Modelling stress levels based on physiological responses to web contents

ISIAKA, Fatima

Modelling Stress Levels Based on Physiological Responses to Web Contents

Fatima Isiaka

This thesis is dedicated firstly to Almighty God for his protection and guidiance, to my husband Ibrahim Adamu Mailafiya and son Khalifa Ibrahim Adamu Mailafiya for their companionship, friendship and guidiance. I am most thankful to God for my father Isiaka Adamu, my mother Rakiya Adamu, my sisters Amina, Halima, Maryam, Bilkisu, Sophia (Elai), Zainab, Maimuna, my brothers Awwal and Khalid for their undying support. To all my nieces and nephews. And finally my pretty little daughter Fareeda (Turai) for bearing my absense during the period of this research.

Acknowledgements

My thanks to my supervisor, Dr Kassim Mwintondi for his unflinching support, wonderful patience, and great supervisory aptitudes throughout the period of this research, to my former supervisors Dr Simon Harper, Dr Caroline Jay, my colleagues at the Web Ergonomics lab in University of Manchester for thier kind hardcore drilling and made sure I become an independent researcher, love you guys. To to my colleagues in Shefflied Hallam Unit 12, Ghayda, Sophia, Aramide, ken and the others. Finally to Dr Frances and all the staff in the research admistrative office espesially Tracey, Rachel and Claire for the support they have given to both home and internaltional students, they made sure we all are welcome no matter who you are, religious background and cultural differences. God be with you all.

Abstract

Capturing data on user experience of web applications and browsing is important in many ways. For instance, web designers and developers may find such data quite useful in enhancing navigational features of web pages; rehabilitation therapists, mental-health specialists and other biomedical personnel regularly use computer simulations to monitor and control the bahaviour of patients. Marketing and law enforcement agencies are probably two of the most common beneficiaries of such data - with the success of online marketing increasingly requiring a good understanding of customers' online behaviour. On the other hand, law enforcement agents have for long been using lie detection methods - typically relying on human physiological functions - to determine the likelihood of falsehood in interrogations. Quite often, online user experience is studied via tangible measures such as task completion time, surveys and comprehensive tests from which data attributes are generated. Prediction of users' stress level and behaviour in some of these cases depends mostly on task completion time and number of clicks per given time interval. However, such approaches are generally subjective and rely heavily on distributional assumptions making the results prone to recording errors.

We propose a novel method - PHYCOB I - that addresses the foregoing issues. Primary data were obtained from laboratory experiments during which forty-four volunteers had their synchronized physiological readings - Skin Conductance Response, Skin Temperature, Eye tracker sensors and users activity attributes taken by a specially designed sensing device. PHYCOB I then collects secondary data attributes from these synchronized physiological readings and uses them

for two purposes. Firstly, naturally arising structures in the data are detected via identifying optimal responses and high level tonic phases and secondly users are classified into three different stress levels. The method's novelty derives from its ability to integrate physiological readings and eye movement records to identify hidden correlates by simply computing the delay for each increase in amplitude in reaction to webpages contents. This addresses the problem of latency faced in most physiological readings. Performance comparisons are made with conventional predictive methods such as Neural Network and Logistic Regression whereas multiple runs of the Forward Search algorithm and Principal Component Analysis are used to cross-validate the performance. Results show that PHYCOB I outperforms the conventional models in terms of both accuracy and reliability - that is, the average recoverable natural structures for the three models with respect to accuracy and reliability are more consistent within the PHYCOB I environment than with the other two.

There are two main advantages of the proposed method - its resistance to over-fitting and its ability to automatically assess human stress levels while dealing with specific web contents. The latter is particularly important in that it can be used to predict which contents of webpages cause stress-induced emotions to users when involved in online activities. There are numerous potential extensions of the model including, but not limited to, applications in law enforcement - detecting abnormal online behaviour; online shopping (marketing) - predicting what captures customers attention and palliative in biomedical application such as detecting levels of stress in patients during physiotherapy sessions.

Contents

List of Figures

Chapter 1

Introduction

In many ways, capturing data on user experience of web applications and browsing is very useful. An example of such data are used by web designers and developers in enhancing navigational features of web pages. Also, rehabilitation therapists, mental-health specialists and other biomedical personnel often use computer simulations to monitor and control the bahaviour of patients. Marketing and law enforcement agencies are perhaps two of the most common beneficiaries of such data - with the success of online marketing increasingly requiring a good understanding of customers' online behaviour. For long, Law enforcement agents have also used human physiological measures to determine the likelihood of falsehood in interrogations.

Quite often, online user experience is studied via traditional measures such as task completion time, surveys and comprehensive tests from which data attributes are generated. Prediction of users stress level and behaviour in some of these cases depends mostly on task completion time and number of clicks per given time interval. However, such approaches are generally subjective and rely heavily on user memory and distributional assumptions leading to bias and prone to recording errors, to clearly tackle such problems, there is need for deep understanding of hidden concepts in Human computer interaction and human physiological response.

Various aspects of Human Computer Interaction (HCI) - particularly on web applications - have been widely studied in recent years (Nielsen, 1994). One motivation for such studies has been the need to gain insights into how user

interaction with computing applications may help in monitoring and controlling stress levels (Zhai & Barreto, 2006). Studies have revealed that human physiology reacts to an extensive range of emotional measures and physiological signs of arousal have been known to be correlated to mental occurrences like positive and negative emotions (Filipovic & Andreassi, 2001). Studies have therefore focused on, inter-alia, variations in user attention and on subjective measures of how users respond to different stimuli (Smith *et al.*, 1999). What about collecting the data through objective methods, however, and using the data for modelling? Also, there are stress related algorithm (Marullo & Randall Jr, 2000) with tools that simulate simultaneous users based on task completion time and number of clicks per a given time interval; how would appending physiological attributes of users affect performance when applied? These are part of what we intend to investigate in the thesis.This research is based on applying objective methods and employing an algorithm to determine the connection between the emotional response of users and web content such as video and picture contents.

1.1 Motivation

The human physiological response is very important in HCI, in the respect that it provides physiological measures which can be used to understand how users feel during interaction with different interfaces in applications such as software, games and webpages. The major motivation behind the research is the need to gain more insight in HCI and the human physiological response to the static and dynamic contents of a webpage and also to be able to control stress levels during user interaction by utilising the user attributes and identifying the content that changes stress levels. The understanding of these user attributes depends particularly on an appreciation of the physiological attributes of the users. By definition, the psychophysiological response of humans is the part of physiology that is concerned with the measurement of emotional responses as they relate to behaviour. In this context, behaviour is mostly referred to in the broadest sense to include activities like problem solving, stress reaction to stimuli, perception, etc (Filipovic & Andreassi, 2001; Stern *et al.*, 2001).

Human behaviour can be reduced to six basic emotions, excitement/happiness anxiety, sadness, surprise, anger and disgust (Kim *et al.*, 2004), which can be detected using physiological measuring sensors. For example, a person's emotional state can be extrapolated based on his/her Skin Conductance (SC), which can simply be described as the body's reaction to stress or stimuli.

The purpose behind this research is the need for acquisition: a vision into how user interaction with software applications like webpages, can assist in monitoring and controlling the stress level of users, using physiological measures such as Skin Conductance Response (SCR), Skin Temperature (ST), Pupil dilation (PD) and Eye movement. We can extract the physiological variables from these measures, predict an emotional epitome and correlate this to user interaction by modelling the physiological reaction and determining those components that have the most effect on users in HCI.

Some of the physiological attributes used to extrapolate the emotional state of users in this research, include measures of changes of pupil size, eye movement (saccade and fixation points), electrodermal activity (changes in the electrical activity of the skin surface as a result of sweat induced by users' emotion) - mostly referred to as the Skin Conductance response and, finally, the skin temperature (change in temperature at the extremes of the body).

1.2 Background

It is generally known in the HCI literature that the physiological response of users normally correlates to their task performance, particularly when stress stimuli are applicable to interactive settings (Zhai & Barreto, 2006). One of the areas that deals most with human emotion in order to develop interfaces is the field of Affective Computing. This is a recent area of computing research that arises from processes that relate to or influences emotions (Hudlicka, 2003; Picard, 2000). It describes the importance to HCI of emotional features in order to automatically detect stress in users by developing mechanisms that can make a model or a computer aware of the users' emotional state. This section provides the background of the underlying mechanics.

Emotions are mostly seen as evolving from adaptive values such as fitness to basic life tasks (Ekman, 1992); each emotion has its own unique features which include physiology, antecedent events and signals. Each emotion also have similar characteristics in common with one another, such characteristics can include short duration or occurrence over short term time intervals occurrence, rapid onset or spontaneous reaction, unbidden occurrence, automatic appraisal and coherence among responses. These unique relationships are a product of evolution and differentiate them from other affective phenomena (Ekman, 1992; Ortony & Turner, 1990).

There are presently different types of biosensors that can measure the emotional state of arousal, through the skin conductance, which is a function *f(s)* of electrical changes s of the skin as a result of sweat. These sensors measure the electrodermal activity as it grows higher during states such as interest, attention or nervousness, and lower during states such as relaxation or boredom (Eq 1.1), depending on the task the user is involved in (Uğur, 2013).

$$f(s) = \begin{cases} Stressed\ (High) \\ Relaxed\ (Low) \qquad \forall s \in \mathcal{E} \end{cases} \tag{1.1}$$

where $\mathcal{E}$ can be any emotional arousal state. This expression can be further expanded in (Eq 1.2). The emotional state $f(s)$ can be stress, neutral or relaxed mood which is equivalent to 0, 1 and s, if $\mathcal{E}$ (positive or negetive affect) is substituted in the equation; s is a neutral mood that is neither 1 nor 0.

User experience can be reflected and determined through their attitude or behaviour (Castañeda *et al.*, 2007). A negative attitude towards a complex application shows a poor experience while a positive attitude towards a complex application shows a good experience (Saint-Aimé *et al.*, 2009). The three possible emotional states discussed here that can be experienced by a user during an interactive session could be expressed as stress, relaxation and a neutral mood, and can be demonstrated in the following concept.

$$f(s) = \begin{cases} ((1-\mathcal{E}) \times 0) + (\mathcal{E} \times s) & : \text{ if relaxed (positive mood)} \\ ((1-\mathcal{E}) \times s) + (\mathcal{E} \times 1) & : \text{ if stressed (negative mood)} \\ s & : \text{ if neutral (neutral mood)} \end{cases} \tag{1.2}$$

where

$$\mathcal{E} = \begin{cases} 0 & : \text{ if negative affect} \\ 1 & : \text{ if positive affect} \end{cases}$$

$$with \left| \begin{array}{lcl} f(s) & = & \text{user emotion} \\ \mathcal{E} & = & \text{user experience emotion} \end{array} \right.$$

The way we view or perceive our everyday user applications can be reflected by the way we react to the interfaces every time we log on to an online application system. Most applications, such as a complex website or complex gaming software, have the potential to induce mixed emotions such as anger and frustrations which mostly make the user uncomfortable and dissatisfied at that moment; this kind of reaction from the users induce emotion which is normally termed as **stress**, especially when they are using the application for the first time. If a user is relaxed during an interactive session with an interface then we say the user finds the application less complex and easy to deal with.

Most comprehension of stress in user and understanding user perception is through heuristics methods (Nielsen & Molich, 1990) and direct measures such as administering questionnaires and talking to the users to understand how they perceive interfaces (Nielsen, 1994). Studies such as (Healey & Picard, 2005; Zhai & Barreto, 2006) tend to investigate through non-invasive or non-obtrusive methods with several physiological signals and find ways to visualise and assess the emotional state or stress level of users that will improve the detection of emotion and produce affect detection systems in usability studies.

The main emotional classification states we aim to look at in this research are the 'stressed', 'neutral' and 'relaxed' moods of users. Distinguishing and finding

the similarities between these states of affect can be achieved by conducting an experiment that involves the use of physiological measuring sensors to monitor the users' reaction and collect the data. We will be looking at the stimuli eliciting psychological stress that involve stimuli conditions in the form of webpages with static and dynamic contents; these are laboratory based stimuli situations achieved by deactivating some contents on the webpages and taking note of how this affects the users.

1.3 Research Question and Objectives

The project seeks to answer the following research question:

How can user-generated data be utilised in modelling and simulating human physiological responses to the visual content of the web? To answer this question, the following research objectives are required.

- To provide a thorough review of existing literature on HCI and Human Physiological Responses (HCI-HPR) within an inter-disciplinary context.
- To elicit HCI-HPR related data from sampled users using specialised tools in an ergonomic laboratory.
- To develop an algorithm for determining HCI-HPR associations and explore the potential of HCI-HPR modelling.
- To determine the existence of natural structures using Principle Component Analysis (PCA) and Forward Search (FS).
- To compare the proposed model to existing or standardised techniques such as Neural Networks and Logistic Regression.

1.4 Objective Measures (Step towards Contribution)

This section discusses the measures taken to achieve the set objectives, which is briefly discussed in the following steps:

- Most related work to the area of HCI-HPR would be discussed, some of the problems would be investigated, especially the most persistent issues, such as latency in physiological response and emotion recognition on stimuli interface. These issues will be revisited and discussed in the literature review to identify weaknesses and lapses, which would then be addressed in the project work and thesis.

- An experiment will be conducted using physiological measuring sensor and an eye tracker in an ergonomics labouratory with web users to collect user data where the user attributes would be generated.

- A novel approach to modelling HCI-HPR data will be implemented by developing an algorithm (Physiological correlates to online behaviour (PHYCOB I)) that detects peaks (optimal response) in physiological measures which correspond to user activity and thence to classify stress level and make predictions based on identifying patterns in the data.

- A comparative data analysis will be applied that involves standard techniques such as Neural Networks and Logistic Regression, this will help to validate significance of the proposed model and also its reliability.

- By using PCA and a Forward search algorithm that helps to elicit individual stress levels of users and the corresponding webpages they interacted with, we will be utilising these methods which are distinct from the predictive models and would help to detect natural structures that also validates the proposed model's method.

1.5 Summary of the Thesis

This section contains a summary of the entire thesis, each chapter is summarised to give a brief view of what is contained in the chapters.

- Chapter One: Chapter one contains an introduction to the thesis and talks about how human physiological response is important in the connection

between the emotional response of users and visual online content. It also discusses the objectives and motivations behind the research.

- Chapter Two: Chapter two discussed some related work to this area of research some of the problems were investigated, especially the most persistent issues. These issues were revisited and discussed in the literature review to identify weaknesses and lapses, which will be addressed in the thesis.

- Chapter Three: This chapter discusses the experiments conducted in an ergonomics laboratory with experienced and above average participants which is discussed in the methodology of Chapter three. This involved the pilot and main experiment; similar procedure and precautions were taken for both study. Part of Chapter three also discussed and investigates the prospect of HCI-HPR modelling where an algorithm (PHYCOB I) was developed. Both the methodology, design and implementation of PHYCOB I was discussed in a manner that paves the way for exploring the possibilities of HCI-HPR modelling. The steps in PHYCOB I involves two modules:

  1 Generation of users' attributes from the primary data source (sensors).

  2 Prediction of users' stress levels from these attributes.

  The performance from the predictions is used to compare with the two predictive models, Neural Network and Logistic regression. The Principal component analysis and Forward search algorithm were additional methods for validation; these methods used a different routine for pattern recognition unlike the two predictive models. The process for all the methods were discussed in Chapter three.

- Chapter Four: Chapter Four contains the analysis and results from all the methods used in the thesis. The chapter is in a form of comparative data analysis of all the methods and findings from these methods.

- Chapter Five: Chapter five contains conclusions and discussions of the entire thesis. The achievements, implications, limitations and contributions of

the project work were critical addressed and this gives room for the future work.

1.6 Summary

This chapter has discussed how the human physiological response is important in the connection between the emotional response of users and visual online content. It briefly highlights users' sympathetic system reaction to stress or stimuli, which can be measured through the autonomic nervous system. The motivation behind the research is the need to gain insights into how user interaction with software applications like webpage, can assist in monitoring and controlling the stress level of users, using physiological measures. This informs a research question asking how can user-generated data be utilised in modelling human physiological responses to visual content of the web? To answer this question, certain standardised goals were set to achieve the stated objectives, including exploring the possibilities of HCI-HPR modelling, providing insights into stress-related analyses and developing an algorithm for determining HCI-HPR associations.

Chapter 2

Review of Related Work

In HCI, user experience (UX) is often limited to unconcealed observable behaviour (Bergstrom & Schall, 2014). Methods such as interviews, questionnaires and other investigations mostly rely on users' memory and subjective judgements as a good means of gaining insight into the cognitive and emotional processes of users.

A common finding in cognitive neuroscience (Kretzschmar *et al.*, 2013), states that a person's subjective perception of their behaviour does not always relate to their neural activity. Experiments have shown that people do not always know what is going on inside their minds. For instance, in an eye tracking study that involve reading (Liversedge & Blythe, 2007), objective real-time quantitative measures of eye movements, revealed longer fixation times for reading text with transposed letters as compared to reading normal text despite the fact that readers claimed to spend just a few seconds on text with transposed letters. Also Electroencephalography (EEG) of language processing (Demiral *et al.*, 2008) has concluded that phrases judged as easy to comprehend and highly acceptable sometimes entail a larger processing effort on the part of the readers.

While these findings have shown that visual attention and engagement are not necessarily linked to user perceptions of their experiences, today we struggle to determine the best possible techniques to allow us have a greater understanding of users' unconscious minds as they interact with elements of a visual stimulus like a webpage. The application of physiological measures has evoked enthusiasm in this regard. In the study of physiology and HCI, the main interest is on mechanisms (Filipovic & Andreassi, 2001; Green, 1976); on exactly how a particular function

is performed. Why do reactions to stimulus produce sweat on the skin? Why does the heart palpitates in reaction to an event? If we can simply watch what goes on in the users' mind through their physiological readings and collect quantitative data for modelling, we can come close to determining users' cognitive processes.

The following sections discuss various work done in relation to measuring emotion, including explicit reactions and physiological measures in stress related studies and how they are used as metrics for HCI evaluation by providing insights into users' cognitive states. We first discuss physiological measures in objective perception as related HCI.

2.1 Physiological Response as an Objective Measure.

Users sometimes do not feel comfortable talking about their experiences in a usability study, such as letting people know what they really think or feel about a particular interface. This might be due to the fact that they feel it is socially inappropriate or they feel that they are the problem rather than the interface, this as been noted in older participants (Gross *et al.*, 1997).

Objective measures do not rely on a users' experience or assessment, rather they record and measure time and task completion (Bergstrom & Schall, 2014) as user attributes, one novel approach we applied was adding physiological attributes. Physiological response measurements allow for further collection of objective measures of performance, rather than asking participants if they find a task difficult or if they were surprised or their attention was divided when visual stimuli like dynamic content suddenly appeared on screen. The Skin Conductance Response (SCR) can be used to measure their reaction. Objective Skin Conductance (SC) data, when combined with eye-tracking data, can give a different view of the UX, such as a user experiencing emotional arousal with the sudden appearance of dynamic content (Bergstrom & Schall, 2014). Sometimes users may subjectively rate what they feel as non-excited, not-amused, not-interested, or not-stressed, but their physiological response readings may reveal that at that

point in time emotional responses occurred which indicates an increase in amplitude that signifies excitement, amusement, interest or stress. Data that has been collected from users can be further used for modelling to substantiate such assertion. The next section briefly introduces the physiological measures used in this report.

2.2 UX and Physiological Measures.

Physiological response measures are currently used in UX. This is because users' emotional state form part of physiological characteristics such as the increase in electrical changes of users' skin in response to stimulus (Bergstrom & Schall, 2014), which is observed through the variations or spikes in the physiological readings e.g. in a lie detector study that seeks to investigate the likelihood of falsehood in an interrogation. Research into HCI is currently in a place where it is possible to actually measure emotions and implicit reactions, while users interact with stimuli presented to them. Data can be collected and user physiological responses examined without interrupting data collection both in real time with a delay. In real time, user reaction can simply be observed and certain significant events noted, while delayed analysis allows sessions to be exported and run again to register the response rate in a specified place and compare this with readings observed in real time. This can also tackle the problem of latency (delay) which is often the most important parameter in physiological and eye movement readings (Andreassi, 2000a; Bergstrom & Schall, 2014). Investigations of this aspect are often very limited due to the unforeseen error in latency in physiological readings e.g. SCR. The physiological measures used for the purpose of this study are briefly discussed below, most of the user attributes used for the development of the proposed model are also mentioned.

2.2.1 Skin Conductance Response

The Skin Conductance Response (SCR) provides a functional signal of emotional responses by measuring the Electrodermal (EDA) changes of the skin, caused as a result of sweat (Andreassi, 2000a). This reaction can be measured with electrodes

non-invasively placed on the wrist or palms. The diagram below (Figure 2.1) shows the important components that need to be considered when measuring skin conductance. The latency is the time delay from the onset of the stimulus to an onset or rise time of a response, while the amplitude is the difference between minimum peak response and the maximum peak response at a given time interval. We were mostly interested in what happens at the phasic changes, especially the high tonic phase and peak response in reaction to webpage contents. The threshold is used to distinguish one consecutive peak from another and defines the tonic phase (baseline).

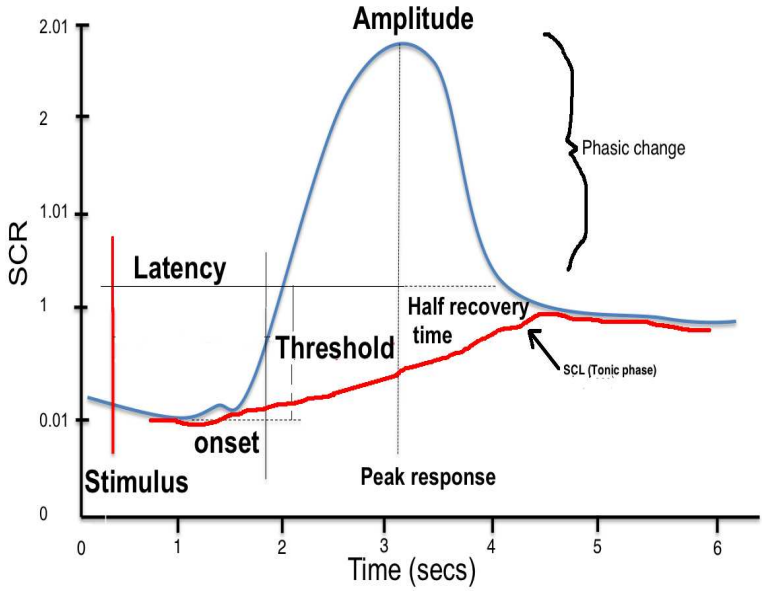


Figure 2.1: SCR with typically computed features

2.2.2 Skin Temperature

The skin temperature (ST) changes according to blood circulation at the surface of the skin (Figure 2.2) through body tissue. In a state of increased emotion, such as interest or stress, muscle fibres contract and cause a stenosis of the vasculature (Kamon *et al.*, 1974; Mindfield, 2014). This leads to a reduction of skin temperature since blood circulation through the tissue is reduced. On the other hand, in a state of moderation and rest, the musculature is compelled to relax, which causes

the vasculature to increase and thus the skin temperature rises. Usually mental stress leads to a lower exterior perfusion and a decrease of skin temperature in areas like the hands. (Mindfield, 2014). The mean skin temperature (MST) for an entire episode can be considered and used as an attribute in terms of user stress classification.

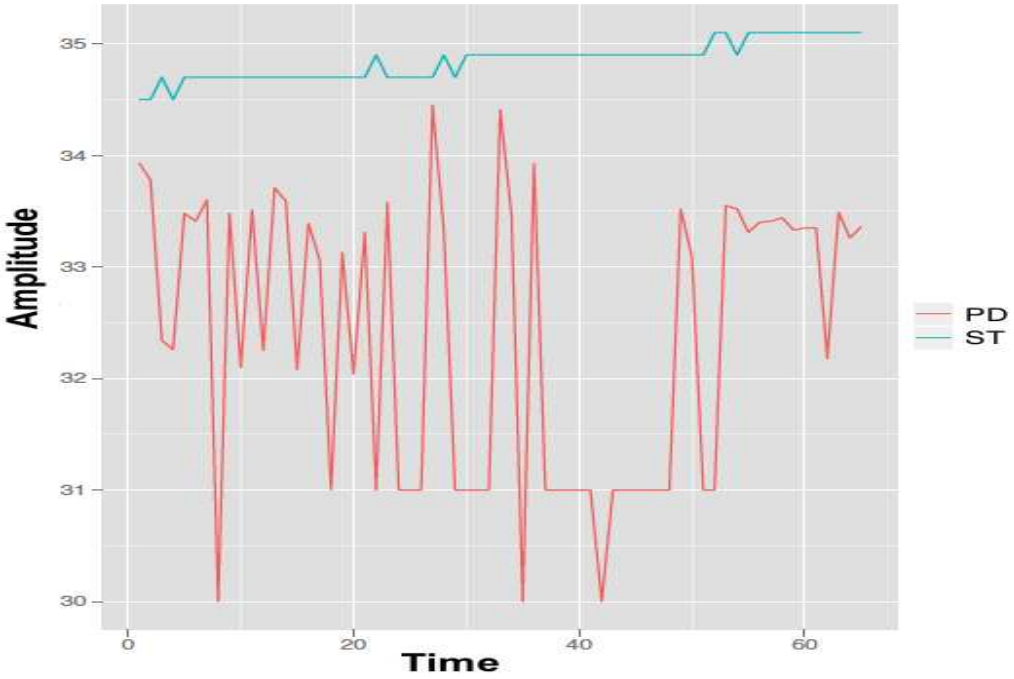


Figure 2.2: Synchronised Pupil-change and Skin temperature measured over time

2.2.3 Pupil Dilation

The pupil dilation does not only reveal changes in light intensity, it is also a measure of underlying cognitive processes (Figure 2.2) as user interacts with visual contents. It provides indices of attention, interest or emotion which are correlated with mental workload and arousal. The variations in pupil change and the average pupil change for a given time interval are considered to be important when measuring eye movement and the behaviour of users in reaction to visual stimuli (Iqbal *et al.*, 2004).

Figure 2.3: Fixations and saccades caused by eye movement

2.2.4 Eye movement

This is the behaviour of the eye during interaction; the eye gaze pattern is a measure of behaviour. The movement of a users' eyes is based on fixations (location of a users' eye gaze), saccades (rapid movements of the eye from one fixation to another), and fixation duration (length of time a user fixates on a particular area) (Figure 2.3) (Bergstrom & Schall, 2014). These parameters are important when modelling HCI-HPR associations, because they are important attributes that contribute to classification of stress level in respect to how users react to webpage content.

2.2.5 Stress Related Studies on Webpages.

There has been quite a lot of work done in modelling users' interaction with web and software applications but few that involves integration of physiological and eye-tracking data for modelling using classification algorithms. Studies (Granka *et al.*, 2004; Russell, 2005) that contribute to the understanding of first impressions of a website tend to apply just the eye tracking method and this also affects traditional usability testing. These studies measure first-time usage of websites.

The participants in the study viewed the home pages of different websites. The results showed that eye-movement data supplement the understanding of what users viewed online, which is a reflection of users' behaviour. Granaka et al. (2004) used eye-tracking to investigate how users interacted with the results page of WWW search engines, show how long it takes most participants to select a document; which took a latency of 7.78 seconds. This was simply to understand their browsing behaviour, when presented with an abstract and links to a webpage but does not provide an insight into their physiological responses.

Physiological measures, when combined with eye tracking can give more insight to these form of studies (Bergstrom & Schall, 2014), by simply synchronising the event data and identifying the correlate between the physiological data and information obtained from the eye tracker. This technique will be further investigated in this report.

Other studies have also worked on this such as those of (Mbipom, 2008; Ward & Marsden, 2003), that investigate how users are drawn to the visual aesthetics of the web and interacts with web content. Some of these studies are directed towards the use of physiological measurements as a means to understand how stressed users become when interacting with web content. The results showed that SCR is an effective tool for identifying the areas of content that cause users most stress by synchronising the SCR data, the observed behaviour data and observed eye movement data. Most of the results were achieved through standard processes. Even though the results obtained were accurate enough, it can sometimes take a long time to process results, especially when larger numbers of participants are involved. This project was aimed towards a similar research area and also to combined data from multimodal sources for modelling.

2.3 Modelling Physiological Data and Cognitive State Assessment

Physiological measures, when combined with human psychology (psychophysiological), can be referred to as the responses of users to mental effects (Andreassi,

2000b) and it is the objective signals that make it possible to establish and understand human emotional processes by observing their bodily or physical changes. This section is mostly concerned with the basics of psychophysiological events and the evaluation of both stress and other cognitive states. The initial stage is to give an insight into other methods that have investigated stress and attempted to arrive at a cognitive assessment of the human body and serves to highlight the basis of psychophysiological measures. This includes their advantages, disadvantages, definition, and methods of application.

Dirican & Gokturk (2011a) and Kramer (1990) stated that subjective measures, physiological measures or objective measures, and performance measures are the three kinds of measurements used in stress and other cognitive states. (Cain, 2007; Dirican & Göktürk, 2011a; Farmer & Brownson, 2003). These are also mentioned with the cognitive load (Dirican & Göktürk, 2011a; Ganglbauer *et al.*, 2009) and mental workload/stress assessment. Although their meaning may show differences, they can also be used in the same context (Dirican & Göktürk, 2011a; Kalyuga, 2008). For instance, cognitive load is a conception that is associated with working memory in cognitive load theory (Dirican & Göktürk, 2011a; Kalyuga, 2008) while on the other hand mental/stress workload is a more challenging concept, with several dimensions attached to it. It is related to task difficulties or potentials, motivation and emotional state of users (Dirican & Göktürk, 2011a; Farmer & Brownson, 2003; Kramer, 1990). In this thesis mental and cognitive workload are referred to as stress. Table 2.1 below shows some physiological measures used in the literature, describing their relative properties as compared to their diagnosticity and sensitivity.

The physiological measures used for this study are chosen based on their strength and weakness and also on their analytical assessment. For SCR, the phasic changes were considered because its time-based sensitivity is reduced in respect to the tonic phase and we are looking for optimal changes at the phasic level. ST has an impact on diagnosis; as stress increases, the temperature decreases as heat moves to the body's core. Eye movements was chosen based on the fact that the eye gaze is a metric for mental interest and it is an appropriate measure for user interface evaluation, user experience and also in usability assessment. The pupil dilation of the eye is directly related to mental workload (stress)

Measures	Defination	Analytical/Assessment	Advantages/Disadvantages
SCR(Ganglbauer *et al.*, 2009; Kramer, 1990; Park, 2009)	Metrics fluctuates in the conductivity of a persons skin	SCR is directly related to arousal. Its also a metric for stress. Its temporal sensitivity is poor, only to tonic changes	reduced sensitivity to noise and less vague than facial electromyogram EMG and heart beat electroencephalogram (ECG)
ST(Riva *et al.*, 2003; Sappenfield *et al.*, 2013)	Measures the changes in temperature on the extremities of the body	As stress increases the temperature in the extremities decreases as heat moves to the bodys core	Gives accurate readings of body electrical resistance. Slow response time
Eye Movements (Allanson & Fairclough, 2004; Dirican & Göktürk, 2011a)	measures gaze, saccades and fixations	Eye duration provides useful information about task performances. It is a linear measure of interest.	an accurate measure of usability testing.
Pupil Diameter(Allanson & Fairclough, 2004; Ganglbauer *et al.*, 2009; Kramer, 1990)	measure changes in pupil size and light intensity	It is correlled with mental workload. Register general changes in knowledge processing, responds to mental processes.	Very difficult to apply in real-life settings since eye respond to different light conditions but can be accurately recorded with appropriate eye tracking sensors.
Respiration (Allanson & Fairclough, 2004; Park, 2009).	used as a metric for task performance.	Also applied in measure of arousal.	Affects SCR and heart rate measures.
Heart Rate (HR) (Allanson & Fairclough, 2004; Park, 2009)	Measures the changes in heart beat (bpm)	responsive to cognitive demands, mental workload,attention and correlated with arousal, assess positive or negative emotion of interest and experience	Interpreting signals related to the user interaction process is challenging.
Blood Pressure (BP) (Park, 2009)	Measures pressure exerted by circulating blood upon the walls of blood vessels	BP increase at active coping and patterns of ECG, and interprets differences between reactions of users during challenging tasks	Applied in the evaluation of important user interfaces.
Electroencephalography (EEG) (Ganglbauer *et al.*, 2009; Kivikangas *et al.*, 2011)	Measures and records the electrical activity of the brain	Alertness and task performance. Responsive to phasic and tonic changes.	Responsive to both physical and electrodes artifacts, though not very condusive for user experience evaluation.

Table 2.1: Physiological Measures and Mental Processes (Dirican & Göktürk, 2011a)

and responds to emotional states (Ganglbauer *et al.*, 2009; Kramer, 1990), and is useful when the apparatus used combines eye movement with changes in pupil size.

Physiological measures have several advantages, with one of the most common being that they are more objective compared to behavioural measures. They are also continuous and time-varying, and when combined with behavioural measures Insko (2003), can lead to boundless opportunities for better design decisions in respect to user interface design. Among these attributes obtained for modelling, the best are used for further analysis, since a good model depends on the best features or variables and also on observations or instances that are fit to be modelled.

2.4 Emotional Response to Stimuli

A stimulus event evokes a specific functional reaction in a person's organs or tissue that arouses activity or energy (Mandryk *et al.*, 2006b). To model HCI-HPR relations, the stimuli that evoke certain reactions in a person are also considered as one of the parameters used, although the least squares fit adopted in PHYCOB I is best placed to decide if it is suitable enough to be considered.

A user can respond to different presented stimuli (visual stimulus) and express certain behavioural attributes that evaluators may associate to the users' opinion of the visual contents (Nielsen & Molich, 1990; Vermeeren *et al.*, 2010). Emotions are judged by behaviour (Sauer & Sonderegger, 2009), and can be modelled. For instance, information detection and emotion recognition are major aspects in affective computing, and can allow for automatic identification of stress in users. This can lead to the development of analytical models that contribute to assisting users in browsing complex, interactive websites or user-friendly software applications with less hassle, serving to eliminate stress-inducing interfaces (Davis, 1990; Kolakowska *et al.*, 2013; Vasalou *et al.*, 2004) and in order to accomplish such an ambition a proper comprehension of the browsing behaviour or psychology of the user is needed (Chen *et al.*, 2000; Skadberg & Kimmel, 2004) when interpreting stress.

Research on controlling the stress levels of users has been based on the use of psychophysiological measurements (Andreassi, 2000b; Mandryk *et al.*, 2006a; Mandryk & Atkins, 2007). Stress can be evaluated in different ways, for example areas such as emotion recognition through computational modelling, model stress by either applying classification algorithms (Picard *et al.*, 2001) or a 2-D space defined by arousal and valence (pleasure) (Lang *et al.*, 1993) is used to detect and predict different emotional arousal state.

