University of Reading

Determining Event Times From Complex Control Room Data Using Physiological Indicators

Engineering Doctorate Thesis

Joshua Fraser Eadie

April 16, 2021

TECHNOLOGIES FOR THE SUSTAINABLE BUILT ENVIRONMENT CENTRE
& THALES UK RESEARCH AND TECHNOLOGY

Declaration

I confirm that this research contribution is based on my own work, and that all externally sourced material has been properly and fully acknowledged.

I grant powers of discretion to the University of Reading to allow this thesis to be copied in whole or in part without further reference to me. This permission covers only single copies made for study purposes, subject to normal conditions of acknowledgement.

Joshua Fraser Eadie

Acknowledgements

No thesis is truly a solo effort and I’d like to thank all the people that helped get this work, and me, over the line throughout the years of this EngD.

I’d like to thank my project supervisors Rachel Craddock, Slawomir Nasuto and Victor Becerra for their patience, dedication and incredibly wide-ranging expertise. Their individual strategies and perspectives were crucial in making me a better researcher than I was in 2014! I’d also like to thank Thales Research and Technology for sponsoring the project.

The Technologies for The Sustainable Built Environments (TSBE) centre was a truly life-changing place for me, from the people that ran it, to the other researchers to the events and support. I’d like to thank Jenny Berger for her tireless efforts in shepherding the EngD cohort through thick and thin, particularly at the end when I found out I would be the last student in the centre. She boosted, defended and pulled us all up when we had our regular wobbles. I’ll never forget the action-stations face she pulled when I barged into her office demanding an overwhelming expensive piece of software for this work; but she knew it was needed and knew what had to be done. There are doubtless hundreds of tales like this through her tenure in her post and I hope she’s aware of the dramatically positive impact she has had on the world as a result. I’d like to thank Dr Daniel Williams who has gone through the journey from mentor to friend to business partner to best man over the course of this project, without him and Dr Yu-Chun Pan’s rock-steady, no-holds-barred feedback and support, I would never have made it this far and most certainly wouldn’t be who I am today. The entire cohort at the TSBE centre made research fun and I will never forget being involved with so many passionate people with such a diverse range of expertise that made for some of the most intellectually stimulating years in the research room, out at events and of course, down the pub. I’d like to specifically thank Dr’s Nicholas Hollely, Steven Lowe, Saadia Ansari, Thomas Chung, Alan Halford, Katie Dawkins and Mitch Curtis. Whether we were working problems through on the

whiteboard or discussing the increasingly acute points of detail of some esoteric reference, whether it be research or classic British sitcoms, they all made this work possible and enjoyable. I'd also like to thank Dr Christos Halios, though not part of the centre, became my most valuable resource in the last days at the University when all other TSBEers had graduated – you really made my research difficulties feel infinitely more possible when you were around.

I'd like to thank my parents, who never falter in their belief in me and never hesitate to support me in any direction I go, who have been instrumental in my career so far – I continue to test their resolve in me, they are yet to let me down in any regard. My family, large, unwieldy and who would never miss the slightest chance to mock and poke me – are my best friends, they create the most useful tool in my arsenal by being my safety net, whilst laughing at me the whole way of course, and I wouldn't have it any other way.

My partner, now wife, Kimberley Morton-Eadie, or Tunk, has never faltered in her promise to support me no matter what. My ever-present confidante has sacrificed a great deal to get me through this project. She knows the trials and insecurities I have gone through on this journey and without her being there everyday to explain to me yet again that the world will not in fact end if I fail, I would not have made it. Her patient and welcoming smile is always there at the end of all things and I am not sure of what I would ever possibly accomplish without her.

Finally, I would like to dedicate this Thesis to my grandfather Frank Ward – the *proper* Doctor of the family. He took a keen interest in the human elements of this work and always had time on a Wednesday to dive into exceptional detail of my progress over fish and red wine, all whilst offering insight and being jealous at "how bloody small ECG machines are now". He is sorely missed by everyone who knew him, and I regret that he will not get to read this work, I would love nothing more than to hear his thoughts on it.

Abstract

Control room environments are present throughout many areas of commercial and industrial infrastructure, from air traffic control and harbour masters to chemical plant control and military operations. These control rooms work using a human-in-the-loop system in which a human operator monitors the data from sensors in the environment they are responsible for and make decisions and take action to maintain efficiency and safety. Though humans possess a natural aptitude for spotting patterns and anomalies in complex data, the majority of safety and process control errors are still human errors. These errors are most often as a result of 'cognitive overload' – a state in which the operator is presented with more information than they can effectively process cognitively in real time.

Machine learning is often employed to automate some of the tasks of the human operator to reduce their cognitive workload to reduce errors. The performance of machine learning systems relies on obtaining large volumes of labelled data; which is an expensive and time consuming process that is currently performed by experts manually reviewing complex data and providing labels.

The work presented here focusses on automating the labelling process by assessing the cognitive state of the operator using objective measures and using these measures to detect when events occurred in complex data in order to provide labels. The research assesses pupil diameter, echocardiogram (ECG) and novel mouse movement metrics to determine their suitability as classifiers that can automatically detect when events occurred. Results demonstrate that pupil diameter performs better than ECG as a physiological classifier. When event types are analysed separately, results demonstrate that all measures developed correctly recall between 75% - 95% of longer event types whereas performance for shorter events types correctly recalls between 14% for mouse measures, 31% ECG measures and 78% for pupil measures.

Contents

Table of Figures

Table of Tables

Common Abbreviations

PD	Pupil diameter
TEPR	Task-evoked pupil response
ANN	Artificial neural network
NASA-TLX	National Aeronautics and Space Administration-task load index
FMRI	Functional magnetic resonance imaging
GSR	Galvanic skin response
EEG	Electroencephalogram
ECG	Electrocardiogram
ISA	Instantaneous Self-Assessment Workload Scale
RR	R peak of the QRS complex
fNIRS	Functional near-infrared spectroscopy
PPG	Photoplethysmogram.
AVNN	Mean of all normal sinus to normal sinus inter-beat intervals (NN)
RMSSD	Root mean square of successive differences between normal heartbeats
SDNN	Standard deviation of the NN (R-R) intervals
pNNx	Proportion of NNx (the number of pairs of successive NNs that differ by more than x ms) divided by total number of NNs.e.g. NN50= the number of pairs of successive NNs that differ by more than 50msl; NN20= the number of pairs of successive NNs that differ by more than 20 ms

Mean RR	Mean average of RR intervals
Mean HR	Mean average heart rate
HRV	Heart rate variance
HRVTRI	Heart rate variance triangular index
LF	Low-frequency component of HRV
MF	Mid-frequency component of HRV
HF	High-frequency component of HRV
LF/HF Ratio	Ratio of low-frequency to high-frequency HRV components
IBI	Inter-beat interval
EMG	Electromyography
ECG MAD	Electroencephalogram median absolute deviation
TCD	Transcranial Doppler sonography
NASA-TLX	National aeronautics and space administration-task load index
MCH	Modified cooper-harper scale
EDA	Electro-dermal activity

Chapter 1: Introduction

1.1. The Engineering Doctorate

The Engineering Doctorate (EngD) is a PhD-level research degree with a particular focus on industrial application. Designed for students with a desire to persue a career in industry, the EngD draws project motivation from actual industrial need; often directly from the sponsoring company of the project itself. With the contributions to knowledge also being a direct contribution to the industrial requirement set out by the company.

Structured into a four-year course, the EngD combines research with 2 years of taught modules in both technical and business subjects whilst working closely with a sponsoring company. The program focusses on developing research engineers with a strong technical and applied industrial skillset, whilst also completing a project that offers an original contribution to knowledge in an academic subject area.

The sponsoring company for this EngD was Thales Research and Technology UK and it was undertaken at The Technologies for Sustainable Built Environments Centre at The University of Reading.

1.2. Human Operators and Control Rooms

Industrial processes often have elements that are either controlled, or performed directly, by a human. The reason for inclusion of a human in these processes, or *human-in-the-loop* systems, is usually to add a level of human pattern recognition and understanding to what might otherwise be an automated system based on preprogrammed and inflexible rules. Humans have a natural aptitude for spotting anomalies and deriving meaning and context from quite complex data; making humans invaluable assets in the field of safety and control processes such as air traffic control and port management. These control rooms are found in many areas of industry but all have the same basic structure; sensors in the environment transmit data back to an interface that is viewed by a human operator. This human operator then makes key decisions based on this data and relays instructions or commands back to the environment.

A major concern in these industries is human error, given the pivotal role that operators play in the safety and efficiency of any control-based process. Many strategies have been implemented in an effort to reduce human error in these fields. Given this, studies have shown human error to be responsible for significant numbers of accidents. In aviation, one such study showed that air traffic control skill-based errors (errors due to memory lapses or attention failures) were responsible for, or involved in, 82% of accidents (Pape, Wiegmann and Shappell, 2001). Another control-based study showed that human error was responsible for 70% of accidents in chemical process industries (Lees, 2016).

A key risk that faces the human operator is that of *data overload*; a situation in which they are presented with more information than they can effectively process, cognitively, in real time. When experiencing data overload an operator is prone to mistakes that can lead to safety and efficiency decreases.

In order to alleviate data overload, various levels of automated processing are applied to the incoming data to offload more rudimentary tasks from the human, allowing them to focus on more important, decision influencing data. The way in which the automation processes data must be intuitive, not impose risk of missed key data and must not burden the operator with further mental workload. In order to create such automations, large volumes of data history are taken from the control rooms. Algorithms are then applied in an attempt to autonomously detect anomalies, create alarms where danger is apparent or impending, or generally search for important information in the incoming data stream. Difficulties arise when the data scientists behind these algorithms have poor understanding of the situational requirement of the operators. To address this, operators are often made part of the process of designing the automations – retrieving relevant data from operators for this is known as “expert data extraction”. This process can be difficult, often due to the intricate subtleties that operators are able to see and understand naturally in the data. Operators often also struggle to articulate their decision-making process. Accurately and effectively describing these events in a manner that can be automated is a complex task.

This labelling process is extremely expensive, time-consuming and often inaccurate, given that different operators have differring opinions on what consitutes an event. This issue of data labelling for control room scenarios provides the key focus for this research.

1.3. Monitoring Operators

Events that occur within control rooms that cause cognitive load to rise are of particular interest, as automation applied to these events will directly reduce the cognitive load on the operator.

In the past few decades, it has been shown that the cognitive load of a human can be objectively measured using sensors recording their physiological output. Certain cognitive processes are linked to the autonomic nervous system. These processes affect some biological signals that can be measured, thus enabling the inference of the cognitive state of the human based on these signals.

Some work has been done to estimate the cognitive state of operators during their work using such sensors that measure cognitive load objectively. These indices are part of systems that intend to form a closed-loop control room system interface in which measures are taken to reduce the cognitive load of the operator in real-time.

Thales Research and Technology has many industrial elements in the airspace and defense sectors that rely of control room operators. The Patterns of Life Team at Thales use machine learning and other big data analytics tools to create meaning from large datasets. These tools are subject to the difficulties of the data labelling problem. Thales are therefore investigating through this project, whether an automatic data labelling system could be made for labelling these complex, safety critical control room environments. Filtering through large data sets that can span several weeks at times is a very time consuming process and poses a significant challenge even to the operator that was there when the data was recorded. A system of significant benefit to Thales and all big data research teams would be a method to determine the times at which events of interest occurred in complex data sets.

1.4. Research Objectives and Thesis Structure

For this work, we have an overall research objective: *To determine if the times at which events of interest occurred within complex control room scenarios can be retrieved without manual intervention from the operator.*

This thesis manuscript will design a piece of research to meet this objective and present the results herein. The structure of this thesis is as follows: In this chapter, we will perform a comprehensive literature review to determine the state of the art in terms of cognitive load measurement and control room applications. We will assess the differing methods used to obtain and process metrics of cognitive load and develop a broad understanding of the conclusions made from these findings. We will then identify the gaps in literature that we will form research questions to address. In chapter 2, i.e. literatrue review, we develop our research design. We will use the knowledge obtained from the literature review to create a hypothesis and experiment to validate the hypthothesis. We will design an experiment that will be used to gather data to be analysed. We will also outline which methods we will use and which signals we will measure to obtain our data sets. We also design a series of trials to both train and validate our analysis. In chapter 4, i.e. analysis, the techniques used for analysing the data gathered from the experiments will be described in detail. Chapter 4 will also discuss the methods used to properly assess the performance of the techniques when applied to the gathered data. In chapter 5, i.e., Results and Discussion, we present the results of the analysis performed on the data gathered from the experiments. The results will then be discussed in the context of our research aim and subsequent research questions outlined in the methodology. Finally, in chapter 6, the conclusion, we will present the conclusions of the work. The limitations of the methods used and scope for future work will also be discussed in this chapter.

Chapter 2: Literature Review

2.1. Introduction

In this chapter we will examine literature to assess the present state of the art and research pertaining to our project scope. We will first examine examples from industry-led projects that are focused on the data labelling problem in context of control rooms. We will then assess the state of the art for assessing cognitive load, addressing multiple methods and discussion in the framework of an extensive review of studies for eye-related measures, followed by measures derived from heart-related psychophysiological signals. The chapter will conclude with a discussion of the research examined and a framework for the research to be undertaken in this project. Further chapters will reference specific conclusions and research discussed in this section.

2.2. Online Data Labelling in Control Rooms

The patterns of life team at Thales Research seeded this project from a piece of research called SeeCoast. Developed to extend the US Coast Guard Port Security system, SeeCoast is a learning program that can accept input from the operators whilst they work to develop a comprehensive model of ‘normal behaviour’. The system uses a heavily modified version of the *Fuzzy ARTMAP* neural network classifier – SeeCoast can learn this model of normalcy either unsupervised or supervised in what they call a hybrid approach. The modifications made to the Fuzzy ARTMAP were to establish it as an anomaly detection method as supposed to a pattern recognition method. The challenge being that anomaly detection, by definition, has very few labelled examples for training. To overcome this, the system provided an interface to the operators that allowed them to confirm that particular anomalies discovered by the system were indeed anomalous from their experience. The labels then provided to the data were also not definite at the point of confirmation, to allow for operator mistakes or flexibility of the definition of anomalous. This flexibility allows the model of normalcy to develop constantly develop over time.

The developers of SeeCoast explain their implementation of the system and the role of the operators in the application context. The approach continues to tune the system to the operator rather than the operator to the system. Difficulties that they acknowledge are those of false alarms due to misleading data e.g. radars confusing the bow and stern of a large ship and two separate ships.

SeeCoast is a specifically tailored system, to a specific application, with a specific goal (anomaly detection). Given this, however, SeeCoast does address the issue of effectively applying the operator's knowledge to the automation and by doing so in an online fashion, vastly reducing the development time and cost of development. By engaging the operators during their tasks, there is no requirement for them to recall and explain their skills in an accurate fashion at a later time after the event has occurred. There is a clear advantage to extracting this expert information whilst the operator is performing their normal tasks. This "online" labelling approach does have the drawback of adding yet another responsibility to the operators workload, in high-stress environments, this may lead to increased cognitive load.

In the context of this work, the SeeCoast system intends to support the development of automated processes by making the labelling of data by field experts occur in real time as supposed to "post-hoc". The system is an early attempt in industry to close the gap between automated system development using large, unlabeled datasets and overly-intrusive real-time alarm systems that are often found in control room scenarios. Though it clearly shows the potential value in online expert data extraction systems; there clearly exists a further gap in the system that *automates* the input process, preventing the operator from having further duties that could exacerbate the cognitive load problem during highly demanding moments of the task.

2.3 Cognitive Load

A key factor cited in affecting operator performance is that of *Mental Workload* or *Data overload*, which in literature, appears synonymous with *cognitive overload, information overload* and *overload of mental effort or mental strain* and can be described in many different ways – subtle differences in definition are often due to the context being used. One such definition:

" *Information overload occurs when the amount of input to a system exceeds its processing capacity. Decision makers have fairly limited cognitive processing capacity. Consequently, when information overload occurs, it is likely that a reduction in decision quality will occur."* (Speier, Valacich and Vessey, 2007)

Work done in the area of data overload is widely ranging in domain, from the neuroscience perspective, tackling the physical factors that contribute to the problem (Chen, 2011), to the empirical measurement of the phenomenon using both objective and subjective measures (Hart and Staveland, 1988; Haapalainen *et al.*, 2010). A report by the United States Air Force Research Laboratory attempts to characterise the problem in the context of technological approaches. They provide three characterisations of cognitive workload:

"*1 – As a clutter problem where there is 'too much data': therefore, we can solve data overload by reducing the number of data bits that are displayed.*

2 - As a 'workload bottleneck' where there is too much to analyze in the time available: therefore, we can solve data overload by using automation and other technologies to perform activities for the user or to cooperate with the user during these activities.

3 - As a problem in 'finding the significance in data' when it is not known a priori what data from a large data field will be informative: therefore, we can solve data overload through model-based abstractions and representations. (Woods, 1984; Vincente and Rasmussen, 1992; Zhang and Norman, 1994) – better organizing that data to help people extract meaning despite the fact that what is informative depends on context."(Woods *et al.*, 2012)

Woods et al. also discuss the issues that attempts to 'solve' data overload using more technology. A feedback loop is created in which technological solutions create further problems that also require more technology to solve etc. Fittingly, they refer to this approach 'a little more technology will be enough'. Woods cites empirical studies (Norman, 1990a; Woods, 1993; Sarter, Woods and Billings, 1997) demonstrating that new systems almost always have surprising consequences or fail. Other examples of technological approaches failing include quotes from major figures in the space industry claiming that 'alarms and flashing lights' are useless, panning the approach to data overload problems to date. Conclusions such as these show clear motivation for the development of such systems as the SeeCoast system mentioned previously, that attempt to reduce the necessary input from an operator in real-time in order to develop a warning system that can more accurately produces "warnings" without further distracting the operator from their duties.

Cognitive load is clearly a multidimensional concept, as a determination of an individual's own subjective experience of a task's pressures, it can also differ significantly between individuals performing the same task. Measuring cognitive load is naturally also a multidimensional task; it can be measured both subjectively and objectively, with subjective measures more directly engaging an individual with their interpretation of load and objective measures inferring this through analysis of changing physiological responses.

2.3.1. Subjective Measures

It has been suggested that when asking an individual to assess the *workload* they experienced with a quantitative scale, their responses are limited and inaccurate. The nature of their responses are inclusive, or exclusive, of particular meaning based on what the individual deems to be relevant markers of *workload* at the time. These types of self-assessment also do not distinguish between differing factors that cause load e.g. *workload* caused by time pressure or by the stressful conditions under which the task was performed. These challenges lead to the development of the NASA Task Load Index (NASA-TLX). The objective of the NASA-TLX system was to create a workload rating scale that was sensitive to differing task types and sources of workload whilst remaining insensitive to inter-individual differences in workload perception. The NASA-TLX system uses six component scales, each one assessed by the individual to achieve an overall workload rating (Hart and Staveland, 1988). These scales are:

- Mental Demand - How much mental activity was required? Was the task mentally demanding?
- Physical Demand - Was the task physically demanding?
- Temporal Demand - Was the task slow or fast? How much time pressure did you feel?
- Performance - How successful were you at the task? How satisfied were you with your performance?
- Effort - How hard did you have to work (physically and mentally) to accomplish the performance you gave?

