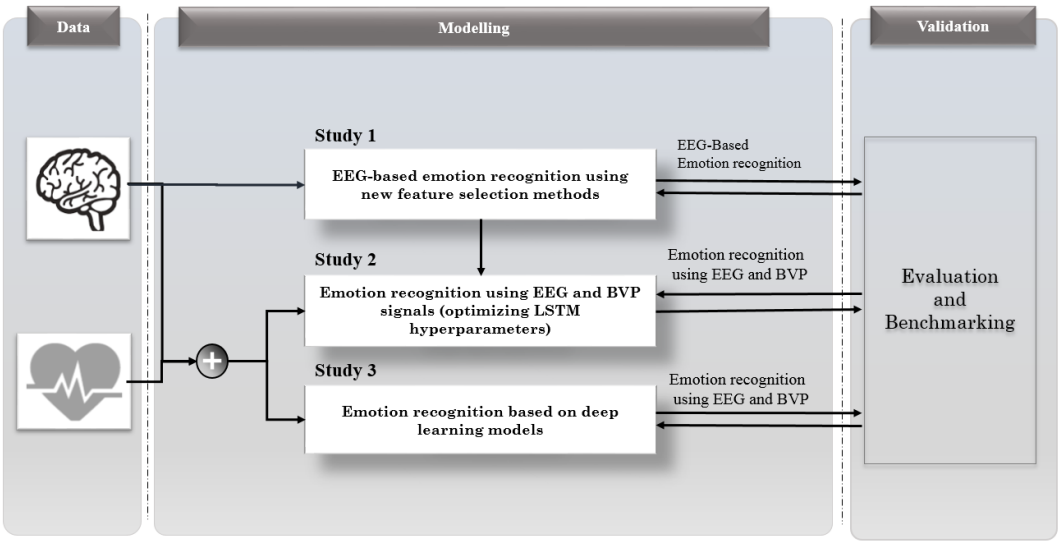


Figure 1.1. The framework of this research

The collected data included EEG signals, BVP signals and profile data. Moreover, in this research, two other public datasets (MAHNOB and DEAP) are also used. The modelling step contains a set of learning algorithms to improve the performance of emotion recognition. In the first model (Study 1), evolutionary algorithms for feature selection algorithms are proposed to improve the performance of EEG-based emotion recognition based on the selected salient set of features. Study 2 focused on optimizing LSTM hyperparameters and finding the best possible LSTM networks to maximize the performance of emotion recognition using wearable physiological sensors (EEG and BVP signals). In study 3, we proposed a new framework based on temporal multimodal deep learning models to fuse EEG and BVP signals and compared the performance of the proposed models with the findings from Study 2.

1.5 CONTRIBUTION OF THIS THESIS

This research extends knowledge, method, and techniques in the field of affective computing. Each contribution and the related publications are listed in the following section.

1) **Investigate the performance of evolutionary algorithms for feature selection techniques using EEG-based emotion recognition:** This research starts with a the state-of-the-art systematic review of feature extraction from EEG signals, and proposes a new framework using evolutionary computation algorithms (Ant Colony Optimization (ACO), Simulated Annealing (SA), Genetic Algorithm (GA), Particle Swarm Optimization (PSO) and Differential Evolution (DE)) to find the salient feature sets and channels and overcome the high-dimensionality problem. The proposed framework has been extensively evaluated with two public datasets using an EEG sensor with 32 channels and our new dataset collected from wireless EEG sensors with 5 channels. The results confirm that evolutionary algorithms can effectively support feature selection to identify the best EEG features and the best channels to maximize performance over a four-quadrant dimensional emotion (HA-P, HA-N, LA-P and LA-N) classification problem. These findings are significant for informing future development of EEG-based emotion classification because low-cost mobile EEG sensors with fewer electrodes are becoming popular in many new applications. Moreover, the combination of time and frequency features consistently showed more efficient performance, compared to using only time or frequency features. The most frequently selected source of features (i.e. the EEG channels) were analysed using the five evolutionary algorithms and weighted majority voting. The electrodes in the frontal and central lobes were shown to be activated more during emotions based on two public datasets, confirming the feasibility of using a lightweight and wireless EEG sensor (Emotiv Insight) for four-quadrant emotion classification.

   Related publication:

   1. B. Nakisa, M. N. Rastgoo, D. Tjondronegoro, and V. Chandran, “Evolutionary Computation Algorithms for Feature Selection of EEG-based Emotion Recognition using Mobile Sensors,” Expert Systems with Applications, 2017. (Q1 Journal, Published)

2) **Optimizing LSTM hyperparameters to improve the performance of emotions recognition based on physiological signals (EEG and BVP signals):** This research is the first study that proposed the use of hyperparameter optimization techniques in the context of affective computing. We proposed a new framework to automatically optimize LSTM hyperparameters (batch size and number of hidden neurons) using the DE algorithm. In this study, we evaluate and compare the proposed framework with other state-of-the-art hyperparameter optimization methods (Particle Swarm Optimization, Simulated Annealing, Random Search and Tree-of-Parzen-Estimators (TPE)) using our dataset collected from wearable sensors (EEG and BVP signals). Experimental results demonstrate that optimizing LSTM hyperparameters significantly improves the recognition of four-quadrant dimensional emotions, with a 14% increase in accuracy. The best model based on the optimized LSTM classifier using the DE algorithm achieved 77.68% accuracy. The results also showed that evolutionary algorithms, particularly DE, are competitive for ensuring the optimized LSTM hyperparameter values.

Related publication:

2. B. Nakisa, M. N. Rastgoo, A. Rakotonirainy, F. Maire and V. Chandran," Long Short Term Memory Hyperparameter Optimization for a Neural Network Based Emotion Recognition Framework" IEEE Access, 2018. (Q1 journal, Published).

3) **End-to-end temporal multimodal learning models based on early and late fusion to classify emotions using raw EEG and BVP data:** We proposed a new framework to fuse physiological signals (EEG and BVP signals) based on multimodal learning approach. This approach is able to improve the performance of emotion recognition based on capturing the non-linear correlation within and across physiological signals over time. The proposed framework used convolutional neural network (ConvNet) and LSTM network in an end-to-end fashion. Using an end-to-end learning approach, the network is trained from the raw data without any a priori feature extraction. To our knowledge this is the first work that applies such an end-to-end temporal multimodal fusion model for emotion recognition based on EEG and BVP signals. The proposed framework is investigated based on early and late fusions approaches. The performance of the

two temporal multimodal deep learning models with different window sizes are evaluated and compared with handcrafted features. The results show that temporal multimodal deep learning models can slightly outperform the accuracy of models using handcrafted features in recognizing four-quadrant dimensional emotions on a dataset collected from wireless wearable sensors. Moreover, the accuracy of the proposed framework based on early fusion is higher than based on late fusion.

Related publication:

3. B. Nakisa, M. N. Rastgoo, A. Rakotonirainy, F. Maire and V. Chandran," Automatic Emotion Recognition Using Temporal Multimodal Deep Learning " Expert Systems with Applications, 2018. (Q1 Journal, To be submitted)

1.6 SIGNIFICANCE

The methods developed in this thesis collectively deliver better algorithms and maximize the use of wireless wearable physiological sensors to recognize four-quadrant dimensional emotions. This research is significant in a number of ways:

1. It enhances techniques based on the traditional emotion classification approaches,

   1.1 to introduce a new feature selection method for EEG-based emotion recognition

   1.2 to propose a new framework to optimize LSTM hyperparameters and find an optimized LSTM classifier using EEG and BVP signals to improve emotion classification.

2. It extends the knowledge of the impacts of advanced machine learning techniques (deep learning) on emotion classification

   2.1 to temporally fuse EEG and BVP signals based on temporal multimodal deep learning models using ConvNets LSTM networks.

Overall, this research provides new, more accurate methods for emotion classification based on EEG and BVP signals. The research in this thesis

supports potential applications in a wide range of areas, including human–computer interaction (HCI), intelligent tutoring, mental-health care and transportation safety. The use of miniaturized wearable sensor technology using the proposed methods assists in the development of an emotionally intelligent HCI system which can sense and respond appropriately to user's emotional states. An emotion classification system can be applied to improve mental health care and to detect fatigue and stress (Leape, Fong, & Ratwani, 2016). In the classroom, detecting negative emotional states of students based on portable physiological sensors can help to improve student learning experiences and enhance their performance (Harrison, 2013). In the field of transportation safety, detecting different emotions such as stress, anger and fatigue can help to issue a warning to the driver of a vehicle before a possible crash.

1.7 THESIS OUTLINE

This section provides an overview of how the thesis is organized.

Chapter 1 presents the introduction. It briefly describes the background, problems, aims, objectives and significance of this research.

The literature review is presented in *Chapter 2*. This is a review of existing works on emotion classification using physiological signals. This chapter also summarizes and discusses the research gaps related to methods and techniques of feature extraction and selection, learning models, and fusion models. In addition, the research gaps based on the current literature are also summarized.

Chapter 3 presents a review of the state-of-the-art EEG features and proposes a new framework using evolutionary algorithms to automatically find the salient set of EEG features (Paper 1). *Chapter 4* presents a framework to optimize LSTM hyperparameters using DE algorithms and reveals an improvement in the performance of emotion classification (Paper 2). In *Chapter 5*, a new framework for the fusion of EEG and BVP signals using deep learning techniques (ConvNet LSTM networks) to temporally capture the emotional related information within and across modalities and automatically classify dimensional emotions (Paper 3).

Chapter 6 concludes the dissertation with a summary of the research and possible future directions.

Chapter 2: Literature Review

2.1 INTRODUCTION

This chapter provides an overview of the background to the research aim, which is to develop techniques to improve the performance of emotion classification using portable wearable physiological sensors. In addition, this chapter aims to provide a more comprehensive overview of the research methodologies undertaken as the part of this program of research than is typically possible within the scope of a journal article or conference paper methods section.

Section 2.2 provides an overview of different emotion models, emotion elicitation, annotation, ground truth and hot to measure emotional states using physiological signals. Section 2.3 provides an overview of emotion classification using physiological signals. Section 2.4 reviews the current methods for feature selection in the affective computing domain. Section 2.5 introduces the advanced classifiers that are applied to emotion classification systems using physiological signals and presents the current research gaps in hyperparameter optimization for the classifiers. Section 2.6 presents multimodal deep learning models based on end-to-end learning and the current gap in knowledge, to establish the motivation for this research. At the end of this chapter is a summary of the key limitations that will be addressed, and each of the subsequent chapters will provide further insights into the current literature with regard to the specific problems addressed.

2.2 SCIENTIFIC PERSPECTIVE ON EMOTION

2.2.1 Emotion Models

Emotions are an integral aspect of human life. In the past decades, many basic studies have been conducted in an attempt to establish a unique definition for emotion and to define a unique set of basic emotions that is acceptable among all theorists, but

as yet, a universal set of basic emotions has not been defined. Some theorists believe that emotion cannot be detected directly and emotion recognition is possible by expressive behaviour, physiological indices, self-assessment and context. Based on the psychologists' notion, the onset of an emotion is associated with stimuli, and feeling and emotions do not happen in isolation. According to the psychologists. there are two major models: categorized models (Ekman, 1992), and dimensional models (Russell, 1980).

Several theorists have conducted experiments to recognize basic emotions and have proposed some sets of categorized models. A theory of emotion is proposed by Darwin (1965) and then interpreted by Tomkins (1962). Tomkins proposed that the basic emotions consist of nine emotions: interest-excitement, enjoyment-joy, surprise-startle, anger-rage, disgust, dissmell, distress-anguish, anger-rage, and fear-terror. It is believed that these nine basic emotions play an important role in optimal mental health.

Another theory which is now widely used is the Ekman model. Ekman and his colleagues decided that there are six basic emotions, sadness, happiness surprise, fear, anger and disgust that are universally distinguishable by facial expression. Ekman (1976) used pictures from the Pictures of Facial Affect (POFA) collection as a stimulus. This collection consisting of images of actors portraying these six basic emotions plus a neutral emotion is universally cross-culturally recognizable. Many theorists and psychologists proposed other emotions in their sets of basic emotions that are different from the six basic emotions proposed by Ekman. Some of them categorized emotions into small groups (Gray, 1985; James, 1884; Mowrer, 1960; Panksepp, 1982; Watson et al., 1925; Weiner & Graham, 1984), which focused more on general emotions like fear or rage (as negative emotions) and happiness or love (as positive emotions), while other psychologists focused on more details and classified emotions into bigger sets. A summary of some of the basic emotion models is shown in Table 2.1.

Table 2.1: Summary of Categorized Emotions Models

Reference	**Emotions**
(Ekman & Oster, 1979)	Fear, sadness, happiness, anger, disgust, and surprise

(Arnold, 1960)	Anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, sadness
(Panksepp, 1982)	Expectancy, rage, fear, panic
(Tomkins, 1962)	Surprise, interest, joy, rage, fear, disgust, shame, and anguish.
(Johnson-Laird & Oatley, 1989)	Happiness, sadness, fear, anger, and disgust
(Frijda, 1986)	Desire, happiness, interest, surprise, wonder, sorrow
(Gray, 1985)	Rage and terror, anxiety, joy
(Izard, 1977)	Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise
(James, 1884)	Fear, grief, love, rage
(McDougall, 2003)	Anger, disgust, elation, fear
(Weiner & Graham, 1984)	Sadness, happiness
(Mowrer, 1960)	Pain, pleasure
(Watson, 1925)	Fear, love, rage

On the other hand, some theorists believe that emotions do not exist as discrete circuits and they believe that the category-based paradigm has some limitations. One of these proposed limitations is that affective states involved in everyday life are too complex to be well represented by a limited number of discrete categories, but unfortunately, augmenting the number of possible labels complicates the annotation process and lowers inter-annotator agreement (Cowie & Cornelius, 2003). Therefore, they propose another method which is called dimensional emotion. In this model, it is proposed that emotions arise from two or three independent physiological systems one of which is related to valence (from a pleasant state to an unpleasant state), another is related to arousal (from a calm state to an excited state) and the last but not the least is related to power (intensity of emotions). The aim of dimensional models is to conceptualize human emotion in two or three dimensions.

In two dimensional models, categorized emotions can be determined by arousal and valence. Arousal is the vertical axis and valence is the horizontal axis and the centre of the circle represents the medium level of arousal and a neutral valence. Earlier psychologists argued that many emotional states that more easily recognizable are placed on valence dimensions from bad, unpleasantness, to good.

Another widespread dimensional model, the "Circumplex model" proposed by Russell (1980), has two dimensions, namely, arousal and valence (Figure 2.1). In this model, all emotions can be placed on a circumplex of this space and at any level of arousal and valence.

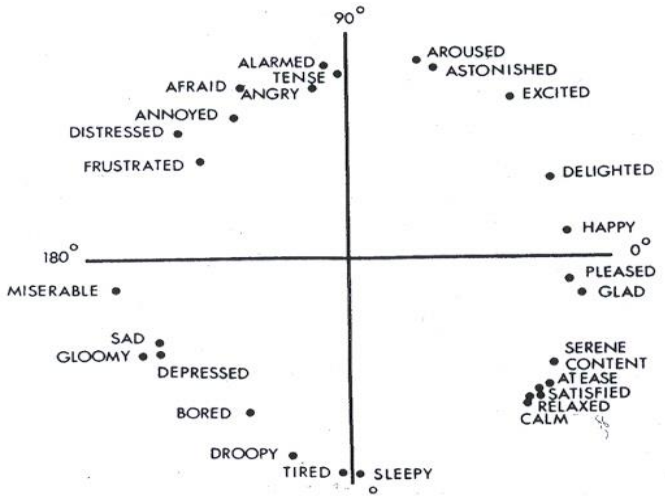


Figure 2.1. Circumplex model of emotion.

A number of researchers have shown that the affective states in everyday interactions between people are non-basic and subtle like depression; therefore, a single label may not reflect the complexity of the affective state in our daily interactions. As a result, the model chosen for this study is the dimensional emotion model.

2.2.2 Emotion Elicitations, Annotation and Ground Truth

Generally, emotion is a body reaction which is an internal and external response towards any stimulus which is deliberately induced or naturally elicited (Douglas-Cowie et al., 2011; Valstar & Pantic, 2010). There are two types of stimulus, namely, induced expression and naturalistic expressions. Induced emotions refer to types of emotions caused by stimuli that are deliberately chosen to induce different emotions in subjects, while naturalistic expressions refer to natural situations and stimuli that are out of control. On the other hand, instead of evoking emotions in subjects, some

studies have used posed expressions which means specific emotions are intentionally expressed by selected subjects in a controlled laboratory environment.

There are number of limitations related to naturalistic and posed expressions for emotion recognition and due to the difficulties in collecting and annotating these in a real environment, these expressions are outside the scope of this research. The induced expressions are very popular in the emotion recognition research; however, Picard, Vyzas, and Healey (2001) stated that the main concern of emotion elicitation is to choose an appropriate stimulus to induce that specific emotion. Different stimuli, such as events, images (Lang, Bradley, & Cuthbert, 2008), music (Lichtenstein, Oehme, Kupschick, & Jürgensohn, 2008) or movies (Soleymani, Pantic, & Pun, 2012), have been used to trigger emotions. But the subject's awareness of the purpose of the experiment might have an impact on the reliability of the elicited data.

Another challenge of emotion detection is in determining the ground truth of the emotion elicitation trial or target emotion (Kim et al., 2004) through annotation. Obviously, it is hard to determine the ground truth for emotion, since there is no clear definition of emotions, but the best way to determine or annotate felt emotion during the experiment is the subjective rating of emotional trials or self-report. Subjective ratings of emotional trials, for example, are used widely by researchers (Bradley & Lang, 1994). Advocates of self-report believe that participants are in a privileged position to assess and annotate their emotions (Larsen & Fredrickson, 1999). Self-reports can be collected from participants using questionnaires or specially designed tools. The Self-Assessment Manikin (SAM) is one such tool which is designed to assess subjects' emotional experience (Bradley & Lang, 1994).

This self-report tool has been used effectively to measure emotional responses in a variety of situations including reactions to pictures, sounds and other stimuli (Bradley & Lang, 1994). Another way to determine ground truth is to use an expert annotator, based on a frontal video of the participants face. This is useful for continuous emotion recognition. In this type of annotation, the annotator to conduct continuous annotations of the participant's facial expressions. The annotators would be able to determine the arousal and valence of the user's emotion continuously using FEELTRACE software (Cowie et al., 2000). In this study, I will use the self-assessment Manikin (SAM) to identify basic emotions which are collected based on

the perceived emotion, and then I will classify and map different emotions into the quadrants of the four-quadrant dimensional model (High Arousal-Positive emotion, High Arousal-Negative emotion, Low Arousal-Positive emotion, Low Arousal-Negative emotion).

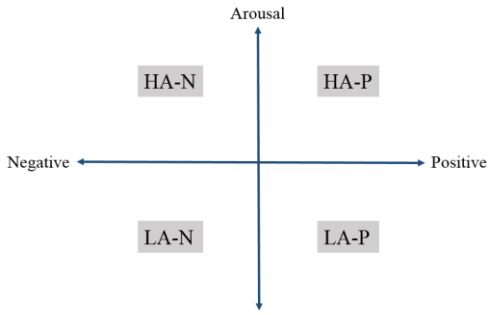


Figure 2.2. The emotions categorized into four-quadrant dimensional emotions

2.2.3 Measuring emotional states using physiological signals

Over the past few decades, research has shown that human emotions can be monitored through physiological signals like EEG, BVP and GSR (Koelstra et al., 2010) and physical indicators such as facial expression (Hossain & Muhammad, 2017). Physiological signals offer some advantageous in comparison with physical modalities such as their suitability for inner feelings and insusceptibility to social masking of emotions (Kim, 2007). In other words, physiological signals are constantly emitted and controlled by the central and autonomic nervous systems which makes it harder for subjects to fake their responses (Peter, Ebert, & Beikirch, 2009). This might be important for some certain applications such as deception detection, lie detection and so on. To understand the inner emotions of humans, emotion recognition methods focus on changes in the two major components of nervous system: the Central Nervous System (CNS) and the Automatic Nervous system (ANS). Physiological responses such as Galvanic Skin Response (GSR), heart activity, respiration, brain activity and skin temperature originating from the CNS and ANS carry information relating to inner emotional states. Among these physiological responses, heart activity and brain activity are good indicators of emotion recognition (Ackermann, Kohlschein, Bitsch, Wehrle, & Jeschke, 2016; Agrafioti, Hatzinakos, & Anderson, 2012).

These physiological signals can be captured using two types of sensors: tethered-laboratory sensors and wireless physiological sensors. Tethered-laboratory sensors, which are normally lab-based, capture strong signals with high resolution. However, they are more invasive and cannot be used for everyday life situations. Wireless physiological sensors can provide a non-invasive and non-obtrusive way to collect physiological signals and can be utilized while the subject is carrying out their daily life activities. However, the resolution of signals collected from these sensors is lower than that of tethered-laboratory sensors.

The activity of the brain, which is measured by EEG signals, can be recorded by either laboratory EEG sensors or wireless EEG headsets. The laboratory EEG sensors with large numbers of channels are usually connected by gel to the scalp which is more invasive and not applicable for outside the lab. Most of the studies in affective computing use the laboratory EEG sensors, whereas, only a few studies used the wireless EEG sensors.

The heart activity can be presented by an electrocardiogram (ECG) or photoplethysmogram (PPG). ECG signals via leads connected to the human chest, arm and leg, presents the fluctuation in heart activity over a period of time. PPG uses a light-based technology to determine the changes in blood volume in blood vessels as controlled by the heart pumping action. PPG is a non-invasive and low-cost technique which has recently been embedded in smart wristbands. The usefulness of these wearable sensors has been proven in applications such as stress prediction (Ghosh et al., 2015; Rastgoo, Nakisa, Rakotonirainy, Chandran, & Tjondronegoro, 2018) as well as emotion recognition (Haag et al., 2004).

2.3 DATASETS

To evaluate the performance of the proposed frameworks, we used two public datasets (MAHNOB and DEAP) and our dataset collected from portable physiological sensors (Emotiv and Empatica E4). In the next subsections, each dataset is described in more details.

2.3.1 MAHNOB Dataset

MAHNOB dataset (Soleymani, Lichtenauer, Pun, & Pantic, 2012) is a multimodal data that is recorded in responses to affective stimuli. This dataset contains a recording of user responses such as face video, physiological signals, eye gaze and audio signals to multimedia content. The Biosemi active II system with active electrodes was used to record physiological signals. Physioloigical signals such as EEG (32 channels), ECG, respiration amplitude and skin temperature were recorded in this dataset. While each participant was watching the videos, EEG signals using 32 channels based on a 10/20 system of electrode placement were collected. The sampling rate of the recording was 1024Hz, but it was down-sampled to 256Hz afterwards. 30 healthy young people, aged between 19 and 40 years old, from different cultural backgrounds, volunteered to participate. Fragments of videos from online sources, lasting between 34.9 and 117s, with different content, were selected to induce 9 specific emotions in the subjects. After each video clip the participants were asked to describe their emotional state using emotional label, arousal, valence, dominance and predictability. The emotional labels included neutral, anxiety, amusement, sadness, joy, disgust, anger, surprise and fear. The arousal levels (1-9 point scales) ranges from calm/bored to excited and the valence level ranges from sad/unpleasant to happy/joy. The dominance scales ranges from without control to in control and predictability levels describes to what extent the sequence of events is predictable for a participants.

In this thesis we only used EEG signals (32 channels) to evaluate the performance of the proposed methods in the next chapters and compare with other datasets.

2.3.2 DEAP Dataset

DEAP (Koelstra et al., 2012) is a multimodal dataset that contains EEG signals with 32 channels, other physiological signals and frontal face video. EEG and other peripheral signals are recorded using a Biosemi ActiveTwo system EEG signals were recorded with a sampling rate of 512Hz. They recorded data from 32 participants aged between 19 to 37 years old. The physiological signals are recorded while the

participants were listening to 40 one-minute music videos. The stimuli were selected to induce emotions in four-quadrant of dimensional emotion model. The experiment started with 2 minutes baseline recording which is asked to relax during this period. At the end of trial (each music video), it is asked to fill the self-assessment for arousal, valence, liking and dominance. After 20 trials (20 music videos), the participants took a short break. Self-assessment Manikin (SAM) is used for this dataset to visualize the scales of the felt emotions between one and nine. The arousal scale ranges from calm/ bored to excited and the valence scales ranges from unhappy/ sad emotions to happy/ joyful. The dominance scale ranges from without control to in control and the liking scale asked the participants' personal liking of the video. This scale measures the participants' tastes, not feeling. In this thesis we only used arousal and valence scales to classify four-quadrant dimensional emotion. Among the physiological signals. We only used EEG signals in the first study (Chapter 3) for evaluation.

2.3.3 Our Dataset

The third dataset was newly collected from 20 subjects, aged between 20 and 38 years old, while they watched video clips. This dataset collected in our lab in a controlled environment. EEG and peripheral physiological signals (BVP, and skin Temperature) were recorded using portable physiological sensors. To collecting the EEG data and other physiological signals (BVP and skin temperature), the *Emotiv Insight wireless* headset and the *Empatica E4 wristband* were used respectively (see figure 2.3). The Emotiv headset contains 5 channels (AF3, AF4, T7, T8, Pz) and 2 reference channels located and labeled according to the international 10-20 system (see figure 3.3). Compared to the EEG recording devices that were used in MAHNOB and DEAP, our data collection used Emotiv wireless sensor, which only captures EEG signals from five channels (instead of 32). The Empatica wristband collects the BVP (heart activity) as well as skin temperature in a less invasive way.

(a) (b)

Figure 2.3. (a) The Emotiv Insight headset, (b) The Empatica E4 wristband.

TestBench software and Empatica Connect were used for acquiring raw EEG and BVP signals from the Emotiv Insight headset and Empatica respectively. Emotions were induced by video clips, used in the MAHNOB dataset, and the participants' brain and heart responses were collected while they were watching 9 video clips in succession. It should be noted that the used video clips derived from MAHNOB dataset induce one specific emotion into human. Therefore, the physiological signals (EEG and BVP signals) in each video clip present one specific emotion.

The participants were asked to report their emotional state after watching each video, using a keyword, arousal and valence levels. The basic emotion keywords such as neutral, anxiety, amusement, sadness, joy or happiness, disgust, anger, surprise, and fear are used in this dataset. The arousal levels ranges from calm to excited and the valence level ranges from unpleasant to pleasant. Before the first video clip, the participants were asked to relax and close their eyes for one minute to allow their EEG and BVP baseline to be determined. Between each video clip stimulus, one minute's silence was given to prevent mixing up the previous emotion. The experimental protocol is shown in figure 2.4.

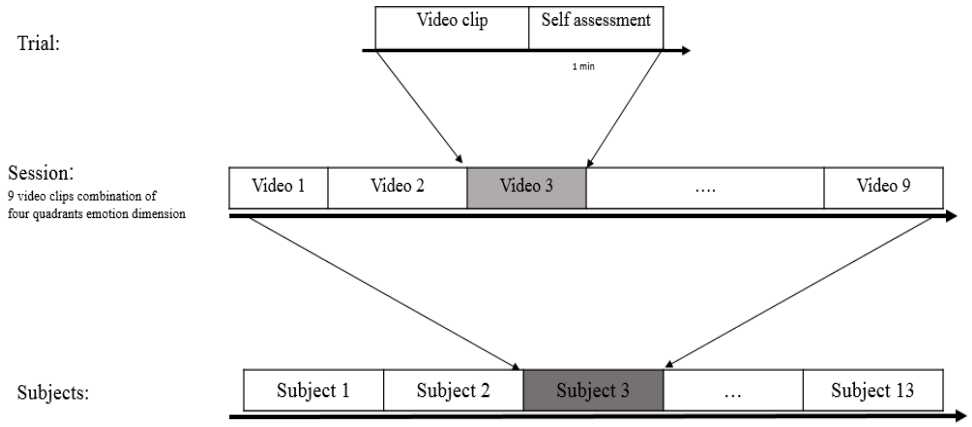


Figure 2.4 Illustration of the experimental protocol for emotion elicitations.

The experiment with this new data allows an investigation into the feasibility of using the Emotiv Insight sensor and Empatica E4 for emotion classification purposes. The expected benefit of these sensors is due to its light-weight, and wireless nature, making it possibly the most suitable for free-living studies in natural settings.

2.4 GENERAL FRAMEWORK FOR EMOTION RECOGNITION USING EEG AND BVP SIGNALS

To recognize different emotions using physiological signals, there are four main steps: pre-processing, feature extraction, feature selection and learning algorithms.

Since the physiological data contains a lot of noise and artefacts, in the pre-processing step the raw data are synchronized and then cleaned. Since the physiological signals in each video clip present one specific emotion, we segmented and labelled the physiological data with the emotion corresponding to the experimental phase in which they occurred.

The raw sensor data may contain out-of-range values and/or missing values. Improper screening of this data prior to the analysis phase can produce misleading results. Thus, pre-processing to ensure the quality of data before running an analysis is important. The collected sensor data is verified against the sensor's standard

distribution or compared with data collected using gold-standard equipment. It is also checked to ensure the sensor is properly calibrated as per the instructions. EEG and BVP signals are usually pre-processed using different noise reduction methods such as Average Mean Reference (AMR) (M. Murugappan et al., 2007), Butterworth Filter (Hettich et al., 2016; Khosrowabadi & bin Abdul Rahman, 2010), and Independent Component Analysis (ICA) (Hsu, 2013).

The next step after pre-processing and noise reduction is feature extraction. The feature extraction step has become an important and often essential step in a machine learning process. In essence, the process tries to determine the most relevant set of features for differentiating affective states through the use of an optimization criterion. To extract features from the raw signal data, first the raw signal is segmented into sequences of consecutive windows, then a set of features is extracted from each window. Some studies have shown that a one-second window size performs well and is sufficient for capturing emotions for emotion recognition purposes (Le & Provost, 2013; Soleymani, Asghari-Esfeden, Fu, & Pantic, 2016).

There are different methods to extract features from the EEG and BVP signals. Handcrafted feature extraction methods are based on statistical techniques and require expert knowledge. To recognize different emotions using BVP signals, some features from the time and frequency domains are used (Akselrod et al., 1981; Gunes & Pantic, 2010; Haag et al., 2004). There are a large number of studies which focus on extracting features from EEG signals (Jenke, Peer, & Buss, 2014) and the extracted features from these signals are generally divided into time, frequency and time-frequency domains.

Feature selection techniques attempt to find a subset of n features out of m features ($m > n$) to enhance the classification performance. Feature selection (feature reduction) helps to reduce the instances of irrelevant features and reduce the effects of high-dimensionality, thus improving the performance of the classifier. From a machine learning point of view, if a system uses irrelevant variables, poor classification accuracy will result.

In the next step, the derived features are used as inputs to a learning algorithm to classify different emotions. In previous studies, algorithms employed a range of conventional classifiers such as ANN, SVM and KNN to advanced classifiers such as Recurrent Neural Network (RNN), Deep Belief Network (DBN) and LSTM, in

emotion recognition using physiological signals (Chen & Jin, 2015; Li, Li, Zhang, & Zhang, 2013; Soleymani et al., 2016). Since physiological signals consist of time-series data with variation over a long period of time and dependencies within shorter periods, to capture the inherent temporal structure within the physiological data and to recognize emotion signatures which are reflected in short period of time, we need to apply a classifier which considers temporal information. To date, most of the studies have focused on recognizing emotions using the conventional classifiers and only a few studies have utilized the advanced machine learning techniques on emotion recognition using physiological signals.

Automated emotion recognition has been improved when different modalities have been used. The fusion of multimodal data can provide additional information and thus an increase in accuracy of the overall result or decision. It has been shown that a combination of physiological signals can improve the performance of emotion classification compared to the use of a single physiological signal. There are several approaches to fuse physiological signals (Brady et al., 2016a; Gunes & Piccardi, 2005).

As a whole, building an emotion classification system to accurately classify different emotions using EEG and BVP signals is a challenging process.

2.5 FEATURE SELECTION METHODS

A multitude of studies have focused on extracting different EEG features from time, frequency and time-domain frequency to accurately classify different emotions. However, a few researchers have investigated the performance of EEG signals in relation to emotion recognition using the combination of time domain, frequency domain and time-frequency domain features. Based on the literature, there is no standard set of EEG features that can be used discriminate different emotions. On the other hand, combining all features from different types of EEG signals may lead to a high dimensionality problem. In addition, not all the features carry significant information regarding emotions and redundant features increase the feature space, making pattern detection more difficult, and increasing the risk of overfitting. It is, therefore, important to identify the salient features that have a significant impact on the performance of the emotion classification model. Feature selection methods have

been shown to be effective in reducing high dimensionality by removing redundant and irrelevant features and maximizing the performance of classifiers. Based on the literature, feature selection can be divided into two models: filter methods and wrapper methods (Figure 2.5). To review these methods, refer to (Alba, García-Nieto, Jourdan, & Talbi, 2007; Kudo & Sklansky, 2000).

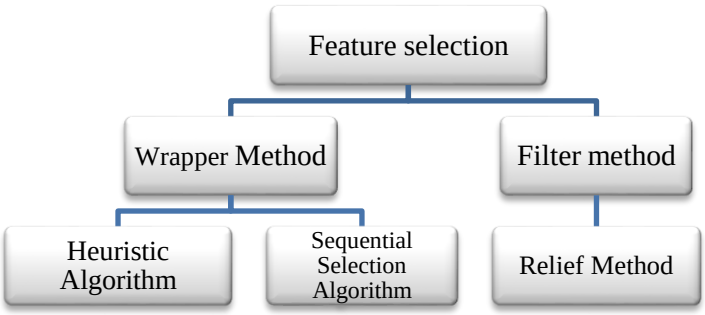


Figure 2. 5: Classification of feature selection methods.

Generally, filter methods are fast due to the fact that they select the most relevant features from the training data and then discard certain features based on the specific threshold. A wrapper method evaluates the quality of each subset of features through a learning algorithm (a classifier, or a clustering algorithm). However, wrapper models are computationally intensive, which restricts their application to huge datasets, where their aim is to improve accuracy. There are more details about the types of feature selection methods in the following sections.

2.5.1 Filter Methods

Filter methods use different ranking techniques, selected due to their simplicity and success in different applications, to order the features. Ranking methods score each feature based on its relevance and use a threshold to remove features below the threshold. Ranking methods are filter methods since they are applied before classification to filter out the less relevant variables. Several publications (John, Kohavi, & Pfleger, 1994; Langley 1994) have presented various definitions and measurements for the relevance of a variable. One definition which will be useful for

the following discussion is that "A feature can be regarded as irrelevant if it is conditionally independent of the class labels". The relevance of features will be measured by different techniques such as the Pearson Correlation Coefficient of the Mutual Information (MI) technique (Battiti, 1994). Some researchers have applied filtering methods to find the most relevant features to discriminate different emotions (Zhang & Zhao, 2008).

The main drawback of this method is that it assumes the features are independent of each other. This can cause two problems:

• Features discarded because they are not individually relevant may become relevant when considered with some other features.

• Features regarded individually as relevant may cause unnecessary redundancies.

2.5.2 Wrapper method

Wrapper methods are a type of feature subset selection methods, whereby, the suboptimal subsets are found by applying heuristic search algorithms. There are a number of search algorithms that can be used to find the subset of variables which maximize the classification performance. For larger datasets, exhaustive search methods can very intensive; thus, a simplified algorithm such as sequential search or an evolutionary algorithm such as Particle Swarm Optimization (PSO) (Kennedy, 2011) or the Genetic Algorithm (Goldberg & Holland, 1988) which yield local optimum results can be employed, as they can produce good results in a reasonable time. In the next section, the wrapper method is classified into Sequential Selection Algorithms and Heuristic Search Algorithms. The Sequential Selection Algorithm starts with an empty set/ full set and adds features until the maximum objective function/classification performance is reached. To speed up the selection, a criterion is chosen which incrementally increases the objective function until the maximum performance is reached with the minimum number of features. The heuristic search algorithms evaluate different subsets to optimize the objective function. Different subsets are generated either by searching in a search space or by generating solutions to the optimization problem. One of the aims of this study is to apply a wrapper

method, particularly a heuristic method to EEG signals to improve the classification performance.

Sequential selection algorithms

The Sequential Feature Selection (SFS) algorithm (Pudil, Novovičová, & Kittler, 1994; Reunanen, 2003) starts with an empty set and adds one feature for the first step, which gives the highest value for the objective function. From the second step onwards the remaining features are added individually to the current subset and the new subset is evaluated. The individual feature is permanently included in the subset if it gives the maximum classification accuracy. The process is repeated until the required number of features are added. This is a naive SFS algorithm since the dependency between the features is not accounted for.