The large amount of data produced from an experiment in which physiological measures are involved requires quantitative approaches to the detection of levels of arousal in users' emotional response to a given task (Arapakis *et al.*, 2009). Some of these algorithms are important in understanding and predicting the affective state of users for the purpose of designing affective systems that emulates human behaviour. Table 2.2 summaries some of the phycophysiological measures of arousal and the results obtained according to the stimulus elicited. There is some experimental support for the idea of using psychophysiological measurement to identify substantial HCI events. Larger phasic changes and high level of tonic arousal in SCR are produced by strong emotions and ST contributes in correct classification of affect states, when action is involved as emotion elicited (Arapakis *et al.*, 2009; Lang *et al.*, 1993). More novel approaches are introduced in this study to re-evaluate available methods, and to investigate the interpretational problems in order to solve viable issues revolving around HCI-HPR. This will require further analysis, re-examination and testing of existing data on both currently available and novel algorithms in order to advance the process depending on the stimuli used. The following sections discuss webpages as the stimulus used in HCI-HPR associations in this research.

2.5 Webpage Contents as Stimulus

The current acceptance rates of web applications have not reached their full possibilities, due to the fact that sign of modern application interfaces fails to satisfy the user interaction requirements of the target users in a variety of respects (Partarakis *et al.*, 2009).

Measures (Signal)	Stimulus and Evaluation	Emotion elicited	Participants	Model and Data analysis technique	Result
GSR (Lanzetta & Orr, 1986)	vocal tone, electric shock, facial expressions	fear and Happiness	60	ANOVA	larger phasic changes and high level of tonic arousal in SCR are produced by strong emotions
ECG and SCR[(Bailenson *et al.*, 2008)	films	amusement and sadness	2 categories	Chi-square, SVM, logistic regression	NP
ECG, SC, ST, EMG (Von Ahn, 2006)	computer based tasks (games)	Autism	3 categories	SVM	83% accuracy
ECG (Calvo & D'Mello, 2010)	imagination/repeating fearful and neutral sentences	neutral and fear	64	ANOVA and Newman-Keulspairwise comparison	heart rate increases during intense experience than neutral during imaginary experience
ECG, SCR, ST, BVP,FEMG (Sinha, 1996)	iimaginary screen improvement	non-aligned, anxiety, enjoyment, action, desolation and annoyance	27	Discriminant function analysis and ANOVA	99% correct classification of affect states
EKG, SC, EMG (J.A. & Picard, (2005)	self elicitation	self elicitation	8 categories	Fisher Projection with SFFS/KNN	81.25% accuracy with 40 features
EKG, SC, EMG (Calvo *et al.*, 2009)	films	amusement and sadness	2 categories	Chi-square, SVM and logistic regression	none
EKG, SC, EMG (Kim *et al.*, 2004)	audio-visual self evaluation	NP	3 and 4 categories	SVM	78% accuracy for 3 categories and 62% for 4 categories
EKG,EMG, SC, ST,BVP (Haag *et al.*, 2004)	pictures	stress arousal, valence	NP	Dimensional model	90-97% accuracy for valence, 63-90% accuracy for arousal
EEG (Heraz & Frasson, 2007)	pictures	valence, arousal and dominance	17	KNN, DT, Bagging	NP

Table 2.2: Work related to physiological measures of arousal to stimuli

More work is being carried out in this direction to provide the means for developing comprehensive web-based and software interfaces that are capable of adapting to significant end user needs. Some of this work requires the development of adaptive algorithms to learn about changes in user interests or emotions (Widyantoro *et al.*, 1999). A flawless user interface (UI) would automatically adapt or change its layout and, web content elements to suit the needs of the users and similarly, allow for users themselves to alter the contents of the UI (Schneider-Hufschmidt *et al.*, 1993). To model HCI-HPR associations based on webpages, there is a need to understand user adaption to these web contents based on their eye movement and pupil dilation as part of the features to model.

Users easily adapt to the less complex applications, due to the cognitive ability easily to familiarise themselves with - friendly and well-designed webpage interfaces, such as those used as a means of information distribution and learning (Brandt *et al.*, 1999; Cooley *et al.*, 1999; Vigo & Harper, 2011).

The information presented in modern day applications, however, is becoming more and more complex, making access to data harder for users. One reason for this is due to intensive ways of concealing information catalogues in most websites (Verschuere *et al.*, 2011). It takes an above average user to adapt quickly to an application's interface if it is advanced and complex (Hackett *et al.*, 2005). To bridge the gap between the rate of stress levels in average users compared to most experienced users requires synchronised events in HCI-HPR and simulation of the process, which is also investigated.

Visually complex applications change the way users view content (Paulson, 2005). Studies have investigated these issues by conducting usability studies with eye tracking (Jacob & Karn, 2003; Poole & Ball, 2006). This provides supportive information on how users perceive the visual presentation of the content of webpages (Fink *et al.*, 1997; Michailidou *et al.*, 2008). Most of the tasks assigned are normally aimed at determining where the users fix their gaze most or on finding out when they reach a page and how long they pay attention to specific areas of a web page. Reactions from the users can relay quantitative information when physiological sensors are part of the equipment used to study and interpret perception.

Eye movement not only registers fixations and saccades in respect to visual contents but also pupil constriction and dilation. These are physiological responses that not only tells of the changes in light intensity, but also communicate user behaviour and reflect an ongoing mental activity (Andreassi, 2000b; Zander *et al.*, 2010). The response is a measure of interest and emotion (Hess, 1965; Lazarus, 1993). Every eye movement can offer interesting information about behavioural events in terms of the expression of emotional stress in humans and animals (Lang *et al.*, 2000).

The following section discusses the physiological measures in HCI investigations in a attempt to review work related to HCI-HPR.

2.5.1 Physiological Measures as a Metric for HCI Evaluation.

Although there could be both advantages and disadvantages in respect to using psychophysiological measures, they provide promising ways for users in trying to understand areas that seek an intuitive faculty for awareness of things beyond the normal perception of users' psychological changes in HCI (Bannon, 1991; Newell & Card, 1985). Sometimes users may appear to be calm, even though he or she may be concealing a considerable amount of stress. To be able to reveal the level of stress a user as undergone, behavioural coding is sometimes combined with physiological measurements (Fabes *et al.*, 1993; Hazlett, 2006) acquired from data acquisition tools (Fabes et al., 1993; Hazlett, 2006) for combining both physiological and behavioural data can give essential information which will provide an insight to users' behaviours or attitudes towards a given interface. A test participant could be tense towards environmental variables, such as the difficulties that a particular software interface presents to the user, or even the behaviour of another person, or the presence of the evaluator. According to Iqbal et al (2004), the key to understanding these processes is the precise integration of physiological measures such as SC, ST, heart rate, muscle tension, neuronal activity, etc. with behavioural data from the subjects.

We live in a universal computing era where humans are surrounded with embedded computing systems containing various interactive devices (Fogg & Tseng,

1999; Rodden *et al.*, 1998). These advances in technology can bring opportunities as well as difficult issues for users. This is due to the fact that the customary methods in a proposed scheme are not sufficient to tackle the needs of the current day data development devices. Dealing with the data produced by some of these devices has become difficult (Dirican & Göktürk, 2011a). Although advanced cognitive abilities enable humans to handle some of these problems (Compeau & Higgins, 1995; Subrahmanyam & Greenfield, 1994), performance is limited in a real sense and so needs to be critically exploited (Fischer, 2001; Macaulay *et al.*, 2006). For this to be possible, HCI studies have tended to focus on producing computing systems that cause less mental workload/stress, a satisfying user experience, a substitute to users' needs and situations and also, more recently, modelling user physiological attributes with task completion time. In order to meet these aims, there is need for a deep understanding of different dimensions that involves user satisfaction, and research has shown that physiological processes, with certain limitations and weaknesses, have the capability to satisfy users' perceptions (Allanson & Fairclough, 2004).

In addition, physiological measures provide an inconspicuous and comprehensive way to evaluate users' emotional and cognitive state based on mind and body relations. In the actual sense, they provide physical signals of the user interaction generated in response to biological changes and measured by specialised tools in real-time (Figure 2.4).

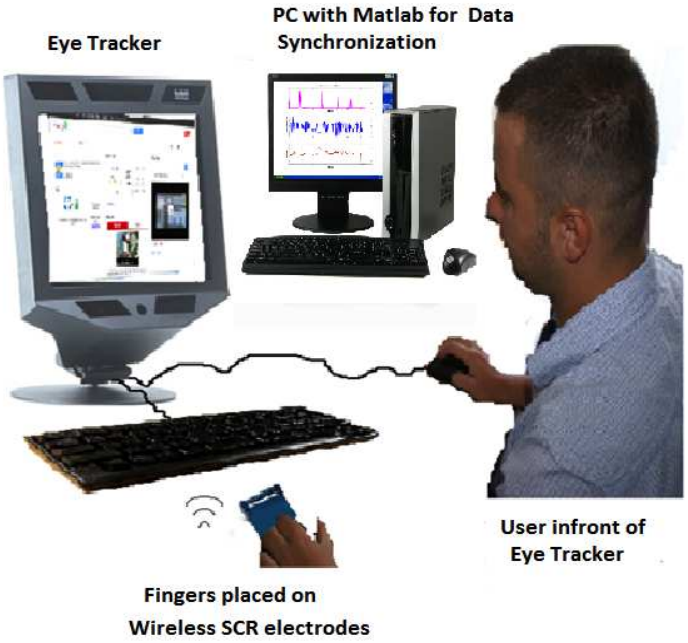


Figure 2.4: Psychological changes measured by specialised apparatus in real-time

2.6 Summary

This chapter has mainly focused on providing a comprehensive review of physiological research related to human cognitive states in HCI. It has discussed the literature concerning physiological measures both in HCI and in other areas. The chapter as also focused on the basics of physiological measures and their relationship with users' cognitive states. A brief introduction concerning the applications of psychophysiological measures in different fields of HCI was also provided.

Chapter 3

Methodology

This chapter discusses the methods adopted for data collection and data analysis. Each section describes one aspect of the process of data acquisition. The initial approach was on HCI usability evaluation testing, based on objective rather than subjective methods. The next approach was modelling the data. The dependent variables (stress levels) referred to in this chapter are the users' affect state and behaviour while the independent variables are the stimulus (webpages) and user attributes from the sensors.

3.1 Methodology description

This section discusses how data will be collected, this will serve as an input to the proposed algorithm (PHYCOB I), it also introduces the predictive models which will be used for comparison and validation purposes.

The participants to be recruited for the pilot study will involve ten students, more participants will be recuited for additional user attributes and data generalisation. The SCR, ST and eye-tracking data were the major data sources that will be captured for data exploration. These data will be systematically synchronised by linking the timestamps of sensors.

The proposed algorithm will involve two modules: the first module will generate the user attributes from the sensors used to measure the physiological response and eye movement of the users, the second module will make predictions on the

users' stress levels from the captured user attributes based on a control system for modelling physiological process of users in reaction to web stimuli.

Methods such as the Neural network and Logistic regression are both predictive models which will be used for comparison. Other methods such as PCA and FS algorithm uses a different form of pattern recognition procedure distinct from the predictive models and will be used as a form of validation technique, for variable reduction and to elicit individual stress level and their corresponding observation. The process for these methods will be discussed in this Chapter and thier comparision with the proposed algorithm will be discribed in Chapter Four.

The following sections discusses the experimental setup, design and implementation of the project work. The steps taken to achieve this is given as follows:

- Experimental setup
- Algorithm Development (Proposed Algorithm)
- Pattern Identification
- Variable Selection
- Model Validation

The initial stage is designing a pilot study for the experiment. This will involve the number of participants to consider to take part, for this, ten participants will be recruited; four female and six male students. The participant eligibility will be based on age restriction and disability. The subjects will be healthy and mature enough to take part. The main study will then involve more participants; participants' involvement will be based mainly on consent and willingness to take part. The experiment will be based on non-intrusive and non-obstructive measures. The equipments to be used includes SCR sensor that measures ST and TOBII eye tracker. The procedure and how this was carried out is discussed in the following sections.

3.2 Experimental Setup: Participants and Equipment

Before the data collection commences, the experiment was assigned reference number CS77, which was approved by the University of Manchester Senate Committee on the ethics of research on human beings.

The pilot study involved ten student participants age between 18-48. Thirty-four participants were later recruited for more data and to generalise participants' involvement; this involved twelve female, twenty-twenty male from workers and regular users of the web, making a total of forty four participants including participants from pilot study age between 18-48 and above. These were recruited through advertisement and recommendations from the University of Manchester. The same procedure used to carry out the experiment for the pilot study is also used for the main experiment.The study took less than 10 minutes. Appendix A.5 shows the information sheet (A.5) and consent form provided for the participants.

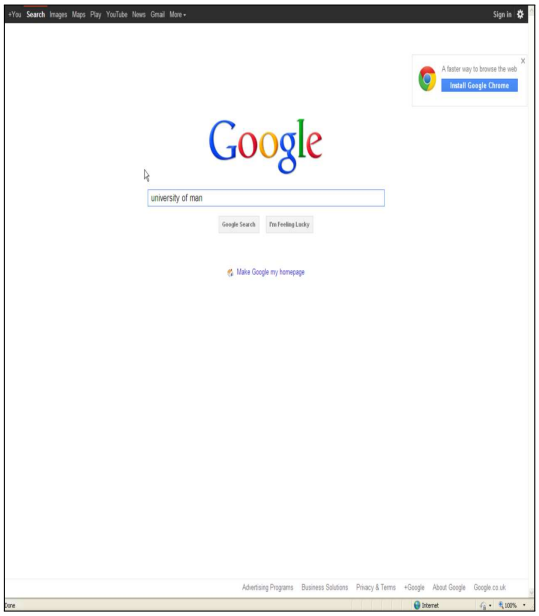
3.2.1 Experimental procedure

Participants recruited for the study were seated, each facing a TOBII 1750 eye tracker. The webpage data and users' eye movements which includes fixations, saccades and pupil dilation were recorded. The participants placed their two middle right fingers or wrist on a wireless SCR Q-sensor (Figure 2.4), leaving the right hand free to perform tasks such as keystrokes and mouse manipulation. The Q-sensor also has the ability to measure the skin temperature in degrees Celsius (O^C). Analysis software embedded with the eye-tracker was used to record the eye movements and fixations. The recorded data was later exported to PHYCOB I for further analysis.

3.2.2 Task

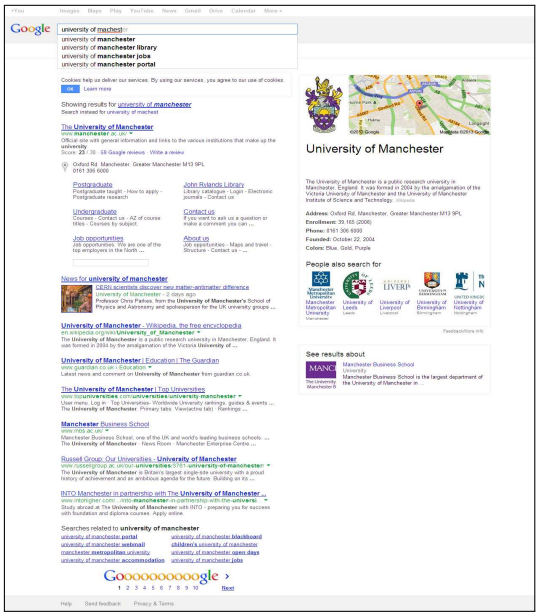
The tasks in the study were designed in the manner that users were accustomed to, this included typing words into text-box content and clicking on icons. The tasks were designed in such a way that participants interacted with static and dynamic web contents by completing six straightforward tasks, each of which was

designed to encourage interaction with an element. Some of these elements were represented as static information, while others were dynamic. The task "search" encouraged the users to interact with static contents, while the task "suggest" encouraged interaction with dynamic contents such as automated lists (ASL). These contents automatically appeared as the users typed, thus assisting them to reach their goal without any stress.

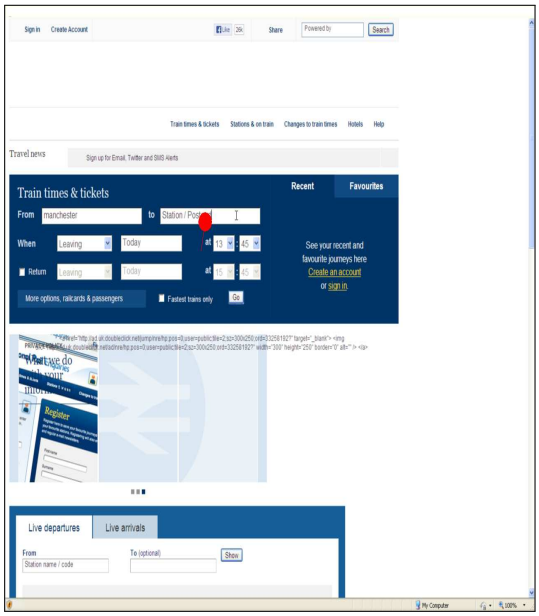


(a) Google Search, with ASL disabled.

(b) Yahoo Portal



(c) Google Suggest with ASL enabled.



(d) National Railway Enquiries search with ASL disabled.

Figure 3.1: Live Websites with Google search, Yahoo Portal, ASL enabled for Google and disabled for National Railway Enquiries.

The aim of the study was to capture users, responses to web stimuli, such as the areas or content of the page where users fixations are mostly seen, correlating to spikes in their SCR, a slight increase in ST and variations in pupil dilation. A short-term interval was used to represent each interactive session on each webpage. The web stimulus includes contents in the National-Rail-Enquiries-Suggest, Yahoo, Google-Search, National-Rail-Enquiries-Search, Google-Suggest and iGoogle pages. The webpages were live websites directly from the Internet, the "Webhp" for Google Search browser was assigned to zero to disable the search textbox while the National Rail Enquiries Search was made static by disabling the search textbox using the AJAX tool (Asynchronous Java and XML) for creating and updating webpages without reloading pages. The webpages were accessed from an external storage device. Figure 3.1 and Figure 3.2 shows the websites used in the study. The tasks that participants performed on the websites, are illustrated in Table 3.1.

(a) iGoogle Search

(b) National Railway Enquiries suggest, with ASL enabled

Figure 3.2: Live Websites with iGoogle search and National Railway Enquires suggest pages with ASL enabled.

Table 3.1: *Table indicating the task allocated to each webpage.*

Stimulus	Task
Google-Search	"Locate Manchester University".
Google-Suggest	"Locate Manchester University".
National-Rail-Enquiries-Search	"Look for a train-route from London to Manchester".
Nationa-Rail-Enquiries-Suggest	"Look for a train-route from London to Manchester".
iGoogle	*"On the CNN.com box, locate news stories".*
	Read the displayed text contents
Yahoo Portal	*"Locate the entertainment, sports, news or stories".*
	"Read the displayed text contents".

The Google-search page[1] is shown in Figure 3.1a, and was designed to see if there could be stress points detected in all task allocated areas or any one of those areas. The Google-suggest[2] provided an automated list on the search engine as the users typed in the text-box content. This task allowed us to observe whether users found the ASLs useful and whether it reduced their stress level.

The National-Rail-Enquiries-Search[3] task in Figure 3.1d is similar to that assigned to Figure 3.1a, but the inputs were more constrained. In the search on this static page, users were limited to direct keywords i.e. searching for a train route from **Manchester** to **London**. The purpose was to focus on the amount of cognitive load placed on users as they typed and to observe the affect on their stress level with ASL disabled during real time and in the algorithm development phase.

[1] www.google.com/webhp?complete=0
[2] www.google.com
[3] www.nationalrail.com

The National-Rail-Enquiries-Suggest task in Figure 3.2b provides suggestions to users by automatically displaying ASL as they typed. In the Google-Suggest tasks, suggestions were made based on popular searches that were refined as the user typed. This page is similar to National-Rail-Enquiries-Search in Figure 3.1d but more dynamic. This allowed for comparison of stress levels detected on the pages.

The task given for the iGoogle-page[1] and Yahoo page[2] in Figure 3.2a and Figure 3.1b were similar to simple browsing, this is to detect if the different dynamic contents found on these pages had any effect on the physiological readings taken or the detected spikes in their SCR results, and whether these increased in amplitude and magnitude in reaction to the presence of the dynamic content.

The participants were asked to read the information sheet and consent form in Appendix A.5 before the experiment. The users commenced with the index page (Figure 3.3), with links to the task allocated webpages for a total time of less than 10 miniutes; interaction with each page was less than 120 seconds. Data was collected objectively without interrupting data collection, and exported to a spreadsheet for analysis in MATLAB with PHYCOB I.

[1] www.google.com/ig
[2] www.yahoo.com

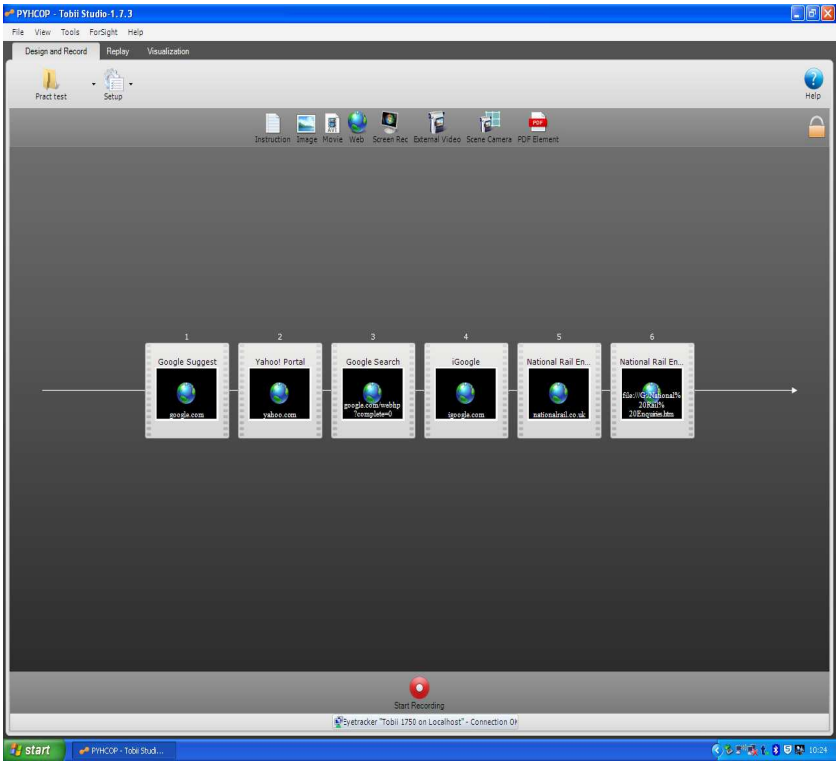


Figure 3.3: Index page of stimulus on Tobii eye-tracker.

The data collection was approached in two ways, *real time* and *delayed*. In real time we were able to investigate and observe the spikes in SCR, variations in PD and increases in ST while users interacted with elements on the webpages. The delayed analysis enabled us to access each session of the experiment, revisit, extract time sessions and register delay in the latency of the physiological readings that correspond to the latency of the eye tracker and for further analysis. This helps to identify a particular instance in time that might have been missed. Most of the behaviourial data such as "Looking", "Interest"and "Clicking", are logged into eye tracker during real time and also during the delayed analysis. Using this method, we were able to obtain real-time feedback without interrupting the collection of data. The quantitative data was exported to a spreadsheet; participants generated six instances each. Short interval processing was adopted for the synchronised HCI-HPR data.

High dimensional datasets were generated using PHYCOB I, which detects

increases in stress level based on the average amplitudes response detected for each webpage and the response duration. The attributes computed were also used to envisage areas of interest and eye movement behaviour of participants (Figure 3.5).

3.3 Proposed Algorithm (PHYCOB I)

The proposed algorithm consist of two modules; the first module computes the user attributes while the second module make predictions from the user attributes.

The steps for the algorithm are stated in Appendix B.3. From the algorithm, to compute the user attributes, X is set as a place holder for the physiological data from the SCR/ST sensor and also PD which is from the eye tracker while Y is the place holder for the eye movement data from the Eye tracker sensor. The main aim is to generate a dataset $Z_{m,p}$ with m instances and p number of attributes. The first step is to set the initial conditions; each participant $i = 1 : 44$ interacts with each webpages $j = 1 : 6$ from which the user attributes are computed and used to update the matrix $Z_{m,p}$ that serve as the secondary data (Figure 4.1). Each user attribute generated uses the correlates of optimal responses to classify the status of participants. The term "correlates" here represents events from sensor and eye tracker that occured at time of optimal response (peaks) of SCR.

To execute the steps in the algorithm and obtain users' attribute, each participant's generated data was considered based on differences in the baseline (b_i). For each person, the baseline is different, thus increases in amplitude (a_i) is computed based on a set threshold given in the algorithm (Appendix B.3). Since the baseline (b_i) for individual users are different, the latency for response time are particularly distinct. The average latency is determined by calculating each delay in a users' SCR's amplitude, which involves taking the time readings at points corresponding to minimum index of high tonic phases of the response signal. This determine the delay for each increase in amplitude of SCR, indicated by the points at the red dashed lines in Figure 3.4.

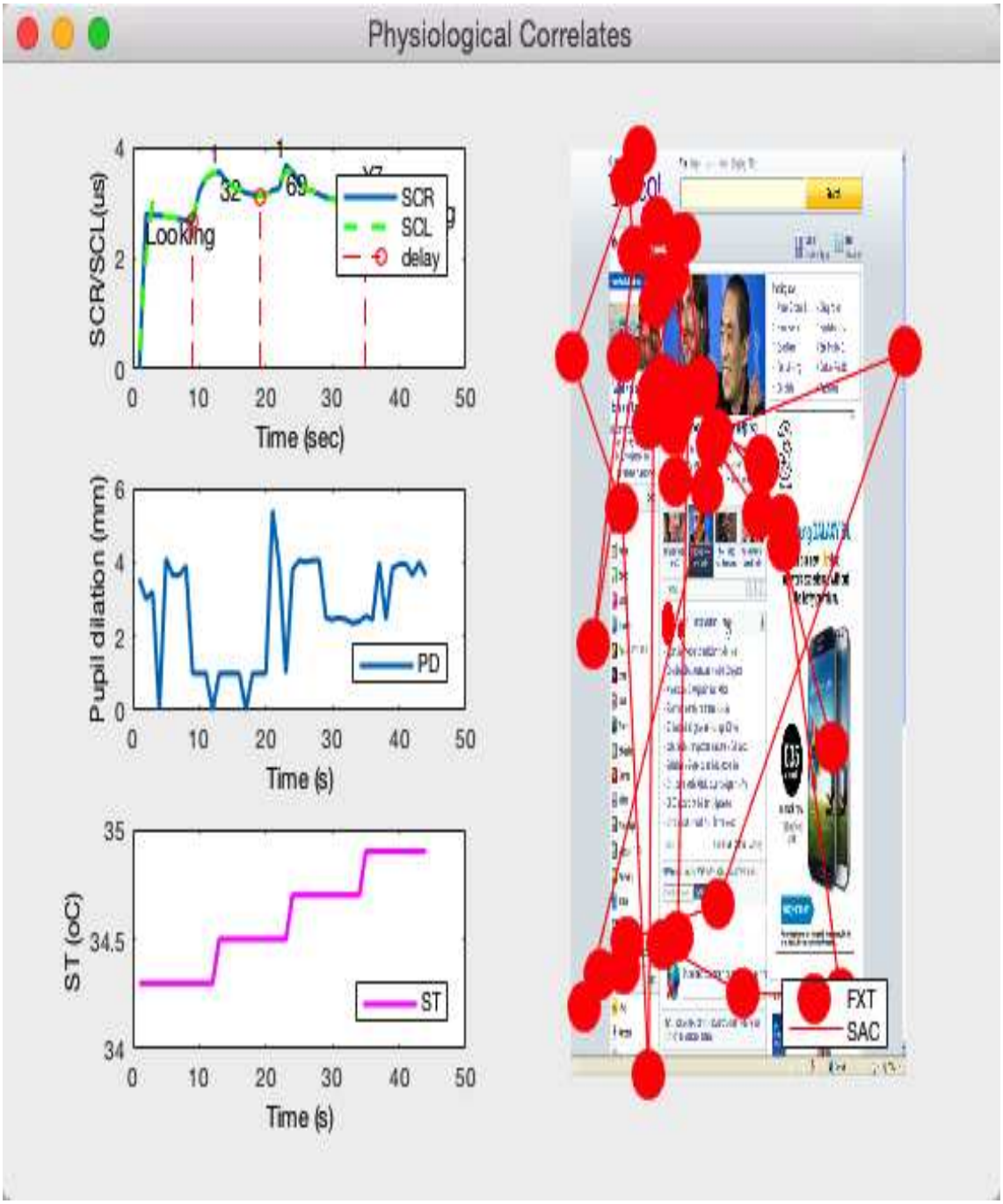


Figure 3.4: Physiological measures with computed delay in SCR in sync with eye movement on webpages

Events correlating to fixation points (XX) and the fixation duration (FD) are extracted. The saccade size (Y_i) was calculated based on the Euclidean distances (Equation 3.5) of mapped fixation x (MPFX) and mapped fixation y (MPFY) from another fixation point.The output is a matrix $\mathcal{Z}_{m,p}$ with m the number of instances or rows and p the number of attributes or column. Data for each participant were based on the number of webpages they viewed; in this case, each participants interacted with six webpage. Other attributes such as the magnitude response (mg_i) and the mean skin temperature (MST) were based on the length of the SCR to each webpage and the average ST of participants during the interaction.

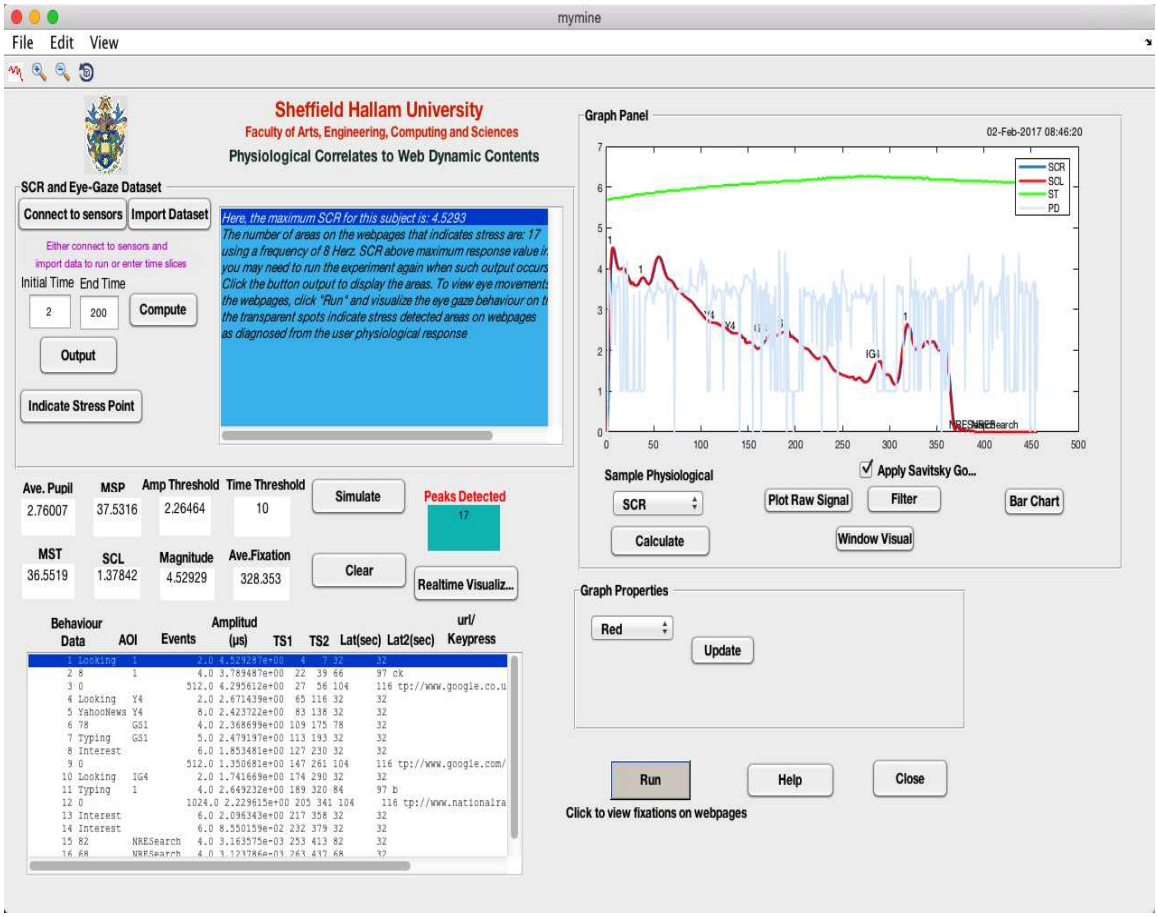


Figure 3.5: Index page showing detected peaks and events corresponding to spikes in physiological responses

3.4 Detecting dynamic contents and Simulation of user attributes

Variables were considered based on significant attributes often mentioned in the literature when considering a multimodal approach, such as combining eye tracking data and physiological measures (Bergstrom & Schall, 2014). The chart in Figure 3.8 shows the breakdown of the attributes. MST is the mean skin temperature measurement of (skin temp1 and skin temp2) measured over time.

The secondary datasets containing users' attributes generated from the sensors was used for the simulation. This process is discussed in this section. Firstly, we discussed process involved in detecting dynamic contents and its physiological correlates.

3.4.1 Detecting dynamic contents

To detect a dynamic content on the webpages, each data point on the user physiological readings (SCR) from the primary data source (sensors) is either termed as stressed, neutral or a relaxed point based on a given threshold:

$$thresh = 0.5(\text{average amplitude} - \text{minimum SCR})$$

this value is chosen because of the differences in baseline level of participants.

If the given mean peak is greater than the threshold and the time interval for a point is greater than 3 seconds (normal time interval for response to appear) (Filipovic & Andreassi, 2001) the participant is characterised as stressed. If the mean peak response is less than the threshold and the threshold is greater than a high level tonic point, the participant is in a neutral mood otherwise the participant is relaxed. The high level tonic point is the optimal response point for a baseline level in the physiological measure (Figure 3.6 and 3.4). For a particular webpage, the highest frequency of affect states in a given interaction is used as the label for the status of a participant.