- Frustration - How irritated, stressed, and annoyed versus content, relaxed, and complacent did you feel during the task?

Each of these component scales is ranked by the participant between 0 – 21. The results are then combined with weights applied to each input to achieve a task load score for the task.

This measure is used widely throughout literature as the benchmark measure for subjective cognitive load. Other techniques for subjectively assessing cognitive load include the Subject Workload Assessment Technique (SWAT) (Reid and Nygren, 1988), Overall Workload scale (OW) (Vidulich, 1987) and the Modified Cooper-Harper (MCH) Scale (Harper and Cooper, 1986). These 4 key methods of subjective ratings scales were compared systematically by Hill et al (Hill *et al.*, 1992); they compared the techniques along four dimensions: operator acceptance, resource requirement and special procedures to determine if a single method was clearly preferable in the settings they were used in – this was primarily in military settings. They conclude that all four methods have unique characteristics that may make them more appropriate for different settings and objectives. OW and MCH are unidimensional scales (the output is measures on a single axis), which though less detailed may aid in the discovery of workload choke points in a control process more quickly. The extra effort required to obtain the more in-depth results produced by the NASA-TLX and SWAT is justified by their ability to more specifically diagnose what precisely is causing the workload in a situation and point toward ways to relieve the excessive workload. Ultimately, it was determined that NASA-TLX and OW were consistently superior in terms of sensitivity and acceptance by the participants of the examined studies.

These methods form the deductive nature of studying cognitive load in operator environments. They are held in direct contrast to the more inductive forms of cognitive load measurement; namely “objective measures” – that we will discuss in detail in the following section. It is not yet clear which methods produce the most *accurate* results, subjective methods are often argued to be a more direct measure as it is obtaining assessed cognitive state from the operators own formulation of cognitive load. These methods have been examined against the objective measures directly in several studies. Fallahi et al compared the NASA-TLX method against measured derived from monitoring the heart directly. They conclude that in the real-world scenario they gathered data from (city traffic control centre), that Heart Rate (HR), root mean square of successive differences (RMSSD), SDNN, ratio of low-frequency components to high-frequency components (LF/HF) and electromyography (EMG) amplitude all were significantly affected by increased traffic density and correlated with results obtained using the TLX (Fallahi *et al.*, 2016). In a contrary study, in a driving simulator, Shakouri et al determined that though the NASA-TLX correlated with the increase of traffic density (and therefore task difficulty, therefore cognitive workload); the same heart-related measures as recorded by Fallahi remained largely unaffected.

This is a timely research topic, contrasting the inductive and deductive methods still has much to investigate, as the examination of the "ground truth" of when an individual is "truly" experiencing increased mental workload appears still to be determined. Presently, there is clearly an operational difference in collecting the necessary data to make these determinations, with subjective measures being assessed post-hoc; there is always an element of memory required for operators to correctly recall their interpretation of the cognitive challenges of a task.

2.3.2. Objective Measures

Measuring cognitive load or stress objectively has been established as a flourishing field of study in the last 30 years. The theory behind measuring cognitive load using direct physiological measures derives from the understanding of the Autonomic Nervous System (ANS). The ANS, comprising the Sympathetic Nervous System and Parasympathetic Nervous

System (SNS and PNS) is largely responsible for the bodies involuntary activities such as variations in heart rate, temperature, sweat production, pupil diameter etc. An increase in stress leads to an increase in activity of the SNS and decrease in the PNS (Sharma and Gedeon, 2012). Measuring these variations is achieved by physically monitoring the outputs of these involuntary actions (e.g. using an Electroencephalogram to measure the voltage of signals sent to the heart or using a camera to measure the diameter of a pupil). These readings are then analysed for correlates to external stressors to the body.

The body of research in measuring cognitive load objectively is highly varied both in objectives and methods. Of the numerous methods of measuring Cognitive Load objectively, the two methods that stand out in prominence for applications in control room scenarios are that of pupillometry and heart related measures; we will assess the literature in these areas across a wide range of studies, and then review the literature from other methods.

Measuring Cognitive Load through Pupil Diameter Methods.

As well as reacting to changes in light, pupil diameter has been shown to react to different types of cognitive process. Observations of a relationship between the difficulty of mental arithmetic problems and the magnitude of the pupil’s dilation during the solution period were first made by Hess and Polt (Hess and Polt, 1964). Subsequent studies confirmed this correspondence in multiple contexts: arithmetic, short-term memory tasks of varying load, pitch discriminations of varying difficulties, standard tests of "concentration", sentence comprehension, paired-associate learning, imagery tasks with abstract and with concrete words and the emission of a freely selected motor response instead of an instructed response (Egeth and Kahneman, 1975). These studies all demonstrated an increase in dilation follows an increase in task demand or difficulty. It should be mentioned also that the underlying mechanism that drives cognitive pupillary effect and is still a topic of active enquiry. Recent studies suggest that pupillary response may be reflecting noradrenergic activity in the brain (Murphy *et al.*, 2014). The exact cognitive measure that pupil diameter is responding to is still not known precisely and remains an area of active enquiry. Many studies have measured pupil diameter against a large variety of tasks designed to manipulate different aspects of cognition such as tracking, switching and inhibition (van der Wel and van Steenbergen, 2018) discuss that studies across these domains show that pupil dilation closely respond to task demands.

Given that pupils react to changes in ambient light – it has also been shown that this effect and the effect on cognitive processes on is not additive. The advice of literature is to carefully control lighting when taking pupil diameter readings (Beatty and Lucero-Wagoner, 2000).

Methods of Measuring the Pupil Diameter.

In initial studies in literature, equipment involving two-way mirrors and cameras was used, with measurements been taken directly from photos. More advanced systems today use computer vision to track and measure the pupil from live video taken from either a stationary camera or a camera mounted on the participant's head. While more accurate in measurement (Marshall, 2002), head mounted eye trackers have been described as cumbersome – indicating some detriment to results gathered from them. Eye trackers that are not physically attached to the participant, by way of a head mount or a chin rest are called *remote eye trackers*. Pupil Foreshortening Error (PFE) is the main methodological concern with remote eye trackers. It refers to the error in pupil diameter measurement that occurs when the eye rotates away from the camera. A geometric model was developed by (Hayes and Petrov, 2016) to correct for this error resulting in a reduced Root Mean Square Error of 82.5%.

Klingner et al demonstrated that it is possible to determine cognitive load through pupil diameter using remote eye trackers (Klingner, Kumar and Hanrahan, 2008); replicating results produced in literature that utilized more complex, accurate equipment. Klinger also notes that they were unable to produce the systematic PFE observed by (Velichkovsky *et al.*, 2000) and suggest that the error is due to the method in which the eye tracker measures pupil diameter. The two main methods of measuring pupil diameter are: counting the number of pixels in the detected pupil and fitting an ellipse to the detected pupil and measuring the length of the major axis of the ellipse. The pupil itself is detected using an infrared light source that is reflected differently by the pupil than the surrounding eye, enabling simple subtraction on the image to classify the pupil's location. Klinger suggests that as the Tobii 1750 eye tracker used in their experiments uses the ellipse-fitting technique, it is not affected by the perspective distortion. Though he does mention that this does not mean that it is free from error.

Analysis of the Change in Pupil Diameter and Cognitive Load.

The link between change in pupil diameter and underlying cognitive processes is field of active enquiry for differing fields of study. Human Factors research utilize this method to optimize layouts in user interfaces or to determine where processes can become stressful so as to edit and change systems to reduce stress on operators that oversee that particular system (such as car drivers, control room operators or pilots). Psychology favours the pupil diameter as a measure to determine the types of cognitive process that have greatest effect on cognitive load and other processes; attempting to delineate which types of physiological response most closely reflect changes in certain specific mental processes.

Our project scope is seeking to apply the theory of objective cognitive load measurement to control room operators to determine times in which cognitive load was high. Resultantly, we will seek to assess the state-of-the-art for this field.

Here, we will employ our own review to assess which applications pupil diameter have been monitored and analysed to assess its correlation with cognitive load.

Papers were gathered from the Google Scholar, Science Direct and Elsevier repositories with no date restrictions. Terms used for the search were “pupil” + “cognitive” then + “load”, “workload” and “overload” the inclusion of these three terms was to include variations in terminology. The paper’s abstracts were then read and manually determined if they were to be included; any papers with coincidental key word matches that were clearly not assessing pupil diameter or cognitive load were excluded.

An initial total of 42 articles were found to fit the criteria for review. They were then first assessed to determine some key characteristics. We first determine whether the article is from an application or was domain-free or domain-independent. These were defined by whether or not the stimulus used to assess cognitive load for the participants was a specific test designed to increase cognitive load. Earlier studies mentioned above show researchers demonstrating the link between cognitive load and pupil diameter empirically using carefully designed tests such as the stroop-colour-word test and the n-back task. For this characterization, we will separate the studies based on the core stimulus test that participants were instructed to do; tasks such as reading, word games, memory tasks, auditory tasks and other categories of task specifically designed to contain mechanisms to only change the mental difficulty were separated from tasks that were either specifically from a domain such as a control room task or other tasks that have components that include in-task objectives such as games, driving, web browsing.

We then include whether other physiological variables were assessed, whether the results were compared to a subjective measure, which analysis technique was used on the pupil dilation data and whether results showed a significant correlate with cognitive load. Table 2.1 a to g presents a summary of the articles reviewed.

Reference	Year	Journal	Domain	Domain-free	Other Measures taken	Subjective method used	Pupil Diameter Method	Significant Correlate Found
(Kahneman and Beatty, 1966)	1966	Science		Auditory listen and report (digit span task).	No	No		Yes
(Jainta and Baccino, 2010)	2010	International Journal of Psychophysiology		Basic arithmetic tasks.	No	no		
(Kin and Epps, 2016)	2006	Computer Methods in Biomedicine		Auditory listen and recall	no	no	TEPR rate analysis	yes
(Pedrot*ti et al*>, 2014)	2014	International Journal of Human-Computer Interaction	Simulated driving task		no	no	Change in Mean PD across trials and TEPR analysis	yes
(Van Gerven *et al.*, 2004)	2004	Psychophysiology		Memory-search task (number memorization).	no	no	Change in Mean PD across trials	yes
(Coyne and Sibley, 2016)	2016	Proceedings of the Human Factors and Ergonomics Society		Auditory listen and report (digit span task).	No	No	Change in Mean PD across trials	yes
(Peysakhovich, Dehais and Causse, 2015)	2015	Procedia Manufacturing	Simulated Piloting Task			NASA-TLX	Change in Mean PD across trials and TEPR analysis	yes
(Privitera *et al.*, 2010)	2010	Journal of Vision	Target detection		No	No	Change in Mean PD across trials	yes

Table 2.1a. List of studies assessed in literature review for examination of effect of cognitive load on pupil diameter changes.

Reference	Year	Journal	Domain	Domain-free	Other Measures taken	Subjective method used	Pupil Diameter Method	Significant Correlate Found
(Kiefer *et al.*, 2016)	2016	Lecture Notes in Computer Science	Map search task		No	No	Change in Mean PD across trials	yes
(Hossain and Yeasin, 2014)	2014	IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops		Vigilance, memory and arithmetic tasks	No	No	Hilbert analytic phase on TEPRs	yes
(Klingner, Kumar and Hanrahan, 2008)	2008	Proceedings of the 2008 symposium on Eye tracking research & applications		Multiplication, short term memory and aural vigilance	No	No	TEPR analysis	No, simply seeking if TEPR could be captured with remote device (which one can).
(Krejtz *et al.*, 2018)	2018	PLoS ONE		Mental arithmetic	No	NASA-TLX	Change in Mean PD across trials	Yes
(Hess and Polt, 1964)	1964	Science		Mental Arithmetic	No	No	Percentage increase in pupil diameter	yes

Table 2.1b. List of studies assessed in literature review for examination of effect of cognitive load on pupil diameter changes.

Reference	Year	Journal	Domain	Domain-free	Other Measures taken	Subjective method used	Pupil Diameter Method	Significant Correlate Found
(Van Acker *et al.*, 2020)	2020	International Journal of Industrial Ergonomics	Manual assembly procedures		No	Likert-scale	Change in Mean PD across trials	no
(Porter, Troscianko and Gilchrist, 2007)	2007	Quarterly Journal of Experimental Psychology		Visual search task	No	No	Change in Mean PD across trials	yes
(Mosaly, Mazur and Marks, 2017)	2017	Ergonomics		Working memory task	no	no	Change in TEPR	yes
(Alnaes *et al*>, 2014)	2014	Journal of Vision	Multiple object tracking task		FMRI	No	Change in Mean PD across trials	yes
(Ren *et al.*, 2014)	2014	Annals of Biomedical Engineering		Stroop colour word test	GSR	Yes	Custom PD change windowing technique	yes
(Peysakhovich, Vachon and Dehais, 2017)	2017	International Journal of Psychophysiology		Arithmetic and n-back task	No	no	Change in Mean PD across trials	yes
(Bhavsar, Srinivasan and Srinivasan, 2016)	2016	Industrial and Engineering Chemistry Research	chemical production facility control room		no	no	custom ‘ steady-state‘ value analysis	“ results promising ”

Table 2.1c. List of studies assessed in literature review for examination of effect of cognitive load on pupil diameter changes.

Reference	Year	Journal	Domain	Domain-free	Other Measures taken	Subjective method used	Pupil Diameter Method	Significant Correlate Found
(Peysakhovic*h et al*>, 2015)	2015	International Journal of Psychophysiology		Short term memory task			TEPR analysis using frequency amplitude	yes
(Hoeks and Levelt, 1993)	1993	Behavior Research Methods, Instruments, & Computers		Listen-respond/read-respond task	No	No	TEPR analysis over trials	Yes
(Wahn *et al.*, 2016)	2016	PLoS ONE	Multiple object tracking task		No	No	Change in Mean PD across trials	yes
(Ferdous, 2014)	2014	Proceedings of the IEEE Visualization Conference	Data visualisation		No	No	Change in Mean PD across trials	Yes
(Denison, Parker and Carrasco, 2019)	2019			Stimulus discrimination and estimation task	No	No	TEPR analysis	yes
(Wong and Epps, 2016)	2016	Computer Methods and Programs in Biomedicine		Digit span task	No	no	TEPR Analysis	yes
(Rozado and Dunser, 2015)	2015	Computer		Mental arithmetic	EEG	no	Change in Mean PD across trials	Yes, reduced error rate in EEG

Table 2.1d. List of studies assessed in literature review for examination of effect of cognitive load on pupil diameter changes.

Reference	Year	Journal	Domain	Domain-free	Other Measures taken	Subjective method used	Pupil Diameter Method	Significant Correlate Found
(Rozado and Dunser, 2015)	2015	Computer		Mental arithmetic	EEG	no	Change in Mean PD across trials	Yes, reduced error rate in EEG workload detection
(Marquart and de Winter, 2015)	2015	PeerJ Computer Science		Mental arithmetic	no	NASA-TLX	Mean pupil diameter, mean PD change, mean PD change rate	yes
(Scharinger, Kammerer and Gerjets, 2015)	2015	PLoS ONE		Reading task	EEG	no	Change in Mean PD across trials	yes
(Lisi, Bonato and Zorzi, 2015)	2015	Biological Psychology		Auditory-distracted visual task	no	no	TEPR analysis	yes
(Hogervorst, Brouwer and van Erp, 2014)	2014	Frontiers in Neuroscience		n-back task	EEG, GSR, respiration, ECG		Change in Mean PD across trials	Yes
(Yan, Wei and Tran, 2019)	2019	International Journal of Industrial Ergonomics	Maritime operation control room		no	NASA-TLX and SWAT	Mean PD and ANN technique	yes

Table 2.1e. List of studies assessed in literature review for examination of effect of cognitive load on pupil diameter changes.

Reference	Year	Journal	Domain	Domain-free	Other Measures taken	Subjective method used	Pupil Diameter Method	Significant Correlate Found
(Yan, Wei and Tran, 2019)	2019	International Journal of Industrial Ergonomics	Maritime operation control room		no	NASA-TLX and SWAT	Mean PD and ANN technique	yes
(Marinescu *et al.*, 2018)	2018	Human Factors	object tracking video game		RR intervals, breathing rate, facial thermography	ISA and NASA-TLX	Change in Mean PD across trials	yes
(Palinko *et al.*, 2010)	2010	Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications	Simulated driving task		No	No	Change in Mean PD across trials	yes
(Mandrick *et al.*, 2016)	2016	Biological Psychology		N-back task	fNIRS, cardiovascular	DP15 scale	Median pupil diameter over trial	yes
(Zhai and Barreto, 2006)	2006	Annual International Conference of the IEEE Engineering in Medicine and Biology		Stroop Color-Word Interference Test	GSR, blood volume pulse, skin temp	No	Change in Mean PD across trials	Yes, above other measures
(Bernhardt *et al.*, 2019)	2019	Applied Ergonomics	Simulated ATC task		EEG	No	Change in Mean PD across trials	yes

Table 2.1f. List of studies assessed in literature review for examination of effect of cognitive load on pupil diameter changes.

Reference	Year	Journal	Domain	Domain-free	Other Measures taken	Subjective method used	Pupil Diameter Method	Significant Correlate Found
(Mallick *et al.*, 2017)	2017	Proceedings of the 2nd Workshop on Eye Tracking and Visualization	Tetris Game		No	No	Change in Mean PD across trials	yes
(Jimenez-Molina, Retamal and Lira, 2018)	2018	Sensors (Switzerland)	Web-browsing task		electro dermal activity, PPG, EEG, temperature	no	Change in Mean PD across trials	yes
(Lecoutre *et al*>, 2015)	2015	PhyCS 2015 - 2nd International Conference on Physiological Computing Systems	Operational video game set up		EEG, HRV	NASA-TLX	Change in Mean PD across trials	Correlates with subjective measures
(Truschzinski *et al.*, 2018)	2018	Applied Ergonomics	Simulated Air Traffic Control task		no	Multidime nsional Mood State Questionn aire	Change in Mean PD across trials	yes
(Puma *et al.*, 2018)	2018	International Journal of Psychophysiology	Air Traffic Control training ask (multitasking system)		EEG	NASA-TLX	Change in Mean PD across trials	yes

Table 2.1g. List of studies assessed in literature review for examination of effect of cognitive load on pupil diameter changes.

From the literature reviewed in Tables 2.1 (a-g).we can determine the most common methods of assessing pupil diameter and cognitive load. Twenty six (62%) of the studies used, at least in part, the change in Mean PD across trials technique; eleven (26%) analysed in some way, the Task-Evoked Pupillary Response (TEPR); one (2.3%) analysed the median change in diameter across trials; one (2.3%) analysed the mean change *rate* of pupil diameter; one (2.3%) utilized an Artificial Neural Network and two (4.7%) used a custom technique.

The Mean Pupil Diameter across Trials.

