Faculty of Information Technology
Monash University

An Accurate and Explainable Multimodal Deep Fusion Network for Affect Recognition

This thesis is presented in partial fulfillment
of the requirements for the degree of
Master of Data Science at Monash University

Jionghao Lin

28287681

Master of Data Science

Supervised By: Dr. Vincent Lee

Year: 2019

Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the work of others has been acknowledged.

Signed by: Jionghao Lin

Date: 26 May 2019

Faculty of information tehcnology
Monash University
Caufield East, VIC, 3145
Australia

Acknowledgements

I would like to express my heartfelt and gratitude to all participants of this project for their professional expertise and dedicated contribution to this thesis.

First, I would like to thank Associate Professor Vincent Lee for his guidance and generous support throughout my research and revision process. His experience in supervising students has been inspiring me to complete the master thesis.

Then, I would like to thank Professor Sharon Oviatt who provided me an opportunity to join her team as research assistant and gave me access to the laboratory and research facilities. Without her precious support, it would not be possible to conduct this research.

Additionally, I would like to thank the rest team's members of Human Centered AI: Prof. Philip Cohen, Dr. Leiming Tian, and Dr. Jarrod Knibbe for their insightful comments and encouragement, widening my research from various perspectives.

Next, I would like to thank Julie Holden for providing many comments about academic writing and moderating this paper to improve the manuscript significantly.

Finally, I am greatly thankful to my parents Wei Lin and Shufen Jiang for their wholehearted support and encouragement.

Abstract

Affective analysis is an emerging research area which helps human to better understand their mental state through human-machine interaction. During the interaction process, bio-signal analysis is the essential work of detecting human affective changes since bio-signals are considered as a representation of physiological responses which are related to human affective states. Machine learning methods to analyse bio-signals are currently promoted as the better way to detect physiological changes, but most empirical works have utilised limited types of bio-signals in affect recognition, which cannot provide precise results. Moreover, these empirical works have mainly deployed traditional machine learning methods rather than deep learning models, which may have the opportunity to improve the classification accuracy.

This research provides a performance comparison between traditional machine learning models and deep learning models on multimodal bio-signals for human mental state classification tasks. The extensive experimental results suggest that deep learning algorithms outperform traditional machine learning algorithms in accuracy and weighted F_1 score for classification tasks in affective analysis.

Furthermore, to improve the explainability of the deep learning model, this research conducted a thorough analysis to understand the contribution of each bio-signal, and how they differ for various affective states. The research work improves the state of the art for emotion recognition from bio-signals and the current understanding of the relationship between affect and bio-signals.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Our daily life is surrounded by many modalities such as human vision, hearing, and smell. The term *modality* refers to a certain type of representation that stores information such as bio-signals, texts and videos [5]. Modalities can be characterised to different types of representation with various statistical properties, in particular sensory modality, which is a primary form of sensation to encode information [34]. Bio-signals are a type of sensory modality which can represent human physiological responses. This modality is increasingly applied to artificial intelligence to make progress in human computer interaction due to its numerical representation and the advancement in wearable sensors which allows convenient measurement. For example, the ECG signal collected by wearable sensors (e.g., Apple watch) is a certain sensory modality, which can interpret human heart activity patterns at many value points. Computers can process this numerical information to infer human mental states and then take action to interact with the human. However, many other signal modalities (e.g., EMG, EDA) are also collected at different sampling rates. When these modalities are processed by one computer simultaneously, techniques are needed for aligning and analysing multiple sources of information.

An emerging field called multimodal machine learning is established, which interprets and reasons about human activity from multimodal information [33]. This field involves signal denoising techniques, multimodal fusion, and machine learning algorithms to process multimodal-multisensory information. One significant application area of multimodal machine learning is affective computing which is broadly defined as an interdisciplinary field of psychology, computer science and signal processing [34]. In the affective computing, affect recognition (or affect detection) is a popular topic which can estimate the human latent mental state (e.g., stress) from machine-readable signals (e.g., bio-signals, video, and audio) to understand the human emotion and provide mental wellbeing support [43]. Most of affect recognition research focuses on analysing video, audio, and text modalities. Nevertheless, using those modalities to analyse human affect is quite limited due to mobility and applicability in human life. In comparison, bio-signals recently have received increasing attention since mobile devices are widely embedded with many types of bio-sensors to record the user's physiological signals continuously [43].

There are many types of time series signals in the real world such as bio-signals, and audio. Most of them have existed in a continuous form which is hard to process. The way to enable computers to process signal data more efficiently is to digitise the continuous signal. A weakness of digitisation is massive noise blended in the digitized process such as quantization error, and instrument errors [35]. To minimise the impact of errors or noise, signal processing techniques are commonly used. Signal processing is a field that analyses, operates, and modifies the time series information by using the theory in mathematics, information, and electronic engineering [43]. The existed signal processing techniques can filter noise in the raw signal at *Time domain* and *Frequency domain* [35]. When computers are utilised for analysing the multimodal signals, the processed signals need to be summarised in a joint representation (or feature vector) for the machine learning algorithms. Multimodal fusion is the field discussing the ways to integrate different modalities into a joint representation. There are two commonly used schemes for integrating multiple modalities: *early fusion* and *late fusion* [34, 45]. Early fusion concatenates the unimodal information into a joint representation for model classification, while late fusion processes the unimodal information individually, and combines the output results from all submodels as a joint representation for the final decision process.

Machine Learning is the study of discovering information from data, which have been applied to affective computing in the last two decades [33]. Machine learning approaches can be subcategorised into Shallow machine learning (SML) algorithms (e.g., Decision Tree, Support Vector Machine, and K-Nearest-Neighbour) and Deep learning algorithms (e.g., Convolutional Neural Networks, and Recurrent Neural Networks). SML algorithms are commonly used to affect recognition task due to their high explainability. However, their performance depends on the feature engineering step, which determines the upper bound of SML classification performance. In contrast, deep learning algorithms can avoid feature engineering steps and typically have better performance than SML models in processing large datasets [25]. In the task of facial recognition, deep learning algorithms even achieve better performance than human judgement [25]. However, many empirical works on affective computing have mainly focused on SML methods due to model simplicity and explainability [43]. In contrast, deep learning models are known as a "black box" to analyse the signals because the models have a complex structure and low explainability [7]. To improve the complex model's explainability, Been [7] has summarised three explainable phases, which are *1) Before building the models*, *2) Building the models*, and *3) After building the models*. Although these phases have been widely applied to the linguistic and visual modality, there has been little application to the study of physiological signals.

1.1 Motivation

This thesis is inspired by a state-of-the-art study in affective computing based on physiological signals by [42]. This study collected multiple types of physiological signals from 15 participants and conducted an experiment on distinguishing stress, amusement, and neutral state recognition using SML models including Random Forest, Adaboosting, and Linear Discriminant Analysis [42]. Although the models achieved admirable results using carefully selected features, the classification performance (accuracy and F_1 score) can be potentially improved using other methods such as deep learning models. However, the main disadvantage of deep learning is the low explainability, which restricts applying deep learning to analyse physiological signals in affect recognition. Therefore, this research is motivated to explore the performance and explainability of a deep learning model for distinguishing stress, amusement, and neutral state from multiple physiological signals.

1.2 Research Methodology

1.2.1 Research Objectives and Questions

This thesis investigates methods to improve the classification accuracy and explainability in affect recognition tasks based on the multimodal-multisensory dataset provided by [42]. The study utilises this dataset and implements the same classification task but focuses on deep learning models and their explainability. By exploiting the empirical works in multimodal deep learning applications, this research builds several models to facilitate both the classification task and the explainable tasks. The research objectives and questions are detailed as follows:

Objective 1: Design several deep fusion networks for affect recognition based on multimodal bio-signals, and select the best one as the proposed model.

RQ 1.1: How to build an accurate deep learning model on the multimodal-multisensory dataset?

RQ 1.2: Which fusion scheme is better for affect recognition on multimodal bio-signals?

Objective 2: Provide experimental evidences to demonstrate that the proposed model has better classification performance than the state of art.

RQ 2.1: To what extent, are the deep learning models better than the state of the art for the same affect recognition task?

Objective 3: Provide the explainability analyses on the classification results of the deep fusion network in three explainable phases.

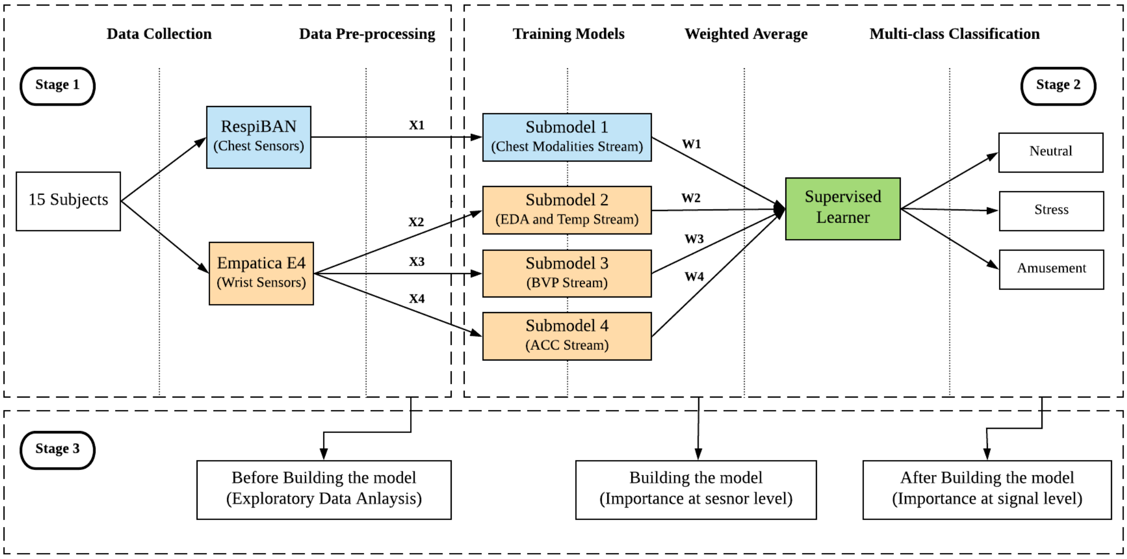
RQ 3.1: What are the various bio-signal patterns unique to stress, amusement, and neutral states?

RQ 3.2: What is the importance of the chest and wrist wearable sensor for affect recognition?

RQ 3.3: What is the importance of different types of signals for affect recognition?

1.2.2 Research Framework

Figure 1.1 displays the operational framework for this research which is divided into three stages. This research framework is based on and extends the data processing chain proposed by [44].



Data Collection
Data Pre-processing
Training Models
Weighted Average
Multi-class Classification
Stage 1
Stage 2
15 Subjects
RespiBAN (Chest Sensors)
Empatica E4 (Wrist Sensors)
X1
X2
X3
X4
Submodel 1 (Chest Modalities Stream)
Submodel 2 (EDA and Temp Stream)
Submodel 3 (BVP Stream)
Submodel 4 (ACC Stream)
W1
W2
W3
W4
Supervised Learner
Neutral
Stress
Amusement
Stage 3
Before Building the model (Exploratory Data Anlaysis)
Building the model (Importance at sesnor level)
After Building the model (Importance at signal level)

Figure 1.1: Research Framework

Stage 1 is based on data collection and pre-processing. Since the dataset has been collected from 15 subjects using chest and wrist sensors [42] in a less control environment, this research starts with the data pre-processing step. There are many techniques involved in pre-processing: outliers processing, signal denoising, data normalisation, and segmentation. These are necessary to ensure the model can work as expected on their dataset. The processed signals will be set as the input for model training and explored by the *Before building the model* phase.

Stage 2 is based on training models, classification, and evaluation. Since the processed signal sampling rates are different, the multimodal fusion scheme is adopted to generate joint representation for the final classifier. Although the benchmark has achieved 80% classification accuracy, it can be potentially improved by using deep learning algorithms. Therefore, each submodel will be built based on the deep learning architecture. The models' hyper-parameters (e.g., layer selection, and batch size) will be tuned to the optimal level, and these models are then evaluated by accuracy and F_1 score using Leave-One-Subject-Out cross-validation.

Stage 3 aims to understand the classification results and explainability. To obtain interpretable results, three explainable phases will be deployed to the previous two stages.

- *Before building the model* is conducted after data pre-processing to distinguish the differences between three affective states in various bio-signals.

- *Building the model* is analysed based on the importance (i.e., stream weights, W_i) at the sensors-level to understand which sensor contributes more to the classification results.

- *After building the model* is explored based on the importance at the signal-level to assess the contribution of each signal to the model performance.

1.2.3 Research Data

This research adopts naturalistic data named WESAD (Wearable Stress and Affect Detection data set) for the experimental simulation retrieved from UCI machine learning repository[1]. The dataset is provided by [42] and contains 15 subjects' data. Each subject wore two wearable sensors: *RespiBAN* and *Empatica E4*. The RespiBAN sensor is used to record the bio-signals elicited from the chest, and the Empatica E4 sensor is used to capture the bio-signals on the participants' non-dominant wrist. The specific number of records for each subject is presented in Appendix A.

These subjects experienced three main activities (baseline, amusement, and stress) to induce the target emotion states. To avoid the effect of the activities' order, the data collection was conducted interchangeably (shown in figure 1.2). Both orders start with the *Baseline* stage which is the control group. The red bar is a checkpoint representing the end of the previous activity, and subjects needed to fill out several types of questionnaires to obtain the ground truth. There was a *Rest* stage following the *Stress* stage to ensure that the result in meditation was unbiased,

[1] UCI machine learning repository https://archive.ics.uci.edu/ml/datasets.php

which had a duration of 10 minutes. Finally, meditation (*Medi*) was the stage to relieve stress and excitement.

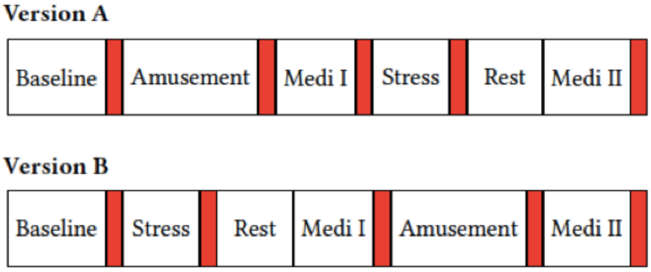


Figure 1.2: Ways to collect bio-signals
Image retrieved from [42]

1.2.4 Experimental Configuration

There are more than 20 millions data records in the WESAD dataset which are processed on a computer with GPU (NVIDIA GeForce RTX 2080 Ti), CPU (Intel(R) Core(TM) i7-9700), and RAM (32 GB). The models are implemented on Python 3.6. The main external packages are Biosppy, Keras, and Sklearn. These packages are introduced as follows.

Biosppy —provides many types of biological signals analytical template including BVP, ECG, Resp, and EDA. The different types of filters (i.e., FIR and Butterworth) and bands (low-pass, high-pass, and band-pass) are encapsulated in this library. Additionally, this library can also extract some basic signal features such as heart rate, respiratory cycle.

Keras —is an open source neural network API written in Python, which is running on top of Tensorflow, CNTK, and Theano. It provides high-level neural network design framework and fast implementation of experiments. Keras also has many extensible functions for the user to define such as the learning rate in optimizers.

Sklearn —is a free machine learning package in Python programming language. It provides a range of classification, regression, and clustering algorithms such as random forest, linear regression, and k-means. It is characterized by a streamlined, uniform, and clean API, and it is easy to implement the machine learning algorithms. Moreover, it also provides data pre-processing tools such as min-max normalization.

Other packages are applied to manipulate the data type and shape (e.g., Numpy, Scipy, and Pandas), and to operate exploratory data analysis (e.g., Matplotlib).

1.3 Thesis Structure

This thesis is organised as follows:

Chapter 1 gives an introduction to this research to explain the needs of deep learning application in affective computing based on bio-signals and the needs of explainability for deep learning algorithms.

Chapter 2 provides a brief review of the literature on human affective states, multimodal fusion schemes, classification algorithms, and model explainability.

Chapter 3 describes the preprocessing methods on an original dataset including removing irrelevant information, denoising raw signals, and data normalisation. The input preparation is discussed in detail for future repeated experiments.

Chapter 4 provides the exploratory data analysis for each signal in stress, amusement, and normal states. All the physiological signals have been analysed by various statistical features to find differences.

Chapter 5 introduces multimodal deep fusion networks to affect recognition based on bio-signals. It is designed to solve the modalities which are collected by the different sampling rate and generate explainable results.

Chapter 6 presents the ways to conduct the experiment and the classification results measured by accuracy and F_1 score. The best model has a significant improvement in both measurements compared with the state of the art. Additionally, the performance of early and late fusions are discussed in this chapter.

Chapter 7 explains the classified results at sensor- and signal-level. Three explainable phases are adopted to analyse the importance of each modality.

Chapter 8 concludes the research, discusses the limitations, and provides recommendation for future works in the field of multimodal machine learning.

Chapter 2

Literature Review

This chapter provides some backgrounds on human affective states, bio-signals, machine learning models, multimodal fusion schemes and explainability for modelling. Section 2.1 introduces the definition of some terms in affective computing. Section 2.2 introduces the sensors that [42] used to collect bio-signals, and the characteristics of bio-signals. Section 2.3 discusses the empirical works on shallow machine learning and deep learning in the field of affective computing based on bio-signals. Section 2.4 presents the commonly used multimodal fusion schemes, and section 2.5 describes the three explainable phases to interpret differences between various states. Finally, evaluation metrics are presented in Section 2.6 for model performance measurement.

2.1 Background of Human Affective States

2.1.1 Emotion model

The affective state is an undirected feeling and usually lasts for a long time. It is defined as a neurophysiological state which is consciously accessible as the nonreflective and simple feeling composed of arousal and valence [40]. In contrast, emotions are the directed and temporal feelings arising from a cognitive process by stimuli which can also be considered as the indicator of affect [40]. Although the empirical works provide many definitions of these two terms, it is still hard to distinguish them clearly [43]. There is a way to quantify the human affect or emotion for computing which is the circumplex model or emotion model. This model is frequently mentioned in the articles on human affective recognition which is normally composed of *categorical models* and *dimensional models* [43]. Categorical models include many discrete categories to represent different emotions such as stress, amusement and anger [43]. As for dimensional models, they can map different emotions into multi-dimensional space [43]. The valence-arousal model (shown in Figure 2.1) was proposed by [40] to vectorise the affective representation in the vector space. The valence axis defines the polarity of experiencing emotion whether it is a positive or a negative feeling [40]. The arousal axis indicates the level of excitement or calm [40].

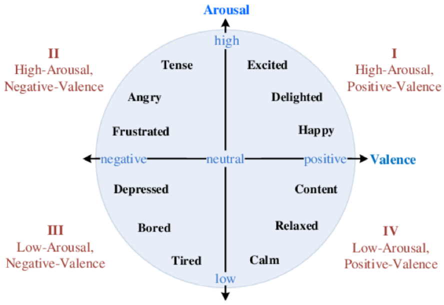


Figure 2.1: Circumplex Model
Image retreved from [40]

2.1.2 Stress and Amusement

Stress and amusement are two types of human affective or mental states. Stress is a term used to describe an experience under pressure [43]. It is often elicited by physiological (external) or psychological (internal) stimuli [43]. From a scientific view, stress is primarily viewed as a physiological response, which can be classified as eustress (positive outcome) and distress (negative outcome) [30]. Eustress has a positive (e.g. motivating) effect on an individual [43]. For example, when a student has prepared well for an exam (the exam is considered as a stimulus), the body can provide more energy as a reaction to the stimulus. This body reaction could motivate students to look forward to solving the questions. In contrast, if the students do not review the class content, they worry about failing the exam and feel anxious to attend, which could be considered as distress (e.g. anxious and worried). On the other hand, amusement is defined as the positive human affective state of experiencing an entertaining event which has a similar arousal level as happy [43]. To measure whether a person is amused or stressed, researchers normally record physiological responses such as sweating, faster breathing, and heart palpitations to infer affective states [43]. The physiological information is recorded by sensors in the form of bio-signals. The description of various bio-signals will be introduced in section 2.2.1.