The SFS and Sequential Forward Floating Selection (SFFS) methods suffer from producing nested subsets since the forward inclusion was always unconditional which means that two highly correlated variables might be included if they each give the highest performance in the SFS evaluation. To avoid the nesting effect, an adaptive version of the SFFS was developed in (Somol, Pudil, Novovičová, & Paclık, 1999; Sun, Babbs, & Delp, 2006). The Adaptive Sequential Forward Floating Selection (ASFFS) algorithm used a parameter r which would specify the number of features to be added in the inclusion phase which was calculated adaptively.

Heuristic method

In comparison with other feature selection methods, evolutionary computational (EC) algorithms show powerful global search capabilities and have been widely accepted as efficient methods for feature selection. EC algorithms can help to overcome the limitations of individual feature selection by assessing the subset of variables based on their usefulness. The main advantage of using EC algorithms to solve optimization problems is the ability to search simultaneously within a set of possible solutions to find the optimal solution, by iteratively trying to improve the feature subset with regard to a given measure of quality. The capability of these algorithms in finding the optimal or near optimal solutions is investigated in different

domains (Nakisa & Rastgoo, 2014; Nakisa, Rastgoo, Nasrudin, & Nazri, 2015; Nakisa, Rastgoo, & Nazri, 2018; Nakisa, Rastgoo, & Norodin, 2014; Niu, Chen, & Chen, 2011; Rastgoo, Nakisa, & Ahmad Nazri, 2015; Rastgoo, Nakisa, & Ahmadi, 2015; Rastgoo, Nakisa, & Najafabadi, 2014; rey Horn, Nafpliotis, & Goldberg, 1994). Five well-known EC algorithms – Ant Colony Optimization (ACO), Simulated Annealing (SA), Genetic Algorithm (GA), Particle Swarm Optimization (PSO) and Differential Evolution (DE) – are widely used for feature selection in various applications, including facial expression-based emotion recognition (Mistry, Zhang, Neoh, Lim, & Fielding, 2016) and classification of motor imagery EEG signals (Baig, Aslam, Shum, & Zhang, 2017).

2.6 CLASSIFICATION METHODS FOR EMOTION RECOGNITION AND THEIR CHALLENGES

One of the important factors in building a reliable emotion classification system is finding the best classifier which can accurately classify different emotions. A variety of classification methods have been employed in the affective computing domain for classifying affective physiological data (see Table 2.2). These classifiers range from simpler classifiers like support vector machine (SVM) with linear kernels, linear discriminant analysis (LDA), and Decision Trees to more complex and advanced classifiers like recurrent neural networks (RNNs), Long short term memory (LSTM) and so on.

Recently, deep learning algorithms have generated a great impact in signals and physiological processing. Different deep architecture models are proposed and applied to physiological signals such as EEG, electromyogram (EMG), and ECG signals and achieved comparable results compared to other conventional methods (Brady et al., 2016b; Y. Lin et al., 2010; W.-L. Zheng, Zhu, Peng, & Lu, 2014a). For example, deep belief network (DBN) is applied to EEG signals to classify three different emotions (positive, negative and neutral) and showed the superior performance of deep models over shallow network (W. Zheng & Lu, 2015; W.-L. Zheng, Zhu, Peng, & Lu, 2014). However, the learning of the DBN algorithm is difficult, especially when the training

data are limited. While having simple learning algorithm, the general Boltzmann machine are very complex to study and very slow to compute in learning.

Another type of deep learning techniques that has led to a significant improvement in recognition accuracy by modelling sequential data is Recurrent Neural Network (RNNs). RNN algorithms are able to elicit the context of observations within sequences and accurately classify sequences that have strong temporal correlation. An RNN is basically artificial intelligence with recurrent topology. In this topology there is no restriction on information flow direction. RNNs are powerful algorithms especially for processing sequential data such as sound, time series data (sensors) or written natural language. In a traditional neural network, it is assumed that all inputs (and outputs) are independent of each other. An RNN is a neural network with cyclic connections, with the ability to learn temporal sequential data. These internal feedback loops in each hidden layer allow RNN networks to capture dynamic temporal patterns and store information. A hidden layer in an RNN contains multiple nodes which generate the outputs based on the current inputs and the previous hidden states.

However, training RNNs is challenging due to vanishing and exploding gradient problems which may hinder the network's ability to back propagate gradients through long-term temporal intervals. This limits the range of context they can access, which is of critical importance to sequence data.

To overcome the gradient vanishing and exploding problem in RNNs training, LSTM networks were introduced (Hochreiter & Schmidhuber, 1997). These networks have achieved top performance in emotion recognition using multi-modal information (Chao, Tao, Yang, Li, & Wen, 2015; Chen & Jin, 2015; Nicolaou, Gunes, & Pantic, 2011; Soleymani et al., 2016).

The LSTM cells contain a memory block and gates that let the information through the connection of the LSTM. There are several connections into and out of these gates. The memory blocks contain memory cells with self-connections storing the temporal state of the network in addition to special multiplicative units called gates to control the flow of information (Hochreiter & Schmidhuber, 1997). Each memory block in the original architecture contains three gates: input gate, forget gate, and output gate. The input gate controls the flow of input activation into the memory cells. The forget gate scales the internal state of the cell before adding it as input to the cell

through the self-recurrent connection of the cell, therefore, adaptively forgetting or resetting the cell's memory. Finally, the output gate controls the output flow of cell activation into the next layer. The LSTM network transition equations are as follows:

$$i_t = \sigma(\omega^{(i)} x_t + U^{(i)} h_{t-1} + b^{(i)}) \quad (3)$$

$$f_t = \sigma(\omega^{(f)} x_t + U^{(f)} h_{t-1} + b^{(f)}) \quad (4)$$

$$o_t = \sigma(\omega^{(o)} x_t + U^{(o)} h_{t-1} + b^{(o)}) \quad (5)$$

$$u_t = tanh(\omega^{(u)} x_t + U^{(iu)} h_{t-1} + b^{(u)}) \quad (6)$$

$$i_t = i_t \odot u_t + f_t \odot c_{t-1}) \quad (7)$$

$$h_t = o_t \odot \tanh(c_t) \quad (8)$$

Where i_t, f_t and o_t are the input, forget and output gates respectively, x_t is the input at time step t and h_{t-1} denotes the function of input vectors that the network receives at time $t-1$, the ω and U terms denote the weight matrixes and b is a bias vector ($b^{(i)}$ is the input gate bias vector), σ is the logistic sigmoid function and $\odot$ is elementwise multiplication.

Although the performance of LSTM networks in classifying different emotions is promising, training these networks depends heavily on a set of hyperparameters that determine many aspects of algorithm behaviour. To find a successful LSTM classifier which can accurately classify different emotions, we need to select optimal values for its hyperparameters to achieve a state-of-the-art result. These hyperparameters range from optimization hyperparameters such as learning rate, number of hidden neurons and batch size, to regularization hyperparameters. These hyperparameters affect the quality of the model and its output and it is therefore essential to find the best possible parameters. Automatic algorithmic approaches range from simple grid searches and random searches to more sophisticated model-based approaches such as Tree-Parzen methods. Evolutionary Algorithms (EA) such as Particle Swarm Optimization (PSO) and Simulated Annealing (SA) have been shown to be very efficient in solving challenging optimization problems (Nakisa, Nazri, Rastgoo, & Abdullah, 2014;

Nakisa, Rastgoo, Nasrudin, & Nazri, 2014; Rastgoo, Nakisa, & Nazri, 2015). Of the EAs, Differential Evolution (DE) has been successful in different domains due to its capability of maintaining high diversity in exploring and finding more solutions compared to other EAs (Baig et al., 2017; Nakisa, Rastgoo, Tjondronegoro, & Chandran, 2017). However, the performance of this algorithm had not been investigated to optimize LSTM hyperparameters.

Table 2.2 Overview of some emotion classification methods that use different classifiers.

References	Physiological signals	Stimuli	Classifier	No. Classes	Results/ Accuracy
(Herbelin et al., 2005)	SC, EMG, ST, RSP, ECG	Actor	KNN	3 Valence	39%–45%
(Kim, André, Rehm, Vogt, & Wagner, 2005)	ECG, PPG, ST, SC	Audio, video and cognitive tasks	SVM	4 Sadness, Anger, Stress, Surprise	78.4%
(Chanel, Kronegg, Grandjean, & Pun, 2006)	EEG, SC, ST, RSP, BVP	IAP	Bayes Classifier	3 Arousal	72%
(Rigas, Katsis, Ganiatsas, & Fotiadis, 2007)	EMG, ECG, RSP, SC	IAPS	K-NN, Random Forest	Happiness, Disgust, and Fear	Happiness: 48% Disgust: 68% Fear: 69%
(Kim, André, & Vogt, 2009)	ECG, EMG, SC, RSP	Music	LDA	2 Arousal Valence	89%

(Lichtenstein et al., 2008)	RSP, ECG, SC, EMG, ST	Films	SVM	2 Arousal	Arousal: 82%
				2 Valence	Valence: 72%
(Jang, Park, Kim, & Sohn, 2012)	EDA, ECG, ST, BVP	Films	LDF, CART, SOM, Naïve Bayes, SVM	4 Sadness, Fear, Stress, Surprise	SVM: 100% LDA: 50.7% CART: 84% SOM: 51.2% Naïve Bayes: 76.2%
(Y.-P. Lin et al., 2010)	EEG	Music	SVM	4 Joy, Anger, Sadness, Pleasure	82.29%
(Li, Huang, Zhou, & Zhong, 2017)	EEG images	Music	RNNs	4 HA-P, HA-N, LA-P, LA-N	75.21%

(Zhang, Zheng, Cui, Zong, & Li, 2018)	EEG, Facial expressions	Films, Images	RNNs	3 Positive Negative Natural	Films: 89.5% Image: 95.4%

2.7 MULTIMODAL FUSION

Automated emotion recognition has been improved with the use of different modalities. The fusion of multimodal data can provide additional information with an increase in accuracy of the overall result or decision. However, detecting emotions using the fusion of physiological signals remains complex and challenging. Differences between sensors ranging from data type and sample rates itself make principled approaches to integrating these signals challenging. To date, there are two main levels of fusion are studies: feature-level fusion or early fusion, and decision-level fusion or late fusion. In early fusion, first different features are extracted from each modality, then all features from different modalities are concatenated to construct the joint feature vector. Finally, the joint feature vector is used to build an emotion recognition model. Decision level or late fusion refers to the approach in which the feature sets of each modality are examined and classified independently then the predicted classes from each modality are fused as a decision vector to obtain the final result. There are several studies that build emotion classification models based on early and late fusion approaches (Caridakis et al., 2007; Gunes & Piccardi, 2005; Wu, Oviatt, & Cohen, 1999; Zheng, Dong, & Lu, 2014). However, this approach is not able to capture the non-linear correlation across data modalities, as the correlation between features in each modality is stronger (Ngiam et al., 2011). This is because, these sort of approaches focus on learning the patterns within each modality separately while giving up learning patterns that occurs simultaneously across multiple data modalities.

Therefore, to build a robust emotion recognition system using multimodal physiological signals, it is essential to propose a multimodal fusion model that can capture and learn the inherent emotional changes within each modality and as well as across them. We believe that a good model for multimodal learning should simultaneously learn a joint representation of multimodal, and temporal structure within each modality.

Recent works have verified the efficiency of deep learning networks to produce useful representations for various kinds of data and obtained state-of-the-art performance (Y. Kim, Lee, & Provost, 2013b; Ngiam et al., 2011; Nguyen, Nguyen, Sridharan, Dean, & Fookes, 2018; Schirrmeister et al., 2017; Sohn, Shang, & Lee,

2014). Multimodal fusion methods have been proposed to jointly learn and explore the highly correlated representation across modalities after learning each channel data with single deep network. However, most of the works based on deep learning techniques in the literature are applied to audiovisual modalities and only a few works proposed multimodal fusion based on deep learning techniques on physiological signals.

It should be noted that physiological signals are inherently temporal in nature, which means that the current pattern in signal is influenced by the previous ones. However, the multimodal networks like deep Autoencoder, Boltzmann Machine could not model the temporal multimodal fusion.

Therefore, to build a robust emotion recognition system using multimodal physiological signals, it is essential to propose a multimodal fusion model that can temporally capture and learn the inherent emotional changes within each modality and as well as across them. We believe that a good model for multimodal learning should simultaneously learn a joint representation of multimodal, and temporal structure within each modality.

2.8 SUMMARY OF CURRENT GAPS IN THE RESEARCH

As identified in this chapter, the key gaps are provided below:

1. Although many features are extracted from EEG signals to enhance the performance of emotion classification, combining different features may cause high dimensionality and reduce the performance of emotion recognition. Deciding on a suitable feature selection method to overcome the problem of high dimensionality and finding the salient set of EEG features to improve emotion recognition is not thoroughly investigated in this domain. This study proposes a new framework to automatically search for the optimal subset of EEG features using evolutionary computation (EC) algorithms and evaluates the performance of the proposed framework on two public datasets (MAHNOB and DEAP) and our dataset collected from portable physiological sensors (Emotiv Insight and Empatica E4) and compared with the latest methods in this domain.

2. Although the performance of LSTM networks in classifying different emotions is promising, training these networks depends heavily on a set of hyperparameters that determine many aspects of algorithm behaviour. Therefore, finding the optimal hyperparameters for the LSTM classifier can improve the performance of emotion classification. This study focuses on optimizing LSTM hyperparameters using DE algorithms and evaluates the performance of the resulting classifiers on emotion classification using physiological signals (EEG and BVP signals). This is the first study that proposes a framework to optimize the LSTM hyperparameters in the affective computing domain and to investigate the performance of the optimized LSTM networks in classifying different emotions. The performance of the proposed framework is evaluated on our dataset collected from portable physiological sensors and compared with other latest works in this domain.

3. It has been shown that building an automatic emotion recognition system using multimodal physiological signals can improve the classification performance. However, the fusion of multimodal physiological signals to build an accurate emotion classification system is challenging. To date, there are several studies in the affective computing domain that focus on proposing different approaches to fuse multimodal data. However, these sort of approaches cannot capture the non-linear correlation across multimodal data, as the correlation between features within each modality is stronger. In order to improve an automatic emotion classification system based on multimodal physiological signals, it is essential to capture the non-linear emotional information within and across physiological signals over time. This study proposes a new framework to fuse physiological signals (EEG and BVP signals) based on a multimodal deep learning approach. The proposed framework uses a convolutional neural network (ConvNet) and an LSTM network in an end-to-end fashion. Using end-to-end learning, the constructed features using ConvNets are trained jointly with the classification step as a single network. Moreover, in an end-to-end learning approach, the network is trained from the raw data without any a priori feature extraction. To our knowledge this is the first work that applies such an end-to-end temporal multimodal fusion model for emotion recognition

based on EEG and BVP signals. The performance of multimodal fusion model using ConvNet LSTM network is evaluated based on early and late fusion levels. The performance of the proposed frameworks are evaluated on our dataset collected from portable physiological sensors and compared with handcrafted features.

Chapter 3: Evolutionary Computation Algorithms for Feature Selection of EEG-based Emotion Recognition using Mobile Sensors (Paper 1)

Bahareh Nakisa[1]

Mohammad Naim Rastgoo[1]

Dian Tjondronegoro[2]

Vinod Chandran[1]

1. Science and Engineering Faculty, Queensland University of Technology, Brisbane, Australia
2. School of Business and Tourism, Southern Cross University, Gold Coast, Australia

Corresponding Author:

Bahareh Nakisa

Science and Engineering Faculty

Queensland University of Technology,

Brisbane, Australia

Phone: [redacted]

Email: Bahareh.nakisa@qut.edu.au

RESEARCH STUDENTS CENTRE
Examination Enquiries: 07 3138 1839
Email: research.examination@qut.edu.au

Statement of Contribution of Co-Authors for Thesis by Published Paper

The following is the suggested format for the required declaration provided at the start of any thesis chapter which includes a co-authored publication.

The authors listed below have certified that:

1. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
2. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
3. there are no other authors of the publication according to these criteria;
4. potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit, and
5. they agree to the use of the publication in the student's thesis and its publication on the QUT's ePrints site consistent with any limitations set by publisher requirements.

In the case of this chapter:

Title and status: **Evolutionary Computation Algorithms for Feature Selection of EEG-based Emotion Recognition using Mobile Sensors (Published at Expert system with applications, Q1 Journal)**

Contributor	**Statement of contribution**

Bahareh Nakisa	**Candidate** Experimental design, conducted the experimental work, performed data analysis, interpreted results, wrote the manuscript
Mohammad Naim Rastgoo	Assisted with paper planning, editing and proof-reading the paper
Dian Tjondronegoro	**Associate Supervisor (External)** Assisted with paper planning, editing and supervise the experimental work and provide advice.
Vinod Chandran	**Associate Supervisor (External)** Assisted with paper planning, editing and supervise the experimental work and provide advice.

Principal Supervisor Confirmation		
I have Sighted email or other correspondence from all Co-authors confirming their certifying authorship. Andry Rakotonirainy		
Name	Signature	Date

3.1 INTRODUCTORY COMMENTS

To build an accurate and reliable emotion recognition system based on physiological signals, one of the main steps is to find the salient features to feed into a classifier particularly for EEG signals. As such, this chapter presents a new framework using evolutionary algorithms for feature selection (study1) to improve the performance of EEG-based emotion recognition based on the selected set of EEG features. To find the salient set of EEG features, in this study, we reviewed and implemented most of the state-of-the-art EEG features (time, frequency and time-frequency). This study, showed the effectiveness of evolutionary algorithms for feature selection to improve the performance of EEG-based emotion classification.

This chapter seeks to address Research question 1: "What impacts do feature selection algorithms have on emotion recognition system?". As such, this chapter seeks to identify which sort of features can contribute more in developing an accurate emotion classification based on EEG signals. The finding of this study will be complementary for the next study (study 2).

Taken from: B. Nakisa, M. N. Rastgoo, D. Tjondronegoro, and V. Chandran, "Evolutionary Computation Algorithms for Feature Selection of EEG-based Emotion Recognition using Mobile Sensors," Expert Systems with Applications, 2017.

Publication status: Available online, 2 Nov 2017

Journal Quality: Expert Systems With Applications is a refereed international journal whose focus is on exchanging information relating to expert and intelligent systems applied in industry, government, and universities worldwide. The journal impact factor is **3.9,** and rank Q1 (SCImago) in Artificial Intelligence, Computer science applications and Engineering.

3.2 ABSTRACT

There is currently no standard or widely accepted subset of features to effectively classify different emotions based on electroencephalogram (EEG) signals. While combining all possible EEG features may improve the classification performance, it can lead to high dimensionality and worse performance due to redundancy and inefficiency. To solve the high-dimensionality problem, this paper proposes a new framework to automatically search for the optimal subset of EEG features using evolutionary computation (EC) algorithms. The proposed framework has been extensively evaluated using two public datasets (MAHNOB, DEAP) and a new dataset acquired with a mobile EEG sensor. The results confirm that EC algorithms can effectively support feature selection to identify the best EEG features and the best channels to maximize performance over a four-quadrant emotion classification problem. These findings are significant for informing future development of EEG based emotion classification because low-cost mobile EEG sensors with fewer electrodes are becoming popular for many new applications.

3.3 INTRODUCTION

Recent advances in emotion recognition using physiological signals, particularly electroencephalogram (EEG) signals, have opened up a new era of human computer interaction (HCI) applications, such as intelligent tutoring (Calvo & D' Mello, 2010; Du Boulay, 2011), computer games (Mandryk & Atkins, 2007) and e-Health applications (Liu, Conn, Sarkar, & Stone, 2008; Luneski, Bamidis, & Hitoglou-Antoniadou, 2008). Automatic emotion recognition using sensor technologies such as wireless headbands and smart watches is increasingly the subject of research, with the development of new forms of human-centric and human-driven interaction with digital media. Many of these portable sensors are easy to set up and connect via Bluetooth to a smart phone or computer, where the data can be readily analysed. These mobile sensors are able to support real-world applications, such as detecting driver drowsiness (Li, Lee, & Chung, 2015), and, more recently, assessing the cognitive load of office workers in a controlled environment (Zhang et al., 2017).

Emotion recognition supports automatic interpretation of human intentions and preferences, allowing HCI applications to better respond to users' requirements and customize interactions based on affective responses. The strong correlation between different emotional states and EEG signals is most likely because these signals come directly from the central nervous system, providing information (features) about internal emotional states. EEG signals can thus be expected to provide more valuable than less direct or external indicators of emotion such as interpreting facial expressions.

Previous works on extraction of EEG features have demonstrated that there are many useful features from time, frequency and time–frequency domains, which have been shown to be effective for recognizing different emotions. A recent study (Jenke, Peer, & Buss, 2014) proposed the most comprehensive set of extractable features from EEG signals, noting that advanced features like higher order crossing can perform better than common features like power spectral bands for classifying basic emotions. However, their experiment did not include a publicly available dataset and cannot be directly compared with our proposed method. Moreover, there is no standardized set of features that have been generally agreed as the most suitable for emotion recognition. This leads to what is known as a high-dimensionality issue in EEG, as not all of these features would carry significant information regarding emotions. Irrelevant and redundant features increase the feature space, making patterns detection more difficult, and increasing the risk of overfitting. It is therefore important to identify salient features that have significant impact on the performance of the emotion classification model. Feature selection methods have been shown to be effective in automatically decreasing high dimensionality by removing redundant and irrelevant features and maximizing the performance of classifiers.

Among the many methods which can be applied to feature selection problems, the simplest are filter methods which are based on ranking techniques. Filter methods select the features by scoring and ordering features based on their relevance, and then defining a threshold to filter out the irrelevant features. The methods aim to filter out the less relevant and noisy features from the feature list to improve classification performance. Filter methods that have been applied to emotion classification systems include Pearson Correlation (Kroupi, Yazdani, & Ebrahimi, 2011), correlation based

feature reduction (Nie, Wang, Shi, & Lu, 2011; Schaaff & Schultz, 2009), and canonical correlation analysis (CCA). However, filtering methods have two potential disadvantages as a result of assuming that all features are independent of each other (Zhang & Zhao, 2008). The first disadvantage is the risk of discarding features that are irrelevant when considered individually, but that may become relevant when combined with other features. The second disadvantage is the potential for selecting individual relevant features that may lead to redundancies.

Evolutionary computation (EC) algorithms can help to overcome the limitations of individual feature selection by assessing the subset of variables based on their usefulness. The main advantage of using EC to solve optimization problems is the ability to search simultaneously within a set of possible solutions to find the optimal solution, by iteratively trying to improve the feature subset with regard to a given measure of quality. Five well-known EC algorithms – Ant Colony Optimization (ACO), Simulated Annealing (SA), Genetic Algorithm (GA), Particle Swarm Optimization (PSO) and Differential Evolution (DE) – are widely used for feature selection in various applications, including facial expression-based emotion recognition (Mistry, Zhang, Neoh, Lim, & Fielding, 2016) and classification of motor imagery EEG signals (Baig, Aslam, Shum, & Zhang, 2017). Baig's study, particularly, has achieved a very high (95%) accuracy using a DE algorithm, but was based on only 5 subjects.

Compared to previous work, this is the first study to identify the best EC-based feature selection method(s), which have not been previously tested on EEG-based emotion recognition. The proposed method is evaluated using two public datasets (DEAP and MAHNOB) and a newly collected dataset using wireless EEG sensors to give a comprehensive review on experimental results in different contexts. In all three datasets, video clips and music were used as stimuli to induce different emotions. In addition, this study investigates the most optimal subset of features within each dataset and identifies the most frequent set of selected channels using the principle of weighted majority voting. These findings are significant for informing future development of EEG-based emotion classification because low-cost mobile EEG sensors with fewer electrodes are becoming popular for many new applications. This paper is organized as follows. Section 3.2 discuss the framework and adjustment

of algorithms for the feature-selection problem. Section 3.3 presents the methodology, including the data collection. Section 3.4 discusses the experimental results, while Section 3.5 provides the conclusion and discusses future work.

3.4 SYSTEM FRAMEWORK

A typical emotion classification system using EEG signals consists of four main tasks: *pre-processing, feature extraction, feature selection* and *classification* (see Fig. 3.1). The first and most critical step is pre-processing, as EEG signals are typically noisy as a result of contamination by physiological artefacts caused by electrode movement, eye movement, muscle activities, heartbeat and so on.

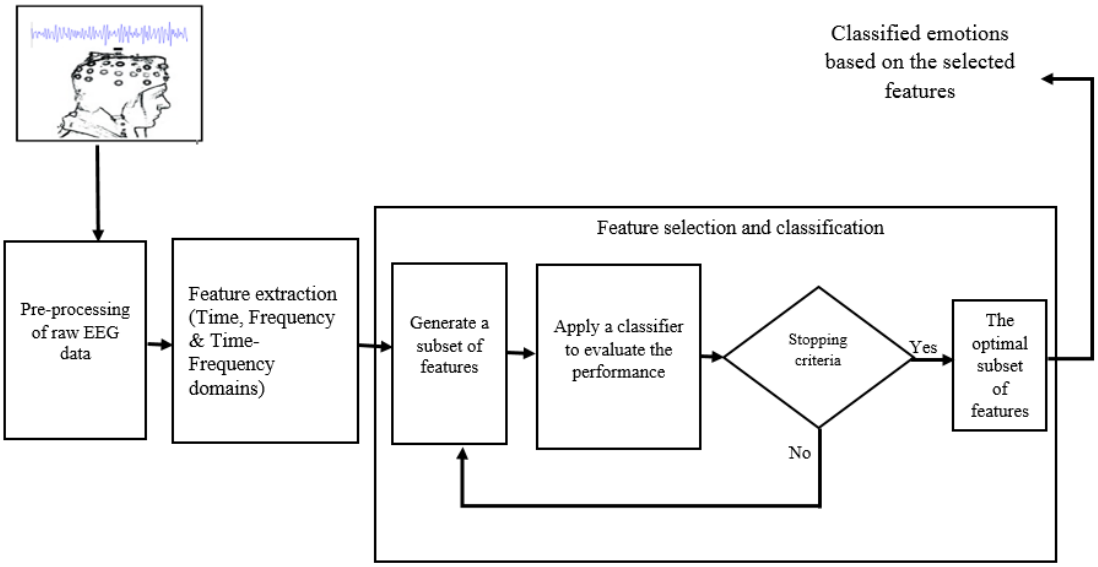


Figure3.1. The proposed emotion classification system using evolutionary computational (EC) algorithms for feature selection.

The artefacts that are generated from eye movement, heartbeat, head movement and respiration are below the frequency of 4 Hz, while the artefacts caused by muscle movement are higher than 40 Hz. In addition, there are some non-physiological artefacts caused by power lines with frequencies of 50 Hz, which contaminate the EEG signal. In order to remove artefacts while keeping the EEG signals within specific frequency bands, sixth-order (band-pass) *Butterworth* filtering was applied to obtain

4–64 Hz EEG signals to cover different emotion-related frequency bands. Notch filtering was applied to remove 50 Hz noise caused by power lines. In addition to these pre-processing methods, independent component analysis (ICA) was used to reduce the artefacts caused by heartbeat and to separate complex multichannel data into independent components (Jung et al., 2000), and provide a purer signal for feature extraction. The purer EEG signals were then passed through a feature extraction step, in which several types of features from time, frequency and time–frequency domains were extracted to distinguish different emotions. Subsequently, all the extracted features from each channel were concatenated into a single vector representing a large feature set. To reduce the number of features used for the machine learning process, EC algorithms are applied iteratively to the different feature sets to find the optimal and most effective set. The classification and feature selection steps were integrated to iteratively evaluate the quality of the feature sets produced by the feature selection against the classification of specific emotions based on experimental results. To evaluate the performance of each EC feature selection algorithm, a probabilistic neural network (PNN) (Specht, 1990) was adopted, as it has been shown to be effective for emotion recognition using different modalities. A PNN is a feed forward network with three layers which is derived from Bayesian networks. In our framework, training and testing of each EC algorithm was conducted using 10-fold cross validation, which helps to avoid overfitting. This process is made possible thanks to PNN's faster training process compared to other classification methods, as the training is achieved using one pass of each training vector rather than several passes. Of these tasks, feature extraction and the integrated feature selection and classification methods represent the most important parts of the framework. After reviewing and evaluating these methods, the key contribution of this paper is to find the optimal strategy for feature selection of high-dimensional EEG-based emotion recognition.

3.4.1 Feature Extraction

EEG features are generally categorized into three main domains: time-, frequency- and time–frequency.

Time-domain features

Time-domain features have been shown to correlate with different emotional states. Statistical features – such as mean, maximum, minimum, power, standard deviation, 1st difference, normalized 1st difference, standard deviation of 1st difference, 2nd difference, standard deviation of 2nd difference, normalized 2nd difference, quartile 1, median, quartile 3, quartile 4 – are good at classifying different basic emotions such as joy, fear, sadness and so on (Chai, Woo, Rizon, & Tan, 2010; Takahashi, 2004). Other promising time-domain feature is Hjorth parameters: Activity, Mobility and Complexity (Ansari-Asl, Chanel, & Pun, 2007; Horlings, Datcu, & Rothkrantz, 2008). These parameters represent the mean power, mean frequency and the number of standard slopes from the signals, which have been used in EEG-based studies on sleep disorder and motor imagery (Oh, Lee, & Kim, 2014; Redmond & Heneghan, 2006; Rodriguez-Bermudez, Garcia-Laencina, & Roca-Dorda, 2013).

All above mentioned features were applied for real-time applications, as they have the least complexity compared with other methods (Khan, Ahamed, Rahman, & Smith, 2011). In addition to these well-known features, we incorporated two newer time-domain features. The first one is fractal dimension to extract geometric complexity, which has been shown to be effective for detecting concentration levels of subjects (Aftanas, Lotova, Koshkarov, & Popov, 1998; Sourina, Kulish, & Sourin, 2009; Sourina & Liu, 2011; Sourina, Sourin, & Kulish, 2009; Wang, Sourina, & Nguyen, 2010). Among the several methods for computing fractal dimension features, the Higuchi method has been shown to outperform other methods, such as box counting and fractal Brownian motion (Yisi Liu & Sourina, 2013). The second newer feature is Non-Stationary Index (NSI) (Kroupi et al., 2011), which segments EEG signals into smaller parts and estimates the variation of their local averages to capture the degree of the signals' non-stationarity. The performance of NSI features can be further improved by combining them with other features, such as higher order crossing features (Petrantonakis & Hadjileontiadis, 2010) that are based on the zero-crossing count to characterize the oscillation behaviour.

Frequency-domain features

Compared to time-domain features, *frequency-domain* features have been shown to be more effective for automatic EEG-based emotion recognition. The power of the EEG signal among different frequency bands is a good indicator of different emotional states. Features such as power spectrum, logarithm of power spectrum, maximum, minimum and standard deviation should be extracted from different frequency bands, namely Gamma (30–64 Hz), Theta (13–30 Hz), Alpha (8–13 Hz) and Beta (4–8 Hz), as these features have been shown to change during different emotional states (Barry, Clarke, Johnstone, Magee, & Rushby, 2007; Davidson, 2003; Koelstra et al., 2012; Onton & Makeig, 2009).

Time–frequency domain features

The limitation of frequency-domain features is the lack of temporal descriptions. Therefore, *time–frequency-domain* features are suitable for capturing the non-stationary and time-varying signals, which can provide additional information to characterize different emotional states. The most recent and promising features are discrete wavelet transform (DWT) and Hilbert Huang spectrum (HHS). DWT decomposes the signal into different frequency bands while concentrating in time, and it has been used to recognize different emotions using different modalities, such as speech (Shah et al, 2010), electromyography (Cheng & Liu, 2008) and EEG (Murugappan et al., 2008). HHS extracts amplitude, squared amplitude and instantaneous amplitude from decomposed signals obtained from intrinsic mode functions, and has been applied to investigate the connection between music preference and emotional arousal (Hadjidimitriou & Hadjileontiadis, 2012). From the Discrete Wavelet Transform, different features can be further extracted to distinguish basic emotions (Murugappan Murugappan, Ramachandran, Sazali, & others, 2010; Muthusamy Murugappan, 2011) , including power, recursive energy efficiency (REE), root mean square, and logarithmic REE. This study adopts all of the abovementioned features to automatically select the most important subset of EEG features that can achieve the optimal classification performance. The features are summarized in Table 3.1.

Table 3.1.The Extracted features from EEG signals

Time Domain	
Minimum	Mean
Maximum	Power
Standard Deviation	1st Difference
Normalized 1st difference	Second difference
Normalized second difference	Hjorth Feature (Activity, Mobility and Complexity)
Non-Stationary Index	Fractal Dimension (Higuchi Algorithm)
Higher Order Crossing	Variance
Root mean Square	Quartile 1, quartile median, quartile 3, quartile4
Frequency Domain	
Power Spectrum Density (PSD) from Gamma, Theta, Alpha, Beta	Mean
Time–Frequency Domain	
Power of Discrete Wavelet Transform (DWT) from Gamma, Theta, Alpha, Beta	
Root Mean Square (RMS) of DWT from Gamma, Theta, Alpha, Beta	Recursive Energy Efficiency (REE) of DWT from Gamma, Theta, Alpha, Beta
Log (REE) of DWT from Gamma, Theta, Alpha, Beta	Abs (log (REE)) of DWT from Gamma, Theta, Alpha, Beta

3.4.2 Feature Selection and Classification

Evolutionary computation (EC) algorithms can be used to select the most relevant subset of features from extracted EEG features. Our study used the five

population-based heuristic search algorithms useful for global searching, mentioned above, namely: Ant Colony Optimization, Simulated Annealing, Genetic Algorithm, Particle Swarm Optimization and Differential Evolution. Although each of these algorithms is different, the common aim among them is to find the optimum solution by iteratively evaluating new solutions. All EC algorithms follow three steps: 1) initialization, where the population of solutions is initialized randomly; 2) evaluation of each solution in the population for fitness value; 3) iteratively generating a new population until the termination criteria are met. The termination criteria could be the maximum number of iterations or finding the optimal set of features that maximize classification accuracy.

Ant Colony Optimization (ACO)

First proposed by Dorigo and Gambardella (1997), ACO is inspired from the foraging behaviour of ant species. It is based the finding the shortest paths from food sources to the nest. In an ant colony system, ants leave a chemical, pheromone, on the ground while they are searching for food (Dorigo, Birattari, & Stutzle, 2006). Once an ant finds a food source, it evaluates the quantity and quality of the food and, during its return trip to the colony, leaves a quantity of pheromone based on its evaluation of that food. This pheromone trail guides the other ants to the food source. Ants detect pheromone by smell and choose the path with the strongest pheromone. This process is performed iteratively until the ants reach the food source.

ACO has been widely applied in many domains such as job scheduling (Blum & Sampels, 2002; Colorni, Dorigo, Maniezzo, & Trubian, 1994), sequential ordering (Dorigo & Gambardella, 1997), graph coloring (Costa & Hertz, 1997), shortest common super sequences (Michel & Middendorf, 1998) and connectionless network routing (Sim & Sun, 2003). Studies have shown the utility of the ACO algorithm for the feature-selection problem (Al-Ani, 2005; Sivagaminathan & Ramakrishnan, 2007). To apply the ACO algorithm to the feature-selection problem, we need to include a path for feature selection algorithms. The path can be represented as a graph, where each node in the graph represents a feature and the edge shows the next feature to be selected. Based on this path, the ants are generated with a random set of features. From

their initial positions, they start to construct the solution (set of features) using heuristic desirability, which denotes the probability of selecting feature *i* by ant *r* at time step *t:*

$$P_i^r(t) = \begin{cases} \frac{\tau(i)^{\alpha}.n(i)^{\beta}}{\sum_{u \in J(r)} \tau(u)^{\alpha}.n(u)^{\beta}} \; if \; i \in J(r) \\ 0 \; otherwise \end{cases} \quad (1)$$

Where, for the ant r, $n(i)$ and $\tau(i)$ are the heuristic information and the pheromone value of feature *i*, and α and β are the parameters which determine the importance of pheromone value and heuristic information respectively. The $n(i)$ and $\tau(i)$ parameters can create a balance between exploration and exploitations, influenced by α and β values. If $\alpha = 0$, then no pheromone information is used and the previous search is overlooked and if $\beta = 0$, then the exploration or global search is overlooked. After constructing the solutions (a set of features) for each ant, the fitness function is applied on each solution to evaluate performance. In this study, a PNN classifier was employed to evaluate the accuracy of each solution, which represents the set of features. Then the pheromone evaporation was applied as follows:

$$\tau(t+1) = (1-\rho) * \tau(t) \quad (2)$$

Where $\rho \in (0,1)$ is the pheromone decay coefficient. Finally, the process stops when the termination criteria are met – either the optimum set of features with highest accuracy or the maximum number of iterations is achieved.