To assess whether or not a particular stressor, whether it be applied or not, is influencing change in pupil diameter, the stimulus scenario is broken into multiple trials (or phases). There will be a baseline phase, in which the participant is presented with no stimulus at all, to determine a control value for the pupil diameter. The experiment then continues with multiple trials of differing perceived cognitive load. In the example of the Stroop-colour word interference test (SCWT), used here in two articles (Zhai and Barreto, 2006; Re*n et al*>, 2014) a word is presented to the participant to read aloud (this is the congruent condition), the word is that of a colour. The participant is then given a differing selection of words to read in which the colour of the word to read is not that being described by the word, this is the incongruent condition designed to interfere with the more automated cognitive task of reading. In our reviewed studies, variations of the SCWT are created with specific trials being designated as congruent and incongruent, the hypothesis being that an incongruent condition will create higher demand on the participant's mental effort, increasing cognitive load and by extension, increase pupil diameter. These hypotheses are tested by comparing the mean pupil diameter of the participant throughout a congruent or incongruent trial and comparing these values against each other and that of the baseline to determine if a significant effect has been observed.

This method is the most prolific in the literature studied. The general approach is to create a stimulus that can be manipulated in such a way as to change the cognition of the participant in some controlled way, create a baseline trial, then trials for varying degrees of load, measure the average pupil diameter of each trial then perform statistical analyses to determine significance. These results can then be compared to another standard for further analysis: in our review, 11 (26%) of studies compared the results from pupil diameter analysis with other physiological measures and 11 (26%) compared the pupil diameter results against some subjective measure such as those discuss previously. There is often correlation between pupil diameter and subjective measures, in our review, every study demonstrated correlates with subjective measures.

The Task-Evoked Pupillary Response (TEPR)

The Task-Evoked Pupillary Response is an indicator of brain processing that underlies the dynamic aspect of human cognition (Beatty and Lucero-Wagoner, 2000). The TEPR, similar to that of the Event-Related Potential (ERP) in EEG study, is a response of the pupil that occurs shortly after a task is started and subsides quickly when the task is complete (Kahneman and Beatty, 1966). In literature, the TEPR is modelled as a peak amplitude that is reached some time after an onset stimulus with reference to a baseline value. The TEPR requires precise knowledge of when the stimulus occurred and knowledge of the baseline pupil size pre-stimulus. The TEPR is examined as an average response measured over multiple trials, given inter-individual differences in pupil size, this makes normalization of the data by way of reference to a baseline absolutely critical. The discovery of the baseline value of the pupil is often achieved by programming in some rest or calm state just prior to the onset stimulus. This approach has been reviewed for the assumed nature of a “rest state” over stimulus scenarios that consist of multiple events sequentially, it was determined that the baseline should be re-established before every TEPR for accurate analysis, as the baseline pupil level shifts incrementally throughout sustained activity (Mosaly, Mazur and Marks, 2017).

Custom Techniques

Of the literature reviewed here, two articles used a custom technique to derive cognitive load from pupil diameter. One study utilized the stroop-colour word interference test as a stimuli to stress the participant, then using pupil diameter, attempt to delineate between congruous and incongruous segments of the task post-hoc (Ren *et al.*, 2014). In this study they split the data into pairs for each participant, congruous and incongruous. These pairs have three features extracted from them; the mean PD value, the max PD value and the difference between the Walsh coefficient after Walsh transform based on the PD value at the onset of each segment. These features were then used as features for a classifier to determine if any given segment was "congruent" or "incongruent", which was achieved with a successful classification rate of 85% - it should be mentioned the data was first filtered to only include segments from participants that already showed significant differences in PD values between segments.

Another study that developed a custom method used a chemical production plant control room simulator as its stimulus. They focus on the continuous change of PD as a basis for their analysis; hypothesizing that the pupil will reach a 'steady-state' after rising from the onset of increased workload and again after falling. They define this steady state by sliding a 3 second window over the data, and defining the window as 'steady' if the pupil size obtained is $-0.005fW < \mu W < 0.005fW$, where fW is the range of pupil size values within the window and μW is the mean value of PD for the window. They then compare the time taken to reach this state from the onset of the task (Bhavsar, Srinivasan and Srinivasan, 2016).

Domains of Stimulus Scenarios

Of the literature reviewed, 18 (43%) used a participant stimulus scenario that was either from an applied domain or specifically not a controlled cognition task. The stimulus tasks ranged widely; 5 (28%) were from control rooms, 5 (28%) from game set ups, 2 (11%) from driving simulators, the remainder included aircraft piloting, target detection set up, map reading, manual assembly tasks, data visualization and web browsing (5.5% each). This body of literature does demonstrate that pupil diameter has been considered for use measuring cognitive load in applied scenarios.

Of all the reviewed articles here, only 1 failed to find a significant correlate with task difficulty and cognitive load (Van Acker *et al.*, 2020). This study used head-mounted eye trackers whilst the participants were instructed to build increasingly complicated physical structures on a desk. The nature of this task was to determine if pupil diameter could be implemented in a non-laboratory setting without screen-based interfaces. The results determined that although the subjective measures correlated significantly with task difficulty, pupil diameter did not, perhaps shedding light on the infancy of the field in more deployable work scenarios. It could also simply demonstrate the limitations of the applied fields in which this method can be deployed; being only usable in more controlled and less physically active settings such as control rooms.

Twenty eight percent of the literature applied pupil diameter as a cognitive load measure in control room settings, ranging from air traffic control to maritime operations to chemical production facilities. Puma et al increased the difficulty of the task in four incremental loads, the study is primarily about discerning where the ceiling may exist is terms of EEG’s ability to discriminate between these levels of difficulty and PD was included as a comparative measure. The PD did increase significantly for the first 3 tasks but not significantly between task 3 and 4, the tasks being clearly separated as trials and mean PD values compared across each (Puma *et al.*, 2018). Truschzinski et al programmed a custom air traffic control task, with the participants being presented with a top-down display of aircraft that crossed paths, the objective being to prevent collisions. The difficulty of the task was tuned by increasing the frequency in which aircraft appeared. The experiment was conducted in stages, in between each stage the participant would complete a questionnaire to determine their mental state after the task. The stages were considered "low conditions" and "high conditions" they measured PD mean average 1.5s after events logged in the simulator such as a crash or an aircraft appearing. It was determined that mean PD was significantly higher during the high conditions than the low conditions (Truschzinski *et al.*, 2018). Bhavsar et el used PD as a cognitive load measure to model the responses of control room operators and specific events of interest. The six events of interest were manually programmed events in a chemical production facility control room. They hypothesize that these abnormal events of interest would peak the cognitive load of the operators, as they have to respond to the issue, discover the cause and rectify it before the balance of the system they are maintaining fails. The conclusions from the work are unclear but they claim that the results demonstrate that PD is a valid measure of real-time cognitive load in control room operators, though it is not established how this is achieved as a real time metric (Bhavsar, Srinivasan and Srinivasan, 2016).

There were five articles reviewed here that used a form of video game as the stimulus scenario to determine cognitive load using pupil diameter. Lecoutre et al simply uses an existing Playstation video game (Rayman Origins) as the stimulus scenario, their assumption being that the game has built-in methods of objectively increasing the difficulty for them to compare their measures (EEG, HRV and PD) against. They claim that the set up; an office chair and flat screen tv make "it close to a real-life situation" (Lecoutre *et al.*, 2015). The results show a significant correlation between PD, the subjective measure and game difficulty. They also collected performance data from the game to determine correlation, this data consisted of collected items within the game set up; this measure did not show any correlation with the other measures. This could be a relevant measure of performance data and cognitive load or simply a limitation given then incredibly specific measure of performance chosen, which has no potential to be transferred to another domain other than Rayman games for the Playstation. Marinescu et al designed a custom video game in which participants had to find and 'shoot' balls as they fell down the screen. The game was designed so that the difficulty could be manipulated in two ways; firstly, the number of target balls could be increased, secondly, the method of finding the target ball. In the easy condition, the target balls were red, whereas other falling balls were different colours and in the hard condition, all the balls were the same colour but were numbered, with the target balls being odd numbers, introducing another cognitive element to the tracking task (Marinescu *et al.*, 2018). Mallick et al demonstrated that PD correlated with game difficulty whilst playing Tetris (Mallick *et al.*, 2017). Wahn et al measured PD over multiple trials over multiple days for a multiple object tracking task. Multiple objects would appear on the screen, the target objects would be highlighted, then un-highlighted and scrambled, the correct targets then need to be found by the participant. They conclude that pupillometry provides a viable metric for precisely assessing attentional load and task experience in visuospatial tasks (Wahn *et al.*, 2016).

Two of the papers reviewed assessed pupil diameter as a measure of cognitive load during driving tasks. Palinko et al had participants follow a lead car whilst driving normally; as they were driving they were to play "20 Questions" verbally with a questioner. The task was interrupted by playing the "last letter" (LL) word game at various intervals. They hypothesize that cognitive load would be greater during the interruption of the LL game – which was verified by way of mean pupil diameter change during these tasks (Palink*o et al*>, 2010). Pedrotti et al focused more on the challenge of the driving itself for their stimulus task. Their stimulus to increase cognitive load used a lane change task, which was then staged for increased difficulty by including a distracting beep that indicated their driving was poor (which occurred regardless), and then also by the presence of observers that were ostensibly assessing their driving. They used an artificial neural network to classify which trial represented which level of difficulty based on their pupil diameter – which was achieved with approximately 83% accuracy (Pedrotti *et al*., 2014).

Summary of Pupil Diameter Literature

It is very clear from the literature that pupil diameter (PD) represents a valid measure of cognitive load and has been well established as a measure that can be deployed in applications outside of stringent lab conditions. The literature reviewed demonstrates PD as a measure still under investigation with regards to the specific cognitive process's it is reflecting. This being said, its use in applications to demonstrate significant change with respect to certain stimulus is certainly promising for use in future technologies. For this project, we seek to use a physiological measure to detect events in control room situations; the PD is certainly a viable measure to assess given its simple implementation, lack of interference to the operator and clear response to changing scenarios based on mental load.

2.3.3 Analysis of Cognitive Load through Heart Rate Measures

Due to its direct influence on the Autonomic Nervous System (ANS), cardiac signals can also reflect cognitive load (Cohen, Janicki-Deverts and Miller, 2007). Though it is well known that an increase in physical effort such as exercise can increase the heart rate, other metrics of cardiac measurement can reveal variances that relate to cognitive processes.

Signals from the nervous system can be measured using skin mounted electrodes on the torso. This Echocardiogram (ECG) signal is characteristic of the polarisation and depolarisation of the heart during beats. The resulting waveform is can be analysed for a pattern of repeating morphological patterns known as QRS complexes. The peak of the complex is the R point, accurately measuring the time between these peaks gives the R-R interval. Heart Rate Variability or HRV is the variability of successive R peaks in ECG signals taken from the heart (see figure 2.1).

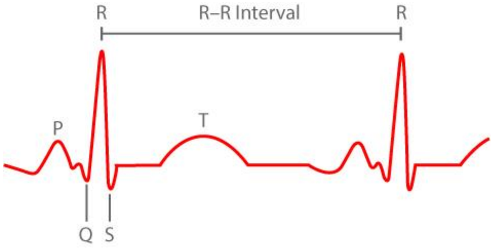


Figure 2.1. Illustration of QRS Complex and R-R Interval (Science, 2017).

The number of R peaks that occur within a minute is referred to as Beats per Minute (BPM) or Heart Rate (HR) and is the most common form of heart beat related measure. This creates an average R-R interval over the course of a minute. HRV is most frequently analysed in two forms: in the time domain and the frequency domain. Three major frequency bands are usually examined for HRV:

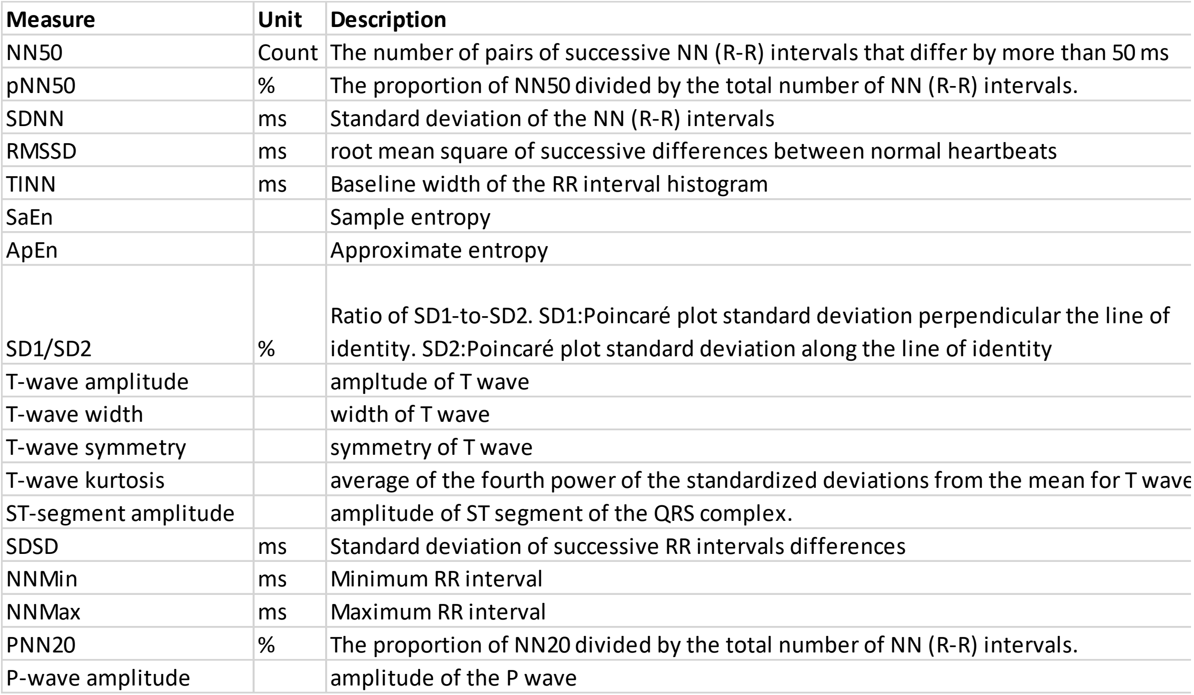
1 - The low frequency band (from 0.02 to 0.06 Hz). The energy in this band mainly comes from vasomotor activity involved in the regulation of body temperature and from slow linear and nonlinear trends in heart rate.

2 - The mid-frequency band (from 0.07 to 0.14 Hz). The energy in this band comes from mechanisms involved in the short-term regulation of arterial pressure.

3 - The high-frequency band (from 0.15 to 0.50 Hz). The energy in this band mainly reflects the effects of respiratory activity on the cardiac interval signal. (Aasman, Mulder and Mulder, 1987).

There are many different time-domain measures of HRV, one review counted 18 different time domain measures across the literature reviewed. These measures are detailed in Table 2.2.

Table 2.2. Common HRV Time Series Measures with unit and descriptions (Shaffer and Ginsberg, 2017).



Measure	Unit	Description
NN50	Count	The number of pairs of successive NN (R-R) intervals that differ by more than 50 ms
pNN50	%	The proportion of NN50 divided by the total number of NN (R-R) intervals.
SDNN	ms	Standard deviation of the NN (R-R) intervals
RMSSD	ms	root mean square of successive differences between normal heartbeats
TINN	ms	Baseline width of the RR interval histogram
SaEn		Sample entropy
ApEn		Approximate entropy
SD1/SD2	%	Ratio of SD1-to-SD2. SD1:Poincaré plot standard deviation perpendicular the line of identity. SD2:Poincaré plot standard deviation along the line of identity
T-wave amplitude		ampltude of T wave
T-wave width		width of T wave
T-wave symmetry		symmetry of T wave
T-wave kurtosis		average of the fourth power of the standardized deviations from the mean for T wave
ST-segment amplitude		amplitude of ST segment of the QRS complex.
SDSD	ms	Standard deviation of successive RR intervals differences
NNMin	ms	Minimum RR interval
NNMax	ms	Maximum RR interval
PNN20	%	The proportion of NN20 divided by the total number of NN (R-R) intervals.
P-wave amplitude		amplitude of the P wave

Extensive literature outlines the relationship between each of the measures outlined above and various bodily processes. In the review paper, these measures were used to assess their response to mental workload (Tao *et al.*, 2019). In their review, IBI (inter-beat interval), pNN50, SDNN and RMSSD were respectively the most commonly used measures in the time-domain, with IBI showing a significant decrease with the increase in cognitive load. The review also examined frequency domain features and found that the ratio between the low frequency (LF) and high frequency (HF) or LF/HF ratio was the most widely used frequency-domain measure, with 75% of studies reporting sensitivity to cognitive load.

Extensive literature review list

In this section we will examine the literature to assess the state of the art and relative popularity of these measures across multiple studies. We will use the same criteria as with the pupil diameter review. Search terms were “heart” + “cognitive” then “load”, “workload” and “overload” to account for variability in terminology.

Reference	Year	Journal	Domain	Domain-free	Other Measures taken	Subjective method used	Cardiovascular Method	Minimum Window Length	Result
(Pereir*a et al*>, 2017)	2017	Computer Methods and Programs in Biomedicine		Trier Social Stress Test (TSST)	No	No	AVNN, RMSSD, SDNN and pNN20	50s	AVNN +50s window shows reliable correlate
(Mansikka, Simola, *et al*>, 2016)	2016	Applied Ergonomics	Flight Simulator		no	no	Mean RR SDNN, Mean HR, RMSSD, NN50, pNN50 HR	-	Measures sensitive to task difference even if performance did not
(Hidalgo-Muñoz *et al.*, 2018)	2018	International Journal of Psychophysiology	Flight Simulator		EEG		SDRR, RMSSD, pNNx	-	HR sensitive to increased cognitive load
(Tattersall and Hockey, 2006)	2006	Human Factors: The Journal of the Human Factors and Ergonomics Society	Flight Simulator		no	SWAT	Mean HR, MR, HF	4 min	MF only metric that significantly correlates
(Luque-Casado *et al.*, 2016)	2016	Biological Psychology		N-back task		NASA-TLX	RMSSD, pNN50, HR V/HF	4 min	HRV sensitive to attention load more than cognitive load

Table 2.3a. Literature reviewing impact of cognitive load on various psychophysiological signals.

Reference	Year	Journal	Domain	Domain-free	Other Measures taken	Subjective method used	Cardiovascular Method	Minimum Window Length	Result
(Ryu and Myung, 2005)	2005	International Journal of Industrial Ergonomics	Aircraft Landing Simulator tracking task	Arithmetic task	Eye blink, EEG	NASA-TLX	MF	5 min	Significant change for tracking task but not arithmetic
(Nickel and Nachreiner, 2004)	2004	Human Factors: The Journal of the Human Factors and Ergonomics Society		AGARD-STRES battery	respiration	two-level sequential judgment	0.1hz power HRV	128s	HRV only discriminates between work and rest
(Ding *et al.*, 2020)	2020	Ergonomics		Mental arithmetic	EMG, EDA, respiration	NASA-TLX	HF/LF ratio	2 min implied	No significant difference in task difficulty
(Mansikka, Virtanen and Harris, 2019)	2019	Ergonomics	Flight simulator task		no	NASA-TLX, MCH	Inter-beat-interval	3 min	Can differentiate between rest/task and low-med/high difficulty but not low/med difficulty
(Zhang *et al*>, 2014)	2014	2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society		Go/No-go visual reaction test, Stroop test, fast counting test, speed run test, visual	EEG, GSR	no	max, min, mean, median, SDNN, SDSD, NN50, pNN50, RMSSD	-	Sensitive, but significantly less that EEG and less than GSR

Table 2.3b Literature reviewing impact of cognitive load on various psychophysiological signals.