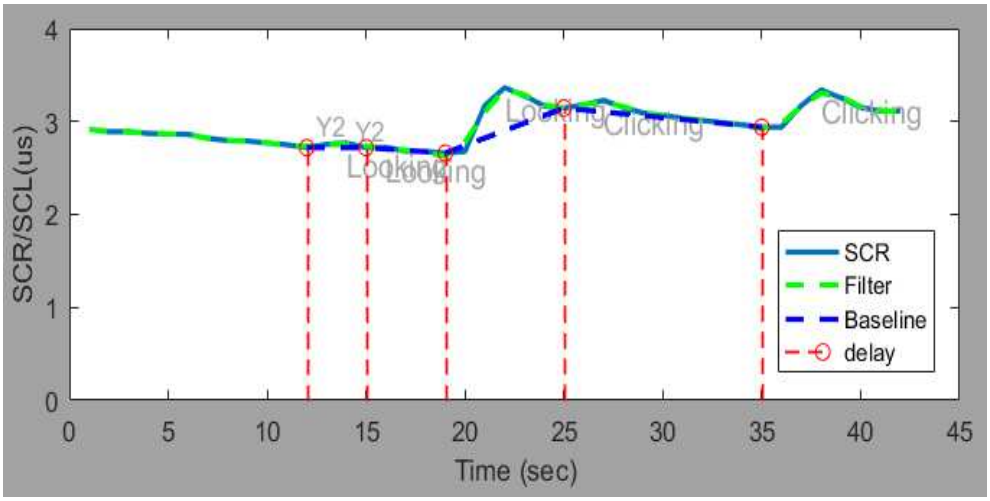
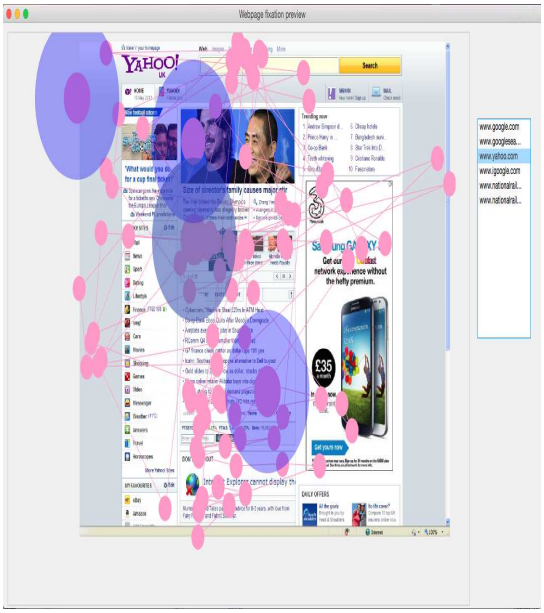


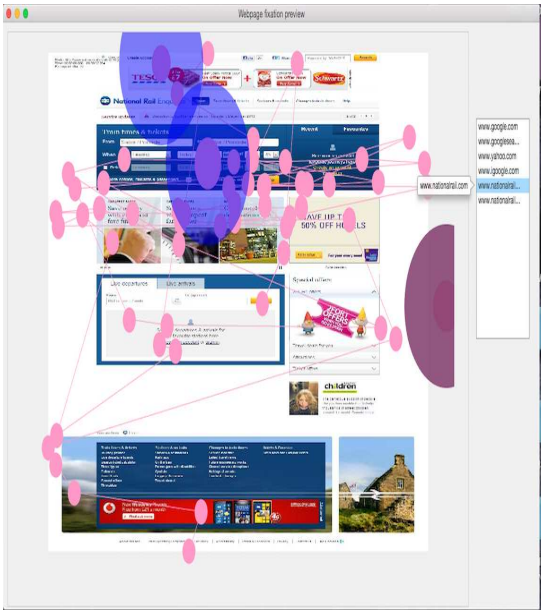
Figure 3.6: Detected peaks, correlating AOI and user behaviour

The predicted affect state obtained for the PHYCOB I model fit (Eq 3.6) of the users is mapped to the webpages and are seen on the areas (AOI) correlating to the task allocated positions when we run the process or click on "simulate" button on the interface (Figure 3.5). These are possible positions where dynamic contents are detected that are responsible for spikes in physiological response and changes in stress levels. Example of such occurrence is shown in Figure 3.7. For instance a user response appeared while "looking" at AOI labelled "Y2" on the yahoo and increases/decreases in amplitude while "Clicking"(Figure 3.6). Control can be directed from an eye tracker to PHYCOB I, that reruns the eye movement on the webpages and detects the stress point to identify a web content.

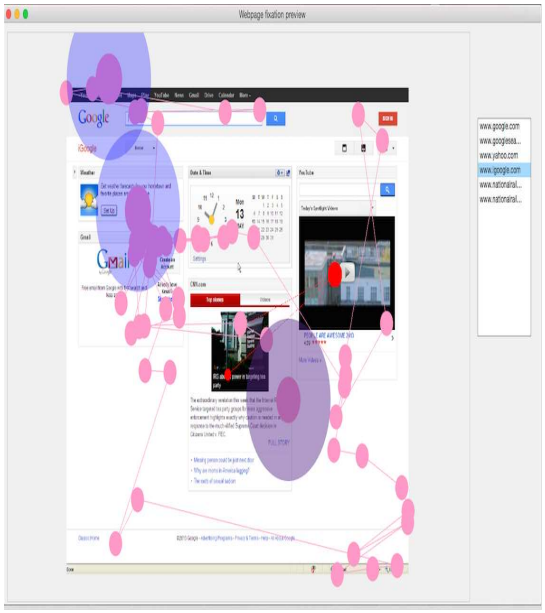


(a) Yahoo page with detected stress points

(b) Google page with detected stress points and a relaxed point



(c) National rail enquiry page with detected stress points and a neutral point



(d) iGoogle page with detected stress points and a neutral point

Figure 3.7: Areas on webpages with detected stress points of users.

3.4.2 Simulation of users' attributes

The input matrix are the physiological data X and eye movement data Y, the output resulted into the secondary dataset:

$$\mathcal{Z}_{m,p} = \begin{pmatrix} Z_{i,j} & \dots & Z_{1,p} \\ \vdots & \ddots & \vdots \\ Z_{n,1} & \dots & Z_{n,p} \end{pmatrix} \text{ and } Z_{i,j} = Z_{1,1} \tag{3.1}$$

where primary data:

$$\begin{pmatrix} X_{i,j} & \dots & X_{1,p} \\ \vdots & \ddots & \vdots \\ X_{m,1} & \dots & X_{m,p} \end{pmatrix} \times \begin{pmatrix} Y_{i,j} & \dots & Y_{1,p} \\ \vdots & \ddots & \vdots \\ Y_{m,1} & \dots & Y_{m,p} \end{pmatrix} = \mathcal{Z}_{m,p}, \;\; X_{i,j}, Y_{i,j} = X_{1,1}, Y_{1,1} \tag{3.2}$$

This is developed for the purpose of modelling, the length m differs in all datasets, since some participants took less than the time given to complete the tasks.

Two thousand and fifty instances of these were simulated based on the covariance ($C = cov(\mathcal{Z}_{m,p})$ of the original dataset (M1) by using a union of random numbers $M_i = U([0,1])(m * p)$; the number of rows is equal to the dimension of the cholesky decommposition (Harbrecht *et al.*, 2012; Yu *et al.*, 2009) of the covariance matrix C such that outcomes $M_i, \dots M5$ closely match the original dataset Z (Algorithm 1 with some R syntax).

Data: input dataset$Z_{m,p}$; no. of instances m; no. of attributes p;
Result: outpu matrix M_i
n = 2500 ▷ no. of instance to generate
if (is.matrix($Z_{m,p}$)) **then**
$M_1 = Z_{m,p}$
while $n = 2500$ **do**
  for $i = 2 : 5$ **do**
    $n \leftarrow solve(n)$
    $C = cov(Z_{m,p})$ ▷ covariance of matrix
    $L = chol(C)$ ▷ cholesky decomposition of covariance
    $vars = dim(L)[1]$ ▷ dimension of decomposed matrix
    $t = t(L)$ ▷ transpose decomposed matrix
    $R = t(t)\% * \%matrix(rnorm(vars * n), nrow = vars, ncol = n)$ ▷ generate new matrix from random numbers
    $R = t(R)$ ▷ transpose new matrix
    $M_i = matrix(R)$ ▷ generate new datasets
    goto loop.
    close;
    goto top
  end
end

Algorithm 1: Simulation of users' attributes

By observing and making predictions from this, stress areas on webpages are detected (Figure 3.7) this gains more insight into HCI-HPR associations, such as the contents that caused stress reaction.

The stressed, neutral and relaxed mood are indicated by the transparent blue, purple, and red spot on the webpages e.g. a participant experienced stress emotion while looking at ASL on google page, national rail enquiry page and picture content on yahoo page; a neutral mood is seen on two pages while a relaxed mood is on a google page.

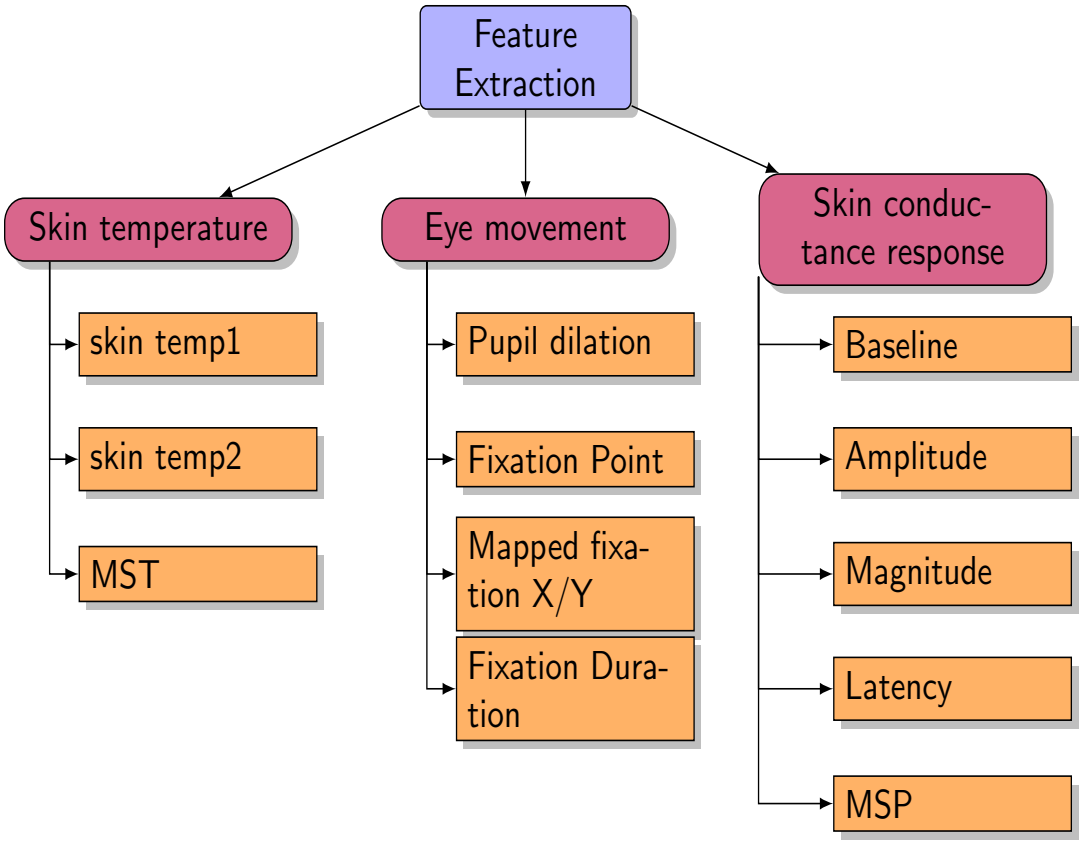


Figure 3.8: Feature extraction breakdown

3.4.3 Application of filters to physiological responses

The Savitsky filter (Savitzky & Golay, 1964) (Eq 3.3) was applied for removal of noise and other artefacts on the physiological measures. The baseline (SCL) was estimated based on the point interpolation moving average technique (Eq 3.4). Optimal response on the SCR were detected based on a given threshold that corresponds to a participant's response at onset and half recovery time of SCR.

$$(\mathbb{X}_k) = \frac{\sum_{i=-n}^{n} \mathbb{X}_k}{\sum_{i=n}^{n} A_i} \tag{3.3}$$

k are the data points of the physiological measures

$\mathbb{X}_{k+1} = A_0 + A_1 \text{ x } + A_2\ x^2 + \ldots + A_n\ x^n$

A_i = Coefficients (Scalers)

x_i = Scaler Variables

This is applied to the physiological signal $\mathbb{P}_k$ such that:

$$\mathbb{X}_k = T_k(\frac{d^{2n+1}\mathbb{P}_k}{dt^{2n+1}}) \tag{3.4}$$

$\mathbb{X}_k$ is the resulted datapoints by resampling the raw physiological signal $\mathbb{P}_k$, taking a window size or polynomial order of $2n+1$ in $\mathbb{P}_k$, for each time interval T_k. Accordingly, raw signal still maintained its shape, with clear peaks. Each physiological measure undergoes this process depending on how noisy the data is. The eye movement data obtained includes the PD and fixations captured by the eye tracker. The derived variable is the saccade size $\mathcal{D}$ that gives the euclidean distance between two fixation points (x_n, y_n) and (x_m, y_m):

$$\mathcal{D} = ((x_1 - y_1)^2 + ... + (x_m - y_m)^2)^{0.5} \tag{3.5}$$

where x_n, y_n are fixation points on the vertical plain of a webpage and x_m, y_m are the fixations on the horizontal plain. The synchronised data are exported and written to a spreadsheet (Figure 3.9).

If the SCR falls below the median range, the participant is at the tonic phase and hence nothing is actually happening, he/she is relaxed and a high tonic phase point is detected. Once there is an increase in amplitude that exceeds the threshold level, the participants is considered stressed. Given the nature of the task, stress is the most predicted outcome of the affective state detected at the peaks. Hence we integrate between "Stress", "Neutral", and "Relaxed" state of the users, looking to minimise the effect of stress. The SCR is the physiological measure that serves as the major constant response for this case, given its ability for detecting spontaneous and evoked reaction, as shown in the literature regime (Baumgartner *et al.*, 2006; Mavratzakis *et al.*, 2016). PHYCOB I detects each point of increase in amplitude and detects the maximum SCR (peak) within a certain interval that correspond to a particular event.

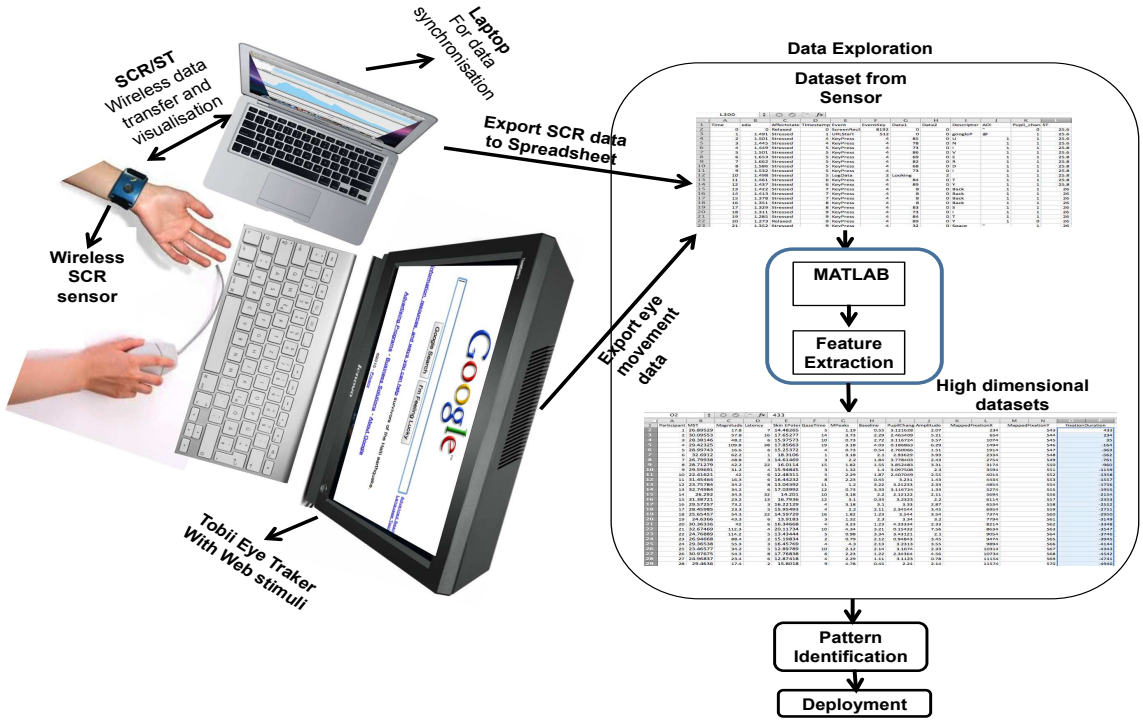


Figure 3.9: Graphical representation of method description

3.4.4 The PHYCOB model fit

To get a significant accuracy and a prediction focus close to the original class label, the model adopts the concept of physical processes on dynamic systems for modelling the physiological processes; a least squares technique applied to control systems. The final dataset $\mathcal{Z}_{m,p} = M_i$ is used for the modelling. For user attributes on data saved from the sensors (Figure 3.10), the entire system is represented by the expression in Eq 3.6, this is the PHYCOB I model fit to the data with a default prediction focus four minutes from the originally stress levels of the user (class labels).

$$\frac{dx}{dt} = ax(t) + bu(t) \tag{3.6}$$
$$y(t) = cx(t) + du(t) + 4k$$

where y(t) is the response variables (stress levels) that determine coefficients of

physiological reactions with computed variables $u(t)$ which represents the data matrix $\mathcal{Z}_{m,p}$, each input variable has p-values less than the default critical point (0.05). This point is chosen because the average p-values of the user attributes is cIose to this critical value. The model adopts a discrete time identified model fit representing the set of values the physiological reaction processes can take in response to dynamic contents of the web. The primary data sources (Figure 3.10) are measured in frequency (Herz) and contains both categorical and numeric variables that contributes to making predictions on the single response output using multiple inputs (MISO). The model is tested on the different polynomial order between 1 to 10 of the state space x(t). The best performance from these sets are selected.

An identity state space object is created with a as the state space matrix, that is, the possible set of values the process can take, b the input matrix containing the variables, c the output matrix, d the transformation matrix and a disturbance matrix k. Also y(t) is used to represent other response variables like the eye movement (fixations on webpages). The performance from the predictions is compared to other predictive models.

	EDA	Skin Temperature	Affectstate	Timestamp	Event	EventKey	Data1	Data2	Descriptor	AOI	Pupil Change	FixationIndex	Timestamp	FixationDuration	M
19	1.3000	34.6000	Stressed	13	LogData	2	Looking			1	3.3200	19	9879	259	
20	1.3287	34.6000	Stressed	14	LogData	2	Looking			1	1	20	10112	219	
21	1.3163	34.6000	Stressed	15	LogData	2	Looking			1	1	21	10331	239	
22	1.2994	34.6000	Stressed	15	KeyPress	4	85	0	U	1	4.1800	22	10571	140	
23	1.2794	34.6000	Stressed	15	KeyPress	4	78	0	N	1	1	23	10710	279	
24	1.2581	34.6000	Stressed	16	KeyPress	4	73	0	I	1	1	24	10989	140	
25	1.2374	34.6000	Stressed	17	LogData	5	Typing			1	1	25	11986	140	
26	1.2191	34.6000	Stressed	17	LogData	5	Typing			1	1	26	12126	120	
27	1.2046	34.9000	Stressed	18	LogData	5	Typing			1	1	27	15774	179	
28	1.2287	34.6000	Stressed	18	KeyPress	4	73	0	I	1	1	28	15954	179	
29	1.2047	34.9000	Stressed	18	LogData	5	Typing			1	3.3700	29	16133	538	
30	1.2022	34.9000	Stressed	19	KeyPress	4	86	0	V	1	1	30	19263	179	
31	1.1970	34.9000	Stressed	19	LogData	5	Typing			1	1	31	19443	379	
32	1.1946	34.9000	Stressed	19	KeyPress	4	69	0	E	1	1	32	19822	159	
33	1.1970	34.9000	Stressed	19	KeyPress	4	82	0	R	1	1	33	19981	279	
34	1.2055	34.9000	Stressed	19	KeyPress	4	83	0	S	1	1	34	20420	20	
35	1.2180	34.9000	Stressed	20	LogData	5	Typing			1	3.4600	35	22752	100	
36	1.2330	34.9000	Stressed	20	LogData	5	Typing			1	1	36	22852	139	
37	1.2479	34.9000	Stressed	21	KeyPress	4	85	0	U	1	1	37	22991	797	
38	1.2588	34.9000	Stressed	22	LogData	5	Typing			1	1	38	23789	339	
39	1.2632	34.9000	Stressed	22	KeyPress	4	8	0	Back	1	1	39	24128	199	
40	1.2600	34.9000	Stressed	22	KeyPress	4	73	0	I	1	1	40	24328	797	
41	1.2502	35.1000	Stressed	22	KeyPress	4	84	0	T	1	3.3300	41	25125	200	
42	1.2372	35.1000	Stressed	23	KeyPress	4	89	0	Y	1	3.2200	42	25324	1077	
43	1.2234	35.1000	Stressed	23	LogData	5	Typing			1	3.3700	43	26401	598	
44	1.2119	35.1000	Stressed	23	LogData	5	Typing			1	3.4500	44	26999	259	
45	1.2061	35.1000	Stressed	24	KeyPress	4	40	0	Down	1	3.5100	45	27258	239	
46	1.2075	35.1000	Stressed	24	LogData	5	Typing			1	3.5800	46	27498	239	
47	1.2181	35.1000	Stressed	25	KeyPress	4	40	0	Down	1	3.4000	47	27737	199	
48	1.2379	35.1000	Stressed	25	KeyPress	4	40	0	Down	1	3.4600	48	27936	239	
49	1.2657	35.1000	Stressed	25	KeyPress	4	40	0	Down	1	3.2100	49	28175	239	
50	1.2997	35.1000	Stressed	26	KeyPress	4	38	0	Up	1	3.1200	50	28414	359	
51	1.3349	35.1000	Stressed	26	LogData	5	Typing			1	3.4300	51	28774	398	
52	1.3683	35.3000	Stressed	26	KeyPress	4	13	0	Return		3.3300	52	29172	239	
53	1.3960	35.3000	Stressed	26	URLEnd	1024	0	0	http://www.g...		3.5100	53	29411	219	
54	1.4166	35.3000	Stressed	26	URLStart	512	0	0	http://www.g...		3.6900	54	29631	359	
55	1.4286	35.3000	Stressed	27	URLEnd	1024	0	0	http://www.g...		3.5300	55	29990	259	
56	1.4319	35.3000	Stressed	27	URLStart	512	0	0	http://www.g...		3.4600	56	30249	379	

Figure 3.10: Recorded primary data from SCR and eye tracker sensor

3.5 Data exploration

The data exploration stage involves an instructive search used in modelling the data, this is to form a true analysis from the information collected. Data are assembled in a controlled manner in large quantity. For true analysis to be achieved, the disorganised volume of data needs to be constricted to obtain accuracy.

Two main techniques were employed to retrieve significant data from large datasets with disorganised classes (Kaski, 1997; Zuur *et al.*, 2010): the manual and automatic methods. The manual method is mostly the data exploration, for which Principle Component Analysis (PCA) and FS were employed; while the automatic method involves data mining or use of predictive modelling. For both cases, we used the R package for analysis, because of its capabilities and environment for statistical computing and graphics that provides a wide variety of techniques for linear, nonlinear and time series analysis.

The PHYCOB tool was used initially to specify the class values of each of the instances in our dataset based on the peaks detected or increases in amplitude, mean skin temperature (MST) and fixations. The categorical data were not included in the first stage, a normally distributed response parameter was simulated. In order to determine whether or not the data could be modelled and also to identify the instances that could be labelled, PCA and FS algorithms were adopted, with the residuals visualised using stalactite plots (Atkinson, 1994; Mwitondi & Said, 2011). The steps for the FS were based on the Mahalanobis distances illustrated in Appendix B.2, Algorithm 1. This was used to identify the natural structures the dataset represents. Compared to Atkinson's, the initial stage for the first principle method is applying the least median square (LMS) of the linear model to evaluate the attributes by solving the non-linear reduction task.

3.6 Pattern Identification

Patterns were identified based on the statistical information extracted from the datasets. This involved grouping the measurements or observations collected from participants and defining them in an appropriate multidimensional space.

This section describes the process adopted in classifying the dataset based on the statistical information or structural patterns we extracted to form the patterns recognised. The resulting best structures are then used for labelling based on the statistical regularity of the patterns. The structure identified was based on the structural interrelationships of the variables of the datasets. The stalactite plot was used to determine if any known patterns exist in the dataset: the plot was used to determine the outlying cases from dataset consisting of 264 instances, capturing 15 attributes/features as described in Appendix B.1 and to test performace with predictive models, datasets with more than a thousand instance was also simulated from the original data. For the FS, 164 instances were used as a training set for the model. The remaining instances (2^{nd} dataset) served as a form of test data for the linear model fit. The two processes employed for pattern identification were, first, detecting the variability of the components using PCA and then, detecting outliers with masking effect using stalactite plot; fitting a linear model to determine the best fit. These outlying cases were then removed and the model was fitted again to test if this improved or reduced the performance of the model. The section below describes these processes.

3.6.1 Forward Search (FS) Algorithm for Identifying Natural Structures

To determine the natural structures, a stalactite plot was used to represent the output of the FS algorithm. This was used as a supplementary tool to identify the natural structures and also used to determine outlier cases or observations in the dataset. It is based on Mahalanobis distances Eq 3.7 (Mahalanobis, 1936) computed from the covariance and mean of the estimated subset of the dataset.

$$D_m(x) = (x - \alpha)^T \, S^{-1}(x - \alpha)^{0.5} \tag{3.7}$$

where $x = (x_1, x_2, \ldots, x_N)^T$ set of observations, mean $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_N)^T$ and covariance matrix S.

The goal is mostly to decrease the hidden outliers that arise from the data sample. The rationale behind this is that given a set of m observations for estimation, to determine mean and covariance, the next observation $m + 1$ to be

used for assessment for the next step are probably the $m + 1$ negligible spaces. Observations were included in the subset used for valuation, say for a value m, and as m increases, these observations were excluded.

The output plot vividly demonstrates the progress of the set of outliers on the set of observations. Originally m was usually chosen as $q + 1$ where q is the number of attributes, i.e. the smallest number allowed for computing the Mahalnobis distances. The cutoff point commonly used for defining outliers in FS algorithm is the expected value of n size random attributes having a q degrees of freedom on the chi-squared distribution of the set of attributes, which is given approximately as the threshold:

$$threshold = X_P((n - 0.5)/n, p) \tag{3.8}$$

To model the relationship between the dependent simulated variable and the predictor variables (physiological parameters), it is necessary to assume that the relationship between the dependent simulated data and the physiological parameters of the in linear model are normally distributed. The relationship is modelled through unobserved random variables which contribute to noise in the linear association among the simulated variables and the regressors, which takes the form:

$$y_i = \beta_1 x_{i1} + \cdots + \beta_n x_{in} + \varepsilon_i = {x_i}^T \beta + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3.9}$$

where T is the transpose, such that $x_j^T \beta$ is the inner product of x_j and β, of a linear equation.

The following section discusses the main processes for feature selection.

3.7 Variable Selection

The purpose of variable selection is to select the best group of variables or subset of predictors that explains the data in a simplified way, i.e., redundant predictors were removed. This reduces the effect of outlying cases, reduces overfitting and can affect model performance.

Among several reasonable solutions for a phenomenon, the simplest is best (Posada & Crandall, 2001; Scott, 2015). Relating this to regression model, this

indicates that the most simplified model that fits the data is the most suitable for modelling. Redundant predictors may add noise to the estimations of other outlying cases we were interested in (Hector *et al.*, 2015; Khan *et al.*, 2007). Also collinearity among predictors causes too many variables trying to do the same job (Hector *et al.*, 2015; van Havre *et al.*, 2015); this is mostly referred to as overfitting. For example, the amplitude of a response in SCR is the same as a peak in the SCR, since both register a significant response to stimuli (Bach *et al.*, 2010). For such a case one of the two will have to be excluded.

Prior to variable selection, we defined outlying cases and influential points and the masking effect with the stalactite plot. Exclusion of these cases was done temporarily and other transformations such as fitting both standardised and non-standardised form of the data were carried out to validate and justify the best data model for the methods.

When selecting variables, one important issue that needs to be adhered to is respecting the hierarchy of data. This is important, since some models have a natural hierarchy (Baayen *et al.*, 2008; Friston, 2005). The summarised hierarchical model is given in Figure 3.11, with the lowest level of user stress starting at the top.

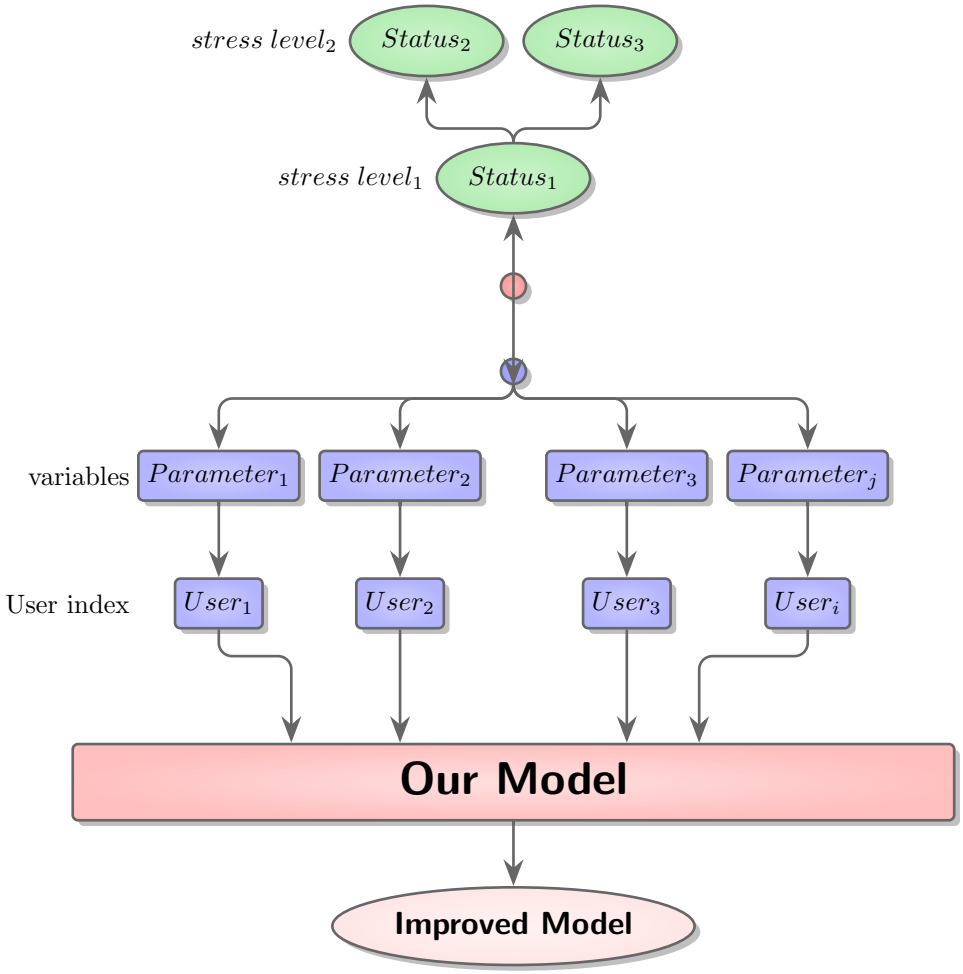


Figure 3.11: Summaried model hierachy

In most choice of models, introducing quadratic terms implies lower order terms are retained before considering the higher order terms of the same properties. This is to avoid the interpretation of results depending on the choice of scaling. Removing a first order term may result in the hypothesis that the predicted response is symmetric and equal to zero. This would not be meaningful and not considered in most cases, since it makes sense when this hypothesis is valid and justifies removal of the lower order (Chatterjee & Hadi, 2015).

Also, if the models have interactive terms, two interacting terms are removed before considering their single nature. A combined elimination associates to both the linear and quadratic surface. The removal of the interacting terms compares to a surface that is aligned with coordinate axes. In most cases, this is hard to interpret (Scott, 2015). The procedure adopted for variable selection for a simplified/generalised model is based on various forms of the linear model, through PCA and also the stepwise/criterion method in FS which is discussed in the following sections.

3.7.1 Stepwise and Criterion procedure

This procedure is suitable for the forward/backward elimination, which takes into account the variables with the smallest p-values and eliminates predictors with p-values greater than the critical point (α_{crit}). The procedure looks at all likely subsets of mutually descriptive variables and locates the model that has the best fit to the data, this is also the criterion based procedures. All the variables are fitted and the best choice is based on the Adjusted R^2, Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC).

1. Adjusted R^2 - is the difference between the total sum of square (TSS) and the residual sum of squares (RSS) to one, given that $R^2 = 1 - RSS/TSS$.

2. BIC/AIC: This allows for larger models to fit better and to have smaller RSS with more parameters or variables. The best choice balances the fit with the model size. A smaller model is best choice for AIC while BIC penalises larger models.

All choices of the variable selection criterion were considered and compared, and the best was considered as the model for FS.

3.8 Model Validation

Validation is an essential part of the modelling to compare the performance of the different modelling techniques adopted. To demonstrate the major contributions,

this thesis approaches the steps to evaluation and validation of the models in the following stages:

3.8.1 Principal Component Analysis (PCA)

The PCA is used here as an ordination method to reduce the dimensionality of the datasets, this is done by creating different key explanatory variables which are the principal components (PCs). Each of the components justifies most of the variance obtained in the dataset. All the components are independent and quadratical. The components detected are used to establish the contributions of each of the variables in the data to each of the principal analyses. The steps involve standardising the variables to the same scale and avoiding some variables becoming dominant or outlying due to their large dimensional units.

The main aim of PCA in respect to our data is to identify hidden or strong patterns in the original dataset and reduce the dimensionality, or eigenvectors, of the data by excluding noise and redundancy in the data, as well as to identify correlated variables. The rationale is to determine the number of possible components to retain for summarising the information in the data, compute the coordinates, the characteristics, attributes and contribution so the variables. The contributions and quality of the retained individual interactions with webpages, were also calculated, the correlation circle of PCA interpreted and the PCA predicted in order to validate the performance of the methods used for comparative data analysis. The FS algorithm was also used to validate the outcome of the PCA.