Simulated annealing (SA)

First proposed by Kirkpatrick, Gelatt, and Vecchi (1983), SA is inspired from metallurgy processes. It is based on selecting the best sequence of temperatures to achieve the best result. The algorithm starts from the initial state of high temperature,

then iteratively creates a new random solution and opens a search space widely to slowly decrease the temperature to a frozen ground state. SA is used for different domains such as job shop scheduling (Suresh & Mohanasundaram, 2006), clustering (Bandyopadhyay, Saha, Maulik, & Deb, 2008) and robot path planning (Zhu, Yan, & Xing, 2006).

For feature selection, SA iteratively generates new solutions from the neighborhood and then the fitness function of the new generated solution is calculated and compared with the current solutions. A neighbor of a solution is generated by selecting a random bit and inverting it; if the fitness function of the new solution is better than the current solution, then the new solution will be replaced for the next iteration. Otherwise, it will be accepted based on the Metropolis condition which states that, if the difference between the fitness function of the current solution and the new solution is equal or higher than zero, then a random number δ will be generated between [0,1]. And then if the Boltzmann's function (Eq. 4) value is higher than δ, the new generated solution will be accepted for the next iteration.

$$\exp(\Delta E/T) \geq \delta \tag{4}$$

After all iterations at each temperature are complete, the next temperature state is selected based on a temperature updating rule (equation 5). This process continues iteratively until the termination criteria are reached, namely a fixed number of iterations or until no further improvement is observed.

$$T_n = \propto^N T_i \tag{5}$$

Where:

- T_d is the new decreasing temperature state.
- $\propto$ is the cooling ratio.

- N is the number of iteration in each temperature state.
- T_i is the initial temperature state.

Genetic Algorithm (GA)

First proposed by Goldberg and Holland (1988), GA is inspired by natural selection. The algorithm aims to find the (near) optimal solution for chromosomes to continue surviving, based on stochastic optimization. It has been applied to find the optimal solutions for job scheduling problems (Gonçalves, de Magalhães Mendes, & Resende, 2005; Karthigayan, Rizon, Nagarajan, & Yaacob, 2008; Tuncer & Yildirim, 2012; J. Yang & Honavar, 1998)

Continuing with the analogy, the algorithm contains a set of chromosomes, which are represented in binary form, with operators for fitness function, breeding or crossover, and mutation. Each of these binary chromosomes is represented as a solution and these solutions are used to generate new solutions. Initially, the chromosomes are created randomly to represent different points in the search space. The fitness of each chromosome is evaluated and the chromosomes with better fitness value are more likely to be kept for the next generation (as a parent). New chromosomes are then generated using a pair of the fittest current solutions through the combination of successive chromosomes and some crossover and mutation operators. The crossover operator replaces a segment of a parent chromosome to generate a new chromosome, while the mutation operator mutates a parent chromosome into a newly generated chromosome to make a very small change to the individual genome. This mutation process helps in introducing randomness into the population and maintaining diversity within it. Otherwise, the combination of the current population can cause the algorithm to become trapped in the local optima, unable to explore the other search space. Finally, the newly generated chromosomes are used for the next iterations. This process continues until some satisfactory criteria are met.

To apply the GA algorithm to the feature-selection in high-dimensional settings, each chromosome is represented by a binary vector of dimension m, where m is the total number of features. If a bit is 1, then the corresponding feature is included, and if

a bit is 0, the feature is not included. The process of GA for feature selection problem is the same as GA. The process terminates when it finds the subset of features with highest accuracy or reaches the maximum number of iterations.

Particle Swarm Optimization (PSO)

First proposed by Eberhart and Kennedy (1995), PSO is inspired by the social behaviors of bird flocking and fish schooling. The algorithm is a population-based search technique similar to ACO (see 2.4.1). PSO was originally applied to continuous problems and then extended to discrete problems (Kennedy & Eberhart, 1997). Due to its simplicity and effectiveness, this algorithm is used inn different domains such as robotics (Nakisa et al., 2014; Rastgoo et al., 2015) and job scheduling (Sha & Hsu, 2006; G. Zhang, Shao, Li, & Gao, 2009).

The algorithm is similar to GA, as it consists of a set of particles, which resemble the chromosome in GA, and a fitness function. Each particle in the population has a position in the search space, a velocity vector and a corresponding fitness value which is evaluated by the fitness function. However, unlike GA, PSO does not require sorting of fitness values of any solution in any process, which may be a computational advantage, particularly when the population size is large.

To apply PSO to feature-selection problems, the first step is initialization. At each iteration, the population of particles spread out in the search space with random position and velocity. The fitness value of each particle is evaluated using the fitness function. The particles iterate from one position to another position in the search space using the velocity vector. This velocity vector (equation 6 is calculated using the particle's personal best position (P_{best}), the global best (g_{best}) and the previous velocity vector. The particle's personal best value is the best position that the particle has visited so far and the global best (g_{best}) is the best visited position by any particle in the population. These two values can be controlled by some learning factors. The next particle's position will be evaluated through the previous position and the calculated velocity vector (as described in equations 6 and 7).

$$v_i^{t+1} = \omega v_i^t + c_1 r_1 (P_i^t - x_i^t) + c_2 r_2 (G^t - x_i^t) \quad (6)$$

$$x_i^{t+1} = x_i^t + v_i^{t+1} \quad (7)$$

where v_i^t and x_i^t are the previous iteration's velocity vector and the previous particle's position respectively, ω is the inertia weight, c_1, c_2 are learning factors and r_1, r_2 are random numbers that are uniformly distributed between [0, 1].

Differential Evolution (DE)

Another stochastic optimization method is Differential Evolution (DE), which has recently attracted increased attention for its application to continuous search problems (Price, Storn, & Lampinen, 2006). Although its process is similar to the PSO algorithm, for unknown reasons it is much slower than PSO. Recently, its strength has been shown in different applications such as strategy adaptation (A. K. Qin, Huang, & Suganthan, 2009) and job shop scheduling (Pan, Wang, & Qian, 2009). Most recently this algorithm has shown promising performance as a feature-selection algorithm for EEG signals in motor imagery applications (Baig, Aslam, Shum, & Zhang, 2017).

DE algorithm represents a solution by a D-dimensional vector. A population size of N with a D-dimensional vector is generated randomly. Then a new solution is generated by combining several solutions with the candidate solution, and these solutions are evolved using three main operators: mutation, crossover and selection. Although the concept of solution generation is applied in the DE algorithm in the same way as it is applied in GA, the operators are not all the same as those with the same names in GA.

The key process in DE is the generation of a trial vector. Consider a candidate or a target vector in a population of size N of D-dimensional vectors. The generation of a trial vector is accomplished by the mutation and crossover operations and can be summarized as follows. 1) Create a mutant vector by mutation of three randomly selected vectors. 2) Create a trial vector by the crossover of the mutant vector and the target vector. When the trial vector is formed, the selection operation is performed to keep one of the two vectors, that is, either the target vector or the trial vector. The

vector with better fitness value is kept, and is the one included for the selection of the next mutant vector. This is an important difference since any improvement may affect other solutions without having to wait for the whole population to complete the update.

3.5 EXPERIMENTAL METHOD

We conducted extensive experiments to determine if evolutionary computation algorithms can be used as effective feature selection processes to improve the performance of EEG-based emotion classification and to find the most successful subset of features. Based on the experiments, we determined which features are generally better, across all stimuli and subjects.

The experiments were based on the goal of classifying four-class emotions based on EEG signals, using the most successful subset of features generated from five different EC algorithms (ACO, SA, GA, PSO and DE). The experiments used two public datasets (MAHNOB and DEAP), which contain EEG signals with 32 channels. In addition, a new dataset of EEG signals with only 5 channels was used for comparison purposes. In order to decrease the computation time, the experiments used data from 15 randomly-selected subjects. We simplified the dimensional (arousal-valence based) emotions into four quadrants: 1) High Arousal-Positive emotions (HA-P); 2) Low Arousal-Positive emotions (LA-P); 3) High Arousal-Negative emotions (HA-N); 4) Low Arousal-Negative emotions (LA-N)).

As described in Section 2.1, some noise reduction techniques, including Butterworth, notch filtering and ICA, were applied to the EEG signals to remove artefacts and noise. After noise reduction, a variety of EEG features from *time, frequency and time–frequency domains* were extracted from a one-second window with 50% overlap (45 features in total from each window). For the experiments using the MAHNOB and DEAP datasets, we extracted features from all 32 EEG channels. Hence, the total number of EEG features for each subject was $45 \times 32 = 1440$ features. To reduce the high dimensionality of the feature space and to improve the performance of EEG-based emotion classification, the five EC algorithms were applied. Finally, a PNN classifier with 10-fold cross validation was used to evaluate the performance of the generated set of features. This means that the integrated feature selection and

classification process was iteratively processed 10 times on each dataset with a different initial population from the 15 subjects. The average over all 10 runs was calculated as a final performance. The whole experimental software was implemented using MATLAB, and the following settings were applied:

- For the ACO algorithm: number of ants = 20, evaporation rate = 0.05, initial pheromone and heuristic value = 1 (Section Ant Colony Optimization (ACO)).
- For the SA algorithm: initial temperature = 10, cooling ratio = 0.99, maximum number of iteration in each temperature state = 20 (Section Simulated annealing (SA)).
- For the GA algorithm: crossover percentage = 0.7, mutation percentage = 0.3, mutation rate = 0.1, selection pressure = 8 (Section Genetic Algorithm (GA)).
- For the PSO algorithm: construction coefficient = 2.05, damping ratio = 0.9, particle size = 20. For the DE algorithm: population size = 20, crossover probability = 0.2, lower bound of scaling factor = 0.2, upper bound of scaling factor = 0.8 (Section Particle Swarm Optimization (PSO)).

3.6 DESCRIPTION OF DATASETS

3.6.1 MAHNOB (Video)

MAHNOB (Soleymani, Lichtenauer, et al., 2012) contains a recording of user responses to multimedia content. In the study, 30 healthy young people, aged between 19 and 40 years old, from different cultural backgrounds, volunteered to participate. Fragments of videos from online sources, lasting between 34.9 and 117s, with different content, were selected to induce 9 specific emotions in the subjects. After each video clip the participants were asked to describe their emotional state using a keyword such as neutral, surprise, amusement, happiness, disgust, anger, fear, sadness, and anxiety. While each participant was watching the videos, EEG signals using 32 channels based on a 10/20 system of electrode placement were collected. The sampling rate of the recording was 1024Hz, but it was down-sampled to 256Hz afterwards. Only the signal parts recorded while the participants watched the videos were included in the analysis

and the annotation part was left out. To decrease computational time, our experiment randomly selected 15 of the 30 recorded participants.

3.6.2 DEAP (music)

DEAP (Koelstra et al., 2012) contains EEG signals with 32 channels and other physiological signals, recorded from 32 participants while they were listening to 40 one-minute music videos. The self-assessment in this dataset was based on the Arousal-Valence, like/dislike, and dominance and familiarity levels. To decrease computational time, our experiment used the EEG signals from 15 of the 32 participants.

3.6.3 New Experiment Dataset (Video)

The third dataset was newly collected from 13 subjects, aged between 20 and 38 years old, while they watched video clips. To collecting the EEG data, the *Emotiv Insight wireless* headset was used (see figure 3.2). This Emotiv headset contains 5 channels (AF3, AF4, T7, T8, Pz) and 2 reference channels located and labeled according to the international 10-20 system (see figure 3.3). Compared to the EEG recording devices that were used in MAHNOB and DEAP, our data collection used Emotiv wireless sensor, which only captures EEG signals from five channels (instead of 32). Based on the experimental results, we will determine if the use of wireless sensor can become a viable option for future studies that require subjects to move around freely.

Figure 3.2. The Emotiv Insight headset.

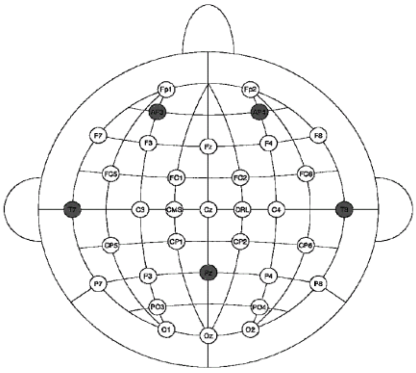


Figure 3.3. The location of five channels used in Emotiv sensor (represented by the black dots, while the white dots are the other channels normally used in other wired sensors).

We used TestBech software for acquiring raw EEG data from the *Emotiv* headset while a participant was watching videos. Emotions were induced by video clips, used in MAHNOB dataset, and the participants' brain responses were collected while they were watching 9 video clips in succession. The participants were asked to report their emotional state after watching each video, using a keyword such as neutral, anxiety, amusement, sadness, joy or happiness, disgust, anger, surprise, and fear. Before the first video clip, the participants were asked to relax and close their eyes for one minute to allow their baseline EEG to be determined. Between each video clip stimulus, one minute's silence was given to prevent mixing up the previous emotion. The experimental protocol is shown in figure 3.4.

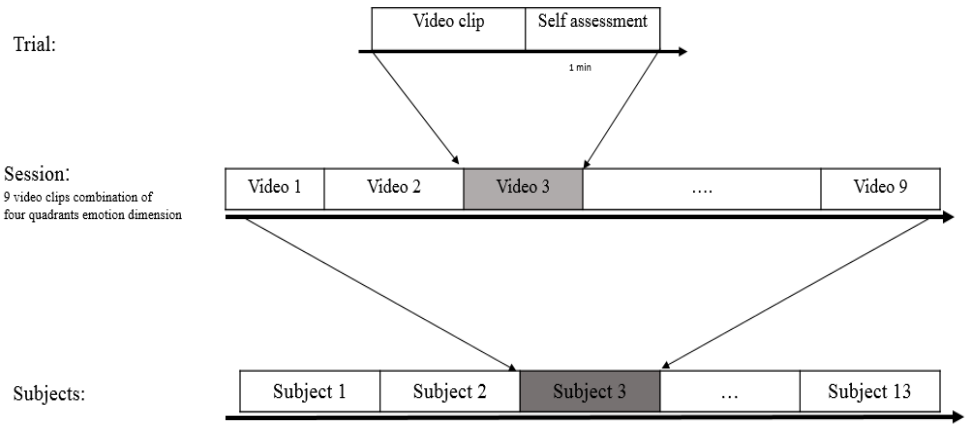


Figure 3.4. Illustration of the experimental protocol for emotion elicitations.

To ensure data quality, we manually analyzed the signal quality for each subject. Some EEG signals from the 5 channels were either lost or found to be too noisy due to the long study duration, which may have been caused by loose contact or shifting electrodes. As a result, only signal data from 11 (5 female and 6 male) out of 13 participants is included in this dataset. Despite this setback, the experiment with this new data allows an investigation into the feasibility of using the Emotiv Insight sensor for emotion classification purposes. The expected benefit of this sensor is due to its light-weight, and wireless nature, making it possibly the most suitable for free-living studies in natural settings.

3.7 EXPERIMENTAL RESULTS AND DISCUSSION

3.7.1 Benchmarking Feature Selection Methods

The performance of EC algorithms was assessed based on the optimum accuracy that can be achieved within reasonable time frames for convergence. Each algorithm was tested based on its ability to achieve the best subset of features within a limited time. As mentioned in Section 3.3, the total number of features generated from 32 channels was 1440. In order to find the minimum subset of features to maximize the classification performance, we empirically limited the selected number of features to 30 out of 1440. Based on our findings, this number of features not only maximizes the performance of the proposed model, but can also keep the computational cost sufficiently low. In addition to feature size, the computational complexity of ECs algorithms is dependent on the number of iterations required for convergence, which may need to be obtained by a trial-and-error process. Despite EC algorithms are computationally expensive, they are less complex than full search and sequential algorithms. Moreover, the computational complexity of the proposed feature selection method is less important, as long as it is possible to achieve an acceptable result and complete the step in a reasonable time. This is due to the fact that the process of proposing the most optimized feature selection method is only conducted during development and training stages.

Table 3.2 presents the relative performance of each EC algorithm, providing average accuracy $\pm$ standard error and processing time for 30 selected features over 10 runs with random starts on three datasets (MAHNOB, DEAP and our dataset). Processing time was determined by Intel Core i7 CPU, 16GB RAM, running windows 7 on 64-bit architecture. Based on the average processing time across three datasets, ACO takes the most amount of processing time (106.2 h), while not delivering the highest accuracy (92%, 60% and 60% for MAHNOB, DEAP and our dataset respectively). In contrast, DE gives the highest accuracy (96%, 67% and 65% for MAHNOB, DEAP and our dataset respectively), while the processing time (76.4 h on average) is lower than others and only higher than that of SA (but SA gives the lowest accuracy). Based on the average accuracy across all datasets (representing the overall performance), the best result is achieved when DE is applied, followed by PSO and GA. From 25 to 100 iterations, the mean accuracy rate of PSO and GA improved by 11% over the MAHNOB dataset and 14% and 8% over the DEAP dataset respectively. However, the improvement rate for DE was lower (7% for MAHNOB and 4% for DEAP). Similarly, the improvement rate for ACO and SA was only 2% and 5% over MAHNOB respectively. This phenomenon is most likely due to PSO and GA's diversity property (i.e. its capability in searching for solutions more widely), while ACO and SA are more likely to be trapped in local optima from the early iterations and seem to converge faster than the other algorithms. This is one of the common problems among EC algorithms, known as the *premature convergence* problem.

Further analysis of the diversity property of all algorithms is provided in Figure 3.5, showing that the proposed system reached an acceptable result based on the peak performance within 100 iterations within an acceptable processing time period. This graph depicts the challenging part of EC algorithms, when these algorithms become trapped in local optima and fail to explore the other promising subsets of features. Generally, these algorithms converge to local optima/ global optima after some iterations and their performance remains steady without any improvement. In this study, it is shown that as the number of iterations increased, the performance of DE, PSO and GA increased and they found better solutions with higher accuracy, while ACO and SA converged in the early iterations and failed to improve their performance. DE had the best convergence speed and founds better solutions. All DE, PSO and GA

algorithms came very close to the global optima in the MAHNOB dataset in all runs, but the other two algorithms (SA and ACO) usually failed to explore and stagnated into the local optima, due to less diversity in the nature of these algorithms.

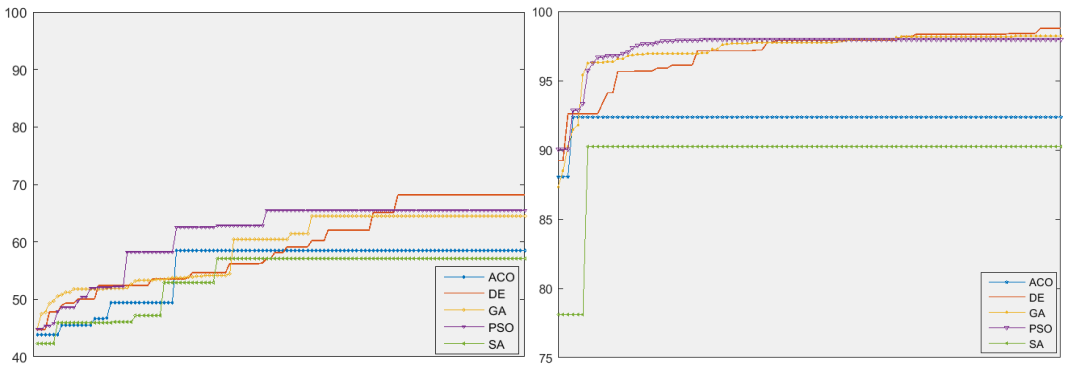
In terms of comparing results across the different datasets, the results from MAHNOB are significantly better than DEAP. This can be explained by the fact that video is a more effective stimulus to induce different emotions, compared to music.

Table 3.2. The average accuracy of EC algorithms over three datasets (MAHNOB, DEAP, our dataset).

FS method	No. Iteration	(MAHNOB)		(DEAP)		(Our dataset)		Average time across datasets
		Time (h)	Accuracy (%)	Time (h)	Accuracy (%)	Time (h)	Accuracy (%)	
PSO	15	12.5	82.94263 ∓2.865	12.9	51.18436 ∓4.23412	11.7	59.16438 ∓4.07415	
	25	20.9	85.5012 ∓2.5847	21.5	51.64421 ∓3.06233	19.5	60.84009 ∓5.77746	
	45	37.5	95.30953 ∓2.0627	38.7	54.07454 ∓2.36788	35.4	60.41786 ∓2.81327	
	100	**83.9**	**96.58661 ∓1.83694**	**86.1**	**65.31437 ∓3.22760**	**78.4**	**61.48601 ∓6.59215**	**82.8 h**
ACO	15	16	89.63667 ∓1.54467	16.2	6.09645 ∓1.22374	15.3	53.58217 ∓1.99010	
	25	26.6	89.96537 ∓1.45023	27.2	56.7229 ∓1.87711	25.5	54.4585 ∓1.57451	
	45	48	90.7626 ∓1.32020	48.8	58.23569 ∓1.54648	45.9	55.58698 ∓1.27219	
	100	**106.6**	**91.97158 ∓1.14934**	**110**	**59.25536 ∓1.39521**	**102**	**55.58698 ∓5.94841**	**106.2 h**
GA	15	11.9	84.85624 ∓2.21359	12.3	50.12445 ∓2.98451	11.8	57.62701 ∓2.89420	
	25	19.7	86.26149 ∓1.97594	20.6	55.91216 ∓2.63143	19.6	59.59106 ∓2.67428	
	45	35.6	94.99567 ∓1.93795	37.1	57.98351 ∓2.42171	35.4	57.58815 ∓3.60488	
	100	**79**	**97.11983 ∓1.41752**	**82.5**	**63.63564 ∓3.17107**	**71.3**	**61.237212 ∓2.16085**	**77.6 h**
SA	15	10.9	83.11863 ∓3.11444	11.2	44.65506 ∓2.73391	10	49.04851 ∓1.94359	
	25	18.1	84.54847 ∓5.37894	18.7	48.6937 ∓3.62394	16.6	51.30788 ∓2.31224	
	45	32.7	86.86808 ∓3.89385	33.7	52.20574 ∓5.76376	30	51.14426 ∓3.10572	

	100	**72.7**	**89.01175 ∓2.31283**	**'5**	**55.25314 ∓2.64210**	**66.6**	**52.04453 ∓2.95477**	**71.4 h**
DE	15	11.6	83.01443 ∓1.9758	11.9	60.92632 ∓2.92837	10.8	56.14765 ∓2.73955	
	25	19.4	89.67716 ∓2.00478	19.8	63.59828 ∓2.17161	18	59.11505 ∓3.03241	
	45	35.1	92.99567 ∓1.85698	35.7	64.59933 ∓2.85288	32.4	63.04303 ∓2.19556	
	100	**77.8**	**96.97023 ∓1.89385**	**9.3**	**67.47447 ∓3.38984**	**72.1**	**65.043028 ∓3.19556**	**76.4 h**

This hypothesis is further justified by the results from our new dataset, which was obtained using the same video stimuli as MAHNOB. Using only 5 EEG channels (instead of 32), the overall performance appears to be very similar to that of the DEAP dataset (slightly lower), which confirms the feasibility of using the wireless and light-weight EEG sensor (albeit of lower accuracy).



(a) (b)

Figure 3.5. Performance of different algorithms for DEAP (a) and MAHNOB (b) databases.

3.7.2 Frequently-Selected Features

To investigate the general quality features for emotion recognition based on two shared datasets (MAHNOB and DEAP), i.e. which feature selected by 5 EC feature-selection methods are repeated more frequently than the others and performed more successfully in emotion classification, we computed the occurrence number of each of the 45 features in the best generated features by each EC algorithm. In this case, we put all the best generated subset of features using 5 EC algorithms tested against each

dataset (MAHNOB, DEAP) together and then generated a set of features regardless of the channels (since each channel has the set of 45 features and in order to find the most repeated features we do not need to consider the channels). Then we provided figures to show the occurrence weight of each feature for each dataset.

The most frequent time-domain features from the MAHNOB dataset are maximum, 1st difference, 2nd difference, normalized 2nd difference, band power, HOC, mobility and complexity. Among the frequency-domain features, PSD from alpha, beta, and gamma are repeated more than the other frequency bands. Of the time frequency domain features, *Rms_Theta*, *REE_Beta*, *power_Gamma*, and *power_Theta* are selected more often than the others (see Figure 3.6).

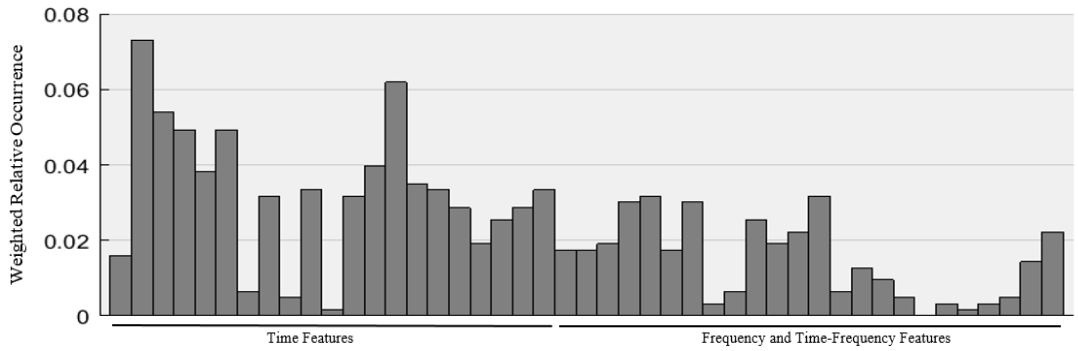


Figure 3.6. The weighted relative occurrence of features over the MAHNOB dataset.

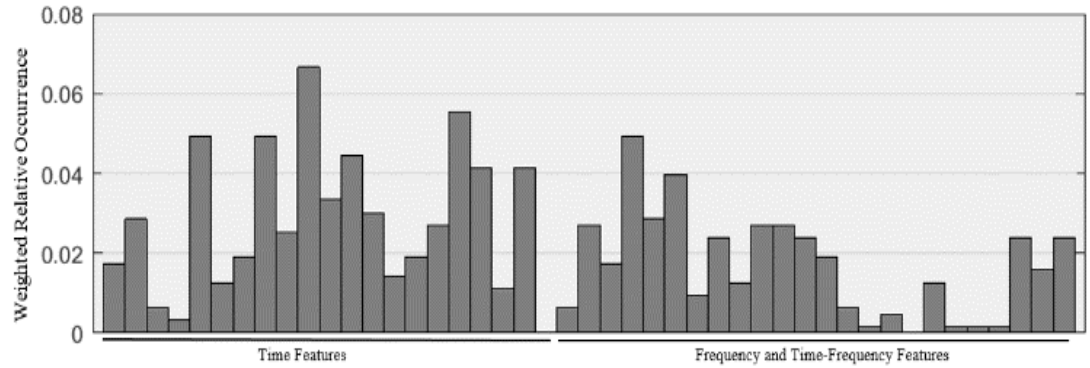


Figure 3.7. The weighted relative occurrence of features over the DEAP dataset.

Similarly, as shown in Figure 3.7, the most repeated features using the 5 EC algorithms tested against the DEAP dataset are mostly selected from the time domain such as: maximum, 1st difference, 2nd difference, normalized 2nd difference, median, mobility, complexity, HOC. Features from the frequency domain, such as PSD from

Alpha and Gamma, are repeated more than the other frequency bands. Among the time–frequency features, rms_Theta and rms_Gamma, power_Gamma, power_Theta and power_Alpha are more frequent than the others.

For the two datasets (collectively), the most common frequently selected features are: maximum, 1st difference, 2nd difference, normalized 2nd difference, complexity, mobility, HOC, PSD from Gamma and Alpha, rms_Theta, power_Gamma and power_ Alpha. However, results suggest that REE and LogREE from different bands from the time–frequency features are less efficient in emotion classification, since their relative occurrence is lower than the others. The relative occurrence of power features from DWT is higher than the other features in this domain, but not as high as PSD features from the frequency domain. It should be noted that the combination of time and frequency domain features is more efficient, since EC algorithms find the more successful and efficient subset of features by the combination of these features.

3.7.3 Channel Selection

We investigated the most frequent set of selected channels (i.e. electrodes usage) by the combination of EC algorithms via the principle of weighted majority voting. Having the best subsets of features from each 10 runs of the EC algorithms, we then considered the task of building a vector representing the importance of channels. The importance of channels can be represented as follows:

$$w_c = \sum_{i=1}^{k} \sum_{j=1}^{n} \alpha_i * f_c \, , 0 < \alpha_i < 1 \, , \tag{8}$$

Where κ, n are the number of algorithms and the number of runs in each algorithm respectively (10 runs for each algorithm is considered for this study). α_i represents the average weight of i_{th} algorithm over all runs which is dependent on the accuracy of classifiers over 10 runs. It means that at each run the performance of each EC algorithm is collected and then, based on the average performance over all 10 runs, an average weight (α_i) is allocated to each EC algorithm. This number is multiplied

by f_c, which is the total number of selected features for c_{th} channel. w_c represents the weight of channel c. Figures 3.8, 3.9 and 3.10 show the plots of w_c based on the experiments using DEAP and MAHNOB (32 channels), and our dataset with 5 channels respectively. Darker boxes are channels with higher weight (w_c), which indicates the most repeated channels among the EC algorithms.

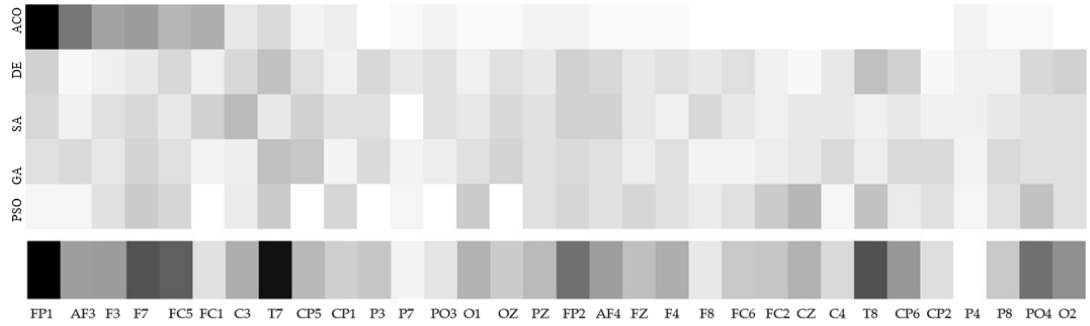


Figure 3.8. Average electrode usage of each EC algorithm within 10 runs on the DEAP dataset is specified by darkness on each channel.

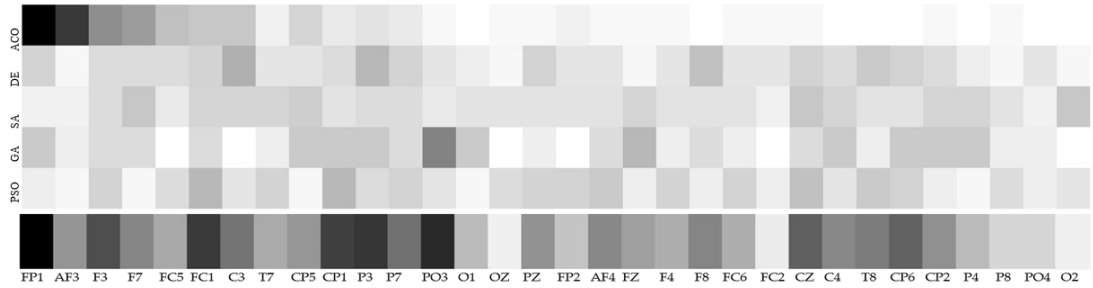


Figure 3.9. Average electrode usage of each EC algorithm within 10 runs on the MAHNOB dataset is specified by darkness on each channel.

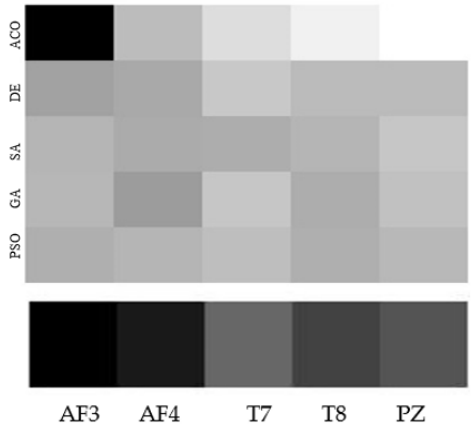
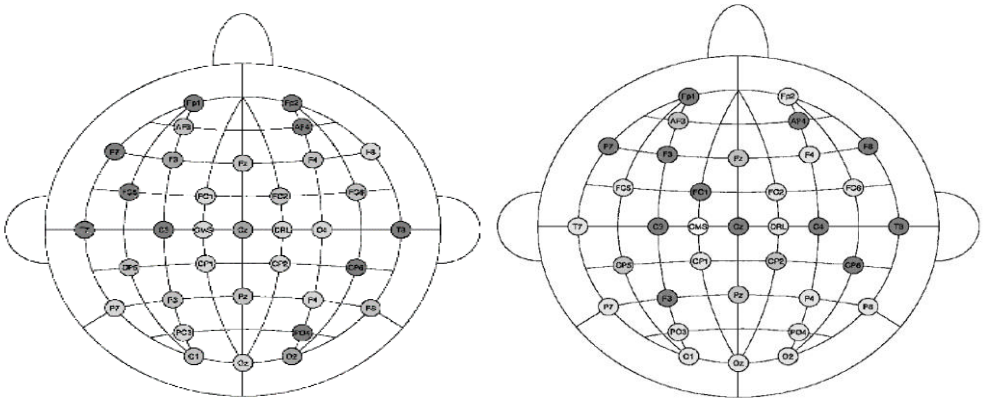


Figure 3.10. Average electrode usage of each EC algorithm within 10 runs on our dataset is specified by darkness on each channel.

Based on the DEAP dataset, the average accuracy shows that the DE algorithm achieves better accuracy, followed by the PSO and GA algorithms. Therefore, their average weight is higher than others. Based on the obtained results, FP1, F7, FC5, AF4, CP6, PO4, O2, T7 and T8 are the prominent electrodes among all other channels. Although the average weights of ACO and SA are lower, the number of selected features from the frontal lobe channels (FP1, F7, FC5) are big enough to compensate and increase the value of w_c. T7 and T8 are also salient, since these channels were frequently selected by PSO, GA and DE, which have bigger average weights. Moreover, features selected from PO4 and O2 are shown to give acceptable accuracy using these feature selection methods. Based on these findings, electrodes located over the frontal and parietal-partial lobes are generally favored over the occipital lobes (see Figure 3.11a).

Based on the MAHNOB dataset, the most selected features using the ACO and GA algorithms are from the frontal lobes, which increase the w_c value of these channels (FP1, FC1, F3, F7, and AF4). In addition to those channels from the frontal lobe, the channels CP1, CZ CP6, C3, T8, C4 and Cp2 are also highly selected by most of the feature-selection methods. We can conclude that most of the prominent channels using the 5 EC algorithms over the MAHNOB dataset are selected from the frontal and central lobes (Figure 3.11a). Based on the most optimal subsets of features from each of the 10 runs of the EC algorithms, the most frequently selected channels are CZ, CP6, C3, T8, C4, CP2 from the frontal and central lobe.

Across DEAP and MAHNOB datasets, the electrodes located on the frontal and central lobes, such as FP1, AF4, CZ, T8, are found to be most salient. This is aligned with the results from our dataset, which found that the electrodes of frontal and central lobes were the most activated for four quadrant emotion classification (as shown previously in Figure 3.3). This confirms that emotion classification using the mobile and wireless *Emotiv Insight* sensor is feasible.



(a) (b)

Figure 3.11. The average electrode usage of each EC algorithm within 10 runs on (a) DEAP and (b) MAHNOB (a) dataset. On the greyscale, darkest nodes indicate the most frequently used channel.

3.7.4 Comparison with Other Works over MAHNOB and DEAP Datasets

A final experiment compared the best-tuned configuration of our system against some state-of-the-art methods. To this end, DE and PSO were used for feature selection and PNN as a classifier. Experimental results are shown in Table 3.3, indicating the classification accuracy for different emotion classification methods. While it only shows the DE results to represent the proposed method (as DE yields the highest performance based on the experiments). EC-based feature selection after 100 iterations is consistently better than not using any feature selection.

Table 3.3. Comparison of our recognition approach with some state-of-the-art methods.