Reference	Year	Journal	Domain	Domain-free	Other Measures taken	Subjective method used	Cardiovascular Method	Minimum Window Length	Result
(Heine *et al.*, 2017)	2017	Applied Ergonomics	Driving simulation	N-back test	no	NASA-TLX	20 time and freq domain features	6 min	General response to increase cognitive load but no feature totally discriminatory
(Hogervorst, Brouwer and van Erp, 2014)	2014	Frontiers in Neuroscience		N-back test	EEG,GSR, respiration, pupil diameter		RMSSD, HF, MF	2 min	RMSSD barely significant, HRV worst performing of all objective measures.
(Shakouri *et al.*, 2018)	2018	International Journal of Industrial Ergonomics	Driving simulator		no	NASA-TLX	RMSSD, HF, LF, HF/LF ratio	Not clear.	HRV largely not affected by more challenging environments
(Sun *et al.*, 2012)	2012	Mobile Computing, Applications, and Services		Stroop-colour word test during physical activity	GSR	no	Mean RR, SDRR, Mean HR, SDHR, RMSSD, pNN50, LF, HF, LF/HF ratio	60secs	Mean HR and Mean RR sensitive, the rest too affected by movement.
(Glenn F. Wilson, 2002)	2002	The International Journal of Aviation Psychology	Real piloting task		EEG eye blinks, electro dermal activity	Custom	LF, HF	2 min	No significant changes
(Matthews *et al.*, 2017)	2017	Personality and Individual Differences	Unmanned Ground vehicle control room		EEG,TCD, fNIRS, eye fixation	DSSQ	IBI and “ HRV”	-	No significant correlate found

Table 2.3c. Literature reviewing impact of cognitive load on various psychophysiological signals.

Reference	Year	Journal	Domain	Domain-free	Other Measures taken	Subjective method used	Cardiovascular Method	Minimum Window Length	Result
(Haapalain*en et al*>, 2010)	2010	-		Elementary cognitive tasks	EEG, Pupil diameter, GSR, heat flux	custom	IBI, ECG MAD	-	ECG MAD significant change with cognitive load
(Mandrick *et al.*, 2016)	2016	Biological Psychology		n-back tasks	fNIRS, Pupil diameter	DP15 Scale	RMSSD	-	Significant decrease with task difficulty
(Tjollen*g et al*>, 2017)	2017	Applied Ergonomics	Driving simulator	N-back task	no		IBI, RMSSD	100secs	Significant correlation with task difficulty
(Lecout*re et al*>, 2015)	2015	PhyCS 2015	Operational video game		EEG, pupil diameter	NASA-TLX	LF/HF ratio	-	HRV trends with task difficulty but not significantly
(Fallahi *et al.*, 2016)	2016	Applied Ergonomics	City traffic control room		EMG	NASA-TLX	LF, HF, Mean HR, SDNN, RMSSD, pNN50, LF/HF Ratio	5 min	Significant change found in all measure with increase task demand.
(Haapalain*en et al*>, 2010)	2010	-		Elementary cognitive tasks	EEG, Pupil diameter, GSR, heat flux	custom	IBI, ECG MAD	-	ECG MAD significant change with cognitive load

Table 2.3d. Literature reviewing impact of cognitive load on various psychophysiological signals.

Domains of Stimulus Scenarios

Across our review, 12 (57%) of the total studies reviewed in Tables 2.3 (a-d) used domains from industrial settings or stimulus' that were not tasks specifically designed to alter cognitive load.

Four (19%) of the studies examined the HRV of pilots during a simulated flight situation. Mansikka et al reported that both the time and frequency measures used were capable of significantly differentiating between different types of piloting task even when the performance data was insignificant. Though they note further research is required, this demonstrates that the measures used here were sensitive to a notion that was not able to be detected by an outside observer, as the pilots performance did not change enough to determine which task they found more difficult (Mansikka, Virtanen, *et al.*, 2016). Hidalgo et al and Tattersall and Hockey reported different findings for their studies both performed in a flight simulator, with Hidalgo (Hidalgo-Muñoz *et al.*, 2018) reporting that mean RR was sensitive to increased cognitive load and Tattersall and Hockey (Tattersall and Hockey, 2006) reporting that only the MF component of the HRV was sensitive to changes in task difficulty. A further study by Mansikka demonstrated that IBI was sensitive and showing the difference between pilots at rest and performing a simple task and the difference between a simple/medium difficulty task and a very difficult task but not between simple and medium (Mansikka, Virtanen and Harris, 2019). This could demonstrate that the tasks that one wishes to differentiate between must be of significant difference in mental difficulty in their own right.

Three (14%) of the reviewed studies employed a driving simulator as a stimulus to determine the cardiac response to driving tasks. The results of these scenarios show a consistency in that there is no clear sign that the measures respond to task difficulty. Heine et al assessed 20 different measures across driving tasks of increasing difficulty and determined that though the measures generally trended with cognitive load, none of the measures were totally discriminatory between different types of task (Heine *et al.*, 2017). Shakouri et al determined no significant correlates between task difficulty and the HRV measures the calculated (Shakouri *et al.*, 2018). The study by Tjolleng et al was the only study to conclude a significant change in HRV with task difficulty, but it should be mentioned that they adjusted the task dificulty by introducing a verbal n-back task to the driver to increase task difficulty, so not measuring the driving difficulty directly (Tjolleng *et al.*, 2017).

Two (9.5%) of the studies used a control room scenario as the stimulus for their studies. Matthews et al used a simulated control room set up for an unmanned ground vehicle system in a military environment, with different preprogrammed scarios used to increase cognitive load. The simulator involved a number of complex tasks for the participant to moniotor and respond to appropriately, the study was set up predominately to study EEG though HRV among other psycophyioslogical measures were also examined. They conclude that there was not significant relationship between the HRV and the tasks, but do offer an explaination that given the highly complex and demanding nature of the task, the results may indicate stress-driven emotion-regulation, reflecting CNS-ANS integration (Matthews *et al.*, 2017). This may show a potential limitation of using HRV as a specific marker of task dissagregation when the environment has a large number of simultaneous tasks of high complexity such as surveilling a complex environment such as this. Another study monitored control room operators that were monitoring a city traffic centre, they noted that across all measures taken for HRV, they found a significant change between the low-density and high-desnity traffic conditions (Fallahi *et al.*, 2016). The two studies are not really comparable in terms of results given the significantly different nature of the stimuli and study set up. This being said, the trafic monitoring study's differing task difficulties were far more broad in scope and significantly longer-form in terms of time (the particiapnts were monitored over 12 hour periods).

The study by Lecoutre et al was also found in the review for pupil diameter as both of these psychophysiological measures were taken. In this operation video game set up, HRV did not correlate with the subjective measure taken or with the task difficulty whereas the pupil diameter measure correlated significantly with both. The authors suggest the small number of participants (8) may explain the results (Lecoutre *et al.*, 2015).

Metrics Used In Review

Across our review, we found no prevailing metric consistently used for either a specific measure of cognitive load or task difficulty, the tally of metrics used is shown in table 2.4.

Table 2.4. Tally of metrics used within Literature cases.

Measure	Count
RMSSD	10
HF	7
pNN50	5
LF/HF Ratio	5
SDNN	4
Mean HR	4
IBI	4
LF	4
MF	3
Mean RR	2
NN50	2
SDRR	2
pNNx	1
AVNN	1
pNN20	1
SDHR	1
SDSD	1
ECGMAD	1
HRVTRI	1

The most used metric was the RMSSD, followed by the HF, pNN50, LF/HF ratio etc. There was no reasoning offered for the use of any of the measures in the review; this could be due to the speculative element of the research in this field at this time. As the technology to assess HRV becomes more accessible, researchers are applying it to a wider variety of scenarios, selecting a large number of metrics to apply to see which is the most sensitive. The lack of consistency in the discussion of the selection of the measure is noted in some cases as an element to be examined in future research, though this scattered approach could be a systematic limitation in applying this research in this way.

HRV Methodology Discussion

As mentioned briefly above, there is little consistency in the methodology of using HRV as a stress or cognitive load metric across the literature surveyed; measures are either picked arbitrarily or without explanation and their implementation variables are not applied or controlled with any consistency. This being said, there is clearly evidence that HRV is a valid measure of cognitive load across the literature as a whole with most studies demonstrating a correlation with the controlled difficulty of a task. These findings are consistent with a 2015 review of HRV as a measure of stress assessment (Castaldo, Melillo and Pecchia, 2015). The authors also found consensus among the literature across at least seven HRV measures that changed consistently with mental stress. They also recommend that studies are defined clearly in terms of length of HRV measures and the study protocol and to use statistical tests of results consistently. A further paper also criticizes the lack of standardization and assumptions behind studies that use ultra-short HRV features (<5min) in place of short term features (nominally 5 mins) (Pecchia *et al.*, 2018). The paper then demonstrates through literature review, the large heterogeneity in present studies and outlines the pitfalls in the statistical methods used to justify conclusions when assumptions about the validity of significance are made. For example, several studies claimed that as there was significant correlation between results gathered using short-term features and the same gathered with ultra-short term they were valid surrogates of each other. This conclusion is flawed due to the nature of HRV data being non-normally distributed, therefore the parametric significance test is invalid, the author notes the data first requires log-transformation.

Summary of HRV Literature

The literature clearly demonstrates that HRV is a valid and sensitive measure of cognitive load. The subject is clearly very timely and heterogeneous in its application and domains. The literature reviewed here follows a certain trend, which is that studies start with the assumption that HRV is sensitive to cognitive load, then apply multiple measures of HRV in a domain scenario to determine the link between certain cognitive tasks and their HRV response. This approach seems speculative in the literature assessed that used a specific application domain. These human-centric scenarios are often investigated to optimize processes and the success in literature from psychological and medical research demonstrating the value of HRV as an objective measure of cognitive processes is clearly of interest in this developing field. As such, it appears to be applied in an inconsistent manner, which may affect the validity of the results when viewed in a narrower context. For example, many studies simply seek statistical differences between scenario A and scenario B and derive conclusions about the underlying cognitive processes, when it is clear that there is still a lot of research to do to differentiate the myriad of potential cognitive processes using a single psychophysiological method. This being said, for many of the applications this technology could apply to, this relative variance may be enough to draw some conclusions, such as when a control room task is becoming generally more difficult for a particular operator or a driving task is becoming too challenging for a particular driver. In these applications, a deeper understanding of the specific mental processes may not be necessary to conclude that the task requires some sort of intervention to reduce overall cognitive load.

2.3.4. Other Physiological Measures

In this section, we will discuss literature relating to other psychophysiological measures of cognitive load.

The Brain

For a more direct measurement of cognitive load (given it occurs in the brain), neuroimaging techniques can be used to attain cognitive load values. Both EEG (electroencephalogram) and fMRI (functional Magnetic Resonance Imaging) have been used to assess cognitive load. The time resolution of EEG makes it possible to observe complex behaviors as they occur, EEG also can be applied outside of specialist laboratories that require large machines and teams of technicians (Gevins *et al.*, 2007). Multiple studies have found correlations between EEG and cognitive load (Klimesch, 1999; Ryu and Myung, 2005; Berka *et al.*, 2007; Trejo *et al.*, 2007; Anderson *et al.*, 2011; Brouwer *et al.*, 2012; Das *et al.*, 2014; Hogervorst, Brouwer and van Erp, 2014). Some studies show a relatively consistent relationship between cognitive load and a suppression of the lower alpha band frequency. Other studies have shown there to be a relationship between the alpha and theta bands and cognitive load (Klimesch, 1999; Das *et al.*, 2014).

Devices that measure cognitive load can often be expensive and inherently impractical for this some real world applications e.g. functional Magnetic Resonance Imaging (fMRI) machines take up the better part of two rooms, require technical support staff and restrict the person inside to limited space and movement. Given this, there is yet to be a consistent, robust technique to obtain a cognitive load metric using EEG across literature. EEGs also require a participant to be restricted to a piece of sensitive equipment that limits movement. Wireless, low-cost EEGs do exist but are in their infancy and claim to be able to produce a metric of cognitive load (Das *et al.*, 2014) (Berka *et al.*, 2007) though the positive literature is limited and not thorough in its investigation of the claims made by the developers of their proprietary "black box" metrics of cognitive load.

"Our results indicate that, although most of the measures point toward the same direction, the B-Alert metrics fails to give a clear indication of the mental workload state of the participants. The use of the B-Alert workload index alone is not precise enough to assess an operator mental workload condition with certainty. Further evaluations of this measure need to be done." (Lecoutre *et al.*, 2015)

Functional near-infrared spectroscopy (fNIRS) is a technique for measuring cortical blood flow. A small device containing infrared LEDs and infrared sensors is attached to the forehead. The infrared LEDs shine IR light through the skin, this light is then reflected off the blood and is detected by the IR sensors. The detected fluctuations in optical density result from metabolic changes in the brain (Ferrari and Quaresima, 2012). These changes have been linked with changes in cognitive load, though not definitively (Fishburn *et al.*, 2014).

The Skin

Galvanic Skin Response (GSR) is the term used to describe the minute changes in the skin conductivity that change due to some external or internal stimulus that often relates to physiological arousal. This measurement, usually taken from the fingers, measures the electrical conductivity at the skin surface, which changes due to the activity of the sweat glands that are controlled by the sympathetic nervous system and as such, respond to internal stimuli (such as stress) as well as external factors (such as heat). Though the majority of work in the field concentrates on determining mental state and emotional response from GSR, some recent studies have performed experiments to determine a link between GSR and cognitive load (Shi *et al.*, 2007; Nourbakhsh *et al.*, 2012).

Mouse Activity and Cognitive Processes

The use of basic interface devices such as mouse and keyboard has also been shown to provide some insight into the mental state of a human user. One such study allows the user to 'find Waldo' (a popular children's book activity) (Handford, 1987) by searching through an image on a screen with the mouse. Using data from the mouse activity, such as click rate, zoom and pan statistics etc. they infer cognitive traits of the individual using the mouse such as extraversion and neuroticism. The algorithm was also able to predict within a 62 – 83% accuracy how fast the user would solve the puzzle.

Mouse gesture information has been used as part of a multimodal approach to evaluate a user's emotional state (Ko, 2013; Kaklauskas, 2015). Varying degrees of accuracy were achieved, but these results usually included the fusion of other data to aid in the recognition process. A version of the Dynamic Time Warping (DTW) algorithm was used to process the gestures from the mouse and touch screen information (Schuller, Lang and Rigoll, 2002). Though the algorithm was developed for use in speech recognition applications (Itakura and Umezaki, 2005) it has been adapted for use with mouse data; a gesture being defined by a single stroke (a click and drag action) with the left mouse button being the delimiter.

2.4. Summary

In this chapter, we assessed literature pertaining to the relationship between cognitive load and physiological signals in applied scenarios. Thales Research and Technology have an industrial requirement to assess the data in control room situations automation solutions. An issue in this field is that of labelled data, which is rare and difficult to acquire given the complex nature of control rooms. There is some research into this field by way of an online labelling system for a machine learning model in the SeeCoast study (Rhode*s et al*>, 2007). This work outlines the importance of cognitive load in control room situations and the need for identifying events that may cause increased cognitive load as they can lead to decreased operator performance.

This relationship between operator performance and cognitive load can also be examined in a more formal manner by way of subjective ratings scales. Scales such as the NASA-TLX and SWAT scale have been used in a wide variety of applications to assess a participant’s cognitive state or level of mental effort during a certain task. The fundamental issue behind this approach is that the surveys must be conducted post-hoc, requiring an element of memory and recall on the part of the participant. For the purposes of identifying when events of interest occur during a certain time frame, this method wouldn’t be appropriate – if you are asking the operator in the field a set of psychometric questions, you could simply ask when the events occurred. These methods are most often used as a general measure of task difficulty for specific, discrete task timespans. Given the infancy of the objective methods, subjective methods are held as the “gold standard” by which objective measures are often benchmarked against.

The literature also clearly demonstrates the potential in the field of assessing cognitive load objectively using psychophysiological methods. There is a wide variety of signals that can be correlated with cognitive load, the results of which are varied. The field of objective cognitive load measurement is very timely especially given the more recent technological advancements that allow for the devices that measure these signals to be far more accessible and usable in practical environments. Naturally, this has lead to a recent flurry in applying these objective methods in control room contexts amongst other real-world domains given the link between cognitive load and operator performance. Of the myriad of measures for objective cognitive load, some lend themselves to be more appropriate for use in applied scenarios than others; EEG and fMRI still require somewhat cumbersome equipment that has a sensitive and extended set up time, whereas remote eye tracking has clearly demonstrated its ability to measure variance in pupil diameter and correlate this significantly with cognitive load. Pupil diameter has also been widely deployed in driving simulators which typically have significantly more physical movement than a simple computer interface, movement being a component that often requires control or at least some level of filtering to separate the component of mental effort from physical exertion component of other psychophysiological signals (Sun *et al.*, 2012).

The main body of literature reviewed here focusses on modelling the nature of psychophysical response to certain tasks within a domain, controlling some aspect of the task to artificially increase perceived task difficulty and measure the difference in responses of various signals. To determine if there is significant change in a particular measure, tasks are often separated to assess average response e.g. difference between rest and easy task, rest and difficult task, easy task and difficult task etc. The results from these experiments allow researchers to confirm whether the signals do indeed respond in these various scenarios, but there is minimal research using these signals to determine when these events occurred in a timeframe in which many different events occurred. The precise knowledge of the time in which the events occurred is either preprogrammed or manually determined. The example of the Task-Evoked Pupil Response (TEPR) requires millisecond-precise knowledge of the onset of a stimulus in order to properly determine the response; this also needs to be repeated over a series of identical trials. The trials must also refer back to some known state in which there is no stimulus or a rest period in which responses are compared against for significance.

There is much heterogeneity in the literature regarding conclusions about specific cognitive process' respective psychophysiological response, there is clearly much future work regarding linking specific responses to specific mental processes. This being said, a fair conclusion from the broad literature is that, in general, changes in cognitive load elicit changes in certain psychophysiological measures with significant correlation.

The desired output of this project is the timestamps at which events occurred during a control room setup, with the input being passively recorded psychophysiological signals. The majority of the literature uses these timestamps as inputs to determine the nature of the recorded psychophysiological signals.