2.2 Background of Bio-signals

2.2.1 Sensors and Bio-ignals

This section provides an overview of sensory modalities to inference human affect. The wearable sensor is a device to sample the continuous signals into a digital representation. In the process of conversion, the noise in the signal is unavoidable. Therefore, using a high-quality sensor is necessary to reduce the percentage of noise in the signal. Schmidt et al. [42] utilized *RespiBAN Professional*[1] for collecting chest signals and *Empatica E4*[2] for wrist signals. The details for these sensors are as follows:

1) RespiBAN —It is a wearable chest sensor which is placed around the participant chest shown as Figure 2.2. which can collect six various bio-signals in total, and the sampling rate for these signals is at 700 HZ.

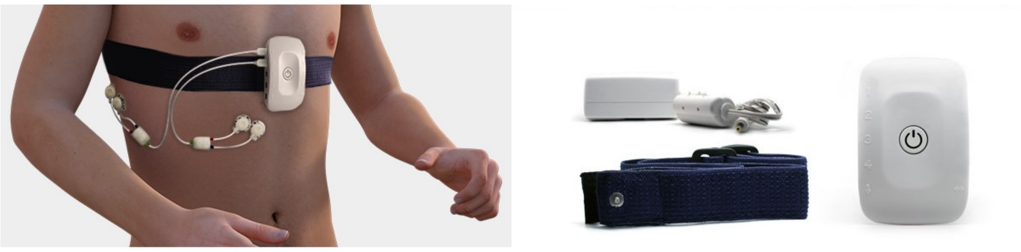


Figure 2.2: Placement of the RespiBAN on human torso.
Image retrieved from https://www.biosignalsplux.com/en/respiban-professional

2) Empatica E4 —It is a wearable wrist sensor that can collect the signals ACC (32HZ), BVP (64HZ), and EDA and Temp (4HZ). Subjects wear the Empatica E4 on the wrist.

The physiological changes in bio-signals are quite related to the affective states so most studies utilised bio-signals for affective recognition task [43]. The bio-signals collected by above two sensors are introduced as follows:

Electrocardiogram (ECG) can illustrate the human heart activity, which is considered as a reliable source for heart disease diagnosis. It can also analyse human mental states like stress. The standard sensor for ECG signal has three electrodes (shown in Figure 2.3) placed on the subject torso, and the minimum sampling rate is 50 HZ [31].

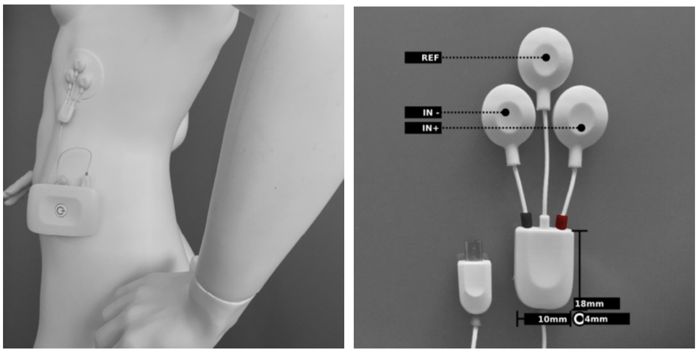


Figure 2.3: ECG signal collection by RespiBAN
Image retrieved from https://www.biosignalsplux.com/en/

[1] https://www.biosignalsplux.com/en/respiban-professional
[2] https://www.empatica.com/research/e4/

Electromyography (EMG) can present the activity cycle, amplitude and other features of the muscle. Commonly, researchers deploy a pair of electrodes to the skin on the muscle such as the shoulder (shown in Figure 2.4). When the muscle cells are activated, the surface electrodes can record the muscle physiological changes. The minimal sampling rate is 1000 HZ [43].

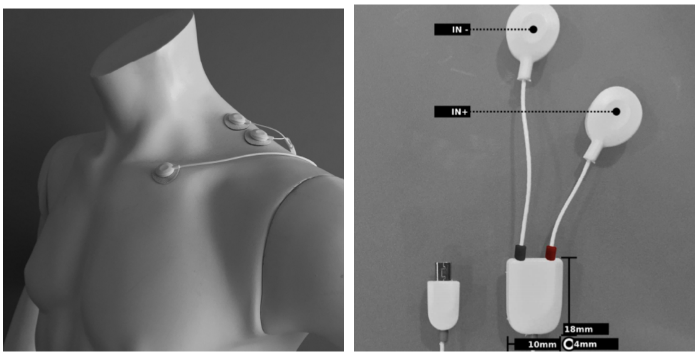


Figure 2.4: EMG signal collection by RespiBAN
Image retrieved from https://www.biosignalsplux.com/en/

Respiration (RESP) is utilized to record human inhalation and exhalation activity. Commonly, a subject is equipped with a chest belt (shown in Figure 2.5) measuring the respiration pattern directly. The minimal sample rate for RESP sensor is 31 HZ [17].

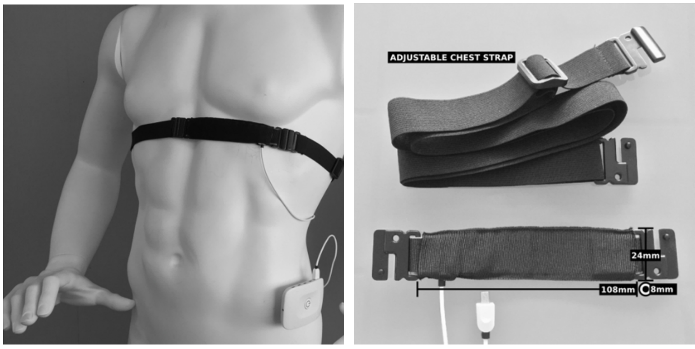


Figure 2.5: RESP signal collection by RespiBAN
Image retrieved from https://www.biosignalsplux.com/en/

3-axis Accelerometer (ACC) can record human movement in 3-dimensional space (X, Y, Z), which can measure physical activity, the range of motion, and shocks. These activities are provided as context information to recognize affective states. The intensity level of activities is estimated for stress detection instead of classifying different activities [36].

Temperature (TEMP) records the suitable temperatures for the human body. It is measured by the temperature-dependent resistors. Due to low-frequent changes of the body temperature, 1 HZ sampling rate is sufficient [43].

Electrodermal activity (EDA) is commonly utilized to measure the activity of electrical changes at the surface of the skin placed at the sweat gland such as wrist or palm (shown in Figure 2.6). The minimum sampling rate to acquired EDA signal is at 31 HZ [17].

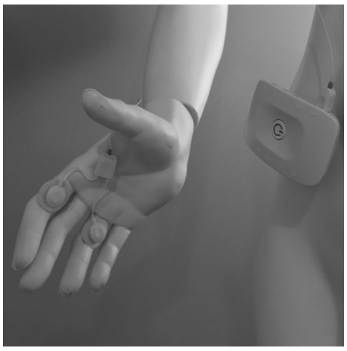


Figure 2.6: EDA signal collection by RespiBAN
Image retrieved from https://www.biosignalsplux.com/en/

Blood Volume Pulse (BVP) is widely used to measure the heart rate based on the volume of blood and is embedded in many heart rate variability biofeedback systems and apps [43].

2.3 Machine Learning in Affective Computing

2.3.1 Shallow Machine Learning Models

Classification task in affective computing can be performed by statistical test (e.g., ANOVA) or machine learning algorithms [43]. In comparison, machine learning methods, especially for SML, are used more frequently than statistical test in bio-signals to detect affective states [28, 42, 46]. For example, Lisetti and Nasoz [28] used K-Nearest Neighbor, Linear Discriminant Analysis with ECG, EDA, and Temp to detect six affective states; Katsis et al. [21] used Support Vector Machine with EDA, ECG, EMG, Resp to detect stress, disappointment, and euphoria. The details about SML algorithms on bio-signals for affective detection task are summarised in Table 2.1.

Author	Signals	Classification	Detected States	Performance
Lisetti and Nasoz [28]	ECG, EDA, Temp	KNN, LDA, NN	Sadness, amusement, fear, surprise, anger, frustration	Acc: 84%
Healey and Picard [17]	ACC, EDA, HR, audio	LDF	Level of Stress	Acc: 97%
Katsis et al. [21]	ECG, EDA, EMG, Resp	SVM, ANFIS	Stress, disappointment, euphoria	Acc: 79%
Schmidt et al. [42]	ECG, EDA, Temp, ACC, BVP, Resp,EMG	DT, RF, KNN, LDA, Adaboost	Stress, amusement, neutral	Acc: 80% F1-score: 74%

Table 2.1: Summary of traditional machine learning on bio-signals.
Abbreviations: Classification Accuracy (ACC), Decision Tree (DF), Random Forest (RF), K Nearest Neighbour (KNN), Linear Discriminant Analysis (LDA), NN (Neural Networks), Linear Discriminant Function (LDF), Support Vector Machine (SVM), Adaptive Neuro-fuzzy Inference system (ANFIS), Logistic Regression (LR), Multilayer Perceptron (MLP), Naive Bayes (NB), Bayesian Network (BN), Funtion Tree (FT)

Nevertheless, SML heavily relies on feature extraction, which focuses on the knowledge inspired features but ignores some other potential features [15, 32]. In other words, it requires domain knowledge to construct a set of features as the input for the model training but the domain knowledge might not cover all the significant features. Some unexplored significant features can determine the upper bound of the model performance. On the other hand, deep learning algorithms can avoid the feature engineering step. Many works on deep learning based applications have also obtained breakthrough progress such as computer vision, and speech recognition [8, 12, 38]. However, deep learning algorithms are still not the main role in the field of bio-signal based affect recognition [2]. To narrow the gap, this research will adopt deep learning algorithms which are expected to obtain better classification performance than the state of the art on the affect recognition task.

2.3.2 Deep Learning Models

Deep learning model consists of multiple layers and each layer can learn a vector representation at a different complexity. The fundamental of deep learning is neural networks [25]. Generally, standard neural networks (NN) are composed of many *fully connected layers* (or *dense layers*) which connect all the neurons with the previous layer and classify the results [25]. The fully connected layer is the basic layer in the neural network which is used for classification. Before classifying results, fully connected neural networks normally need convolutional layers to extract features vector from data [25]. For time series data, recurrent layers are considered more effective in processing the information than the convolutional layers [8]. The details of convolutional and recurrent neural networks are introduced as follows:

Convolutional Neural Network (CNN) is composed of convolutional layers, pooling layers, and fully connected layers [24]. Convolutional layers are designed to extract features from signals and then feed the features to fully connected layers for classification (shown in figure 2.7).

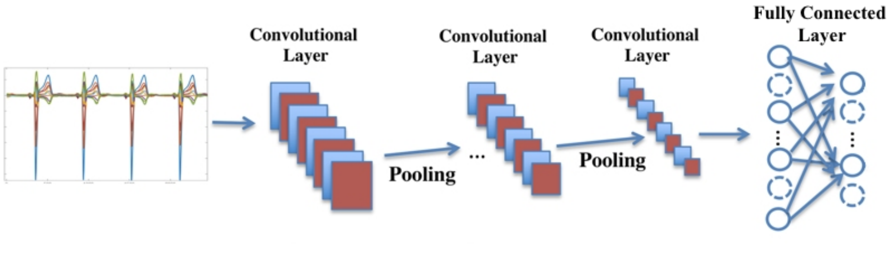


Figure 2.7: Convolutional neural networks for signal
Image retrieved from [4]

- *Convolutional Layer:* It is used to slide across each bio-signal with filters that can extract features from inputs [4]. Features are continuously extracted and compressed, which means that the original features are concentrated with the increasing number of convolutional layers. Therefore, high-level features are generated in the last layer of convolution.

- *Pooling Layer:* It is used to perform subsampling on the input feature map from the convolutional layer and simplifies the network computing complexity to prevent overfitting and reduce computational intensity [4]. The pooling layer is optional in the CNN network.

Recurrent Neural Network (RNN) is normally used to process and predict sequential data-related tasks such as speech recognition, and machine translation [48]. RNN has been proved to outperform CNN at processing sequential data because RNN considers sequential order and context [22]. A typical RNN (shown in Figure 2.8) can deliver the processed information from the previous state to the current state [10]. Based on the RNN structure, it could remember the history information, and use this information to determine the output. Additionally, there is an improved version of RNN called Long Short Term Memory (LSTM), which can capture long-term dependencies better than ordinary RNN and solve the gradient vanishing problem in RNN [10]. The memory in LSTM is called a cell. Internally, the cell determines the reservation of information in memory. Then, the current memory and the input from the previous state are combined. LSTM has been proven to be more effective in capturing long-term dependencies in time series data processing [10].

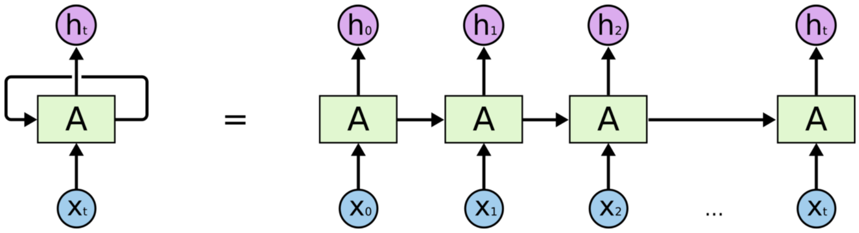


Figure 2.8: Recurrent Neural Network
Image Retrieved from http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Hybrid Deep Neural Networks (HDNN) is a model framework that combines various functional layers. The previous introduction summarised that CNN is usually used for image processing, and RNN is for time series processing. Based on the respective advantages, Bashivan et al. [6] found that the hybrid deep learning model achieved high accuracy in stress detection tasks.

Deep learning models have been proved to outperform SML methods in many fields [6, 25], but there are few empirical works applying deep learning algorithms to human mental state recognition based on bio-signals due to low explainability [43]. Moreover, most of these empirical study focused on individual signals (shown in

Author	Input Bio-signals	Classification Methods
Acharya et al. [3]	ECG	9-layer CNN to categorize ECG beats
Yildirim [48]	ECG	LSTM to classify ECG
Kiranyaz et al. [23]	ECG	CNN to classify ECG in real time
Su et al. [47]	Blood Pressure	LSTM to predict Blood Pressure

Table 2.2: Summary of deep learning models on bio-signal related task

table 2.2). Compared to multimodal bio-signals, the unimodal result is less reliable [43]. Therefore, this research will explore the classification performance of deep learning models using multimodal bio-signals while improving their explainability.

2.4 Multimodal Fusion Schemes

When the model input consists of the representations from multiple bio-signals at different sampling rates, multimodal fusion schemes can be deployed in model structure to integrate these signals for classification [34]. Since the deep learning model is normally stacked with multiple functional layers, it allows multiple bio-signals fusion at any stage. There are three basic levels of multimodal fusion, namely data, feature, and decision level. Figure 2.9 takes two modalities as an example for the integration in the deep learning model, and the details of the basic fusion levels or schemes are introduced in the following sections.

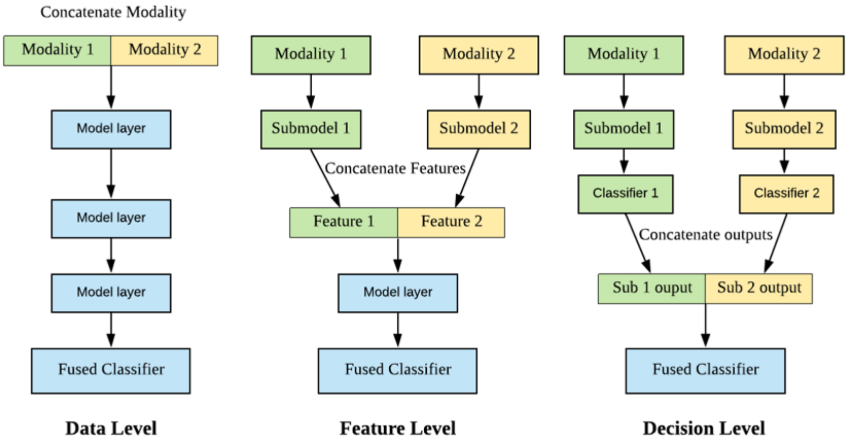


Figure 2.9: Basic Fusion Strategies for two modalities

2.4.1 Data Level

Fusion in data level is also called signal-level fusion which is similar to the multivariate tabular data. As shown in the left of Figure 2.9, multiple modalities are concatenated at the first stage of the model diagram, and then the representation is fed to the model's first layer [33]. The main advantage of this scheme is that it has less information loss compared with the other two schemes. The main drawback is that it is failed to integrate modalities with different input shapes.

2.4.2 Feature Level

Fusion in feature level is also called early fusion, multiple modalities are integrated before being fed to the classifier [33]. As shown in the middle of Figure 2.9, two modalities are fed into submodels (without classifiers) separately and then concatenate the feature vectors before the classifier. The main advantage of this scheme is that it can compress the information which is more efficient for the streaming process than the other schemes. The main drawback is that it simply concatenates the features from multiple modalities which lose the original data structure.

2.4.3 Decision Level

Fusion in decision level is also called late fusion [33]. As shown in the right of Figure 2.9, multiple bio-signals are trained individually on each submodel (with a classifier), and then the outputs from the separated classifiers are concatenated as the input for the final stacking classifier. The learning process of late fusion is completed in two steps: 1) training the submodel classifiers, and 2) training the final fused classifier. The main advantage of this scheme is that it can lower the voting weight for the weak classifier. The main drawback of this scheme is that the computational cost is much higher than the other two schemes.

2.4.4 Hybrid Fusion Scheme

The hybrid fusion scheme combines the above three levels to the model pipeline, which is expected to take advantages of each scheme for better classification performance [33].

2.5 Explainability for Modelling Tasks

Deep learning has achieved leading performance in many classification tasks. However, a big issue in deep learning is that it is hard to explain the variables or features which influence the results [14]. A model which is hard to explain influence or correlation from various features is referred as a ”black-box” (e.g., deep learning models) [13]. Explainability is also an important evaluation metric of machine learning algorithms [1, 26]. Explainability is a model ability to interpret important features or variables in machine learning tasks [1, 13, 14].

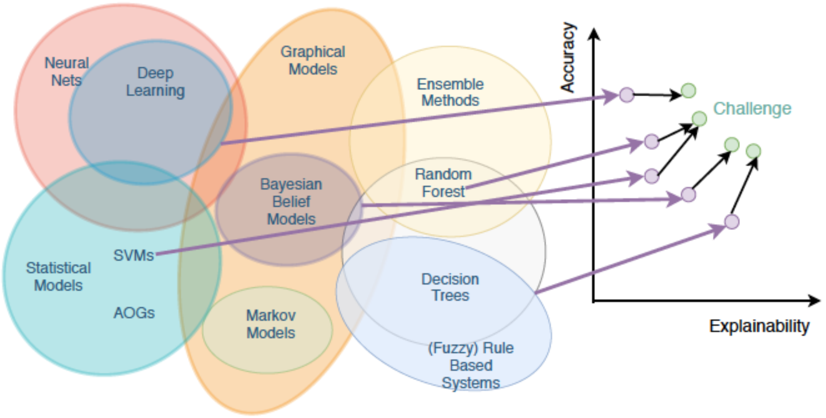


Figure 2.10: Accuracy Vs. explainability
Image retrieved from [13]

Researchers use the word interpretability and explainability interchangeably. Currently, model interpretability and accuracy are the criteria to measure the model performance [13, 14, 19]. Nevertheless, they are hard to keep a balance (shown in Figure 2.10). For example, the traditional ML method such as decision tree has decent interpretability but generally low accuracy. On the other hand, deep learning models have high accuracy in general but low interpretability. The current challenges are to obtain higher explainability. There are three explainable phases summarised by [7] which are introduced in the following sections.