Method	Extracted Features (No.)	Feature Selection methods	Classifier	No. classes	Accuracy	Dataset
(Y. Zhu, Wang, & Ji, 2014)	Statistical features (5)	-	SVM	2 Arousal Valence	Arousal: 60.23 % Valence :55.72 %	MAHNOB
(Candra et al., 2015)	Time-frequency feature (2)	-	SVM	4 Sad Relaxed Angry Happy	60.9 ± 3.2%	DEAP

(Feradov & Ganchev, 2014)	Short term energy (2)	-	SVM	3 Negative Positive Neutral	62 %	DEAP
(Ackermann et al., 2016)	Statistical features (number not specified)	mRMR	SVM & Random Forest	3 Anger, Surprise Others	Average accuracy:55 %	DEAP
(Kortelainen & Seppänen, 2013)	Frequency-domain features (number not specified)	Sequential feed-forward selection (SFFS)	KNN	2 Arousal Valence	Valence: 63 % Arousal: 65 %	MAHNOB
(Menezes et al., 2017)	Frequency domain and Time domain (11)	-	SVM	Bipartition Arousal & Valence	Arousal=69% Valence=88%	DEAP
				3Triple on Arousal &Valence	Arousal=58% Valence=63%	
(Yin, Wang, Liu, Zhang, & Zhang, 2017)	Frequency and Time domain features (16)	Transfer Recursive Feature eliminati on (T-RFE)	LSSVM	2 Arousal Valence	Arousal: 78 % Valence: 78 %	DEAP
Our proposed method	Frequency, Time and Time–Frequency domain features (45)	EC algorithms	PNN	4 HA-P HA-N LA-P LA-N	DEAP: 67.474 ± 3.389 % MAHNOB: 96.97 ±1.893 % Our Dataset: 65.043028 ∓3.195 %	DEAP, MAHNOB & Our Dataset

The comparisons show that although Yin et al. (2017) achieved promising accuracy (about 78 %), their feature-selection method was Recursive Feature

Elimination, which is computationally expensive. In addition, this method was used on two-class classification (Arousal and Valence), while our method was able to achieve similar accuracy (Maximum 71% (67.474 ± 3.389 %) on DEAP dataset) on four-class classifications (HA-P, LA-P, HA-N, LA-N). The performance of their method was tested on one specific dataset (DEAP) with on mode of stimuli (music), whereas our proposed method was tested on two public datasets (DEAP and MAHNOB) plus a newly collected dataset using mobile sensors (Emotiv Insight), across two different modes of stimuli (music and video).

Ackermann et al. (2016) classified three different emotions (anger, surprise and others) using Support Vector Machine (SVM) and Random Forest classification systems, which are trained by a smaller number of features. They only applied the Minimum Redundancy Maximum Relevance (mRMR) feature-selection method to eliminate less useful features. Their evaluation of the results on the DEAP dataset shows that the performance of the proposed method using the SVM classifier (average of 55%) is more robust and successful when the number of selected features is between 80 and 125.

Menezes et al., (2017) extracted some limited features from time and frequency domains and applied SVM method to classify each emotion dimension (Arousal and Valence), into two and three-class (Bipartition and Tripartition). This model has been tested on DEAP dataset. Although, the performance of their method in Bipartition is promising, the extracted features using SVM classifier could not classify Tripartition-class significantly.

In comparison with other latest studies, our proposed EEG-based emotion classification framework has shown state-of-the-art performance for both the DEAP and MAHNOB datasets, which confirms the value of integrating EC algorithms for feature selection to improve classification performance.

3.7.5 Towards The Use of Mobile EEG Sensors for Real World Applications

The reliability and validity of mobile EEG sensors has been tested in earlier studies (Duvinage et al., 2013; Stytsenko, Jablonskis, & Prahm, 2011). These studies found that, while mobile sensors are not as accurate as a wired and full-scale EEG

device, they can be used in noncritical applications. Some recent studies into the benefits of mobile EEG sensors on different domains have demonstrated an acceptable reliability (Leape et al., 2016; Lushin, 2016; Y. Wu, Wei, & Tudor, 2017; F. Zhang et al., 2017).

In our study, a mobile EEG sensor (Emotiv Insight) was used to recognize four-quadrant dimensional emotions while watching video clips. The proposed method uses different EC algorithms for feature selection applied to three different datasets (DEAP, MAHNOB and our collected dataset), and the result of 65% accuracy shows the validity of using mobile EEG sensors in this domain. In addition, we compared the salient channels in two datasets using full-scale EEG devices (32 channels) with the collected dataset using the mobile EEG sensor (5 channels). The result shows that the most frequent channels from the two public datasets (DEAP and MAHNOB) were from the frontal and central lobes, the same channels detected in our four-quadrant emotion recognition using five EC algorithms. This comparison confirms the feasibility of using mobile EEG sensor for emotion classification. However, the performance based on the MAHNOB dataset (96.97%) is still significantly higher than that of our new dataset (65.04%) despite using the same stimuli. This confirms that more progress is needed to improve the results for mobile EEG sensors. This paper paves a way for future sensor development in selecting the correct channels and features to focus on the most important electrodes.

3.8 CONCLUSION AND FUTURE WORK

In this study, we propose the use of evolutionary computation (EC) algorithms (ACO, SA, GA, PSO and DE algorithms) for feature selection in an EEG-based emotion classification model for the classification of four-quadrant basic emotions (High Arousal-Positive emotions (HA-P), Low Arousal-Positive emotions (LA-P), High Arousal-Negative emotions (HA-N), and Low Arousal- Negative emotions (LA-N)). Our experiments have used two standard datasets, MAHNOB and DEAP, obtained using an EEG sensor with 32 channels, and a new dataset obtained using a mobile sensor with 5 channels. We have reported the performance of these algorithms with different time intervals (different iterations) – 10, 25, 45 and 100 iterations – to

demonstrate the benefits of using EC algorithms to identify salient EEG signal features and improve the performance of classifiers.

Of the EC algorithms, DE and PSO showed better performance over every iteration. Moreover, the combination of time and frequency features consistently showed more efficient performance, compared to using only time or frequency features. The most frequently selected source of features (i.e. the EEG channels) were analyzed using the 5 EC algorithms and weighted majority voting. The electrodes in the frontal and central lobes were shown to be more activated during emotions based on DEAP and MAHNOB datasets, confirming the feasibility of using a lightweight and wireless EEG sensor (Emotiv Insight) for four-quadrant emotion classification.

Despite the promising results, this paper has identified an important limitation due to the premature convergence problem in EC algorithms, particularly DE, PSO and GA. Therefore, for future work, it is worthwhile exploring development of new EC algorithms or modifying the existing ones to overcome this problem and improve the classification performance accordingly.

3.9 ACKNOWLEDGEMENTS

We thank copy editing services provided by Golden Orb Creative, NSW, Australia. This Work is supported by QUT International Postgraduate Research Scholarship (IPRS).

Chapter 4: Long Short Term Memory Hyperparameter Optimization for a Neural Network Based Emotion Recognition Framework (Paper 2)

Bahareh Nakisa[1]

Mohammad Naim Rastgoo[1]

Andry Rakotonirainy[2]

Frederic Maire[1]

Vinod Chandran[1]

1. School of Electrical Engineering and computer Science, Queensland University of Technology, Brisbane, QLD, Australia
2. Centre for Accident Research & Road Safety-Queensland, Queensland University of Technology, Brisbane, QLD, Australia

Corresponding Author:

Bahareh Nakisa

Science and Engineering Faculty

Queensland University of Technology,

Brisbane, Australia

Phone: [redacted]

Email: Bahareh.nakisa@qut.edu.au

RESEARCH STUDENTS CENTRE
Examination Enquiries: 07 3138 1839
Email: research.examination@qut.edu.au

Statement of Contribution of Co-Authors for Thesis by Published Paper

The following is the suggested format for the required declaration provided at the start of any thesis chapter which includes a co-authored publication.

The authors listed below have certified that:

1. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
2. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
3. there are no other authors of the publication according to these criteria;
4. potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit, and
5. they agree to the use of the publication in the student's thesis and its publication on the QUT's ePrints site consistent with any limitations set by publisher requirements.

In the case of this chapter:

Title and status: **Long Short Term Memory Hyperparameter Optimization for a Neural Network Based Emotion Recognition Framework (Published at IEEE Access, Q1 Journal)**

Contributor	Statement of contribution
Bahareh Nakisa	**Candidate** Experimental design, conducted the experimental work, performed data analysis, interpreted results, wrote the manuscript

Mohammad Naim Rastgoo	Assisted with paper planning, editing and proof-reading the paper
Andry Rakotonirainy	**Principal Supervisor** Assisted with paper planning, editing and supervise the experimental work and provide advice.
Frederic Maire	**Associate Supervisor** Assisted with paper planning, editing and supervise the experimental work and provide advice.
Vinod Chandran	**Associate Supervisor (External)** Assisted with paper planning, editing and supervise the experimental work and provide advice.

Principal Supervisor Confirmation		
I have Sighted email or other correspondence from all Co-authors confirming their certifying authorship.		
_____	_____	_____
Name	Signature	Date

4.1 INTRODUCTORY COMMENTS

In order to build emotion recognition system based on physiological signals, another main issue is to use an appropriate classifier. This is due to the fact that physiological signals are characterized by non-stationarities and nonlinearities. In fact physiological signals consist of time-series data with variation over a long period of time and dependencies within shorter periods. To capture the inherent temporal structure within the physiological data and recognize emotion signature which is reflected in short period of time, we need to apply a classifier which considers temporal information (Soleymani et al., 2016; Wöllmer et al., 2013). LSTM networks is a successful network for classifying physiological signals. However, tunning and optimizing this network to provide an acceptable performance is challenging. This chapter focused on optimizing LSTM hyperparameters and tunning LSTM classifier to improve emotion classification based on EEG and BVP signals (study 2).

This chapter seeks to address the Research Question 2: "How can we optimize the performance of emotion classification system based on physiological signals? ". In this study, we propose a new framework to optimize LSTM hyperparameters using DE algorithm. We demonstrate that optimizing LSTM hyperparameters result in significant improvement in emotion classification based on EEG and BVP signals compared to other hyperparameter optimization methods. The finding of the preceding chapter (best possible EEG features) is used in this study. In addition the performance of the proposed model in this chapter is compared with the preceding chapter (Study 1).

Taken from: B. Nakisa, M. N. Rastgoo, A. Rakotonirainy, F. Maire and V. Chandran," Long Short Term Memory Hyperparameter Optimization for a Neural Network Based Emotion Recognition Framework" IEEE Access, 2018.

Publication Status: Published 03 Sep 2018

Journal Quality: IEEE Access is an award-winning, multidisciplinary, all-electronic archival journal, continuously presenting the results of original research or

development across all of IEEE's fields of interest. The journal impact factor is **3.557** and rank Q1 (SCImago) in Computer science and Engineering.

4.2 ABSTRACT

Recently, emotion recognition using low-cost wearable sensors based on Electroencephalogram (EEG) and Blood Volume Pulse (BVP) has received much attention. Long Short Term Memory (LSTM) networks, a special type of Recurrent Neural Networks (RNNs), have been applied successfully to emotion classification. However, the performance of these sequence classifiers depends heavily on their hyperparameter values and it is important to adopt an efficient method to ensure the optimal values. To address this problem, we propose a new framework to automatically optimize LSTM hyperparameters using Differential Evolution (DE). This is the first systematic study of hyperparameter optimization in the context of emotion classification. In this study, we evaluate and compare the proposed framework with other state-of-the-art hyperparameter optimization methods (Particle Swarm Optimization, Simulated Annealing, Random Search and Tree-of-Parzen-Estimators (TPE)) using a new dataset collected from wearable sensors. Experimental results demonstrate that optimizing LSTM hyperparameters significantly improve the recognition rate of four-quadrant dimensional emotions with a 14% increase in accuracy. The best model based on the optimized LSTM classifier using the DE algorithm, achieved 77.68% accuracy. The results also showed that evolutionary computation (EC) algorithms, particularly DE, are competitive for ensuring optimized LSTM hyperparameter values. Although DE algorithm is computationally expensive, it is less complex and offers higher diversity in finding optimal solutions.

4.3 INTRODUCTION

Automatic emotion recognition using miniaturized physiological sensors and advanced mobile computing technologies has become an important field of research in Human Computer Interaction (HCI) applications. Wearable technologies, such as wireless headbands and smart wristbands, record different physiological signals in an unobtrusive and non-invasive manner. The recorded physiological signals such as Electroencephalography (EEG), Blood Volume Pulse (BVP) and Galvanic Skin Response (GSR) are able to continuously record internal emotional changes. The application of these technologies can be found in detecting driving drowsiness (Li et

al., 2015), and more recently in assessing the cognitive load of office workers in a controlled environment (Zhang et al., 2017).

However, building a reliable emotion classification system to accurately classify different emotions using physiological sensors is a challenging problem. This is due to the fact that physiological signals are characterized by non-stationarities and nonlinearities. In fact physiological signals consist of time-series data with variation over a long period of time and dependencies within shorter periods. To capture the inherent temporal structure within the physiological data and recognize emotion signature which is reflected in short period of time, we need to apply a classifier which considers temporal information (Soleymani et al., 2016; Wöllmer et al., 2013).

In recent years, the application of Recurrent Neural Networks (RNNs) for human emotion recognition has led to a significant improvement in recognition accuracy by modelling temporal data. RNNs algorithms are able to elicit the context of observations within sequences and accurately classify sequences that have strong temporal correlation.

However, RNNs have limitations in learning time-series data that stymied their training. Long Short Term Memory networks (LSTM) are a special type of RNNs that have the capability of learning longer temporal sequences (Hochreiter and Schmidhuber, 1997). For this reason LSTM networks offer better emotion classification accuracy over other methods when using time-series data (Kim et al., 2013; Tsai et al., 2017; Wöllmer et al., 2013, 2008).

Although the performance of LSTM networks in classifying different problems is promising, training these networks like other neural networks depend heavily on a set of hyperparameters that determine many aspects of algorithm behaviour. It should be noted that there is no generic optimal configuration for all problem domains. Hence, to achieve a successful performance for each problem domain such as emotion classification, it is essential to optimize LSTM hyperparameters. The hyperparameters range from optimization hyperparameters such as number of hidden neurons and batch size, to regularization hyperparameters (weight optimization).

There are two main approaches for hyperparameter optimization: manual and automatic. Manual hyperparameter optimization has relied on experts, which is time

consuming. In this approach, the expert interprets how the hyperparameters affect the performance of the model. Automatic hyperparameter optimization methods removes expert input but are difficult to apply due to their high computational cost. Automatic algorithmic approaches range from simple Grid search and Random search to more sophisticated model-based approaches. Grid search, which is a traditional hyperparameter optimization method, is an exhaustive search. Based on this approach, Grid search explores all possible combination of hyperparameters values to find the global optima. Random search algorithms, which are based on direct search methods, are easy to implement. However, these algorithms are converged slowly and take a long time to find the global optima.

Another type of random search methods with ability to converge faster and find an optimal/ near optimal solution in an acceptable time is Evolutionary computation algorithms (EC). EC algorithms such as Particle Swarm Optimization (PSO), and Simulated Annealing (SA) have been shown to be very efficient in solving challenging optimization problems (S.-M. Chen & Chien, 2011; Rastgoo, Nakisa, & Nazri, 2015). Among ECs, Differential Evolution (DE) has been successful in different domains due to its capability of maintaining high diversity in exploring and finding better solutions compared to other ECs (Baig et al., 2017; Nakisa et al., 2017).

In this study, we introduce the use of DE algorithm to optimize LSTM hyperparameters and demonstrate its effectiveness in tuning LSTM hyperparameters to build an accurate emotion classification model. This study focuses on optimizing the number of hidden neurons and batch size for LSTM classifier.

The performances of the optimized LSTM classifiers using the DE algorithm and other state-of-the-art hyperparameter optimization techniques are evaluated on a new dataset collected from wearable sensors. In the new dataset, a light-weight wireless EEG headset (Emotiv) and smart wristband (Empatica E4) are used which meet the consumer criteria for wearability, price, portability and ease-of-use. These new technologies enable the application of emotion recognition in multiple areas such as entertainment, e-learning and the virtual world.

To understand different emotions, there are two existing models of emotions: categorized model and dimensional model. Categorized model is based on a limited

number of basic emotions which can be distinguished universally (Ekman, 1992). On the other hand, dimensional emotion divides the emotional space into two to three dimensions, arousal, valence and dominance (Candra et al., 2015; Ramirez and Vamvakousis, 2012; Sourina and Liu, 2011). In this study, we categorize different emotional states based on arousal and valence into four quadrants: 1-High Arousal-Positive emotions (HA-P); 2-Low Arousal-Positive emotions (LA-P); 3-High Arousal-Negative emotions (HA-N); 4-Low Arousal-Negative emotions (LA-N).

In summary, the contribution of this study is:

- A new emotion classification framework based on LSTM hyperparameters optimization using the DE algorithm. This study aims to show that optimizing LSTM hyperparameters (batch size and number of hidden neurons) using DE algorithm and selecting a good LSTM network can result in accurate emotion classification system.
- Performance evaluation and comparison of the proposed model with state-of-the-art hyperparameter optimization methods (PSO, SA, Random search and TPE) using a new dataset collected from wireless wearable sensors (Emotiv and Empatica E4). We show that our proposed method surpass them on four-quadrant dimensional emotion classification.

The rest of this paper is structured as follows: Section II presents an overview of related work. The methodology is presented in Section III and experimental results in Section IV. Finally, Section V discusses conclusion and suggestions for future work.

4.4 RELATED WORK

Emotions play a vital role in our everyday life. Over the past few decades, research has shown that human emotions can be monitored through physiological signals like EEG, BVP and GSR (Koelstra et al., 2010) and physical data such as facial expression (Hossain and Muhammad, 2017). Physiological signals offer several advantages over physical data due to their sensitivity for inner feelings and insusceptibility to social masking of emotions (Kim, 2007). To understand inner human emotions, emotion recognition methods focus on changes in the two major

components of nervous system; the Central Nervous System (CNS) and the Autonomic Nervous system (ANS). The physiological signals originating from these two systems carry information relating to inner emotional states.

Gathering physiological signals can be done using two types of sensors: tethered-laboratory and wireless physiological sensors. Although tethered-laboratory sensors are effective and take strong signals with higher resolution, they are more invasive and obtrusive and cannot be used in everyday situations. Wireless physiological sensors can provide a non-invasive and non-obtrusive way to collect physiological signals and can be utilized while carrying out daily activities. However, the resolution of collected signals from these sensors are lower than tethered-laboratory ones.

Among the physiological signals, EEG and BVP signals have been widely used to recognize different emotions, with evidence indicating a strong correlation between emotions such as sadness, anger, surprise and these signals (Haag et al., 2004; K. H. Kim et al., 2004). EEG signals, which measure the electrical activity of the brain, can be recorded by electrodes placed on the scalp. The strong correlation between EEG signals and different emotions is due to the fact that these signals come directly from the CNS, capturing features about internal emotional states. Several studies have focused on extracting features from EEG signals, with features such as time, frequency and time-frequency domains used to recognize emotions. In a recent study (Nakisa et al., 2017), we proposed a comprehensive set of extractable features from EEG signals, and applied different evolutionary computation algorithms to feature selection to find the optimal subset of features and channels. The proposed feature selection methods are compared over two public datasets (DEAP and MAHNOB) (Koelstra et al., 2012; Soleymani et al., 2012), and a new collected dataset. The performance of emotion classification methods are evaluated on DEAP and MAHNOB datasets, using EEG sensors with 32 channels, and compared with the new collected dataset, used a lightweight EEG device with 5 channels. The result showed that evolutionary computation algorithms can effectively support feature selection to identify the salient EEG features and improve the performance of emotion classification. Moreover, the combination of time domain and frequency domain features improved the performance of emotion recognition significantly compared to using only time, frequency and time-frequency domains features. The feasibility of using the lightweight and wearable EEG

headbands with 5 channels for emotion recognition is also demonstrated. The use of everyday technology such as lightweight EEG headbands has also proved to be successful in other non-critical domains such as game experience (McMahan et al., 2015), motor imagery (Kranczioch et al., 2014), and hand movement (Robinson and Vinod, 2015).

Another physiological activity which correlates to different emotions is Blood Volume Pulse (BVP). BVP is a measure that determines the changes of blood volume in vessels and is regulated by ANS. In general, the activity of ANS is involuntarily modulated by external stimuli and emotional states. There are some studies that have investigated BVP features in emotional states and the correlation between them (Haag et al., 2004; A. M. Khan & Lawo, 2016; Kazuhiko Takahashi, 2004). BVP is measured by Photoplethysmography (PPG) sensor, which senses changes in light absorption density of skin and tissue when illuminated. PPG is a non-invasive and low-cost technique which has recently been embedded in smart wristbands. The usefulness of these wearable sensors has been proven in applications such as stress prediction (Ghosh et al., 2015) as well as emotion recognition (Haag et al., 2004).

Research on wearable physiological sensors to assess emotions has focused on building an accurate classifier based on the advanced machine learning techniques. Recently, advanced machine learning techniques have achieved empirical success in different applications such as neural machine translation system (Qin, Shinozaki, & Duh, 2018), stress recognition using breathing patterns (Cho, Bianchi-Berthouze, & Julier, 2017) and emotion recognition using audio-visual inputs (Chang & Lee, 2017; Chang, et al., 2017). Among these machine learning techniques, RNNs as dynamic models, have achieved state-of-the-art performance in many technical applications (Ebrahimi Kahou, Michalski, Konda, Memisevic, & Pal, 2015; Graves, 2012; Graves, Mohamed, & Hinton, 2013; Yao et al., 2015; Zoph & Le, 2016), due to their capability in learning sequence modelling tasks.

An RNN is a neural network with cyclic connections, with the ability to learn temporal sequential data. These internal feedback loops in each hidden layer allow RNN networks to capture dynamic temporal patterns and store information. A hidden layer in an RNN contains multiple nodes, which generate the outputs based on the current inputs and the previous hidden states. However, training RNNs is challenging

due to vanishing and exploding gradient problems which may hinder the network's ability to back propagate gradients through long-term temporal intervals. This limits the range of contexts they can access, which is of critical importance to sequence data. To overcome the gradient vanishing and exploding problem in RNNs training, LSTM networks were introduced (Hochreiter and Schmidhuber, 1997). These networks have achieved top performance in emotion recognition using multi-modal information (Chao et al., 2015; Chen and Jin, 2015; Nicolaou et al., 2011; Soleymani et al., 2016).

The LSTM cells contain a memory block and gates that let the information go through the connection of the LSTM. There are several connections into and out of these gates. Each gate has its own parameters that need to be trained. In addition to these connections, there are other hyperparameters such as the number of hidden neurons and batch size that need to be selected. Achieving good or even state-of-the-art performance with LSTM requires the selection and optimization of the hyperparameters, and there is no generic optimal configuration for all problem domains. It also requires a certain amount of practical experience to set the hyperparameters. Hence, LSTM can benefit from automatic hyperparameters optimization to help improve the performance of the architecture. This study focuses on optimizing the number of hidden neurons and batch size for LSTM networks.

Hyperparameter optimization can be interpreted as an optimization problem where the objective is to find a value that maximizes the performance and yields a desired model. Sequential model-based optimization (SMBO) is one of the current approaches that is used for hyperparameter optimization et al., 2011; Hutter, et al., 2011). One of the traditional approaches for hyperparameter optimization is grid search. This algorithm is based on an exhaustive searching. However, this algorithm takes a long period of time find the global optima. It has been shown that Random search can perform better than Grid search in optimizing hyperparameters (J. Bergstra & Bengio, 2012). Strategies such as Tree-based optimization methods, sequentially learn the hyperparameter response function to find the promising next hyperparameter combination.

Recently, several libraries for hyperparameter optimization have been introduced. One of the libraries which provides different algorithms for hyperparameter optimization for machine learning algorithms is the Hyperopt library

(J. Bergstra, et al., 2015). Hyperopt can be used for any SMBO problem, and can provide an optimization interface that distinguishes between a configuration space and an evaluation function. Currently, a three algorithms are provided: Random search, Tree-of-Parzen-Estimators (TPE) and Simulated Annealing (SA).

Random search, which is based on direct search methods, is popular because it can provide good predictions in low-dimensional numerical input spaces. Random search first initializes random solutions and then computes the performance of the initialized random solutions using a fitness function. It computes new solutions based on a set of random numbers and other factors, and evaluates the performance of the new solutions using the specified fitness function. Finally, it selects the best solution based on the problem objective (minimization or maximization). Although this algorithm is easy to implement, it suffers from slow convergence.

TPE, which is a non-standard Bayesian-based optimization algorithm, uses the tree-structure Parzen estimators for modelling accuracy or error estimation. This algorithm is based on non-parametric approach. Tree-based approaches are particularly suited for high-dimensional and partially categorical input spaces and construct a density estimate over good and bad solutions of each hyperparameter. In fact, TPE algorithm divides the generated solutions into two groups. The first group contains the solutions that improve the current solutions, while the second group contains all other solutions. TPE tries to find a set of solutions which are more likely to be in the first group.

SA method is another useful optimization method, which is inspired by the physical cooling process. There is a gradual cooling process in SA algorithm. This algorithm starts with random solutions (high temperature). This method iteratively generates neighbour solutions given the current solutions and accept the solutions with higher accuracy (lower energy). In the early iteration, the diversity of algorithm in generating new solutions is high. As the number of iterations increases the temperature decreases and the diversity of new solutions decreases. This process makes the algorithm effective in finding the optimum solutions.

Besides SMBO approaches, there are also existing strategies to optimize hyperparameters based on ECs (Kuremoto,et al., 2012; K. Liu, Zhang, & Sun, 2014;

Papa et al., 2015). EC algorithms such as Particle Swarm Optimization (PSO), and Differential Evolution (DE) are beneficial because they are conceptually simple and can often achieve highly competitive performance in different domains (Nakisa et al., 2014; Nakisa et al., 2014; Rastgoo et al., 2015). The advantage of using these algorithms is that while they are generating different solutions with high diversity, they can converge fast. Generating different solutions with high diversity enable the algorithms to find the potential regions with the possibility of optimum solutions. And the fast convergence property makes the algorithms to focus on the potential regions and find the optimal/ near optimal solutions in an acceptable time. It has been shown that PSO is efficient in finding the optimal number of input, hidden nodes and learning rate on time series prediction problems (Kuremoto et al., 2012). Although some studies have applied EC algorithms for hyperparameter optimization problems, especially on deep learning algorithms, this is the first study that utilizes DE to optimize LSTM hyperparameters on emotion recognition.

4.5 METHODOLOGY

In this section, the proposed approach to optimize LSTM hyperparameters using DE algorithm is presented. To optimize LSTM hyperparameters for emotion classification, three main tasks, pre-processing, feature extraction, optimizing hyperparameters of LSTM classifier are considered (see Fig. 1). In this study, we used the most effective pre-processing techniques as well as an effective set of features to keep the processes before classification to the best practice based on previous studies. We then focused on optimizing the performance of the LSTM with optimally selected hyperparameters. The proposed approach is evaluated on our dataset collected using wearable physiological sensors, which is described in the following section. The expected advantages of these sensors include their light-weight, and wireless nature, and are considered highly suitable for naturalistic research.

4.5.1.1 Description of the New Dataset

The dataset used in this study was collected from 20 participants, aged between 20 and 38 years, while they watched a series of video clips. To collect EEG and BVP signals, Emotiv Insight wireless headset and Empatica E4 were used respectively (see Fig. 4.2). This Emotiv headset contains 5 channels (AF3, AF4, T7, T8, and Pz) and 2 reference channels located and labeled according to the international 10-20 system (see Fig. 4.3).

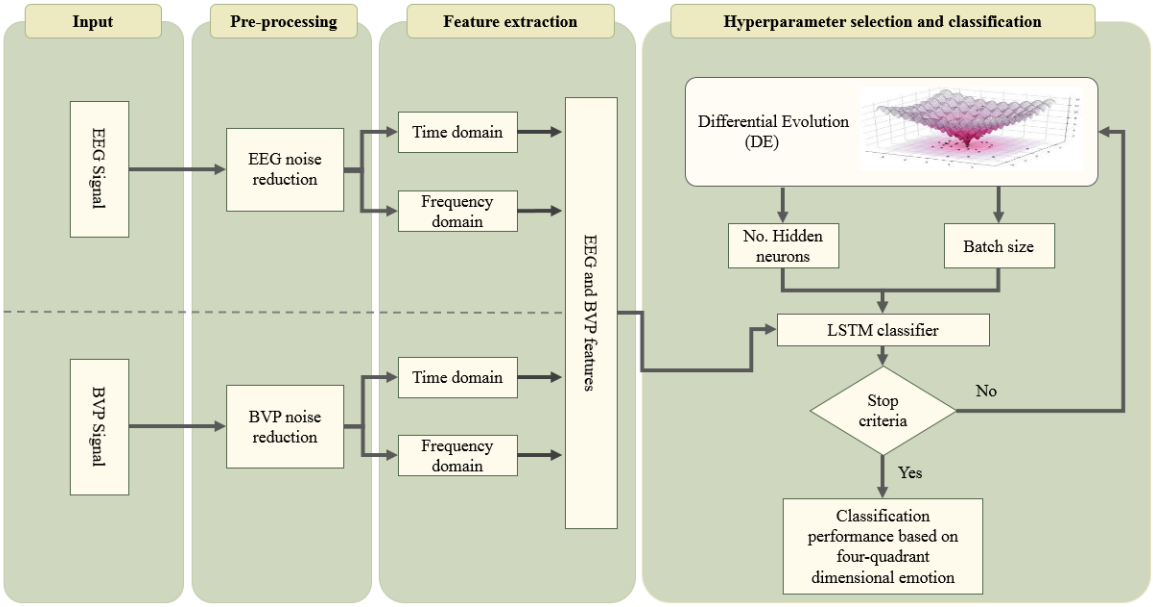


Figure 4.1. Framework to optimize LSTM hyperparameters using DE algorithm for emotion classification.

TestBench software and Empatica Connect were used for acquiring raw EEG and BVP signals from the Emotiv Insight headset and Empatica respectively. Emotions were induced by video clips, used in the MAHNOB dataset, and the participants' brain and heart responses were collected while they were watching 9 video clips in succession. The participants were asked to report their emotional state after watching each video, using a keyword such as neutral, anxiety, amusement, sadness, joy or happiness, disgust, anger, surprise, and fear. Before the first video clip, the participants were asked to relax and close their eyes for one minute to allow their baseline EEG and BVP to be determined. Between each video clip stimulus, one minute's silence

was given to prevent mixing up the previous emotion. The experimental protocol is shown in Fig. 4.4.

To ensure data quality, the signal quality for each participant was manually analyzed. Some EEG signals from the 5 channels were either lost or found to be too noisy due to the long study duration, which may have been caused by loose contact or shifting electrodes.

(a) (b)

Figure 4.2: (a) The Emotiv Insight headset, (b) the Empatica wristband.

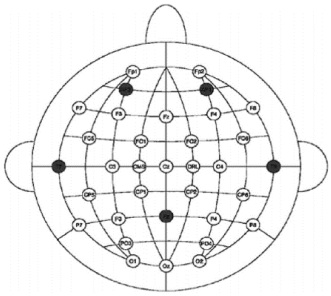


Figure 4.3. The location of five channels in used in emotive sensor (represented by black dot).

As a result, signal data from 17 (9 female and 8 male) out of 20 participants were included in the dataset. Despite this setback, the experiment allows an investigation into the feasibility of using Emotiv Insight and Empatica E4 for emotion classification purpose.

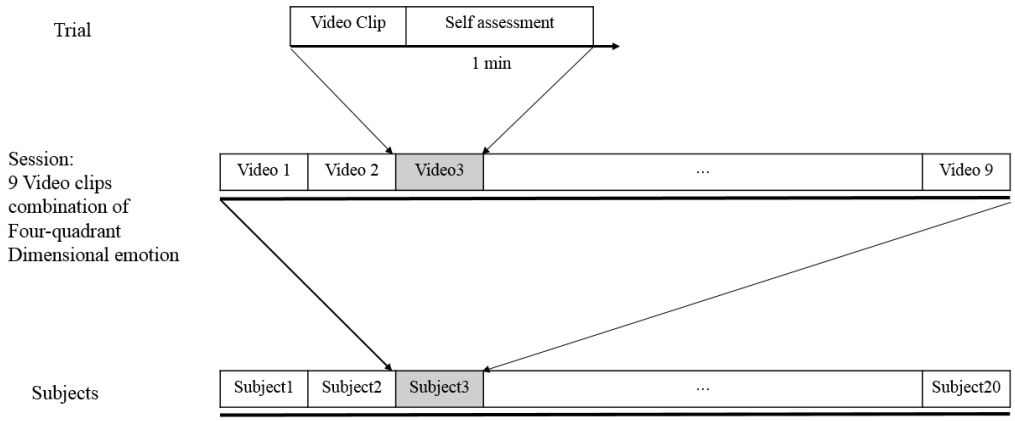
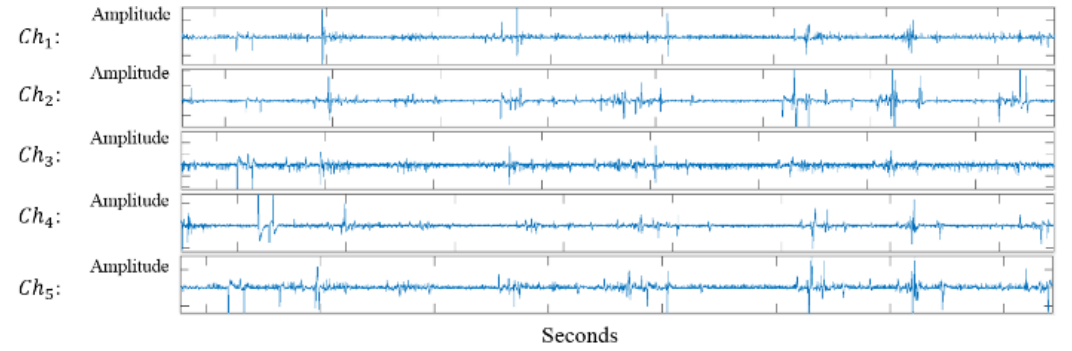


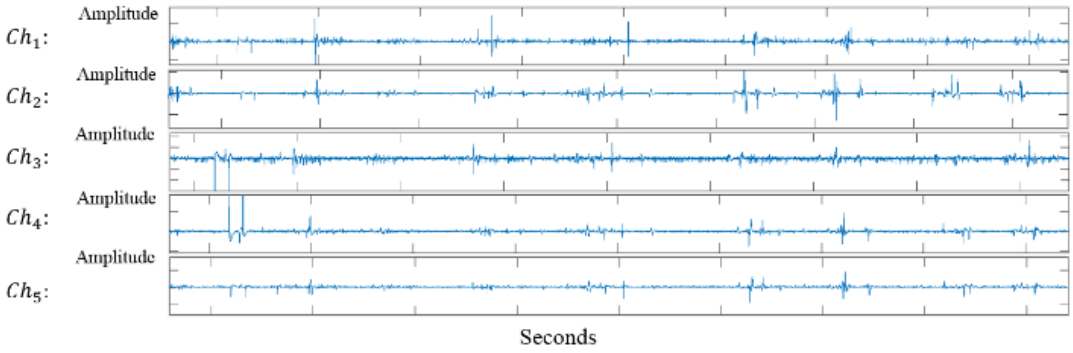
Figure 4.4. Illustration of the experimental protocol for emotion elicitations with 20 participants. Each participant watched 9 video clips and were asked to report their emotions (self-assessment).

4.5.2 Pre-Processing

To build the emotion recognition model, the critical first step is pre-processing, as EEG and BVP signals are typically contaminated by physiological artefacts caused by electrode movement, eye movement or muscle activities, and heartbeat. The artefacts generated from eye movement, heartbeat, head movement and respiration are below the frequency rate of 4Hz, while those caused by muscle movement are higher than 40Hz. In addition, there are non-physiological artefacts caused by power lines with frequencies of 50Hz, which can also contaminate the EEG signals. In order to remove artefacts while keeping the EEG signals within specific frequency bands, sixth-order (band-pass) Butterworth filtering was applied to obtain 4-64Hz EEG signals and cover different emotion-related frequency bands. Notch filtering was applied to remove 50Hz noise caused by power lines. In addition to these pre-processing methods, independent component analysis (ICA) was used to reduce the artefacts caused by heartbeat and to separate complex multichannel data into independent components (Jung et al., 2000), and provide a purer signal for feature extraction. To remove noise from the BVP signal, a 3 Hz low pass-Butterworth filter was applied. Fig. 4.5, 4.6 show the EEG and BVP signals before and after pre-processing task.