In the context of the industrial work described at the beginning of this chapter, we can clearly see a gap in the literature combining the research in the fields of objective cognitive load and online data labelling systems. We discussed earlier in section 0 that the clear drawback of online labelling systems was the need for the operator to manually label events as they occur in the system, potentially increasing their cognitive load further at a time when it is likely to already be peaking. Other studies demonstrated that cognitive load could be used to infer when cognitive load was peaking during events of differing difficulty, remotely and with no input required from the operator. The gap between an online labelling system and measuring cognitive load objectively presents the gap in literature that will be the domain of this work. Bringing the objective cognitive load detection methods to control room scenarios to label data in real time, as supposed to "post-hoc" or using subjective methods. The value of such systems can be applied to many other contexts such as some that have been used as test bed in this literature review, such as in the contexts of video games, driving, piloting and other control rom scenarios.

3. Methodology

3.1. Introduction

As discussed in the literature review, the current state of literature in this field focuses heavily on characterising cognitive load using psychophysiological signals. There is little work on determining *when* events of interest occurred within a continuous session of control room activity. The closest body of research to this topic would be in examples in which the physiological output of an operator during a challenging event was compared against normal operating conditions to determine the difference in the signals. In some cases, this led to the development of a classifier that would determine the state of the operator during events of interest, but this is limited to lengthy periods of time in which task difficulty had sufficiently increased (Fa*llahi et al*., 2016). There exists a gap in the work for determining smaller scale events of interest that occurred at unknown times in a continuous task setting in which participants were required to perform multiple duties. This area is of interest to Thales Research and Technology's patterns of life department.

In this chapter, we will outline our research objectives in the context of the literature reviewed and design a research method to assess the questions drawn from the objectives. The research will entail the standard methods set out by previous studies regarding psychophysiological signals but with a focus on retrieval of events of interest, rather than on the characterisation of the events by linking them to more nuanced cognitive processes.

3.2. Research Objectives

This project aims to determine if is possible to identify events of interest from a control room scenario without relying on asking the operators to label the data from the control situation either in real time or post hoc. Such labelling procedures are a large problem in the field of machine learning as generally, the quality and accuracy of a machine learning system is largely dictated by the volume of available labelled data. Obtaining these labels is still a manual process in most use cases. In simple scenarios where there is limited complexity in the model and the labels, this can be a relatively simple process, if not time consuming. Take an example where the objective is to train a model to determine if a person is walking or running based on accelerometer data. The output is a simple concept understood by most people, running is a clearly defined notion and can be determined by most observers, they simple need to watch and note the times in which the individual wearing the accelerometer was running and this, along with an aligned signal from the accelerometer, forms that labelled data set. A model can then be developed to iterate its parameter based on how correct it is until optimum accuracy is achieved. This process becomes significantly more difficult when the environments relative complexity increases. In the case of control room scenarios, the operators are highly trained and skilled people that are experts in interpreting the highly complex data they are receiving which often requires interpretation that takes many years to be able to understand. Transferring this knowledge base to a labelled data set becomes more complex as the pool of people available to label dataset of this nature is very small and therefore difficult and expensive to access. This problem is further compounded by the 'unnatural' way in which this data requires labelling. Operators are typically trained to observe data and make decisions in real time that directly affect the data they are observing. For example, if an air traffic controller spots a potential collision, their actions to mitigate the crash will affect how other aircraft move and operate in the space, potentially creating scenarios that require further actions to mitigate.

Given the high stakes environments these specialists often work in, it is not feasible to add extra tasks to their workload as they are performing them. This limits the potential for operators to label the data in real time. Equally, the comfort of these operators is often optimized to ensure their full mental resources can be applied to their tasks and observations without distraction. This limits the nature of intrusion of devices and observers to their live working environments.

Strategies for optimizing the environment of operators begin with human factors regarding their working practices and physical environment e.g. strict shift patterns to reduce fatigue and ergonomic design of their equipment. The fundamental objective of these measures is ultimately to optimize the operator's cognitive load. As we have found through our literature review, cognitive load is a complex mental process that is made up of different types of mental process (Woods *et al.*, 2012). There exists a relationship between operators task performance and their cognitive load (Pape, Wiegmann and Shappell, 2001). Resultantly, research has focused on measuring operator's cognitive load to establish if this attentional resource can be quantified. These methods have been robustly established using subjective measures (Hill *et al.*, 1992), most often a formalized questionnaire taken after a task that characterizes the mental effort of the operator beyond a simple one-dimensional scale of subjective difficulty that can be applied across multiple individuals accounting for their inter-individual subjective experience of the same event. These methods are still used as the gold standard of cognitive load measurement given the infancy in their objective counterparts.

Objective measures have present a potential solution to the issues of post-hoc measurement of cognitive load. Several psychophysiological measures have been developed over recent decades that aim to correlate physical signals from the body with cognitive load. Though the link between subconscious mental processes and some of these outputs has been observed for centuries (think heart rate increases and eye dilation without the presence of physical exertion or light respectively) they were only truly empirically established in the latter half of the 20th century. With a more recent flurry of work resurging in recent years due to the increasing availability of previously cumbersome and expensive sensors required to measure these signals. These sensors have now been deployed in environments such as control rooms to determine if they are capable of responding significantly to changes in task difficulty. This approach of measuring cognitive load objectively clearly has many advantages over other methods. It enables researchers and stakeholders to gather large amounts of data across their respective domains to draw high-level conclusions about the cognitive nature of the work performed by these operators and to make potential adjustments to their working environment as discussed before but armed with information that has not been either accessible before or even understood given the subconscious nature of the processes being assessed. Research into applying the conclusions drawn from the body of work in this field is very timely and consensus is still to be obtained on the precise nature of the cognitive processes being reflected by these signals. One system found in literature has attempted to integrate the new technology into a control room setting in a closed loop system. The system designed by Aricò et al measures cognitive load in real time using EEG and adjusts the workload interface to the operator in real time to increase or decrease the cognitive load (Aricò *et al.*, 2016). The process aims to optimize the available cognitive resources across a team of air traffic controllers to ensure average cognitive load is not too high or too low.

Given our original research objective: *To determine if the times at which events of interest occurred within complex control room scenarios can be retrieved without manual intervention from the operator*. We can see the developments in literature certainly demonstrate that the use of psychophysiological can be used to determine the variance in the cognitive state of an operator. The literature clearly demonstrates that there exists at least broad correlations between cognitive load and psychophysiological signals, even though the precise nature of these correlations is still to be determined. Given their established nature however, we can change our objective accordingly: *To determine if the times at which events of interest occurred within complex control room scenarios can be retrieved without manual intervention from the operator by analysing their psychophysiological signals.*

This research does not seek to further the understanding of the relationship between psychophysiological signals but to utilise their present correlations to determine when events occur in a control room setting. This will form the basis for an automatic labelling system for control room environments, with the subconscious signals from the operators providing labels regarding the time at which events that are abnormal occur.

The literature in its present state outlines significant work in the field of determining the nature of the physiological responses to certain stimuli, often in controlled cognitive tests such as the Stroop test. There has been work in more applied domains also, with these domains often designing custom environments to control stimuli presented to participants. The work focuses on characterisation of signal responses, with significance being determined using statistical tests averaged over a series of repeated trials and multiple individuals. In our work, we seek to establish the *times* at which the stimulus events occurred rather than the nature of the events. The research presented here has a potential use case in being applied in real world scenarios so the factors affecting those scenarios will be considered here.

In real world scenarios where labelling would need to occur, events can take place at times unknown to the operators, these events can also often be totally unexpected and require immediate action. As a result, averaging results over several trials would not be possible in a real world scenario. Though there may be occasions in which multiple operators are working simultaneously, it is rarely on the same exact data, they will have separate duties. Given the times at which these events occur, their potential recurrence and repeatability in unknown, multiple trials over identical events would not be a scenario that would likely occur in a real world scenario. Establishing significance over a multitude of trials would empirically demonstrate the nature of the response to a specific task, but not whether that response could be reliably repeated if the event occurred during an extended timeframe in a control room context, it also limits the search to events that have been previously understood by the operator. As a result, participants in this research will be treated as individuals and the focus will be on normalising their signals for examination as a whole.

The nature of the stimulus scenarios found in literature also followed an established design. The design of the environments was to establish a normal working task for an operator, then adjust the task in some way as to increase the difficulty to simulate higher cognitive load. The hypothesis of these works is that increased difficulty in tasks will yield an increase in cognitive load that will be observed in the psychophysiological signals. These hypotheses can be tested by verifying the participants cognitive load using subjective measures for comparison, but often rely on the assumption that their adjustments to the normal tasks to increase difficulty are indeed increasing cognitive load. This assumption is undermined by the subjective nature of the tasks and the relative skills of the operator. If an individual is more skilled at a task, their response will not necessarily reflect the intended difficulty rise response, as was demonstrated by Bernhardt et al when studying engagement metrics across air traffic controllers with difference skill and experience levels (Bernhard*t et al*>, 2019). Without a precise and quantifiable understanding of an individual’s skill level at any given task, this assumption is difficult to research without. It also impacts the overall methodology of averaging different individuals together as higher and lower performers will skew the data assuming the skill levels are not normally distributed, which is difficult to determine empirically.

We therefore present our hypothesis:

An event of interest that differs from the *normal* procedures of a control room operator will induce a change in cognitive load that can be used to label the time at which the event occurred.

3.3. Research Design

To test our hypothesis we design an experiment and analysis procedure. We will outline this design in this section.

Based off work from literature, we will design a control room simulator to act as our stimulus in which events of interest and normal operating procedures will occur. We will design events of interest to insert into the simulator at times unknown to the operators. The operators will run the system whilst their psychophysiological signals are being measured. These signals will then be analyzed to determine if the events of interest can be determined using only the psychophysiological data.

The success criteria of this experiment is determined as a function of precision and recall. In the labelling process, accuracy is the most important factor. A label that is inaccurate can skew the results of a model. For our system, accuracy is measured in both recall and precision – with recall being defined in equation 1.

$$Recall = \frac{Total\ True\ Positives}{Total\ True\ Positives + Total\ False\ Negatives} \qquad \text{Eq. 1}$$

Recall is the proportion of total events that are identified. This measure reflects the system's performance in terms of number of events classified. A system with a low recall is one that did not identify many of the events that actually occurred. The other aspect of accuracy measured is precision, formally defined in equation 2.

$$Precision = \frac{Total\ True\ Positives}{Total\ True\ Positives + Total\ False\ Positives} \qquad \text{Eq. 2}$$

Precision represents the proportion of events classified correctly from the total number identified. A system with low precision may have identified many events, but not correctly.

Both of these measures reflect aspects of accuracy that are important yet different. A system with high recall and low precision generates too many false positives which is not optimal for labelling. Equally, a system with low recall and high precision simply does not generate enough labels. There is a balance between making lots of predictions and being often wrong and making very few predictions but being correct. This balance is most often represented as the F1-score; a weighted average of the recall and precision results; formally defined in equation 3.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad \text{Eq. 3}$$

The F1 score will be used to determine the accuracy of the labelling system devised for this research. The system design of the experiment is shown in Figure 3.1.

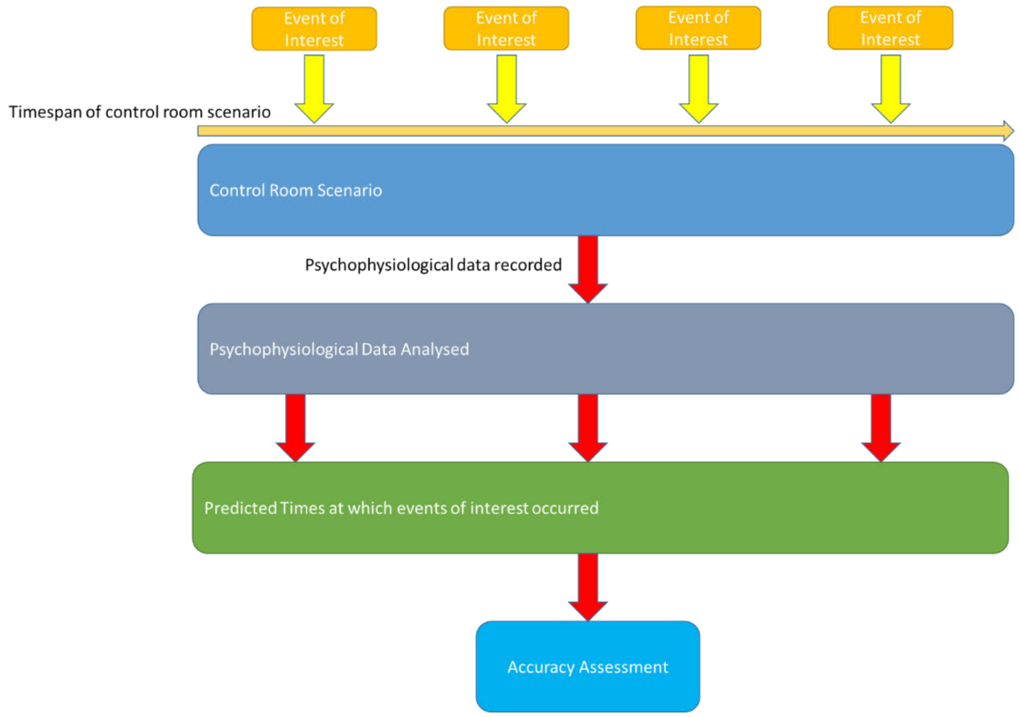

Event of Interest
Event of Interest
Event of Interest
Event of Interest
Timespan of control room scenario
Control Room Scenario
Psychophysiological data recorded
Psychophysiological Data Analysed
Predicted Times at which events of interest occurred
Accuracy Assessment

Figure 3.1. Experiment System Design Diagram.

Figure 3.1. shows the events of interest occurring at certain points within the timespan of the control room scenario. The psychophysiological data is captured from the participant as they are using the simulator. This data is then analysed to determine the cognitive load of the participant and the times at which there is significant change are recorded. These times are then compared to the actual times the events occurred, generating our accuracy. This accuracy will be used to assess our hypothesis as to whether it is possible to consistently identify the timestamps of events of interest from control room scenarios using only psychophysiological data.

3.4. Simulator Design

In this section we will discuss the design of the simulation space. Given the practical restrictions of this project, a real world control scenario was not used. An appropriate analogue was designed based on the types of simulators used in the literature to assess psychophysiological signals.

A common theme amongst simulators is air traffic control-type games. These tasks involve the tracking of moving objects on a screen and ensuring the objects do not collide and are properly managed and controlled until they leave the "airspace". Studies such as (Truschzinsk*i et al*, 2018), see figure 3.2, used such a system to determine that air traffic control tasks of differing difficulty do impact cognitive load using both subjective and objective measures.

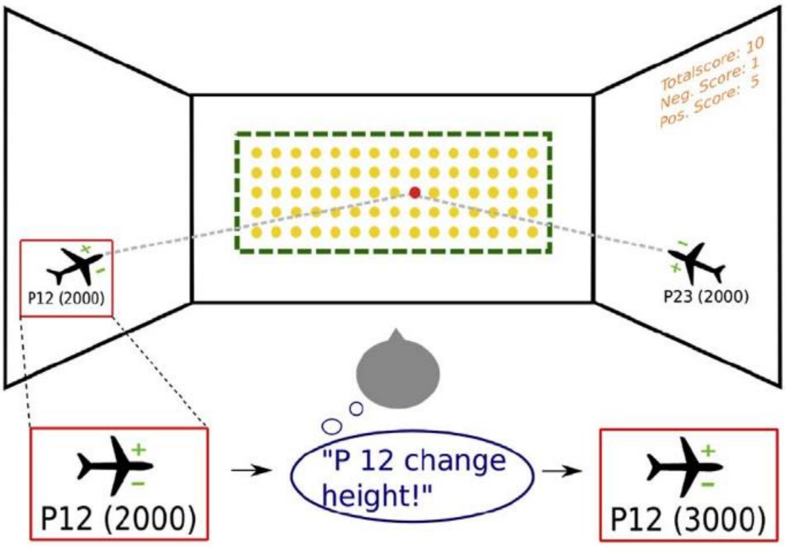


Figure 3.2. Schematic representation of the air traffic control task employed by Truschzinski (Truschzinsk*i et al*, 2018).

The difficulty of the task is increased by also increasing the frequency at which the aircraft appear at the edges of the screen. The tasks are run separately, so one trial would be high frequency, then a separate trial would be at a lower frequency, the labels for each were “high difficulty” and “low difficulty” respectively with no further nuance during the period of time the task runs for (4mins per task).

A similar tracking style task is employed by Marinescu as an object tracking task (see figure 3.3).

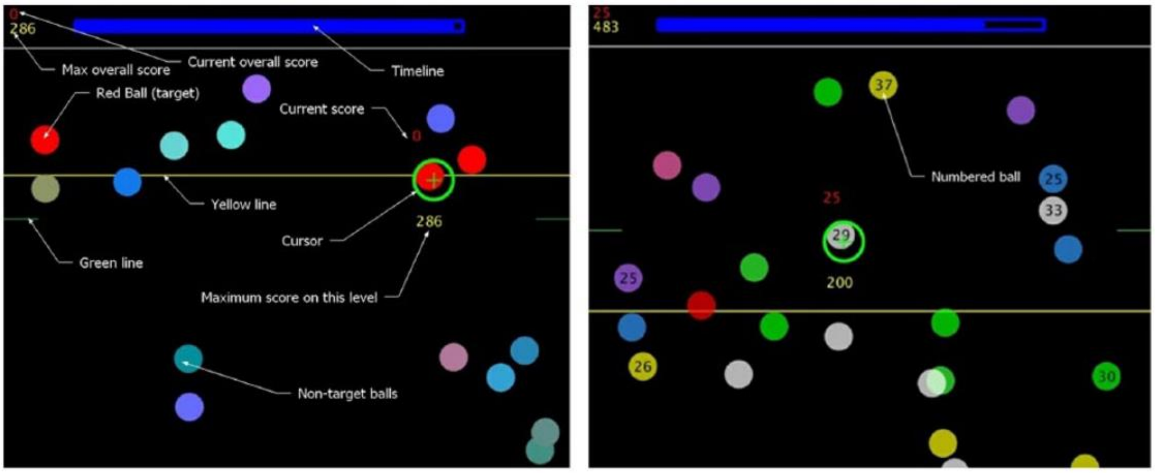


Figure 3.1. Object tracking task by Marinescu (Marine*scu et al*>, 2018).

The objective of the task is to handle the correct targets as they move down the screen, again with the difficulty being adjusted by increasing the number of targets appearing at the edge of the screen (Marine*scu et al*., 2018).

These systems are simple for participant operators to learn and use, but also require the operators to engage in elements of vigilance, timing, tracking and scheduling. A game style environment is also used elsewhere in literature as game metrics by their very nature allow for easy adjustment of task difficulty and performance measurement (Mallick *et al.*, 2017; Taub *et al.*, 2017). Fundamentally, the task should be complex as to have multiple, simultaneous real-time requirements on the operator in a continuous fashion.