3.8.2 Classification models

The classification models used for comparison includes Logistic Regression and Neural Network , as discussed below:

3.7.2.1 Logistic Regression

The purpose of using logistic regression for evaluation is based on the nature of the dataset (i.e. that they are not highly correlated with one another). The

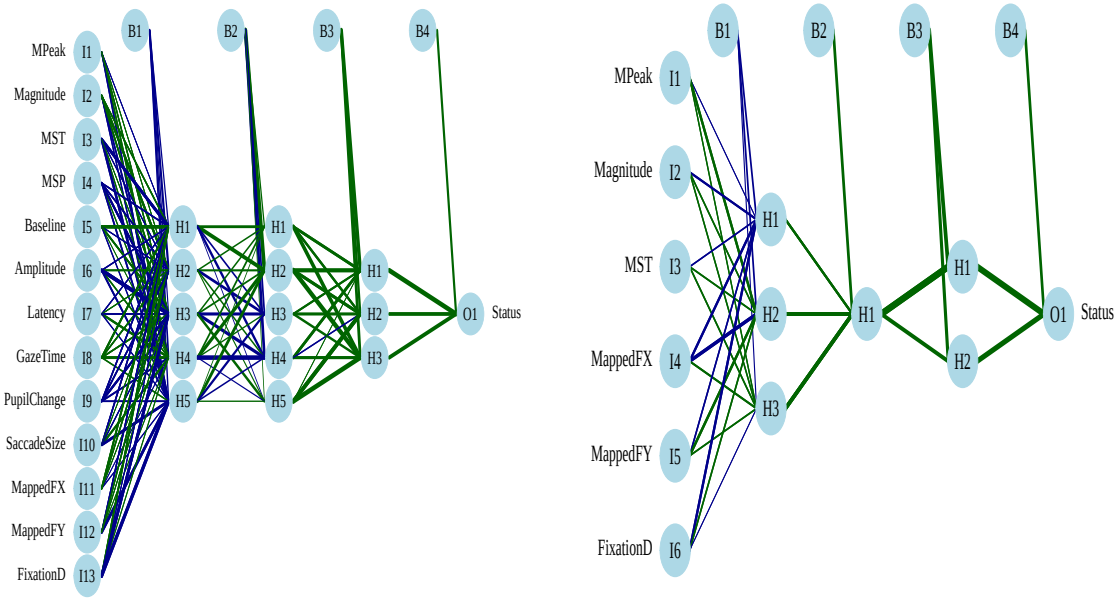
choice of this model is based on the fact that it is reliable for predicting problems arising from the continuous or measured data of regression model. This predicts the outcome of the response given to a set of inputs (physiological attributes), where only binary outputs are given. So for this case, a user can either be stressed or relaxed. The logistic regression is also a case of a linear model with a similar relationship to linear regression.

To select the best model for evaluation, we used the Forward, Backward and Stepwise method of feature selection already discussed in section 3.7.1, for different splits. The model with the least error was selected and compared with the PHYCOB I model.

3.7.2.2 Neural Network

The rationale for using a neural network as a form of evaluation to PHYCOB I is based on the fact that given the nature of our dataset (dispersed distribution) from physiological attributes, it takes these attributes as inputs and predicts the output with a set of specified hidden layers or neurons. This helps to solve issues arising from complex problems such as speech recognition and handwriting recognition.

Different hidden layers or network architecture were specified for the purpose of selecting the model with least error (Figure 3.12) for validation purposes. Since our neural network is a classifier, the role of thumb is to set the number of neurons to be between the number of a given output to the number of a given input, for each split containing different training and test sets.



(a) Neural network with primary attributes

(b) Neural network with important attributes

Figure 3.12: Neural network architechtures with original data attributes and simplified version.

The size of the hidden layers/neurons depends on their optimal size. To get a decent performance we set the hidden layer configuration using the following rules:

- the number of hidden layers to be equal to one or more.
- the number of neurons to be equal to and not greater than the total neurons in the input and output layers.

For these we set the hidden layer for different splits to be one with sizes between 1-15 (14 attributes and 1 for the response variable). The three hidden layers comprised total neurons in each layer to be less than 15 nodes. The best performance with the least error was selected for evaluation.

The reliability of PHYCOB I was tested by comparing its performances with the predictive models. The receiver operating characteristics (ROC) were used to

visualise the performance of the predicted values.This indicates the performance of the classifier models since its discrimination threshold is varied, thereby providing tools to select possibly optimal models and to remove suboptimal ones autonomously from the class distribution and thus to reach a diagnostic for decision making.

The reliability of the features selected were tested by comparing the data generated by PHYCOB I with another feature selection method such as PCA, stepwise/criterion procedure, and also with the model obtained by simulation of a dataset with the same properties of the original model. The learning error curve was used to select the model with least error and this was compared with the proposed model.

3.8.3 Cross Validation

The common method for cross-validation in classification models is leave-one-out cross validation, as mentioned in literature. This is to address the problem of overfitting in the data by using random multiple splits. To visualise the outputs and results the packages in *R* programming was utilised, as mentioned before, on account of its capabilities for data mining, statistical analysis and graphical representation. The result of the output are illustrated and discussed in Chapter Four.

3.9 Summary

This chapter discusses the methods adopted for data collection and data analysis. The experimental setup was first discussed, in terms of recruiting participants for the experiment and the collection of data using an objective method. The chapter introduced the algorithm for detection of peaks that correlates to events, as well as the method used in classifying and generating a high dimensional dataset for modelling. Data exploration was conducted by using the PCA and FS algorithm in Appendix B.2 to identify patterns, detect outlying cases and masking effect or missing data. Stalactite plots were used to visualise these cases and the regression model was applied in feature selection. Other existing techniques such as Neural

Network and Logistic Regression were adopted for validation and evaluation. The purpose of validation was also to compare the performance of the modelling techniques to determine the best model and best fit. The ROC was also used for this purpose.

Chapter 4

Design and Implementation of PHYCOB I

This Chapter discusses the design and implementation of the PHYCOB I model. It talks briefly about the programming language used to design the model and also discussed its implementation, model fit and outcome of the model.

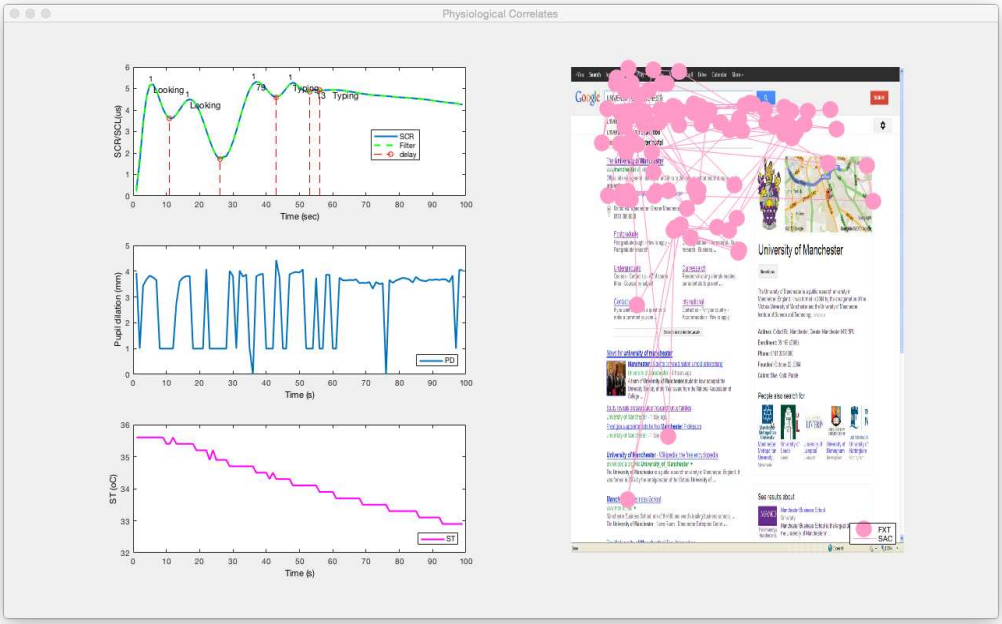
4.1 Design of PHYCOB I

The PHYCOB I (Appendix B.3) was developed in MATLAB and its graphics user interface was used to include controls for easy access and user friendly capabilities during analysis. It can generate secondary data (Figure 4.1) of more than a thousand instances of users interaction. These secondary data are formed from the computed user attributes obtained from the physiological measuring sensor. The purpose of PHYCOB I (Figure 3.5) is to import primary data generated by these sensors to perform the relevant compilation of the data attributes and prepare them for modelling. Figure 4.3 shows an example window that accessed and produce the outcomes of users' interaction to six webpages in a session, an example of which is in Figure 4.2. For this case each participant interactes with six webpages, therefore each participant generates six intances (Figure 4.1), the seven different colored outlines on the dataset shows seven participant's generated data during interaction. For the forty-four participants, a total of 264 instances was produced.

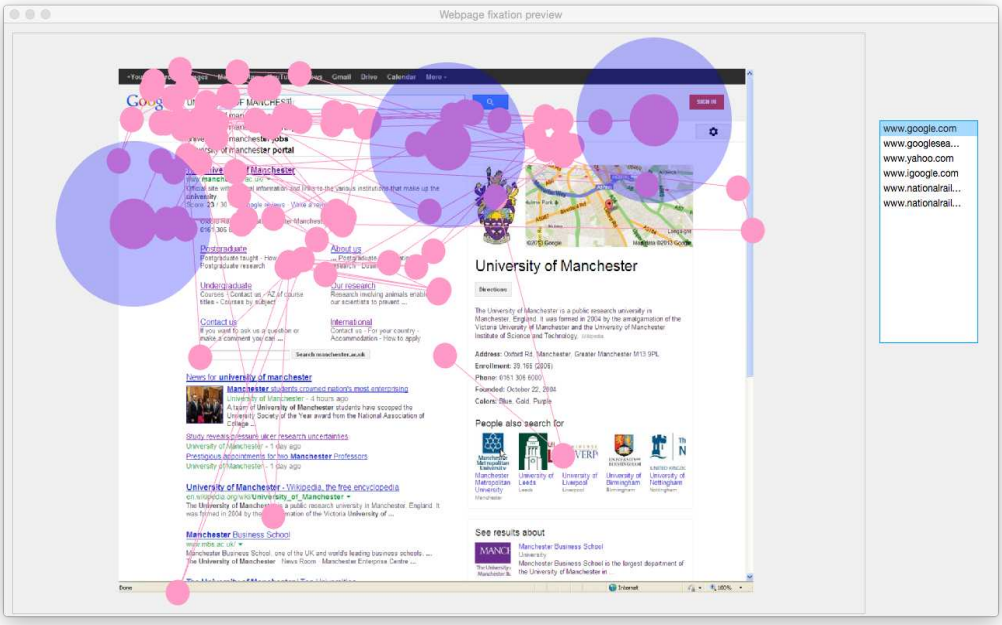
PHYCOB I integrates physiological readings and eye movement behaviour to produce a single interface (Figure 4.2) where the stress points on the webpages can be seen. For example a participant felt stressed while looking at ASL on AOI(1) and looking off screen from the Google-suggest page which appeared as blue transparent dots on Figure 4.2b. The result of the spikes in the physiological readings generates an intergrated interface with the users' affect state located on the webpages. The status of the user is derived from the computed secondary data. These computed parameters were obtained from physiological readings that correspond to eye movement and fixations on a webpage. The increase and decrease in amplitude of SCR correspond to user activity. The average peaks, latency and amplitude were computed for the SCR; likewise the mean pupil dilation (PD) and the mean skin temperature (MST) of users' responses to the different webpages mentioned in Chapter Three.

	URL	MST	Magnitude	latency	MSP	Gaze_Time	MPeaks	Baseline	PupilChange	Amplitude	MFixationX	MFixationY
1	Google	34.8971	10.7287	6	14.0748	63.5000	1.3543	1.2334	2.6957	1.3273	506.2016	334.4032
2	Yahoo	35.9750	8.2364	46	27.4668	187.5000	0.8492	0.8477	3.2635	0.4253	559.1905	520.6905
3	GoogleSearch	36.7444	6.4058	6	14.2973	312.5000	0.7815	0.7434	2.4878	0.1474	672.0714	449.9206
4	iGoogle	37.3903	4.4164	18	18.5543	437.5000	0.5048	0.4688	3.0963	0.2727	561.3333	400.8413
5	National Rail Suggest	37.3528	5.4649	7	14.9497	562.5000	0.6834	0.5812	2.2692	0.4964	650.8095	480.1190
6	National Rail Search	36.8320	2.0563	4	13.8436	702	0.0591	0.0027	2.7726	0.6988	395.2258	343.9161
7	Google	34.9094	19.6158	11	15.5108	61.5000	2.7886	2.5892	3.0155	0.6230	458.5000	235.5750
8	Yahoo	33.3127	20.3709	3	12.1582	181.5000	2.7760	2.7071	3.3527	0.1620	507.2131	382.7541
9	GoogleSearch	32.0255	21.0650	2	11.4618	302.5000	2.9047	2.7754	3.0747	0.3600	440.6803	642.5164
10	iGoogle	31.2036	20.4831	49	26.8819	423.5000	3.0500	2.6080	3.3384	0.4420	502.1721	388.8361
11	National Rail Suggest	30.6291	21.0965	15	15.3480	544.5000	2.9923	2.7463	3.1120	0.4150	387.9426	438.4180
12	National Rail Search	30.5465	14.7612	5	12.9288	679	1.0634	0.8789	1.8411	3.2398	394.6779	419.4899
13	Google	34.3693	41.3053	6	15.1138	93	5.0274	3.9356	2.7130	4.9721	459.6393	258.5628
14	Yahoo	32.0989	39.4740	34	22.2463	276	4.1758	4.0425	3.2659	0.6399	483.7838	577.8865
15	GoogleSearch	30.8378	44.7626	13	15.1891	460	4.8129	4.2529	2.7674	1.7295	465.2216	389.9946
16	iGoogle	30.6000	45.7294	48	26.5369	644	4.7278	4.6126	3.2429	1.0107	380.9459	449.1730
17	National Rail Suggest	31.8756	48.4082	25	19.7408	828	4.9559	4.1722	2.4376	2.3467	448.4108	260.8486
18	National Rail Search	31.9120	49.7737	27	21.4861	1.0255e+03	2.6974	2.3747	2.6738	5.5462	480.7594	553.8679
19	Google	34.7629	25.0902	7	15.0387	54	3.3733	3.0491	2.3934	3.3532	468.1238	398.4095
20	Yahoo	32.9609	27.0516	18	17.2179	159	3.5698	3.2653	3.5411	0.6929	504.3271	639.2430
21	GoogleSearch	31.6375	26.9031	16	16.0867	265	3.3772	3.1574	2.9089	0.6225	421.3458	455.9813
22	iGoogle	30.8656	27.8680	4	11.8157	371	3.5222	3.3596	2.7588	0.5814	346.8318	512.6449
23	National Rail Suggest	30.5500	25.4805	11	13.9450	477	3.2314	3.1329	3.2123	0.2849	315.3084	585.4766
24	National Rail Search	30.6022	20.1381	3	12.4389	596.5000	1.4099	0.9261	2.8739	3.7145	382.3209	397.9328
25	Google	26.8186	19.4901	3	10.7103	90.5000	2.2140	2.0872	3.8144	2.3123	511.3876	422.9719
26	Yahoo	26.7227	20.6236	6	10.9568	268.5000	2.2119	2.1727	3.7816	0.1476	399.6278	544.0722
27	GoogleSearch	26.6523	20.2500	8	11.5832	447.5000	2.1762	2.1385	3.7205	0.0975	471.8889	449.4556
28	iGoogle	26.9386	19.8465	2	9.6798	626.5000	2.1338	2.1108	3.8678	0.1007	555.6722	384.1556
29	National Rail Suggest	27.2932	19.4488	13	13.4826	805.5000	2.0781	2.0585	3.9203	0.1545	521.3111	544.4722
30	National Rail Search	27.5726	20.3857	23	17.5985	1000	1.1368	1.0325	3.6732	2.2227	431.8436	358.2701
31	Google	34.8971	10.7287	6	14.0748	63.5000	1.3543	1.2334	2.6957	1.3273	506.2016	334.4032
32	Yahoo	35.9750	8.2364	46	27.4668	187.5000	0.8492	0.8477	3.2635	0.4253	559.1905	520.6905
33	GoogleSearch	36.7444	6.4058	6	14.2973	312.5000	0.7815	0.7434	2.4878	0.1474	672.0714	449.9206
34	iGoogle	37.3903	4.4164	18	18.5543	437.5000	0.5048	0.4688	3.0963	0.2727	561.3333	400.8413
35	National Rail Suggest	37.3528	5.4649	7	14.9497	562.5000	0.6834	0.5812	2.2692	0.4964	650.8095	480.1190
36	National Rail Search	36.8320	2.0563	4	13.8436	702	0.0591	0.0027	2.7726	0.6988	395.2258	343.9161
37	Google	34.7629	25.0902	7	15.0387	54	3.3733	3.0491	2.3934	3.3532	468.1238	398.4095
38	Yahoo	32.9609	27.0516	18	17.2179	159	3.5698	3.2653	3.5411	0.6929	504.3271	639.2430

Figure 4.1: Generated secondary data for each participant's interaction with six webpages



(a) Synchronised physiological responses with eye movement on webpages



(b) Detected stress points on webpage

Figure 4.2: Physiological responses of a user in sync with fixations made by eye movement behaviour on a webpage and the integrated interface

4.2 Implementation

This section discusses the implementation of the PHYCOB I model. Figure 4.3 shows the window of dataset generated from the input data (Figure 3.10) of participants collected from physiological measuring sensors. The interface used to import these data is shown in Figure 3.5. The main purpose of this interface is to indicate example data generated from some participants. The PHYCOB I algorithm processes the physiological parameters for each webpage interaction and the data generated for each participant was used to update the users' data. The total data generated for all participants give rise to a secondary dataset for modelling which is discussed in the proceeding sections. We later tested the nature of our datasets by detecting natural structures or components arising from the data. The performance and model reliability of PHYCOB I was implemented by using the original data and multiple simulations of datasets so as to compare with neural network and logistic regression models in Chapter 5. PCA and FS was used to determine strong patterns, help validate the choice of models used for comparison and also validate the performance of models.

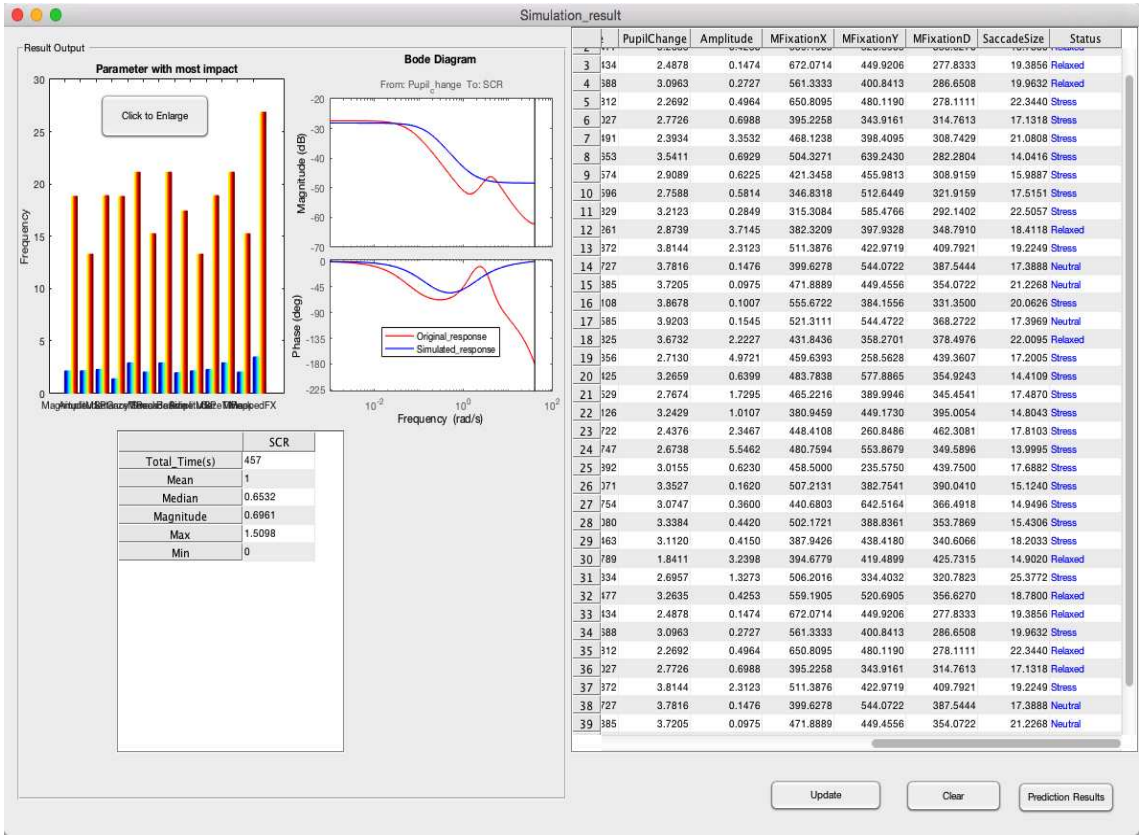


Figure 4.3: Data generated for participants interaction with webpages

4.3 Outcome of the PHYCOB model fit

The PHYCOB I was run on multiple polynomial order, the best three outcomes are polynomial order 1-3 of the model fit, running on order 2 and 3 generates similar phasic change compared to runing on order 1. The generated effect from order 2 gives the best outcome in terms of possibility of a high accurate and significate performance with a probability of 0.74 (Figure 4.5), the predicted and original response of MappedFX and MappedFY lies in the same catersian coordinate and close together; this implies precision in prediction focus (four seconds from original response); this is the default in the control system. The bode plot in Figure 4.4 shows the model fit for each of the user attributes. From Figure 4.3, the bode for the simulated and original response shows a normalised variation in data of the simulated response compared to the original response

from pupil dilation to SCR, this a perfect representation of a real life physiological reactions of user interaction to webpage's dynamic contents .

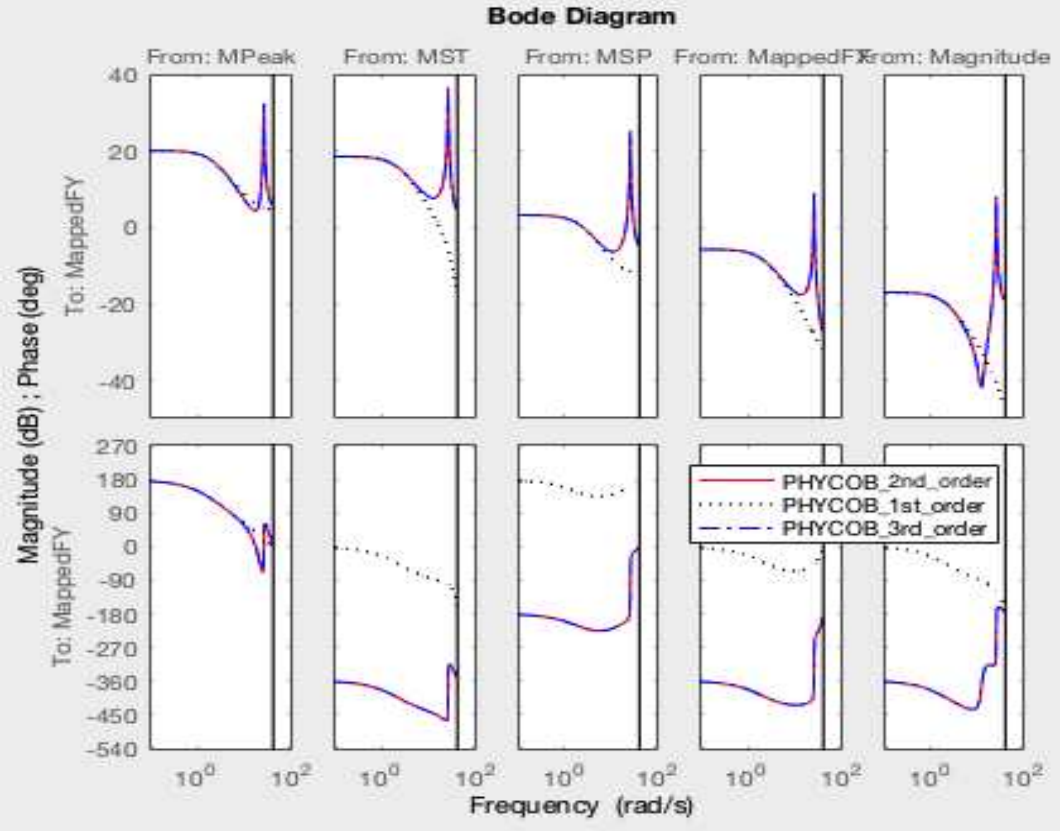
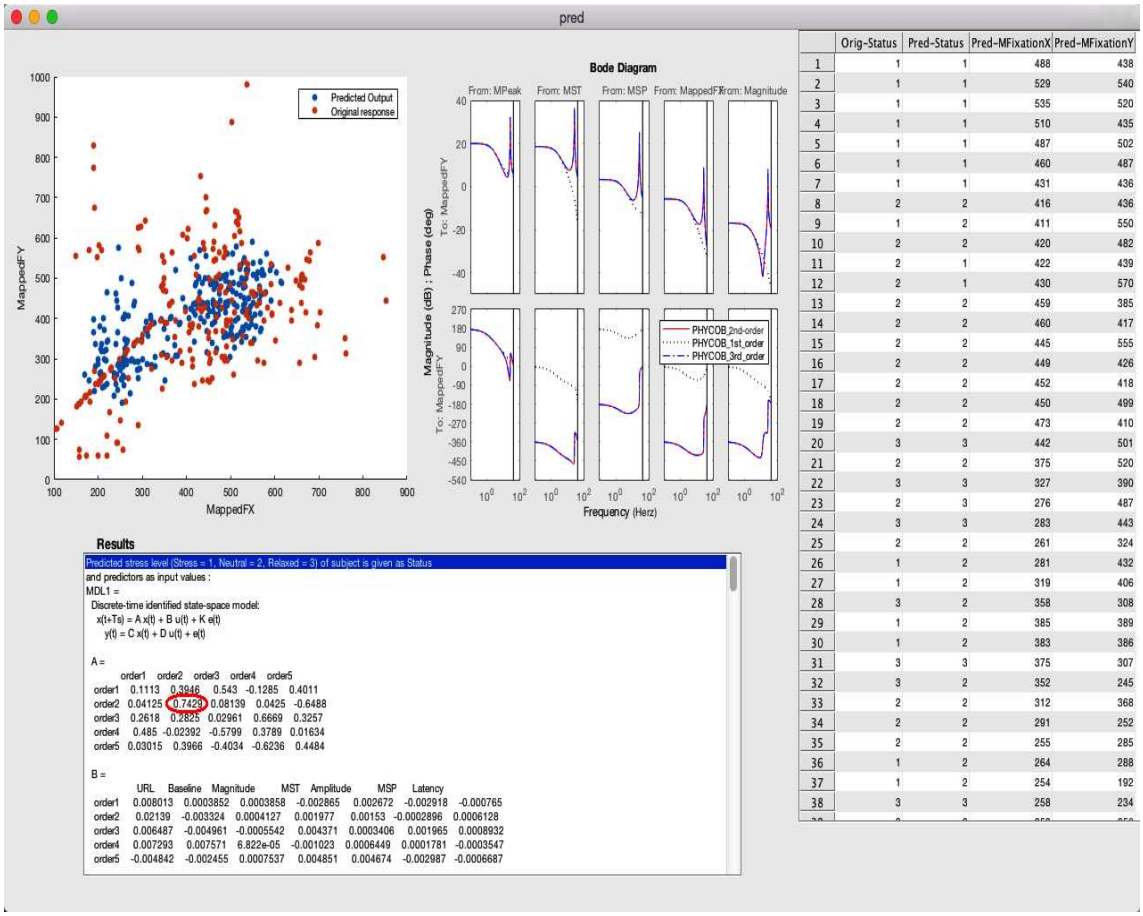


Figure 4.4: Phasic changes of optimal parameters on multiple runs

Figure 4.5 indicates a window that displays bode plot of user attributes, estimated coefficients of the users attributes, original/predicted user status (stress levels: 1 = Stressed, 2 = Neutral and 3 = Relaxed) and corresponding predicted MappedFX and MappedFY of the webpage. The learning curve (Figure 5.1c) on the model for cross validation is used to visualise the error in the learning process of the model.



	Orig-Status	Pred-Status	Pred-MFixationX	Pred-MFixationY
1	1	1	488	438
2	1	1	529	540
3	1	1	535	520
4	1	1	510	435
5	1	1	487	502
6	1	1	460	487
7	1	1	431	436
8	2	2	416	436
9	1	2	411	550
10	2	2	420	482
11	2	1	422	439
12	2	1	430	570
13	2	2	459	385
14	2	2	460	417
15	2	2	445	555
16	2	2	449	426
17	2	2	452	418
18	2	2	450	499
19	2	2	473	410
20	3	3	442	501
21	2	2	375	520
22	3	3	327	390
23	2	3	276	487
24	3	3	283	443
25	2	2	261	324
26	1	2	281	432
27	1	2	319	406
28	3	2	358	308
29	1	2	385	389
30	1	2	383	386
31	3	3	375	307
32	3	2	352	245
33	2	2	312	368
34	2	2	291	252
35	2	2	255	285
36	1	2	264	288
37	1	2	254	192
38	3	3	258	234

Figure 4.5: Predictions for three predicted response parameters of PHYCOB I and a bode plot comparing different order for the model

The variables with the most impact (Figure 4.6), which were obtained from the covariance of the dataset, has p-values less than critical value (p = 0.05 default) with a 95% confidence interval. These are the most significant variables that contribute in classifying the stress levels (Status) for predictions. The variables are the MappedFX, GazeTime, Saccadesize, MST and MPeak (Appendix B.1). The predictions from the model fit (Equation 3.6) are made on the class labels and on example response variables like the eye fixations (MappedFX and MappedFY), which indicate possible areas of interest (AOI) of users on the webpages. The resulting predictions serve as possible coordinates for AOI that can affect the status or stress level of users on the webpages they interacted with. The model

fit option is based on least squares, which automatically normalises the data. The predictions for stress levels as the output response is used to represent the model's relibility when compared with other models (Figure 4.5).

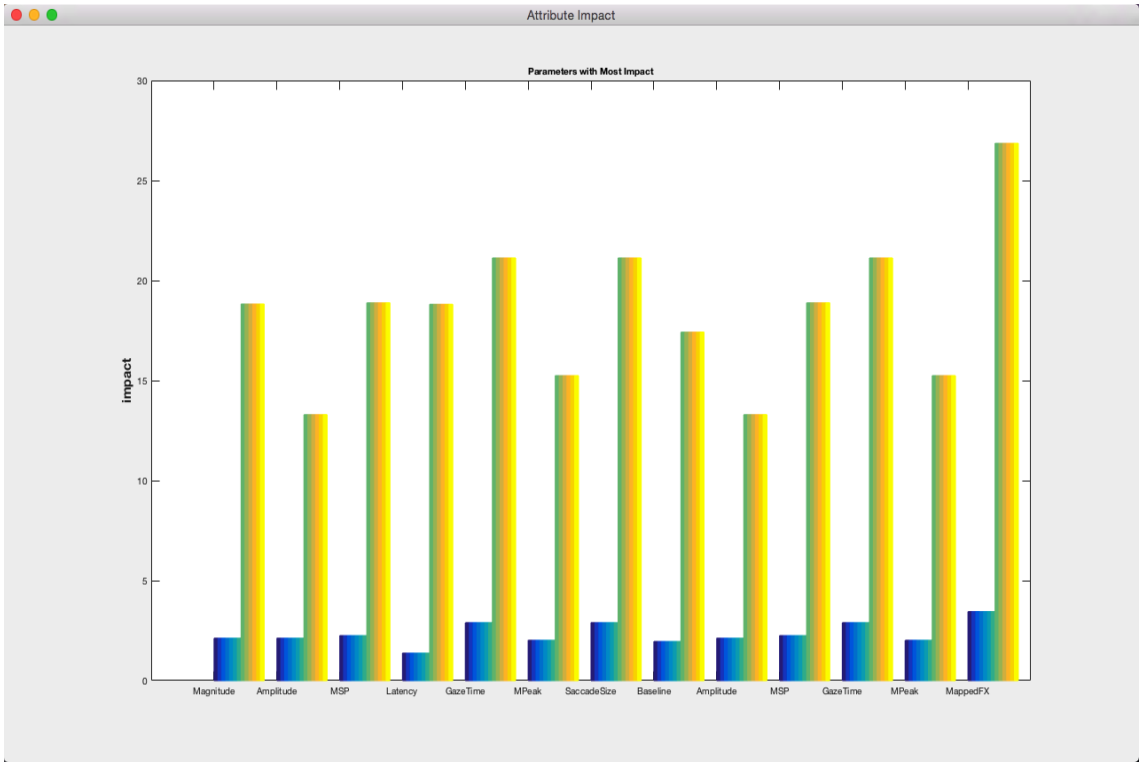


Figure 4.6: Graph indicating parameters with most impact for PHYCOB I

4.4 Summary

This Chapter discussed the design and implementation of PHYCOB I, which mostly talks about the significant windows used to generated users' attributes that was prepared for modeling. The model fit is based on a system control dynamics using and testing on different polynomial order of the differential equation used to define the model fit. The best attributes that clearly defines the model on the best order, are seen to be MappedFX, GazeTime, Saccadesize, MST and MPeak which are the users' attributes that contributes to determing the correlates of users' physiological response to the webpage's contents.

Chapter 5

Comparative Data Analysis and Contribution

This chapter presents a comparative data analysis and outlines the contributions of the project work. The initial stage for data analysis was discussed in Chapter Three. This involves computation of attributes with the PHYCOB I model that helps to classify the status of users based on the mean peak response, average amplitude, response duration detected for each webpage and make predictions based on these attributes. One of the focuses of this chapter is comparing its performance to other standard methods. FS and PCA were both used as validation methods for the class labels. These validations are compared across the Neural Network, Logistic Regression and PHYCOB I models. The steps towards implementation, contribution, method of analysis, data visualisation, model results and performances is discussed in the succeeding sections.

5.1 Comparing the Performance of PHYCOB I To Other Classification Models

To compare PHYCOB I to Neural Network and Logistic Regression, each of the model was trained on a sequence of sub-samples and tested on the remaining part of the data with test splits as $90\%, 80\%, 70\%, 75\%, 50\%, 35\%, 25\%, 20\%$ and 10% using the true class label (stress levels) for all iterations.Table 5.1 shows the best

accuracy of all training sets. An optimal Logistic Regression model was selected from runs involving the Forward, Backward and Stepwise models and likewise for the Neural Network model on different schematic structures, the set of different network architectures has already been described in section 3.7.2.2 of Chapter 3, and includes different hidden layers and nodes selected between 1 and 15 on multiple
training sets used to define the learning curves (Figure 5.1). The choices were based on the Receiver Operating Characteristics (ROC) in Figure 5.4. For each split, the training/testing set $\alpha = 10\%, 30\%, 50\%, \text{and } 75\%$ were used since it indicates the best performance for all splits. The performance result shows that PHYCOB I has high accuracy on all sets of simulated data, except at simulated sets $M3$ and $M4$ (Table 5.1 and 5.3), where it shows the worst performance of all the training sets. The dataset simulated for this was based on random normal values and multivariate data, using the stress level as the true class for the data on all splits.