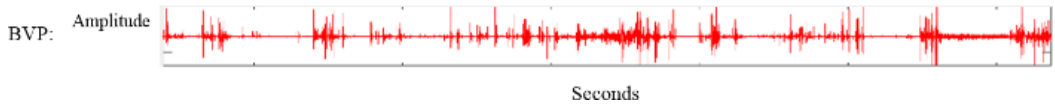


(a)

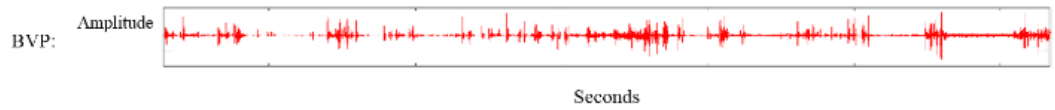


(b)

Figure 4.5: (a) Raw EEG signals before pre-processing, (b) EEG signals after pre-processing



(a)



(b)

Figure 4.6: (a) Raw BVP signal before pre-processing, (b) BVP signal after pre-processing.

4.5.3 Feature Extraction

EEG feature extraction

In our previous paper (Nakisa et al., 2017), the combination of time-domain and frequency-domain features showed more efficient performance compared to only time-domain, frequency-domain and time-frequency-domain features. Therefore, in

order to optimize the process of the feature extraction in this study, we have extracted features from time-domain and frequency-domain of EEG signals. In addition, one-second window size with 50% overlap is considered for feature extraction. It has been shown that this window size performs well on and is sufficient for capturing emotion recognition (Le & Provost, 2013; Soleymani et al., 2016).

Time-domain features have been shown to correlate with different emotional states. Statistical features – such as mean, maximum and minimum values, power, standard deviation, 1st difference, normalized 1st difference, standard deviation of 1st difference, 2nd difference, standard deviation of 2nd difference, normalized 2nd difference, quartile 1, median, quartile 3, quartile 4 can help to classify basic emotions such as joy, fear, sadness (Chai, Woo, Rizon, & Tan, 2010; K. Takahashi, 2004). Other promising time-domain features are Hjorth parameters: Activity, Mobility and Complexity (Ansari-Asl, Chanel, & Pun, 2007; Horlings, Datcu, & Rothkrantz, 2008). These parameters represent the mean power, mean frequency and the number of standard slopes from the signals and have been used in EEG-based studies on sleep disorder and motor imagery (Oh, Lee, & Kim, 2014; Redmond & Heneghan, 2006; Rodriguez-Bermudez, Garcia-Laencina, & Roca-Dorda, 2013). These features have been applied to real-time applications, as they have the least complexity compared with other methods (M. Khan, Ahamed, Rahman, & Smith, 2011). In addition to these well-known features, fractal dimension (Higuchi method), which represents the geometric complexity are employed in this study (Aftanas et al., 1998; Olga Sourina & Liu, 2011; Wang et al., 2010). Non-stationary Index (NSI) segments EEG signals into smaller parts and estimates the variation of their local average to capture the degree of the signals' non-stationarity (Kroupi, Yazdani, & Ebrahimi, 2011). Compared to time-domain features, frequency-domain features have been shown to be more effective for automatic EEG-based emotion recognition. The power of EEG signals among different frequency bands is a good indicator of different emotional states (Soleymani et al., 2016). Features such as power spectrum are extracted from different frequency bands, namely Gamma (30-64 Hz), Theta (13-30 Hz), Alpha (8-13 Hz) and Beta (4-8 Hz), as these features have been shown to change during different emotional states (Barry, Clarke, Johnstone, Magee, & Rushby, 2007; Davidson, 2003; Koelstra et al., 2012; Onton & Makeig, 2009).

Blood volume pulse (BVP) Feature Extraction

In this study, we employed the BVP signal from the PPG sensor of the Empatica E4 wristband. This sensor measures the relative blood flow in the hands with near infrared light, using photoplethysmography. From the raw BVP signal, we calculated the power spectrum density from three sub-frequency from the range of VLF (0-0.04Hz), LF (0.05-0.15Hz) and HF (0.16-0.4Hz), and the ratio of LF/HF (Akselrod et al., 1981). In addition, typical statistics features such as mean, standard deviation, and variance are calculated from time domain.

4.5.4 Optimizing LSTM Hyperparameters

The main objective is to optimize the hyperparameters of the LSTM classifier using the DE algorithm and achieve better performance on emotion classification. In this study, the parameters in the framework are: batch size and the number of hidden neurons. The DE algorithm starts from the initial solutions (initialize the hyperparameters), which is randomly generated, and then attempts to improve the accuracy of emotion classification model iteratively until stopping criteria is met. The fitness function is LSTM networks which are responsible for performing this evaluation and returning the accuracy of emotion classification.

Differential Evolution (DE)

Proposed by Storn and Price (1996), DE was developed to optimize real parameters and real value functions. This algorithm, which is a population-based search, is widely used for continuous search problems (Price et al., 2006). Recently, its strength has been shown in different applications such as strategy adaptation for global numerical optimization (A. K. Qin et al., 2009) as well as feature selection for emotion recognition (Nakisa et al., 2017). Like Genetic Algorithm, DE also uses the crossover and mutation concepts, but it has explicit updating equation. The optimization process in DE consists of four steps: *initialization, mutation, cross over* and *selection.*

In the first step, *initialization,* the initial population $S_i^t = \{s_{1,i}^t, s_{2,i}^t, \dots, s_{D,i}^t\}, i = 1, \dots, Np$ is randomly generated, where *Np* is the population size and *t* shows the current iteration, which in the *initialization* step is equal to zero. It should be mentioned that for each dimension of the problem space there may be some integer ranges that are constrained by some upper and lower bounds: $s_j^{low} \leq s_{j,i} \leq s_j^{up}$, for j=1, 2, ..., D, where D is the dimension of the problem space. As this study focuses on optimizing the number of hidden neurons and batch size for LSTM networks, therefore, the dimension of the problem is two (D=2). Each vector forms a candidate solution s_i^t, called target vector, to the multidimensional optimization problem. In the second step, *mutation step*, three individual vectors $s_{j,p}^t, s_{j,r}^t, s^t{}_{j,q}$ from population S_i^t are selected randomly to generate a new donor vector v using the following mutation equation:

$$\mathrm{v}_{j,i}^t = s_{j,p}^t + F_i * (s_{j,r}^t - s_{j,q}^t) \quad (1)$$

Where F_i is a constant from [0, 2] denotes the mutation factor. It should be noted that the value of $\mathrm{v}_{j,i}^t$ for each dimension is rounded to the nearest integer value.

In the third step, *crossover step*, the trial vector is computed from each of D dimension of target vector s_i^t and each of D dimension of donor vector v_i^t using the following binomial crossover. In this step, every dimension of the trial vector is controlled by c_r, *crossover rate*, which is a user specified constant within the range [0, 1).

$$\mathrm{u}_{j,i}^t = \begin{cases} v_{i,j}^t \; if \; r_i \leq c_r \; or \; j = J_{rand} \\ s_{i,j}^t \; otherwise \end{cases} \quad (2)$$

Where J_{rand} is a uniformly distributed random integer number in range [1, D], and r_i is a distributed random number in range [0, 1]. In the final step, *selection step*, the target vector s_i is compared with trial vector u_i, and the one with higher accuracy value or lower loss function is selected and admitted to the next generation. It should be noted that in this study the process of selection is achieved by LSTM algorithm, and the vector with the better fitness value is kept. The last three steps continue until

some stopping criteria is met. The following Pseudo-code shows the steps of the DE algorithm as a hyperparameter optimization method. The following Pseudo-code shows the steps of the DE algorithm as a hyperparameter optimization method.

Pseudo-code for the DE algorithm with LSTM classifier

Define the size of the population *NP*, *D* dimension of problem, crossover rate c_r, scale factor *F*.

Initialization: Initialize the population $S_i^{t=0} = \{s_{1,i}^t, s_{2,i}^t, \dots, s_{D,i}^t\}$, $i = 1, \dots, Np$ *which each individual uniformly distributed in the range*$[s^{low}, s^{high}]$

While the termination criteria is not met

For each individual, target vector, in the population NP

Mutation: Select three individual from the population randomly and generate a donor vector $\mathrm{v}_{\mathrm{i}}^{\mathrm{t}}$ using the following mutation equation:

$$\mathrm{v}_{\mathrm{j,i}}^{\mathrm{t}} = \mathrm{s}_{\mathrm{j,p}}^{\mathrm{t}} + \mathrm{F_i} * (\mathrm{s}_{\mathrm{j,r}}^{\mathrm{t}} + \mathrm{s}_{\mathrm{j,q}}^{\mathrm{t}})$$

Crossover: Compute the trial vector for the *i*th target vector $\mathrm{u}_{j,i}^{t+1}$:

$$\mathrm{u}_{j,i}^{t} = \begin{cases} v_{i,j}^t & if \;\; r_i \le c_r \; or \;\; j = J_{rand} \\ s_{i,j}^t & otherwise \end{cases}$$

Selection: Apply LSTM classifier as fitness function *f* and evaluate s_i^t and u_i^t:

If $f(s_i^t) \le f(\mathrm{u}_i^t)$ *then* $s_i^{t+1} = \mathrm{u}_i^t$

Else $s_i^{t+1} = s_i^t$

End For

End While

Conventional LSTM

In this work, we applied a conventional LSTM network with two stacked memory block as a fitness function of the DE algorithm. LSTM network evaluates the performance of the emotion classification system based on the obtained values for the number of hidden neurons and batch size. The memory blocks contain memory cells with self-connections storing the temporal state of the network in addition to special multiplicative units called gates to control the flow of information (Hochreiter &

Schmidhuber, 1997). Each memory block in original architecture contains three gates: input gate, forget gate, and output gate. The input gate controls the flow of input activation into the memory cells. The forget gate scales the internal state of the cell before adding it as input to the cell through the self-recurrent connection of the cell, therefore adaptively forgetting or resetting the cell's memory. Finally the output gate controls the output flow of cell activation into the next layer.

4.6 EXPERIMENTAL RESULTS

We conducted extensive experiments to determine if the DE algorithm can be used as an effective hyperparameter optimization algorithm to improve the performance of the emotion classification using EEG and BVP signals. We compared the performance of the optimized LSTM by the DE algorithm with other well-known hyperparameter optimization algorithms.

The experiments focused on classifying the four-quadrant dimensional emotions (HA-P, LA-P, HA-N and LA-N) based on the optimized LSTM. The performance of the proposed system is evaluated on the collected data from 17 participants, while they were watching 9 video clips. It should be mentioned that the length of each video clip varied and in order to prepare data for LSTM network with variable length, we transferred data with the same length. In this case, we considered the maximum length and applied the zero-padding to pad variable length.

Before applying zero-padding, noise reduction techniques such as Butterworth, notch filtering and ICA were applied. After applying noise reduction and zero-padding, different features (time-domain, frequency-domain) from one second window with a 50% overlap were extracted from 5 EEG channels and a BVP signal. Twenty-five features from each window of each EEG channel and 8 features from each window of BVP signal were extracted.

Then, these features were concatenated to form a vector (EEG+BVP=25×5+8×1= 133 features) and feed into LSTM network. The LSTM network consisted of two stacked cells developed in TensorFlow. The number of input nodes corresponded to the number of extracted features from EEG and BVP signal (133 features) and the number of output corresponds to the number of target classes (4

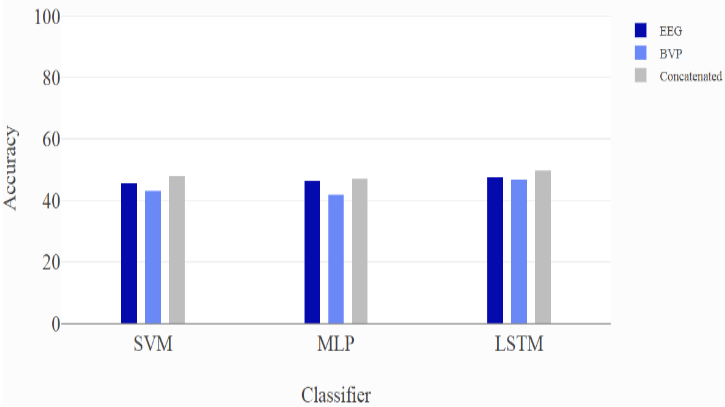
classes). The LSTM network used the following settings: lambda_loss_amount = 0.0015, learning rate= 0.0025, training epoch=10, number of inputs=133.

The results is produced using 17 leave-one-subject out cross validation. In this method one participant is used for testing and the remaining participants are used for training. Then the classification model was built for the training dataset and the test dataset was classified using this model to assess the accuracy. This process was repeated 17 times using different participants as test dataset. The 17 leave-one-subject out cross-validation trial is repeated 5 times to obtain a statistically stable performance of the recognition system. The overall emotion recognition performance (accuracy/ valid loss) is obtained by averaging the result of 5 times cross-validation trials.

4.6.1 Fusion of EEG and BVP

In this section, we assess the performance emotion classification using EEG signals, BVP signal and their fusion. To evaluate the performance of the fusion of EEG and BVP signals, we applied feature level fusion.

Fig. 4.7 presents the performance of each modality as well as the fusion of these two modalities using three well-known classifiers in emotion recognition (SVM, Multilayer Perceptron (MLP) and LSTM). In the previous sections, we chose to have a LSTM network with two stacked cells. The number of hidden neurons in this configuration is set to a quarter of the input layer of the features (133/4=33), and the batch size is set to 100. The result showed that the performance of LSTM classifier based on the fusion of EEG and BVP signals is better compared to SVM and MLP classifiers. It also showed that EEG signals from Emotiv are performing better than BVP signal generated from Empatica E4, and the performance of fusion of both sensors is higher than the single ones.



Classifier	EEG	BVP	EEG+BVP
SVM	**45.5%**	43.1%	47.9%
MLP	**46.2%**	41.9%	47.1%
LSTM	**47.5%**	46.8%	**49.6%**

Figure 4.7. The performance (classification accuracy) of each signal and their fusion at feature level using different classifiers.

4.6.2 Evaluating LSTM Hyperparameters Using DE and Other Methods

The performance of the LSTM classifier optimized by the DE algorithm is evaluated in this section. It is also compared with other existing methods (PSO, SA, Random search and TPE) based on the optimum accuracy and valid loss achieved within reasonable time. To apply Random search, TPE and SA, Hyperopt library was used, and the DE algorithm is developed in Python language programming. The following settings are applied on each hyperparameter optimization algorithm:

For the DE algorithm, the crossover probability is set to 0.2, bounds for batch size is between 1 and 271, bounds for the number of hidden neurons are set between 1 and 133 and the population size is equal to 5. For the SA algorithm with Hyperopt library, anneal.suggest algorithm is selected, bounds for batch size are placed between [1- 271] and the number of hidden neurons are set between [1- 133]. For the Random search with Hyperopt library, Random.suggest algorithm is chosen. Batch size is selected from the range of 1 and 271, and the number of hidden neurons are ranges from 1 to 133. For the TPE search with Hyperopt library, tpe.suggest is used, bound for batch size is between [1- 271], and the number of hidden neurons are ranges from 1 to 133. In order to find an optimized LSTM to classify emotions more accurately,

we ran each hyperparameter optimization algorithms with three different iterations (50, 100 and 300 iterations). These iterations assisted the algorithms to explore and find more solutions (maximum 1500 solutions=5×300, 5=population size, 300=number of iterations). In this experiment, each hyperparameter optimization algorithm is tested based on its ability to achieve the best values for the number of hidden neurons and batch size for LSTM classifier to improve the performance of the proposed system.

Table 4.1 presents the performance of all hyperparameter optimization methods, providing processing time, average accuracy ± standard deviation, best accuracy, average loss value± standard deviation, and the best loss found in different iterations. The presented mean accuracy, average time, mean loss value and their standard deviation are the result of averaging 5 times cross validation trials (17 one-leave-subject-out cross validation). Processing time is determined by Intel Core i7 CPU, 16 GB RAM, running windows 7 on 64- bit architecture.

Based on the average time, the processing time for SA algorithm was lower than the others, while its performance was lower than PSO and the DE algorithms and higher than Random search and TPE algorithms. In contrast, the DE took the most processing time over all iterations, however its performance was higher than all other algorithms.

Since the goal of this study is to introduce an optimized LSTM model that can accurately classify different emotions the computational time to build such an optimized classification model is less important, as long as it can achieve an acceptable result. In fact, finding the optimal classifier is done offline, and the time (358 hours for 17 cross fold validation at 300 iterations) to do this is within reasonable development time for such projects and will be considerably less with higher performance, parallel or cloud computing. Based on the best accuracy, LSTM classifier using the DE algorithm achieved significant result (77% accuracy) at 300 iteration, although its computational cost is expensive.

Table 4.1. The average accuracy, time, valid loss and best loss of different hyperparameter optimization methods

Hyperparameter optimization	No. iteration	Time	Accuracy (mean± std)	Best Accuracy	Valid loss (mean± std)	Best loss
Random search	50	25 h	53.47 ±15.9	67.35	22.47 ± 6.8	1.9
	100	58 h	59.50 ±12.9	68.59	21.68 ± 6.5	**1.1**
	300	174 h	59.21 ±16.1	68.63	21.39 ±11.9	1.5
TPE	50	29 h	58.16 ±12.1	64	24.48 ±5.1	11.3
	100	57 h	59.95 ±12.1	68.5	19.64 ±5.8	5.2
	300	172 h	60.86 ±14.46	68.9	17.45 ±8.9	3.1
SA	50	23 h	50.17±17.4	59.65	21.72±5.5	1.96
	100	47 h	60.98±12.4	68.65	21.05±5.1	1.73
	300	143 h	62.35±12.06	69.1	20.76±9.1	1.91
PSO	50	25 h	61.32 ±10.7	67.65	16.21 ±2.3	5.43
	100	55 h	**63.99±9.4**	69.98	14.84 ±2.6	3.44
	300	167 h	62.61±11.7	69.65	10.97 ±4.9	2.4
DE	50	59 h	**62.77±10.1**	**68.4**	**14.33±1.3**	**1.2**
	100	119 h	63.91±12.9	**73.36**	**12.65±2.8**	**1.14**
	300	358 h	**67.52±10.3**	**77.68**	**10.57±3.5**	**1.16**

It also demonstrated that the best achieved accuracies of all the other hyperparameter optimization algorithms are less than 70%, and this confirms the ability of the DE algorithm in finding a better solution. Based on the mean accuracy, the performance of DE algorithms followed by PSO and SA algorithms from the early iterations, was significantly better than Random search and TPE algorithms. From the 50 to 100 iterations, the improvement rate for DE and PSO algorithms was 3%, however, after 100 iterations the average accuracy of PSO algorithm decreased while for DE it improved slightly. This phenomenon is most likely due to the DE diversity property and its capabilities in searching and exploring more solutions. However, ECs did not achieve a significant result due to their convergence property and trapping in local optima (Rakshit et al., 2016; Udovičić et al., 2017). Trapping in local optima is

a common problem among ECs. Performance differences across different hyperparameter optimization methods were also tested for statistical significance using a two-way repeated measure ANOVA. The results showed that mean accuracy for each iteration (50, 100, 300 iterations) differed significantly between all hyperparameter optimization methods (F-ratio= 36.225, ρ-value<0.00001).

Further analysis of all algorithms and the relative distribution (accuracy) is shown in Fig. 4.8. From Fig. 4.8, it is apparent that as the number of iterations increased, the accuracies of LSTM classifier using all hyperparameter optimization methods are improved (the maximum accuracy is improved by about 17%). However, the median accuracy of the DE algorithm over all three iterations is higher than all the other algorithms, particularly in 300 iterations. Results showed that in 50 iterations, the median accuracies of LSTM using DE and PSO algorithms were higher than 60%, while the median accuracies of the other three algorithms (TPE, Random search and SA algorithms) were lower than 60%. Fig. 8 also shows that as the number of iterations increased, no significant improvement using TPE, Random search were found. However, the median accuracy of LSTM using DE was improved by 7%. The result confirmed that LSTM hyperparameter optimization using DE followed by PSO algorithms is more robust compared to TPE and Random search.

The changing process of algorithms within 300 iterations are presented in Fig. 4.9. We plot the average test accuracy of each algorithm over 17 cross validation trials. Based on the figure, the performance of all hyperparameters in early iterations are increasing and close to each other. However, after some iterations, the accuracy of TPE, SA and PSO algorithms remain steady with slight improvements. It is shown that as the number of iterations increased, the performance of DE improved more compared to other algorithms. It can find better solutions with higher accuracy, while PSO and SA algorithms stagnated to local optima due to less diversity in the nature of these algorithms.

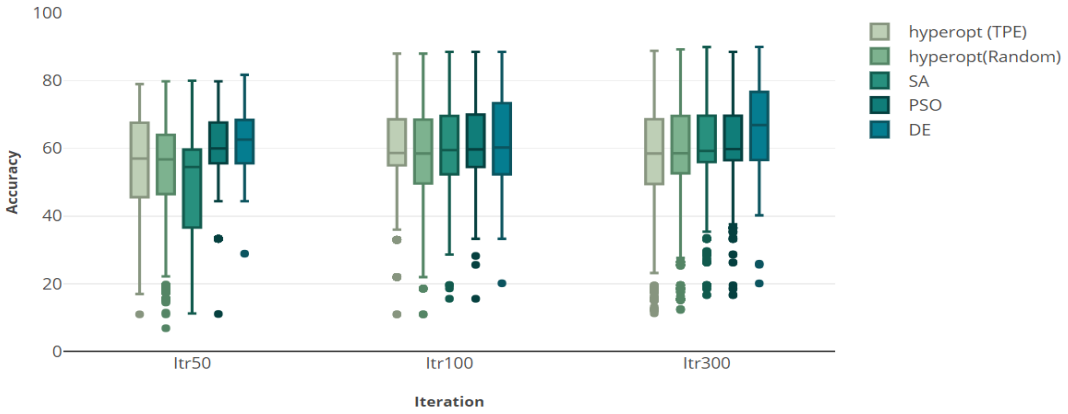


Figure 4.8. Box plots showing the distribution accuracy of the proposed system optimized by DE, PSO, SA, Random search and TPE algorithms in three different configurations on our collected dataset.

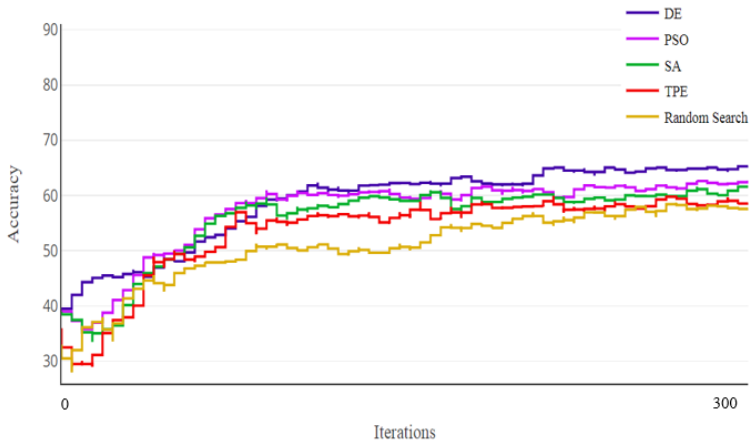


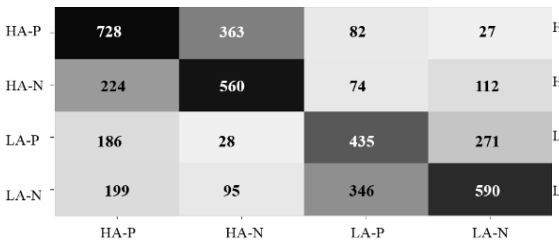
Figure 4.9. The average accuracies of the hyperparameter optimization methods over 300 iterations.

To show that how the optimized LSTM network can classify four-quadrant dimensional emotion and demonstrate the connection between the best achieved performances of the tuned LSTM model and different affects, confusion matrices of such models are presented (Tables 4.2. (a)- (f)). Table 4.2(a) presents the confusion matrix of LSTM without using any hyperparameter optimization algorithm. Using this approach the batch size and the number of hidden neurons are randomly selected. Tables 4.2(b)-(f) show the confusion matrices of the optimized LSTM classifiers using DE, PSO, SA, Random search and TPE algorithms as hyperparameter optimization techniques respectively. The provided confusion matrices show the best tuned LSTM model using the mentioned hyperparameter optimization algorithms at 300 iterations.

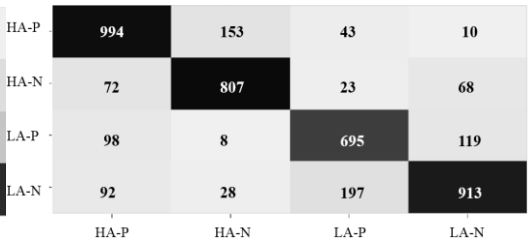
Based on Table 4.2(a), recognizing HA-P, HA-N and LA-N emotions are more difficult using the LSTM algorithm without any hyperparameter optimization algorithm. It is shown that recognizing HA-N is more confusing with HA-P and vice versa. LA-P quadrant is mostly misclassified as HA-P and LA-N, and LA-N is misclassified as LA-P and HA-P.

However, the optimized LSTM using the DE algorithm could classify all four-quadrant dimensions better than the other LSTM classifiers without and with hyperparameter optimization algorithm(s). The achieved confusion matrix from the best tuned LSTM classifier using the DE algorithm demonstrate that DE algorithm is able to find better values for the LSTM hyperparameters and improve the performance of emotion classification. Based on Table 4.2(b), the performance of the LSTM classifier using the DE in recognizing HA-P and HA-N is better than LA-P and LA-N. Although, the LSTM using the DE algorithm could classify these two quadrants better than the other hyperparameter optimization, it is still difficult to classify these two quadrant accurately. From the other Tables (4.2(c)-(f)), we can generally observe that recognizing LA-N using the optimized LSTM classifiers using SA, Random search and TPE is more difficult than the other three quadrants (HA-P, HA-N and LA-P), and this quadrant is mainly misclassified as LA-P and HA-P.

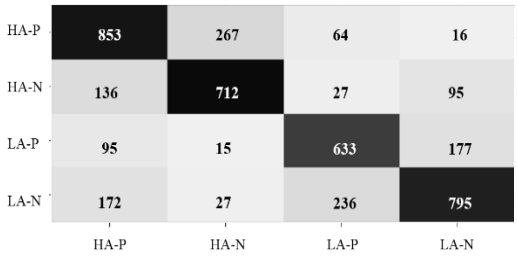
Table 4.2: The emotional confusion matrices corresponding to the LSTM models (a) without Hyperparameter optimization, the best achieved LSTM models (b) using DE algorithm, (c) using PSO algorithm, (d) using SA algorithm, (e) using Random search algorithm, (f) using TPE algorithm.



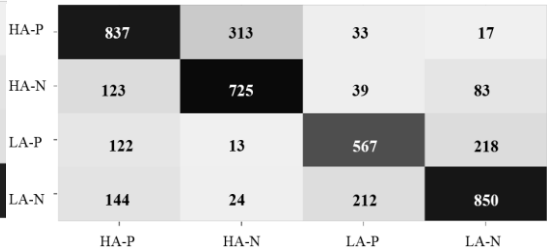
(a)



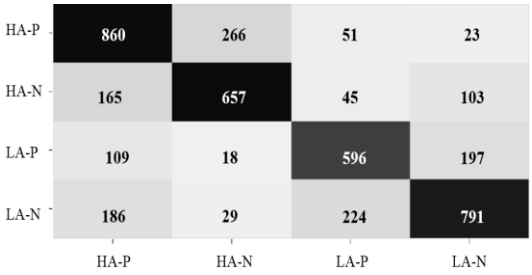
(b)



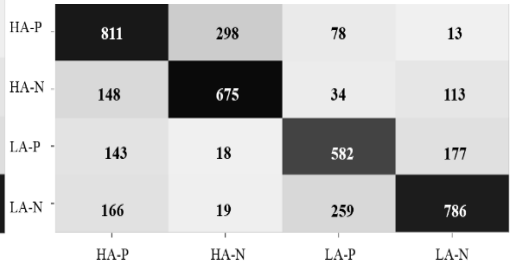
(c)



(d)



(e)



(f)

4.6.3 Comparison with Other Latest Works

The final experiment compared the performance of our proposed framework (the optimized LSTM by DE) against some other state-of-the-art studies. Experimental results are shown in Table 4.3, indicating the average classification accuracy for different emotion classification methods based on different number of classes as well as different classifiers. In addition, we compared the performance of emotion classification methods using tethered-laboratory and wireless physiological sensors. Tethered-laboratory sensors are wired sensors with high quality signals which are usually used for clinical purpose, but they are not deployable for open natural environments.

We used wireless physiological sensors in this study and these types of sensors can be used for real-life situations. The comparison shows that the average performance of both tethered-laboratory and wireless physiological sensors in two-class emotion classification is significant. However, the performance of EEG signals (tethered-laboratory and wireless) are better than BVP signal. The result also shows that SVM and decision tree can classify emotions into two-class better than the other classifiers. In three-class emotion classification, Rakshit et al. (2016) achieved higher

accuracy using tethered-laboratory BVP signal, while the performance of EEG signals, tethered-laboratory and wireless, is low. Moreover, Table 4.3 shows that the obtained accuracies by tethered-laboratory EEG sensor and wireless EEG sensor seem similar.

Candra et al. (2015) classified four different emotions (sad, relaxed, angry and happy) using SVM classifier. The result indicated that the average performance of four-class emotion classification using EEG signals (tethered-laboratory) is promising, which is better than three-class emotion classification. In our previous work (Nakisa et al., 2017), we achieved promising accuracy (65% accuracy on average) using the only wireless wearable EEG sensor. It should be stated that the average achieved performance using wireless EEG sensor can compete with tethered-laboratory EEG sensor with higher resolution.

In comparison with other works, our proposed system (optimized LSTM by DE) has shown the-state-of-the-art performance in classifying four-quadrant dimensional emotions using wireless wearable physiological sensors. The average achieved accuracy confirms that fusion of EEG and BVP signals along with the optimized classifier can help in classifying four-class emotions more accurately. Since the goal of this work is to find an optimized LSTM classifier with high performance, thus, the performance of the best model is also presented in the Table 4.3. The optimized LSTM classifier using DE algorithm (best model) achieved 77.68% accuracy in classifying four-quadrant dimensional emotions.

Table 4.3. Comparison of our approach with some latest works.

Method	**Classifier**	**No. classes**	**Accuracy**	**Sensor**
Sourina & Liu, (2011)	SVM	2 (Arousal, Valence)	90%	EEG (tethered-laboratory)
Ramirez & Vamvakousis, (2012)	SVM(RBF)	2 (Arousal, Valence)	Arousal: 78% Valence: 80%	EEG (Wireless)
Udovičić, Đerek, Russo, & Sikora, (2017)	SVM, KNN	2 (Arousal, Valence)	Arousal: 67% Valence: 70%	BVP& GSR (tethered-laboratory)

Xu, Hübener, Seipp, Ohly, & David,(2017)	J48, Naïve, KNN, SVM	2 (Arousal, Valence)	J48: 60% Naïve: 56% KNN: 52% SVM: 20%	BVP (tethered-laboratory)
Ragot, Martin, Em, Pallamin, & Diverrez, (2017)	SVM	2 (Arousal, Valence)	Arousal: 65% Valence: 69%	BVP (tethered-laboratory)
			Arousal: 65% Valence: 70%	BVP (Wireless)
Rakshit et al., (2016)	SVM	3 (Happy, Sad, Neutral)	83%	BVP (tethered-laboratory)
Ackermann et al., (2016)	SVM, Random forest	3 (Anger, Surprise, other)	55%	EEG (tethered-laboratory)
Feradov & Ganchev, (2014)	SVM	3 (Negative, Positive, Neutral)	62%	EEG (Wireless)
Candra et al., (2015)	SVM	4 (Sad, Relaxed, Angry, Happy)	60.9±3.2	EEG (tethered-laboratory)
Nakisa et al., (2017)	PNN	4 (HP-A, HA-N, LA-P, LA-N	65.04±3.1	EEG (Wireless)
Our work	LSTM	4 (HP-A, HA-N, LA-P, LA-N)	**66.92±9.3** **(Best model: 77.68%)**	EEG & BVP (Wireless)

4.7 CONCLUSION

In this study, we presented a new framework based on the use of a DE algorithm to optimize LSTM hyperparameters (batch size and number hidden neurons) in the context of emotion classification. The performance of the proposed framework is evaluated and compared with other state-of-the-art hyperparameter optimization algorithms (PSO, SA, Random search and TPE) over the new collected data using lightweight physiological signals (Emotiv and Empatica E4), which can measure EEG and BVP signals. This performance was evaluated based on four-quadrant dimensional emotions: High Arousal Positive emotions (HA-P), Low Arousal-Positive emotions (LA-P), High Arousal-Negative emotions (HA-N), and Low Arousal- Negative emotions (LA-N).

The results show that fusion of EEG and BVP signals provided higher performance for classifying four-quadrant dimensional emotions. In addition, the performance of the proposed system using different hyperparameter optimization methods was compared with different time intervals, 50, 100 and 300 iterations. The results demonstrated that that the average accuracy of the system based on the optimized LSTM network using DE and PSO algorithms improved over every time interval. The average accuracy of the proposed framework based on the optimized LSTM network using the DE is compared with other latest works and the result demonstrated that finding good values for LSTM hyperparameters can enhance emotion classification significantly. The better optimized LSTM classifier using the DE algorithm achieved 77.68% accuracy for four-quadrant emotion classification. However, the performance of the best achieved models using the other hyperparameters optimization algorithms is lower than 70% accuracy. This finding confirms the ability of the DE algorithm to search and find better solution compared to the other state-of-the-art hyperparameter optimization methods.

It should be noted that after a number of iterations (100 iterations) the performance of the system using EAs (PSO, SA and DE) did not change significantly. This could be due to the occurrence of premature convergence problem which may cause the swarm to be trapped into local optima and unable to explore other promising solutions. Therefore, it is recommended to explore a new development of EA

algorithms to overcome the premature convergence problem and further improve emotion classification performance. In addition, since the processing time of DE is more expensive, this can be reduced using parallel and/or cloud computing.

Chapter 5: Automatic Emotion Recognition Using Temporal Multimodal Deep Learning (Paper 3)

Bahareh Nakisa[1]

Mohammad Naim Rastgoo[1]

Andry Rakotonirainy[2]

Frederic Maire[1]

Vinod Chandran[1]

1. School of Electrical Engineering and computer Science, Queensland University of Technology, Brisbane, QLD, Australia

2. Centre for Accident Research & Road Safety-Queensland, Queensland University of Technology, Brisbane, QLD, Australia

Corresponding Author:

Bahareh Nakisa

Science and Engineering Faculty

Queensland University of Technology,

Brisbane, Australia

Phone: [redacted]

Email: Bahareh.nakisa@qut.edu.au

RESEARCH STUDENTS CENTRE
Examination Enquiries: 07 3138 1839
Email: research.examination@qut.edu.au

Statement of Contribution of Co-Authors for Thesis by Published Paper

The following is the suggested format for the required declaration provided at the start of any thesis chapter which includes a co-authored publication.

The authors listed below have certified that:

1. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
2. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
3. there are no other authors of the publication according to these criteria;
4. potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit, and
5. they agree to the use of the publication in the student's thesis and its publication on the QUT's ePrints site consistent with any limitations set by publisher requirements.

In the case of this chapter:

Title and status: **Automatic Emotion Recognition Using Temporal Multimodal Deep Learning (Submitted to Expert System with Applications, Q1 Journal)**

Contributor	Statement of contribution
Bahareh Nakisa	**Candidate** Experimental design, conducted the experimental work, performed data analysis, interpreted results, wrote the manuscript
Mohammad Naim Rastgoo	Assisted with paper planning, editing and proof-reading the paper

Andry Rakotonirainy	**Principal Supervisor** Assisted with paper planning, editing and supervise the experimental work and provide advice.
Frederic Maire	**Associate Supervisor** Assisted with paper planning, editing and supervise the experimental work and provide advice.
Vinod Chandran	**Associate Supervisor (External)** Assisted with paper planning, editing and supervise the experimental work and provide advice.

Principal Supervisor Confirmation		
I have Sighted email or other correspondence from all Co-authors confirming their certifying authorship.		
_____	_____	_____
Name	Signature	Date

5.1 INTRODUCTORY COMMENTS

Fusion of physiological modalities is an essential approach to obtain more accurate emotion classification model while unimodal data cannot represent a full understanding of subject's emotional state. Most of studies have used fusion techniques to build emotion classification model based on multimodal data. One of the common fusion techniques is to concatenate features from each modality and form one feature vector to solve the classification problem. However, these sort of approaches are not able to capture the non-linear correlation across data modalities, as the correlation between features within each modality is stronger. The non-linear emotional information across modalities can provide complementary information for emotion classification. Therefore, to build an automatic emotion classification system based on multimodal physiological signals, it is essential to capture both emotional information within and across physiological signals over time.