We present a simulation environment loosely based on a video game called *Flight Control* which simulates a subset of the tasks performed by a traffic management control room. The operator is presented with a birds eye view of an airport with three landing zones for different types of aircraft (large fixed wing, small fixed wing and helicopters) represented by different coloured circles. The circles appear at the edge of the screen at random positions; the operator must click and drag a path for the circle to follow at a set speed, directing it to the appropriate landing zone (see figure 3.4).

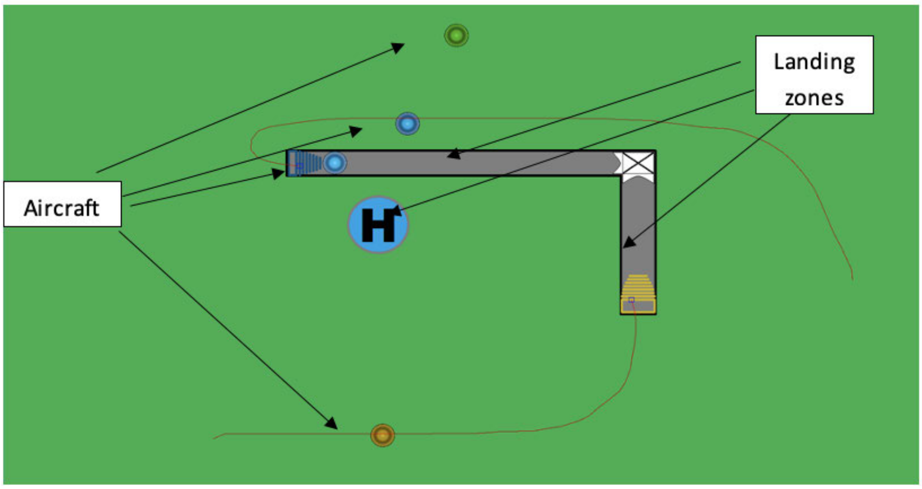


Figure 3.4. Snapshot of the proposed simulator showing aircraft represented as circles and the landing zones.

As this path is created by the operator, it is drawn on the screen as a red line. The objective of the task is to safely land aircraft at their respective landing zone. The trajectories and starting positions of the aircraft are random so a key element of the task is to prevent collisions. The operator does have the ability to edit and change flight paths. Once an aircraft reaches the appropriate landing zone, it fades away or "lands", if aircraft collide, a red circle appears around the crash zone and both aircraft disappear in a "crash".

The game is simple to operate, only requiring the participant to have basic mouse control skills. The basic functions can be mastered in less than 5 minutes.

To use the simulator as a stimulus scenario for our research purposes, we must establish what is *normal* operating procedure and what is *abnormal.* The game was tested to determine what frequency of appearing aircraft was required to ensure the task required constant vigilance and attentional load, but was possible to do without crashing any aircraft. Normal operating procedures were determined by trialling the simulator on 5 participants recruited from the Technologies for the Sustainable Built Environment for 2 minutes a trial at increasing appearing aircraft frequencies. It was determined that a rate of 10 aircraft per minute was the level at which all participants could manage the system without any crashes and reported the task was consistently achievable. It should be noted that of the 5 participants (4 male, 1 female, age 21 – 37, mean 31) one claimed to have significant gaming experience, 4 claimed to have minimal to no experience. The objective of these trials was simply to determine if the game could be learned quickly and what frequency of aircraft appearing was the optimal for defining *normal* simulator experience. A normal operating procedure for the participants would be this task maintained consistently any given length of time.

The simulation space also requires features that allow the task difficulty to be adjusted. To do this, the frequency of appearing aircraft is increased, similar to the simulations from literature.

We now define elements of the simulator that can be adjusted to create *abnormal* operating conditions by varying certain parameters of the task that will not be expected by the participant but will not pose such a change to the *normal* operating procedures a fundamental game mechanics such that they cannot handle the tasks. These changing elements will constitute the *events of interest* that we hypothesise will require a significant increase in cognitive load that will be detectable from the base cognitive load such that the times at which these events occurred will be recoverable using only psychophysiological data. We see in literature that events such as these are clearly separated into different trials and participants operate through a single trial containing a single event of interest, with the whole trial receiving the label of that event. By designing the experiment in this way, it is possible to determine if a significant difference in physiological output exists on average across trials of different difficulty, but by definition requires the start and end times of the event to be known such that the average can be calculated. It being established that increases in cognitive load to yield changes in psychophysiological output, this experiment puts multiple events of interest into a single trial, at different times, such that the average cognitive load measurement cannot be determined during an event period. As in the real world, the time at which the event occurred would not be known and is indeed the desired output of the system proposed here.

We therefore design 4 events of interest. The events differ in the variable of the fundamental game mechanic that is being altered; the frequency of aircraft appearing, the speed of an aircraft, cloud cover blocking the view of the operator and an aircraft that doesn't not respond to attempt to control it by the operator. In our simulation, these *abnormal events of interest* would be difficult to determine from the raw simulator data without reviewing some playback of the simulator trial, the discovery of which would therefore require the operator to label manually – replicating the lengthy task that forms the background problem of this research. For example, to determine when an event happened, you would need to ask the operator, who would not be able to precisely recall the exact time of the event throughout the course of the trial and would require them to watch the replayed footage of themselves to determine. The events of interest are split into two typographies: long-form and short form. The long form events lasting significantly longer periods of time than the short-form.

3.4.1. Abnormal Event 1 – Traffic Surge (long form)

This abnormal event represents the kind found in literature: a sharp increase in the frequency of aircraft appearing on the screen. This spreads the mental resources over more targets and increase the rate at which they must decide on paths and prevent crashes. The operator must respond by increasing the rate at which they create paths and the shapes the paths take to maintain order in the screen (see figure 3.5). This task will last for 75 seconds, though it may take longer for the operator to land the increased number of aircraft that appeared during that time.

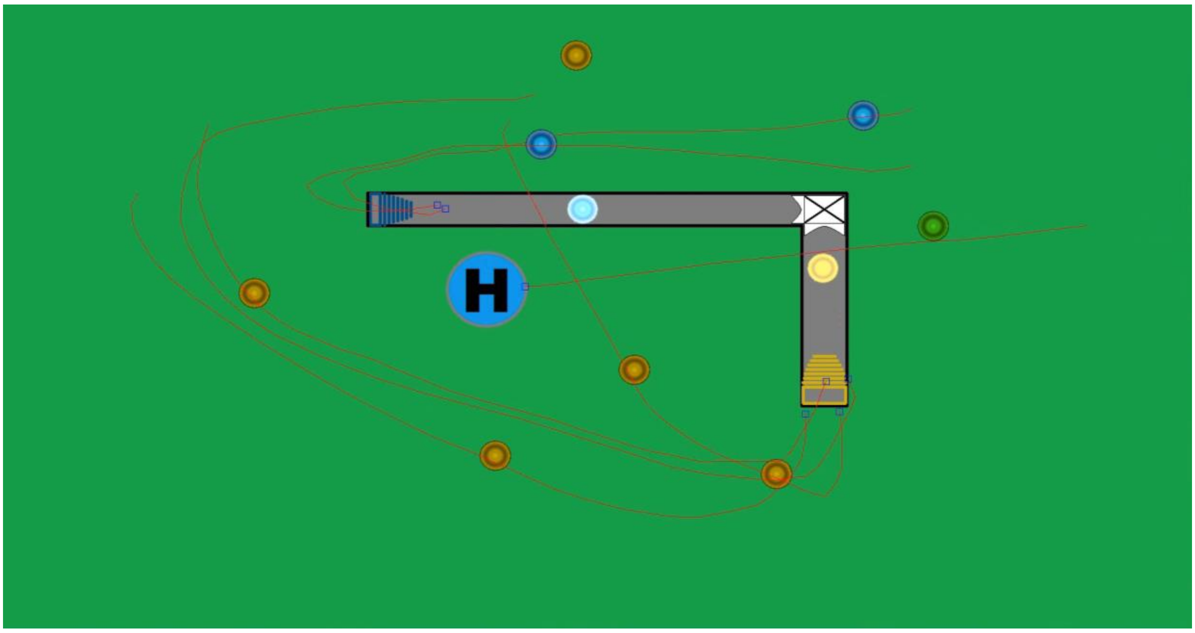


Figure 3.5. Example screenshot of traffic surge event showing increased number of aircraft.

3.4.2. Abnormal Event 2 – Cloud Cover (long form)

The second long form event involves a cloud drifting across the screen, obscuring the view of the operator (the aircraft will travel under the cloud out of sight, but still maintain their trajectories). The operator must respond by ensuring there are no colliding paths under cloud or by steering the path around the cloud as it moves across the screen. At no other point in the trial will clouds appear so the operator will not have any preparation but must react appropriately and in a timely fashion to maintain order in the airspace.

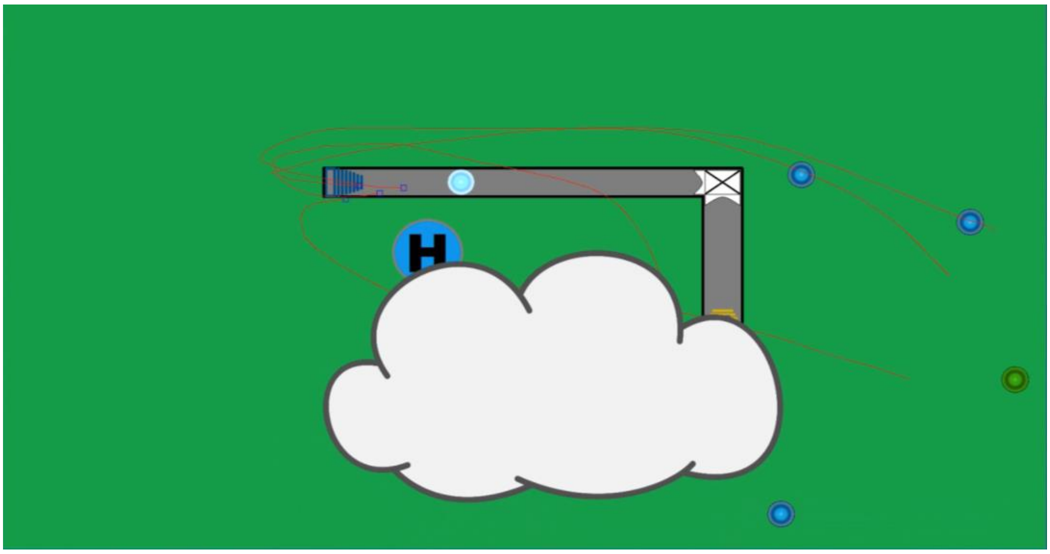


Figure 3.6. Example Screenshot of cloud cover event.

Figure 3.6 shows a screenshot of the cloud cover event, demonstrating the loss of visibility in the simulator. The cloud takes 30 seconds to cross the screen.

3.4.3. Abnormal Event 3 – Speeding Aircraft (short form)

For this event, one aircraft will appear as normal, but travelling at 3 times the speed of all the other aircraft (which all travel at a constant speed). The operator must respond by "catching" the aircraft with mouse and dragging a new path for it, at which point the aircraft will slow to normal speed. Catching this aircraft requires faster mouse movements and response from the operator, which diverts attention away from the other aircraft on screen. The operator needs to react in time to prevent the speeding aircraft crashing (see figure 3.7).

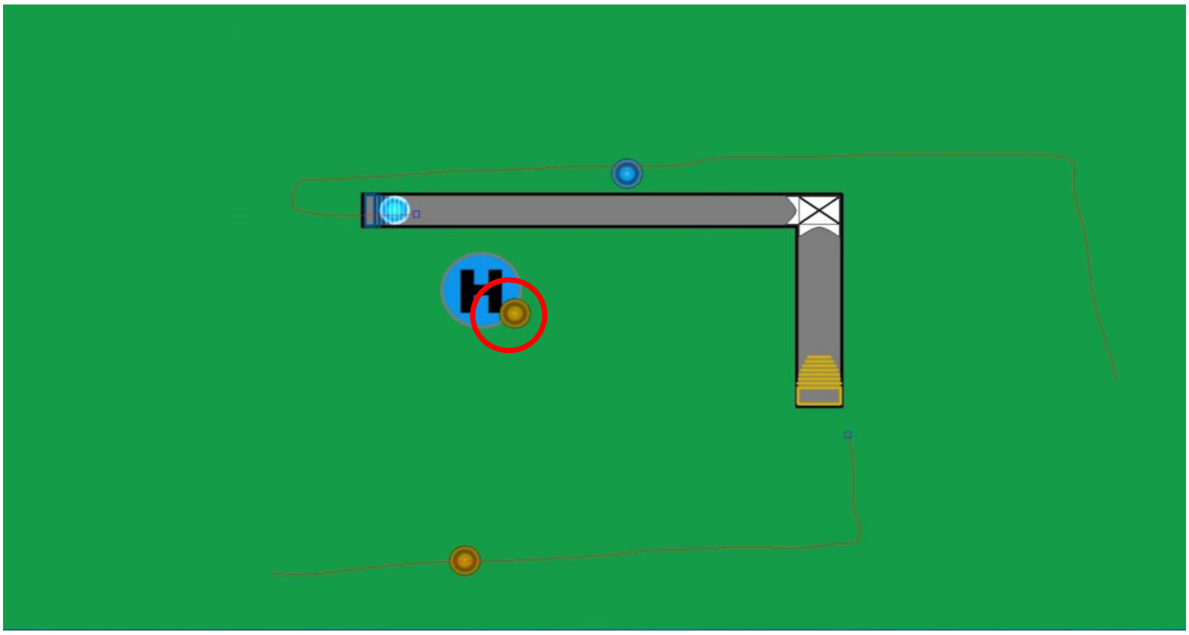


Figure 3.7. Example of Speeding Aircraft (Circled in red).

This event will take as long as it takes the operator to catch the aircraft; theoretically, this could be achieved in 2 -3 seconds, making it the shortest event of the four abnormal events.

3.4.4. Abnormal Event 4 – Unresponsive Aircraft (Short form)

The abnormal aircraft event tricks the operator by not allowing a path to be created from the aircraft. It will simply continue on its initial trajectory until it has left the screen. The operator will need to recognise that this aircraft cannot be controlled and adapt by changing the paths of relevant nearby aircraft to compensate for the aircraft that cannot be controlled. This task will take as long as it takes for the unresponsive aircraft to leave the screen (approximately 10-13 seconds depending on the trajectory).

3.4.5. Simulator Design Summary

In this section we have outlined the design for our simulation environment. Similar in design to those in literature, it allows for adjustment of the difficulty of the task. We have also designed four events of interest, of differing types and lengths. The simulator has a simple interface that can be picked up quickly, but still require constant vigilance to complete the task successfully. The events of interest represent changes in the way the operator must handle the task, we hypothesise that the tasks will yield a change in cognitive load such that the times at which they occurred can be determined from the psychophysiological data being recorded from the participants as they use the simulator.

The events of interest are not designed to alter specific elements of cognition to be demonstrated though characterisation of psychophysiological signals – that is outside the scope of this work. They are designed to represent a change in task that challenges the operator in such a way that their relative cognitive load will change. The increase in task difficulty represents the standard tasks difficulty change found in literature, which has been demonstrated as yielding a significant change in cognitive load. Here we present tasks of significantly shorter length to determine if the change in cognitive load they yield is enough to be detected from a baseline – as far as we are aware this is the first piece of research to investigate such short term events in relation to psychophysiological signals, which usually are determined over much longer periods in discrete repeated trials.

3.5. Signal Selection

As the operators use the simulator space designed above, their psychophysiological signals will be measured for analysis to determine when the events of interest occurred without manually consulting the operators either during the task or post-hoc. In this section, we outline the signals chosen to measure and the equipment used to measure them.

3.5.1. Practical Signal Measurement

This project aims to determine if an automatic labelling system would be possible given the present state of sensors and understanding of psychophysiological signals in a control room scenario. As such, for this work, the primary factor when selecting the appropriate signals to measure from the operators was practicality in a real world scenario. This means the equipment used to measure the signal must pose minimal interference to the comfort of the operator performing their duties.

As discussed in the literature review, there are some methods of measuring cognitive load objectively that, though demonstrate success in measuring cognitive load, are not appropriate for deployment in a control room. As such, fMRI was eliminated as a signal as this measure requires room-sized equipment that requires the participant to remain as still as possible to get clear images of the brain.

Pupil diameter readings are most accurate when measured using a chin-mount and camera set up, this would be impractical for a control room even though the operator would still be able to move slightly more than inside an fMRI. Given this, low-cost eye trackers have been used in significant amount of recent literature as devices for measuring pupil diameter as a measure of cognitive load with significant correlation with subjective results, confirming their validity as tools for measurement in applied scenarios (Coyne and Sibley, 2016; Vlas*tos et al*>, 2020). Of the two main types of device in this low cost bracket, head mounted and remote eye trackers, the remote eye tracker poses the most attractive in terms of practicality. As it is a completely non-contact device it can’t interfere with the operator on a physical level, in which they might accidentally knock or adjust the head mounted device for comfort, invalidating some results. There has also been some work to suggest that head mounted eye trackers are cumbersome and can distract users to the point of negatively affecting results (Marshall, 2002). Remote eye trackers allow for a relatively wide range of motion and have been successfully deployed in driving simulator environments in which physical movement constitutes a fundamental part of the activity, the effects of which have not negatively affected the results of studies that use them.

EEG devices have also demonstrated their ability to accurately assess cognitive load in control rooms and are also used in other domains to assess emotion states amongst other cognitive processes. Though EEG devices have shown potential in examples such as those posed by Arico et al, in which operators performed air traffic control duties whilst attached to EEG caps. Practically speaking however, EEG caps still require significant setup as the individual electrodes require soaking and calibration, which is a lengthy process. The EEG is also a very sensitive device that requires the user to be relatively still not to mention tethered to a computer. Though wireless EEG devices are now available, their value as devices capable of reliably measuring cognitive load still remains under question (Lecoutr*e et al*., 2015).

HRV has clearly demonstrating prominence in literature as a measure of cognitive processes. Though there is still research to do on the exact processes the changes in HRV are measuring, there is a consistent body of evidence to suggest that changes in cognitive load yield changes in HRV. HRV derived from ECG can now be achieved using low-cost wireless devices. These devices require minimal set up and do not restrict he movement of the individual as they perform their duties.

3.5.2. Chosen Psychophysiological Signals

For this research, two psychophysiological measures were chosen to assess cognitive load. Two methods were chosen so that they can be compared for accuracy in determining events of interest.

Pupil Diameter

Pupil diameter was chosen because of its inherent practical benefits when measured using a remote eye tracker and seems a natural choice for a control room context given its low intrusion. As the stimulus scenario is entirely based on a computer screen, a screen-mounted eye tracker was a good fit for the project.

For this project, the Tobii X2-60 eye tracker was chosen. Tobii devices are ubiquitous in literature and come with a separate compute unit that performs some initial calculations automatically such as blink recognition. The Tobii devices also use the ellipse fitting method of measuring pupil diameter which has been shown to be more resistant to the pupil foreshortening error (Klingner, Kumar and Hanrahan, 2008) discussed in the literature review. The specifications of the device are listed in Table 3.1.