2.5.1 Before building the models

After filter the noise from the raw signals, these signals can be plotted to analyse some statistical features such as mean, maximum and minimum values. This step is named *Exploratory data analysis* which aims to explore the intuitive patterns, check the research assumptions, and observe the correlation between variables and results visually [7]. The exploratory data analysis is completed before feeding the data to models.

2.5.2 Building the models

It includes three types of approaches to interpret models [7]. *Rule-based approach* such as decision tree lists a set of rules to explain the decision process [13]. *Feature-based approach* can assign weights to each feature to explain the important features [7]. *Case-based approach* can interpret the results based on the similarity around the analysing target [7]. In this research, the feature-based approach is utilised to explain the importance of sensor-level.

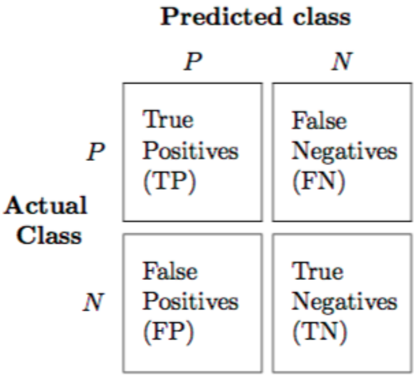
2.5.3 After building the models

There are three approaches in this method: *Ablation experiment*, *Surrogated models*, and *Investigation on hidden models* [7, 27]. Ablation test or sensitivity analysis can help to explain the important features [7]. The surrogate model is an approach to find a high explainability model to mimic the 'black-box' model training on the same predictors and outputs, and then use the surrogate model to explain the 'black-box' model [27]. Investigating hidden layers is the approach which usually applies to images. Since CNN can extract features from the raw data, the high-level features can be visualised in the last convolutional layers [7]. For after building the models phase, the ablation experiment is utilised to explain the importance at signal-level.

Most empirical explainable studies deployed explainable methods on images and text instead of time series data [7, 27]. Therefore, this research will explore the explainability of affect recognition using deep learning models with time series data collected as bio-signals.

2.6 Model Evaluation Metrics

After tuning the parameters for the proposed models, the next step is to check the model performance for multi-classification tasks. There are two commonly used methods to evaluate models in signal processing: classification accuracy and F_1 score [42].



		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Figure 2.11: Binary classification confusion matrix
Image retrieved from http://rasbt.github.io/mlxtend/user_guide/evaluate/confusion_matrix/

- *Classification Accuracy* is defined as the percentage of predicting the correct results (shown in equation 2.1)which includes true positive and true negative (shown in Figure 2.11). For the classification of the affective state, accuracy is used to record the corrected prediction for stress, amusement, and neutral states [42].

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{2.1}$$

- F_1 score is considered as a robust metric for unbalanced dataset classification (shown in equation 2.4), which is defined as the harmonic mean of recall (shown in equation 2.2) and precision (shown in equation 2.3).

$$\text{Recall} = \frac{\text{Number of true predicted labels}}{\text{Total number of labels}} \tag{2.2}$$

$$\text{Precision} = \frac{\text{Number of true predicted labels}}{\text{Total number of predicted labels}} \tag{2.3}$$

$$F_1\text{score} = \frac{2 \cdot \text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}} \tag{2.4}$$

Since the state of the art utilised these two metrics to evaluate the model performance, this research will continue to use them to measure the proposed model.

2.7 Chapter Summary

This chapter has summarised the background of the research and discussed the research gaps in the empirical works on affect recognition tasks. The literature [42, 43] claim that the human physiological responses are related to the biological signals. One of the ways to analyse the correlation is to use machine learning algorithms. Basically, there are two subcategories in machine learning: *Shallow Machine Learning* and *Deep Learning*. In terms of shallow machine learning, it highly relies on the feature engineering step and requires domain knowledge to find effective features. On the other hand, deep learning algorithms can avoid the step to design the features and provide generally high classification performance on a large dataset. However, deep learning is considered as the 'black-box' model due to their low explainability which is still a big challenge in this community. To improve the explainability, there are three explainable phases proposed by previous works. The details about implementing an experiment on affective detection and explainability are demonstrated in the following chapters.

Chapter 3

Data Pre-processing

Chapter 3 describes the data preprocessing works including outliers processing, signal denoising, data normalisation, and data segmentation. Data pre-processing is considered as a crucial step in the experiment. First, the target records should be selected since there are many irrelevant labels such as meditation. Subsequently, the outliers and data noise should be processed to avoid their negative influences on model performance. Then, since the values for each signal are different, the model normalisation needs to be applied to transform the signals to the same scale which enables effective model training. After processing the signals based on the above steps, these signals are synchronised and segmented into many aggregated windows as the model input. The details of these steps are introduced in this chapter.

3.1 Data Description

This research simulates the experiment on WESAD dataset (Wearable Stress and Affect Detection dataset), which was collected by Schmidt et al. [42] for affective recognition based on bio-signals. There are 15 subjects in WESAD dataset, and the file size for each subject is presented in Appendix A. Additionally, Schmidt et al [42] provided the note for the participant's personal information (e.g., age, gender, and height), study pre-requisites (e.g., Did you drink coffee today? Are you a smoker?), and additional notes[1]. The participants' age ranges from 24 to 35, and the average age is 27.5. There are 12 males and 3 females. The chest and wrist signals were stored in each subject's file separately. In the participant file, it contains the chest and wrist bio-signal, labels and participant's information. The RespiBAN sensor was worn on the subject's torso collecting all signals in sampling rate 700 HZ. On the other hand, the Empatica E4 sensor was worn on the participant's non-dominant wrist. The sampling rate for various bio-signals is different[2]. Then, the ground truth

[1] Record the Unexpected behaviour happened in experiment

[2] ACC in 32 HZ; BVP in 64 HZ; EDA and Temp in 4 HZ

was encoded as *not defined/transient* - 0, *baseline* - 1, *stress* - 2. *amusement* - 3, *meditation* - 4, *meaningless* - 5/6/7. Baseline (or neutral), amusement, and stress are extracted as the targets in this research. These three target labels were obtained in the following conditions:

Baseline condition: Subjects were sitting or standing at a table and reading provided magazines.

Amusement condition: Subjects watched a set of funny video clips. These clips were selected from the corpus by [41].

Stress condition: Subjects were asked to give a public presentation or have a mental arithmetic task.

These target labels are imbalanced in each subject's data file. The proportion is approximately 5 (Baseline): 3 (Amusement): 2 (Stress).

3.2 Outliers Processing

The information recorded in the additional notes presents that some subjects have some special conditions which are different from the typical subjects [42]. For example, subject 6 had completed many interviews before the stress test, and the sensors of subject 2 and 17 were not fully attached. The details of this information were attached in Appendix B. Considering these note information, this research does not include them in the testing set, but in the training set. Therefore, the test set consists of subject 4, 7, 9, 10, 11, 13, and 14.

3.3 Signal Denoisng

Normally, raw signals contain too much noise. Take 10 seconds Resp signal from a subject as an example (shown in 3.1), it presents the noise around the line. There are two commonly used ways to denoise signal: *Time domain* filtering and *Frequency domain* filtering [35]. The time domain method utilises many types of sliding windows to smooth the signal. On the other hand, the frequency domain method filters the noise frequency components. This research focuses on the latter method to filter the noise.

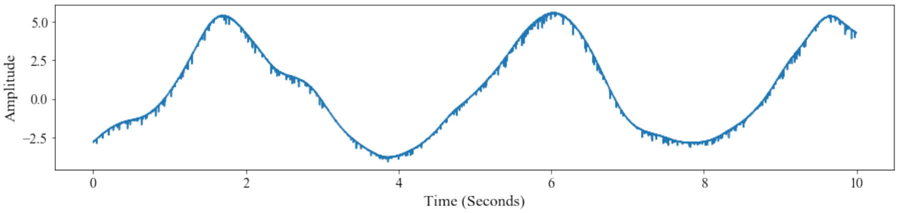


Figure 3.1: Unprocessed Resp signal from a subject

To filter the noise in the frequency domain, it is important to first observe the frequency components. *Forward Fourier Transformation* is utilised to transform the signal from time domain representation to frequency components [37] (See Equation 3.1). After transformation, the noise and signal are displayed in separated frequency components (shown in Figure 3.2).

$$X(w) = \sum_{t=0}^{+\infty} x(t)e^{-jwt} \tag{3.1}$$

Where x(t) is a function of time; X(w) is a function of frequency, and the function values of X(w) are complex numbers.

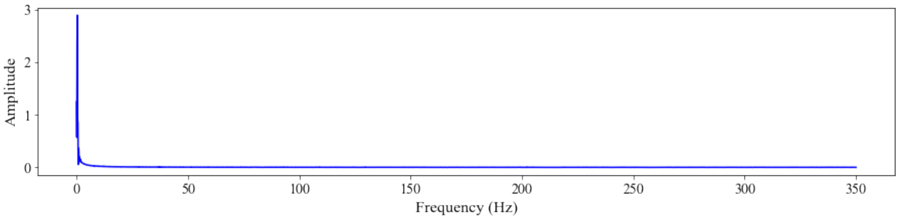


Figure 3.2: Unprocessed Resp signal's frequency domain from the subject

In the frequency domain, filters can suppress the unexpected signal frequency components to pass through the system and then obtains the desired signal frequency components [37]. The filter type can be classified based on the range of signal frequency in the pass-band (e.g., low pass, and high pass); on the frequency response (e.g., Butterworth filter). In this research, the following filters are used in frequency domain denoising.

- **Finite impulse response (FIR)** —It is a filter that the impulse responses have a finite duration. It can be designed to be a linear phase which is more stable and easier to analyse compare with the infinite impulse response system.
- **Butterworth**: It is designed to flatten the frequency response in the pass-band and obtain a smoothed frequency response.
- **Lowpass**: It allows the low frequency (lower than the cutoff frequency[3]) signal to go through the filter but attenuate the high-frequency signal components.
- **Highpass**: It allows the high frequency (higher than the cutoff frequency) signal to pass the filter but attenuate the low-frequency signal components.
- **Bandpass**: It allows the signal to pass the defined range, and attenuate the frequency out of the pass-band[4].

[3] Cutoff frequency is the threshold that the signal in the frequency domain beyond that level will not be passed.

[4] pass band is the band allowing the frequency to pass in a certain band.

The selection of filter, band type and cutoff frequency for various bio-signals follows the empirical works [9, 42, 43].

- **ECG Pre-processing**: ECG is processed by the FIR filter and bandpass with the passband from 3 HZ to 45 HZ [9].
- **BVP Pre-processing**: BVP is processed by the Butterworth filter and bandpass with the passband from 1 HZ to 8 HZ to filter [9].
- **RESP Pre-processing**: Resp is processed by the Butterworth and bandpass with the pass band from 0.1 HZ to 0.35 HZ [42].
- **EDA Pre-processing**: EDA is processed by the low-pass band to filter the frequency components higher than 5 HZ. According to [43], physiological plausible changes normally happen on the low-frequency domain in the EDA signal.
- **ACC Pre-processing**: ACC signal is processed by the Butterworth filter with the low-pass band to filter the frequency higher than 5 HZ [43].
- **EMG Pre-processing**: EMG signal is processed by the Butterworth filter and highpass band to filter the frequency lower than 100 HZ [9].

After filtering the noise in frequency domain for Resp signal, *Inverse Fourier Transformation* is then used to transform the processed frequency components back to the time domain representation [37] (See Equation 3.2). Finally, the smoothed Resp signal is presented in Figure 3.3.

$$x(t) = \frac{1}{2\pi} \sum_{w=0}^{+\infty} X(w) e^{jwt} \tag{3.2}$$

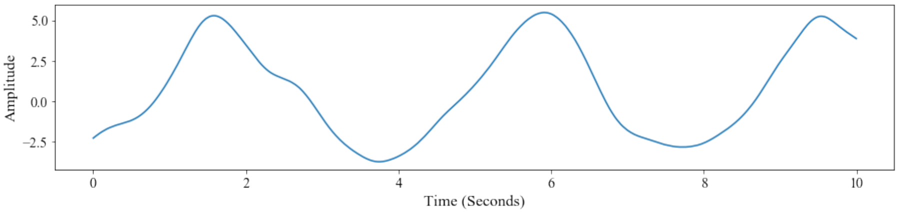


Figure 3.3: Processed Resp signal from a subject

Based on the literature of bio-signal pre-processing, the processed and unprocessed signals for all chest modalities in different affective states are presented in Appendix C.

3.4 Data Normalisation

Min-max normalisation was applied across different signals to build the same scale for model training (shown in Equation 3.3). The values were converted to the range from 0 to 1.

$$Min - Max\ scores = \frac{x - min(x)}{max(x) - min(x)} \tag{3.3}$$

Where min(x) is the minimum value of a signal; max(x) is the maximum value of the signal.

3.5 Data Segmentation

Some human activity patterns normally appear on the short time scales so a window size of less than 5 seconds is commonly used [16, 18, 42]. In this research, the window size is defined in 1 second with the window shift of 0.25 seconds. Each 1 second window is composed of all chest and wrist bio-signals (shown in Table 3.1). The sampling rate means the number of points that are collected in 1 second. For example, all chest-based bio-signals were sampled at 700 HZ so there are 700 points with 8 signals (ACC contains 3 axes) in a window. As for wrist-based bio-signals, the sampling rates are BVP 64 HZ, ACC 32 HZ, and EDA & Temp 4 HZ, which shows that the number of points is different in 1 second window. Additionally, the number of windows for each subject is attached in Appendix A

Signal	Sampling Rate	Number of Points (1 sec)
All Chest Signals	700 HZ	700
BVP	64 HZ	64
ACC	32 HZ	32
Temp and EDA	4 HZ	4

Table 3.1: Number of points in 60 seconds windows

3.6 Chapter Summary

Chapter 3 has described the data preprocessing methods in terms of extracting the target records for the classification task, outliers processing to improve the model generalisability, signal denoising to improve the data fidelity, data normalisation to build the common scale for each signal, and data segmentation for model input shape. After processing the signals, exploratory data analysis, which is the first phase of explainability methods, will be discussed in the Chapter 4 to compare some features from different modalities in the baseline, stress, and amusement state.

Chapter 4

Exploratory Data Analysis

This Chapter introduces exploratory data analysis to compare the differences in various affect states. In statistics, exploratory data analysis is a way to analyse the dataset and conclude the characteristics by visualisation. The goals of exploratory data analysis are to explore the intuitive patterns from the data, check the assumptions, determine the relationship between each signal, and estimate the relation between explanatory and outcome variables. Normally, the exploratory analysis does not include formal statistical testing and inference analysis. Thus, the signal plots and specific feature values are presented for observing the patterns.

4.1 Explanatory Features

The explanatory features are used to represent the characteristics or patterns from the processed data [42, 43]. The 10 features selected from [42] are adopted to analyse the signals patterns in this research.

- *Mean* —The mean value is a measure of central tendency [39], which is equal to the sum of all the values in the dataset divided by the number of values in the dataset (shown in Equation 4.1).

$$\overline{X}_{signal} = \frac{1}{N}\sum_{n=1}^{N} X_n \tag{4.1}$$

- *Median* —The median is also the measure of central tendency which can be found in the middle score for the dataset that has been arranged in order of magnitude [39]. Compared with the mean measurement, the advantage of the median is that it is influenced less by outliers in the data distribution.

- *Standard deviation* —The standard deviation is used to quantify the variation of a set of data values [39]. The way to calculate the standard deviation is shown in Equation 4.2.

$$\sigma_{signal} = \sqrt{\frac{\sum_{n=1}^{N}(X_n - \overline{X}_{signal})^2}{N}} \tag{4.2}$$

- *R Peak (RP)* —The maximum amplitude in the R wave (shown in 4.1) is called R peak amplitude or R peak [20].

- *RR Interval (RRI)* —The interval (shown in 4.1) derived from the two consecutive R wave [20].

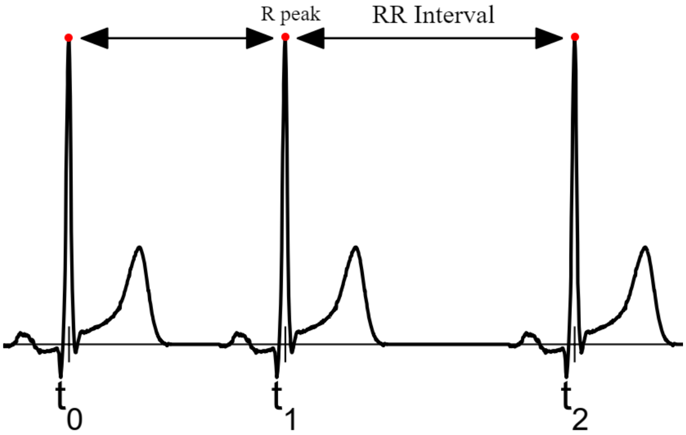


Figure 4.1: R peaks and RR intervals in ECG

- *Zero cross points* —This feature can be used to analyse the duration of respiratory [42].

- *Signal power* —Power is defined as the consumed energy in unit time. In the discrete signal domain, the calculation is shown as 4.3.

$$P_{signal} = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} |x(n)|^2 \tag{4.3}$$

- *Signal Correlation* —Signal cosine similarity is normally used to measure the correlation for homogeneous signals such as ACC X, Y, and Z or heterogeneous signals such as ECG and Resp [43]. The equation 4.4 is used to calculate the signal cosine correlation.

$$Cosine\ Similarity = \frac{A \cdot B}{||A|| \cdot ||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \tag{4.4}$$

- *Slope and Intercept* —These two features are obtained by fitting a linear regression in a defined window. Slope and intercept are used to analyse the trend and changes in the window over time [42].

In the following plots, subject 7 (S7), who has no additional note, is selected as a representative for comparing the feature differences between baseline, stress, and amusement in different signals. The X-axis represents the index of the point in the time series for each state, and the Y-axis represents the actual value for each signal at a certain timestamp.

4.2 Chest signal analysis

4.2.1 ECG

ECG signal presents a periodical pattern (shown as Figure 4.2), which consists of peaks and wave interval. Figure 4.2 shows that there are more R peaks in stress than the other two states which can deduce the mean of RR interval is smaller than the other two. Compared baseline with amusement state, there are more R peaks in baseline state than amusement state. Therefore, the number of R peaks per minute and mean of RR interval are selected to observe the ECG signal.