This chapter seeks to address Research Question 3: "How can physiological signals be fused to capture temporal emotional changes and improve emotion recognition?". In this chapter, we propose a new framework to fuse physiological signals based on multimodal learning approach. This approach is able to improve the performance of emotion recognition based on capturing the non-linear correlation within and across physiological signals over time. The proposed framework used convolutional neural network (ConvNet) and LSTM network in an end-to-end fashion. The proposed framework is investigated based early and late fusion models. The performance of the proposed models in this chapter is compared with the proposed model (handcrafted-features) from the previous chapter (Chapter 4).

Taken from: B. Nakisa, M. N. Rastgoo, A. Rakotonirainy, F. Maire and V. Chandran," Automatic emotion recognition Using Temporal Multimodal Deep Learning " Expert Systems with Applications, 2018.

Publication status: Under Review

Journal Quality: Expert Systems With Applications is a refereed international journal whose focus is on exchanging information relating to expert and intelligent systems applied in industry, government, and universities worldwide. The journal impact factor is **3.9,** and rank Q1 (SCImago) in Artificial Intelligence, Computer science applications and Engineering.

Copyright: The publisher of this article (ELSEVIER B.V.) stated that authors can use their articles in full or in part to include in their thesis of dissertation (provided that this is not to be published commercially).

5.2 ABSTRACT

Emotion recognition using miniaturized wearable physiological sensors has emerged as a revolutionary technology in different applications. However, detecting emotions using the fusion of physiological signals remains complex and challenging. Differences between sensors ranging from data type and sample rates itself make principled approaches to integrating these signals challenging. In the fusion of physiological signal it is essential to consider that how successfully the fusion approaches are able to capture the information contained within and across modalities. Moreover, physiological signals consist of time-series data, it becomes imperative to consider their temporal structures during the fusion process. In this paper, we focus on the fusion of multimodal information from electroencephalography (EEG) and blood volume pulse (BVP) signals. To best of our knowledge, most of studies in the literature extract the features from different modalities and then combined them by simply concatenating them into a long feature vector. In this paper, we proposed a temporal multimodal deep learning model to capture the non-linear correlation within and across modalities and improve the performance of emotion classification. We study the proposed model using different fusion levels (early and late). Specifically, we use convolutional neural network (ConvNet) long short-term memory (LSTM) to fuse EEG and BVP signals to jointly learn and explore the highly correlated representation across modalities after learning each modality with a single deep network. To validate the effectiveness of the multimodal fusion model based on ConvNet LSTM network, we performed experiments on a dataset collected from smart wearable sensors (Emotiv and Empatica wristband) and compared with handcrafted features. In addition, we evaluated the performance of the multimodal fusion models with different window sizes to find the most appropriate window size for an accurate emotion classification model. The experimental results show that the temporal multimodal deep learning models based on early and late fusion successfully classified four-quadrant dimensional emotions and achieved 71.61% and 70.17% accuracy respectively. The achieved performance using the proposed models also outperformed the handcrafted feature extraction method.

5.3 INTRODUCTION

Recent advances in automatic emotion recognition using miniaturized physiological sensors and advanced mobile computing technologies has become an important field of research in Human Computer Interaction (HCI) applications like computer game (Mandryk & Atkins, 2007), e-health applications (C. Liu, Conn, Sarkar, & Stone, 2008; Luneski, Bamidis, & Hitoglou-Antoniadou, 2008) and road safety (Nugraha, Sarno, Asfani, Igasaki, & Munawar, 2016). The key benefit of wearable technologies such as lightweight wireless headbands and smart watches is that they can be utilized while carrying out our daily life activities. These lightweight sensors record physiological signals like Electroencephalograms (EEG), blood volume pressure (BVP), skin conductance and skin temperature in an unobtrusive and non-invasive manner. Among the diversity of physiological signals, EEG and BVP are allowing the inference of human cognitive and emotional states. There are some studies that indicate a strong correlation between emotions such as sadness, anger, surprise and such physiological signals (Haag et al., 2004; K. H. Kim, Bang, & Kim, 2004).

Recent researches in building automatic emotion recognition have shown that the use of multimodal data can substantially improve the performance of classification (Haq & Jackson, 2011; Soleymani et al., 2016; W.-L. Zheng, Zhu, et al., 2014b). It has been shown that different types of data can be used to describe the same phenomenon. The data from multiple sources are correlated and there are some emotional related information across them which can provide complementary information. In order to exchange such information, it is important to capture the correlation between modalities with a compact set of latent variables. However, learning the latent emotional information between heterogeneous physiological data like EEG and BVP signals is a challenging problem. This is due to the fact that physiological signals are heterogeneous time-series and there are some emotional structures within and across modalities over time.

To date, several approaches are presented to solve multimodal fusion problem. One of the approaches is concatenating features data from each modality to form one feature vector and use it to solve classification problems. However, this approach is

not able to capture the non-linear correlation across data modalities, as the correlation between features in each modality is stronger (Ngiam et al., 2011). This is because, these sort of approaches focus on learning the patterns within each modality separately while giving up learning patterns that occurs simultaneously across multiple data modalities.

Therefore, to build a robust emotion recognition system using multimodal physiological signals, it is essential to propose a multimodal fusion model that can capture and learn the inherent emotional changes within each modality and as well as across them. We believe that a good model for multimodal learning should simultaneously learn a joint representation of multimodal, and temporal structure within each modality.

Recently, deep architecture and learning techniques have been shown to be effective in capturing non-linear correlation across multimodal data like audio-visual (Y. Kim, Lee, & Provost, 2013) and obtained state-of-the-art performance (Ngiam et al., 2011; Sohn, Shang, & Lee, 2014). Multimodal fusion methods have been proposed to jointly learn and explore the highly correlated representation across modalities after learning each channel data with single deep network. It should be noted that physiological signals are inherently temporal in nature, which means that the current pattern in signal is influenced by the previous ones. However, the multimodal networks like deep Autoencoder, Boltzmann Machine could not model the temporal multimodal fusion.

To address these challenges, we present temporal multimodal deep learning models with aim of fusing EEG and BVP signals to improve the performance of emotion classification.

Figure 5.1 shows a simple illustration of the temporal multimodal fusion model. The raw EEG and BVP signals are segmented into consecutive windows. In each window (time slice) the EEG signals and BVP signals are jointly learned using multimodal networks. The learnt joint representations across modalities at different windows are directly connected from start to end, which make the current window learn based on the previous one.

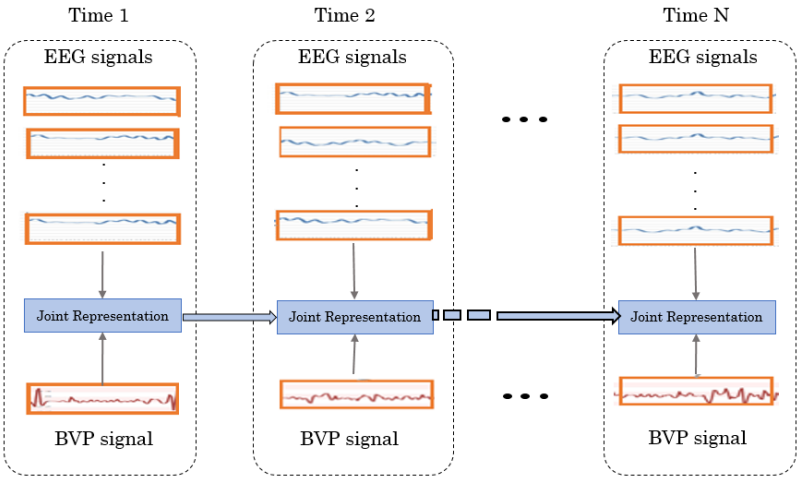


Figure 5.1. This model is proposed to demonstrate temporal multimodal fusion. The EEG channels and BVP signal are segmented into windows. The sequence of window from each channel is fed into a deep learning network and the output forms the joint representation across modalities. The generated joint representation based on the current window depends on the previous ones.

To date, in order to build an automatic emotion recognition based on the conventional models first features are extracted from physiological signals. The output of these networks are combined together and fed into an LSTM networks to find the emotional states. Using the existing approaches, each modality is trained on an individual network and the results are fed into a classifier. However, our system in this study is trained in an end-to-end fashion. Using end-to-end learning, the constructed features using ConvNets are trained jointly with the classification step as a single network. Moreover, in end-to-end learning approach, the network is trained from the raw data without any a priori feature extraction.

To our knowledge this is the first work that applies such an end-to-end temporal multimodal fusion model for emotion recognition based on EEG and BVP signals. The proposed models are based on convolutional neural networks long-Short Term Memory (ConvNets LSTM) network. The temporal multimodal fusion models based on ConvNets LSTM networks help in fusing EEG and BVP signals temporally to capture the temporal emotional structures within and across the modalities.

Furthermore, we investigate two approaches to fuse information across temporal domain: early and late fusion.

In this study, we categorize different emotional states based on arousal and valence into four quadrants (Fig. 5.2): 1-High Arousal-Positive emotions (HA-P); 2-Low Arousal-Positive emotions (LA-P); 3-High Arousal-Negative emotions (HA-N); 4-Low Arousal-Negative emotions (LA-N).

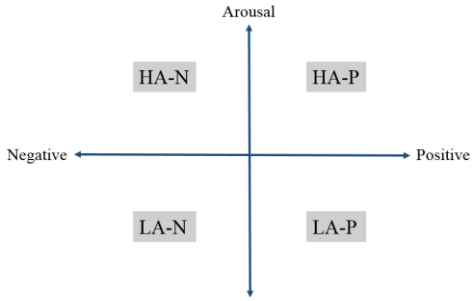


Figure 5.2. Categorized emotions into four quadrant dimensional emotion states.

In summary, the contribution of the proposed framework are as follows:

- We study two temporal multimodal deep learning models based on early and late fusion approaches using ConvNets LSTM in an end-to-end manner. The goal of these two models is to build an emotion classification model which can capture the temporal patterns within and across EEG and BVP signals and improve the performance of emotion classification based on four-quadrant dimensional emotions.
- The performance of the two temporal multimodal fusion models with different window sizes using sliding window strategy are evaluated and compared with non-temporal multimodal deep learning models using trial-wise strategy. In the trial-wise training the whole duration of a raw physiological signal per each video clip, called trial, as inputs and the corresponding trial emotion label as target are used for training.
- We also show that the performance of temporal multimodal fusion models can outperform the accuracy of handcrafted features extraction method for recognizing four-quadrant dimensional emotions on a dataset collected from wireless wearable sensors (Emotiv and Empatica wristband).

The rest of this paper is organized as follows: In section 2, the theoretical background is provided and the previous studies related to our proposed systems are reviewed. Section 3 presents the proposed methods. In section 4, we evaluate the performance of our systems based on our collected dataset for emotion recognition using wearable sensors.

5.4 BACKGROUND AND RELATED WORK

Recently automatic human emotion recognition using physiological signals like EEG, BVP and GSR (Koelstra et al., 2010) and physical data such as facial expression (Hossain & Muhammad, 2017) is increasingly became the subject of HCI applications. Physiological signals offer several advantages over physical data due to their sensitivity for inner feelings and insusceptibility to social masking of emotions (J. Kim, 2007). To understand inner human emotions, emotion recognition methods focus on changes in the two major components of nervous system; the Central Nervous System (CNS) and the Automatic Nervous System (ANS). The physiological signals originating from these two components carry information relating to inner emotional states.

Gathering physiological signals using miniaturized wearable sensors can provide non-invasive way to recognize different emotions continuously. Moreover, lightweight wearable sensors can be utilized while carrying out our daily life activities. Among different physiological signals, EEG and BVP signals have been widely used to recognize different emotions, with evidence indicating a strong correlation with different emotions (Haag et al., 2004; K. H. Kim et al., 2004). EEG signals, which measure the electrical activity of the brain, can be recorded by electrodes placed on the scalp. The strong correlation between EEG signals and different emotions is due to the fact that these signals come directly from the CNS, capturing features about internal emotional states. EEG-based emotion recognition systems have often had improved results when different modalities have been used (Chanel et al., 2006; Koelstra et al., 2010; K. Takahashi, 2004). Among the many peripheral physiological signals, BVP is a good indicator of recognizing different emotions (Haag et al., 2004;

A. M. Khan & Lawo, 2016; Kazuhiko Takahashi, 2004). BVP signal, which is measured by Photoplethysmography (PPG) sensor, indicates the blood flow rate controlled by heart bumping activity and is regulated by ANS. In general, the activity of ANS is involuntarily modulated by external stimuli and emotional states. Although its accuracy is considered lower than that of electrocardiograms (ECGs), due to its simplicity, it has been used to develop wearable biosensors in non-clinical applications such as detecting cognitive load of office workers in a controlled environment (F. Zhang et al., 2017).

5.4.1 Emotion Recognition Framework

Generally, to build an automatic emotion recognition system there are three main steps that should be considered: pre-processing, feature extraction and emotion classification. In the first main step, pre-processing, the noise and artefact are removed from the raw physiological data to prepare the data for modelling. In the next step, a set of features from the purer signals are extracted. In the final step, the extracted features are fed into the classifier to classify different emotions.

One of the most challenging steps in the pipeline of automatic emotion classification is feature extraction. Generally, there are two main approaches for feature extraction: handcraft feature extraction and deep learning techniques. To date, most of the reported approaches to recognize different emotions rely on extracting handcrafted features. This process accomplished either by using some conventional feature extraction algorithms or taking advantage of human expert knowledge. Several studies have focused on extracting features from EEG signals, identifying useful features such as time, frequency and time-frequency domains which can be used to recognize emotions. In a recent study (Nakisa et al., 2017), we proposed a comprehensive set of extractable features from EEG signals, and applied different evolutionary algorithms as feature selection to find the best possible subset of features and channels. There are some studies that focus on extracting features from time and frequency domain from BVP signal (Akselrod et al., 1981; Hui & Sherratt, 2018; Rani, Liu, Sarkar, & Vanman, 2006). Time domain features such as mean, standard deviation, variance from peak have shown a promising performance in recognizing

different emotions. It has been shown that the power spectrum density from three sub-frequencies: VLF (0-0.04 Hz), LF (0.05-0.15Hz) and HF (0.16-0.4Hz) and the ratio of LF/HF can distinguish different emotions accurately.

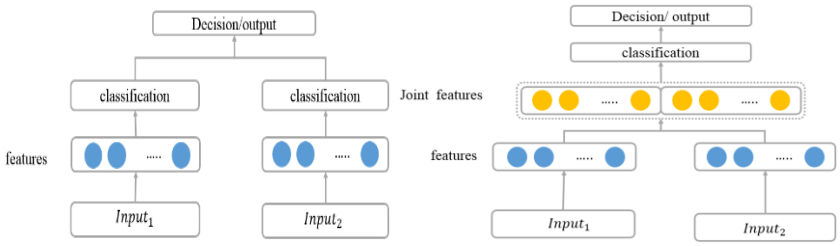
It should be noted that the performance of the emotion recognition model significantly depends on the quality of extracted features. Therefore, extracting most representative and critical features is always desirable. However, extracting suitable features using expert-knowledge (handcrafted feature extraction) is time-consuming and ad-hoc and the extracted features are not always robust to the variations such as noise, scaling. In addition, extracting features from physiological signals is challenging and requires a deep knowledge and expertise.

Recently, deep learning (DL) methods have increasingly emerged to solve the existing challenging recognition problems. DL methods are actively applied in multidimensional signal processing due to their state-of-the-art performance and strong capabilities in constructing reliable features in different fields such as speech recognition (Hinton et al., 2012) and time-series data analysis (Cecotti & Graser, 2011; Y. Zheng, Liu, Chen, Ge, & Zhao, 2014) . In emotion recognition, DL technologies have been studied to develop models of affect more reliable and accurate than the popular feature extraction-based affective modelling (Kahou et al., 2016; Y. Kim et al., 2013; Martinez, Bengio, & Yannakakis, 2013). There are several DL methods such as deep belief networks, recurrent neural networks and convolutional neural networks (ConvNet) that have been utilized in different domains to overcome various classification problems. Among the DL methods, ConvNet has been successfully used for constructing strong and suitable features for different problems. The strong feature learning capabilities of ConvNets make them an ideal and suitable choice for multidimensional signal processing applications. ConvNets consist of convolutional layers which can learn local pattern from the raw input. ConvNets are artificial neural networks that can learn local patterns in data by using convolutions as their key component. ConvNets vary in the number of convolutional layers, ranging from shallow architectures with just one convolutional layer such as in a successful speech recognition (Abdel-Hamid et al., 2014) over deep ConvNets with multiple consecutive convolutional layers (Krizhevsky, Sutskever, & Hinton, 2012) to very deep architectures with more than 1000 layers as in the case of the recently developed

residual networks (He, Zhang, Ren, & Sun, 2016). ConvNets with different layers can first extract local, low-level features from the raw input and then increasingly more global and high level features in deeper layers. There are some studies that applied ConvNets with different number of layers on physiological signals to classify different emotions (S. Chen & Jin, 2015; Martinez et al., 2013; Ringeval et al., 2015). One of the attractive properties of ConvNets that was leveraged in many previous applications is that they are well suited for end-to-end learning, i.e., learning from the raw data without any a priori feature extraction. End-to-end learning might be especially attractive in physiological signal decoding, as not all relevant features can be assumed to be known a priori. However, this technique are not well exploited in emotion recognition using wearable physiological signals particularly fusion of EEG and BVP signals.

5.4.2 Multimodal Learning to Recognize Emotions

Automated emotion recognition have had improved when different modalities have been used. The fusion of multimodal data can provide surplus information with an increase in accuracy of the overall result or decision. To date, there are mainly two levels of fusion are studies by researchers: early fusion, and late fusion (see Fig. 5.3). In early fusion, first different features are extracted from each modality, then all features from different modalities are concatenated to construct the joint feature vector. Finally, the joint feature vector are used to build the affective recognizer.



(a) (b)

Figure 5.3. Different fusion models. (a) Early fusion for temporal and non-temporal data. (b) Late fusion for temporal and non-temporal fusion.

One of the major advantages of early fusion is the detection of correlated features generated by different sensor signals that could improve recognition accuracy. However, the main drawback of the fusion based on this model is a little control over the contribution of each feature set from each modality on the final result and the augmented feature space can imply a more difficult classifier design, large training sets are typically required. Moreover, the obtained features from different modalities are different in many aspects like vector rate, therefore, multimodal learning in this level sometimes is difficult. There are some studies which showed that early level fusion can improve the performance of emotion recognition using different modalities (Caridakis et al., 2007; Gunes & Piccardi, 2005; W.-L. Zheng, Dong, & Lu, 2014).

Late fusion refers to the approach that the feature sets of each modality are examined and classified independently then the achieved results from each modality are fused as a decision vector to obtain the final result. The benefit of using late fusion in compare to early fusion is that it is easier to combine asynchronous data. Another advantage of using this fusion model is that every modality utilize its best classifiers which are suitable for the task. This may help to increase the performance of the decision. Late fusion is most commonly found in speech and gesture combination (L. Wu, Oviatt, & Cohen, 1999). However, it is almost certainly incorrect to use late fusion in real-time approaches and consider each modality independently and combine them at the end. In the real-time environment people display audio, video and tactile interactive signals in complementary and redundant information.

To accomplish a mutual analysis of multimodal which resembles human processing of such information, the input should be processed and learned temporally in a joint feature space and according to a context dependent model. To jointly learn data of different modalities and obtain state-of-the-art performance, temporal fusion is proposed.

Using this method the raw signal from each modality is segmented into some consecutive windows with a fixed size and the degree of overlap. Then each window from each modality is learned using multimodal networks. The fusion level depends on the designed multimodal networks. The learned joint representation across

modalities at different windows are directly connected from start to end, which makes the current window learned based on the previous ones. Recently, some studies focused on fusing different modalities using temporal fusion strategy for image and audio modalities (Hu & Li, 2016; Karpathy et al., 2014; Yu Liu, Chen, Peng, & Wang, 2017; X. Yang et al., 2017).

Physiological signals carries information in a sequence of time and they are temporal in nature, therefore, the influence of human emotion in physiological changes may misinterpreted if the temporal pattern of those changes is ignored. Thus, fusing physiological signals by considering time sequence can help in capturing temporal patterns in the fused data and obtain mutual information from multimodalities. This technique are not well exploited in emotion recognition using physiological signals using ConvNets LSTM based on end-to-end fashion.

5.5 MODELS

In this section, we describe the proposed temporal multimodal learning models based on early and late fusion models. The proposed models are based on ConvNets LSTM networks. These models are evaluated on our collected dataset using wearable physiological sensors (Emotiv and Empatica E4), which record EEG and BVP signals.

We first provide the description of the dataset in Section 3.1. Afterwards, Data preparation for temporal fusion is described (Section 3.2). Then we present two new frameworks for temporal multimodal learning based on early and late fusion models (Sections 3.3 and 3.4). Next the ConvNet architecture designed for this study is presented in Section 3.5.

5.5.1 Description of Dataset

The dataset used in this study was collected from 20 subjects, aged between 20 and 38, while they watched video clips. To collect EEG and BVP signals, Emotiv Insight wireless headset and Empatica E4 were used respectively (see Fig. 5.4). This Emotiv

headset contains 5 channels (AF3, AF4, T7, T8, and Pz) and 2 reference channels located and labeled according to the international 10-20 system (see Fig. 5.5).

(a) (b)

Figure 5.4. (a) The Emotiv Insight headset, (b) the Empatica wristband.

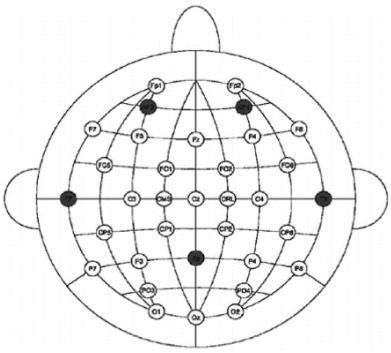


Figure 5.5. The location of five channels is used in emotive sensor (represented by black dot).

We used TestBench software and Empatica Connect for acquiring raw EEG and BVP signals from the Emotiv Insight headset and Empatica respectively. Emotions were induced by video clips, used in MAHNOB dataset, and the participants' brain and heart responses were collected while they were watching 9 video clips in succession. The participants were asked to report their emotional state after watching each video, using a keyword such as neutral, anxiety, amusement, sadness, joy or happiness, disgust, anger, surprise, and fear. Before the first video clip, the participants were asked to relax and close their eyes for one minute to allow their baseline EEG and BVP to be determined. Between each video clip stimulus, one minute's silence was given to prevent mixing up the previous emotion. The experimental protocol is shown in Fig. 5.6.

To ensure data quality, we manually analyzed the signal quality for each subject. Some EEG signals from the 5 channels were either lost or found to be too noisy due

to the long study duration, which may have been caused by loose contact or shifting electrodes. As a result, only signals data from 17 (9 female and 8 male) out of 20 participants are included in this dataset. Despite this setback, the experiment with this new data allows an investigation into the feasibility of using Emotiv Insight and Empatica E4 for emotion classification purpose.

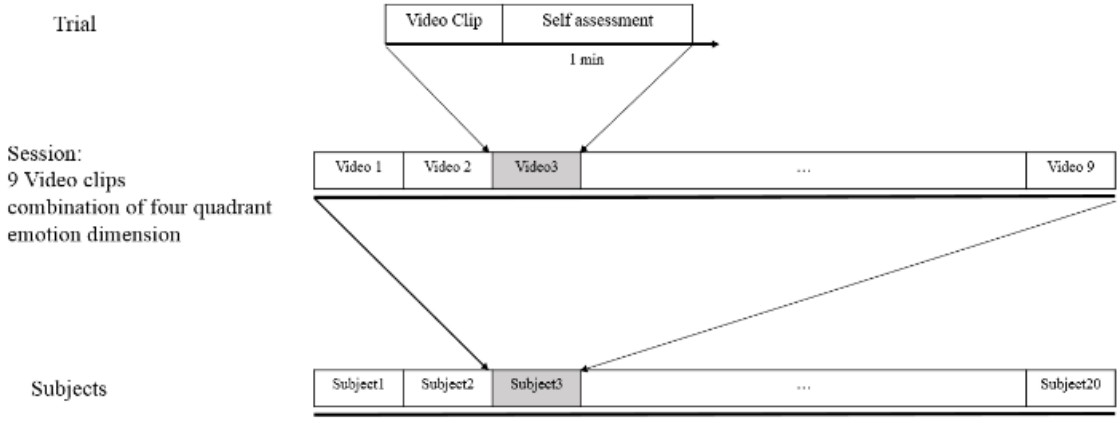


Figure 5.6. Illustration of the experimental protocol for emotion elicitations with 20 participants. Each participant watched 9 video clips and were asked to report their emotions (self-assessment).

The expected benefit of these sensors are due to their light-weight, and wireless nature, making it possibly the most suitable for free-living studies in natural settings.

5.5.2 Data Preparation for the Proposed Models

To prepare data, it is assumed that we are given 9 trials per subject as each subject watched 9 video clips. Each trial is labelled with different classes. There are 6 channels for each trial, 5 EEG channels and one BVP channel. To prepare the data for the purpose of temporal fusion, we use the sliding window strategy on each channel per each trial. Using this strategy, we applied the sliding window and create the set of consecutive windows with a fixed size and the degree of overlap as time slice of the trial. Let us denote 6-channel inputs as sequences of length *T*, namely

$$EEG\_ch_1 = (ch_1^1, \dots, ch_1^{t-1}, ch_1^t, \dots, ch_1^T),$$

$EEG\_ch_2 = (ch_2^2, \dots, ch_2^{t-1}, ch_2^t, \dots, ch_2^T)$,

…,

$EEG\_ch_5 = (ch_5^1, \dots, ch_5^{t-1}, ch_5^t, \dots, ch_2^T)$,

$BVP = (BVP^1, \dots, BVP^{t-1}, BVP^t, \dots, BVP^T)$

Where $\mathrm{ch}_1^t, \dots, \mathrm{ch}_5^t, \mathrm{BVP}^t$ denote the window of $\mathrm{EEG\_ch}_1, \dots, \mathrm{EEG\_ch}_5, \mathrm{BVP}$ at time slice t. All of the windows are the new training data examples for our model and will get the same labels as their original trials. In this study we segment each channel into different window sizes (2sec, 3sec, 5 sec and 10-sec long) sequence windows and 50% overlap. It should be noted that before applying sliding window strategy on each modality, some pre-processing techniques are applied on both EEG and BVP signals to provide the purer signals. The pre-processing techniques such as band-pass filtering (6^{th}order Butterworth filtering), Notch filtering and ICA are applied on EEG signals. To remove noise and artefact from BVP signals, a 3Hz low-pass Butterworth filter is applied. In addition, we normalize our data with zero mean and unit variance.

5.5.3 Temporal Multimodal Learning Based On Early Fusion

In this section, temporal multimodal learning model based on early fusion is presented. The proposed model aims at fusing the temporal physiological signals into the joint representation sequence at early level (after ConvNets). In this section the architecture of the proposed model is described (see Fig 5.7). Specifically, our model consist of input layer, ConvNets, feature map, early fusion and classifier.

Input. Temporal multimodal learning model strongly depends on its inputs. To apply the temporal multimodal learning, the sliding window strategy is applied on each EEG and BVP channels. Using this strategy, 6-channel physiological signals are segmented into consecutive windows with a fixed window size and a degree of overlap. The window at time t from each channel is considered as an input to be fed into ConvNets for training.

ConvNets. For each channel, the input, the sliced window from each channel at time t, is fed into the 2-block feature extractor. This feature extractor learns hierarchical features through convolution, activation, normalization and max-pooling layers. Since in this study physiological signals are used, thus we should apply 1D convolution layer.

There are more details about the ConvNet architecture in Section 3.5. Based on this architecture, we separate EEG channels and perform individual ConvNets on each of them as well as BVP channel, at window t. The output of ConvNet from each channel at time t is the corresponding feature map.

Feature map. If $ConvNet_{ch1}^{t}$ denotes the ConvNet for EEG_ch$_1$ and if FM_{ch1}^{t} denotes the corresponding feature maps at window *t*, then:

$$FM_{ch1}^{t} = ConvNet_{ch1}^{t}(\mathrm{ch}_1^{t})$$

In order to achieve temporal ConvNet learning for each channel both current input and its history are considered. Specifically, at window t the recent per-modality history ($ConvNet^{t-1}$) is appended to the current window to obtain the feature maps representation. The prepared feature maps at time *t* from each EEG channel are concatenated to form the EEG joint representative feature map at time *t*.

Early fusion. In this layer, the EEG joint representation feature map and BVP feature map at time *t* are combined to create one vector feature map from all the channels for time *t.*

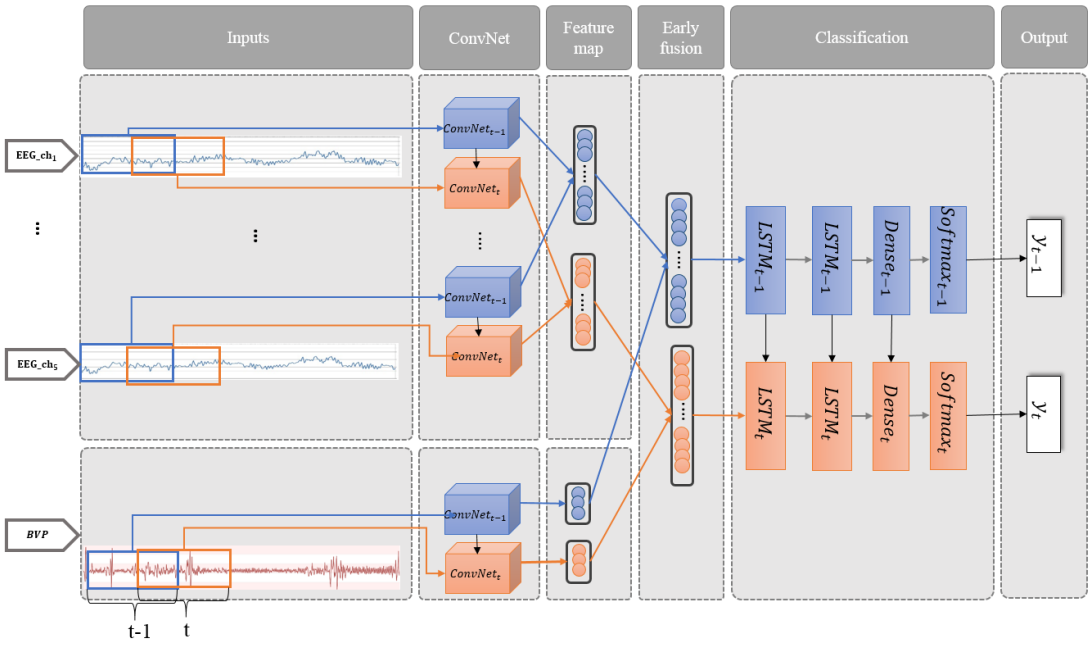


Figure 5.7. The overall pipeline of the temporal multimodal learning based on early fusion using ConvNets LSTM networks in an end-to-end learning fashion. The input of this system is the EEG (5 EEG channels) and BVP (one channel) signals, which are segmented into consecutive windows with a fixed size and the degree of overlap (50% overlap). The output of this model is four-class dimensional emotions (HA-P, HA-N, LA-P and LA-N). Each window at time *t* from each 6-channel are fed into individual two-block ConvNets to extract feature maps. The output of feature maps from channels over the window *t* are concatenated to form the joint representation from all the channels for that window. The joint representation at time slice *t* are fed into the two-layer LSTM followed by a dense layer and soft-max layer for emotion classification.

Classification. In this layer, the two-layer LSTM networks followed by a dense and Softmax layers are used to model the overall temporal dynamics of the multimodal feature representation at time *t.* It should be noted that an LSTM network consist of hidden state or memory, which can help in storing its previous hidden layers. Thus, the output of the $LSTM^t$ is computed using the current state as well as the previous hidden states (*t-1*), which can capture the temporal pattern of the previous joint representations. Although the extracted features from EEG and BVP signals might not be synchronized or delayed due to the nature of physiological signals, the LSTM memory blocks as hidden units have access to past information, which renders it ideal for processing data with different sample rates.

We should emphasize that this architecture not only try to learn the temporal pattern from each channel individually, but it also learns the temporal patterns from across modalities using the joint representations. This architecture is fully trained in an end-to-end manner and does not require any explicit feature extraction.

5.5.4 Temporal Multimodal Learning Based on Late Fusion

This section presents temporal multimodal learning model based on late fusion. Based the proposed architecture, EEG channels and BVP signal are temporally fused based on the late level fusion (after dense layer). The architecture of this model consists of input, ConvNet, feature map, classification, late level fusion and output layers (see Fig. 5.8). The input, ConvNets and the feature map layers in this architecture are the same as temporal multimodal learning model based on early fusion.

Input. First the windows from each 6 channels (5 EEG channels and 1 BVP signal) at time t are fed into the ConvNets.

ConvNets. For each channel, the input, the sliced window from each channel at time t, is fed into the 2-block feature extractor. The architecture of ConvNet in this architecture is the same.

Feature map. The feature map of each channels are generated by individual ConvNets. The feature maps of each EEG channels are concatenated to form a joint representative for EEG modality. The feature maps from each modality are fed into the two-layer LSTM networks followed by a dense layer.

In *late fusion layer*, the higher level feature maps generated from each modality (EEG and BVP signals) at time t are concatenated to form a joint representative layer from both modalities. The joint representative layer at time t is fed into a dense and a Softmax layers to classify emotions.

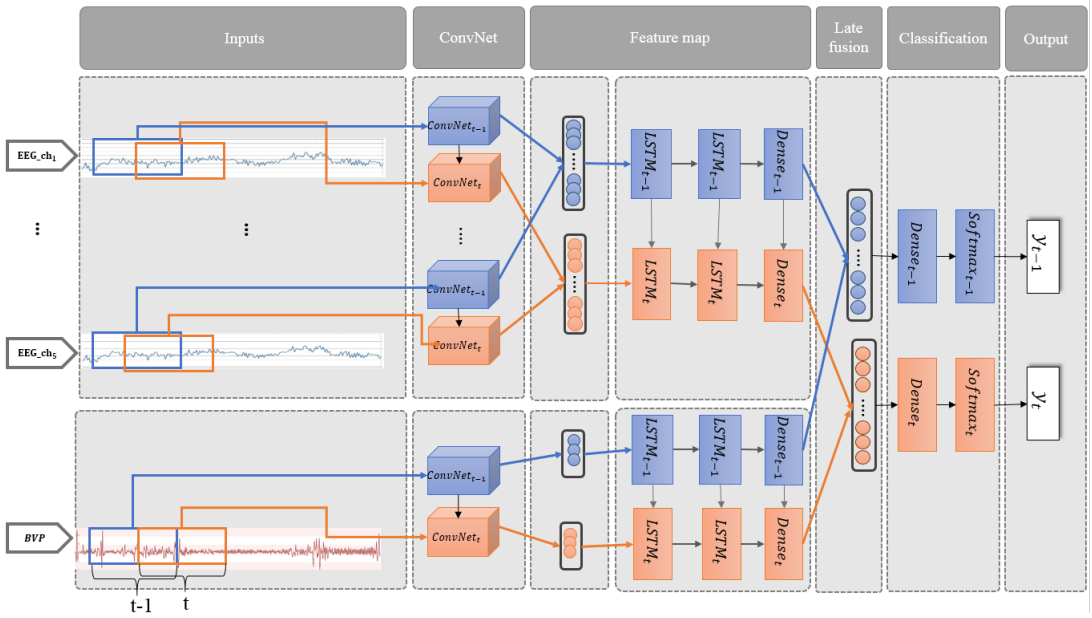


Figure 5.8. The overall pipeline of the temporal multimodal learning based on late fusion using ConvNets LSTM networks in an end-to-end learning fashion. The input of this system is the EEG (5 EEG channels) and BVP (one channel) signals, which are segmented into consecutive windows with a fixed size and the degree of overlap (50% overlap). The output of this model is four-class dimensional emotions (HA-P, HA-N, LA-P and LA-N). Each window at time *t* from each 6-channel are fed into individual two-block ConvNets to extract feature maps. The output of feature maps from EEG channels over the window t are concatenated to form the joint representation. The joint respresetation at time slice t from each modality (EEG and BVP) are fed into the two-layer LSTM followed by a dense layer. The output of dense layer at time *t* from two modalities are combined to create a joint representative and then is fed into a dense layer followed by a Softmax layer for emotion classification.