Table 5.1: Performance of models on multiple simulaitons

Stress Level												
Sim	PHYCOB I Test set				Logistic R Test set				Neural Net Test set			
	10%	30%	50%	75%	10%	30%	50%	75%	10%	30%	50%	75%
Performance												
M1	0.86	0.90	0.84	0.84	0.80	0.85	0.85	0.81	0.79	0.86	0 67	0.61
M2	0.70	0.87	0.79	0.84	0.85	0.84	0.81	0.81	0.85	0.78	0.80	0.85
M3	0.33	0.63	0.43	0.47	0.66	0.61	0.54	0.58	0.74	0.58	0.53	0.60
M4	0.40	0.61	0.58	0.53	0.16	0.49	0.51	0.51	0.20	0.49	0.51	0.52
M5	0.60	0.59	0.65	0.66	0.56	0.60	0.59	0.50	0.65	0.63	0.61	0.53

For test sets with high performance models the average number of predicted stress levels with the true and false positive class for the three models is shown in the table below. The PHYCOB model mostly agrees with the FS (section 5.2.2) in terms of detecting all possible class labels in their actual stress levels. These

predicted classes were obtained by projecting each model training set on the test set (new data) i.e 30% of the original data.

Table 5.2: Matrix table indicating average number of true positive and false positive predicted classes for the three models

Models	Predicted class				
			Neutral	Relaxed	Stressed
PHYCOB I	Actual	Neutral	24	0	1
		Relaxed	1	19	0
		Stressed	0	0	78
Logistic R	Actual	Neutral	24	0	1
		Relaxed	2	18	0
		Stressed	0	0	78
Neural Network	Actual	Neutral	23	1	1
		Relaxed	1	18	1
		Stressed	1	0	77

Table 5.3: Performance of Testing/Training set on multiple simulations

Splits			10%	30%	50%	75%
PHYCOB I	M1	Train Accuracy	0.86	0.90	0.84	0.84
		Test Accuracy	0.79	0.82	0.84	0.74
	M2	Train Accuracy	0.70	0.87	0.79	0.84
		Test Accuracy	0.86	0.76	0.75	0.83
	M3	Train Accuracy	0.33	0.63	0.43	0.47
		Test Accuracy	0.36	0.35	0.45	0.43
	M4	Train Accuracy	0.40	0.61	0.58	0.53
		Test Accuracy	0.36	0.55	0.46	0.47
	M5	Train Accuracy	0.60	0.59	0. 65	0.66
		Test Accuracy	0.56	0.55	0.56	0.57
Logistic R	M1	Train Accuracy	0.80	0. 85	0.85	0.81
		Test Accuracy	0.78	0.80	0.84	0.84
	M2	Train Accuracy	0.85	0. 84	0.81	0.81
		Test Accuracy	0.87	0.75	0.75	0.73
	M3	Train Accuracy	0.66	0.61	0.54	0.58
		Test Accuracy	0.56	0.65	0.55	0.53
	M4	Train Accuracy	0.16	0.49	0.51	0.51
		Test Accuracy	0.26	0.45	0.56	0.47
	M5	Train Accuracy	0.56	0.60	0.59	0. 50
		Test Accuracy	0.54	0.55	0.52	0.51
Neural Network	M1	Train Accuracy	0.79	0.86	0. 67	0.61
		Test Accuracy	0.76	0.81	0.64	0.64
	M2	Train Accuracy	0.85	0.78	0.80	0.85
		Test Accuracy	0.86	0.75	0.75	0.78
	M3	Train Accuracy	0.74	0.58	0.53	0.60
		Test Accuracy	0.76	0.75	0.65	0.53
	M4	Train Accuracy	0.20	0.49	0.51	0.52
		Test Accuracy	0.26	0.35	0.46	0.57
	M5	Train Accuracy	0.65	0.63	0.61	0.53
		Test Accuracy	0.54	0.65	0.62	0.54

Table 5.4: Aggreate of performance in splits

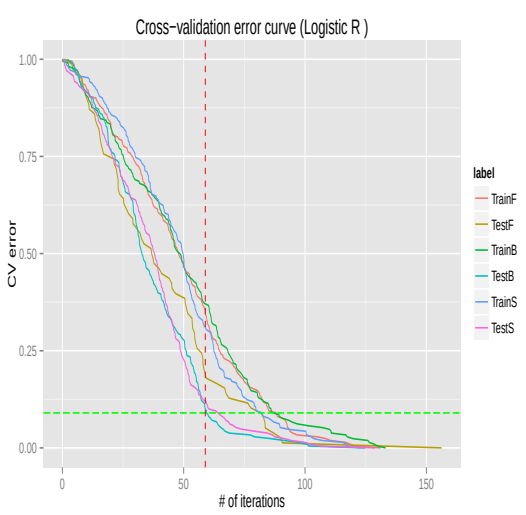
Splits	10%	30%	50%	75%	Average
PHYCOB I	58	72	66	67	65.75
Logistic R	61	68	66	64	64.75
Neural Net	64	67	62	62	63.75

Table 5.3 shows the best performance using different testing and training splits in each simulation, and the aggregates in Table 5.4 show that PHYCOB I is close to and exceeds the other models with 1% on average. The results for the logistic regression and neural network are used to plot the learning curve in Figure 5.1, where the models with least error are selected and compared with PHYCOB I.

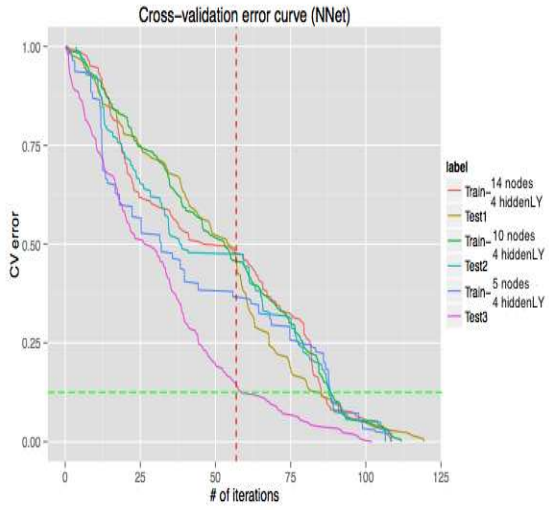
5.1.1 Overfitting

Overfitting normally occurs when a model is extremely complex (Diebold, 2015; Scheres & Chen, 2012), and also if there is large number of possible outcomes. To avoid this we resort to multiple splits of the data. An example is the Neural Network showing different validation errors in Figure 5.1. The errors are due to slight variations in data as a result of noise and other constraints which were not modelled during the training routine such as differences in feature selection and number of observations simulated.

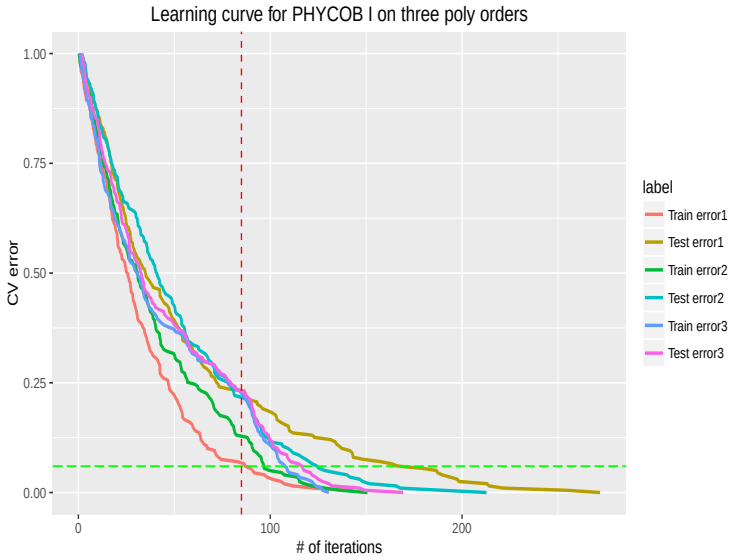
The models were tested on unseen test set (30% of 2500 simulated data) to get some idea of whether overfitting had occurred or not. As explained earlier, different datasets were used during the training process to determine this. For each dataset the test set was extracted and used on each algorithm. To generate a learning curve, the models were fitted at a sequence of dissimilar training set sizes and the validation and training error of the models was computed.



(a) Learning curves for Training/Testng sets for Logistic Regression different iterations



(b) Learning curvse for Trainng/Testng sets for Neural Network on different architecture



(c) Learning curvse for PHYCOB I on three best orders $(4th, 3th, 1st)$

Figure 5.1: Cross-validation error curve for Logistic Regression and Neural Network from the split with best performnce

The optimal models for the Logistic Regression and Neural Network with the least errors are then compared to the PHYCOB I's best output . The performance

of PHYCOB I using 70% training gives a significant performance of 0.90. This implies that the model learns more with a higher number of training sest than test sets compared to the other models. Even with high test sets, there is still an indication of highly significant performance and the model seems to resist overfitting due to regularisation by taking a smooth function of the variables.

The cross-validation error of the training/test set for Logistic regression gives the least error between 0.08 and 0.13 at the 54th iteration while the error training/testing set for the Neural network model gives the least error of 0.14 at the 59th iteration. The learning curve shows that the model error decreases as the number of training sets increases. The variables that were optimal for each of the forward, backward and stepwise methods for logistic regression indicate the "Mean Peak"(MPeak), "MST"and "MappedFX" of the webpages as being the best parameters for the optimal model with p-values less than the critical value of 0.05 while for Neural network the best features includes MPeak, MST and SaccadeSize (Figure 5.2), by recursive matrix multiplication for all hidden layers. The relationship between these variables can be seen in the PCA chart in Figure 5.6, indicating their independency to each other, which contributes to the high accuracy of these models.

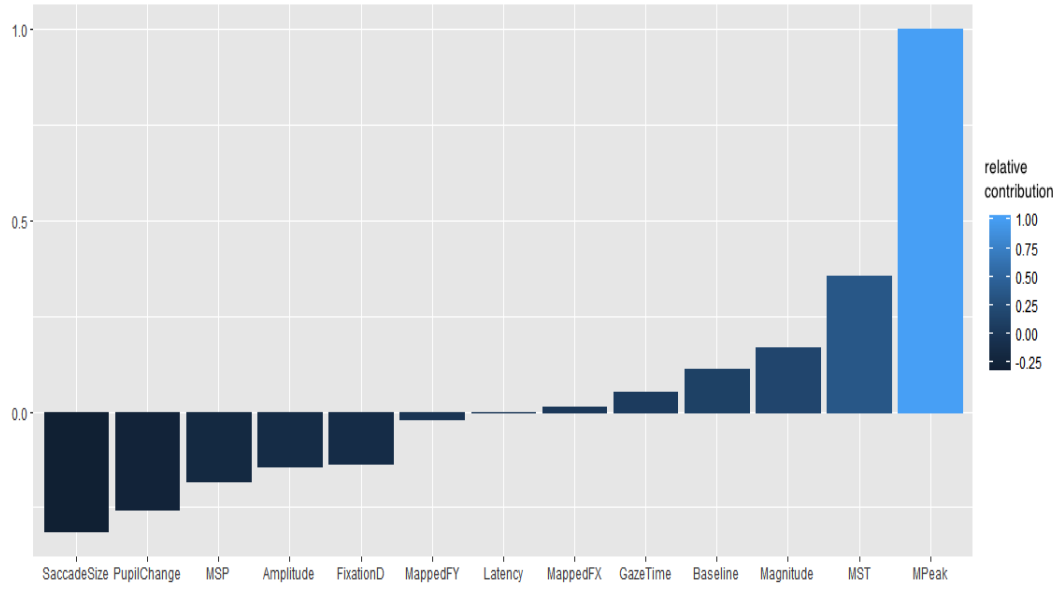
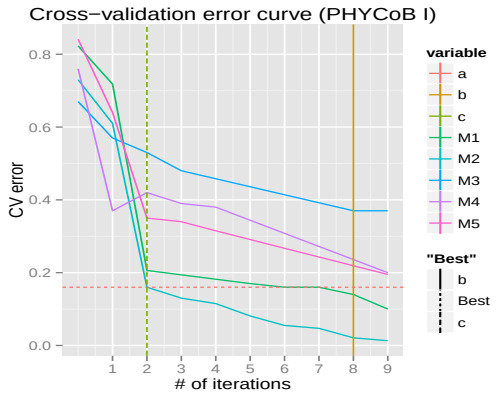
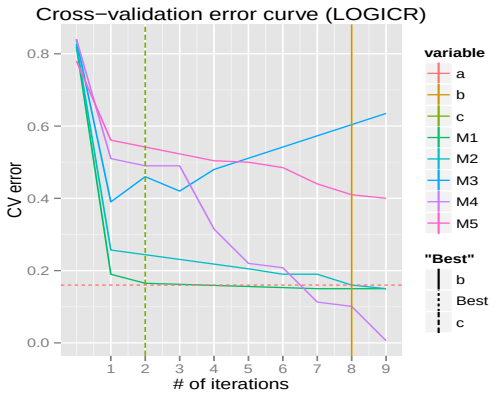


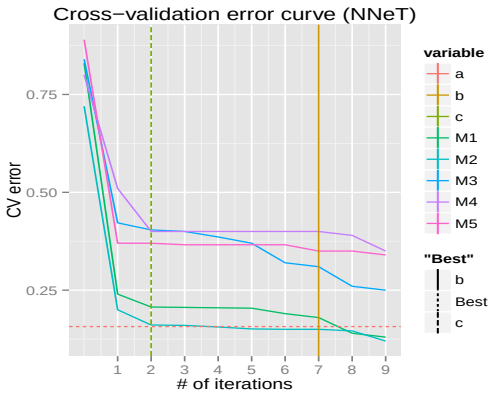
Figure 5.2: Feature importance for Neural network



(a) Error for multiple simulations for PHYCOB I



(b) Error for multiple simulations for Logistic Regression



(c) Error for multiple simulations for Neural network

Figure 5.3: Cross-Validation error for the three models at different simulations of training/testing sets

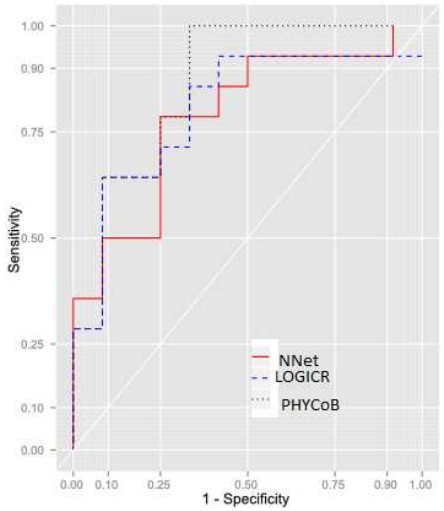
In Figure 5.3, the lines represented by “a”, “b” and “c”are parameters that represent the models with least error and maximum error. The dashed lines denotes least error and iteration steps for the best model. With random normal simulated data, the performance of the models appeared significantly low since there are more sets of simulated data compared to the original datasets. The simulated set “M2”(standardised form of the original data) had the least error (0.2−0.1%) at the second iteration (80% split) for all models but higher validation

error when there was less training set data. The optimal logistic regression model was selected from runs involving the multiple simulated data. The optimal models for the training/test sets using the backward method of logistic regression and training/test sets with four hidden layers and five nodes in the Neural Network have the best performance and are used to compare with PHYCOB I.

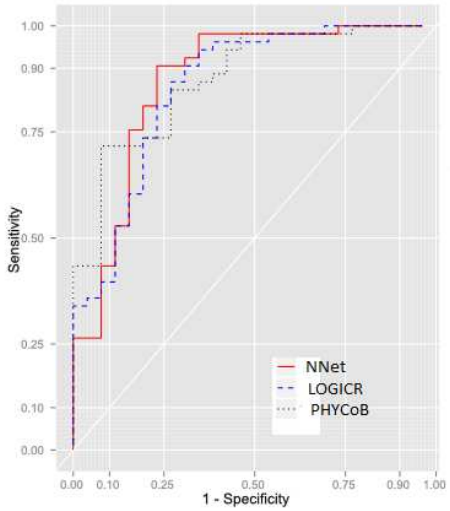
The diagnostics from the comparison in Figure 5.4 show that PHYCOB I is significant enough to indicate the relevance of the model in terms of classification of stress status relating to physiological measures and interaction to the dynamic contents of visual stimuli.

The optimal models for logistic regression and Neural Network M1 at 30% test set are also compared to the PHYCOB I, as shown in Table 5.1 and Figure 5.4.

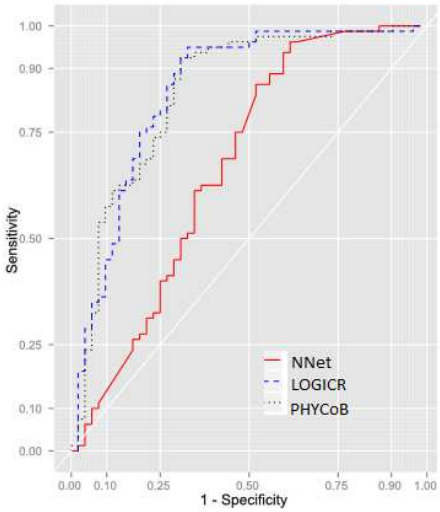
In Figure 5.4 it can be seen that PHYCOB I outperforms the two algorithms used at both 70% train set and 75% test set, which indicates less training error for all the simulated sets $(Mi, \ldots, M5)$ that it was trained on. As the training set size increases, it reduces over-fitting at the second iteration for all simulated sets for PHYCOB I, and increase the error rate at the eighth iteration for logistic regression and at the seventh iteration for the Neural Network. The Neural Network model is trained based on different hidden nodes, as mentioned earlier for each set of training data due to its complexity. The least error and validation performance is 0.1, which indicates higher performance compared to both logistic regression and the neural network models.



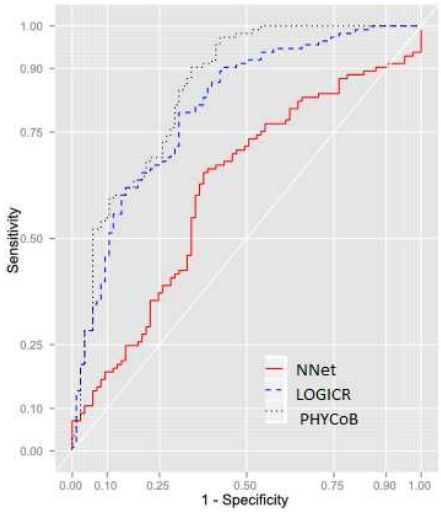
(a) Test set $\alpha = 10\%$



(b) Test set $\alpha = 30\%$



(c) Test set $\alpha = 50\%$



(d) Test set $\alpha = 75\%$

Figure 5.4: Diagnostic plot for models with stress level as class labels

5.2 Verification of the Methods with PCA and FS

Results from implementations of the three predictive models and the selection of optimal choices were discused above in section 5.1. In the next exposition PCA

and FS are used to verify the performance ranking of the three models.

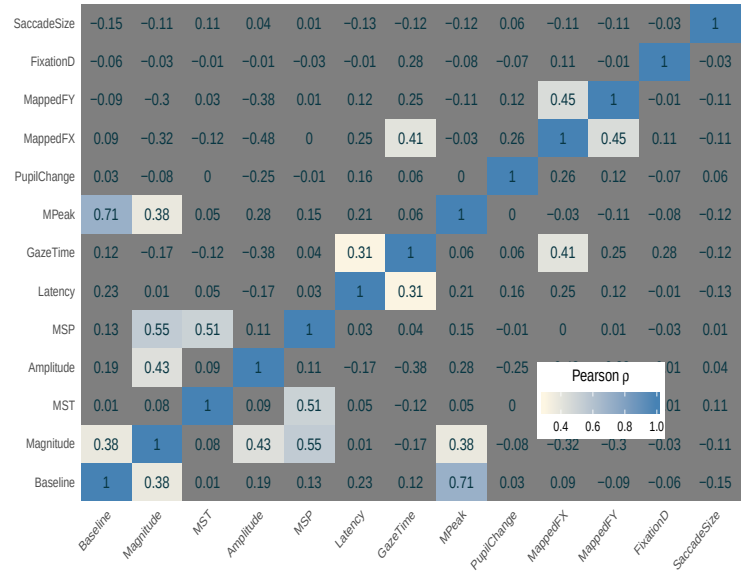
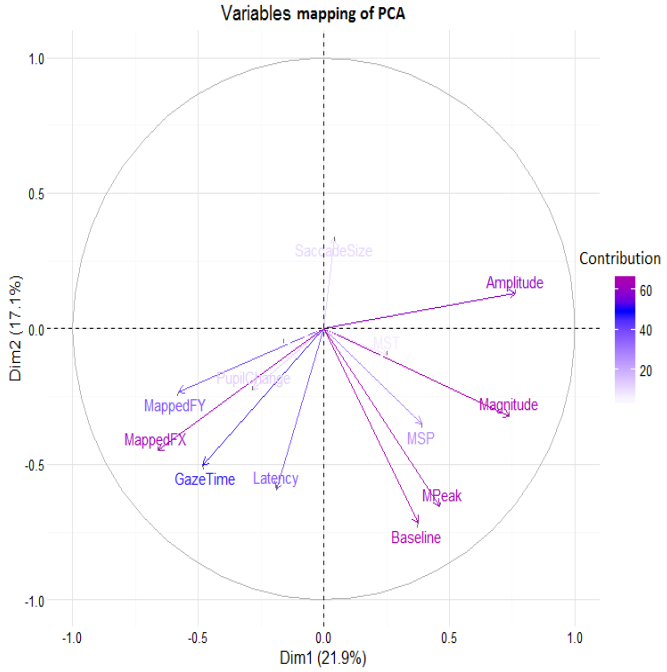


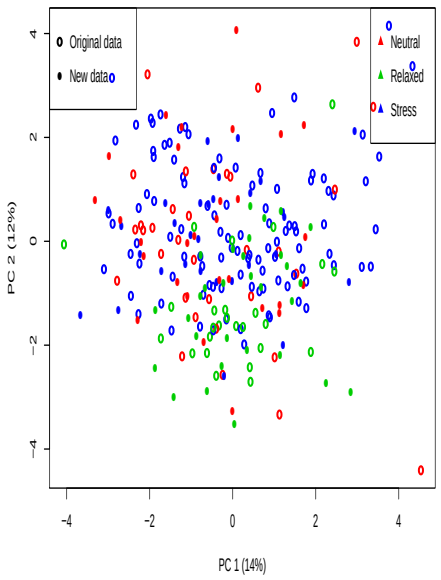
Figure 5.5: Heatmap showing correlations of attributes

5.2.1 Verification by PCA

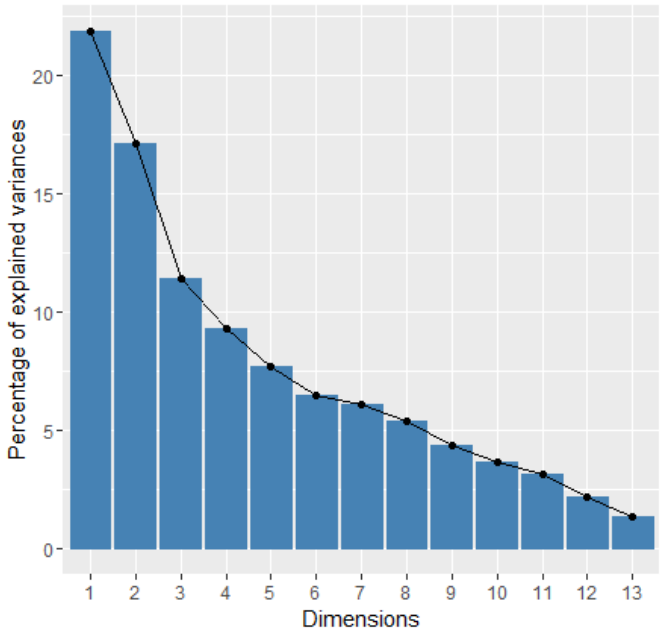
PCA was used to detect strong patterns and verify the performance ranking of the methods including PHYCOB I. To detect the natural structures, it was first determined if there was any correlation among the variables. The correlation heatmap in Figure 5.5 indicates that few of the variables from the original data were strongly correlated, which indicates an overlap in the influence of some of the variables, which accouts for the data variability. Applying the PCA will help disentangle the correlation and also help to detect natural structures that may or may not be similar to the original data. Figure 5.6a and Figure 5.6c shows that 13 components were detected, this equals the number of variables from the original data using class labels for the iterations. The variance plot (Figure 5.6c) shows that the variation in the principal data components (predictive attributes) is high and has a dispersed distribution pattern. The figure also indicates different variations in each predictive component.



(a) PCA of data using stress levels as class labels



(b) Projection of new data onto the PCA coordinate basis



(c) Variance plot showing variations in components

Figure 5.6: Detected PCA components and variance of data

Table 5.5: Eigenvalues of PC with the percentage and cumulative percentage of variance

PCs	Eigenvalue	% of variance (POV)	Cumulative % of variance (CPV)
comp 1	2.8901363	21.9318178	21.93182
comp 2	2.2412783	17.1406020	39.47242
comp 3	1.1874710	9.1343922	48.60681
comp 4	1.1545720	8.8813228	57.48813
comp 5	0.9669600	7.4381537	64.92629
comp 6	0.9206566	7.0819736	72.00826
comp 7	0.8477417	6.5210904	78.52935
comp 8	0.7068345	5.4371882	83.96654
comp 9	0.6672025	5.1323270	89.09887
comp 10	70.5673294	4.3640722	93.46294
comp 11	0.3932258	3.0248135	96.48775
comp 12	0.3271781	2.5167543	99.00451
comp 13	0.1294140	0.9954923	100.00000

To validate based on predictions, new datasets were created for each of the three class labels. These are random data based on the characteristic properties of the original data by computing the correlated/covarying random numbers that an R package provides. The PCs were predicted by projecting the new data onto the PCA coordinate basis, Figure 5.6b shows the final graph of the data when projected on new data or test data. The PCs lie in similar area of the plot as the original dataset, with a total performance of 91%; 64% for the first PC and 27% from the second PC. This closely relates to the performance of PHYCOB I in respect to the original data (M1) in Table 5.1 of Section 5.1. This also demonstrates the generalisation of the proposed model and can be used on other datasets with the same characteristic properties.

Table 5.5 shows the eigenvalues, i.e. the amount of variation retained by each PC . The first PC corresponds to the direction with the maximum amount of variation in the datasets as shown in Figure 5.6a, -60.6% of the variances contained in the data are retained by the first two PCs. The loadings are the correlation between the variable and the PCs. The correlation circle helps to

visualise the most correlated variables. The Mpeak with Baseline, MappedFX with MappedFY are refered to in this case. The closer the variables are to the correlation circle, the better it represents variable impact and the more important it is to interpret the components.

From Figure 5.6a, the most significant variables for the PC1, are MappedFX ($r = 0.74, p = 2.2E-30$) and Amplitude ($r = -0.75, p = 1.04E-30$). The most significant variables for the PC2 are : Magnitude ($r = 0.84, p = 1.01E-45$) and Mpeak ($r = 0.73, p = 1.6E-28$).

These variables happen to correspond with the variables with the most impact in the PHYCOB I model. This further validates the PHYCOB I model in terms of reliability and significance of the model. In addition the dominant patterns in the original, identified in PCA, contribute to the resistance of over-fitting seen in PHYCOB I.

Table 5.5 also shows the proportion of the variance each principal components captures, with the cumulative proportion being computed to output. From the table, eleven PCs are needed to capture 96% of the variance in the original data. From the loadings, Figure 5.6a show variables that can serve as inputs to both the FS, Neural Network and Logistic Regression, these are variables closer to the circle of correlation where the data is scaled to the original data.

The total percentage of the two sets of principle components (PC1 and PC2) is 39.0% of the datasets, variability, which might not be a respectable sum for the total variation. This is due to the fact that the data are distributed and not clustered in groups.The variables MappedFY and MappedFX in Figure 5.6a are strongly correlated; and together with other variables reveal a discrete pattern in the data.

The PCA results suggests that four components be retained for summarising the results, these are components with eigenvalues greater than 1 (Table 5.5). The first two account for 39.0% of the variability of the dataset from PHYCOB I. The loadings in Figure 5.6a and 5.6c indicate parameters MappedFX, MSP, Magnitude and Amplitude, at least two variables of which are true for PHYCOB I. A formal interpretation of this implies that to detect or recognise any potential relationship between "stress levels" and these variables, the areas to where the users, gaze was directed i.e. AOI, and where fixations (MappedFX) were located,

correspond to the magnitude of their physiological response and also to increases in amplitudes of SCR that help to define a users' stress status. This can also be further verified using the FS algorithm in Section 5.2.2.

4.2.1.1 Contributions and quality of elements to PC

This section briefly discusses the contributions of the four PCs selected from Table 5.5, in respect to summarising the information in the data (Table 5.6). These contributions are also extended to the quality of both the variables and observations, by looking at their correlation to these components.

Table 5.6: Eigenvalues of PC with the precentage and cumulative percentage of variance

	eigenvalue	variance.percent	cumulative.variance.percent
PC1	2.8901363	21.9318178	22.23182
PC2	2.2412783	17.1406020	39.47242
PC3	1.1874710	9.1343922	48.60681
PC4	1.1545720	8.8813228	57.48813

The correlation between variables and principal components is shown in the table below. The Magnitude, Amplitudes, MappedFX, MappedFY of the users' responses are highly correlated to PC1, which explains the majority of the data's variability.

Table 5.7: Correlation of variables to principle components

	PC1	PC2	PC3	PC4
Baseline	0.32298929	-0.547013491	0.44074598	-0.03614859
Magnitude	0.68169307	-0.269305214	-0.49689603	-0.08255919
MST	0.23670110	-0.696060043	-0.09891487	0.18526217
Amplitude	0.78261875	0.099684299	0.19235189	-0.05661740
MSP	0.40229978	-0.681550073	-0.51846547	0.11336428
Latency	-0.33863004	-0.534797112	0.20927330	-0.06140924
GazeTime	-0.42116328	-0.499458543	0.21014509	-0.21925359
MPeak	0.42262360	-0.475270412	0.47930843	0.08443821
PupilChange	-0.52846149	-0.140272754	-0.20588394	0.39535161
MappedFX	-0.67824872	-0.385824669	-0.02550196	-0.04643184
MappedFY	-0.60634996	-0.300776547	-0.12788957	-0.02986667
FixationD	0.06733977	-0.007967062	0.05710121	-0.59468121
SaccadeSize	0.10191412	0.127059497	0.26633311	0.71979178

To determine the contributions of the principal components to the data, or contributions of significant variables to the four principal components used to summarise the data, the result table below from the PCA was used. The amplitudes of the user responses had the highest contribution (20.47) to the dataset and also 0.78% significance to the principal components (Table 5.7).

Table 5.8: Contributions of variables to principal components

	PC1	PC2	PC3	PC4
Baseline	3.486787	12.640109	16.0804323	0.1137692
Magnitude	15.531983	3.063686	20.4386435	0.5934345
MST	1.872620	20.466716	0.8099239	2.9882394
Amplitude	20.471499	0.419767	3.0627668	0.2790884
MSP	5.409396	19.622317	22.2515716	1.1189081
Latency	3.832658	12.081834	3.6253391	0.3283291

The most important or most contributing observations are indicated in Figure 5.8. The sum of each row is 1, since we considered the 13 numeric components. The sum of all the contributions per column is 100. The cut-off to determine

the elements that have the most impact is the average of the total contributions by the elements, which is indicated by the red dashed line on the graph. All components above the red dashed line can be said to have a distinctive attribute and characteristic possessed by each users to the dataset, which is related to stress or the presence of contents that induces stress.

The webpages with static contents, such as the National Rail Enquiry (NR1)and Google page (GoogleSearch1) with deactivated contents, has a high contribution to PC1, while dynamic pages like Yahoo and iGoogle also have a high contribution to the PC1. This implies that most individual stress level are increased during interaction to these webpages, given the high quality of these pages when compared to others.

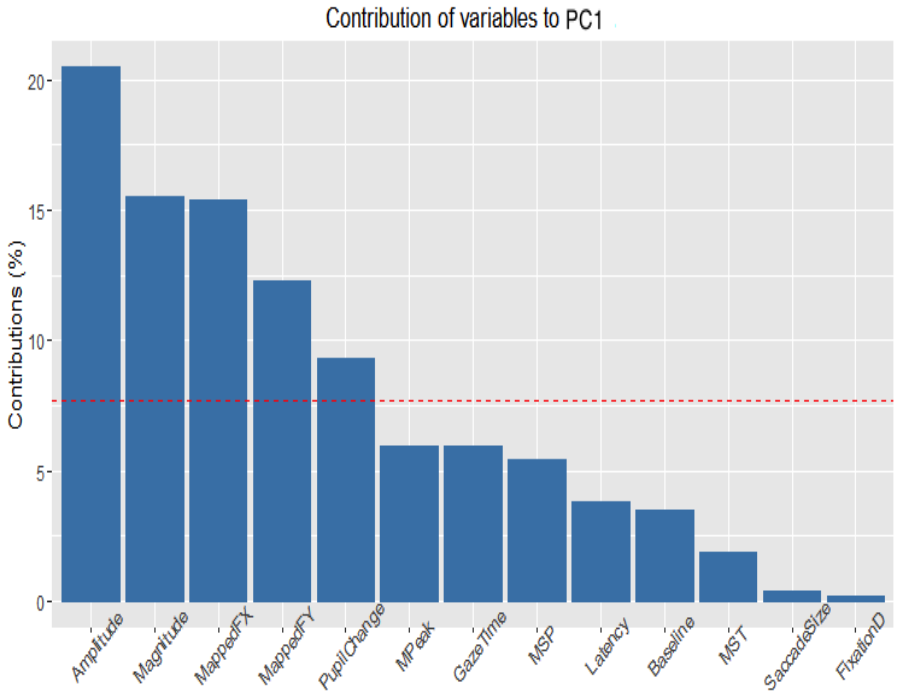


Figure 5.7: Percentage of Contributions of variables to PC1

The coefficient (0.78) of the contribution of the most significant variable shows the validity of the performance for the three models. From Table 5.1, where $M1$ represents the original data, the highest performances of these three models are

0.90, 0.85 and 0.86, which are within the range, or higher than, the coefficient value of the contribution made by the variable to the PCA. Hence the choice of the models is valid.