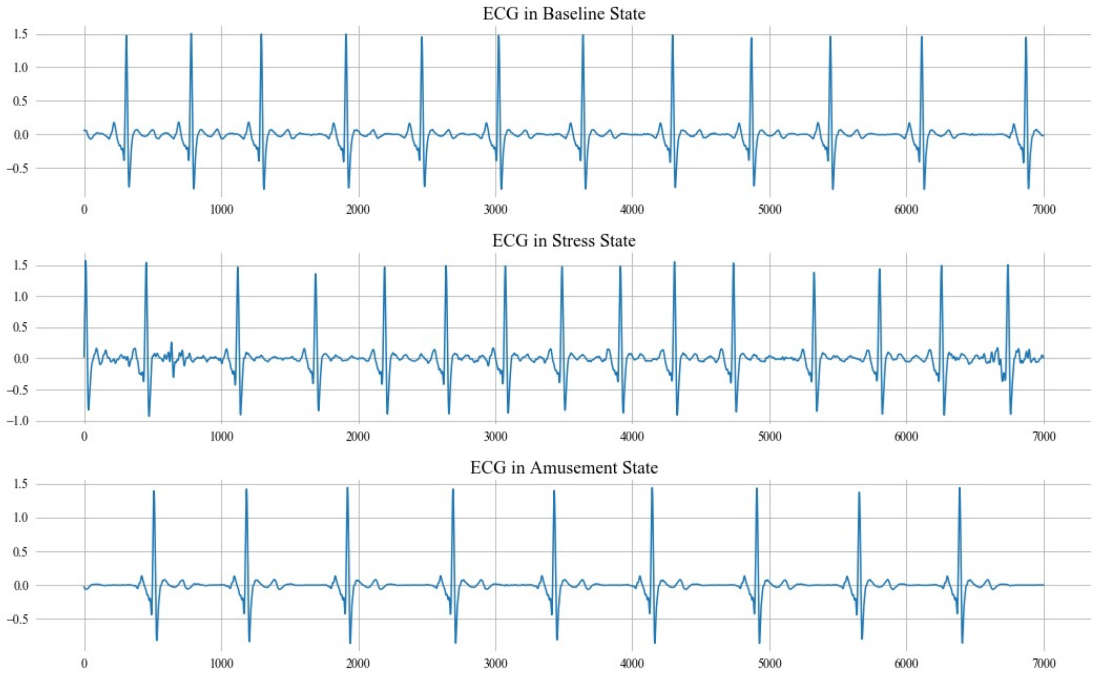


Figure 4.2: ECG signal Comparison

The number of R Peaks[1] and mean of RR interval for the S7's ECG signal features are shown in Table 4.1, and the stress state presents the most number of R peaks per minute.

4.2.2 Chest Temp

The body temperature is in a certain range for each state (as shown in Figure 4.3). Figure 4.3 shows that the temperature in the stress state is slightly higher than the

[1] The R Peak is calculated by `biospp_ecg`

Affective State	No. RPs (per mins)	Mean of RRI
Baseline	71	595
Stress	84	500
Amusement	58	720

Table 4.1: ECG explanatory features for subject 7

other two states. Comparing baseline with amusement state, the body temperature in the amusement could be higher than the baseline. Both baseline and amusement states present a stable trend. The trend can be analysed by fitting a linear regression on the Temp signal. Therefore, mean and linear regression parameters are selected for observation.

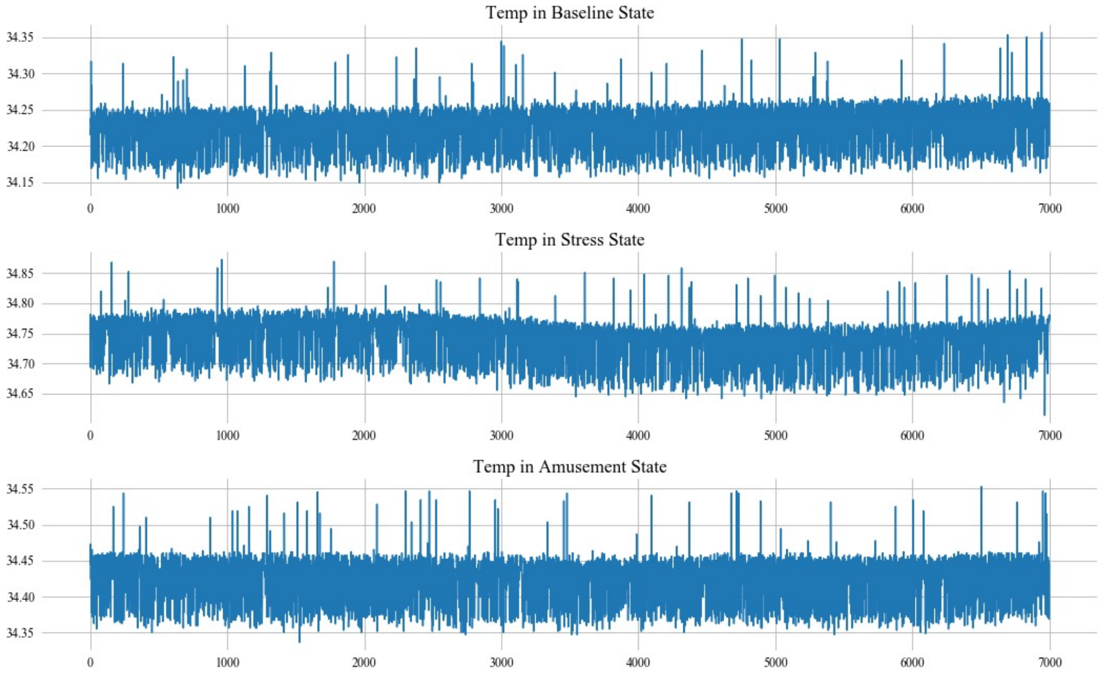


Figure 4.3: Chest Temp signal Comparison

Affective State	Mean of Temp	Slope	Intercept
Baseline	34.48	3.4×10^{-7}	34.34
Stress	34.91	5.8×10^{-7}	34.78
Amusement	34.44	-5.5×10^{-8}	34.44

Table 4.2: Temp explanatory features for subject 7

The mean of temperature[2], slope and intercept [3] for S7's Temp signal features are shown in Table 4.2, which shows that the slope, intercept and mean of temperature in stress are slightly higher than the other two states.

[2] The mean value is calculated by `numpy.mean`

[3] The slope and intercept are calculated by `sklearn.linear_model.LinearRegression`

4.2.3 Chest ACC

There are three lines in chest ACC signal plot representing the three dimensions of movements (shown as Figure 4.4), which displays that these lines keep fluctuating in baseline and stress state but are stable in amusement state. Therefore, the correlation for each axis in the different state is explored.

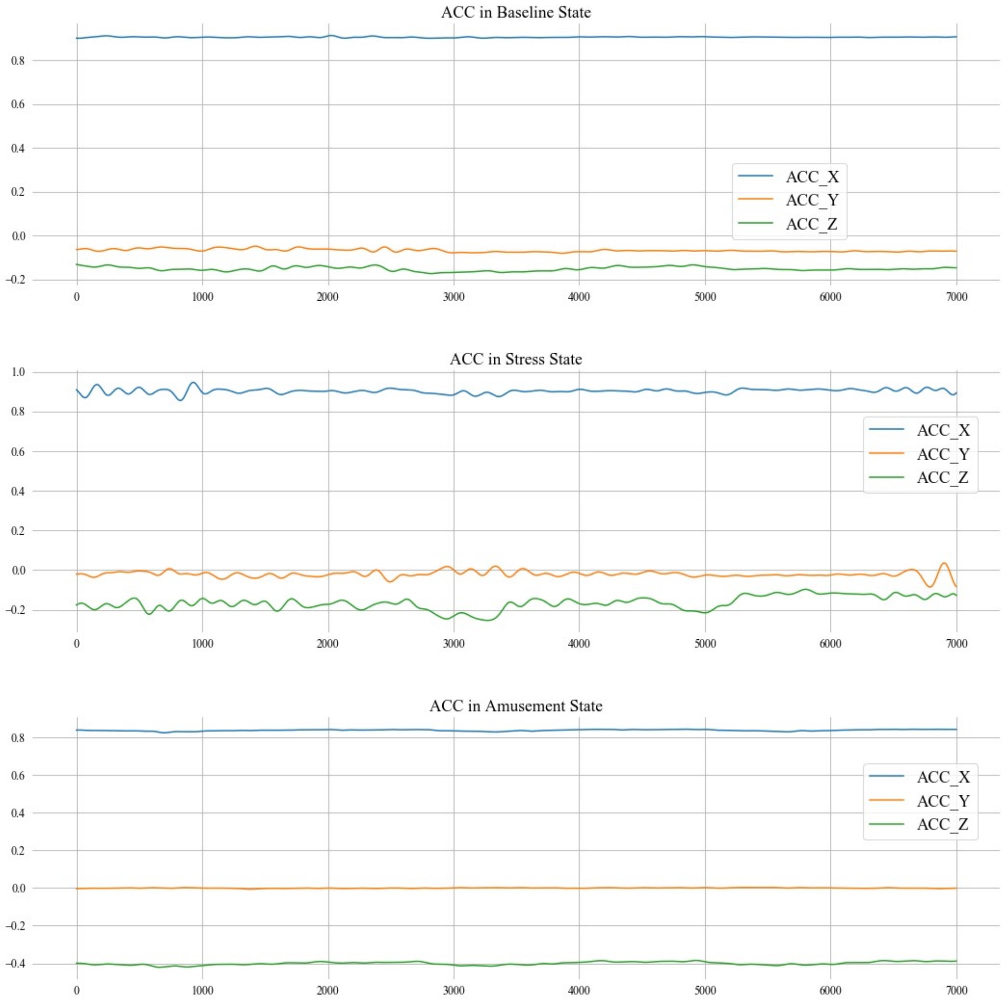


Figure 4.4: Chest ACC signal Comparison

Affective State	Cosine XY	Cosine XZ	Cosine YZ
Baseline	-0.91	-0.96	0.85
Stress	-0.86	-0.98	0.84
Amusement	-0.54	-0.99	0.52

Table 4.3: Chest ACC explanatory features for subject 7

The signal correlation[4] for the S7's ACC signal features are shown in Table 4.3, which presents that X and Z axis are highly correlated in each state.

[4] The correlation is calculated by `scipy.sptial.distance.cosine`

4.2.4 Resp

The respiratory presents a smoothing wave shape in a certain range for each state (as shown in Figure 4.5), which displays that the duration in stress state is longer than the other two states. In other words, subject 7 took breathing more frequent in baseline and amusement states than the stress state. In terms of signal power, the stress state might be higher than the other two. The number of zero cross points per 1 minute and signal power are selected to observe.

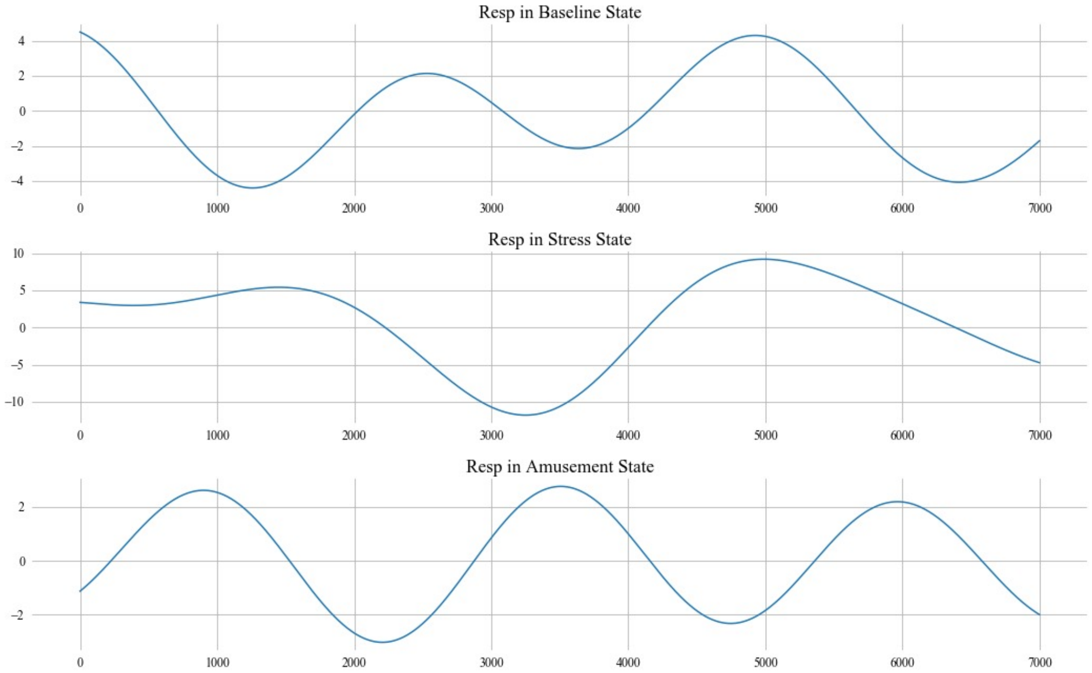


Figure 4.5: Respiratory signal Comparison

Affective State	No. Zero points (per mins)	Power
Baseline	32	2.74×10^6
Stress	24	6.18×10^6
Amusement	33	7.56×10^5

Table 4.4: Respiratory explanatory features for subject 7

The result (shown in Table 4.4) presents that The number of zero points[5] (per minutes) in stress state is lower than the other two. However, the Resp power in stress state is much higher than the other two. It indicates that the subject takes deep breathing in the stress state.

4.2.5 EMG

The EMG signal normally contains much noise which is hard to remove. Figure 4.6 presents that the amplitude fluctuated range in amusement state could be slightly

[5] The number of zero cross points is calculated by biosppy.resp

smaller than the other two states. Additionally, the median metric can be used to find the representative central value. Therefore, median and power are selected to measure and describe the EMG signal.

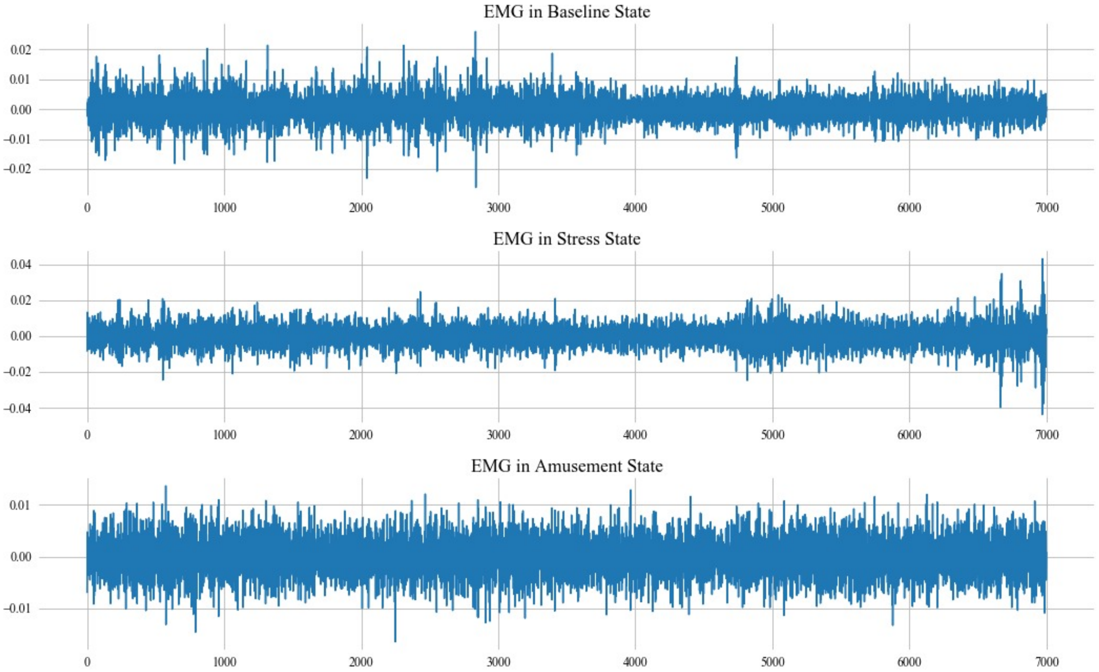


Figure 4.6: EMG signal Comparison

Affective State	Median	Power
Baseline	-1.69×10^{-5}	42.07
Stress	-7.19×10^{-6}	27.50
Amusement	6.56×10^{-5}	4.94

Table 4.5: EMG explanatory features for subject 7

The mean and power values for the S7's Resp signal features are shown in Table 4.5. The results present that the EMG's power in amusement state is lower than the other two, and the median feature in amusement state is the positive value.

4.2.6 Chest EDA

The EDA signal is a measure of sweat excreted by the human body, which is quite related to human affective states [43]. Figure 4.7 presents a periodical falling pattern of EDA for baseline and amusement state. However, the stress state displays a different trend compared with the other two states. Additionally, the variability of sampling points in the stress state seems more complex than the other two. Therefore, mean, standard deviation, slope and intercept are selected to observe the EDA signal.

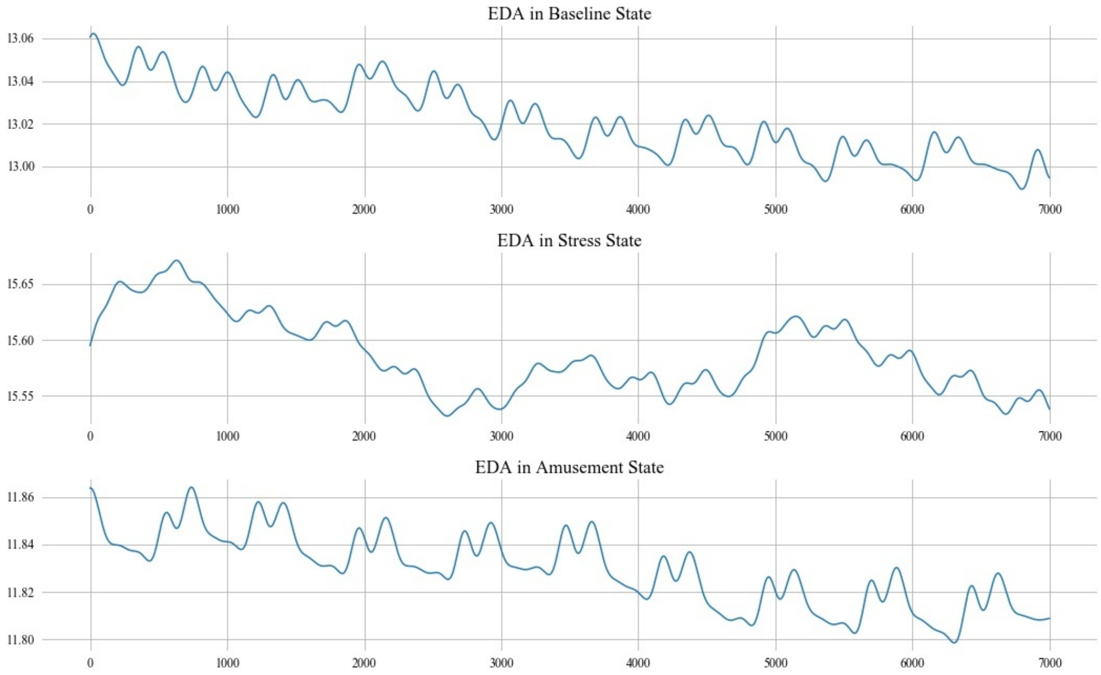


Figure 4.7: Chest EDA signal Comparison

Affective State	Mean	Std	Slope	intercept
Baseline	11.50	0.49	-1.66×10^{-6}	12.19
Stress	18.07	1.49	1.14×10^{-5}	15.51
Amusement	11.77	0.15	-1.72×10^{-6}	11.99

Table 4.6: EDA explanatory features for subject 7

The result (shown in Table 4.6) presents that the four explanatory features' values of EDA signal in stress state are higher than the other two which indicates that the skin electronic activity at the chest is more active in stress state than the others.

4.3 Wrist signal analysis

The sampling rate for wrist signals is different from the chest signals. The chest signal points are visualised in 7000 points (10 seconds) which can present a basic view of signal changes. However, the sampling points for wrist signals might not be enough to view the signal changes due to their lower sampling rate. Hence, the following plots deploy around 700 sampling points for each signal.

4.3.1 Wrist ACC

Based on the nature of ACC, the physical activity in the wrist is more active than the chest (shown as Figure 4.8). The wave for each axis changes frequently in baseline and stress state. However, these axes are quite stable in amusement state. Since the temporal changes for each axis may affect other axes, the correlation metric is also applied to wrist ACC.