5.5.5 ConvNet Architecture for the Raw Physiological Signals

Generally, the convolutional neural network is used for many learning tasks on natural signals like image and audio. ConvNets often learn local non-linear features and present higher-level features as a composition of lower-level features. It consists of convolutional layers, which can produce lower-level features using a set of learnable filters, some multiple layer of processing, which can represent the higher-level features. In addition, many ConvNets use pooling layer to reduce the spatial size of representation and amount of parameters and computation in the network, and hence control overfitting. The presented ConvNet in this study is composed of two convolutional-max-pooling blocks (see Table 5.1.).

Each block is constituted by a convolutional layer, an Exponential Linear Unit (ELU), a batch normalization and a max-pooling layer. Convolution Layer convolves the input (window t) or the previous layer's output with the set of filter (K) to be learned. It capture the temporal information using trainable filters with the fixed small size. The output of each filter is computed according to

$$y = frame^t * K + b$$

Where b is the bias term, and the $*$ is convolution operator. The activation function Exponential Linear Units (ELU), that maps the output of previous layer by the function of the $ELU(x) = alpha * (\exp(x) - 1)\ x < 0, ELU(x) = x\ x \geq 0$; (iii) a batch normalization layer that normalize the value of different feature maps in the previous layer; (iv) a max pooling layer, which finds the maximum feature maps over a range of local neighbourhood.

Table 5.1. Two-block of ConvNets architecture.

ConvNet
Convolutional Layer: Filter=20,kernel size=(10,1), stride=2
Exponential Linear Units(ELU): Alpha=0.1
Batch Normalization+ Dropout (0.15)
Max-Pooling: Pool-size=(2,1), stride=2
Convolutional Layer: Filter=20,kernel size=(10,1), stride=2
Exponential Linear Units(ELU): Alpha=0.1
Batch Normalization+ Dropout (0.15)
Max-pooling: Pool-size=(2, 1), stride=2

5.6 EXPERIMENTAL RESULTS

For the quantitative evaluation, we used the collected dataset using wearable physiological signals (EEG and BVP signals) to analyse human affective states. With

investigating the temporal multimodal learning models based on early and late fusion models and compared with non-temporal multimodal learning models and handcraft feature extraction methods, our experimental results show that the proposed temporal multimodal learning models are effective in building an automatic human emotion recognition system using EEG and BVP signals.

To evaluate the performance of the two proposed models, first the efficacy of sliding window strategy is investigated. In this regards, we evaluated and compared the performance of the temporal multimodal models based on early and late fusion approaches with different window sizes (Section 4.1). Moreover, these two models are also evaluated based on non-temporal strategy. In non-temporal strategy, the input layer is based on the trial-wise strategy instead of sliding window strategy.

In Section 4.2, the confusion matrices of the best achieved temporal multimodal models in classifying four-quadrant dimensional emotions are presented. Lastly, in Section 4.3, the best average performance of both models which are based on the ConvNets LSTM networks are compared with the conventional handcraft feature extraction method.

5.6.1 Experimental Setup

We conducted extensive experiments to determine if the proposed temporal multimodal learning models based on early and late fusion can be used as an effective fusion methods for automatic emotion classification using EEG and BVP signals. The experiments focused on classifying the four-quadrant dimensional emotions (HA-P, LA-P, HA-N and LA-N). The results is produced using 17 leave-one-subject out cross validation (LOSO), which is subject-independent. In this method one participant is used for testing and the remaining participants are used for training. Then the classification model was built for the training dataset and the test dataset was classified using this model to assess the accuracy. This process was repeated 17 times using different participants as test dataset. The Physiological signals (EEG and BVP signals) are segmented into consecutive windows with a fixed size and the degree of overlap. In this study, we have chosen different window size to evaluate the performance of the

proposed temporal models using different window size. However, the degree of overlap is fixed and the raw physiological signals are segmented with 50% overlap.

Before the segmentation, some noise reduction techniques such as Butterworth, notch filtering and ICA were applied. The proposed multimodal learning models (early and late fusion) based on temporal and non-temporal approaches use two-layer LSTM networks followed by a dense layer with 100 and 20 hidden states respectively. To train our models, we used learning batches of 10 sequences. We also perform early-stopping on validation set. The above configuration is set as a good configuration, which yielded the minimum loss in the training set. The training is performed for 30 iterations.

5.6.2 Comparison of Temporal and Non-Temporal Models Based On Early and Late Fusion

In this section the effectiveness and performance of the two temporal multimodal learning models (early and late fusion) are evaluated based on different window sizes and compared with the non-temporal multimodal learning models. The aim of this comparison is to present the efficacy of sliding window strategy on emotion classification. To evaluate the efficacy of sliding window strategy on emotion classification using multimodal learning models, we investigated the performance of the two temporal multimodal learning models using different window sizes (2-sec, 3-sec, 5-sec and 10-sec long). Since the goal of this study is to build an automatic emotion classification model with the potential to be applied for real-time applications, the small window sizes for this study is selected.

The architectures of Non-temporal multimodal learning models are the same as temporal ones, only the input layer in these architecture is different. The input layer in the non-temporal models is based on the trial-wise strategy. In trial-wise training the whole duration of a raw physiological signals per each video clip, called trial, as inputs and the corresponding trial label as target are used for training the ConvNets. In this study, the EEG channels and BVP signals for different video clips are used for training. In our data collection, we are given 9 trials per 17 subjects.

Therefore, the whole number of training and testing samples is 9 * 17=153. It should be mentioned that the length of each video clip varied and in order to prepare data for ConvNets with variable length, we transferred data with the same length. In this case, we considered the maximum length and applied the zero-padding to pad variable length.

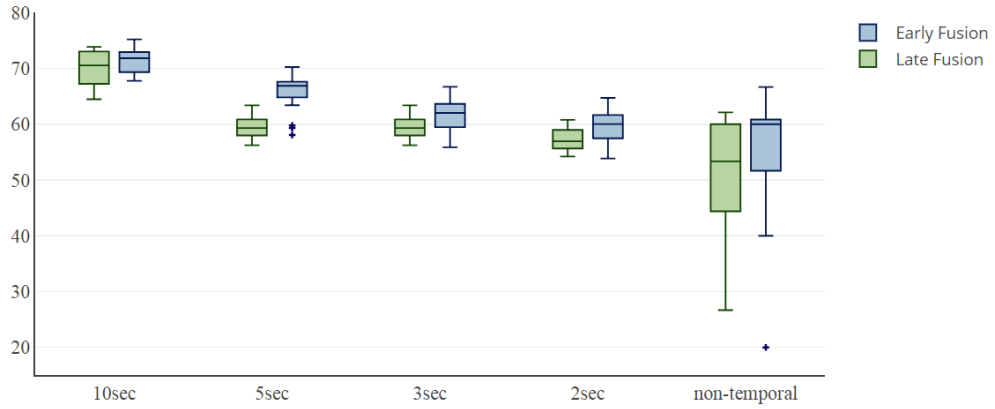


Figure 5.9. Box plots showing the accuracy distribution of the temporal multimodal learning model with different window sizes and non-temporal multimodal learning model based on early and late fusion.

Figure 5.9 presented the distribution accuracy of the two multimodal learning models based on non-temporal (trial-wise) and temporal with different window sizes, 2-sec, 3-sec, 5-sec and 10-se, at 30 iterations.

Based on Fig. 5.9, as window sizes increased the performance of emotion classification based on both temporal multimodal learning models are improved, and the best performance achieved at 10-sec long window size. It also shows that the both temporal models using ConvNets can capture and learn better spontaneous patterns from the bigger window size.

From the figure is can be seen that the performance of all temporal models with different window sizes are higher than non-temporal models. It shows that the sliding window strategy is essential for building an accurate emotion classification model. Moreover, the overall accuracy of temporal multimodal models based on early fusion

is slightly better than late fusion models. This means that not concatenating features at early level by capturing the correlated information across modalities can improve the performance of emotion classification. The average performance of temporal and non-temporal models, providing average accuracy ± standard deviation, average loss value ± standard deviation are presented in Table 5.2.

Table 5.2. The average performance of temporal multimodal learning models with different window size and non-temporal models.

Fusion Models	**Temporal model**								**Non-temporal model**	
	10-sec		**5-sec**		**3-sec**		**2-sec**			
	Accuracy	**Valid Loss**	**Accuracy**	**Valid Loss**	**Accuracy**	**Valid Loss**	**Accuracy**	**Valid Loss**	**Accuracy**	**Valid Loss**
Early fusion	**71.61± 2.71**	**0.62± 0.08**	65.5± 3.3	0.74± 0.03	61± 2.7	0.81± 2.4	56 ± 3.4	0.93± 0.07	55.07± 4.3	0.96± 0.13
Late fusion	70.17± 3.7	0.63± 0.10	64.4± 3.7	0.74± 0.09	59.4± 1.9	0.87± 2.4	55.9± 3.4	0.94± 0.05	52.28± 4.6	0.98± 0.15

Based on Table 5.2, the overall accuracy of the multimodal learning based on early fusion is higher than the late fusion based on both temporal and non-temporal approaches. It also shows that the performance of the both temporal models based on early and late fusion with window size bigger than 3 sec are significantly increased.

This not only confirms the efficacy of sliding window strategy in improving emotion classification using multimodal learning models, but also shows that there is no generic value for window size to be used for achieving an acceptable performance on emotion classification.

Since the goal of this study is propose an automatic emotion classification based on the temporal multimodal learning model with the potential to be applied for real-time applications, it is essential to choose the small window size which can result in high accuracy. In this regard, selecting window size with 10-sec and 5 sec long for classifying four-quadrant dimensional emotions could be an acceptable window sizes.

However, the performance of the model using 10-sec window size is more accurate than using 5-sec window size.

5.6.3 Evaluation Temporal Multimodal Learning Based On Early and Late Fusion

The best achieved performance of two proposed models with 10-sec window size to classify four-quadrant dimensional emotions are evaluated and compared in this section. Fig. 5.9 and 5.10 show the confusion matrices of four-quadrant dimensional emotions based on the temporal multimodal learning models based on early and late fusion with 10-sec window size. The provided confusion matrices show the average performance achieved by two temporal multimodal learning models at 30 iterations.

Figure 5.10. Temporal multimodal deep learning model based on early fusion.

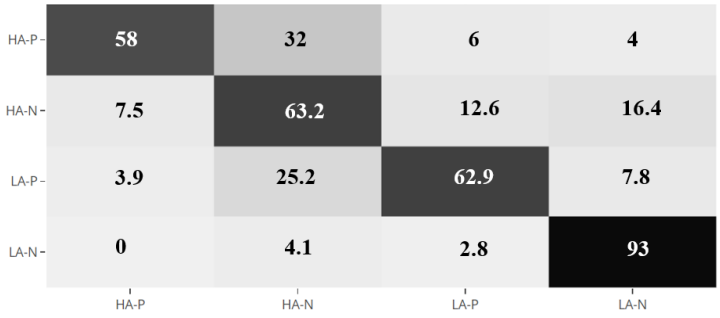


Figure 5.11. Temporal multimodal deep learning model based on late fusion.

Based on Fig. 5.10, recognizing high-arousal negative (HA-N) is more difficult than the other three quadrant emotions. It is also shown that recognizing HA-N is more

confusing with HA-P. LA-P quadrant is mostly misclassified as HA-N and HA-P. Among all four-quadrant dimensional emotions LA-N is classified more correctly and the performance of the built model in recognizing this quadrant is better than the others.

Fig. 5.11 shows that the performance of the built model using late fusion model in recognizing HA-N and LA-P is better than HA-P quadrant. HA-P quadrant is mostly misclassified as HA-N quadrant. It is also shown that the performance of this architecture in recognizing LA-P is as good as temporal multimodal learning model based on late fusion.

5.6.4 Comparison of Temporal multimodal models with handcrafted feature approach

In our previous study, we analysed the performance of the conventional handcraft feature extraction models using EEG and BVP signals. In the study, we proposed a framework to classify four-quadrant dimensional emotions accurately using an optimized LSTM classifier [21]. To compare the performance of the best achieved models with the conventional handcraft feature extraction method, we used the obtained performance from our previous study. Table 5.3 presents the average performance of the temporal models using ConvNets LSTM based on early and late fusion and the built model using the handcraft feature extraction approach. The table shows that the performance of temporal multimodal models using both early and late fusions are outstanding and could surpass the emotion classification using conventional handcraft feature extraction.

Table 5.3. The comparison of the best average performance of temporal multimodal models based on the early and late fusion and the conventional feature extraction model.

Models	Feature extraction	Classifier	Fusion	Accuracy
Nakisa et al. (2018)	Time and Frequency domain Features (25 features)	LSTM	Early fusion	66.92 ± 9.3
Our proposed models	ConvNets	LSTM	Early fusion	**71.61 ± 2.71**

Late Fusion	70.17 ± 3.7

5.7 CONCLUSION

In this study, we proposed two new frameworks using temporal multimodal learning models based on early and late fusion in the context of emotion recognition. The proposed temporal multimodal learning models are based on ConvNet LSTM networks using end-to-end fashion. The performance of the proposed models are evaluated on a collected dataset using wireless wearable sensors (Emotiv and Empatica wristband). In this dataset, the physiological signals of 17 participants during watching 9 video clips are recorded. In order to apply temporal multimodal learning models on physiological signals, sliding window strategy is utilized. Using this strategy the raw physiological signals are segmented into consecutive window with a fixed size and the degree of overlap. In this study we investigated the performance of the proposed models with different window sizes and compared with the non-temporal multimodal learning models using trial-wise strategy. In trial-wise training the whole duration of a raw physiological signals per each video clip, called trial, as inputs and the corresponding trial label as target are used for training. The performance of the temporal multimodal learning models using early and late fusion is higher than the multimodal learning models based on non-temporal strategy 71.61± 2.71 and 70.17± 3.7 vs. 55.07± 4.3 and 52.28± 4.6 respectively. Moreover, the average accuracy of temporal multimodal learning models based on early fusion with different window sizes are higher than the model based on late fusion model. The results showed that the temporal multimodal models based on early fusion at bigger window sizes are better than smaller window sizes. In this study the best achieved window size for multimodal learning based on EEG and BVP signals is 10-sec long at 71.61 ± 2.71 accuracy.

Chapter 6: Conclusions

6.1 INTRODUCTORY COMMENTS

This thesis contributed to the development methods and models to improve emotion recognition systems using portable wearable physiological signals (EEG and BVP signals). In this thesis, three main issues are addressed towards building robust and high performance emotion recognition systems using physiological signals: (1) the efficiency of evolutionary algorithms for the feature selection process to improve the performance of EEG-based emotion classification, (2) optimizing LSTM hyperparameters to enhance the performance of emotion classification using EEG and BVP signals, and (3) fusing EEG and BVP signals using ConvNet LSTM techniques to automatically maximize emotion classification based on raw physiological signals.

Three main studies were conducted to address the aforementioned issues and improve the performance of emotion classification. In Study 1, a comprehensive review of the-state-of-the-art EEG features was presented. A new framework using evolutionary algorithms was proposed to overcome the problem of high dimensionality and improve the performance of emotion classification. This study evaluated the performance of the proposed feature selection model on three different datasets (MAHNOB, DEAP and our dataset). The reliability and validity of mobile EEG sensors are also tested in this study.

Study 2 presented a framework to optimize LSTM hyperparameters using a DE algorithm and to maximize the performance of emotion classification using the fusion of EEG and BVP signals. The results of this study showed that the classifier with the automatically selected LSTM hyperparameters outperformed the classifier with manually selected hyperparameter values. This is the first study systematic study of hyperparameter optimization in the context of emotion classification. The effectiveness of the proposed method was evaluated on a dataset collected from portable wearable physiological sensors and compared with other well-known hyperparameters optimization methods.

In Study 3, we focused on the fusion of EEG and BVP signals and proposed a new framework to not only automate the process of fusion, but also to improve the performance of emotion classification. The proposed framework is based on a ConvNet LSTM network. Two different structures were proposed to fuse EEG and BVP signals: early fusion and late fusion. The study evaluated and compared the performance of the two multimodal fusion models using the dataset captured from portable physiological sensors.

This chapter discusses the summary of achievements from the three studies and the strengths and limitations of the research are outlined. Finally, the thesis is concluded with a number of recommendations for future studies.

6.2 SUMMARY OF ACHIEVEMENTS

Research Question 1: **What impacts do feature selection algorithms have on emotion recognition system?**

Since there is no standard set of EEG features to effectively classify different emotions, combining different features can lead to a high-dimensionality problem and reduce the performance due to redundancy and inefficiency. To overcome the problem of high-dimensionality, this study (Study 1) proposed a new framework to automatically search for the most salient set of EEG features using EC algorithms. This study comprehensively reviewed the state-of-the-art EEG-based feature extraction methods. EC algorithms can help to overcome the limitations of individual feature selection by assessing the subset of variables based on their usefulness. The main advantage of using EC algorithms to solve feature selection optimization problems is the ability of these algorithms to iteratively search within a set of possible solutions to improve the feature subset using a given measure of quality to find the optimal solution. The proposed framework has been extensively evaluated using two public datasets captured using an EEG sensor with 32 channels (MAHNOB and DEAP) and our new dataset collected from wireless EEG sensors with 5 channels.

The results confirm that evolutionary algorithms can effectively support feature selection to identify the best EEG features and the best channels to improve the classification performance of a four-quadrant dimensional emotion (HA-P, HA-N, LA-P and LA-N) classification problem. The results showed that, among the EC algorithms applied for feature selection, ACO is both computationally more expensive than the others and it did not achieve the highest accuracy. However, DE algorithms followed by PSO algorithms provided the best results and found the most salient subset of EEG features and thus, improved the performance of emotion recognition. The computational cost of DE and PSO algorithms is less than that of other EC algorithms. The results showed that the combination of time and frequency domain features is more efficient, since EC algorithms find the more successful and efficient subset of features by a combination of these features. Moreover, we investigated the set of channels (i.e. electrodes usage) most frequently selected by the combination of EC algorithms via the principle of weighted majority voting. The results showed that the electrodes in the frontal and central lobes were more highly activated, confirming the feasibility of using lightweight and wireless EEG sensors to capture dimensional emotion. In this study, a mobile EEG sensor (Emotiv Insight) was used to recognize four-quadrant dimensional emotions while the subject was watching video clips, and the result with a 65% accuracy shows the validity of using mobile EEG sensors in this domain. However, it also showed that more progress is needed to improve the performance of emotion recognition using wearable sensors. This study paves a way for future sensor development in regard to selecting the correct channels and features to focus on the most important electrodes.

Research Question 2: **How can we optimize the performance of emotion classification system based on physiological signals?**

Choosing an appropriate classifier to effectively classify different emotions using physiological signals is still a challenge in this domain. This is due to the fact that physiological signals are characterized by non-stationarities and nonlinearities. In fact, physiological signals consist of time-series data with variation over a long period of time and dependencies within shorter periods. To capture the inherent temporal structure within the physiological data and to recognize emotion signatures which are

reflected in short period of time, we need to apply a classifier which considers temporal information (Soleymani et al., 2016; Wöllmer et al., 2013). Long Short Term Memory networks (LSTM) are a special type of RNN that have the capability of learning longer temporal sequences (Hochreiter and Schmidhuber, 1997). For this reason LSTM networks offer better emotion classification accuracy over other methods when using time-series data (Kim et al., 2013; Tsai et al., 2017; Wöllmer et al., 2013, 2008).

Although the performance of LSTM networks in classifying different problems is promising, training these networks like other neural networks depends heavily on a set of hyperparameters that determine many aspects of algorithm behaviour. It should be noted that there is no generic optimal configuration for all problem domains. Hence, to achieve a successful performance for each problem domain, such as emotion classification, it is essential to optimize the LSTM hyperparameters.

The second contribution of this thesis, presented in Chapter 4 (Study 2), proposes the use of a novel framework to optimize LSTM hyperparameters using a DE algorithm for emotion recognition based on EEG and BVP signals. This is the first study that utilizes a DE algorithm to optimize LSTM hyperparameters for emotion recognition. The performance of the LSTM optimized by the DE algorithm was evaluated on our dataset collected from wearable sensors (Emotiv and Empatica) and compared with other well-known hyperparameter optimization algorithms like Random search, SA and TPE.

It should be noted that the proposed method used features from time and frequency domains (findings from the preceding chapter). The results of this study first demonstrated that the fusion of EEG and BVP signals performed better compared to non-fused EEG and BVP signals. Moreover, the results showed that the performance of the DE algorithm in finding the most appropriate values for LSTM hyperparameters is better than other hyperparameter optimization methods (PSO, SA, TPE and Random Search algorithms). The results showed that the average accuracy of the DE algorithm followed by that of the PSO algorithm was higher than all other hyperparameter algorithms. Although the performance of the DE algorithm is the highest, it took the most processing time over all iterations. Since the goal of this study is to introduce an optimized LSTM model that can accurately classify different emotions, the computational time required to build such an optimized classification model is less

important, as long as it can achieve an acceptable result. In fact, finding the optimal classifier is done offline and the time (358 hours for 17 cross fold validations at 300 iterations) taken to do this is within reasonable development time for such projects and will be considerably less with higher performance, parallel or cloud computing. The LSTM classifier using the DE algorithm achieved 77% accuracy at 300 iterations. In comparison, the accuracies of all of the other hyperparameter optimization algorithms were less than 70%, which confirms the ability of the DE algorithm to find a better solution.

Findings from this study suggest the performance of the LSTM network in classifying emotions is highly associated with its hyperparameter values. It showed that optimization of the LSTM network by DE algorithm significantly improved the performance of the emotion recognition system compared to the performance when the LSTM hyperparameters were selected randomly. Although this method is computationally expensive, the process of optimization is done offline and only the output, which is the optimized LSTM network, will be used for real-world (online) applications.

Research Question 3: **How can physiological signals be fused to capture temporal emotional changes and improve emotion recognition?**

Many studies in the domain of AC focused on multimodal fusion techniques to build an accurate emotion classification model. One of the common fusion techniques is to concatenate features from each modality and form one feature vector to solve the classification problem. However, these sorts of approaches are not able to capture the non-linear correlation across data modalities, as the correlation between features within each modality is stronger. The non-linear emotional information across modalities can provide complementary information for emotion classification. Therefore, to build an automatic emotion classification system based on multimodal physiological signals, it is essential to capture both emotional information within and across physiological signals over time.

The third contribution, presented in Chapter 5, is about proposing a new framework to fuse physiological signals based on a multimodal learning approach

(Study 3). The presented framework is able to improve the performance of emotion recognition by capturing the non-linear correlation both within and across physiological signals over time. The proposed framework used convolutional neural networks (ConvNets) and LSTM networks in end-to-end fashion. Using end-to-end learning, the constructed features using ConvNets are trained jointly with the classification step as a single network. Moreover, in an end-to-end learning approach, the network is trained from the raw data without any a priori feature extraction. To our knowledge, this is the first work that applies such an end-to-end temporal multimodal fusion model for emotion recognition based on EEG and BVP signals. The temporal multimodal fusion models based on ConvNet LSTM networks can fuse EEG and BVP signals temporally to capture the temporal emotional structures within and across the modalities. We investigate two approaches to fuse information across the temporal domain: early fusion and late fusion.

The performance of the two temporal multimodal fusion models with different window sizes is evaluated using the sliding window strategy and compared with non-temporal multimodal deep learning models using a trial-wise strategy. In the trial-wise training, the duration of the raw physiological signal per video clip, called a trial, is the input and the corresponding trial emotion label is the target used for training. The performance of the proposed models in this chapter is compared with that of the proposed model from the previous chapter (Chapter 4).

The performance of the proposed model was studied based on both early and late fusion models with different window sizes. The results showed that the accuracy of the model based on early fusion in all different window sizes (2-sec, 3-sec, 5-sec and 10-sec long) is higher than such model based on late fusion. As the size of window increased, the emotion classification performance based on temporal multimodal deep learning models (both early and late fusion models) improved, with the best performance achieved at a window size of 10-sec. This not only confirms the efficacy of the sliding window strategy in improving emotion classification using multimodal learning models, but also shows that there is no generic value for window size that can be used to achieve high performance in regard to the emotion classification problem.

The proposed models were also compared to models based on a trial-wise strategy. In trial-wise training, the duration of the raw physiological signals per video

clip, called a trial is the input and the corresponding trial label is the target used for training the ConvNets. Findings from this comparison showed that model based on trial-wise training did not result in acceptable performance for emotion classification. Moreover, the proposed models also outperformed the methods using handcrafted feature extraction.

Although the average performance of the proposed method using deep learning techniques in Chapter 5 is better than the proposed technique based on the handcrafted features in Chapter 4, the handcrafted features are extracted from smaller window size (1-sec window size) while the best achieved accuracy using deep learning techniques is achieved with 10-sec window size. This shows that the handcrafted features are more suitable for real-time applications while the deep learning technique remove expert knowledge and is able to learn features automatically.

All the proposed models were evaluated on a dataset collected from wireless wearable sensors (Emotiv and Empatica wristband).

6.3 LIMITATIONS AND FUTURE WORK

This thesis proposed a different framework to build robust and reliable emotion classification models using advanced machine learning techniques. While the findings of this research are undoubtedly encouraging, a number of limitations should be acknowledged.

Although the performance of the EC algorithms as feature selection methods, in Chapter 3 was promising, these algorithms suffer from premature convergence problems. Therefore, for future work, it is worthwhile exploring the development of new EC algorithms or to modify the existing ones to overcome this problem and improve the classification performance accordingly.

While Study 2 (Chapter 4) provided evidence to suggest such a framework can be used to find appropriate LSTM hyperparameter values to improve emotion classification, it is recommended that the performance of this algorithm be evaluated on other classifiers and in other contexts. Moreover, in this study we only investigated

the effectiveness of the DE algorithm on two hyperparameters (batch size and number of hidden neurons), but it would be worthwhile to explore the efficiency of the proposed framework on other hyperparameters.

In Chapter 5, the proposed temporal multimodal fusion models using early fusion and later fusion achieved acceptable performance, however, providing more complex ConvNets with more layers may provide better emotion classification performance. It is also recommended that this algorithm be applied to different datasets to evaluate its effectiveness for emotion classification.

Although this thesis proposed diverse machine learning algorithms to enhance the performance of emotion classification methods using wearable physiological sensors, the proposed models were evaluated on emotion classification based only on the subject watching a movie or listening to music. Therefore, additional investigation is required to check the suitability of the proposed models for other contexts like driver fatigue monitoring.

Although the proposed methods in Chapter 3 and 4, feature selection using EC algorithms and LSTM hyperparameter optimization, are computationally expensive, these methods are only proposed for the offline data analysis where accuracy is more important than the computational cost. In the age of cloud computing, the process of training can be deployed in cloud to decrease the response time. The outcome the built models can be exported and used for online emotion classification systems. For example, the optimized LSTM classifier, the outcome of Chapter 4, can be utilized for real-time in-vehicle affective intelligent system. The proposed models are suitable for real-time affective intelligent system due to the small selected window sizes (1-10-sec long).

The performance of all the proposed frameworks in this thesis were evaluated on a dataset with only 20 subjects, but evaluating the performance of the proposed methods on a dataset with larger samples is recommended to confirm their effectiveness.

The presented program of research supports the applicability of the proposed methods using portable physiological sensors in the context of affective computing. Future research should seek to evaluate incorporating of these methods into the smart

sensors (wristband and headset), specifically in the context of mental health monitoring. Then the proposed algorithms could be evaluated over a long period of time to investigate the suitability of the proposed sensor-based methods on individuals in a day-to-day environment. In fact the proposed models are based on subject-independent to be used; however, it is recommended to improve the recognition rate by transferring the subject-independent way to subject-dependent way.

The proposed algorithms that detect different emotions can also be incorporated into different contexts such as e-health monitoring, mental health care, intelligent tutoring or playing games. In view of these additional applications, the algorithms should be improved for faster and more efficient output and deployed in the cloud.

Bibliography

Abdel-Hamid, O., Mohamed, A., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(10), 1533–1545.

Ackermann, P., Kohlschein, C., Bitsch, J. Á., Wehrle, K., & Jeschke, S. (2016). EEG-based automatic emotion recognition: Feature extraction, selection and classification methods. *2016 IEEE 18th International Conference on E-Health Networking, Applications and Services (Healthcom)*, 1–6. https://doi.org/10.1109/HealthCom.2016.7749447

Aftanas, L. I., Lotova, N. V., Koshkarov, V. I., & Popov, S. A. (1998). Non-linear dynamical coupling between different brain areas during evoked emotions: an EEG investigation. *Biological Psychology*, *48*(2), 121–138.

Agrafioti, F., Hatzinakos, D., & Anderson, A. K. (2012). ECG pattern analysis for emotion detection. *IEEE Transactions on Affective Computing*, *3*(1), 102–115.

Akselrod, S., Gordon, D., Ubel, F. A., Shannon, D. C., Berger, A. C., & Cohen, R. J. (1981). Power spectrum analysis of heart rate fluctuation: a quantitative probe of beat-to-beat cardiovascular control. *Science*, *213*(4504), 220–222.

Al-Ani, A. (2005). Feature subset selection using ant colony optimization. *International Journal of Computational Intelligence*. Retrieved from https://opus.lib.uts.edu.au/handle/10453/6181

Alba, E., García-Nieto, J., Jourdan, L., & Talbi, E.-G. (2007). Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms. *2007 IEEE*

Congress on Evolutionary Computation, 284–290. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4424483

Ansari-Asl, K., Chanel, G., & Pun, T. (2007). A channel selection method for EEG classification in emotion assessment based on synchronization likelihood. *Signal Processing Conference, 2007 15th European*, 1241–1245. Retrieved from http://ieeexplore.ieee.org/abstract/document/7099003/

Arnold, M. B. (1960). *Emotion and personality*. Retrieved from http://psycnet.apa.org/psycinfo/1960-35012-000

Baig, M. Z., Aslam, N., Shum, H. P. H., & Zhang, L. (n.d.). Differential Evolution Algorithm as a Tool for Optimal Feature Subset Selection in Motor Imagery EEG. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2017.07.033

Baig, M. Z., Aslam, N., Shum, H. P., & Zhang, L. (2017). Differential Evolution Algorithm as a Tool for Optimal Feature Subset Selection in Motor Imagery EEG. *Expert Systems with Applications*. Retrieved from http://www.sciencedirect.com/science/article/pii/S0957417417305109

Bandyopadhyay, S., Saha, S., Maulik, U., & Deb, K. (2008). A simulated annealing-based multiobjective optimization algorithm: AMOSA. *IEEE Transactions on Evolutionary Computation*, *12*(3), 269–283.

Barry, R. J., Clarke, A. R., Johnstone, S. J., Magee, C. A., & Rushby, J. A. (2007). EEG differences between eyes-closed and eyes-open resting conditions. *Clinical Neurophysiology*, *118*(12), 2765–2773.

Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, *5*(4), 537–550.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(Feb), 281–305.

Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., & Cox, D. D. (2015). Hyperopt: a python library for model selection and hyperparameter optimization. *Computational Science & Discovery*, *8*(1), 014008.

Bergstra, J. S., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems*, 2546–2554.

Blum, C., & Sampels, M. (2002). Ant colony optimization for FOP shop scheduling: a case study on different pheromone representations. *Evolutionary Computation, 2002. CEC'02. Proceedings of the 2002 Congress on*, *2*, 1558–1563. Retrieved from http://ieeexplore.ieee.org/abstract/document/1004474/

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.

Brady, K., Gwon, Y., Khorrami, P., Godoy, E., Campbell, W., Dagli, C., & Huang, T. S. (2016a). Multi-modal audio, video and physiological sensor learning for continuous emotion prediction. *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 97–104. ACM.

Brady, K., Gwon, Y., Khorrami, P., Godoy, E., Campbell, W., Dagli, C., & Huang, T. S. (2016b). Multi-modal audio, video and physiological sensor learning for continuous emotion prediction. *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 97–104. ACM.

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.

Candra, H., Yuwono, M., Handojoseno, A., Chai, R., Su, S., & Nguyen, H. T. (2015). Recognizing emotions from EEG subbands using wavelet analysis. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 6030–6033. https://doi.org/10.1109/EMBC.2015.7319766

Caridakis, G., Castellano, G., Kessous, L., Raouzaiou, A., Malatesta, L., Asteriadis, S., & Karpouzis, K. (2007). Multimodal emotion recognition from expressive faces, body gestures and speech. *IFIP International Conference on Artificial Intelligence Applications and Innovations*, 375–388. Springer.

Cecotti, H., & Graser, A. (2011). Convolutional neural networks for P300 detection with application to brain-computer interfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(3), 433–445.

Chai, T. Y., Woo, S. S., Rizon, M., & Tan, C. S. (2010). Classification of human emotions from EEG signals using statistical features and neural network. *International*, *1*, 1–6. Retrieved from http://eprints.uthm.edu.my/511/

Chanel, G., Kronegg, J., Grandjean, D., & Pun, T. (2006). Emotion assessment: Arousal evaluation using EEG's and peripheral physiological signals. In *Multimedia content representation, classification and security* (pp. 530–537). Retrieved from http://link.springer.com/chapter/10.1007/11848035_70

Chang, C.-M., & Lee, C.-C. (2017). Fusion of multiple emotion perspectives: Improving affect recognition through integrating cross-lingual emotion

information. *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, 5820–5824. IEEE.

Chang, C.-M., Su, B.-H., Lin, S.-C., Li, J.-L., & Lee, C.-C. (2017). A bootstrapped multi-view weighted Kernel fusion framework for cross-corpus integration of multimodal emotion recognition. *Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on*, 377–382. IEEE.

Chao, L., Tao, J., Yang, M., Li, Y., & Wen, Z. (2015). Long short term memory recurrent neural network based multimodal dimensional emotion recognition. *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, 65–72. ACM.

Chen, S., & Jin, Q. (2015). Multi-modal dimensional emotion recognition using recurrent neural networks. *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, 49–56. ACM.

Chen, S.-M., & Chien, C.-Y. (2011). Solving the traveling salesman problem based on the genetic simulated annealing ant colony system with particle swarm optimization techniques. *Expert Systems with Applications*, *38*(12), 14439–14450.

Cheng, B., & Liu, G.-Y. (2008). Emotion recognition from surface EMG signal using wavelet transform and neural network. *Proceedings of The 2nd International Conference on Bioinformatics and Biomedical Engineering (ICBBE)*, 1363–1366. Retrieved from http://www.paper.edu.cn/en_releasepaper/downPaper/200706-289

Cho, Y., Bianchi-Berthouze, N., & Julier, S. J. (2017). DeepBreath: Deep learning of breathing patterns for automatic stress recognition using low-cost thermal imaging in unconstrained settings. *2017 Seventh International Conference on*

Affective Computing and Intelligent Interaction (ACII), 456–463. https://doi.org/10.1109/ACII.2017.8273639

Colorni, A., Dorigo, M., Maniezzo, V., & Trubian, M. (1994). Ant system for job-shop scheduling. *Belgian Journal of Operations Research, Statistics and Computer Science*, *34*(1), 39–53.

Costa, D., & Hertz, A. (1997). Ants can colour graphs. *Journal of the Operational Research Society*, *48*(3), 295–305.

Cowie, R., & Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication*, *40*(1), 5–32.

Cowie, R., Douglas-Cowie, E., Savvidou*, S., McMahon, E., Sawey, M., & Schröder, M. (2000). "FEELTRACE": An instrument for recording perceived emotion in real time. *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*. Retrieved from http://www.isca-speech.org/archive_open/speech_emotion/spem_019.html

Darwin, C. (1965). *The expression of the emotions in man and animals* (Vol. 526). Retrieved from https://books.google.com.au/books?hl=en&lr=&id=IYJ9RH5PShwC&oi=fnd&pg=PR9&dq=The+expression+of+the+emotions+in+man+and+animals+&ots=TUUh_zdluF&sig=tWeN9d9-NShEg-1djHAHFRZNHNs

Davidson, R. J. (2003). Affective neuroscience and psychophysiology: toward a synthesis. *Psychophysiology*, *40*(5), 655–665.

Dorigo, M., Birattari, M., & Stutzle, T. (2006). Ant colony optimization. *IEEE Computational Intelligence Magazine*, *1*(4), 28–39.

Dorigo, M., & Gambardella, L. M. (1997). Ant colony system: a cooperative learning approach to the traveling salesman problem. *IEEE Transactions on Evolutionary Computation*, *1*(1), 53–66.