Table 3.1. Specifications of the Tobii X2-60 eye tracker.

Sample rate	60 Hz (±1 Hz)
Accuracy	0.4°
Precision	0.34°
Mount type	On screen, stand
Screen size	Up to 25" when mounted (16:9)
Operating distance	40 – 90 cm
Head movement	50 x 36 cm

To assess if the pupil diameter readings responded to the simulation space, a small trial was run in which the number of aircraft on screen was changed dynamically to see if the pupil diameter cognitive load response correlated.

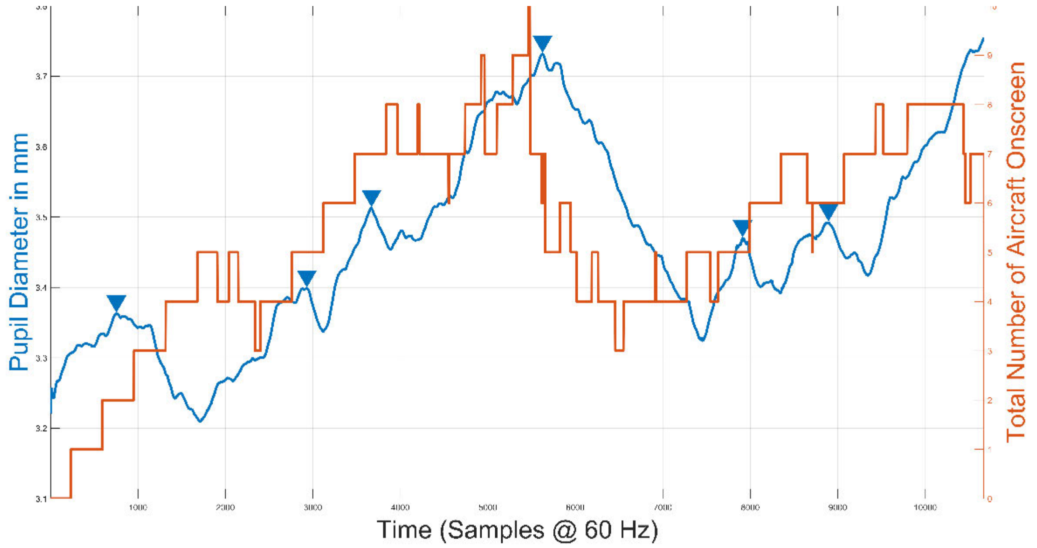


Figure 3.8. Initial trial results to map pupil diameter an cognitive load throughout the simulator experience.

We can see from figure 3.8 a clear correlation between number of aircraft in the simulator and the pupil diameter response, confirming that the change in difficulty in the simulator does effect the pupil diameter response and that this effect can be measured on our chosen equipment. Using this measure, we derive a sub research question, Research Question 1: *Can pupil dilation data be used to identify the times at which events occur in an operator scenario?* We will assess this question through accuracy of our event detection analysis in a later chapter.

HRV through ECG

The second measure chosen for this work was the HRV derived from the electroencephalogram. HRV has been demonstrated to determine cognitive load in operators throughout literature (Tattersall and Hockey, 2006; Haapalainen *et al*>, 2010; Sun>*et al*>, 2012). HRV has also been able to differentiate between different tasks and tasks of differing difficulty in a number of applications, though the measure has received criticism for not being consistent in its results of determining which form of cognitive load it is measuring. It was decided that as the nature of the cognitive process being determined was not the outcome measured here but simply the relative change in HRV, it would be a suitable measure given the large body of research supporting its use and the availability of unobtrusive sensors.

Some low cost devices claim to be able to derive HRV from a simple, single lead, wireless device, but it was chosen to use a five lead ECG monitor for its accuracy. The ECG monitor used in this project is the Shimmer Sensing wireless ECG monitor (see figure 3.9). A five-lead ECG device, the Shimmer communicates over Bluetooth and has a sampling rate of up to 1025hz.

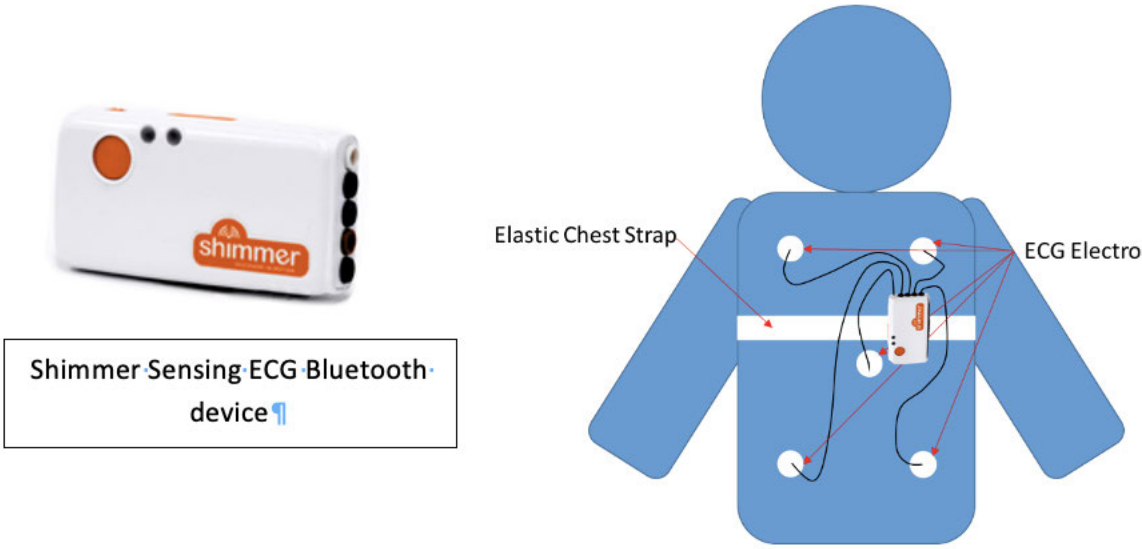


Figure 3.9. The Shimmer Sensing ECG device and typical set up.

The small size and portable nature of the device makes it unobtrusive and does not interfere with an operator’s task making it a good choice for deployment in this research. For this measure, we can pose research question 2:*Can HRV data be used to identify the times at which events occur in an operator scenario?* We will assess this question through accuracy of our event detection analysis in a later chapter.

Other Measures

Other measures considered for use were that of galvanic skin response and fNIRS. Ideally, all potential signals would be measured in a simple and unobtrusive way but due to limitations in budget and given that a GSR device may interfere with the use of a mouse. The two measures above were chosen. This study could be repeated with other measures as potential future work to assess their suitability for assessing our research objective.

3.5.3. Mouse Data

As the simulator environment uses a computer mouse as a primary input device, it was decided to capture positional data from the mouse to analyse alongside the psychophysiological methods. The data captured from the mouse was the positional coordinates and the click state of the mouse similar to studies surveyed in the literature. There is some evidence to suggest that mouse movement data can be used to infer the cognitive state of the user and given the mouse is already being used, its data will also be assessed. This gives us research question 3: *Can mouse movement data be used to identify the times at which events occur in an operator scenario?* We will assess this question through accuracy of our event detection analysis in a later chapter.

3.6. Experiment Design

To gather data from our operators using our simulation environment, we devise two experiments. These experiments will generate the data to be processed by our analysis to determine the accuracy of the system in terms of identifying when the events of interest took place in the simulation environment.

3.6.1. Physical Set Up

To gather data from a participant operator, they first had the ECG device affixed to their chest. The electrodes were placed as shown in Figure 3.10.

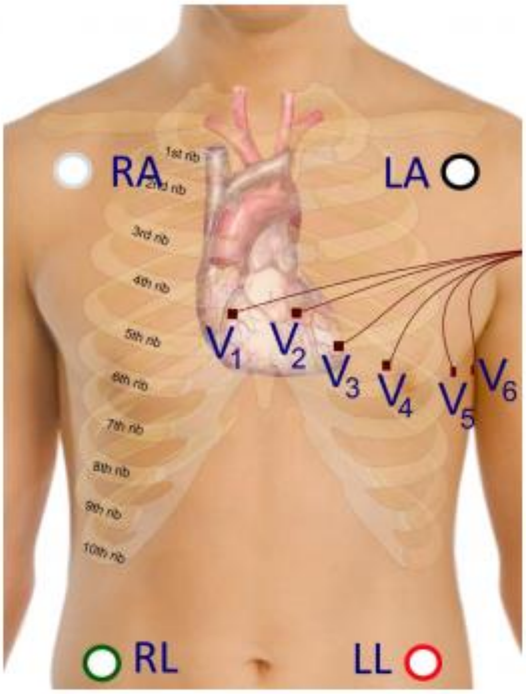


Figure 3.10. Placement of leads of shimmer ECG device.

The central lead was placed at position v2. The rough position was provided to the participant and the electrode affixed to accommodate clothing in a comfortable fashion. The ECG device was then powered on and paired to the central PC.

Once the ECG was affixed, the participant was invited to sit on a chair in front of a desk that had the central PC upon it. The participant was then asked to adjust the position of the chair to suit their comfort. The chair, desk, PC and participant were all situated in a 4m by 4m room inside the Technologies for The Sustainable Built Environment centre at the University. The room had the windows covered to prevent changes in light during the experiment, the internal lights were left on. The participant was then asked to sit in front of the central pc and adjustments were made to ensure that they could both see the screen properly and that the eye tracking device could see their pupils within the optimum viewing angle and distance. The participant was then instructed to adjust the mouse position and sensitivity to their desired comfort. The physical set up of the experiment can be seen in Figure 3.11.

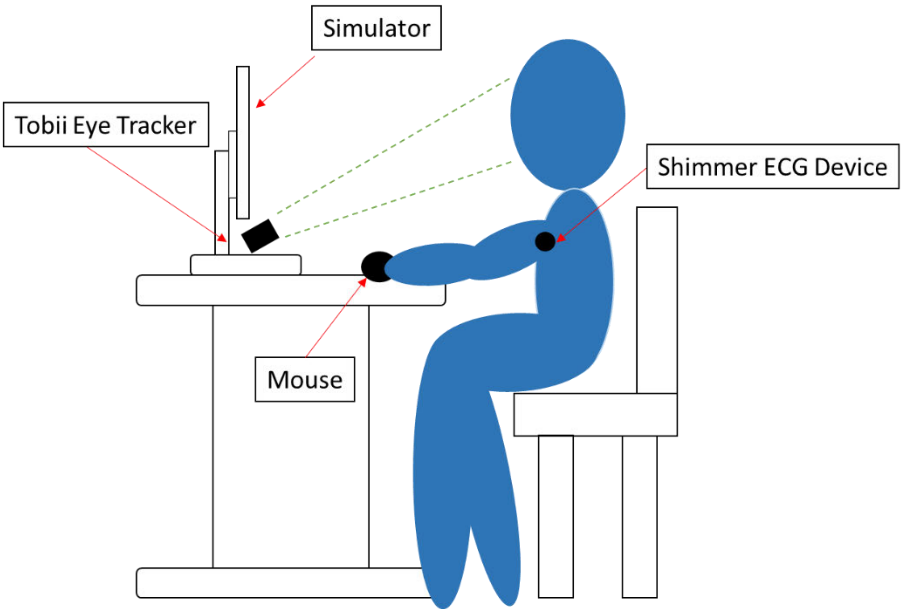


Figure 3.11. Physical set up of experiment.

3.6.2. Preparation

The participant was then shown how to use the simulator and given 30 minutes to familiarise themselves with its operation. For this process, the simulator was fixed at an incoming rate of aircraft at 10 per minute, as would become the *normal* operating procedure of the simulation. The participants were not informed about the events of interest or shown how or when they would occur. When the participant was comfortable with the controls and usage of the simulator, they were informed that they would operate the simulator for the length of the experiment. The eye tracker was calibrated using a 9-point calibration procedure. The software used (iMotions) provides a calibration score, the calibration procedure was repeated until the score was “excellent”.

3.6.3. Simulation Trials

It was decided that two runs of the simulation would be run per participant, on separate days. Trial 1 differed from trial 2, as a repeat trial may allow for a participant to remember the nature, timing and order of the events and as such be prepared for their arrival. As this would not occur in a real world scenario, the timing of the events of interest were changed between trials. The two runs on the simulator were 20 minutes and 10 minutes respectively. The 20 minute run was split into two data sets. This provides us with three 10 minute data sets; one training set and two validation sets per participant. Given that the first validation set is recorded in the same run as the training set and the second is recorded on a separate day, this allows us to compare if the training parameters determined from trial one will allow for accurate determination of event times in trial 2.

The two trials differ in both length and complexity. The second trial will contain a less regular sequence of events of interest and twice the number of events. The second trial will also be the only trial with the unresponsive aircraft event. As this event is not seen in the first trial, it will not be recognised by the participants. This will enable testing of whether the system will be able to identify events of interest that are new and different and not optimised in the training set.

Trial 1

The first trial lasted 20 minutes. The trial is separated at the ten minute mark and the two halves, through completed in one run, are treated as separate trials (see figure 3.12).

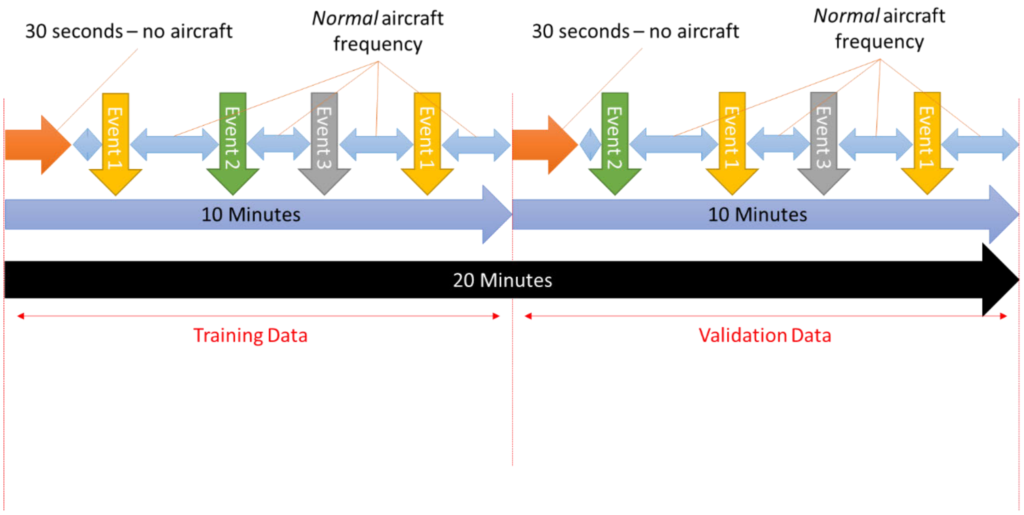


Figure3.12. Order of events and structure of Trial 1.

Figure 3.12 shows the order of events for the 20 minute first trial. The first 30 seconds of the simulator contain no aircraft. This was to allow for the participants pupils to settle from the transition from the calibration window to the simulator screen. Events then occurred with at least a 1 minute gap of *normal* aircraft frequency between them, this would allow for the cognitive load response to settle between events. The first trial run provides this separation and regularity to provide optimum conditions for cognitive load to spike during the events and settle between – this is of course not reflective of a real world scenario where events could take place closer together and with no regularity – these factors are included in the second trial. The first trial contains only the speeding aircraft event, traffic surge event and cloud cover event. The first ten minutes of the trial are then repeated, including the 30 seconds of no aircraft, the order of the events is also changed to avoid the participants predicting the nature of the trial. As the intentional difference between trial 1 and 2 is to make trial 2 a more complex run of the simulator, the non-responsive aircraft was added to the second trial only.

Trial 2

The second trial lasts 10 minutes, recorded on a separate day to the first, the second trial is made up entirely of validation data (See figure 3.13).

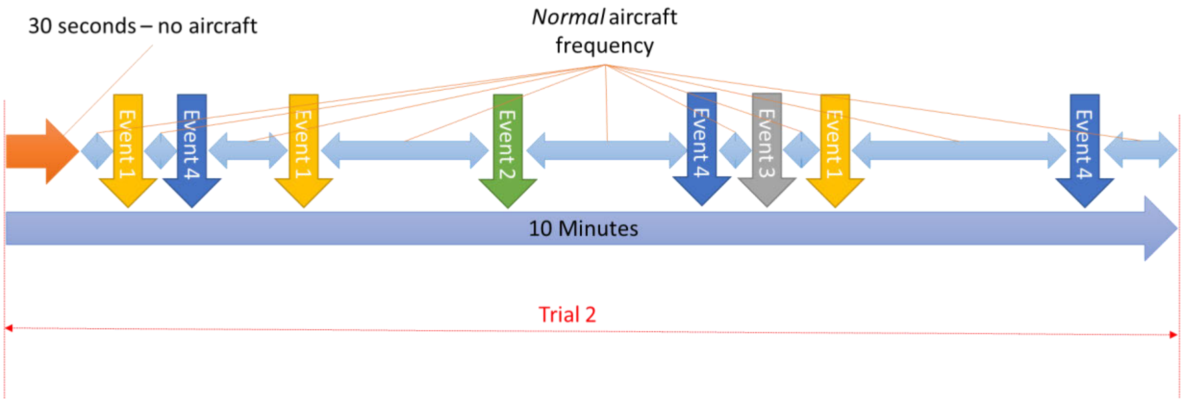


Figure 3.13. Order of events and structure of Trial 2.

The events in the second trial appear as a function of the number of aircraft landed, not at specific set times. The second trial contains eight events total, twice the frequency of the training set and first validation set. This will determine if the system is capable of detecting events if they occur with high frequency and much closer together than the regularly separated events of the first trial. The second set also contains the unresponsive aircraft event. This event will be new to the participant at this stage; this will allow us to test if a previously unknown event can be detected even if it was not seen in the training data set.

3.7. Study Participants

In total, 30 participants were recruited for this study. Their ages ranged from 19-58, mean age of 33, fifteen females and fifteen males. All participants were healthy, with normal or corrected-to-normal vision. They were recruited from the Reading, Berkshire area and were compensated £20 in Amazon vouchers for their time. Ethical approval was sought and approved by the University of Reading ethical committee.

3.7.1. Limitations of Participant Characterisation

The participants were first asked if they had significant issues with their vision that would affect their normal pupil dilation. They were also asked if they knew of any health conditions or medication they were taking that would affect their heart rhythms or pupil responses. Though none of the participants answered yes to these exclusionary questions, it is accepted that this knowledge may have not been known to them given that they are not medically trained. For future work, exclusionary factors should be well defined, enabling potential participants to answer with confidence whether they know of any exclusionary factors.

Equally, a limitation of this study is that a deeper characterisation was not made to control for other common factors that may affect psychophysiological signals such as quality of sleep, caffeine intake or menstrual cycle.

3.7.2. Data Processing

The two trials outlined in the previous section will generate 3 sets of data; the training data set (the first half of trial 1) and 2 validation sets, one with a simpler arrangement of events and one more complex with a previously unseen event. As the trials are running, the data collected from the participant is shown in Table 3.2.