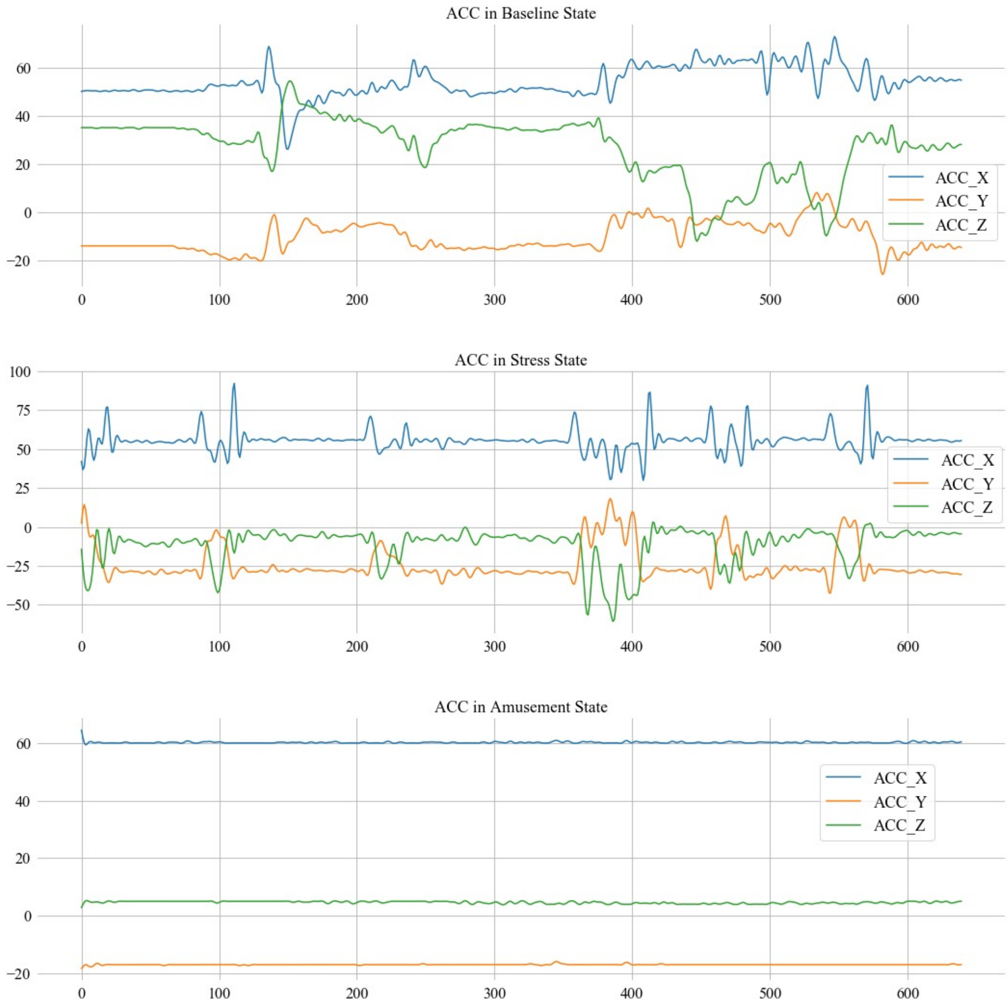


Figure 4.8: Wrist ACC signal Comparison

Affective State	Cosine XY	Cosine XZ	Cosine YZ
Baseline	-0.88	0.79	-0.78
Stress	-0.83	-0.56	0.28
Amusement	-0.99	-0.99	-0.99

Table 4.7: Wrist ACC explanatory features for subject 7

The signal correlation for the subject 7's wrist ACC signal features is shown in Table 4.7, which presents that X, Y and Z axis are highly correlated in amusement state. In the stress state, the cosine similarity between Y and X are lower than the other two features.

4.3.2 BVP

The BVP signal is related to the changes of the blood volume in vessels which normally can be measured by the hand or wrist sensor [43]. Figure 4.9 presents three different trend of lines. In terms of amusement state, the BVP might display a periodical wave over time. However, the baseline and stress state do not have clear patterns. The variability of BVP in the stress state seems more complex than the baseline state. Additionally, the amplitude range for each state are also quite different, which can be measured by power. Therefore, standard deviation and power are proposed to observe the difference between each state.

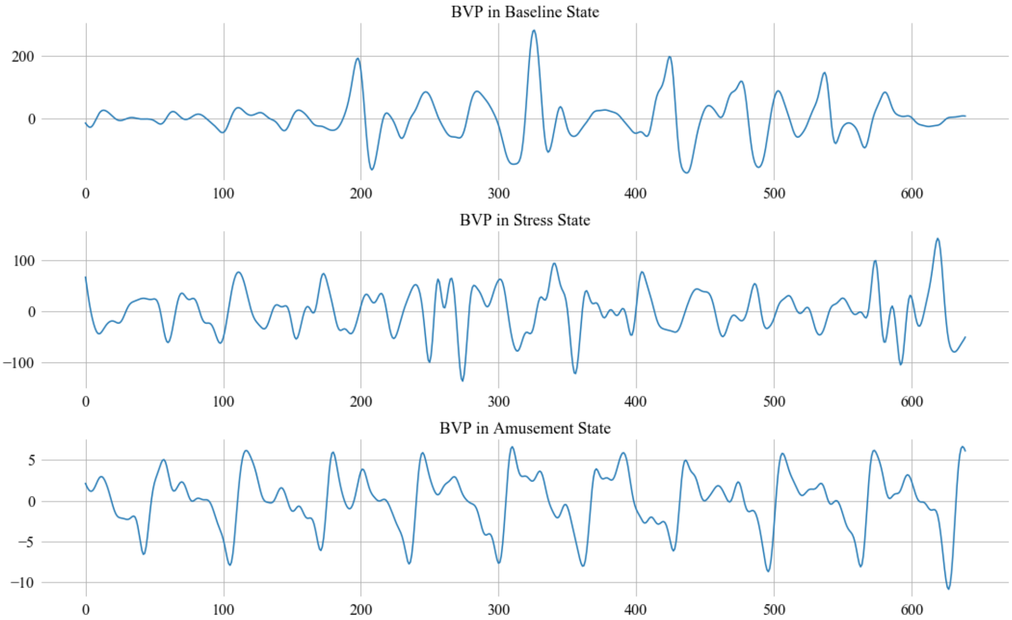


Figure 4.9: BVP signal Comparison

Affective State	Std	Power
Baseline	17.32	2.28×10^7
Stress	40.95	6.86×10^7
Amusement	4.30	4.39×10^5

Table 4.8: BVP explanatory features for subject 7

Table 4.8 presents that The standard deviation and power of BVP signal in stress state are higher than the other two, which indicates that the heart activity of subject 7 is more intense.

4.3.3 Wrist EDA

Compared with the chest EDA plot (shown as Figure 4.7), wrist EDA signal presents different patterns. The baseline and amusement state display a decreasing trend (shown as Figure 4.10). In contrast, the magnitude of wrist EDA signal seems to keep increasing in the stress state. It indicates that the excreted sweat from the human wrist in stress state is more than the other two which is corresponding to the stress response [43]. The differences between baseline and amusement are the magnitude and variability. Therefore, the same explanatory features used in chest EDA are also applied to wrist EDA.

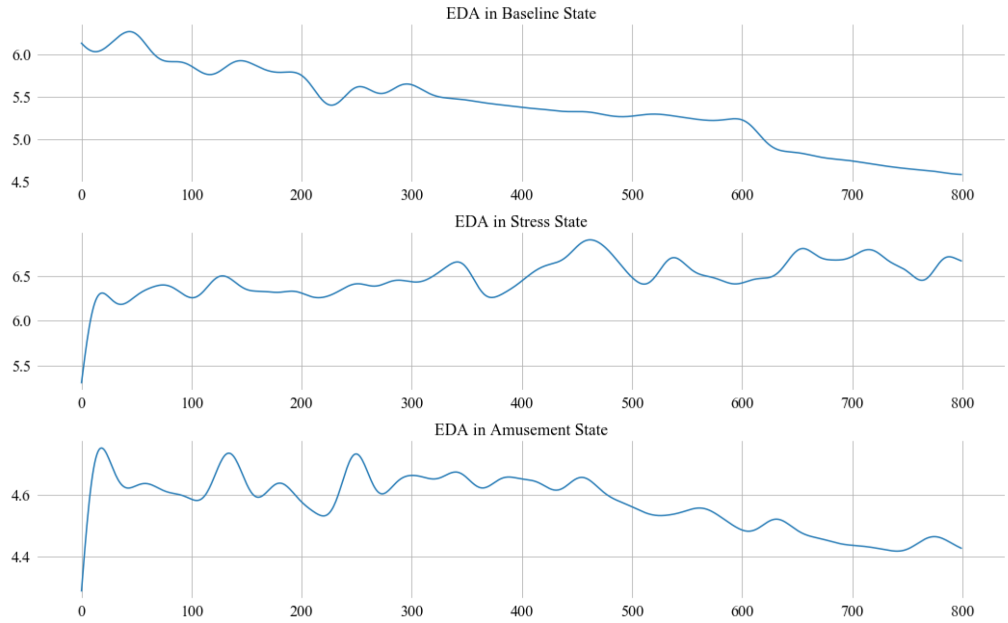


Figure 4.10: Wrist EDA signal Comparison

Affective State	Mean	Std	Slope	Intercept
Baseline	4.66	0.46	-2.88×10^{-4}	5.34
Stress	6.79	0.29	3.52×10^{-4}	6.34
Amusement	4.46	0.15	3.04×10^{-4}	4.69

Table 4.9: Wrist EDA explanatory features for subject 7

Table 4.9 presents that the values of mean, slope and intercept from wrist EDA signal in stress state are higher than the other two. It also indicates that the skin electronic activity at the wrist is more active in stress than the other two states.

4.3.4 Wrist Temp

Compared with the chest Temp plot (shown as Figure 4.3), the wrist Temp signal presents a more clear trend for each state (shown as Figure 4.11). The temperature in the baseline state keeps increasing generally. In terms of stress and amusement state, their lines are decreasing over time. Therefore, the mean, slope, and intercept are also used to describe wrist's Temp.

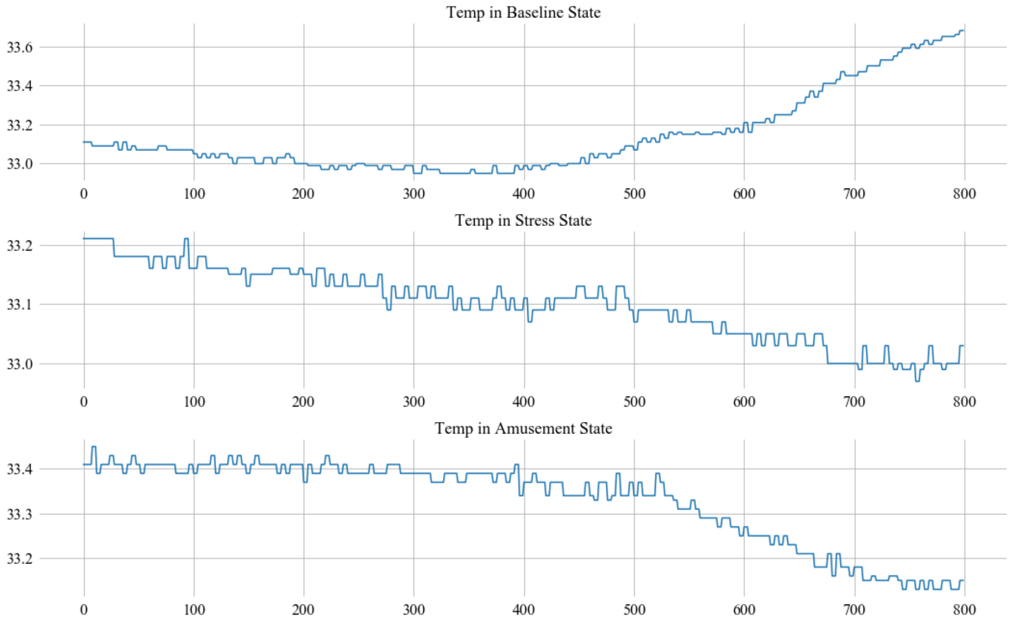


Figure 4.11: Wrist Temp signal Comparison

Affective State	Mean of Temp	Slope	Intercept
Baseline	33.96	2.14×10^{-4}	33.45
Stress	32.88	-2.66×10^{-4}	33.22
Amusement	33.23	-2.89×10^{-4}	33.44

Table 4.10: Wrist Temp explanatory features for subject 7

The results (shown in Table 4.10) presents that the slope, intercept and mean of temperature in stress are lower than the other two states. It indicates that the subject 7's wrist temperature is generally lower than the other two states and also decreasing stably over time.

4.4 Chapter Summary

This chapter has explored the trend and patterns for each signal in various states by exploratory data analysis. Some plots present the trend or periodical patterns which can distinguish the differences between various states intuitively. However, some signals such as EMG are hard to find patterns, which needs several explanatory features to quantify the differences. After analysing signals by their features respectively, the differences between various states are discussed. Then, the next chapter will introduce the model to classify these states.

Chapter 5

Models for Affect Detection

Existing approaches to affect recognition are either statistical analysis (ANOVA) or machine learning [43]. The focus in this research is to use machine learning model, in particular deep learning models, to recognise neutral, stress and amusement state. In this chapter, we will demonstrate the structure of the proposed deep learning fusion model. In particular, the deep learning models consists of various layers including Convolutional Layer, and Full Connected Layer. In Chapter 3, the data segmentation is described using 1 second window with shift sliding 0.25 seconds. The sampling rate for the chest sensor and wrist sensor is different which determines the variety of model's input shape. In terms of chest modalities, all the signals were sampled at 700 HZ. These signals can be fused at the data level as a matrix 700 × 8 for a window processed by a single stream deep learning model. In comparison, the wrist's signals need to be processed by a multi-stream deep learning model to solve the problem of incompatible input shape.

5.1 Model for Chest Sensor

Since all signals collected from chest were sampled at 700 HZ, these signals can be integrated at data level to feed a single stream deep learning model (shown as Figure 5.1). Two 1D convolutional layers are adopted to extract features and then the LSTM layer is added to process the time series information. Subsequently, a 1D convolutional layer is deployed again to refine the feature vector from LSTM output. The high-level features are flattened into a vector which is then fed to fully connected neural networks for classification. Since this is multi-classes classification tasks, Softmax is selected as the activation function in the output layer.

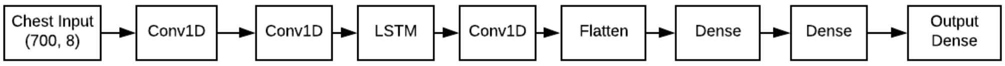


Figure 5.1: Chest Model

5.2 Model for Wrist Sensor

The wrist's signals were sampled at different rates (BVP, 64 HZ; Wrist ACC, 32 HZ; Wrist EDA & Temp, 4 HZ) so the model's input shape for these signals is also different. Therefore, these signals need to be fed into different submodels and then concatenate their feature vectors for the fully connected neural networks to classify results. The proposed model structure is adopted early fusion scheme which is shown in Figure 5.2, and the submodels' description is introduced as follows:

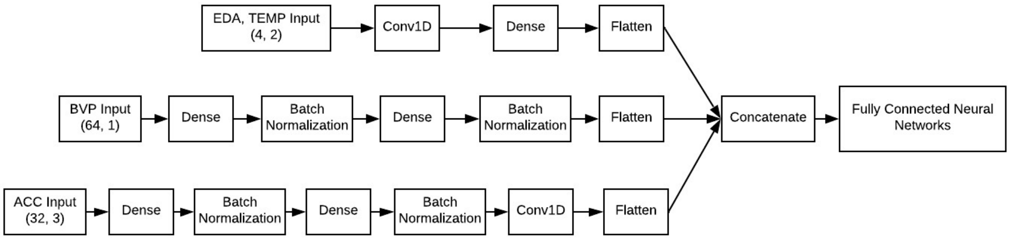


Figure 5.2: Wrist Model

Wrist EDA and Temp submodel —The input shape for EDA and Temp are 4 × 2, which does not have plenty of data points compared with other submodels in a batch. To avoid overfitting, this submodel is designed in a shallow structure. After tuning the hyper-parameters, the optimal architecture consists of one convolutional layer and one dense layer as shown in Figure 5.2.

BVP submodel —The input shape for BVP is 64 × 1 which has more data points in a batch than EDA and Temp. However, there is only one signal for BVP submodel. After tuning the hyper-parameters, the optimal architecture for BVP submodel consists of dense layers and batch normalisation layer alternately as shown in Figure 5.2.

Wrist ACC submodel —The input shape for ACC is 32 × 3. The number of sampling points for the ACC signal's input is plentiful compared with the other two parts, and there are three variables (X, Y, and Z movements) for every input matrix. After tuning the hyper-parameters, the architecture for ACC submodel is similar to the BVP submodel but there is one more 1D convolutional layer added to extract the features again.

After flattening the feature vectors from each stream, these vectors are then merged to be a super-vector which is fed to the fully connected neural networks for classification.

5.3 Main Fusion Models

After building the models for chest and wrist modalities, the next step is to explore the main fusion model combining these two streams. There are three basic fusion levels which have been introduced in section 2.4 including data-, feature-, and decision-levels. The chest signals are trained in data-level. The wrist signals can be trained using early fusion (feature level) or late fusion (decision level). To obtain an optimal main fusion model, early and late fusion are implemented in the experiment.

5.3.1 Early Fusion Scheme

Early fusion merges separated channel's features at the concatenated layer before classification [45]. As shown in Figure 5.3, the extracted features from each channel are concatenated horizontally into a joint representation. Then, the representation is fed to the fully connected neural network for classification [11, 45]. The model is compiled in the categorical cross-entropy loss function with Adam optimiser in the batch size 32 with 5 epochs.

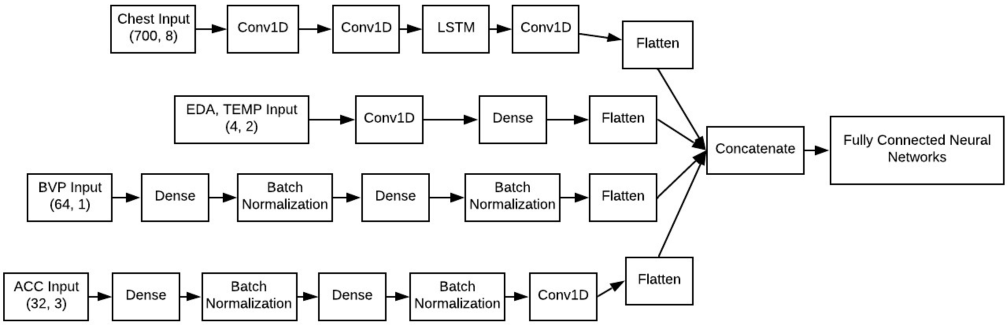


Figure 5.3: Early fusion Model

5.3.2 Late Fusion Scheme

Late fusion trains each submodel separately to obtain the classified results from the submodels, and these classified results are integrated into a joint representation to train the final classifier [45]. In contrast to early fusion, late fusion focuses on the strength of individual model [11]. As shown in Figure 5.4, there are four submodels with output dense layers which can generate the classified results. These results are concatenated to a joint representation which is then fed into the final supervised learner. The final supervised model can be any machine learning algorithms to classify the results. In this research, two supervised learners are adopted: Random Forest and Single-Layer Neural Network. All models in late fusion are compiled in the categorical cross-entropy loss function with Adam optimiser in the batch size 32 with 5 epochs.