Douglas-Cowie, E., Cox, C., Martin, J.-C., Devillers, L., Cowie, R., Sneddon, I., … others. (2011). The HUMAINE database. In *Emotion-Oriented Systems* (pp. 243–284). Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-15184-2_14

Du Boulay, B. (2011). Towards a motivationally intelligent pedagogy: how should an intelligent tutor respond to the unmotivated or the demotivated? In *New perspectives on affect and learning technologies* (pp. 41–52). Retrieved from http://link.springer.com/chapter/10.1007/978-1-4419-9625-1_4

Duvinage, M., Castermans, T., Petieau, M., Hoellinger, T., Cheron, G., & Dutoit, T. (2013). Performance of the Emotiv Epoc headset for P300-based applications. *Biomedical Engineering Online*, *12*(1), 56.

Eberhart, R., & Kennedy, J. (1995). A new optimizer using particle swarm theory. *Micro Machine and Human Science, 1995. MHS'95., Proceedings of the Sixth International Symposium on*, 39–43. Retrieved from http://ieeexplore.ieee.org/abstract/document/494215/

Ebrahimi Kahou, S., Michalski, V., Konda, K., Memisevic, R., & Pal, C. (2015). Recurrent neural networks for emotion recognition in video. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 467–474. ACM.

Ekman, P., & Friesen, W. V. (n.d.). Pictures of facial affect. 1976. *Palo Alto, CA: Consulting Psychologists*.

Ekman, Paul. (1992). *Are there basic emotions?* Retrieved from http://psycnet.apa.org/journals/rev/99/3/550/

Ekman, Paul, & Oster, H. (1979). Facial expressions of emotion. *Annual Review of Psychology*, *30*(1), 527–554.

Feradov, F., & Ganchev, T. (2014). Detection of Negative Emotional States from Electroencephalographic (EEG) signals. *Annual Journal of Electronics*, *8*, 66–69.

Frijda, N. H. (1986). *The emotions*. Retrieved from https://books.google.com.au/books?hl=en&lr=&id=QkNuuVf-pBMC&oi=fnd&pg=PR11&dq=The+emotions&ots=BKHah_1pYs&sig=ey7-q738x_sMIaGZJuEbn0yGXro

Ghosh, A., Danieli, M., & Riccardi, G. (2015). Annotation and prediction of stress and workload from physiological and inertial signals. *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, 1621–1624. IEEE.

Goldberg, D. E., & Holland, J. H. (1988). Genetic algorithms and machine learning. *Machine Learning*, *3*(2), 95–99.

Gonçalves, J. F., de Magalhães Mendes, J. J., & Resende, M. G. (2005). A hybrid genetic algorithm for the job shop scheduling problem. *European Journal of Operational Research*, *167*(1), 77–95.

Graves, A. (n.d.). Supervised sequence labelling with recurrent neural networks. 2012. *ISBN 9783642212703. URL Http://Books. Google. Com/Books*.

Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *Acoustics, Speech and Signal Processing (Icassp), 2013 Ieee International Conference on*, 6645–6649. IEEE.

Gray, J. A. (1985). A whole and its parts: Behaviour, the brain, cognition and emotion. *Bulletin of the British Psychological Society*. Retrieved from http://psycnet.apa.org/psycinfo/1986-00192-001

Gunes, H., & Pantic, M. (2010). Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions (IJSE)*, *1*(1), 68–99.

Gunes, H., & Piccardi, M. (2005). Affect recognition from face and body: early fusion vs. late fusion. *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, *4*, 3437–3443. IEEE.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*(Mar), 1157–1182.

Haag, A., Goronzy, S., Schaich, P., & Williams, J. (2004). Emotion recognition using bio-sensors: First steps towards an automatic system. In *Affective dialogue systems* (pp. 36–48). Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-24842-2_4

Haq, S., & Jackson, P. J. (2011). Multimodal emotion recognition. In *Machine audition: principles, algorithms and systems* (pp. 398–423). IGI Global.

Harrison, T. (2013). *The Emotiv mind: Investigating the accuracy of the Emotiv EPOC in identifying emotions and its use in an Intelligent Tutoring System*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Herbelin, B., Benzaki, P., Riquier, F., Renault, O., Grillon, H., & Thalmann, D. (2005). Using physiological measures for emotional assessment: a computer-aided tool for cognitive and behavioural therapy. *International Journal on Disability and Human Development*, *4*(4), 269–278.

Hettich, D. T., Bolinger, E., Matuz, T., Birbaumer, N., Rosenstiel, W., & Spüler, M. (2016). EEG Responses to Auditory Stimuli for Automatic Affect Recognition. *Frontiers in Neuroscience*, *10*, 244.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., … Sainath, T. N. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*(6), 82–97.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Horlings, R., Datcu, D., & Rothkrantz, L. J. (2008). Emotion recognition using brain activity. *Proceedings of the 9th International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing*, 6. Retrieved from http://dl.acm.org/citation.cfm?id=1500888

Hossain, M. S., & Muhammad, G. (2017). An emotion recognition system for mobile applications. *IEEE Access*, *5*, 2281–2287.

Hsu, W.-Y. (2013). Independent Component analysis and multiresolution asymmetry ratio for brain–computer interface. *Clinical EEG and Neuroscience*, 1550059412463660.

Hu, D., & Li, X. (2016). Temporal multimodal learning in audiovisual speech recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3574–3582.

Hui, T. K., & Sherratt, R. S. (2018). Coverage of Emotion Recognition for Common Wearable Biosensors. *Biosensors*, *8*(2), 30.

Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2011). Sequential model-based optimization for general algorithm configuration. *International Conference on Learning and Intelligent Optimization*, 507–523. Springer.

Izard, C. E., & Izard, C. E. (1977). *Human emotions* (Vol. 17). Plenum Press New York.

James, W. (1884). II.—What is an emotion? *Mind*, (34), 188–205.

Jang, E., Park, B., Kim, S., & Sohn, J. (2012). Emotion classification based on physiological signals induced by negative emotions: Discriminantion of negative emotions by machine learning algorithm. *Proceedings of 2012 9th IEEE International Conference on Networking, Sensing and Control*, 283–288. https://doi.org/10.1109/ICNSC.2012.6204931

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

John, G. H., Kohavi, R., Pfleger, K., & others. (1994). Irrelevant features and the subset selection problem. *Machine Learning: Proceedings of the Eleventh International Conference*, 121–129. Retrieved from https://books.google.com.au/books?hl=en&lr=&id=cEqjBQAAQBAJ&oi=fnd&pg=PA121&dq=Irrelevant+features+and+the+subset+selection+problem&ots=E1sxqhD2EH&sig=TboZIRhN0MGI4ZUwXdJBomv_L_w

Johnson-Laird, P. N., & Oatley, K. (1989). The language of emotions: An analysis of a semantic field. *Cognition and Emotion*, *3*(2), 81–123.

Jung, T.-P., Makeig, S., Humphries, C., Lee, T.-W., Mckeown, M. J., Iragui, V., & Sejnowski, T. J. (2000). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, *37*(2), 163–178.

Kahou, S. E., Bouthillier, X., Lamblin, P., Gulcehre, C., Michalski, V., Konda, K., … others. (2016). Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, *10*(2), 99–111.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1725–1732.

Karthigayan, M., Rizon, M., Nagarajan, R., & Yaacob, S. (2008). Genetic algorithm and neural network for face emotion recognition. In *Affective Computing*. IntechOpen.

Kennedy, J. (2011). Particle swarm optimization. In *Encyclopedia of machine learning* (pp. 760–766). Retrieved from http://link.springer.com/10.1007/978-0-387-30164-8_630

Kennedy, J., & Eberhart, R. C. (1997). A discrete binary version of the particle swarm algorithm. *Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation., 1997 IEEE International Conference on*, *5*, 4104–4108. Retrieved from http://ieeexplore.ieee.org/abstract/document/637339/

Khan, A. M., & Lawo, M. (2016). Recognizing Emotion from Blood Volume Pulse and Skin Conductance Sensor Using Machine Learning Algorithms. *XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016*, 1297–1303. Springer.

Khan, M., Ahamed, S. I., Rahman, M., & Smith, R. O. (2011). A feature extraction method for realtime human activity recognition on cell phones. *Proceedings of 3rd International Symposium on Quality of Life Technology (isQoLT 2011). Toronto, Canada*. Retrieved from

https://pdfs.semanticscholar.org/8aaa/9aa902b524992324fcb359ee4f01beeabdfe.pdf

Khosrowabadi, R., & bin Abdul Rahman, A. W. (2010). Classification of EEG correlates on emotion using features from Gaussian mixtures of EEG spectrogram. *Information and Communication Technology for the Muslim World (ICT4M), 2010 International Conference on*, E102–E107. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5971942

Kim, J. (2007). *Bimodal emotion recognition using speech and physiological changes*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.476.3361&rep=rep1&type=pdf

Kim, J., André, E., Rehm, M., Vogt, T., & Wagner, J. (2005). Integrating information from speech and physiological signals to achieve emotional sensitivity. *Ninth European Conference on Speech Communication and Technology*.

Kim, J., André, E., & Vogt, T. (2009). Towards user-independent classification of multimodal emotional signals. *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, 1–7. IEEE.

Kim, K. H., Bang, S. W., & Kim, S. R. (2004). Emotion recognition system using short-term monitoring of physiological signals. *Medical and Biological Engineering and Computing*, *42*(3), 419–427.

Kim, Y., Lee, H., & Provost, E. M. (2013a). Deep learning for robust feature generation in audiovisual emotion recognition. *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 3687–3691. IEEE.

Kim, Y., Lee, H., & Provost, E. M. (2013b). Deep learning for robust feature generation in audiovisual emotion recognition. *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 3687–3691. IEEE.

Kirkpatrick, S., Gelatt, C. D., Vecchi, M. P., & others. (1983). Optimization by simulated annealing. *Science*, *220*(4598), 671–680.

Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., … Patras, I. (2012). Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, *3*(1), 18–31.

Koelstra, S., Yazdani, A., Soleymani, M., Mühl, C., Lee, J.-S., Nijholt, A., … Patras, I. (2010). Single trial classification of EEG and peripheral physiological signals for recognition of emotions induced by music videos. *International Conference on Brain Informatics*, 89–100. Springer.

Kortelainen, J., & Seppänen, T. (2013). EEG-based recognition of video-induced emotions: Selecting subject-independent feature set. *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 4287–4290. https://doi.org/10.1109/EMBC.2013.6610493

Kranczioch, C., Zich, C., Schierholz, I., & Sterr, A. (2014). Mobile EEG and its potential to promote the theory and application of imagery-based motor rehabilitation. *International Journal of Psychophysiology*, *91*(1), 10–15.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 1097–1105.

Kroupi, E., Yazdani, A., & Ebrahimi, T. (2011). EEG correlates of different emotional states elicited during watching music videos. In *Affective Computing and*

Intelligent Interaction (pp. 457–466). Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-24571-8_58

Kudo, M., & Sklansky, J. (2000). Comparison of algorithms that select features for pattern classifiers. *Pattern Recognition*, *33*(1), 25–41.

Kuremoto, T., Kimura, S., Kobayashi, K., & Obayashi, M. (2012). Time Series Forecasting Using Restricted Boltzmann Machine. *ICIC (3)*, 17–22. Retrieved from http://link.springer.com/content/pdf/10.1007/978-3-642-31837-5.pdf#page=41

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). International affective picture system (IAPS): Affective ratings of pictures and instruction manual. *Technical Report A-8*. Retrieved from http://www.citeulike.org/group/13427/article/7208496

Langley, P., & others. (1994). Selection of relevant features in machine learning. *Proceedings of the AAAI Fall Symposium on Relevance*, *184*, 245–271. Retrieved from http://www.aaai.org/Papers/Symposia/Fall/1994/FS-94-02/FS94-02-034.pdf

Larsen, R. J., & Fredrickson, B. L. (1999). *Measurement issues in emotion research.* Retrieved from http://psycnet.apa.org/psycinfo/1999-02842-003

Le, D., & Provost, E. M. (2013). Emotion recognition from spontaneous speech using Hidden Markov models with deep belief networks. *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, 216–221. https://doi.org/10.1109/ASRU.2013.6707732

Leape, C., Fong, A., & Ratwani, R. M. (2016). Heuristic Usability Evaluation of Wearable Mental State Monitoring Sensors for Healthcare Environments. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*,

60, 583–587. Retrieved from http://journals.sagepub.com/doi/abs/10.1177/1541931213601134

Li, G., Lee, B.-L., & Chung, W.-Y. (2015). Smartwatch-based wearable EEG system for driver drowsiness detection. *IEEE Sensors Journal*, *15*(12), 7169–7180.

Li, K., Li, X., Zhang, Y., & Zhang, A. (2013). Affective state recognition from EEG with deep belief networks. *Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on*, 305–310. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6732507

Li, Y., Huang, J., Zhou, H., & Zhong, N. (2017). Human Emotion Recognition with Electroencephalographic Multidimensional Features by Hybrid Deep Neural Networks. *Applied Sciences*, *7*(10), 1060.

Lichtenstein, A., Oehme, A., Kupschick, S., & Jürgensohn, T. (2008). Comparing two emotion models for deriving affective states from physiological data. In *Affect and Emotion in Human-Computer Interaction* (pp. 35–50). Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-85099-1_4

Lin, Y., Wang, C., Jung, T., Wu, T., Jeng, S., Duann, J., & Chen, J. (2010). EEG-Based Emotion Recognition in Music Listening. *IEEE Transactions on Biomedical Engineering*, *57*(7), 1798–1806. https://doi.org/10.1109/TBME.2010.2048568

Lin, Y.-P., Wang, C.-H., Jung, T.-P., Wu, T.-L., Jeng, S.-K., Duann, J.-R., & Chen, J.-H. (2010). EEG-based emotion recognition in music listening. *Biomedical Engineering, IEEE Transactions on*, *57*(7), 1798–1806.

Liu, C., Conn, K., Sarkar, N., & Stone, W. (2008). Online affect detection and robot behavior adaptation for intervention of children with autism. *IEEE Transactions on Robotics*, *24*(4), 883–896.

Liu, K., Zhang, L. M., & Sun, Y. W. (2014). Deep Boltzmann machines aided design based on genetic algorithms. *Applied Mechanics and Materials*, *568*, 848–851. Retrieved from https://www.scientific.net/AMM.568-570.848

Liu, Yisi, & Sourina, O. (2013). Real-time fractal-based valence level recognition from EEG. In *Transactions on Computational Science XVIII* (pp. 101–120). Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-38803-3_6

Liu, Yu, Chen, X., Peng, H., & Wang, Z. (2017). Multi-focus image fusion with a deep convolutional neural network. *Information Fusion*, *36*, 191–207.

Luneski, A., Bamidis, P. D., & Hitoglou-Antoniadou, M. (2008). Affective computing and medical informatics: state of the art in emotion-aware medical applications. *Studies in Health Technology and Informatics*, *136*, 517.

Lushin, I., & others. (2016). *Detection of Epilepsy with a Commercial EEG Headband*. Retrieved from http://www.doria.fi/handle/10024/123588

Mandryk, R. L., & Atkins, M. S. (2007). A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. *International Journal of Human-Computer Studies*, *65*(4), 329–347.

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

McDougall, W. (2003). *An introduction to social psychology*. Retrieved from https://books.google.com.au/books?hl=en&lr=&id=CSjxe-t6qm4C&oi=fnd&pg=PA1&dq=An+introduction+to+social+psychology&ots=WXl4KwMUaE&sig=dm0Z2d_mwsZHv6jyJKQubBw1_sA

McMahan, T., Parberry, I., & Parsons, T. D. (2015). Evaluating Electroencephalography Engagement Indices During Video Game Play. *FDG*.

Menezes, M. L. R., Samara, A., Galway, L., Sant'Anna, A., Verikas, A., Alonso-Fernandez, F., … Bond, R. (2017). Towards emotion recognition for virtual environments: an evaluation of eeg features on benchmark dataset. *Personal and Ubiquitous Computing*, 1–11.

Michel, R., & Middendorf, M. (1998). An island model based ant system with lookahead for the shortest supersequence problem. *International Conference on Parallel Problem Solving from Nature*, 692–701. Retrieved from http://link.springer.com/chapter/10.1007/BFb0056911

Mistry, K., Zhang, L., Neoh, S. C., Lim, C. P., & Fielding, B. (2016). *A Micro-GA Embedded PSO Feature Selection Approach to Intelligent Facial Emotion Recognition*. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7456259

Mowrer, O. (1960). *Learning theory and behavior*. Retrieved from http://doi.apa.org/index.cfm?fa=search.exportFormat&uid=2005-06665-000&recType=psycinfo&singlerecord=1&searchresultpage=true

Murugappan, M., Rizon, M., Nagarajan, R., Yaacob, S., Zunaidi, I., & Hazry, D. (2007). EEG feature extraction for classifying emotions using FCM and FKM. *International Journal of Computers and Communications*, *1*(2), 21–25.

Murugappan, Murugappan, Ramachandran, N., Sazali, Y., & others. (2010). Classification of human emotion from EEG using discrete wavelet transform. *Journal of Biomedical Science and Engineering*, *3*(04), 390.

Murugappan, Muthusamy. (2011). Human emotion classification using wavelet transform and KNN. *Pattern Analysis and Intelligent Robotics (ICPAIR), 2011*

International Conference on, *1*, 148–153. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5976886

Nakisa, B., Nazri, M. Z. A., Rastgoo, M. N., & Abdullah, S. (2014). A Survey: Particle Swarm Optimization Based Algorithms To Solve Premature Convergence Problem. *Journal of Computer Science*, *10*(9), 1758–1765.

Nakisa, B., Rastgoo, M. N., Nasrudin, M. F., & Nazri, M. Z. A. (2014). A Multi-Swarm Particle Swarm Optimization with Local Search on Multi-Robot Search System. *Journal of Theoretical and Applied Information Technology*, *71*(1). Retrieved from http://www.jatit.org/volumes/Vol71No1/15Vol71No1.pdf

Nakisa, B., Rastgoo, M. N., & Nazri, M. Z. A. (2018). Target searching in unknown environment of multi-robot system using a hybrid particle swarm optimization. *Journal of Theoretical and Applied Information Technology*, vol. 96, no.13, pp. 4055-4065.

Nakisa, B., Rastgoo, M. N., & Norodin, M. J. (2014). Balancing exploration and exploitation in particle swarm optimization on search tasking. *Research Journal of Applied Sciences, Engineering and Technology*, vol. 8, no. 12, pp. 1429–1434.

Nakisa, B., Rastgoo, M. N., Rakotonirainy, A., Maire, F., & Chandran, V. (2018). Long Short Term Memory Hyperparameter Optimization for a Neural Network Based Emotion Recognition Framework. *IEEE Access*, vol. 6, pp. 49325–49338, doi: 10.1109/ACCESS.2018.2868361.

Nakisa, B., Rastgoo, M. N., Tjondronegoro, D., & Chandran, V. (2017). Evolutionary Computation Algorithms for Feature Selection of EEG-based Emotion Recognition using Mobile Sensors. *Expert Systems with Applications*, vol. 93, pp. 143–155, Mar. 2017.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 689–696.

Nguyen, D., Nguyen, K., Sridharan, S., Dean, D., & Fookes, C. (2018). Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition. *Computer Vision and Image Understanding*.

Nicolaou, M. A., Gunes, H., & Pantic, M. (2011). Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *Affective Computing, IEEE Transactions on*, *2*(2), 92–105.

Nie, D., Wang, X.-W., Shi, L.-C., & Lu, B.-L. (2011). EEG-based emotion recognition during watching movies. *Neural Engineering (NER), 2011 5th International IEEE/EMBS Conference on*, 667–670. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5910636

Niu, X., Chen, L., & Chen, Q. (2011). Research on genetic algorithm based on emotion recognition using physiological signals. *2011 International Conference on Computational Problem-Solving (ICCP)*, 614–618. IEEE.

Oh, S.-H., Lee, Y.-R., & Kim, H.-N. (2014). A novel EEG feature extraction method using Hjorth parameter. *International Journal of Electronics and Electrical Engineering*, *2*(2), 106–110.

Onton, J. A., & Makeig, S. (2009). High-frequency broadband modulation of electroencephalographic spectra. *Frontiers in Human Neuroscience*, *3*, 61.

Pan, Q.-K., Wang, L., & Qian, B. (2009). A novel differential evolution algorithm for bi-criteria no-wait flow shop scheduling problems. *Computers & Operations Research*, *36*(8), 2498–2511.

Panksepp, J. (1982). Toward a general psychobiological theory of emotions. *Behavioral and Brain Sciences*, *5*(03), 407–422.

Papa, J. P., Rosa, G. H., Costa, K. A., Marana, N. A., Scheirer, W., & Cox, D. D. (2015). On the model selection of bernoulli restricted boltzmann machines through harmony search. *Proceedings of the Companion Publication of the 2015 Annual Conference on Genetic and Evolutionary Computation*, 1449–1450. ACM.

Peter, C., Ebert, E., & Beikirch, H. (2009). Physiological sensing for affective computing. In *Affective Information Processing* (pp. 293–310). Retrieved from http://link.springer.com/chapter/10.1007/978-1-84800-306-4_16

Petrantonakis, P. C., & Hadjileontiadis, L. J. (2010). Emotion recognition from EEG using higher order crossings. *Information Technology in Biomedicine, IEEE Transactions on*, *14*(2), 186–197.

Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *23*(10), 1175–1191.

Price, K., Storn, R. M., & Lampinen, J. A. (2006). *Differential evolution: a practical approach to global optimization*. Retrieved from https://books.google.com.au/books?hl=en&lr=&id=hakXI-dEhTkC&oi=fnd&pg=PR7&dq=Differential+Evolution:+A+Practical+Approach+to+Global+Optimization&ots=c_5DKPKf60&sig=bn7XPVZZGJqyG8Ko-8sdN9h0JD8

Pudil, P., Novovičová, J., & Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters*, *15*(11), 1119–1125.

Qin, A. K., Huang, V. L., & Suganthan, P. N. (2009). Differential evolution algorithm with strategy adaptation for global numerical optimization. *IEEE Transactions on Evolutionary Computation*, *13*(2), 398–417.

Qin, H., Shinozaki, T., & Duh, K. (n.d.). *Evolution Strategy Based Automatic Tuning of Neural Machine Translation Systems*.

Ragot, M., Martin, N., Em, S., Pallamin, N., & Diverrez, J.-M. (2017). Emotion Recognition Using Physiological Signals: Laboratory vs. Wearable Sensors. *International Conference on Applied Human Factors and Ergonomics*, 15–22. Springer.

Rakshit, R., Reddy, V. R., & Deshpande, P. (2016). Emotion detection and recognition using hrv features derived from photoplethysmogram signals. *Proceedings of the 2nd Workshop on Emotion Representations and Modelling for Companion Systems*, 2. ACM.

Ramirez, R., & Vamvakousis, Z. (2012). Detecting emotion from EEG signals using the emotive epoc device. In *Brain Informatics* (pp. 175–184). Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-35139-6_17

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.

Rastgoo, M. N., Nakisa, B., & Ahmad Nazri, M. Z. (2015). A Hybrid of Modified PSO and Local Search on a Multi-Robot Search System. *International Journal of Advanced Robotic Systems*, *12*(7), 86. https://doi.org/10.5772/60624

Rastgoo, M. N., Nakisa, B., & Ahmadi, M. (2015). A Modified Particle Swarm Optimization on Search Tasking. *Research Journal of Applied Sciences, Engineering and Technology*, vol. 9, no. 8, pp. 594-600.

Rastgoo, M. N., Nakisa, B., & Najafabadi, F. S. (2014). Inspiring Particle Swarm Optimization on Multi-Robot Search System. *International Journal on Computer Science and Engineering*, vol. 6, no. 10, pp. 338-342.

Rastgoo, M. N., Nakisa, B., & Nazri, M. Z. A. (2015). A Hybrid of Modified PSO and Local Search on a Multi-robot Search System. *International Journal of Advanced Robotic Systems*, *11*. Retrieved from http://journals.sagepub.com/doi/abs/10.5772/60624

Rastgoo, M. N., Nakisa, B., Rakotonirainy, A., Chandran, V., & Tjondronegoro, D. (2018). A critical review of proactive detection of driver stress levels based on multimodal measurements. *ACM Computing Surveys (CSUR)*, vol. 51, no. 5, pp. 1–35.

Redmond, S. J., & Heneghan, C. (2006). Cardiorespiratory-based sleep staging in subjects with obstructive sleep apnea. *IEEE Transactions on Biomedical Engineering*, *53*(3), 485–496.

Reunanen, J. (2003). Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research*, *3*(Mar), 1371–1382.

rey Horn, J., Nafpliotis, N., & Goldberg, D. E. (1994). A niched Pareto genetic algorithm for multiobjective optimization. *Proceedings of the First IEEE Conference on Evolutionary Computation, IEEE World Congress on Computational Intelligence*, *1*, 82–87. Citeseer.

Rigas, G., Katsis, C. D., Ganiatsas, G., & Fotiadis, D. I. (2007). A user independent, biosignal based, emotion recognition method. In *User Modeling 2007* (pp. 314–318). Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-73078-1_36

Ringeval, F., Eyben, F., Kroupi, E., Yuce, A., Thiran, J.-P., Ebrahimi, T., … Schuller, B. (2015). Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters*, *66*, 22–30.

Robinson, N., & Vinod, A. P. (2015). Bi-Directional Imagined Hand Movement Classification Using Low Cost EEG-Based BCI. *2015 IEEE International Conference on Systems, Man, and Cybernetics*, 3134–3139. https://doi.org/10.1109/SMC.2015.544

Rodriguez-Bermudez, G., Garcia-Laencina, P. J., & Roca-Dorda, J. (2013). Efficient automatic selection and combination of eeg features in least squares classifiers for motor imagery brain–computer interfaces. *International Journal of Neural Systems*, *23*(04), 1350015.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161.

Schaaff, K., & Schultz, T. (2009). Towards emotion recognition from electroencephalographic signals. *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 1–6. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5349316

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., … Ball, T. (2017). Deep learning with convolutional neural networks for brain mapping and decoding of movement-related information from the human eeg. *arXiv Preprint arXiv:1703.05051*.

Sha, D. Y., & Hsu, C.-Y. (2006). A hybrid particle swarm optimization for job shop scheduling problem. *Computers & Industrial Engineering*, *51*(4), 791–808.

Shah, F., & others. (2010). Discrete wavelet transforms and artificial neural networks for speech emotion recognition. *International Journal of Computer Theory and Engineering*, *2*(3), 319.

Sim, K. M., & Sun, W. H. (2003). Ant colony optimization for routing and load-balancing: survey and new directions. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *33*(5), 560–572.

Sivagaminathan, R. K., & Ramakrishnan, S. (2007). A hybrid approach for feature subset selection using neural networks and ant colony optimization. *Expert Systems with Applications*, *33*(1), 49–60.

Sohn, K., Shang, W., & Lee, H. (2014). Improved multimodal deep learning with variation of information. *Advances in Neural Information Processing Systems*, 2141–2149.

Soleymani, M., Asghari-Esfeden, S., Fu, Y., & Pantic, M. (2016). Analysis of EEG signals and facial expressions for continuous emotion detection. *IEEE Transactions on Affective Computing*, *7*(1), 17–28.

Soleymani, M., Lichtenauer, J., Pun, T., & Pantic, M. (2012). A multimodal database for affect recognition and implicit tagging. *Affective Computing, IEEE Transactions on*, *3*(1), 42–55.

Soleymani, M., Pantic, M., & Pun, T. (2012). Multimodal emotion recognition in response to videos. *IEEE Transactions on Affective Computing*, *3*(2), 211–223.

Somol, P., Pudil, P., Novovičová, J., & Paclık, P. (1999). Adaptive floating search methods in feature selection. *Pattern Recognition Letters*, *20*(11), 1157–1163.

Sourina, O., Kulish, V. V., & Sourin, A. (2009). Novel tools for quantification of brain responses to music stimuli. *13th International Conference on Biomedical*

Engineering, 411–414. Retrieved from http://www.springerlink.com/index/m8681026rr545831.pdf

Sourina, Olga, & Liu, Y. (2011). A Fractal-based Algorithm of Emotion Recognition from EEG using Arousal-Valence Model. *BIOSIGNALS*, 209–214. Retrieved from http://ntu.edu.sg/home/eosourina/Papers/OSBIOSIGNALS_66_CR.pdf

Sourina, Olga, Sourin, A., & Kulish, V. (2009). EEG data driven animation and its application. *Computer Vision/Computer Graphics CollaborationTechniques*, 380–388.

Specht, D. F. (1990). Probabilistic neural networks. *Neural Networks*, *3*(1), 109–118.

Srivastava, N., & Salakhutdinov, R. R. (2012). Multimodal learning with deep boltzmann machines. *Advances in Neural Information Processing Systems*, 2222–2230.

Storn, R. (1996). On the usage of differential evolution for function optimization. *Fuzzy Information Processing Society, 1996. NAFIPS., 1996 Biennial Conference of the North American*, 519–523. IEEE.

Stytsenko, K., Jablonskis, E., & Prahm, C. (2011). Evaluation of consumer EEG device Emotiv EPOC. *MEi: CogSci Conference 2011, Ljubljana*. Retrieved from http://www.univie.ac.at/meicogsci/php/ocs/index.php/meicog/meicog2011/paper/view/210

Sun, Y., Babbs, C. F., & Delp, E. J. (2006). A comparison of feature selection methods for the detection of breast cancers in mammograms: adaptive sequential floating search vs. genetic algorithm. *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, 6532–6535. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1615996

Suresh, R. K., & Mohanasundaram, K. M. (2006). Pareto archived simulated annealing for job shop scheduling with multiple objectives. *The International Journal of Advanced Manufacturing Technology*, *29*(1), 184–196.

Takahashi, K. (2004). Remarks on emotion recognition from multi-modal bio-potential signals. *2004 IEEE International Conference on Industrial Technology, 2004. IEEE ICIT '04*, *3*, 1138–1143 Vol. 3. https://doi.org/10.1109/ICIT.2004.1490720

Takahashi, Kazuhiko. (2004). Remarks on SVM-based emotion recognition from multi-modal bio-potential signals. *Robot and Human Interactive Communication, 2004. ROMAN 2004. 13th IEEE International Workshop on*, 95–100. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1374736

Tomkins, S. (1963). *Affect/imagery/consciousness. Vol. 2: The negative affects.* Retrieved from http://www.citeulike.org/group/1484/article/801989

Tomkins, S. S. (1962). *Affect, imagery, consciousness: Vol. I. The positive affects.* Retrieved from http://psycnet.apa.org/psycinfo/1964-02650-000

Tsai, F. S., Weng, Y. M., Ng, C. J., & Lee, C. C. (2017). Embedding stacked bottleneck vocal features in a LSTM architecture for automatic pain level classification during emergency triage. *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, 313–318. https://doi.org/10.1109/ACII.2017.8273618

Tuncer, A., & Yildirim, M. (2012). Dynamic path planning of mobile robots with improved genetic algorithm. *Computers & Electrical Engineering*, *38*(6), 1564–1572.

Udovičić, G., Đerek, J., Russo, M., & Sikora, M. (2017). Wearable Emotion Recognition System based on GSR and PPG Signals. *Proceedings of the 2nd International Workshop on Multimedia for Personal Health and Health Care*, 53–59. ACM.

Valstar, M., & Pantic, M. (2010). Induced disgust, happiness and surprise: an addition to the mmi facial expression database. *Proc. 3rd Intern. Workshop on EMOTION (Satellite of LREC): Corpora for Research on Emotion and Affect*, 65. Retrieved from http://lrec.elra.info/proceedings/lrec2010/workshops/W24.pdf#page=73

Wang, Q., Sourina, O., & Nguyen, M. K. (2010). Eeg-based“ serious” games design for medical applications. *Cyberworlds (CW), 2010 International Conference on*, 270–276. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5656236

Watson, J. B., & others. (1925). *Behaviorism*. Retrieved from https://books.google.com.au/books?hl=en&lr=&id=PhnCSSy0UWQC&oi=fnd&pg=PR10&dq=Behaviorism&ots=tW28rLqCbp&sig=0P9PRDElIwNEJoxPqegDy3f5THg

Weiner, B., & Graham, S. (1984). An attributional approach to emotional development. *Emotions, Cognition, and Behavior*, 167–191.

Wöllmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., & Cowie, R. (2008). Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies. *Ninth Annual Conference of the International Speech Communication Association*.

Wöllmer, M., Kaiser, M., Eyben, F., Schuller, B., & Rigoll, G. (2013). LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, *31*(2), 153–163.

Wu, L., Oviatt, S. L., & Cohen, P. R. (1999). Multimodal integration-a statistical view. *IEEE Transactions on Multimedia*, *1*(4), 334–341.

Wu, Y., Wei, Y., & Tudor, J. (2017). A real-time wearable emotion detection headband based on EEG measurement. *Sensors and Actuators A: Physical*. Retrieved from http://www.sciencedirect.com/science/article/pii/S092442471631192X

Xu, Y., Hübener, I., Seipp, A.-K., Ohly, S., & David, K. (2017). From the lab to the real-world: An investigation on the influence of human movement on Emotion Recognition using physiological signals. *Pervasive Computing and Communications Workshops (PerCom Workshops), 2017 IEEE International Conference on*, 345–350. IEEE.

Yang, J., & Honavar, V. (1998). Feature subset selection using a genetic algorithm. In *Feature extraction, construction and selection* (pp. 117–136). Retrieved from http://link.springer.com/chapter/10.1007/978-1-4615-5725-8_8

Yang, X., Ramesh, P., Chitta, R., Madhvanath, S., Bernal, E. A., & Luo, J. (2017). Deep Multimodal Representation Learning from Temporal Data. *arXiv Preprint arXiv:1704.03152*. Retrieved from https://arxiv.org/abs/1704.03152

Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., & Courville, A. (2015). Describing videos by exploiting temporal structure. *Proceedings of the IEEE International Conference on Computer Vision*, 4507–4515.

Yin, Z., Wang, Y., Liu, L., Zhang, W., & Zhang, J. (2017). Cross-Subject EEG Feature Selection for Emotion Recognition Using Transfer Recursive Feature

Elimination. *Frontiers in Neurorobotics*, *11*. https://doi.org/10.3389/fnbot.2017.00019

Zhang, F., Haddad, S., Nakisa, B., Rastgoo, M. N., Candido, C., Tjondronegoro, D., & de Dear, R. (2017). The effects of higher temperature setpoints during summer on office workers' cognitive load and thermal comfort. *Building and Environment*. Retrieved from http://www.sciencedirect.com/science/article/pii/S0360132317302834

Zhang, G., Shao, X., Li, P., & Gao, L. (2009). An effective hybrid particle swarm optimization algorithm for multi-objective flexible job-shop scheduling problem. *Computers & Industrial Engineering*, *56*(4), 1309–1318.

Zhang, S., & Zhao, Z. (2008). Feature selection filtering methods for emotion recognition in Chinese speech signal. *2008 9th International Conference on Signal Processing*, 1699–1702. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4697464

Zhang, T., Zheng, W., Cui, Z., Zong, Y., & Li, Y. (2018). Spatial-Temporal Recurrent Neural Network for Emotion Recognition. *IEEE Transactions on Cybernetics*, 1–9. https://doi.org/10.1109/TCYB.2017.2788081

Zheng, W., & Lu, B. (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

Zheng, W.-L., Dong, B.-N., & Lu, B.-L. (2014). Multimodal emotion recognition using EEG and eye tracking data. *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, 5040–5043. IEEE.

Zheng, W.-L., Zhu, J.-Y., Peng, Y., & Lu, B.-L. (2014a). EEG-based emotion classification using deep belief networks. *Multimedia and Expo (ICME), 2014 IEEE International Conference on*, 1–6. IEEE.

Zheng, W.-L., Zhu, J.-Y., Peng, Y., & Lu, B.-L. (2014b). EEG-based emotion classification using deep belief networks. *2014 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. https://doi.org/10.1109/ICME.2014.6890166

Zheng, Y., Liu, Q., Chen, E., Ge, Y., & Zhao, J. L. (2014). Time series classification using multi-channels deep convolutional neural networks. *International Conference on Web-Age Information Management*, 298–310. Springer.

Zhu, Q., Yan, Y., & Xing, Z. (2006). Robot path planning based on artificial potential field approach with simulated annealing. *Intelligent Systems Design and Applications, 2006. ISDA'06. Sixth International Conference on*, *2*, 622–627. Retrieved from http://ieeexplore.ieee.org/abstract/document/4021735/

Zhu, Y., Wang, S., & Ji, Q. (2014). Emotion recognition from users' EEG signals with the help of stimulus VIDEOS. *2014 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. https://doi.org/10.1109/ICME.2014.6890161

Zoph, B., & Le, Q. V. (2016). Neural architecture search with reinforcement learning. *arXiv Preprint arXiv:1611.01578*.