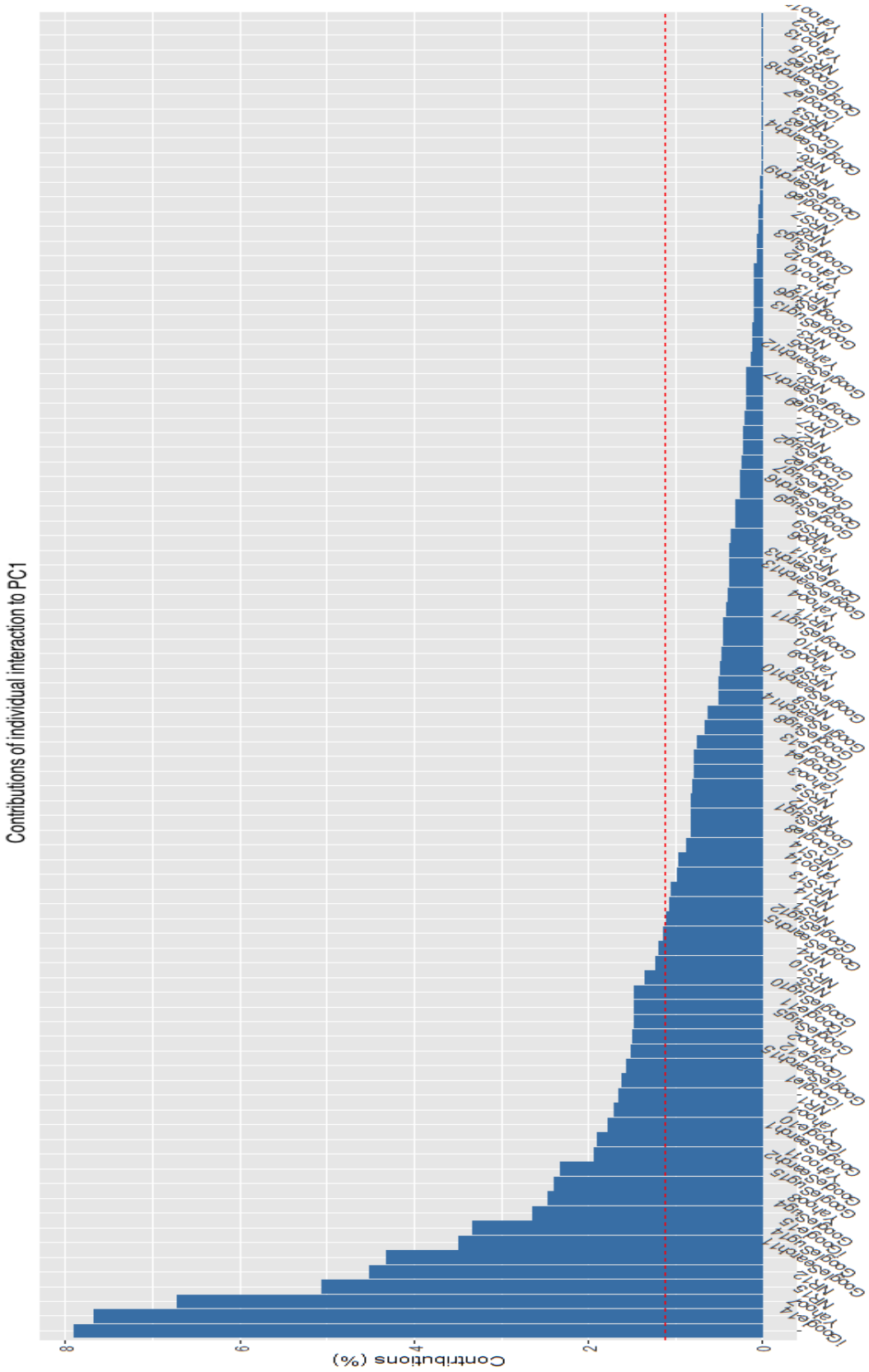


Figure 5.8: Percentage of Contributions of Individual interactions to PC1

5.2.2 Verification by FS

For the purpose of FS, we try to elicit individual observations and their corresponding stress levels. We demonstrate that the pattern detected agrees more with PHYCOB I than with the other methods, starting with the first two variables, original data, standardised data and data with outliers excluded. The secondary data generated by the PHYCOB I model was used to implement the FS algorithm. To detect natural structures based on the data collected, 164 data points were used as a form of training data out of 264 of the original dataset, this is to determine the best fit for the models, since we are also considering the prediction of events. The results of the outlying cases and masking effects were visualised using stalactite plots. The algorithm selects a starting point of minimum residuals that serves as the performance for a given subset. For this particular case, the starting points 5 and 9 produced absolute residual of (0.98). Figure 5.9 shows the residuals of the first two variables used as input to the dataset. The simulated response is based on normalised data with zero mean and a standard deviation of one. The cut-off point normally employed to define an outlier is the highest predictable value for the group (n) of random attributes having a chi-square distribution of p degrees of freedom described in Chapter Three.

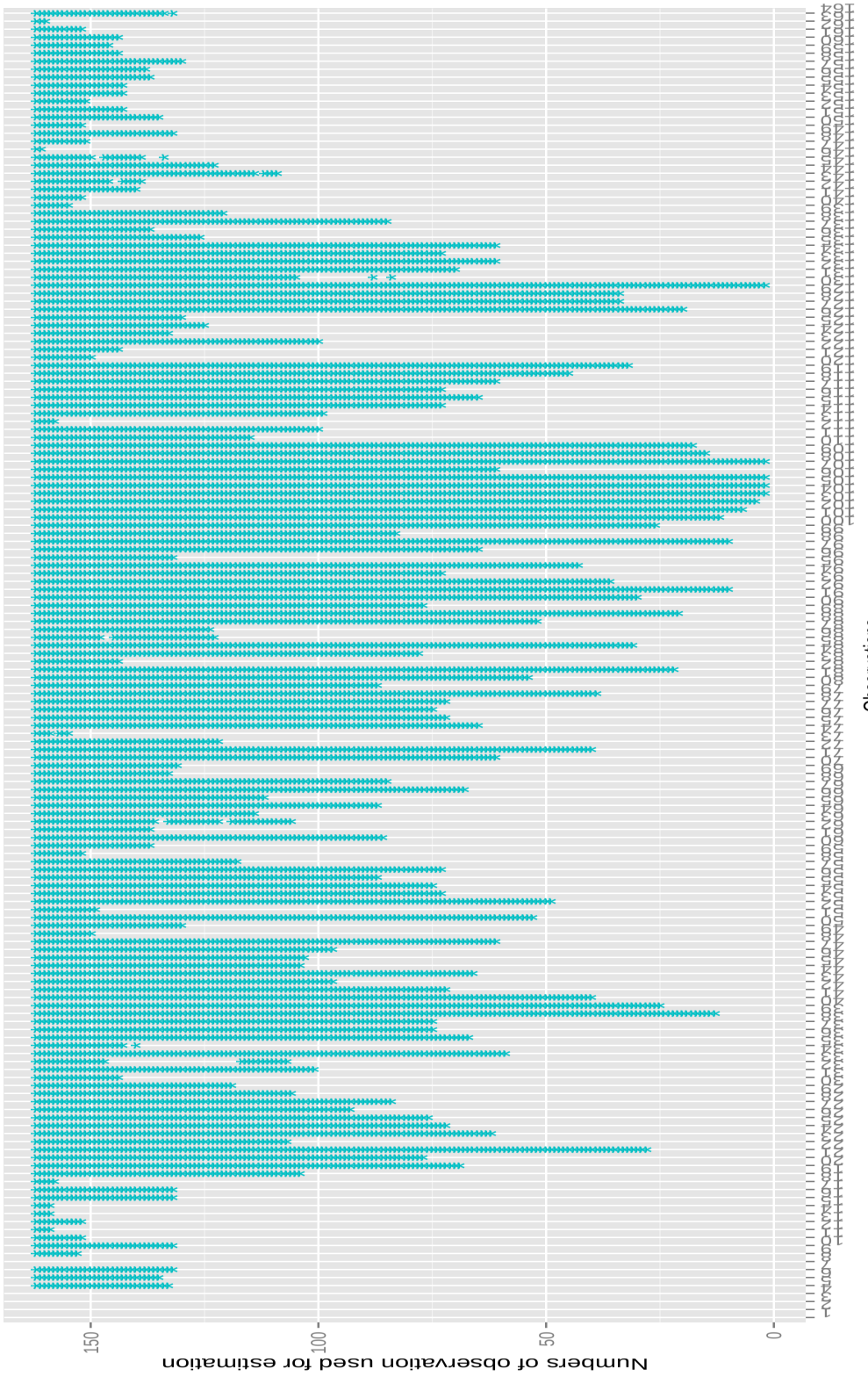


Figure 5.9: Stalactite plot of model with two variables

The masking effect of the outliers is shown by their distance from the full sample size. In Figure 5.9, nearly all the observations are initially termed as outlying given the fact that the residuals are large enough. For $m > 85$, instances 22, 39, 40, 82, 89, 92, 98 were detected as outliers and for $m > 98$, observation 100 through to 106 are also outlying, and also observations 127 and 130. The outlying instances that remain consistent are 100-106. The outlying cases can sometimes depend on how large the residuals and variations of the variables are, to find out whether or not these residuals move with the indexes, observations 100-106 was moved to the top of the datasets (Figure 5.10). These observations matches participant 17's interaction with the National Rail Enquiry suggest/search page (NR17 and NRS17), participant 18's interactions with the Google suggest/search page, Yahoo and iGoogle page (GoogleSug18, GoogleSearch18, Yahoo18 and iGoogle18), as is indicated in Figure 5.11. This can either mean participant 17 and 18 were either stressed or not affected at all by these pages. Section 5.2.3 further discusses the best fit for the stress levels detected.

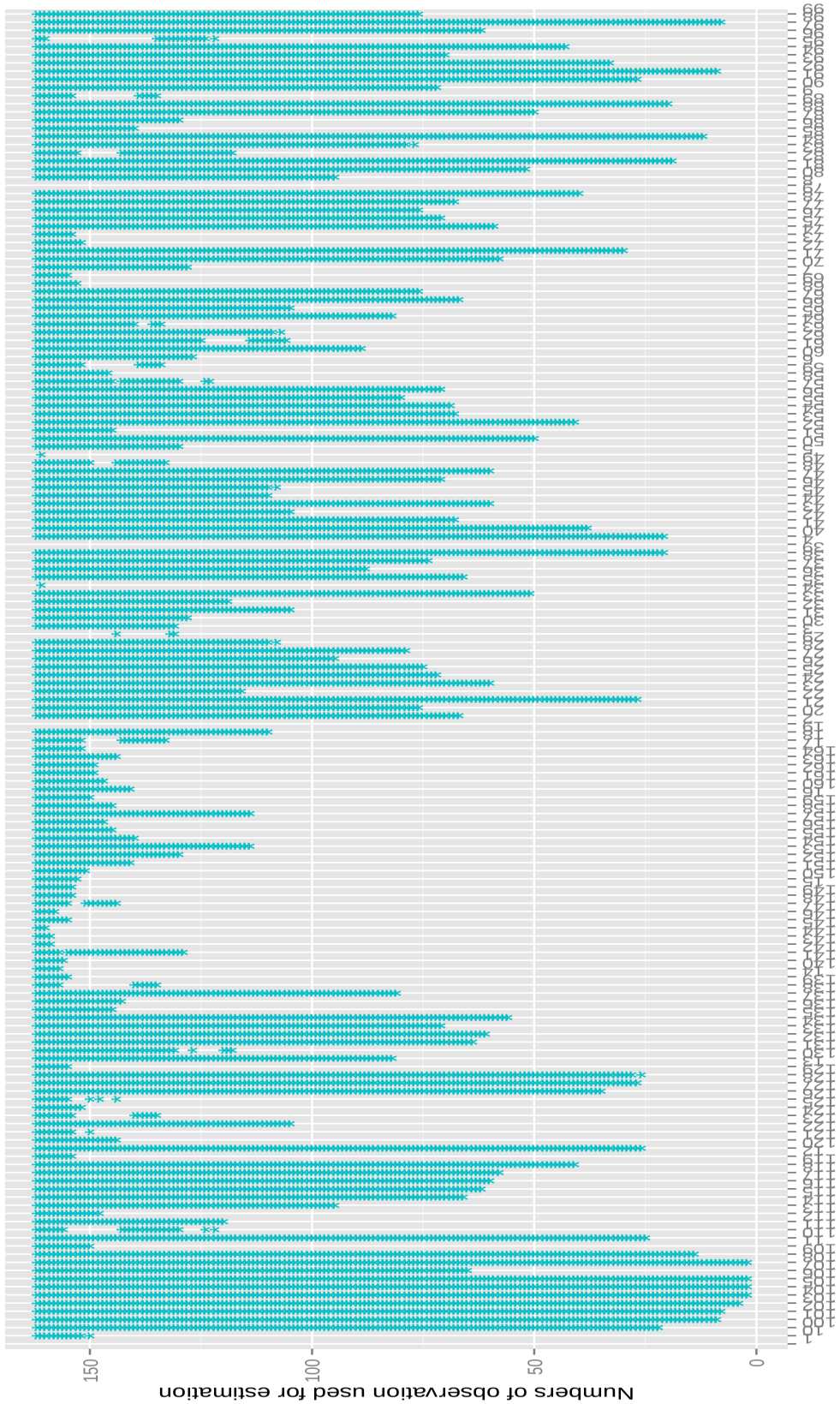


Figure 5.10: Stalactite plot of model with outlying instances moved to the extreme of the dataset

Figure 5.11: Stalactite plot of model with two variable showing participants interaction with webpages as outlying

As the number of input variables increases to contain all datasets, some consistent outlying cases ceased to be outliers (Figure 5.12). This implies that the variations in the dataset were large enough to be the source of such an occurrence, or changes in residual indexing. This can be observed by the orthogonality of the detected variables and the variance plot in Figure 5.6, showing the dependability and nature of the data. To determine if larger variations among variables can cause changes in stalactite output, a standardised version of the data was produced by taking the logarithm of the dataset and compared with the original data. Figure 5.13 shows that minor, but significant differences, exist in the residual plot for both standardised and original data. The first starting point for this, is observation 5 (NR1) with a performance of 89% (absolute residual = 0.11) and the second starting point is 15 (Yahoo3) with a performance of 99% (absolute residual = 1.1).

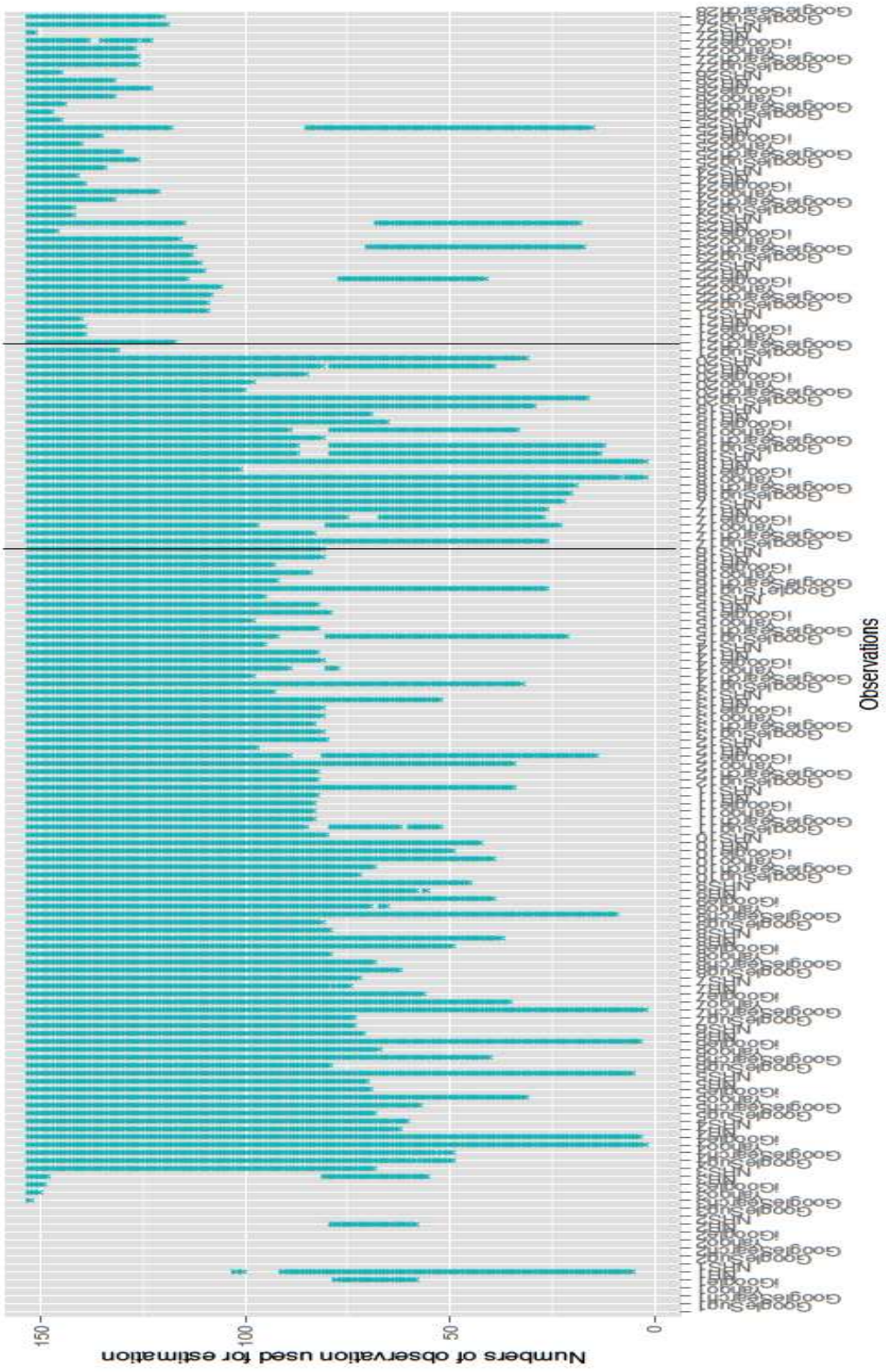


Figure 5.12: Best fit using original data as input

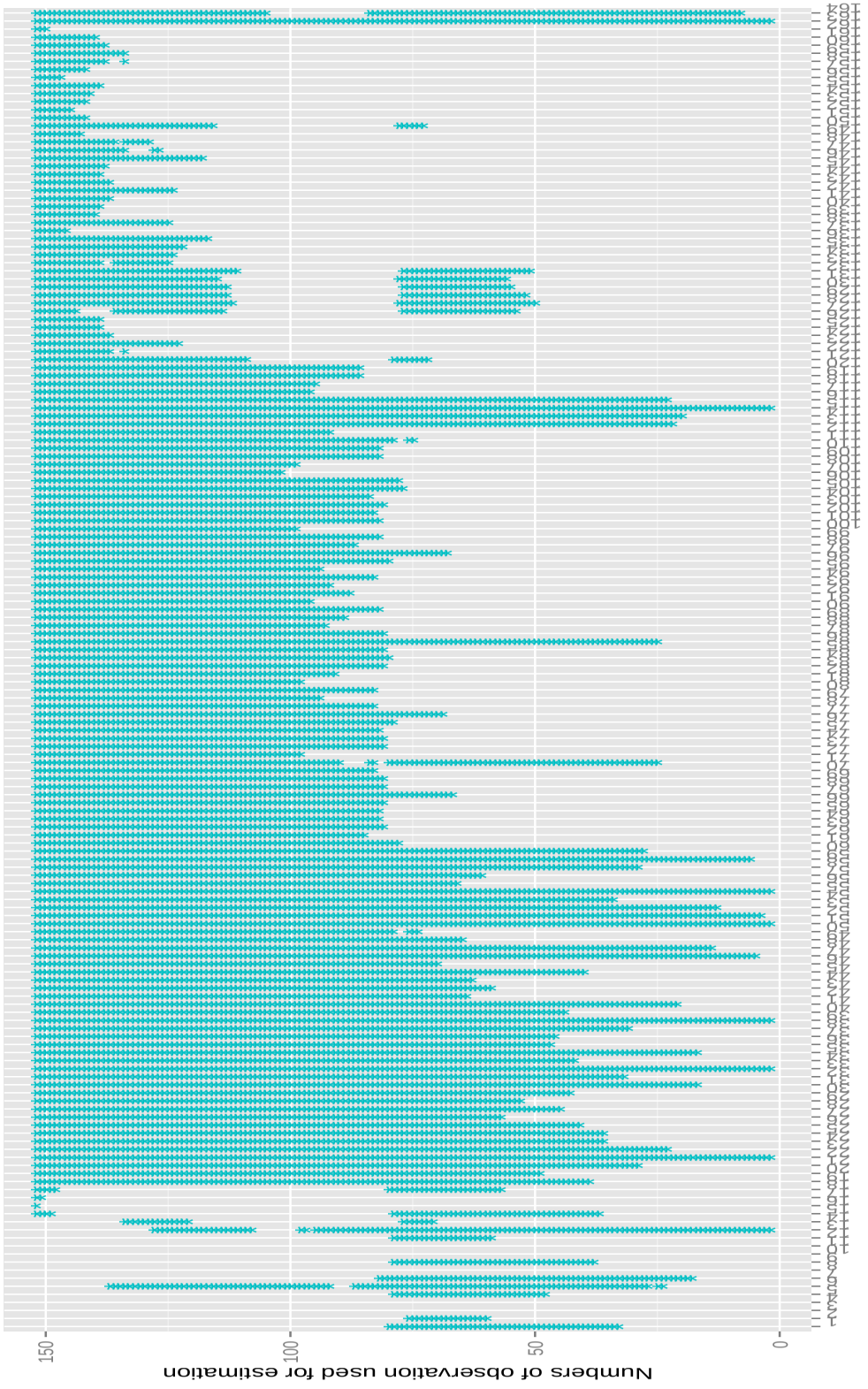


Figure 5.13: Best fit using stardardised data as input

For the remainder of this chapter, the standardised version, original data (normalised data), two variable model and also the outlier excluded version of the dataset was used to represent the input variables. To compare the performance of all models used, the linear model was adopted for the purpose of associating its residual output with the other models used in the report. As mentioned earlier the linear model (regression) is similar to the classification models; the only difference is that the responses are continuous. Given a single value input $x_i \epsilon \mathbb{R}$, and a single real-valued response $y_i \epsilon \mathbb{R}$, a straight line and a quadratic function were fitted to the data for model comparison.

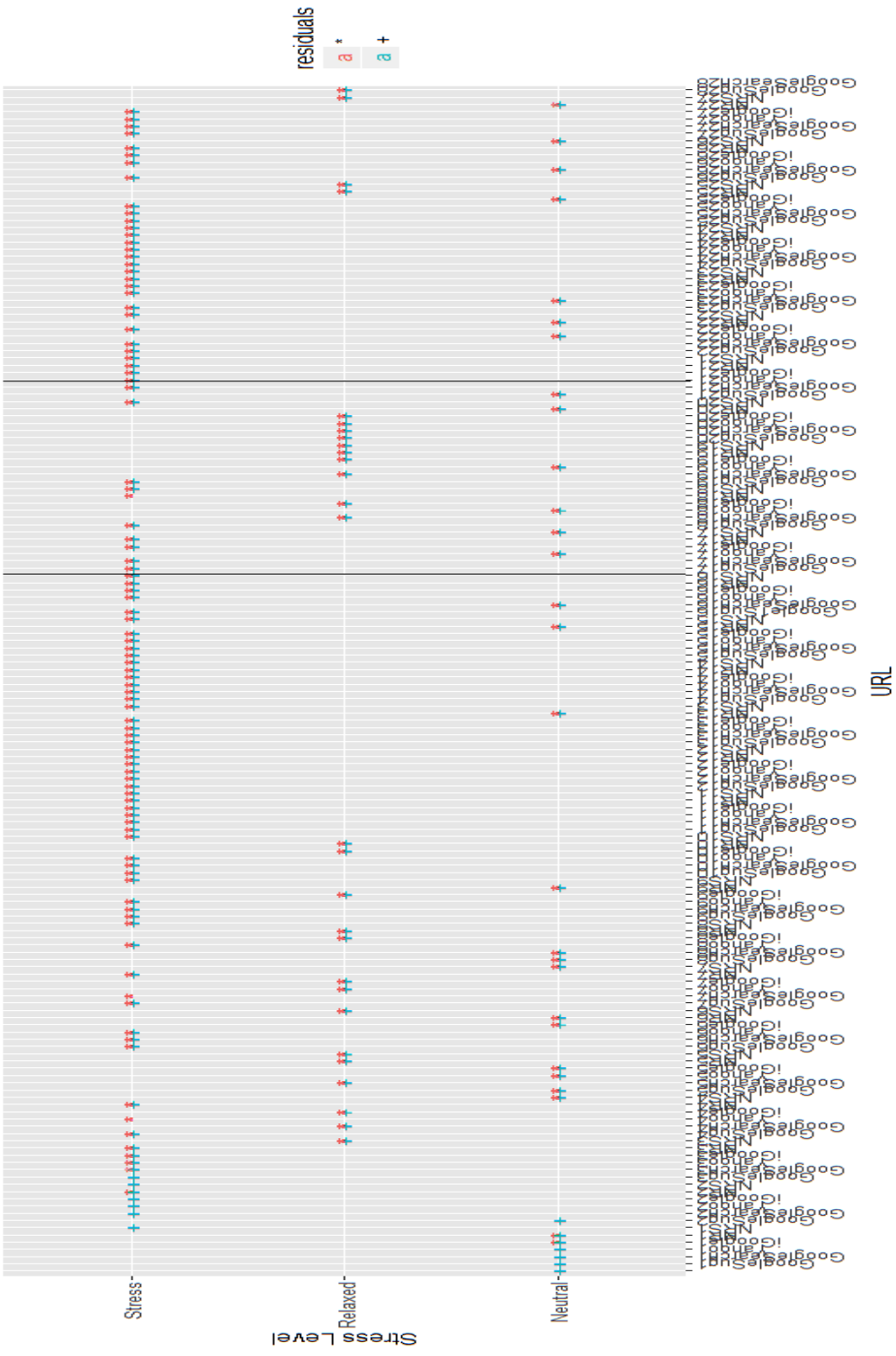


Figure 5.14: Detected stress levels to webpages indicated by adjacent red and green dots from original data of PHYCOB I

The residuals from both the original dataset and the standardised one indicate that for $m > 114$, observations 13, 22, 31, 33, 35, 39, 47, 48, 51, 52, 52, 55, 59, 115, 163 and 164 are outliers (Figure 5.12 and 5.13). While for $m > 41$, observations 7, 53, 113, 114 and 116 are outlying, with most of these outlying cases being detected as stress reactions (Figure 5.14) to both dynamic and static webpages. This gives an adjusted R score of ($r^2 = 0.62, p < 0.05$) with an intercept at 0.21; the fit is based on the parameter with the least p-value (baseline with p = $1.2e^{-13}$). The graph illustrates the R-squared values of the linear model of fitted responses and observed responses. Given that the p-value is too low to indicate variable significance, the covariance is not adequate enough to term the model as significant. This could be due to the presence of the outlying cases, or possibly the model could be further improved by removing the outlying cases and reducing the number of variables.

From the original data, participants 4 and 12, who interacted with the National Rail Enquiry search page (NR) were detected to be the outlying cases, likewise 14 and 18 with the iGoogle page, and 7 and 4 with the Yahoo page. These participants experienced stress during their sessions. The appearance of these instances on the residual and leverage plot in Figure 5.15 also substantiates this further.

Based on findings above and from the literature, sometimes R-squared is not useful in determining the biases of the coefficient estimates of models or to indicate whether a regression model is adequate, and for this reason, we visualize the residuals of the model with stalactite plot, linear model fit and also the classification models used in section 5.1.The leverage plot shows the model is of high-leverage and at a high-standardised residual point.

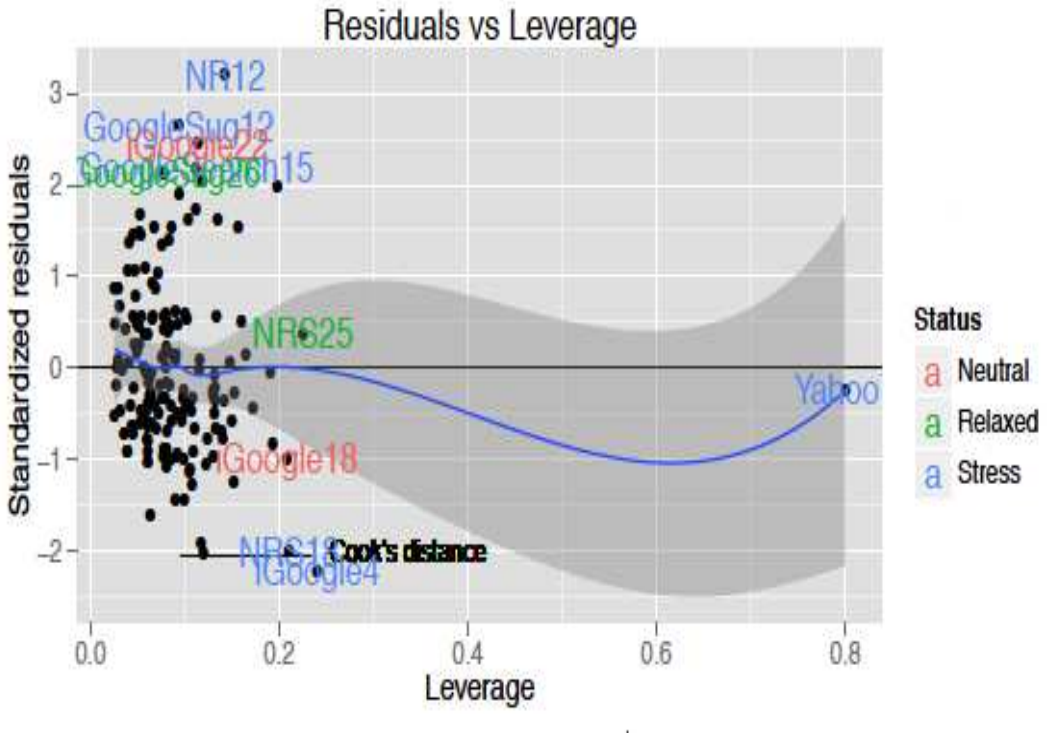


Figure 5.15: Residual and Leverage plot for all data

From the leverage plot, the data seems to be clustered towards the end of the model fit. Generally, this might represent a problem related to the clustering of the dataset, but here the presence of the outlying cases means that it cannot be straightforwardly categorised as a problem with clusters (Fraley & Raftery, 2002; Rocke & Woodruff, 2000), as has already been confirmed by PCA; the data are continuous and time series related. An instance on Yahoo page for participants 4 and 7 shows a special outlier case; this also implies the occurrence of a significant level of stress for these participants, although it could also be due to an error in measurement. To determine if these outlying cases are influential, we considered excluding them from the data.

5.2.3 Outlier Exclusion

The presence of outliers can lead to inflated error, low R-squared in residuals and significant misrepresentations of parameter estimates when a large number

of parameters are involved, although not all outliers are anomalies and also not all anomalies scores turn-out as outlying. Most of the outliers detected were extreme and out of the range of normal data due to their large residuals. The decision to remove them was based on the fact that the R-squared showed a significant score, although the significance is not strong enough to consider the model as substantial, while the p-value is very low, implying that it is important to model all parameters. The stalactite plot was used to identify the extreme outliers and these outlying cases were removed by simply applying choices for outliers greater than the cut-off to be excluded, given has:

$$Z_{new} = Z - (Z_{list} > Z_p \frac{(n - 0.5)}{n})$$

where Z_{list} are the list of outlying cases. Figure 5.16 shows the best plot for data with outliers removed; the starting point for this is for participant 3 on the iGoogle page (iGoolge3) with a mahalnobis squared distance of 5.36. From the plot it also appears that for observation ($m > NR12$), which corresponds to participant 12's interaction with the National Rail Enquiry page, observations NR1, NR4, and NR7 are outlying. It can also be noted that the page's contents which the participant interacted with has a higher number of text-boxes with automated list (ASL) deactivated, meaning that the participants had to type in the text without assistance. This implies that these participants experienced significant stress during their interaction session.

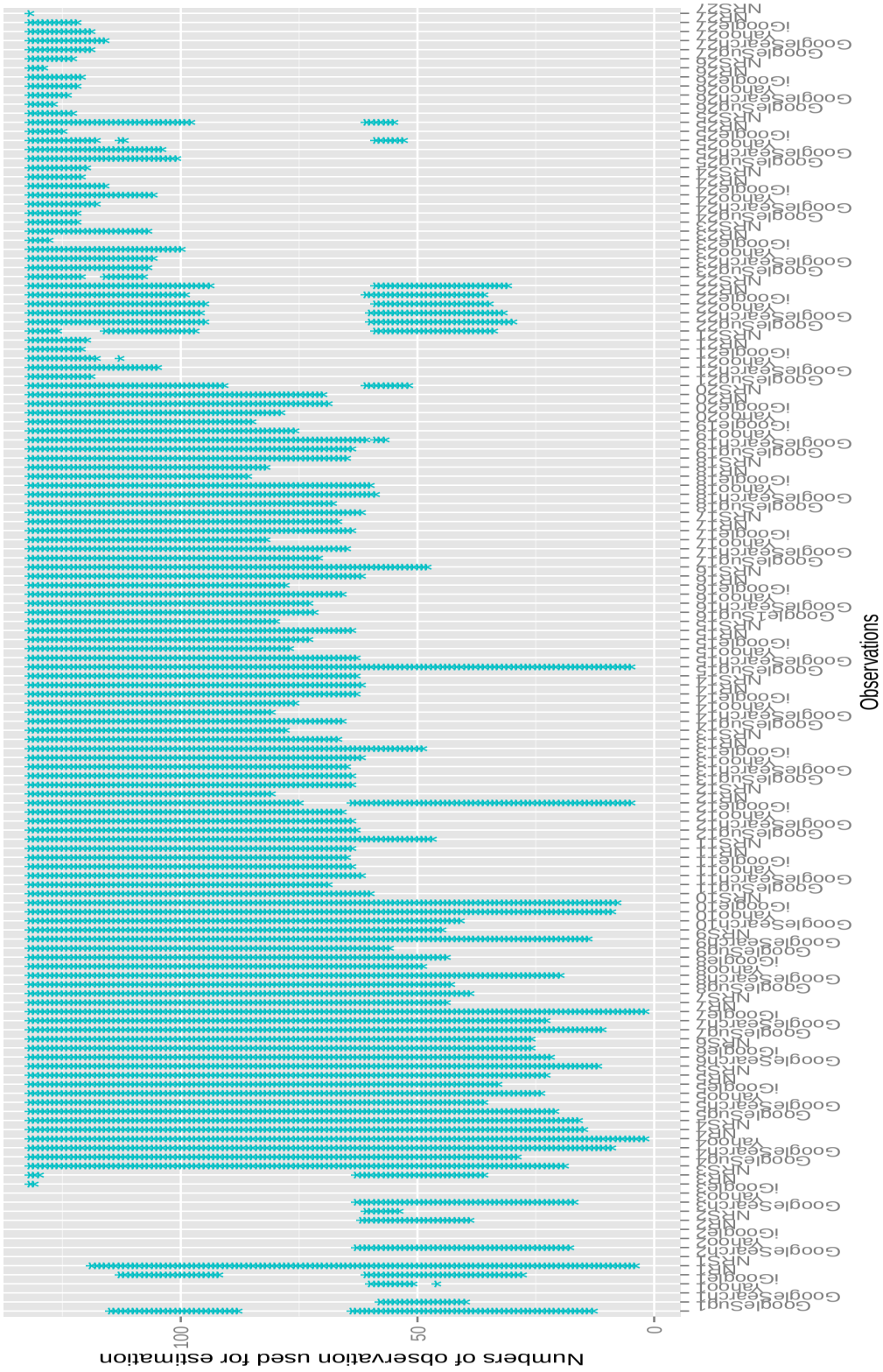


Figure 5.16: Best plot for model with outliers excluded

In addition, participant 12 on the National Rail Enquiry webpage seems to be consistently, an outlier and is also a not affecting the performance of the model, given the adjusted R-squared ($r^2 = 0.73, p < 0.05$). The output shows that the model is very vital. Also, trying to eliminate these cases reduces the R-square and p-value and hence they are dominant.