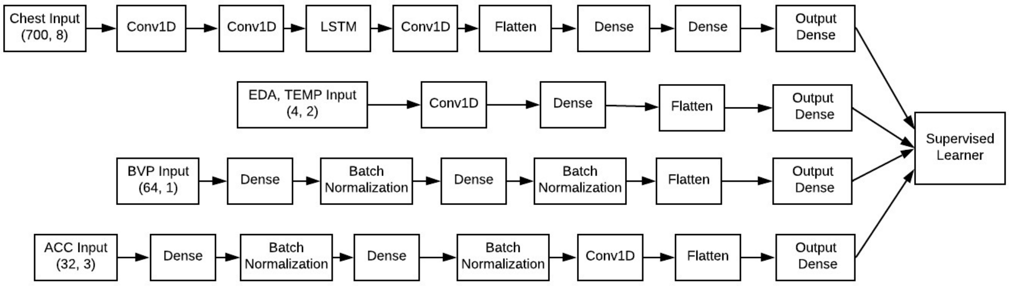


Figure 5.4: Late fusion Model

5.4 Chapter Summary

Chapter 5 has introduced the architecture for chest, wrist and main fusion models. The sampling rate for various signals in the chest channel is the same so a single stream deep learning model is adopted to process the classification task. On the other hand, the wrist signals have different sampling rate. In this research, early fusion is adopted in wrist's model structure for model training. Based on the main fusion model architecture (shown as Figure 5.3 and 5.4), early fusion focuses on feature-level integration which concatenates various features into one super-vector for classification. Compared with late fusion, early fusion scheme only has one learning phase. Thus, it trains faster than late fusion[45]. On the other hand, late fusion focuses on decision-level integration which integrates the classified results or probability from each stream and then classifies the affective states by the final supervised classier [11]. One advantage is that the late fusion scheme is easy to interpret the results from the final classifier [11]. Nevertheless, the main disadvantage of the late fusion approach is that every submodel requires a training process which is quite time-consuming [45]. Additionally, since each submodel is trained separately, late fusion scheme might lose the correlation in mixed feature space [45]. In conclusion, this chapter has demonstrated the model architectures and discussed the strength and weakness of two fusion schemes in the main models. The classification results are reported in Chapter 6 to compare the efficiency of the proposed models and fusion strategies.

Chapter 6

Results and Discussion

Shallow machine learning (SML) includes many algorithms (e.g., Logistic Regression, and Random Forest) to perform classification tasks. However, it heavily relies on pre-defined features which determine the upper bound of classification performance. The state of the art [42] utilised five SML models to detect the three mental states based on the WESAD dataset. Although they achieved competent results, the classification performance can be potentially improved by deep learning models. Several deep learning models have been introduced in Chapter 5, which are expected to learn the features automatically and provide better classification performance. The results by using *All Chest*, *All Wrist* and *All Modalities* are presented in detail in this chapter.

6.1 Evaluation Protocol

The metrics to evaluate model performance are classification accuracy and weighted F_1 score. The weighted F_1 score calculates the scores for each label, and then weight averages the scores to obtain the final score. The weighted F1score is commonly used for label imbalance datasets. Schmidt et al. [42] utilised the Leave-One-Subject-Out (LOSO) cross-validation as the strategy to evaluate their model's classification accuracy and weighted F_1 score. This research continues using the same strategy to measure the classification performance of multimodal deep fusion network models.

Section 3.2 introduced that the WESAD dataset contains 15 subjects in total. 7 subjects of them are set in the testing group and the other 8 subjects are in the training group. In the test group, one subject is selected as a test set, and then the remaining subjects in the test group are integrated with the training group to be a training set. Then, the training set is used to train the model, and the test set is used to evaluate model performance. In conclusion, the diagram of 1 fold LOSO cross-validation is shown in Figure 6.1.

Since there are 7 subjects in the testing group, the model is evaluated by 7 folds

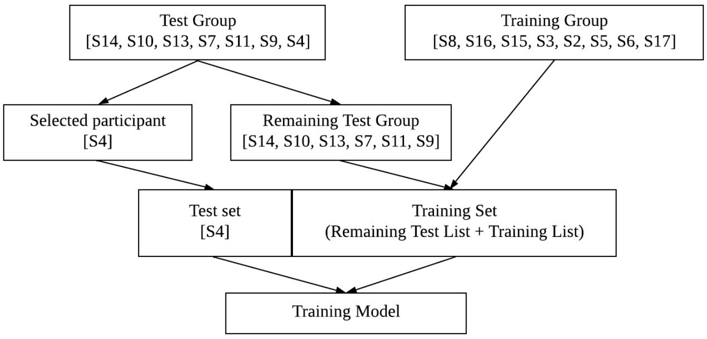

Test Group
[S14, S10, S13, S7, S11, S9, S4]
Training Group
[S8, S16, S15, S3, S2, S5, S6, S17]
Selected participant
[S4]
Remaining Test Group
[S14, S10, S13, S7, S11, S9]
Test set
[S4]
Training Set
(Remaining Test List + Training List)
Training Model

Figure 6.1: One fold LOSO evaluation

LOSO cross-validation for accuracy and weighted f-score. This research defines the 7 folds LOSO cross-validation as a round. To provide less biased results, this research implement 5 rounds (1 trial) 7 folds LOSO cross-validations as shown in Figure 6.2.

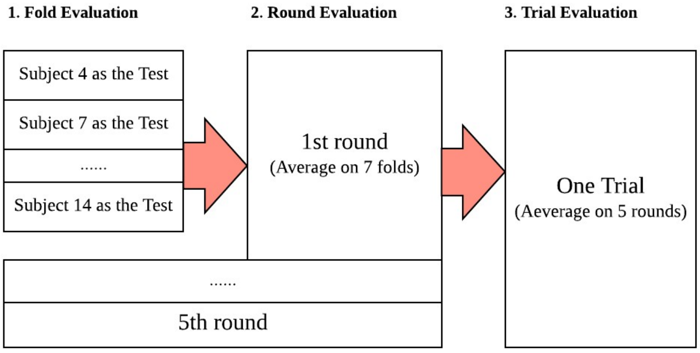


Figure 6.2: Experiment Protocol

6.2 Benchmark Results

Schmidt et al. [42] adopted five SML models: Decision Tree (DT), Random Forest (RF), AdaBoosting (AB), Linear Discriminant Analysis (LDA), and K-Nearest-Neighbour (KNN) to detect the affective states. They utilised 60 seconds as a window with 0.25 seconds as the shift sliding, and the features are extracted from the 60 seconds window to fed the models. Additionally, they defined several terms for evaluation, and this research select *All Chest*[1], *All Wrist*[2], and *All Modalities*[3] for comparison. The results for their work are listed as follows:

Benchmark Classification Accuracy from [42] is shown in Figure 6.3 (left). They obtained 0.77 in *All Chest* by LDA model, 0.75 in *All Wrist* by AB model, and 0.80 in *All Modalities* by AB model.

[1] contains ECG, EDA, EMG, RESP, TEMP, and ACC

[2] contains BVP, ACC, EDA, and TEMP

[3] contains all chest and wrist signals

Benchmark Weighted F_1 score from [42] is shown in Figure 6.3 (right). They obtained 0.73 in *All Chest* by LDA model, 0.64 in *All Wrist* by AB model, and 0.72 in *All Modalities* by LDA model.

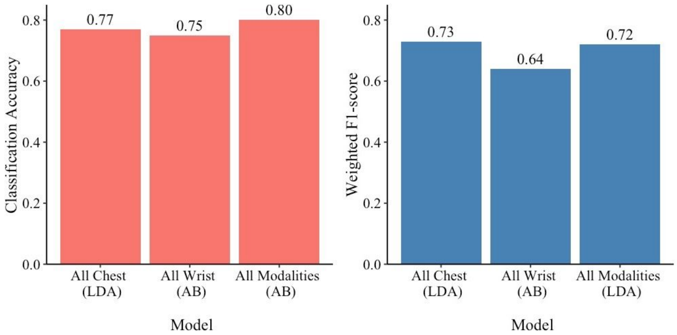


Figure 6.3: Three-class classification accuracy results (left) and weighted F_1 score (right)

6.3 Deep Fusion Network Results

In contrast, this research proposes four deep learning models: Single stream model for *All Chest*, Early Fusion (EF) for *All Modalities*, Early Fusion for *All Wrist*, Late Fusion with Random Forest (LF-RF) for *All Modalities*, and Late Fusion with Neural Network (LF-NN) for *All Modalities*. All deep learning models process the data in 1 second window size with 0.25 seconds sliding stride. The results for classification accuracy and F_1 score are listed as follows:

Classification Accuracy for deep learning models is shown in Figure 6.4 (left). It presents 0.83 in *All Chest* by Single stream model, 0.85 in *All Modalities* by LF-RF, and 0.83 in in *All Modalities* by LF-NN.

Weighted F_1 score for deep learning models is shown in Figure 6.4 (right). It presents 0.81 in *All Chest* by Single stream model, 0.86 in *All Modalities* by LF-RF, and 0.82 in *All Modalities* by LF-NN.

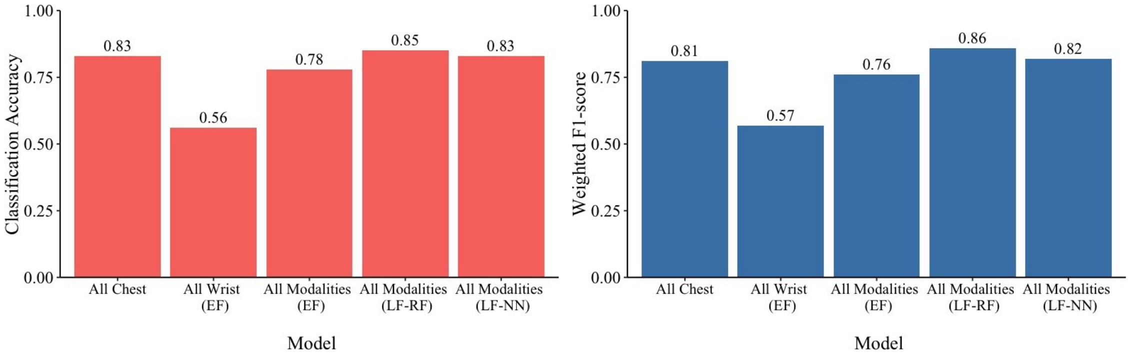


Figure 6.4: Results of Deep Learning model in classification accuracy (left) and weighted F_1 score (right)

6.4 Discussion

There are five models proposed in this research, and their performance is demonstrated in this chapter. The single stream deep learning model (or submodel 1) is trained by 8 layers deep learning architectures based on all chest bio-signals. As shown in Table 6.1, there are 3 convolutional layers to extract the feature vectors from the chest modalities, and 1 LSTM layer to process the time series information. Deep learning models avoid the step of feature engineering, which can be conducted by the convolutional layer automatically. The 4th convolutional layer can extract high-level features to feed the fully connected neural networks for classification so the single stream deep learning model has achieved better classification performance than the state of the art.

Layer	Chest Sensor		Wrist Sensor	
	submodel 1	submodel 2	submodel 3	submodel 4
1	Conv1D	Conv1D	Dense	Dense
2	Conv1D	Dense	BN	BN
3	LSTM	Flatten	Dense	Dense
4	Conv1D	Dense	BN	BN
5	Flatten	-	Flatten	Conv1D
6	Dense	-	Dense	Flatten
7	Dense	-	-	Dense
8	Dense	-	-	-

Table 6.1: Submodels of Late Fusion Models
Abbreviation: BN (Batch Normalization), Conv1D (One dimensional convolutional layer), LSTM(Long short term memory).

In terms of fusion models on all modalities, two late fusion models (LF-NN and LF-RF) have higher classification accuracy than the state of the art in two measurements. These two late fusion models have the same submodel architecture as shown in Table 6.1, but are different at the final supervised classifier. The result presents that the model performance of LF-RF is better than LF-NN. Thus, LF-RF is mainly discussed in this section. Compared with the state of the art [42], the performance of LF-RF and submodel 1 is shown in Table 6.2.

Method	Modalities	Accuracy	Weighted F_1 score
Adaboosting [42]	All Chest	0.80	0.73
LDA [42]	All Chest	0.79	0.74
LF-RF submodel 1	All Chest	**0.83**	**0.81**
Adaboosting [42]	All Modalities	0.80	0.69
LDA [42]	All Modalities	0.76	0.72
LF-RF	All Modalities	**0.85**	**0.86**

Table 6.2: Classification Performance Comparison

As for early fusion, it has faster training stage than the late fusion and can capture the correlation from various modalities which is applicable to deep learning training. However, the performance of early fusion model in *All Wrist* and *All Modalities* is lower than the benchmark. A reason for this result is because the sampling rates of various modalities are quite different, which might not provide expected results in early fusion [34]. The late fusion focuses more on the individual model's strength which can be applied to broader learning tasks since it does not suffer the problems from early fusion [34]. Thus, the late fusion scheme performs better than early fusion on all modalities.

6.5 Chapter Summary

This chapter has introduced the evaluation protocol to obtain less biased results in classification accuracy and F_1 score. The results are compared between shallow machine learning and deep learning models, and the proposed deep learning models have achieved better classification performance on *All Chest* and *All Modalities* than the state of the art. The multimodal fusion schemes are also discussed in this Chapter, and the late fusion scheme has better performance on research task than the early fusion. As described in Section 2.5, the main challenge is to both improve the classification accuracy and explainability. The next chapter will introduce the explainability in the phase of *Building the models*, and *After building the models*.

Chapter 7

Explainability in Affective Recognition

Explainability is considered as a new challenge when deep learning models become popular since many research fields need not only effective models and but also the reasons to explain the classified results [7]. However, the general high accuracy models normally have low explainability and vice versa. To improve the explainability of high-performance models, three explainable phases are introduced by [7] which are *Before building the models*, *Building the models*, and *After building the models* to find the explanation for model classification results. Since the *Late Fusion with Random Forest (LF-RF)* model achieves the highest classification accuracy and weighted F_1 score, this Chapter will discuss the explainable approaches for this model in details.

7.1 Before building the models

Before building the models, the signals' patterns and differences can be identified by exploratory data analysis. The details have been discussed in Chapter 4. All signals from chest and wrist sensors were visualised in a certain period of time, and the intuitive patterns and trends for each state have been illustrated. Additionally, the values of explanatory metrics have also been summarised as tables to explain the differences in various states. Both plots and explanatory variables' results have explained the differences for a subject's physiological responses in various affective states.

7.2 Building the models

In *Building the Models* phase, the sensor-level and wrist signal-level importance are discussed. The model architecture for each stream has been introduced in Chapter

5. This Chapter mainly focuses on the explainability in *Late Fusion with Random Forest (LF-RF)* model and model structure is presented in Figure 7.1. There are four sub-models in LF-RF model. Each sub-model firstly generate the classified scores, and then these scores are set as the input for training the Random Forest model. The weights (i.e., W_i) for each stream classified results are generated after the training process. These weight scores can explain the sensor's or submodel's importance of the classification results.

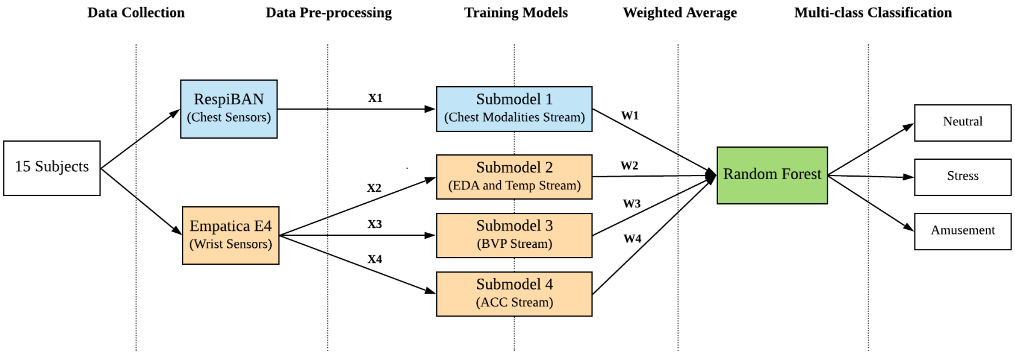


Figure 7.1: Late Fusion with Random Forest (LF-RF) Model

Submodel	Chest Modalities	BVP	ACC (Wrist)	EDA and Temp (Wrist)
Weight Scores	0.595	0.122	0.281	0.001

Table 7.1: Sensor-level Importance

As shown in Table 7.1, the chest and wrist models provide the around 60% and 40% contribution respectively to the correct classification results, which indicates that the modalities and submodel from RespiBAN sensor have higher importance than the Empatica E4 sensor. In the wrist model, the BVP stream takes around 12.2% and the ACC provides 28.1% contribution. However, the wrist EDA and Temp stream contribute 0.1% for the classification which has trivial influences on classification. Since there only 4 data points for EDA or Temp in a window, the data distribution are quite sparse, which leads to the low weight score. In conclusion, the weight scores for the important signals from wrist modalities have been displayed. However, the signal importance from chest modalities is still unknown. To understand the signal-level importance of chest modalities, the next section will introduce the *After Building the Model* approach for each chest signal.

7.3 After building the models

The ablation experiment is an approach in *After Building the Models* phase. The signal-level importance is conducted by the ablation experiment by removing a feature and assessing the effect with the remaining features on the model performance, which can explore how the removal feature affects the model performance [29]. The ablation experiment is an approach from *After building the models* phase, and the algorithm is shown in Algorithm 1.

Algorithm 1 Ablation Experiment

```
Input: A set of Bio-signals
Output: A list of the accuracy scores after feature ablation
ablist = an empty list
n = number of signals
while $n > 0$ do
    abFeat = the type of signal at position n
    remove the abFeat from the set of bio-signals
    p = classification results from training the remaining signals
    save the tuple (name, p) to the ablist
    $n = n - 1$
    put the abFeat signal back to the set of bio-signal
return ablist
```

After removing a certain type of signal out of the full modality set, the model training is also examined five rounds for each ablation to obtain less biased results that are displayed in Table 7.2.

Removed signal	Accuracy	Weighted F_1 score
ACC	0.71	0.68
ECG	0.78	0.76
EMG	0.81	0.78
EDA	**0.65**	**0.63**
Temp	0.73	0.70
Resp	0.82	0.79
All Chest Signals	0.83	0.81

Table 7.2: Chest Signal Ablation Experiment

After removing the EDA signal, the classification accuracy and weighted F_1 score present significant decrease (accuracy from 0.83 to 0.65, and the weighted F_1 score from 0.81 to 0.63). Additionally, removing ACC or Temp also has around 10% deduction. In terms of Resp and EMG signal, the model does not have significant decreasing. Considering EMG, Resp and ACC might be correlated to each other due to human respiratory activity, an additional experiment is conducted which is the same as the ablation experiment but removing a signal combination. The result is shown in Table 7.3, which presents that after removing EMG and Resp, there is still no significant decreasing. Nevertheless, when EMG, Resp, and ACC signals are removed, the classification accuracy and F_1 score are closed to the result of only removing the ACC signal.

Removed signal combination	Classification Accuracy	Weighted F_1 score
EMG and Resp	0.81	0.77
EMG and ACC	0.72	0.68
Resp and ACC	0.72	0.69
EMG, Resp, and ACC	0.72	0.69
All Chest Signals	0.83	0.81

Table 7.3: Addtional experiments on removing set of signals

7.4 Chapter Summary

In Chapter 4, the *Before Building the Models* phase has been discussed by exploratory data analysis to compare the difference between various affective state. In this Chapter, the phase of *Building the Models* and *After Building the Models* are mainly discussed. Since the random forest is adopted as the final supervised model in the LF-RF model, the weight scores for various submodels can be calculated, and the calculated results present that the chest stream takes around 60% contribution for classification, and the other 40% is from wrist model. The signal-level importance for wrist modalities can also be explained due to the LF-RF structure, and the main contributions from wrist's streams are BVP (12.2%) and ACC (28.1%). On the other hand, the sampling rate for all chest modalities are 700 HZ and these modalities are trained in one stream deep learning model. To explain the signal-level importance for chest modalities, the ablation experiment is applied to assess each signal's influence. The result indicates that removing EDA, ACC, Temp signals caused a significant decrease in classification accuracy and weighted F_1 score. Moreover, the additional experiments are also implemented to check the importance for EMG and Resp signals which do not display obvious decreasing.

Chapter 8

Conclusion

This research explores a number of problems in affect recognition by using deep learning models on multimodal bio-signals data, which includes the design of model's architecture, selection of multimodal fusion schemes, and explainability of deep learning models. The proposed models have obtained higher classification accuracy and F_1 score than the state of the art [42] in *All Chest* and *All Modalities* experiments. In all modalities analysis, the *Late Fusion with Random Forest* (LF-RF) model has obtained the best results than other models such as early fusion model in this research. However, high classification performance models normally have low explainability, which has discouraged researchers to use deep learning to affect recognition. Due to the structure of LF-RF, this model can provide the stream weight scores for the sensor-level importance and the wrist signal-level importance. In terms of chest stream, all bio-signals are trained in an 8 layers deep learning model. To explain the chest signal-level importance, the ablation experiment is conducted. Therefore, the achievements of this research are not only to provide a more accurate model than the state of art, but also to improve the explainability for the classified results.

8.1 Contributions

The literature review in this thesis provides related backgrounds and works to give an overview of the field research progress. The main contributions of this research are as follows:

8.1.1 Novel model in affective detection

Chapter 5 describes a novel application of deep learning algorithms to bio-signals for three affective states detection. For these models, convolution, LSTM, and dense layers have been adopted to obtain competent results. Experiments demonstrate that the deep learning model captures a high-level correlation underlying the data, and achieves higher classification accuracy and F_1 score than the state of the art from [42].

8.1.2 Multimodal-multisensory fusion schemes

The application of multimodal fusion to process multimodal multi-sensor bio-signals are described in chapter 5. The early fusion focuses on the feature-level which presents various features concatenated into one super-vector for classification. Compared with late fusion, the early fusion scheme is only required for one learning phase which trains faster than late fusion. However, the classification accuracy and F_1 score are lower than late fusion model in this research. On the other hand, late fusion can analyse and interpret the results after training. Moreover, its classification performance is better than the early fusion scheme. Nevertheless, one disadvantage of the late fusion approach is that every stream requires a standalone supervised learning training process which is quite time-consuming.

8.1.3 Explainability for deep learning in affect detection

The explainability or interpretability for this research has been demonstrated in three different methods *Before building the models*, *Building the models*, and *After building the models*. Chapter 4 has demonstrated the *Before building the models* analysis for every signal in different affective states. This method visualises the general patterns and trends intuitively from the data to find the differences. In terms of the *Building the models* method, Chapter 7 presents the weights for the classified results from every stream sub-model. It shows that the chest sensor contributes more to the correct classified results than the wrist sensor. Chapter 7 describes the ablation experiment for analysing chest modalities to demonstrate the method of *After building the models*, and the results show that the EDA, ACC, Temp, and ECG signals have significant importance.

8.1.4 Publication Plan

The works of multimodal deep fusion network architecture, model performance, and sensor- and signal-level explainability have been submitted to the *ACM International Conference on Information and Knowledge Management (CIKM)*, 2019.

The three phases' explainable results will be extended for submission to the journal of *IEEE Transactions on Affective Computing.*

8.2 Limitations

The limitations of this research are listed as follows:

1. *Instrument Bias* —There are many subjects' data collected in unexpected conditions (shown in Appendix B) such as not fully attached the sensor.

2. *Selection Bias* —All subject are from the European group, and the range of age is from 25-35 years. The training models are limited since the biological signals could vary due to ethnic and age differences.

3. *Inadequate sample size* —The sample size is 15 subjects, which might not be enough for early fusion to achieve better classification performance.

4. *Personalisation* —Since the main fusion model is trained to detect the general pattern, the classification accuracy might be lower for some special cases.

The first three limitations are from the WESAD dataset [42], which cannot be controlled in this research. The fourth limitation is about personalisation which will be explored in future works.

8.3 Future Work

8.3.1 Transfer Learning

The fusion model has been trained on the chest and wrist bio-signals to detect three affective states. When additional types of affects are required to be recognised, the straightforward way is to collect the bio-signals again and label the other affects which are quite expensive and time-consuming. Transfer learning is an emerging field to enable the model adapting to the new cases. After training a model on a task, the model can be transferred to the other similar datasets to learn the patterns, which can reduce the workload in data collection and model training.

8.3.2 Other Multimodal Fusion Schemes

This research has only investigated the multimodal fusion schemes on the early fusion and late fusion. However, it is possible to obtain better classification results by optimising the weighted loss function. The loss function can be concatenated in a weighted way to minimise the training loss together from all the submodel. Additionally, the early fusion scheme can be implemented to concatenate the feature vectors with the weights from submodels, which is also possible to improve the model's classification performance.

References

[1] Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., and Kankanhalli, M. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18* (Montreal QC, Canada, 2018), ACM Press, pp. 1–18.

[2] Acharya, U. R., Fujita, H., Oh, S. L., Hagiwara, Y., Tan, J. H., Adam, M., and Tan, R. S. Deep convolutional neural network for the automated diagnosis of congestive heart failure using ECG signals. *Applied Intelligence 49*, 1 (Jan. 2019), 16–27.

[3] Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., Adam, M., Gertych, A., and Tan, R. S. A deep convolutional neural network model to classify heartbeats. *Computers in Biology and Medicine 89* (Oct. 2017), 389–396.

[4] Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., and Adeli, H. Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. *Computers in Biology and Medicine 100* (Sept. 2018), 270–278.

[5] Baltrušaitis, T., Ahuja, C., and Morency, L. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence 41*, 2 (Feb 2019), 423–443.

[6] Bashivan, P., Rish, I., Yeasin, M., and Codella, N. Learning Representations from EEG with Deep Recurrent-Convolutional Neural Networks. *arXiv:1511.06448 [cs]* (Nov. 2015). arXiv: 1511.06448.

[7] Been, K. Introduction to Interpretable Machine Learning. In *Proceedings of the Tutorial on Interpretable Machine Learning for Computer Vision* (Salt Lake City, Utah, United States, 2018), Conference on Computer Vision and Pattern Recognition.

[8] Belo, D., Rodrigues, J., Vaz, J. R., Pezarat-Correia, P., and Gamboa, H. Biosignals learning and synthesis using deep neural networks. *BioMedical Engineering OnLine 16*, 1 (Dec. 2017).

[9] Carreiras, C., Alves, A. P., Lourenço, A., Canento, F., Silva, H., Fred, A., et al. BioSPPy: Biosignal processing in Python, 2015.

[10] Chao, L., Tao, J., Yang, M., Li, Y., and Wen, Z. Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge* (New York, NY, USA, 2015), AVEC '15, ACM, pp. 65–72.

[11] Dong, Y., Gao, S., Tao, K., Liu, J., and Wang, H. Performance evaluation of early and late fusion methods for generic semantics indexing. *Pattern Analysis and Applications 17*, 1 (Feb. 2014), 37–50.

[12] Faust, O., Hagiwara, Y., Hong, T. J., Lih, O. S., and Acharya, U. R. Deep learning for healthcare applications based on physiological signals: A review. *Computer Methods and Programs in Biomedicine 161* (July 2018), 1–13.

[13] Fernandez, A., Herrera, F., Cordon, O., Jose del Jesus, M., and Marcelloni, F. Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to? *IEEE Computational Intelligence Magazine 14*, 1 (Feb 2019), 69–81.

[14] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys 51*, 5 (Aug. 2018), 1–42.

[15] Hammerla, N. Y., Halloran, S., and Plötz, T. Deep, convolutional, and recurrent models for human activity recognition using wearables. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (2016), IJCAI'16, AAAI Press, pp. 1533–1540.

[16] Healey, J., Nachman, L., Subramanian, S., Shahabdeen, J., and Morris, M. Out of the Lab and into the Fray: Towards Modeling Emotion in Everyday Life. In *Pervasive Computing* (Berlin, Heidelberg, 2010), P. Floréen, A. Krüger, and M. Spasojevic, Eds., Springer Berlin Heidelberg, pp. 156–173.

[17] Healey, J. A., and Picard, R. W. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems 6*, 2 (June 2005), 156–166.

[18] Huynh, T., and Schiele, B. Analyzing features for activity recognition. In *Proceedings of the 2005 Joint Conference on Smart Objects and Ambient Intelligence: Innovative Context-aware Services: Usages and Technologies* (New York, NY, USA, 2005), sOc-EUSAI '05, ACM, pp. 159–163.

[19] Jiang, H., Kim, B., Guan, M., and Gupta, M. To Trust Or Not To Trust A Classifier. In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 5541–5552.

[20] Karpagachelvi, S., Arthanari, M., and Sivakumar, M. Ecg feature extraction techniques-a survey approach. *arXiv preprint arXiv:1005.0957* (2010).

[21] Katsis, C., Katertsidis, N., Ganiatsas, G., and Fotiadis, D. Toward Emotion Recognition in Car-Racing Drivers: A Biosignal Processing Approach. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 38*, 3 (May 2008), 502–512.

[22] Khorrami, P., Le Paine, T., Brady, K., Dagli, C., and Huang, T. S. How deep neural networks can improve emotion recognition on video data. In *2016 IEEE International Conference on Image Processing (ICIP)* (Sep. 2016), pp. 619–623.

[23] Kiranyaz, S., Ince, T., and Gabbouj, M. Real-Time Patient-Specific ECG Classification by 1-D Convolutional Neural Networks. *IEEE Transactions on Biomedical Engineering 63*, 3 (Mar. 2016), 664–675.

[24] LeCun, Y., and Bengio, Y. The handbook of brain theory and neural networks. MIT Press, Cambridge, MA, USA, 1998, ch. Convolutional Networks for Images, Speech, and Time Series, pp. 255–258.

[25] LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature 521* (May 2015), 436.

[26] Lin, J., and Liu, Y. A Neural Network Based Explainable Recommender System. *arXiv:1812.11740 [cs]* (Dec. 2018). arXiv: 1812.11740.

[27] Lipton, Z. C. In machine learning, the concept of interpretability is both important and slippery. *machine learning* (2018), 28.

[28] Lisetti, C. L. e., and Nasoz, F. Using Noninvasive Wearable Computers to Recognize Human Emotions from Physiological Signals. *EURASIP Journal on Advances in Signal Processing 2004*, 11 (Sept. 2004), 929414.

[29] Litkowski, K. Feature Ablation for Preposition Disambiguation. *Technical Report* (2008), 15.

[30] Lu, H., Frauendorfer, D., Rabbi, M., Mast, M. S., Chittaranjan, G. T., Campbell, A. T., Gatica-Perez, D., and Choudhury, T. StressSense: detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing - UbiComp '12* (Pittsburgh, Pennsylvania, 2012), ACM Press, p. 351.

[31] Mahdiani, S., Jeyhani, V., Peltokangas, M., and Vehkaoja, A. Is 50 Hz high enough ECG sampling frequency for accurate HRV analysis? In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Milan, Aug. 2015), IEEE, pp. 5948–5951.

[32] Münzner, S., Schmidt, P., Reiss, A., Hanselmann, M., Stiefelhagen, R., and Dürichen, R. Cnn-based sensor fusion techniques for multimodal human activity recognition. In *Proceedings of the 2017 ACM International Symposium on Wearable Computers* (New York, NY, USA, 2017), ISWC '17, ACM, pp. 158–165.

[33] Oviatt, S., Grafsgaard, J., Chen, L., and Ochoa, X. The handbook of multimodal-multisensor interfaces. Association for Computing Machinery and Morgan Claypool, New York, NY, USA, 2019, ch. Multimodal Learning Analytics: Assessing Learners' Mental State During the Process of Learning, pp. 331–374.

[34] Oviatt, S., Schuller, B., Cohen, P. R., Sonntag, D., Potamianos, G., and Krüger, A., Eds. *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition - Volume 2*, vol. 2. Association for Computing Machinery and Morgan Claypool, New York, NY, USA, 2018.

[35] Prandoni, P., and Vetterli, M. *Signal processing for communications.* Collection le savoir suisse, 2008.

[36] Ramos, J., Hong, J.-H., and Dey, A. K. Stress Recognition - A Step Outside the Lab:. In *Proceedings of the International Conference on Physiological Computing Systems* (Lisbon, Portugal, 2014), SCITEPRESS - Science and and Technology Publications, pp. 107–118.

[37] Rao, K. D., and Swamy, M. *Digital signal processing.* Springer Berlin Heidelberg, New York, NY, 2018.

[38] Ravi, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., and Yang, G.-Z. Deep Learning for Health Informatics. *IEEE Journal of Biomedical and Health Informatics 21*, 1 (Jan. 2017), 4–21.

[39] Ross, S. M. *Introduction to probability and statistics for engineers and scientists.* Academic Press, 2014.

[40] Russell, J. A. Core affect and the psychological construction of emotion. *Psychological Review 110*, 1 (2003), 145–172.

[41] Samson, A., Kreibig, S., Soderstrom, B., Ayanna Wade, A., and J Gross, J. Eliciting positive, negative, and mixed emotional states: A film library for affective scientists. *Cognition and Emotion 30* (12 2015).

[42] Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., and Van Laerhoven, K. Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. In *Proceedings of the 2018 on International Conference on Multimodal Interaction - ICMI '18* (Boulder, CO, USA, 2018), ACM Press, pp. 400–408.

[43] Schmidt, P., Reiss, A., Duerichen, R., and Van Laerhoven, K. Wearable affect and stress recognition: A review. *arXiv:1811.08854 [cs]* (Nov. 2018). arXiv: 1811.08854.

[44] Schutt, R., and O'Neil, C. *Doing Data Science: Straight Talk from the Frontline.* O'Reilly Media, Inc., 2013.

[45] Snoek, C. G. M., Worring, M., and Smeulders, A. W. M. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia - MULTIMEDIA '05* (Hilton, Singapore, 2005), ACM Press, p. 399.

[46] Soleymani, M., Lichtenauer, J., Pun, T., and Pantic, M. A Multimodal Database for Affect Recognition and Implicit Tagging. *IEEE Transactions on Affective Computing 3*, 1 (Jan. 2012), 42–55.

[47] Su, P., Ding, X.-R., Zhang, Y.-T., Liu, J., Miao, F., and Zhao, N. Long-term blood pressure prediction with deep recurrent neural networks. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)* (Las Vegas, NV, USA, Mar. 2018), IEEE, pp. 323–328.

[48] Yildirim, A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification. *Computers in Biology and Medicine 96* (May 2018), 189–202.

A Number of records in Chest and Wrist

Subjects	Chest points	Wrist points (ACC)	Wrist points (BVP)	Wrist points (EDA AND TEMP)	Windows
S2	1484700	67776	135552	8472	8463
S3	1508500	68864	137728	8608	8599
S4	1515501	69184	138368	8648	8639
S5	1551900	70848	141696	8856	8847
S6	1541400	70368	140736	8796	8787
S7	1538601	70240	140480	8780	8771
S8	1546299	70592	141184	8824	8815
S9	1537900	70208	140416	8776	8767
S10	1593900	72768	145536	9096	9087
S11	1559600	71200	142400	8900	8891
S13	1558201	71136	142272	8892	8883
S14	1558901	71168	142336	8896	8887
S15	1563100	71360	142720	8920	8911
S16	1554701	70976	141952	8872	8863
S17	1593200	72736	145472	9092	9083
In Total	23206404	1059424	2118848	132428	132293

Table A.1: Number of data Points

B Abnormal Subjects

Subject 2 Additional Notes
The RespiBAN temperature sensor was not fully attached throughout the entire duration of the study protocol.

Subject 3 Additional Notes
During the baseline condition, the subject was sitting in a sunny workplace.
Subject provided a valence label of 7 after the stress condition, claiming that he was looking forward to the next condition and was therefore cheerful.

Subject 5 Additional Notes
Subject might have fallen asleep during the first meditation.

Subject 6 Additional Notes
Subject claimed that he had a stressfull week and hence the study was rather relaxing for him.
Stress condition / TSST interview part: subject was not really stressed as he encountered many interviews in the weeks before.

Subject 8 Additional Notes
Subject had already a rather stressful day, prior to the study.
Subject felt rather cold in the room where the stress condition was carried out.

Subject 15 Additional Notes
Subject didn't really believe the cover story of the stress condition (TSST).

Subject 16 Additional Notes
Subject felt rather cold in the room where the stress condition was carried out.

Subject 17 Additional Notes
The RespiBAN temperature sensor was not fully attached throughout the entire duration of the study protocol.