Figure 5.17 shows the detected stress level corresponding to the webpages, the red dotted point indicates residuals greater than the cut-off, while the blue dotted points show residuals at the same level and equal to the red dots. All the influential outlying cases turn out to lie within the upper-most level (stress), except for NR1 which is neither stressed nor relaxed (Neutral). From the plot, participant 4 on the Google suggest page (GoogleSug4), representing instance 17, was detected as relaxed given its low residual and lies in the middle level. Likewise, instance 18, which also represents participant 4 on the Google Search page (GoogleSearch4), was detected as stressed. It is very apparent that on pages where ASL is disabled, participants find the task difficult and this increases their stress level. This also responds to the problem in Figure 5.10 in that it detects the stress level of participants 17 and 18, who were among those who interacted with ASL disabled pages.

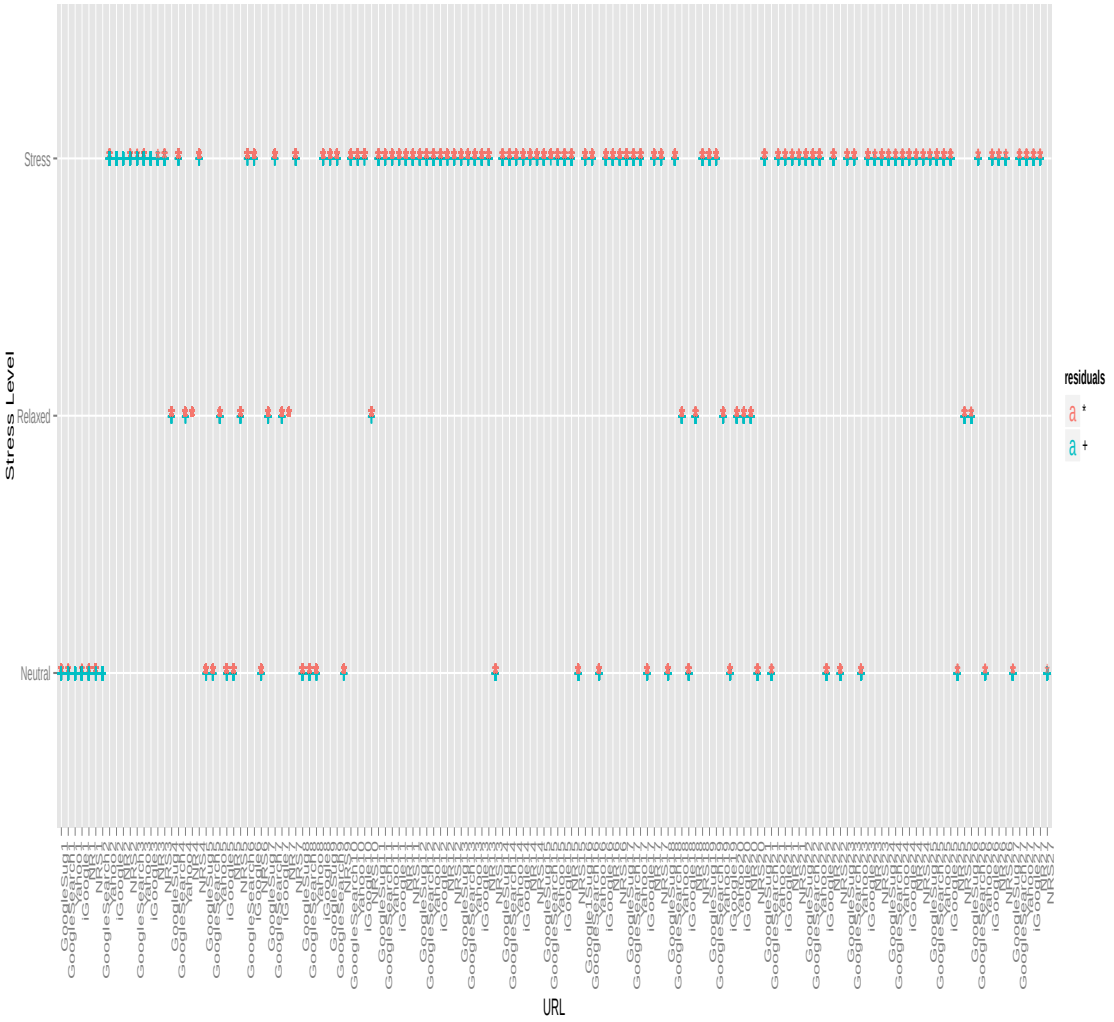


Figure 5.17: Stalactite plot indicating detected stress levels of model with outliers excluded

With the outliers excluded, the data in the linear model fit in Figure 5.18 seems less clustered than that of Figure 5.15, so the choice of the model is still valid. This shows that the data are diverse and requires no clustering techniques, even with the standardised form.

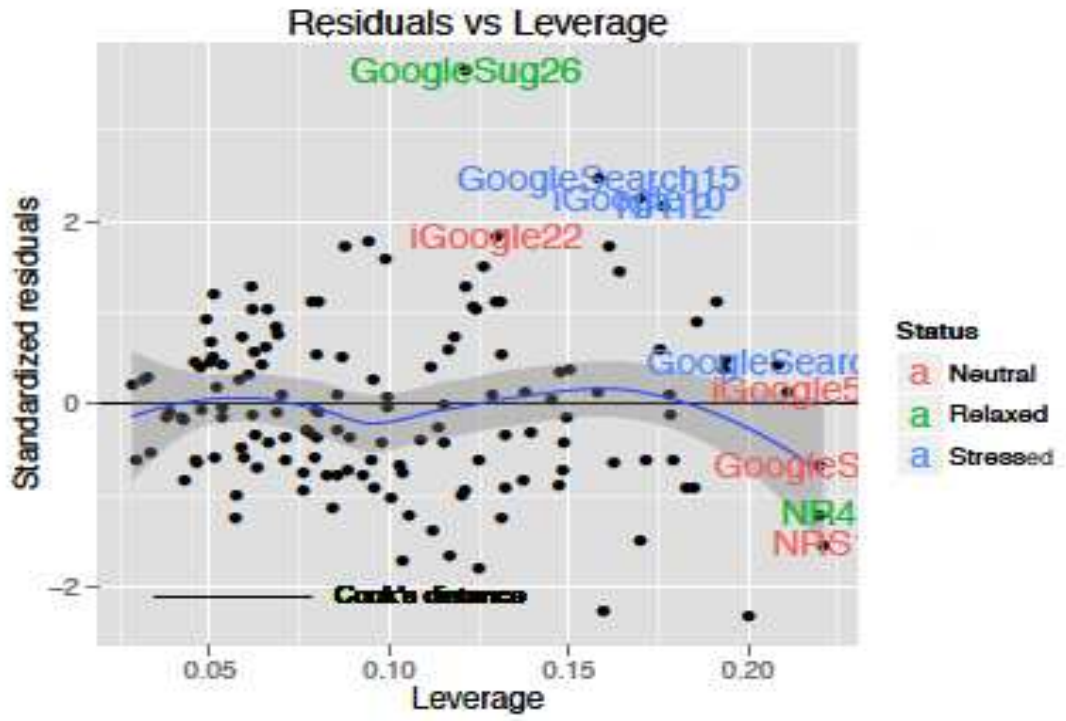


Figure 5.18: Residual and Leverage plot for data with outliers excluded

The two-variable model seems to have the best performance score of all stalactite plots, but it also happens to have the least R-squared (in Table 5.9). The second best solution following this is the original data with all variables included. But the model with all data and outliers removed seemed feasible given its good average performance score and good R-squared value. As mentioned, R-squared does not indicate whether a model fit is adequate (Chatterjee & Hadi, 2015; Faraway, 2014; Nakagawa & Schielzeth, 2013; Pepe *et al.*, 2008); the table simply indicates low and high R-squared values.

The patterns in FS detects all possible observations with their corresponding stress level with more true positives than false positives for all input data used for FS, these patterns agrees more with all the models especially with PHYCOB I, as indicated by the confusion matrix in Table 5.2. Although not all the stalactite plots gives pure class membership, indicated by red dots without green dots, the presence of both dots in the same position reflects pure class membership.

Table 5.9: Table with R^2 and Stalactite performance score

Models	R^2	1^{st} Start Point	2^{nd} Start Point	Ave abs Res	Ave. Perf Score
2-Variable model (MOD2)	0.17	GoogleSug1	GoogleSug2	1.6	99%
All-Data Standard (MOD3)	0.32	NR1	Yahoo3	1.1	94%
All-Data Original (MOD1)	0.62	GoogeSug1	Yahoo3	1.1	94%
All-Data Outier exluded (MOD 4)	0.72	NR1	NR3	0.5	83%

In HCI and psychophysiology on human behaviour, it is naturally probable to have R-squared values less than 50%, because human behaviours are less predictable than physical processes and hence a model with two variables can also be suitable for modelling the HCI-HPR associations (Bayer & Glimcher, 2005; Kurzban & Houser, 2005). The low R-squared of the two variables does not affect the interpretation of significant variables. Sometimes, high R-squared does not automatically suggest a good fit for models (Bayer & Glimcher, 2005; Kurzban & Houser, 2005) as in the relationship between the physiological data and URL in the model fit above. Based on this, the average performance score for residuals in the stalactite plot is also used to determine the best choice of the model. To validate further, we used other classification techniques to determine and evaluate performances in section 5.1.

5.2.4 Model Optimisation and Best Subset Selection

Determining parameters to be selected and excluded in order to improve the model performance, involves stepwise and criterion selection mentioned in Chapter Three. Eliminating or accumulating variables can either lead to perfection or a decline in model performance. This hinges on the p-values and the level of status of the parameters assessed. Carefully selected features could improve model accuracy but adding too many may lead to overfitting the model. This could lead to poor predictive performance on the test data that is intended to be used. Since the thesis is interested in the association of the physiological data (X) and the users' eye movement (Y) conditioning on the variables, the main focus lies in detecting and predicting stress levels from the data collected, using

the hypothesis $H_0 : \beta_j = 0$ and $H_a : \beta_j \neq 0$ where β are the coefficient or residuals sum of squares ($RSS(\beta)$).

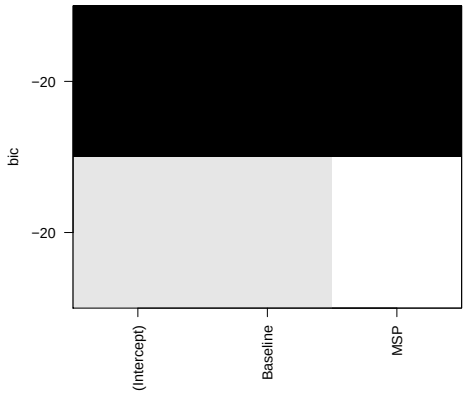
The Table 5.10 shows the outcome from running the process on the models set out in Table 5.9. The forward stepwise method includes both AIC and BIC. To determine the subset for the data, the procedure approximates the response variable y with a constant (intercept) and gradually adds one variable at a time. The one that yields the best accuracy in prediction when added to the subsets is selected.

Table 5.10: Table with aggregates of feature selection method results

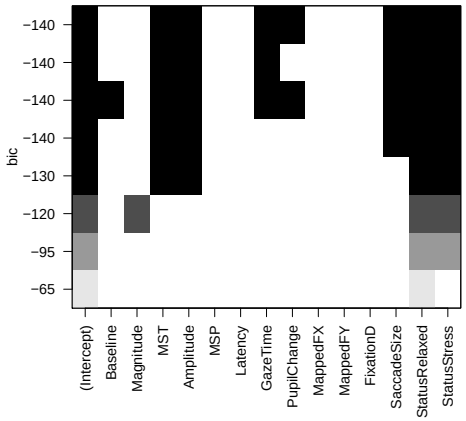
Model	Adj R^2	BIC	AIC	P-value	RSE
Two-Var model (MOD2)	0.19	-20	200.68	$2.8e^{0.8}$	1.83
All Variable Standardised (MOD3)	0.28	-30	-147.26	$1.9e^{-08}$	0.63
All Variable Normalised (MOD1)	0.58	-120	88.6	$2.2e^{-16}$	1.29
All Variable outlier Removed (MOD4)	0.70	-140	332.02	$2.2e^{-16}$	0.3

Based on the AIC selection, the model with two significant variables retained its parameters, since both had equal AIC value, with intercept at 1.76 and a baseline of SCR as the most significant ($p = 1.15e^{-06}$). For the standardised data, the AIC selection excluded eight variables with an amplitude of SCR as the most significant ($p = 3e^{-06}$).

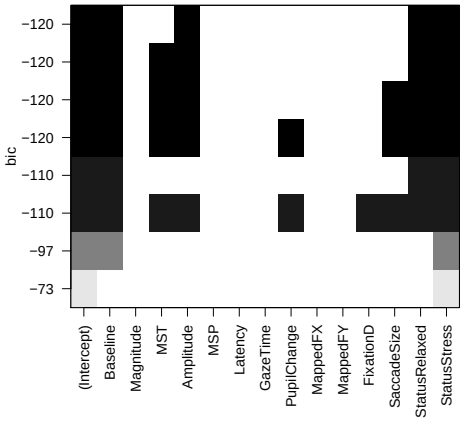
The normalised data has six variables added, excluding eight parameters with baseline ($P = 0.00133$) and status ($P = 3.47e^{-08}$) being the most significant, while the model with outliers removed has nine parameters and excluded five, retaining amplitude as the most important ($p = 0.00193$) along with status ($p = 4.62e^{-10}$). From all indications based on AIC, amplitude, baseline and status appeared to be the most significant in all models which shows the prominence of these attributes in predictions of HCI-HPR relations.



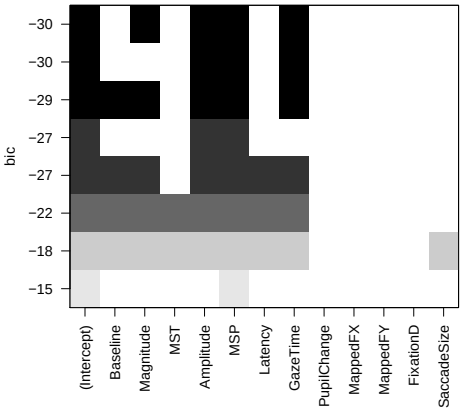
(a) BIC of model with two variables



(b) BIC of model containing outliers



(c) BIC of model with normalised data



(d) BIC of model with Standardised data

Figure 5.19: BIC of the four types of models used as input for FS

In Figure 5.19 the diagram indicates the features and attributes added and excluded based on the BIC selection. The black box indicates attributes added while the white box indicates attributes excluded. The model containing all variables tends to minimise the BIC. The BIC values of models on the y-coordinate indicate the top four models having the same value for data with outliers removed and the normalised data while the first two top models in standardised data had

the same BIC values, this explains the inconsistency in the results. From Table 5.10, the model with all variables and outlier removed, seem to have the least residuals standard error (RSE) which also validates the outputs of Table 5.9.

5.2.5 Performance of Models from Forward Search Algorithm

As part of the validation process, the ROC was also used here, since it contributes also to giving more insight in the data model to use for prediction by the three models (Neural Network, Logistic Regression and PHYCOB I). Based on the predictions obtained for the test set of all models, the diagnostic plot of Figure 5.20 shows that models MOD3 and MOD4 seem to have the same predictive performance. These are the standardised and outlier excluded models. The output here shows that the exclusion of outliers does not affect the performance of results. MOD5 was produced based on taking a smooth function of the attributes and shows the model with the least performance of the datasets.

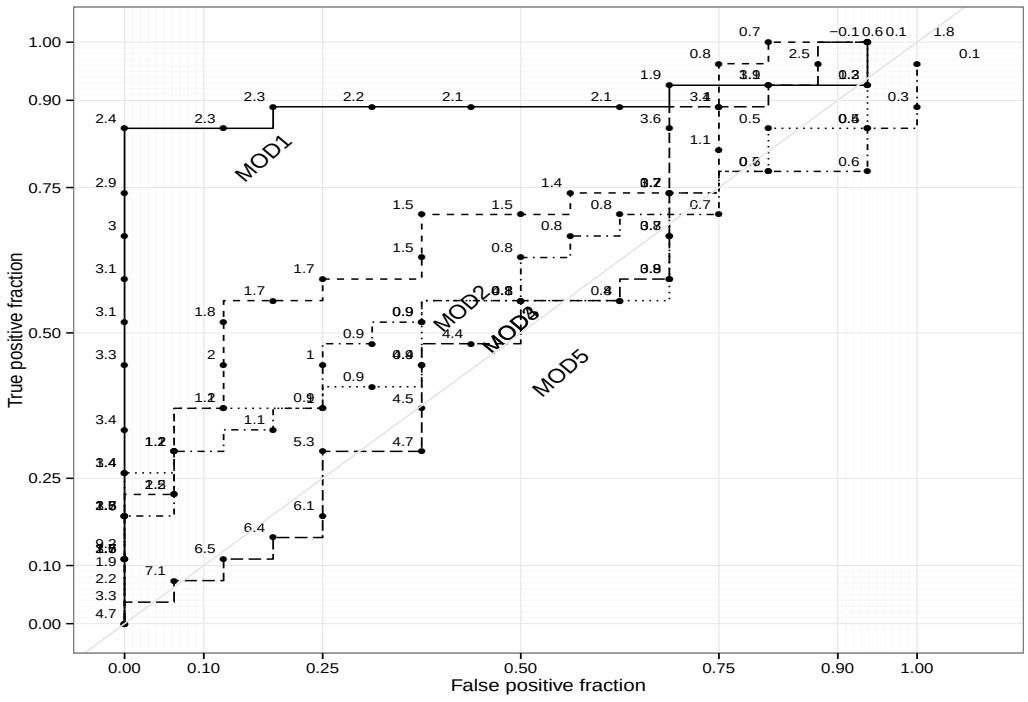


Figure 5.20: Diagnostic plot of dataset models from the forward search algorithm.

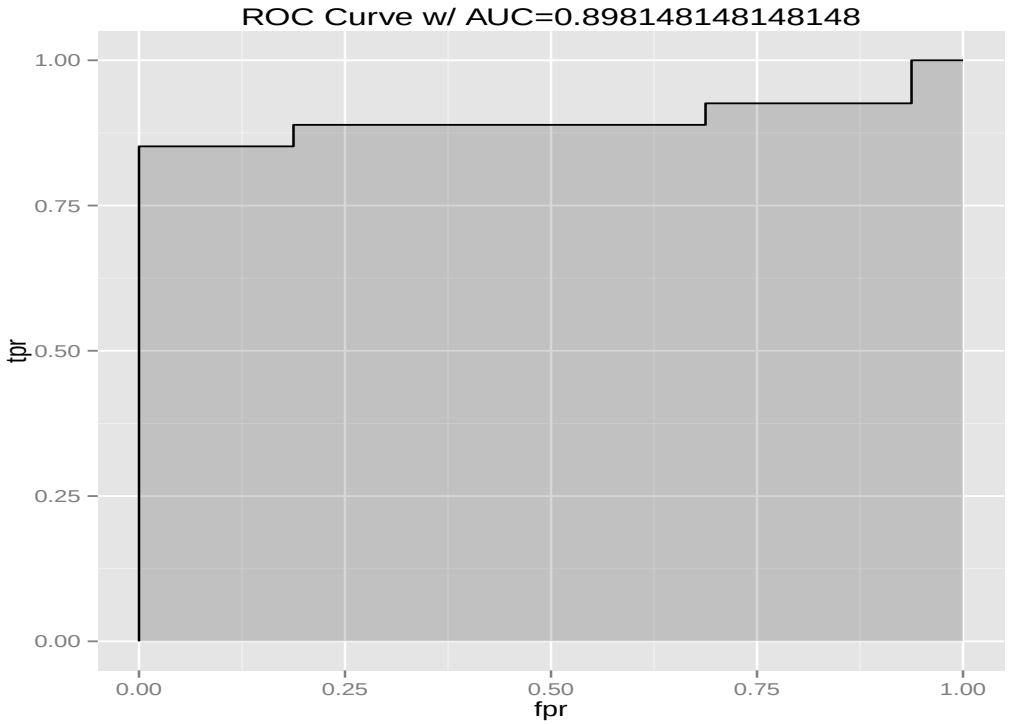


Figure 5.21: Diagnostic plot for the best dataset model from the forward search algorithm.

From the ROC plot, the model MOD1, which corresponds to the original dataset (normalised data) seem to do better than the others at a cut-off of 2.4. The AUC is 0.89 which is close to a very good model and compares closely to the performance of PHYCOB I at the 70% split in Table 5.1 also indicated in summary Table 5.11. From the analysis of all the current models, it appear that MOD1 has one of the lowest p-value as well as an adjusted R-square of 0.58, which is normal for the scaling of human behaviour data. Hence, using a normalised data model can prove a high predictive ability for behaviour data such as where users are likely to look at on webpages or on visual stimuli like games.The most vital part of the process is to be able to predict the stress level of users during interaction based on some physiological attributes. This will assist in effective design decision making in a system interface scheme.

Table 5.11: Connotation of the three methods with PCA/FS

Connotation of methods				
Methods	Feature Selection		General Performance	Ave. Predicted Stress Point
	Optimal Features	Similarity with PCA/FS		
PHYCOB I	4	95%	0.90	78
Logistic Regression	3	80%	0.85	78
Neural Network	3	75%	0.86	77
PCA	4	100%	0.91	78
FS	4	100%	0.89	78

From Table 5.11, it is seen that based on the different methods used for comparison and validation purpose, PHYCOB I agree mostly with PCA and FS in terms of feature selection, general performance and on the average predicted stress levels for all the methods. The Neural network and Logistic regression are two predictive models while PCA and FS are ideal for model reduction and to elicit individual stress levels and their corresponding observation that uses a different form of pattern recognition. Based on this, four most important features were selected by both PCA and FS which closely agrees with the four attributes selected by PHYCOB I with a similarity of 95% compared to the two predictive models. The process validates and justifies the reliability of PHYCOB I for predicting stress levels and for determining individual stress level of users and the webpages they interacted with.

5.3 Contributions of this Research

In the light of the above performances, PHYCOB I outperforms the other models used for comparison and justification purposes. The major contributions to knowledge achieved by this research project is mostly centred on the PHYCOB

I model, whereas the other predictive models mentioned are used for validation purposes. The models were trained and tested on the labelled data described in Chapter 3 and in Appendix B.1 using a hold out cross validation.

As a novel grouping and data organisation algorithm, the performance of PHYCOB I was compared with those of standard techniques such as Logistic Regression and Neural Network by running each of the models under multiple settings. The core contributions of this thesis are outlined and presented in the sections above. The initial stage was to look at the implementation of PHYCOB I, using a validation technique to determine the nature of the generated dataset with PCA and later with the forward search algorithm. In addition, the main idea was also to carry out a comparative performance between the PHYCOB I algorithm and multiple versions of two standard techniques - Logistic Regresson and Neural Network. From the comparative analysis, PHYCOB I outperforms these two standard methods from the best selected models on different network architectures including Neural Network and three feature selection methods of logistic regression.

The attributes on PHYCOB I also agrees with the detected principle components in PCA (Figure 5.6) and the forward search algorithm in respect to the detected stress levels in Figure 5.17. These processes demonstrates the model's significance, reliability and thus the achievements of the research work.

Possible ways to employ PHYCOB I is to use it as a supplementory tool to eye tracking. Currently most eye movement interface in eye tracking can simply visualize users' gaze plot either with fixation points or heatmaps without indicating stress affected areas. PHYCOB I can be embedded to help identify stress points, in order words detect contents that relates to users stress levels.

5.4 Summary

The novel approach to utilising user generated data from HCI-HPR associations is using the PHYCOB I to classify stress levels based on some computed attributes, comparing its performance with standard or existing models such as the Neural Network and Logistic Regression and also using PCA and forward search to detect outlying cases for HCI-HPR generated datasets. PHYCOB I shows highly

significant results on the dataset used. From the results obtained for the best subset using the BIC and AIC on the two variable model, standardised, original data and dataset with outliers removed, it demonstrates that without outliers the data model seems to achieve improved output both in R-square and average percentage performance. Most of the outliers detected were relating to errors in measurement that could lead to bogus and erroneous results. These outliers were removed, and their removal does not affect model performance. The data model with the best performance was used to run the predictive models including PHYCOB I. Considering the human behaviour data scaling involved, models with low R-squared are still acceptable. Since prediction is considered, the models selected for BIC and AIC method can best be used. The original data from PHYCOB I did better in terms of model performance and prediction accuracy, showing the model's reliability.

Chapter 6

Disscussion and Conclusion

This research project set out to explore a method for modelling stress level based on physiological responses to web contents, mostly on user interaction to web applications, which were used has stimuli and led to the development of an algorithm (PHYCOB I) that computes physiological parameters, and which can also serve as secondary indicator of user stress level based on individual interaction with each webpage. The study sought to answer the following research question: How can user generated data be utilised in order to model and simulate human physiological response to the dynamic and static content of webpages?

6.1 Empirical Findings and Contributions

The principal empirical results are chapter specific and were reviewed within the respective chapters (Chapter Four). This section will integrate the contributions and findings to answer the research question; How can user generated data be utilised in order to model and simulate human physiological response to the dynamic and static content of webpages?

6.1.1 A Novel Approach to Determining HCI-HPR Associations

The novel approach we adopted was to determine the physiological correlates of user interaction to webpages by developing an algorithm that also serves as ter-

tiary indicator of stress level in users. To implement the algorithm (PHYCOB I), we first conducted an experiment in both real time and with a delay. In real time, it involved participants who are familier. Data was exported from sensors (SCR and Eye tracking) that measure the physiological response of users, including the SCR, ST, Eye movement (fixations, saccade, pupil dilation), while they interact with six webpages. The PHYCOB model then computes the physiological attributes.

In delay, the PHYCOB model reads in each individual data, applys the Savitsky Golay filter already mentioned in Chapter Three and computes the physiological parameters mentioned in Appendix B.1 from the readings, which help to group the stress status of users to each webpage. PHYCOB I also identifies which physiological attributes have the most affect on user interaction to dynamic and static contents; this happens to be the average peak in the SCR of users (MPeak), mean skin temperature (MST) and mapped fixation on the x-coordinate of the webpages (MappedFX). These attributes were also confirmed in other standard techniques. To test the model's reliability and significance we compared the model to other standard techniques such as Neural network and Logistic regression. The PCA and FS algorithm were used for validation purposes to detect the natural structures arising from the data and to check their consistency with the PHYCOB I. It also outperforms the two standard techniques used for comparison with a 70% train set from multiple simulated splits.

6.1.2 Validating with PCA

The PCA is used both for data visualisation and validation, it uses quadratical or orthogonal alterations to convert the set of observations of some attributes which are correlated to each other, into a set of values of linearly uncorrelated attributes. The set of attributes or detected components are either less or equal to the original data. For this case we found that the detected components were equal to the original data from PHYCOB I. Attributes such as the MappedFX and MappedFY of webpages are strongly correlated. The description of the data obtained from output is of a dispersed nature rather than random or clustered, which justifies the choice of models we used for comparison. Using the stress

status as class labels, the percentage of explained attributes is less than 50% of the dataset's variability, which might not be a respectable sum for the total variable, but in this case is not a highly significant issue given that the resulted principal components is not more than the original data.

6.1.3 Validating with the Forward Search Algorithm

To detect the natural structures arising from the data, we used the forward search algorithm already presented in section 5.2.2 of Chapter Four, which was used for detecting multiple outliers in our multivariate data. These outliers were either errors from computation or errors in measurement, such as zero readings and values whose variations were larger than the others. The computed residuals specifies the masking effect caused by these outliers which was shown by their distances from the full sample size (number of observations). These outliers were removed, to test the performance and reliability of the model. Based on predictions on the original data, on two significant variables, standardised form of the data and also on the data with outliers excluded it was evident that the original dataset did better than the others and the performance can be closely compared to the performance of PHYCOB I using the rule of thumb 70/30 split and an R-square of 0.58 which is feasible, since in HCI and psychophysiology on human behaviours, it is naturally probable to have R-squared values less than 50%, because human behaviour are harder to predict than physical processes. Also, the parameters with the most impact were MPeak, MSP, MappedFX and MappedFY, two of which were already confirmed by PHYCOB I.

6.1.4 Validations with Logistic Regression and Neural Network

To select the best model for comparison, we first applied logistic regression, using the forward, backward and stepwise method of feature selection to obtain the model with least error compared with PHYCOB I; this was to avoid overfitting from both the original data and multiple simulated splits. The least error comes from the model with the backward approach. The variables that were optimal for

each of the forward, backward and stepwise methods for logistic regression, were Mean Peak (MPeak), MST and MappedFX, which are also optimal in PHYCOB I. These were presented in Chapter 4.

Secondly, we applied Neural Networks with different network architectures and with hidden layers and neurons between 1 to 15, which corresponds to the rule of thumb, i.e. using neurons between the number of inputs to the number of targets. Using 4 hidden layers with 5 neurons seemed to be the best option compared to all structures, since it had the least error from multiple splits with a hold out cross validation.

From the general outlook the PHYCOB model outclasses and exceeds the models by 1% performance on average. The significance of the model is demonstrated by its accuracy on forward search, PCA and the two standard techniques.

6.2 Implications

One important approach to HCI is the subjective method of data collection, this includes methods such as surveys, interviews, and think-aloud protocols. These methods have proved to be productive in gaining more insight into user interaction, experience and attributes. Evidence from several studies in the use of physiological measures (Andreassi, 2000b; Dirican & Göktürk, 2011b; Heraz & Frasson, 2007) and in this thesis, seem to point to the fact that these subjective methods might not be enough for predicting and gaining more insights into user experience and interaction.This thesis has used physiological measures in an objective approach to data collection in both real time and with a delay; developing an algorithm that detects the spikes in physiological readings to determine the stress level of user interaction to the dynamic and static content of webpages. The findings and justification of the study suggest that the model is significant enough in terms of grouping stress level and detecting physiological correlations to dynamic and static contents, which could not only be used for webpages, but for other visual stimuli like games.

6.3 Achievements of the project

The main achievements of the project is summariesed in the following steps:

1 **Novel method**: We have proposed a method to understand HCI-HPR associations based on and after a thorough review of existing methods, identified the limits and short comings of existing traditional methods such as suverys and heuristic apporach when dealing with predictions of stress levels and gaining more insights into user experience and interactions through the use of physiological measures.

2 **Novel Algorithm**: We have developed and algorithm (PHYCOB I) for detecting dynamic contents of webpages that changes stress levels. This was achieved by using users' physiological attributes to predict the stress levels and detecting their correlates to task allocated areas on the webpages which represents possible dynamic contents in the area of interest (AOI) of the page.

3 **Novel app**: By assigning controls and making this algorithm event driven for a user friendly capability, we have developed an app for controlling stress levels on webpages by simulation of physiological processes integrated with eye movement data from eye tracker. We can be able to visualise and run user eye movement (fixations and saccades) from the an eye tracker and identify contents that changes stress levels.

3 **Comparison**: Finally, we compared the algorithm with standared techniques and estsblished the model's reliability and significance interms of sentivity and specitivity when compared to Neutral network and Logistic regression. The PHYCOB I model mostly agrees with patterns from PCA and FS based on number of features selected that contributes mostly in the learning process, similarity of these features from the three methods, on the general performance and also on the average predicted stress levels for all simulated splits used for comparison and validation purposes.

6.4 Limitations of the project

Despite the achievements of the project work and as a direct consequence of the method adopted the study encountered a number of limitations, which are presented below.

- Lab work: During data collection some physiological measures sometimes indicated zero readings or low amplitudes (lack of clear peaks). This was tackled by conducting another sessions on the participant or another available user.

- AJAX tool: One of the drawbacks in the study was the use of the AJAX tool to disable the web content such as the ASL of the text box in both Google and National Rail Enquiries websites. Not all browsers support AJAX e.g. Internet explorer, which the eye-tracker uses to deploy it's URL. This issue was tackled by simply limiting ourselves to using Google chrome since it serves as a general browser for most applications and accessing it from a separate server.

- Difficulty of saving the secondary dataset within the system: The drawback of implementing the PHYCOB I model was saving the updated datasets within the environment of the system. In consideration of this point we saved the datasets in a file within the same directory as PHYCOB I, so for each user-generated output the file was updated automatically. The webpages the user interacted with were saved in the image format (static) from the eye tracker rather than by URL, allowing us to conduct analysis with the static pages and monitor the eye movement fixations and saccades.

- The use of synthetic data was also considered due to small sample size which might lead to potential inaccuracies in the results.

- The PHYCOB model has been shown to be effective on the dataset used, though only marginally. It needs to be validated against other datasets.

6.5 Recommendations of Future work

The thesis has covered the objectives and motivations presented in Chapter One, in light of this and to prepare the ground for future work we hope to look at the following:

- **higher dimensional training set and comparison with more predictive models:** other models that are applied to dispersed and clustered data should be considered. This would also be used for validation purpose on the HCI-HPR data and PHYCOB I. Higher training could also alter the performance of the proposed model, when compared to others, e.g. a complex model like the Neural Network. Depending on the model's performance in these tests further mutliple users could be simulated based on added attributes such as physiological parameters.

- **applying more physiological measures:** Simulation of multiple users with more physiological attributes derived from measures like heart rate, electroencephalogram e.t.c. could also help to facilitate the analysis and evaluation protocol of the HCI-HPR generated data. This would help to achieve better design decisions in a short period and improve on the traditional methods.

- **focus on security and performance:** this research opens the way to possible benefits in terms of predicting human behaviour in respect to the visually experience of and internet security by using the tool as an alarm trigger for sending alerts on unauthorised access or abnormal activities online. For performance, load, and stress-test purposes, physiological attributes can also serve as a components in the process of simulating a certain amount of users simultaneously in a given time interval rather than only the task completion time of users.

- **validating on other datasets:** finally the model needs to validated on other datasets to further demonstrate its significance and reliability.

6.6 Conclusion

This thesis has offered an evaluative perspective on an important aspect in HCI by introducing a novel approach to HCI-HPR, using physiological measures integrated with eye tracking to determine physiological correlations that serve as tertiary indicators of the stress levels of participants based on attributes from individual interaction. The benefits of the proposed model have proved to be reliable given the results and findings and the model has offered some solutions to the persistent HCI-HPR association, which may be sustainable in the long-term with further evaluations and validations.

6.7 Summary

This chapterr mostly presents a discussion of the findings, contributions and limitations of the project and as well as potential future directions. In particular, it covers most of the findings and errors from the models used for both comparative and validation purpose. The proposed method provides an automated way of assessing human stress levels when dealing with specific web contents. This is an important achievement in that it can be able to predict what contents on a webpage cause stress induced emotion in users during interaction. This could be applied to other areas like internet security, triggering alarm for unauthorised access or abnormal activities online which is the basis for our future work and also in testing performance.

Appendix A

Participants' Demography

This Appendix contains the demography of participants. This includes the participant's details, and the information sheet and consent form used during the study/experimental set for data collection.

A.1 Participants' Demography Sheet

Participants	Gender	Age	Occupation	Impairment	Internet use	Google	National Rail	iGoogle	Yahoo
P1	Female	24	Student	None	Daily	Daily	Weekly	Never	Rarely
P2	Male	32	Student	None	Daily	Dailly	Rarely	Never	Rarely
P3	Male	34	Student	None	Daily	Daily	Rarely	Daily	Rarely
P4	Male	36	Student	None	Daily	Daily	Weekly	Rarely	Daily
P5	Male	32	Student	None	Daily	Dailly	Weekly	Rarely	Weekly
P6	Female	25	Student	None	Daily	Daily	Monthly	Weekly	Never
P7	Female	27	Student	None	Daily	Daily	Rarely	Never	Daily
P8	Female	34	Student	None	Daily	Dailly	Never	Weekly	Weekly
P9	Male	33	Student	None	Daily	Daily	Monthly	Never	Never
P10	Male	28	Student	None	Daily	Daily	Never	Never	Weekly
P11	Female	36	Worker	None	Daily	Daily	Weekly	Never	Rarely
P12	Male	34	Worker	None	Daily	Dailly	Rarely	Never	Rarely
P13	Male	29	Worker	None	Daily	Daily	Rarely	Daily	Daily
P14	Male	28	Worker	None	Daily	Daily	Weekly	Rarely	Daily
P15	Male	43	Worker	None	Daily	Dailly	Weekly	Rarely	Weekly
P16	Female	36	Worker	None	Daily	Daily	Monthly	Weekly	Never
P17	Female	37	Worker	None	Daily	Daily	Rarely	Never	Daily
P18	Female	38	Worker	None	Daily	Dailly	Never	Weekly	Weekly
P19	Male	39	Worker	None	Daily	Daily	Monthly	Never	Never
P20	Male	40	Worker	None	Daily	Daily	Never	Never	Weekly
P21	Female	34	Older Student	None	Daily	Daily	Weekly	Never	Rarely
P22	Female	36	Older Student	None	Daily	Dailly	Rarely	Never	Daily
P23	Male	38	Older Student	None	Daily	Daily	Rarely	Daily	Rarely
P24	Male	35	Older Student	None	Daily	Daily	Weekly	Rarely	Daily
P25	Male	38	Older Student	None	Daily	Dailly	Weekly	Rarely	Weekly
P26	Male	29	Older Student	None	Daily	Daily	Monthly	Weekly	Never
P27	Female	40	Older Student	None	Daily	Daily	Rarely	Never	Daily
P28	Female	41	Older Student	None	Daily	Dailly	Never	Weekly	Weekly
P29	Female	39	Older Student	None	Daily	Daily	Monthly	Never	Never

A.2 Participants' Demography Sheet

P30	Male	26	Regular User	None	Daily	Daily	Never	Never	Weekly
P31	Male	43	Regular User	None	Daily	Daily	Weekly	Never	Daily
P32	Female	37	Regular User	None	Daily	Dailly	Rarely	Never	Rarely
P33	Male	38	Regular User	None	Daily	Daily	Rarely	Daily	Rarely
P34	Male	34	Regular User	None	Daily	Daily	Weekly	Rarely	Daily
P35	Male	31	Regular User	None	Daily	Dailly	Weekly	Rarely	Weekly
P36	Male	27	Regular User	None	Daily	Daily	Monthly	Weekly	Never
P37	Female	24	Regular User	None	Daily	Daily	Rarely	Never	Daily
P38	Female	33	Regular User	None	Daily	Dailly	Never	Weekly	Weekly
P39	Female	32	Regular User	None	Daily	Daily	Monthly	Never	Never
P40	Male	33	Regular User	None	Daily	Daily	Never	Never	Weekly
P41	Male	36	Regular User	None	Daily	Daily	Weekly	Never	Daily
P42	Female	34	Regular User	None	Daily	Dailly	Rarely	Never	Rarely
P43	Female	31	Regular User	None	Daily	Daily	Rarely	Daily	Daily
P44	Male	32	Regular User	None	Daily	Daily	Weekly	Rarely	Daily

A.3 Information sheet and consent form

The University of Manchester

INFORMATION SHEET AND CONSENT FORM

Ethical approval code CS77

JUNE 2013

Physiological Correlates to Online Behaviour

Introduction

Before making decision to take part in this study, it is most important for you to understand why the research is being carried out and what it would involve. Please take time to read the following information carefully and discuss it with friends, relatives or your personal GP if you wish. Ask us if there is anything that is not clear or if you would like more information. Take your time to decide whether or not you wish to take part.

Background

This is a study to measure the skin conductance response (SCR), skin temperature and eye movement of users as they interact with web stimuli. Understanding the affect states of users during interaction will help us gather information about their browsing behaviour and places on the website they found or experienced stress, relaxed or a neutral mood. We are also attempting to understand user perception and how we can use participants generated data to see how we can control stress level through modelling. Our results may eventually be published in a scientific journal, and may also be reported at scientific meetings.

Procedures

You have been chosen, because you either are a frequent user of the internet or a moderate user, aged between 18 to 40.

A SCR sensor will be placed on your wrist, this sensor also has the capability to monitor your Skin temperature. You will be interacting with webpages on an

1

A.4 Information sheet and consent form

Eye tracker while your SCR is been monitored. Your behaviour data will also be logged into the eye tracking during interaction and data from both sensors are collected simultanoeusly.

Participation in the study is entirely voluntary. It is up to you to decide whether or not to do this. If you do decide to take part, we would ask you to sign a consent form and a copy of this information sheet and the consent will be given to you before participating. If you decide to take part you are still free to withdraw from the study at any time. If you decide not to take part, or to withdraw, you do not have to give a reason for not doing so. The experiment will be stress free and comfortable.

Upon completion of the session we will inform you in more detail about the hypotheses we are testing, and you will have the opportunity to ask further questions. The time spent on the study less than (10 mins). Coffee and snacks is available for participants.

Confidentiality

All data will be coded so that your anonymity will be protected in any research papers and presentations that result from this work. If data is to be recorded that would identify the participant, for example photographs, audio or video, and if there is any intention to use this material in any publication or presentation, a separate release statement should be obtained after the recording has been made.

Finding out about result

If interested, you can find out the result of the study by contacting the researcher Fatima Isiaka or Simon Harper, after the end of the experiments (data 2nd July 2013). Your can contact the former in room LF1 Web Ergonomics Lab (WEL). Her phone number is 07721992160 and her email address is fatima.isiaka@manchester.ac.uk.

(the next page can be used to sign your consent and will be retained by researcher, with participant keeping above information sheet)

2

A.5 Information sheet and consent form

MANCHESTER 1824
The University of Manchester

CONSENT FORM

Study Name: Physiological Correlates to Web Contents

Description:

This study investigates the physiological correlates of users to web stimuli. A SCR sensor will placed on your wrist and given certain task. The task that you are asked to do is a based on Human Computer Interaction: you will be presented with six webpages on an eye tracker, each of this webpages are either static or dynamic. You will be asked to search for "University of Manchester" on the both the static and dynamic search engines. And also search for rail routes from "Manchester" to "London" on the rail way websites containing the static and dynamic textboxes. You will also look for "news" and "Entertainment" on dynamic websites with picture, video, text content and choose the one that is of interest to you. We will observe and save your physiological readings and correlates to the allocated task on the webpage.

Note: You are free to withdraw from the experiment at any time, without having to give a reason.

Eligibility Requirements:

To be eligible to take part in this study you must have no uncorrected visual abnormalities (e.g. colour blindness).

Duration: 10 minutes

Researcher: Fatima ISIAKA 07721992170

Principal Investigator Fatima ISIAKA

Approval Code: CS77

Your signature below indicates that you have understood the information about the CS77 experiment and consent to your participation.

Participant Name	Signature	Date

If you have any concerns related to your participation in this study, please direct them to the University of Manchester Senate Committee on the Ethics of Research on Human Beings, via Simon Harper (simon.harper@man.ac.uk).

3

Appendix B

Index

B.1 List of Acronyms and Attributes

- **Amplitude** This is the difference between the minimum and maximum SCR.
- **Baseline** This refers to the skin conductance level of the participant obtained as: $min_{idx}(SCR)$. min_{idx} is the minimum indices of the SCR.
- **FixationD** This the amount of time a participants fixates on a particular area on the webpage
- **Gaze time** This refers to the time of eye contact to object of interest
- **Latency** This refers to the time between stimulus and an onset in SCR.
- **Magnitude** This is the length of the SCR.
- **MappedFX** The is refers to the mapped fixation point on the vertical plain of the webpage
- **MappedFY** This is the mapped fixation point on the horizontal plain of the webpage.
- **MPeak** This refers to the mean peak responses detected.
- **MSP** This refers to the mean skin potential is also based on MST.

- **MST** This is the mean skin temperature of a participant.
- **PupilChange** The pupil change or pupil dilation is the variations in pupil size it could greater in sizes depend on the rate of emotional response.
- **Saccade size** This is the angular distance between fixations. Its length also varies (Eq 3.5).
- **SCR** Skin Conductance Response
- **Status** This refers to the users stress level.
- **URL** The links to webpages the participants interacted with.

B.2 Forward Search Algorithm

```
library(ggplot2)
library(car)
library(caret)
library(reshape2)

xx = read.csv("data.csv", rownames =1)
x = xx[,1:13]
# x = log(xx[,1:13])

mahal <- function(x, index) {
if (!is.matrix(x)) stop("x is not a matrix")
xbar <- apply(x[index,], 2, mean)
S <- var(x[index,])
S <- solve(S)
xcent <- t(t(x) - xbar)
apply(xcent, 1, function(x) x %*% S %*% x)
}

if (!is.matrix(x)) x <- as.matrix(x)
rn <- rownames(x)
if (is.null(rn)) rn <- 1:nrow(x)
n <- length(x[,1])
p <- length(x[1,])
s <- 1:n
ind <- matrix(0, n-p, n)
ind1 <- 0
thresh <- qchisq((n-0.5)/n,p)

index<-1:(p+1)
for(i in (p+1):n) {
ind1<-ind1+1
if(i==(p+1)) D<-mahal(x,index)
index<-order(D)
index1<-sort(index[1:i])
D<-mahal(x,index1)
index2<-s[D>thresh]
ind[ind1,index2]<-ind[ind1,index2]+1
}

y<-rep(1:(n-p),rep(n,(n-p)))
x<-rep(1:n,(n-p))

labs =paste(row.names(xx))
datf = data.frame(URL = labs, x =x,y = y, Status = xx[,14])
datf$URL <- factor(datf$URL, levels = datf$URL)

par(mai = par("mai") * c(1, 1, 2, 1))
residuals <- c(" ", "*")[as.vector(t(ind[(n-p):1,])) + 1]
p = ggplot(data = datf, aes(x = URL, y))
p1 = p + geom_text( data = datf, aes(URL,y, label = residuals, colour = residuals)) +
ylab("Number of observations used for estimation") #+ geom_point(aes(colour=Status))
p1 = p1 + theme(text = element_text(size=12), axis.text.x = element_text(angle=90, hjust=1),
legend.position = "none") #+ scale_y_reverse()
p1
```

B.3 Physiological correlates X to url $\mathbb{I}$: PHYCoB I algorithm

Data: User physiological data $X, \forall\ SCR, ST, PD \in X$; User eye movement data, $Y\ \forall\ x, y, FD \in Y$; web image/url $\mathbb{I}$

Result: secondary data matrix Z_{mp}, with attributes p, instances m and predictions S_i, N_i,R_i

i = 1;

N = 4 ▷ polynomial order

Z = [ ];

while $i < 44$ **do**

  for $j = 1 : 6$ **do**

    $\mathbb{I}_j \leftarrow X_i$

    $\mathbb{P}_i = [SCR\ \ ST\ \ PD]$

    $\mathbb{P}_i = ((\sum_{i}^{2N+1} \mathbb{P}_1)/2N+1, \ldots, (\sum_{2N+1}^{n(2N+1)} \mathbb{P}_n)/2N+1)$ ▷ filterdata

    determine $pk(X_i) = mean(FINDPEAK(SCR_i, 'threshold', 3))$ ▷ peaks of SCR

    $b(X_i) = mean(minima(SCR_1 \ldots SCR_n))$ ▷ baseline of SCR (SCL)

    compute $mg(X_i) = (\sum_1^n (SCR)^2)^{0.5}$ ▷ magnitude

    $a(X_i) = max(SCR) - min(SCR)$ ▷ mean amplitude

    $MST(X_i) = (\sum_1^n ST)/n$ ▷ mean skin temperature

    $PD(X_i) = (\sum_1^n PD)/n$ ▷ mean pupil size

    while $k = 1 : length(Y_i)$ **do**

      while *input matrix* Y **do**

        $\mathbb{I}_j \leftarrow Y_i$;

        $MPFX(Y_i) = \sum(x_1 \ldots x_k)/k$ ▷ mapped fixation x

        $MPFY(Y_i) = \sum(y_1 \ldots y_k)/k$ ▷ mapped fixation y

        $FD(Y_i) = \sum(FD_i \ldots FD_k)/k$ ▷ fixation duration

      end

    end

    if k = *length(Y);* **then**

      k+1 = 0;

    else

      X X=d(x,y) $= ((x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_k - y_k)^2)^{0.5}$ ▷ fixation points

      compute $saccade(Y_i) = (\sum_{k=1}^{n} (XX)^2)$ ▷ saccade size (euclidean distance)

    end

    ▷ compute stress levels S_i, N_i, R_i

    threshold $= 0.5\{(a(X_i) - min(X_i))\}$

    if $pk(X_i) > threshold\ and\ Time(t_i) \geq 3$ **then**

      stress level = S_i ▷ Stress

      if $pk(X_i) \leq threshold\ and\ pk(X_i) \geq b(X_i)$ **then**

        stress level = N_i ▷ Neutral

      else

        stress level =R_i ▷ Relaxed

        update output Z_{mp};

      end

    else

      return i **return** j

    end

  end

end

while *output matrix* Z_{mp} **do**
compute [R, P] = cov(Z_{mp}) ▷ compute p-values from correlation coefficient of ouput matix
set $X_{IMP} = P < 0.05$ ▷ get matrix with parameters greater than critical value (parameters with most impact)
set $\mathbb{Y}$ = stress levels
$X_{IMP} = iddata(\mathbb{Y}, X_{IMP}, t)$ ▷ create a identity data object with response $\mathbb{Y}$;
, input X_{IMP} ;
and sample time t, using the function iddata in matlab
$X_{IMP} = misdata(X_{IMP})$;
▷ estimating missing data using the function misdata
end
set trainingData = [X_{IMP} $\mathbb{Y}$]
while *trainingData* **do**
▷ create a function that returns an object of class PHYCOBClass
PHYCOB = **function**(a, b, c, d, k, trainingData){
model = get(listparam(a, b, c, d, k, x = trainingData(:, −1), y = trainingData.$\mathbb{Y}$), class = "PHYCOBClass")
return model
}
▷ create a method for function print for class PHYCOBClass
predict.PHYCOBClass = **function** (modelObject){
return (idss(length(modelObject.y)))
} ▷ create a state space model from model object
mdl = $PHYCOB(X_{IMP}, 1 : 10)$ ▷ estimate an identity space model using the function PHYCOB with polynomial orders between 1 to10
predictions = predict(mdl, X_{IMP}, K) ▷ predict the a response using the derived model with horizon K
return predictions
end

References

Allanson, J. & Fairclough, S.H. (2004). A research agenda for physiological computing. *Interacting with computers*, **16**, 857–878. 18, 24

Andreassi, J.L. (2000a). *Psychophysiology: Human behavior & physiological response*. Psychology Press. 12

Andreassi, J.L. (2000b). *Psychophysiology: Human behavior and physiological response*. Psychology Press. 16, 20, 23, 115

Arapakis, I., Konstas, I. & Jose, J. (2009). Using facial expressions and peripheral physiological signals as implicit indicators of topical relevance. In *Proceedings of the 17th ACM international conference on Multimedia*, 461–470, ACM. 20

Atkinson, A. (1994). Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association*, **89**, 1329–1339. 48

Baayen, R.H., Davidson, D.J. & Bates, D.M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, **59**, 390–412. 51

Bach, D.R., Friston, K.J. & Dolan, R.J. (2010). Analytic measures for quantification of arousal from spontaneous skin conductance fluctuations. *International Journal of Psychophysiology*, **76**, 52–55. 51

Bailenson, J.N., Pontikakis, E.D., Mauss, I.B., Gross, J.J., Jabon, M.E., Hutcherson, C.A.C., Nass, C. & John, O. (2008). Real-time classification of evoked emotions using facial feature tracking and physiological responses. *International journal of human-computer studies*, **66**, 303–317. 21

Bannon, L. (1991). From human factors to human actors: The role of psychology and human-computer interaction studies in system design. *Design at work: Cooperative design of computer systems*, 25–44. 23

Baumgartner, T., Esslen, M. & Jäncke, L. (2006). From emotion perception to emotion experience: Emotions evoked by pictures and classical music. *International Journal of Psychophysiology*, **60**, 34–43. 44

Bayer, H.M. & Glimcher, P.W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, **47**, 129–141. 104

Bergstrom, J.R. & Schall, A. (2014). *Eye tracking in user experience design*. Elsevier. 10, 11, 12, 15, 16, 38

Brandt, M.L., Brown, K.E., Dykes, P.J., Lindberg, E.D., Olson, D.E., Selden, J.E., Snyder, D.D. & Walts, J.O. (1999). Computer apparatus and method for providing a common user interface for software applications accessed via the world-wide web. US Patent 5,892,905. 22

Cain, B. (2007). A review of the mental workload literature. Tech. rep., DTIC Document. 17

Calvo, R. & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *Affective Computing, IEEE Transactions on*, **1**, 18–37. 21

Calvo, R.A., Brown, I. & Scheding, S. (2009). Effect of experimental factors on the recognition of affective mental states through physiological measures. In *AI 2009: Advances in Artificial Intelligence*, 62–70, Springer. 21

Castañeda, J.A., Muñoz-Leiva, F. & Luque, T. (2007). Web acceptance model (wam): Moderating effects of user experience. *Information & Management*, **44**, 384–396. 4

Chatterjee, S. & Hadi, A.S. (2015). *Regression analysis by example*. John Wiley & Sons. 52, 103

Chen, H., Wigand, R.T. & Nilan, M. (2000). Exploring web users' optimal flow experiences. *Information Technology & People*, **13**, 263–281. 19

Compeau, D.R. & Higgins, C.A. (1995). Computer self-efficacy: Development of a measure and initial test. *MIS quarterly*, 189–211. 24

Cooley, R., Mobasher, B. & Srivastava, J. (1999). Data preparation for mining world wide web browsing patterns. *Knowledge and information systems*, **1**, 5–32. 22

Davis, E.V. (1990). *Software testing for evolutionary iterative rapid prototyping*. Ph.D. thesis, Monterey, California: Naval Postgraduate School. 19

Demiral, Ş.B., Schlesewsky, M. & Bornkessel-Schlesewsky, I. (2008). On the universality of language comprehension strategies: Evidence from turkish. *Cognition*, **106**, 484–500. 10

Diebold, F.X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of diebold–mariano tests. *Journal of Business & Economic Statistics*, **33**, 1–1. 72

Dirican, A.C. & Göktürk, M. (2011a). Psychophysiological measures of human cognitive states applied in human computer interaction. *Procedia Computer Science*, **3**, 1361–1367. 17, 18, 24

Dirican, A.C. & Göktürk, M. (2011b). Psychophysiological measures of human cognitive states applied in human computer interaction. *Procedia Computer Science*, **3**, 1361–1367. 115

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, **6**, 169–200. 4

Fabes, R.A., Eisenberg, N. & Eisenbud, L. (1993). Behavioral and physiological correlates of children's reactions to others in distress. *Developmental Psychology*, **29**, 655. 23

Faraway, J.J. (2014). *Linear models with R*. CRC Press. 103

Farmer, E. & Brownson, A. (2003). Review of workload measurement, analysis and interpretation methods. *European Organisation for the Safety of Air Navigation*, **33**. 17

Filipovic, S.R. & Andreassi, J.L. (2001). Psychophysiology: Human behavior and physiological response. *Journal of Psychophysiology*, **15**, 210–212. 2, 10, 38

Fink, J., Kobsa, A. & Nill, A. (1997). Adaptable and adaptive information access for all users, including the disabled and the elderly. In *User Modeling*, 171–173, Springer. 22

Fischer, G. (2001). User modeling in human–computer interaction. *User modeling and user-adapted interaction*, **11**, 65–86. 24

Fogg, B. & Tseng, H. (1999). The elements of computer credibility. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 80–87, ACM. 23

Fraley, C. & Raftery, A.E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, **97**, 611–631. 98

Friston, K.J. (2005). Models of brain function in neuroimaging. *Annu. Rev. Psychol.*, **56**, 57–87. 51

Ganglbauer, E., Schrammel, J., Deutsch, S. & Tscheligi, M. (2009). Applying psychophysiological methods for measuring user experience: possibilities, challenges and feasibility. In *Workshop on user experience evaluation methods in product development*, Citeseer. 17, 18, 19

Granka, L.A., Joachims, T. & Gay, G. (2004). Eye-tracking analysis of user behavior in www search. 15

Green, J. (1976). An introduction to human physiology. 10

Gross, J.J., Carstensen, L.L., Pasupathi, M., Tsai, J., Götestam Skorpen, C. & Hsu, A.Y. (1997). Emotion and aging: experience, expression, and control. *Psychology and aging*, **12**, 590. 11

Haag, A., Goronzy, S., Schaich, P. & Williams, J. (2004). Emotion recognition using bio-sensors: First steps towards an automatic system. In *Affective Dialogue Systems*, 36–48, Springer. 21

Hackett, S., Parmanto, B. & Zeng, X. (2005). A retrospective look at website accessibility over time. *Behaviour & Information Technology*, **24**, 407–417. 22

Harbrecht, H., Peters, M. & Schneider, R. (2012). On the low-rank approximation by the pivoted cholesky decomposition. *Applied numerical mathematics*, **62**, 428–440. 41

Hazlett, R.L. (2006). Measuring emotional valence during interactive experiences: boys at video game play. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, 1023–1026, ACM. 23

Healey, J.A. & Picard, R.W. (2005). Detecting stress during real-world driving tasks using physiological sensors. *Intelligent Transportation Systems, IEEE Transactions on*, **6**, 156–166. 5

Hector, A. *et al.* (2015). *The New Statistics with R: An Introduction for Biologists*. Oxford University Press. 51

Heraz, A. & Frasson, C. (2007). Predicting the three major dimensions of the learner's emotions from brainwaves. *International Journal of Computer Science*, **31**. 21, 115

Hess, E.H. (1965). Attitude and pupil size. *Scientific american*. 23

Hudlicka, E. (2003). To feel or not to feel: The role of affect in human–computer interaction. *International Journal of Human-Computer Studies*, **59**, 1–32. 3

Insko, B.E. (2003). Measuring presence: Subjective, behavioral and physiological methods. *Emerging Communication*, **5**, 109–120. 19

Iqbal, S.T., Zheng, X.S. & Bailey, B.P. (2004). Task-evoked pupillary response to mental workload in human-computer interaction. In *CHI'04 extended abstracts on Human factors in computing systems*, 1477–1480, ACM. 14

J.A., H. & Picard, R. ((2005)). Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*, **6**, 156166. 21

Jacob, R.J. & Karn, K.S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. *Mind*, **2**, 4. 22

Kalyuga, S. (2008). *Managing cognitive load in adaptive multimedia learning*. IGI Global. 17

Kamon, E., Pandolf, K. & Cafarelli, E. (1974). The relationship between perceptual information and physiological responses to exercise in the heat. *Journal of human ergology*, **3**, 45–54. 13

Kaski, S. (1997). Data exploration using self-organizing maps. In *Acta Polytechnica Scandinavica: Mathematics, Computing And Management In Engineering Series No. 82*, Citeseer. 48

Khan, J.A., Van Aelst, S. & Zamar, R.H. (2007). Building a robust linear model with forward selection and stepwise procedures. *Computational Statistics & Data Analysis*, **52**, 239–248. 51

Kim, K.H., Bang, S.W. & Kim, S.R. (2004). Emotion recognition system using short-term monitoring of physiological signals. *Medical and biological engineering and computing*, **42**, 419–427. 3, 21

Kivikangas, J.M., Chanel, G., Cowley, B., Ekman, I., Salminen, M., Järvelä, S. & Ravaja, N. (2011). A review of the use of psychophysiological methods in game research. *Journal of Gaming & Virtual Worlds*, **3**, 181–199. 18

Kolakowska, A., Landowska, A., Szwoch, M., Szwoch, W. & Wróbel, M.R. (2013). Emotion recognition and its application in software engineering. In *Human System Interaction (HSI), 2013 The 6th International Conference on*, 532–539, IEEE. 19

Kramer, A.F. (1990). Physiological metrics of mental workload: A review of recent progress. Tech. rep., DTIC Document. 17, 18, 19

Kretzschmar, F., Pleimling, D., Hosemann, J., Füssel, S., Bornkessel-Schlesewsky, I. & Schlesewsky, M. (2013). Subjective impressions do not mirror online reading effort: Concurrent eeg-eyetracking evidence from the reading of books and digital media. *PloS one*, **8**, e56178. 10

Kurzban, R. & Houser, D. (2005). Experiments investigating cooperative types in humans: A complement to evolutionary theory and simulations. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 1803–1807. 104

Lang, P.J., Greenwald, M.K., Bradley, M.M. & Hamm, A.O. (1993). Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology*, **30**, 261–273. 20

Lang, P.J., Davis, M. & Öhman, A. (2000). Fear and anxiety: animal models and human cognitive psychophysiology. *Journal of affective disorders*, **61**, 137–159. 23

Lanzetta, J.T. & Orr, S.P. (1986). Excitatory strength of expressive faces: Effects of happy and fear expressions and context on the extinction of a conditioned fear response. *Journal of Personality and Social Psychology*, **50**, 190. 21

Lazarus, R.S. (1993). From psychological stress to the emotions: A history of changing outlooks. *Annual review of psychology*, **44**, 1–22. 23

Liversedge, S.P. & Blythe, H.I. (2007). Lexical and sublexical influences on eye movements during reading. *Language and Linguistics Compass*, **1**, 17–31. 10

Macaulay, C., Jacucci, G., ONeill, S., Kankaineen, T. & Simpson, M. (2006). The emerging roles of performance within hci and interaction design. *Interacting with computers*, **18**, 942–955. 24

Mahalanobis, P.C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, **2**, 49–55. 49

Mandryk, R., Inkpen, K. & Calvert, T. (2006a). Using psychophysiological techniques to measure user experience with entertainment technologies. *Behaviour & Information Technology*, **25**, 141–158. 20

Mandryk, R.L. & Atkins, M.S. (2007). A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. *International Journal of Human-Computer Studies*, **65**, 329–347. 20

Mandryk, R.L., Inkpen, K.M. & Calvert, T.W. (2006b). Using psychophysiological techniques to measure user experience with entertainment technologies. *Behaviour & Information Technology*, **25**, 141–158. 19

Marullo, F.R. & Randall Jr, D.H. (2000). Automated client-based web server stress tool simulating simultaneous multiple user server accesses. US Patent 6,157,940. 2

Mavratzakis, A., Herbert, C. & Walla, P. (2016). Emotional facial expressions evoke faster orienting responses, but weaker emotional responses at neural and behavioural levels compared to scenes: A simultaneous eeg and facial emg study. *NeuroImage*, **124**, 931–946. 44

Mbipom, G. (2008). Examining the relationship between visual aesthetics and web accessibility: A formative study. *Human Centred Web Lab, University of Manchester*. 16

Michailidou, E., Harper, S. & Bechhofer, S. (2008). Visual complexity and aesthetic perception of web pages. In *Proceedings of the 26th annual ACM international conference on Design of communication*, 215–224, ACM. 22

Mindfield (2014). esence temperature, mindfield biofeedback sytems. *eSence Skin Temperature Handbook*, **1**, 1–12. 13, 14

Mwitondi, K. & Said, R. (2011). A step-wise method for labelling continuous data with a focus on striking a balance between predictive accuracy and model reliability. In *international Conference on the Challenges in Statistics and Operations Research (CSOR)*. 48

Nakagawa, S. & Schielzeth, H. (2013). A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, **4**, 133–142. 103

Newell, A. & Card, S.K. (1985). The prospects for psychological science in human-computer interaction. *Human-computer interaction*, **1**, 209–242. 23

Nielsen, J. (1994). Usability inspection methods. In *Conference companion on Human factors in computing systems*, 413–414, ACM. 1, 5

Nielsen, J. & Molich, R. (1990). Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 249–256, ACM. 5, 19

Ortony, A. & Turner, T.J. (1990). What's basic about basic emotions? *Psychological review*, **97**, 315. 4

Park, B. (2009). Psychophysiology as a tool for hci research: promises and pitfalls. In *Human-Computer Interaction. New Trends*, 141–148, Springer. 18

Partarakis, N., Doulgeraki, C., Leonidis, A., Antona, M. & Stephanidis, C. (2009). User interface adaptation of web-based services on the semantic web. In *Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments*, 711–719, Springer. 20

Paulson, L.D. (2005). Building rich web applications with ajax. *Computer*, **38**, 14–17. 22

Pepe, M.S., Feng, Z., Huang, Y., Longton, G., Prentice, R., Thompson, I.M. & Zheng, Y. (2008). Integrating the predictiveness of a marker with its performance as a classifier. *American journal of epidemiology*, **167**, 362–368. 103

Picard, R.W. (2000). *Affective computing*. MIT press. 3

Picard, R.W., Vyzas, E. & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **23**, 1175–1191. 20

Poole, A. & Ball, L.J. (2006). Eye tracking in hci and usability research. *Encyclopedia of human computer interaction*, **1**, 211–219. 22

Posada, D. & Crandall, K.A. (2001). Selecting the best-fit model of nucleotide substitution. *Systematic Biology*, **50**, 580–601. 50

Riva, G., Davide, F. & IJsselsteijn, W. (2003). Measuring presence: Subjective, behavioral and physiological methods. *Ios Press, Amsterdam, The Netherlands*, 110–118. 18

Rocke, D.M. & Woodruff, D.L. (2000). A synthesis of outlier detection and cluster identification. *Submitted manuscript, University of California, Davis*. 98

Rodden, T., Cheverst, K., Davies, K. & Dix, A. (1998). Exploiting context in hci design for mobile systems. In *Workshop on human computer interaction with mobile devices*, 21–22, Citeseer. 24

Russell, M. (2005). Using eye-tracking data to understand first impressions of a website. *Usability News*, **7**, 1–14. 15

Saint-Aimé, S., Le Pévédic, B. & Duhaut, D. (2009). igrace–emotional computational model for emi companion robot. *Advances in Human-Robot Interaction*, 26. 4

Sappenfield, J.W., Hong, C.M. & Galvagno, S.M. (2013). Perioperative temperature measurement and management: moving beyond the surgical care improvement project. *journal of Anesthesiology and Clinical Research*, **2**, 8. 18

Sauer, J. & Sonderegger, A. (2009). The influence of prototype fidelity and aesthetics of design in usability tests: effects on user behaviour, subjective evaluation and emotion. *Applied ergonomics*, **40**, 670–677. 19

Savitzky, A. & Golay, M.J. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, **36**, 1627–1639. 43

Scheres, S.H. & Chen, S. (2012). Prevention of overfitting in cryo-em structure determination. *Nature methods*, **9**, 853–854. 72

Schneider-Hufschmidt, M., Malinowski, U. & Kuhme, T. (1993). *Adaptive user interfaces: Principles and practice*. Elsevier Science Inc. 22

Scott, D.W. (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons. 50, 53

Sinha, R. (1996). Multivariate response patterning of fear and anger. *Cognition & Emotion*, **10**, 173–198. 21

Skadberg, Y.X. & Kimmel, J.R. (2004). Visitors flow experience while browsing a web site: its measurement, contributing factors and consequences. *Computers in human behavior*, **20**, 403–422. 19

Smith, M.J., Conway, F.T. & Karsh, B.T. (1999). Occupational stress in human computer interaction. *Industrial health*, **37**, 157–173. 2

Stern, R.M., Ray, W.J. & Quigley, K.S. (2001). *Psychophysiological recording*. Oxford University Press. 2

Subrahmanyam, K. & Greenfield, P.M. (1994). Effect of video game practice on spatial skills in girls and boys. *Journal of applied developmental psychology*, **15**, 13–32. 24

Uğur, S. (2013). *Wearing embodied emotions: A practice based design research on wearable technology*. Springer. 4

van Havre, Z., White, N., Rousseau, J. & Mengersen, K. (2015). Overfitting bayesian mixture models with an unknown number of components. *arXiv preprint arXiv:1502.05427*. 51

Vasalou, A., Ng, B., Wiemer-Hastings, P. & Oshlyansky, L. (2004). Human-moderated remote user testing: Protocols and applications. In *8th ERCIM Workshop, User Interfaces for All, Wien, Austria*. 19

Vermeeren, A.P., Law, E.L.C., Roto, V., Obrist, M., Hoonhout, J. & Väänänen-Vainio-Mattila, K. (2010). User experience evaluation methods: current state and development needs. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, 521–530, ACM. 19

Verschuere, B., Ben-Shakhar, G. & Meijer, E. (2011). *Memory detection: Theory and application of the Concealed Information Test*. Cambridge University Press. 22

Vigo, M. & Harper, S. (2011). Human adaptation to the web: A proposal for a research programme. *School of Computer Science, Information Management Group, University of Manchester*, 1–34. 22

Von Ahn, L. (2006). Games with a purpose. *Computer*, **39**, 92–94. 21

Ward, R.D. & Marsden, P.H. (2003). Physiological responses to different web page designs. *International Journal of Human-Computer Studies*, **59**, 199–212. 16

Widyantoro, D.H., Ioerger, T.R. & Yen, J. (1999). An adaptive algorithm for learning changes in user interests. In *Proceedings of the eighth international conference on Information and knowledge management*, 405–412, ACM. 22

Yu, H., Chung, C., Wong, K., Lee, H. & Zhang, J. (2009). Probabilistic load flow evaluation with hybrid latin hypercube sampling and cholesky decomposition. *IEEE Transactions on Power Systems*, **24**, 661–667. 41

Zander, T.O., Gaertner, M., Kothe, C. & Vilimek, R. (2010). Combining eye gaze input with a brain–computer interface for touchless human–computer interaction. *Intl. Journal of Human–Computer Interaction*, **27**, 38–51. 23

Zhai, J. & Barreto, A. (2006). Stress detection in computer users based on digital signal processing of noninvasive physiological variables. In *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*, 1355–1358, IEEE. 2, 3, 5

Zuur, A.F., Ieno, E.N. & Elphick, C.S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, **1**, 3–14. 48