C Raw and processed signals for each state

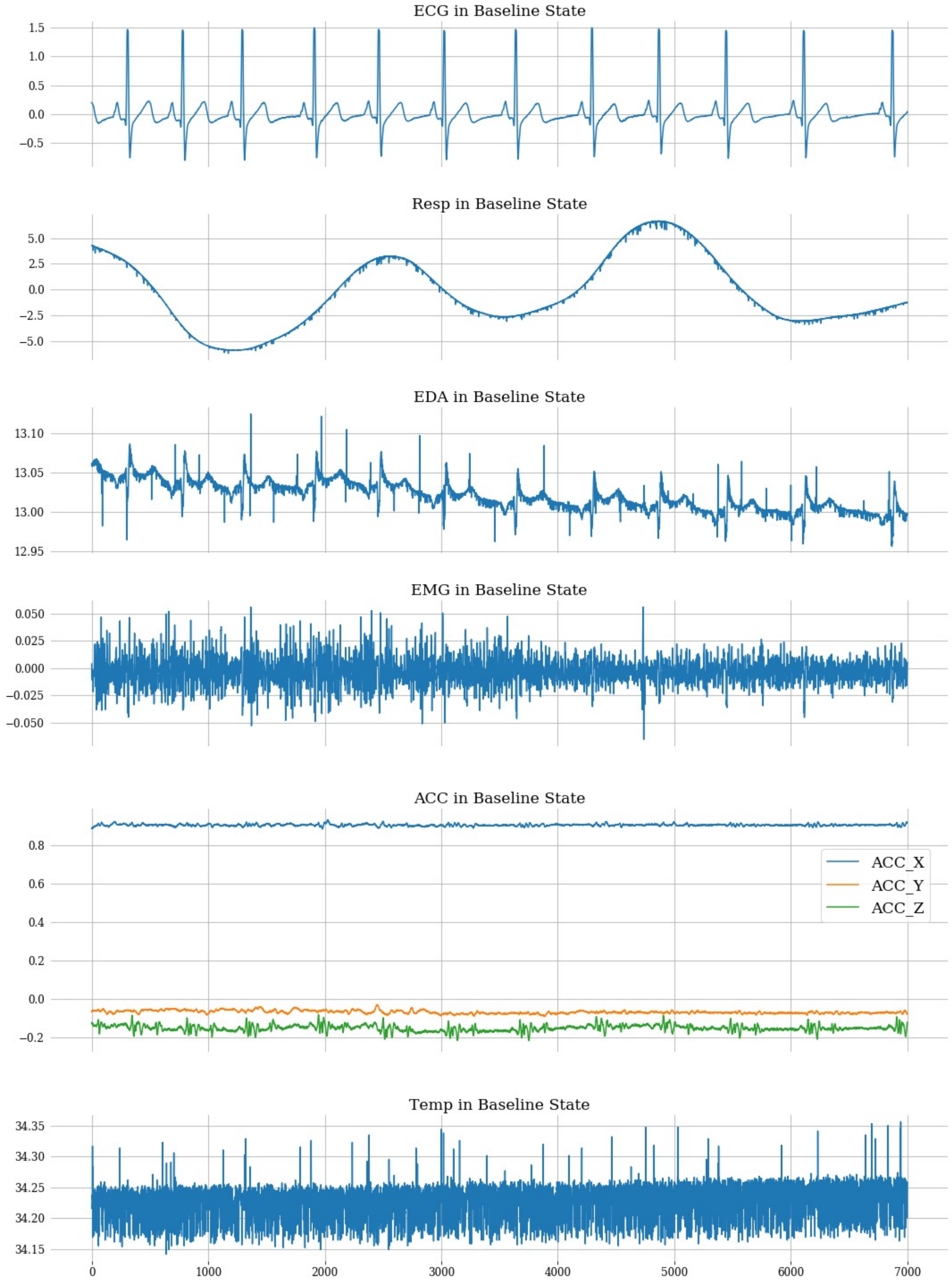


Figure A.1: Subject 7 baseline state without filtering

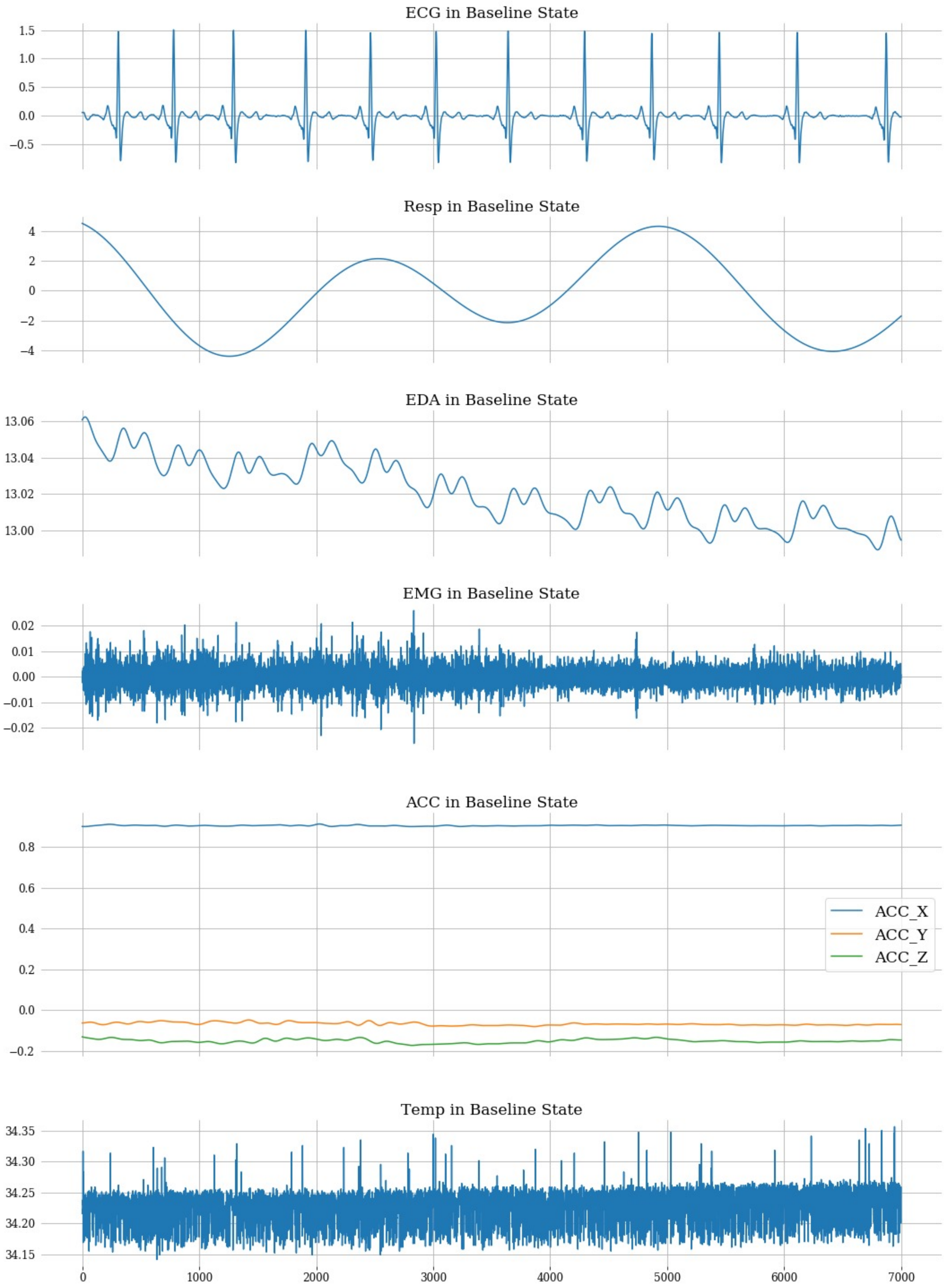


Figure A.2: Subject 7 baseline state with filtering

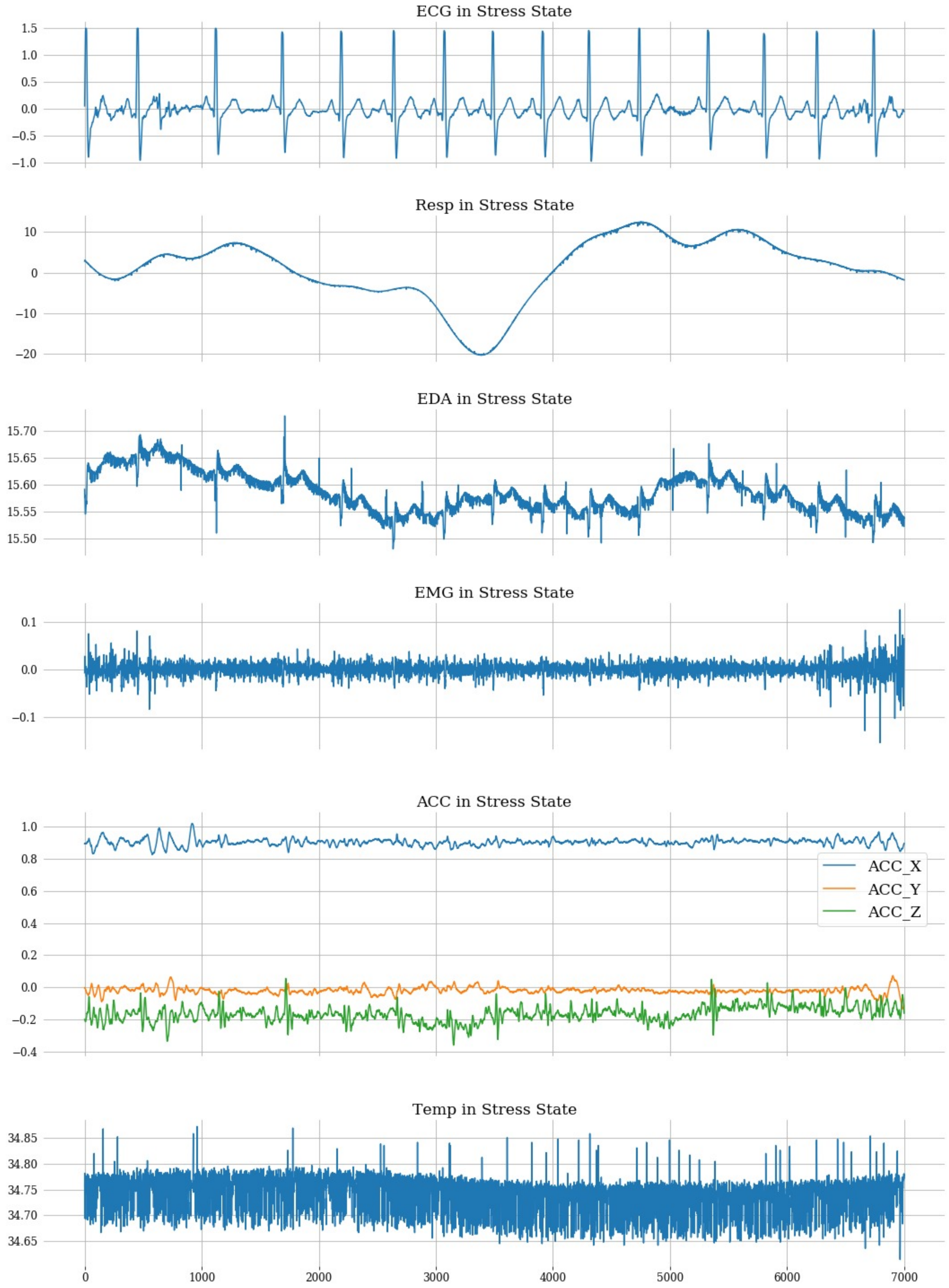


Figure A.3: Subject 7 stress state without filtering

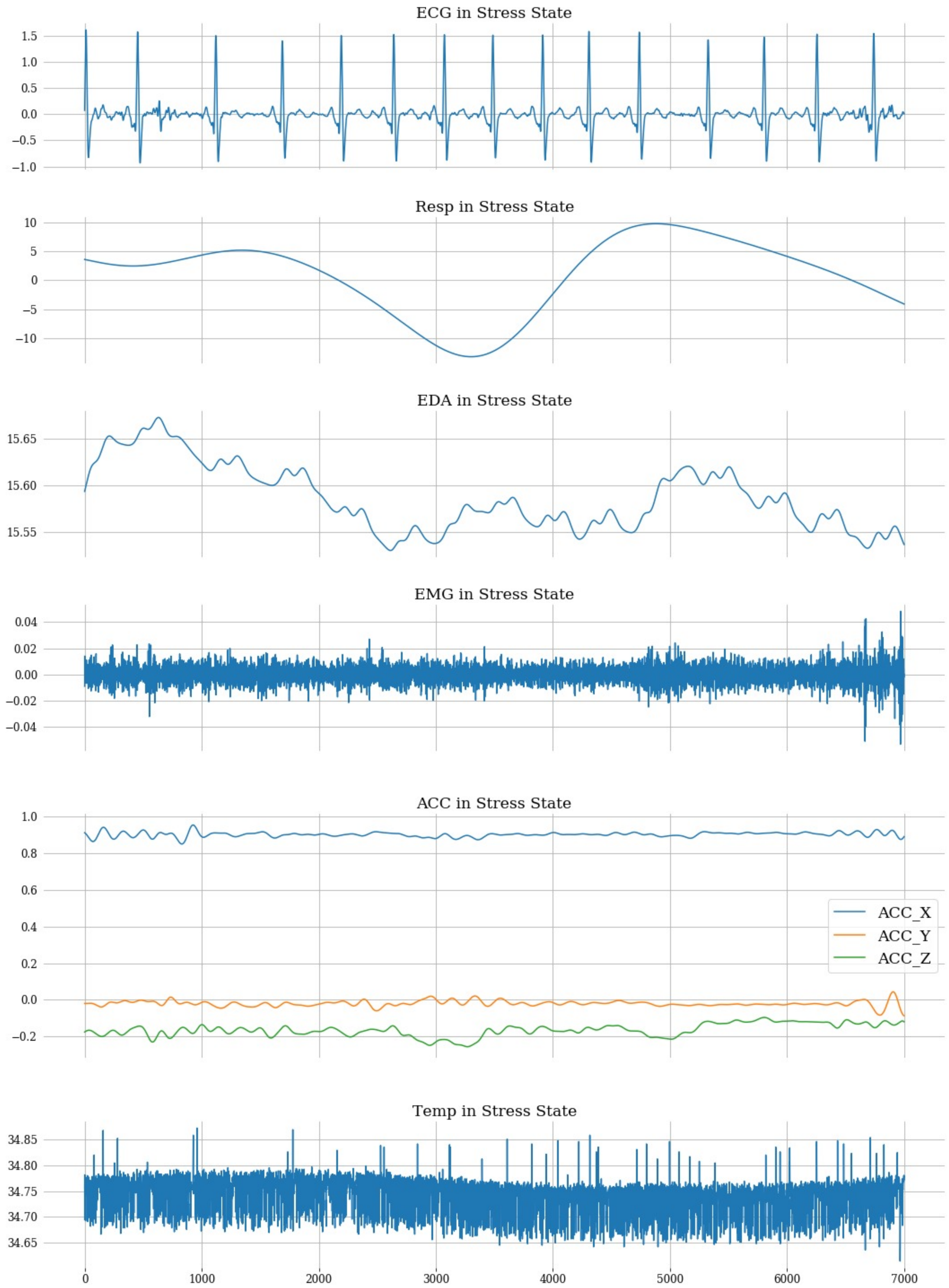


Figure A.4: Subject 7 stress state with filtering

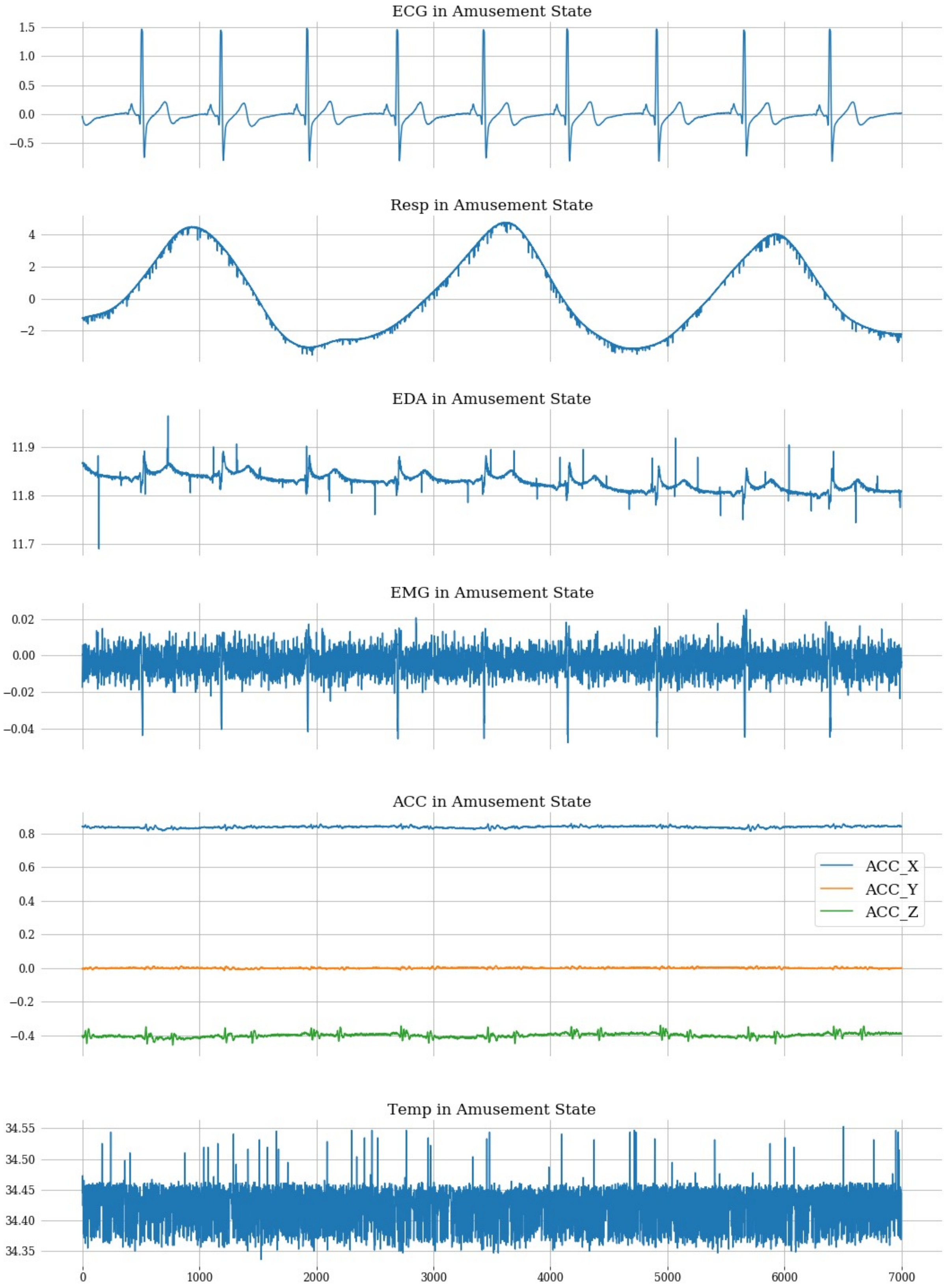


Figure A.5: Subject 7 amusement state without filtering

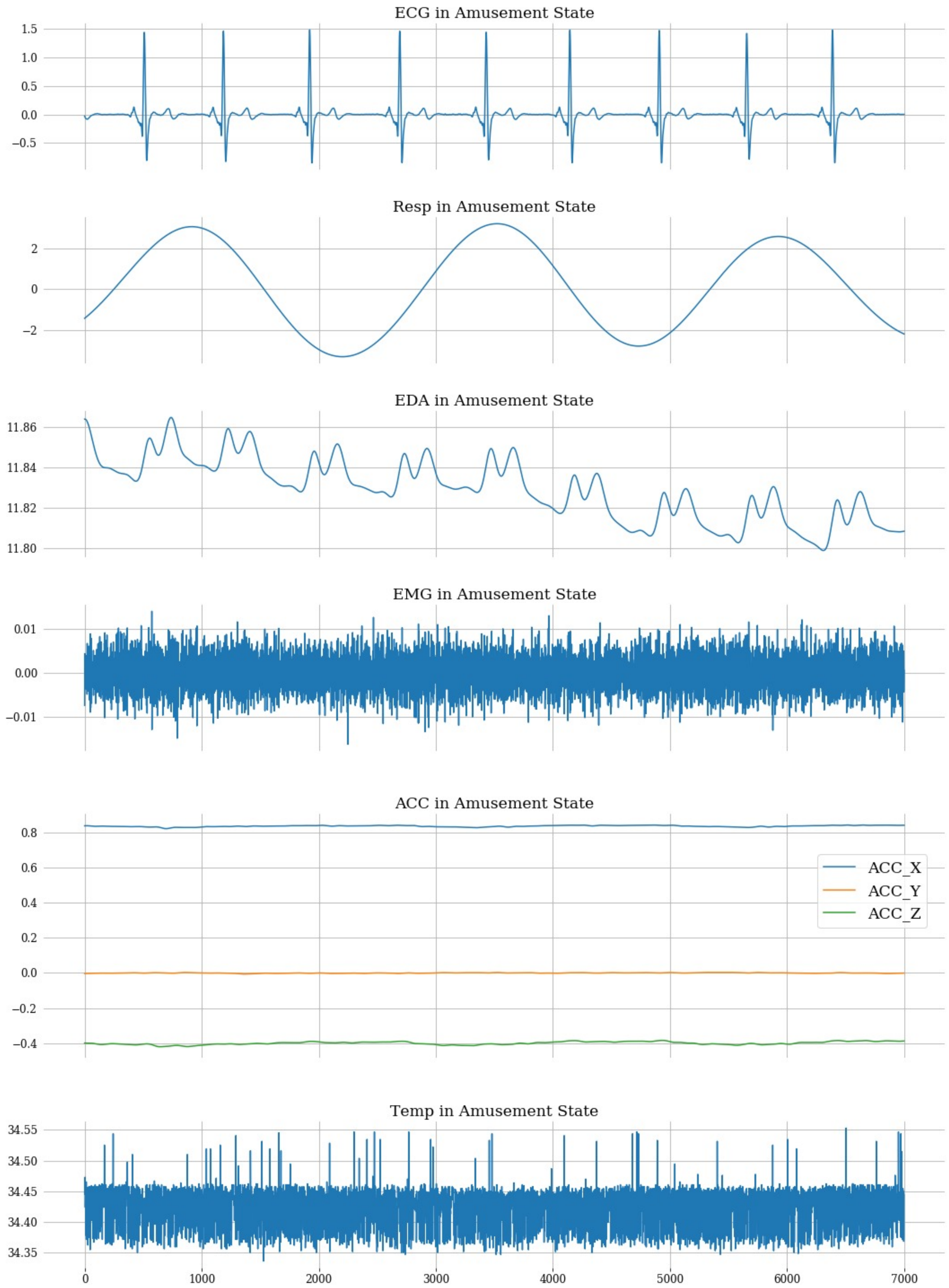


Figure A.6: Subject 7 amusement state with filtering