Table 1.2. List of data type descriptions, units and measurement frequency taken from the participant during the experiment

Data	Unit	Frequency (Hz)
Pupil Size	mm	60
ECG	mV	1023
Mouse X position	pixels	1023
Mouse Y position	pixels	1023
Mouse Click State	"UP" or "DOWN"	1023

The data is recorded using iMotions software. The iMotions software package allows researchers to integrate multiple physiological sensors together whilst presenting participants with a stimulus. The software records all of the sensors at their highest potential frequency and co-references the signals to a master timestamp. The software also provides the calibration sequence for the eye tracker device and records the screen whilst the participant uses the simulation environment, whilst allowing the researcher to overlay statistics such as eye gaze position and pupil diameter. The co-referenced data is then exported to a .csv file to be analysed in the Matlab software package. The details on the analysis of the data are outlined in the Analysis chapter.

3.8. Summary

In this chapter, we have created a research design in order to assess our core research objective: *To determine if the times at which events of interest occurred within complex control room scenarios can be retrieved without manual intervention from the operator.* As a result of reviewing the available literature in this field, we update our research objective: *To determine if the times at which events of interest occurred within complex control room scenarios can be retrieved without manual intervention from the operator by analysing their psychophysiological signals.*

We designed a method to achieve this objective by constructing a controlled simulation stimulus environment in which the operator has to maintain a constant level of vigilance, spatial monitoring, scheduling and timing. The difficulty of the simulation space can be adjusted by increasing the frequency at which the aircraft appear on the screen. We also developed four events of interest that change the nature of the task the operator is performing in a manner intended to create a change in their cognitive load state. The times at which these events occur is the output of the system through analysis of the psychophysiological measures.

Whilst the operators are using the simulator, they will have their psychophysiological signals measured simultaneously. In this chapter, we outlined the available measures and narrowed down HRV and pupil diameter as the most practical signals that demonstrated a consistent relationship with cognitive load in literature. We also determined the devices appropriate to measure these signals. Resultantly, we propose 2 research questions:

Can pupil dilation data be used to identify the times at which events occur in an operator scenario? and

Can HRV data be used to identify the times at which events occur in an operator scenario?

As the simulation environment also contained mouse movement data and there is some evidence to suggest these movements may reflect cognitive activity in users, this data was also considered to determine if events of interest could be identified. As the mouse is the primary input method for the simulator the data can be collected incidentally with no additional requirements from the operator such as affixing the ECG device or remaining in the optimum viewing angle of the eye tracker. Resultantly, we propose our 3rd research question: *Can mouse movement data be used to identify the times at which events occur in an operator scenario?* We will discuss the analysis of the data in the next chapter.

So far in the research, we have explored our research area based on our initial objective and performed a literature review to assess the state of the art in the field. From this, we have identified a gap in the literature to create a novel contribution to the field. We have created research questions and the methods necessary to answer the questions.

Literature has demonstrated that psychophysiological signals are capable of demonstrating changes in cognitive load. The literature has also demonstrated that these measures can be deployed in applied scenarios and characterisation of these signals has demonstrated that they respond significantly to changes in task difficulty. From here, we hypothesised that events of interest in operator scenarios would change the cognitive load of the operator in a manner that can be automatically determined from analysis of psychophysiological signals. We then determined the most practical signals to measure from a human operator given the context of a control room scenario; namely that the measures should not interfere with their comfort or concentration or provide any extra tasks to the operator. We then determined and acquired the necessary and appropriate equipment to measure these signals based on evidence and precedence from literature. We designed a control room scenario using examples in literature to guide the design parameters and created the target events of interest to detect by varying the nature of the task significantly from a predetermined *normal* operating procedure. We then designed a 2-trial experiment that will gather 1 training set and 2 validation data sets. These validation sets differ from each other by providing increased complexity and being performed on a different day and including different events to test the limits of the optimisation of the analysis from the training data.

The next chapter will outline the analysis of the data to be gathered from the experiments.

4. Analysis

4.1. Introduction

In the previous chapter, we outlined the research design. This research design yields an experiment that presents events of interest to human operators during a simulated control room environment. This experiment outputs data that we will analyse to draw conclusions to answer our research questions:

Can pupil dilation data be used to identify the times at which events occur in an operator scenario?

Can HRV data be used to identify the times at which events occur in an operator scenario?

Can mouse movement data be used to identify the times at which events occur in an operator scenario?

The data outputs from the simulation trials are:

- Pupil Diameter (mm)
- ECG (mV)
- Mouse Data (X and Y coordinates, UP and DOWN mouse button)
- Video replay of screen operator used during trial.

In this chapter, we will outline the procedures for analysing the data to produce timestamps that indicate when events of interest occurred.

The timestamp outputs will be generated from the pupil data, the ECG data and the mouse data respectively. The timestamps will be determined though analysis of the time series of these data. These outputs will then be compared against the ground truth to determine the recall and precision of each of the signals.

This chapter will be structured as follows:

- We will start by demonstrating the data preparation procedures for each measure to obtain a time series for each simulation trial. For the pupil data, this will mainly consist of data clean up. For the ECG, we will determine an appropriate metric to derive for use in this research by way of comparing the latency of multiple HRV measures discussed in literature against our training data, we will then select the highest performing measure to generate a time series. For the mouse data, we will derive the speed profile of the mouse movements to generate a time series, we will also generate a second time series by presenting a novel metric derived from shape analysis of the patterns made by the mouse.
- We then present a novel metric for determining time at which the cognitive load measure has changed significantly from a background level. This measure is based on a geometric analysis of the time series, bounded by the assumption that cognitive load will sharply increase then recover based on events that are significantly different enough from normal tasks. This will create a series of peaks in the time series that will be discovered and filtered by optimising the peak finding method by calibrating the parameters on a subset of data.
- We then generate our ground truth to assess the accuracy and precision of our time stamps. This process will be a manual one, requiring us to watch the screen recordings of each trial from every participant and annotate the timestamps. The methods for this process will be discussed in this section.

The process for the data analysis procedure is visualised in Figure 4.1. showing the development of metrics and time series for each of the data streams gathered from the experiments.

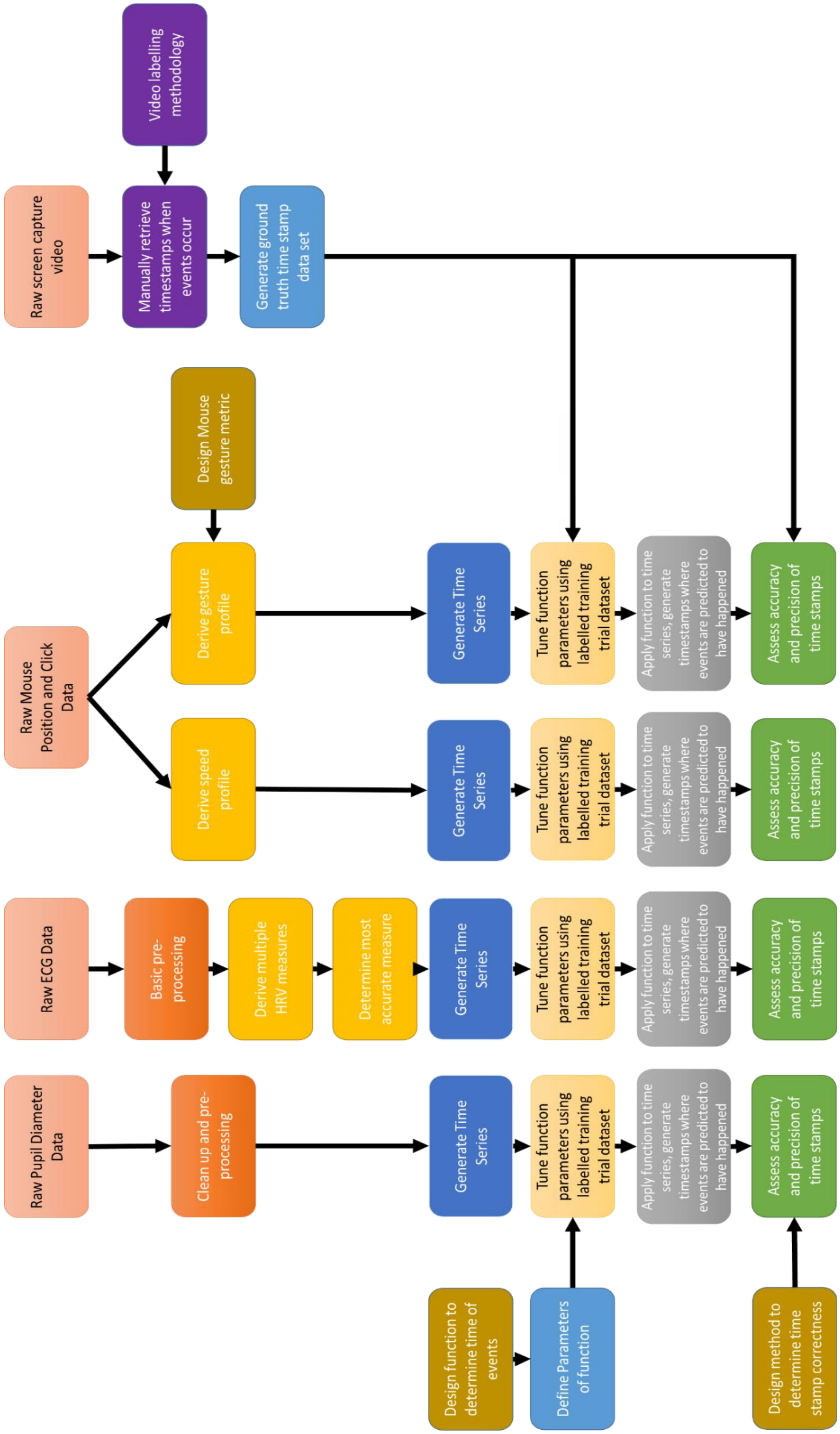

Raw screen capture video
Manually retrieve timestamps when events occur
Video labelling methodology
Generate ground truth time stamp data set
Raw Mouse Position and Click Data
Design Mouse gesture metric
Derive gesture profile
Derive speed profile
Raw ECG Data
Basic pre-processing
Derive multiple HRV measures
Determine most accurate measure
Raw Pupil Diameter Data
Clean up and pre-processing
Generate Time Series
Tune function parameters using labelled training trial dataset
Apply function to time series, generate timestamps where events are predicted to have happened
Assess accuracy and precision of time stamps
Design function to determine time of events
Define Parameters of function
Design method to determine time stamp correctness

Figure 4.1. Diagram showing overview of analysis procedure.

4.2. Pupil Data

In this section, we summarise the methods used to pre-process the pupil diameter data. The raw data from the Tobii device includes automatic blink detection, the process for which is assumed to be accurate for this study as this assumption has been held in other studies using similar Tobii devices.

4.3. Pre-processing

The data is sampled at 60hz by the Tobii X2-60 eye tracker. For each participant, we have three, 10-minute recordings. The first and second recording are both taken from the same 20-minute experiment, but split into two, consecutive, ten-minute recordings. The first of these 10-minute recordings provides our training set. The second half of the first experiment and the 10-minute recording from the second experiment form our testing sets.

4.3.1. Interpolating Blinks

The Tobii X2-60 eye tracker records occurrences of blinks as '-1' values. The device also returns '-1' values for instances when pupil diameter could not be measured; this can be from the head being turned too far from the eye tracker's working viewing angle or from the participant blocking the view of the device, for example by scratching their face. These errors were permissible as the participant was to work comfortably and naturally.

To remove these sections of data with '-1' artefacts, the data was linearly interpolated over the sections that contained '-1' readings. Linear interpolation is the most commonly used method for handling missing data or blinks in literature. An example of this is seen in Figure 4.2.

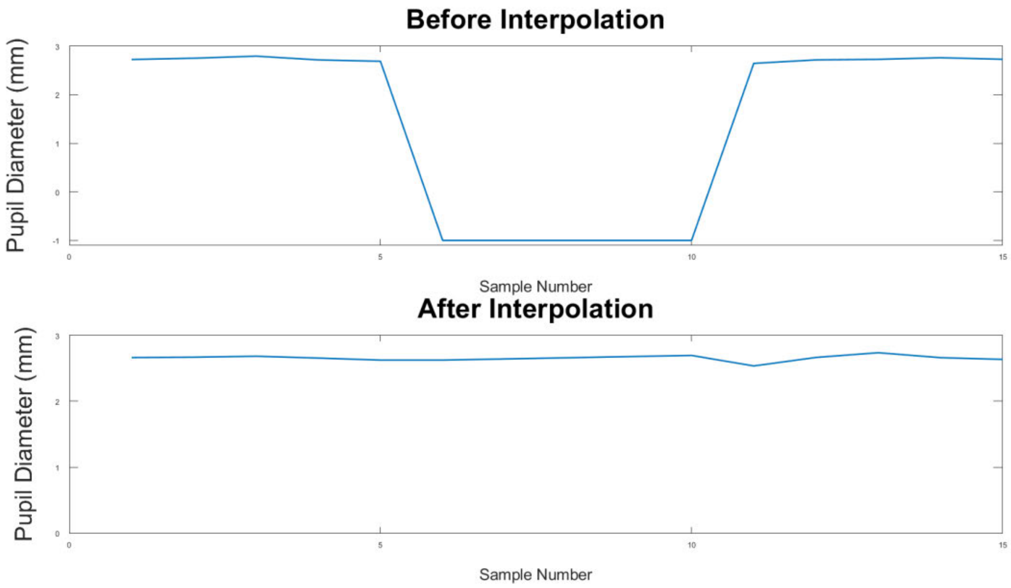
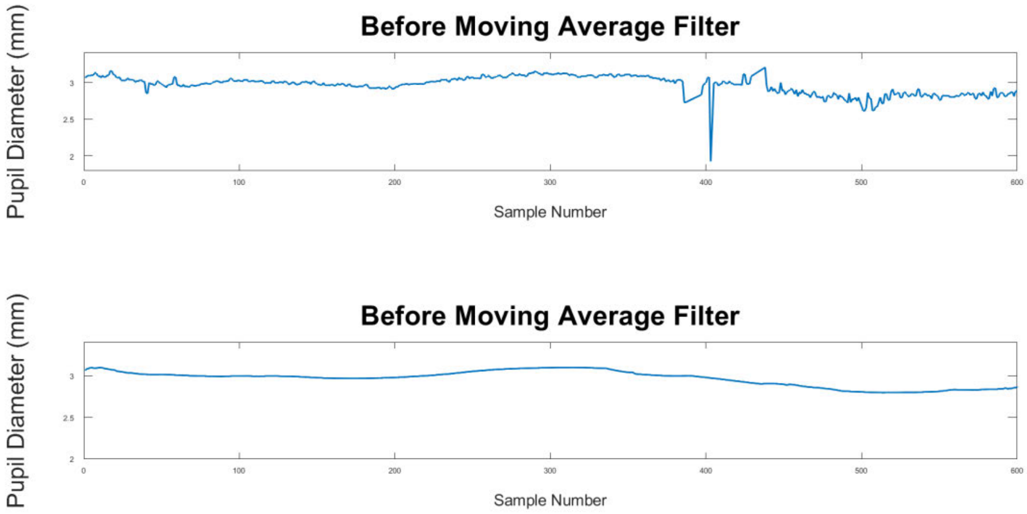


Figure 4.2. Example of pupil data errors being linearly interpolated. Top: data before interpolation. Bottom: data after linear interpolation.

This procedure was applied to all participants' data.

4.3.2. Low Pass Filter

The raw pupil diameter signal contains high-frequency noise. This noise was removed using a simple moving-average filter. Frequencies above 2hz are usually considered noise (Privitera *et al.*, 2010). In addition, some measurements that are recorded contain *physically implausible* results; when a pupil diameter will jump by an implausible factor for a single sample then return. These types of errors are common in non-contact pupil diameter measurement devices and are similarly removed using the moving average filter. An example of this can be seen in Figure 4.3.


Before Moving Average Filter

Before Moving Average Filter

Figure 4.3. Example of moving average filter being applied to pupil diameter data. TOP: raw data before filter applied. BOTTOM: data after filter applied.

4.3.3 Pupil Data Set

The above pre-processing procedure uses standard methods from literature to clean up the raw pupil diameter signal. The resulting data set is three time series; one for the calibration set and two validation sets from each of the simple and complex trials.

4.5. ECG Data

The raw ECG data collected from the experiment comes in the form of a time-series of millivolt values. In this section we will outline the method used to determine notably pronounced peaks in cognitive load from the raw data.

4.5.1. R-R Interval

As discussed in the literature review, a key component of the ECG signal is the QRS complex (see Figure 4.4).

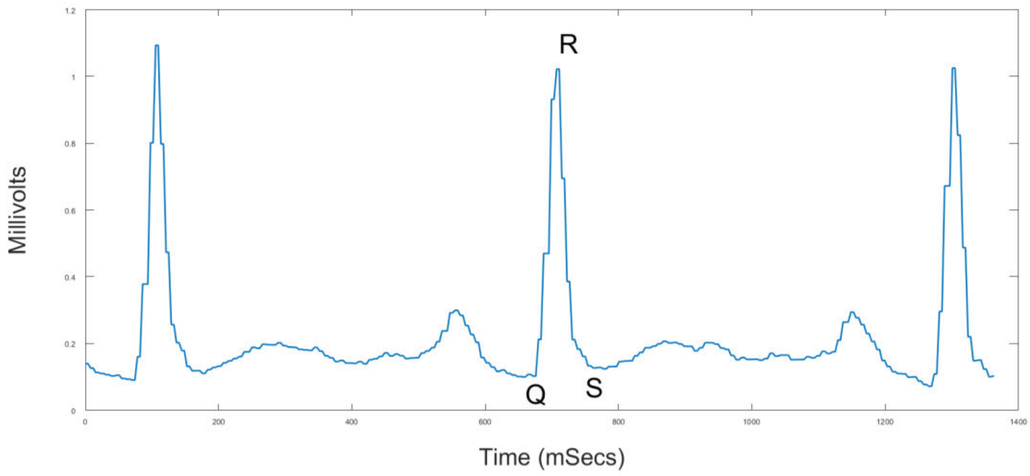


Figure 4.4. Example of QRS complex, points labelled.

The point of interest in this work is the R-R interval, as variances in the R-R interval or Heart Rate Variance, have been shown to reflect cognitive load. In order to measure the R-R interval, the location of the R-peaks must first be determined. This was achieved using Kubios HRV software (Tarvainen *et al.*, 2014). The software uses a variation of the QRS detection algorithm by Pan-Thompkins which is widely used through literature and industry (Pan and Tompkins, 2007). As HRV can be sensitive to errors in R peak location, the raw data can be viewed in the Kubios software and R-peaks that have been missed by the algorithm can be manually inserted. This was performed where necessary for all participants.

4.5.2. HRV Measures

As discussed in the literature review, multiple different metrics of HRV can be derived from the ECG signal. As there is no consensus on which measure provides the most robust indicator of cognitive load, we derive multiple measures and compare results.

We derive two measures of HRV in the time domain and 3 measures in the frequency domain. The measures derived are the RMSSD, mean RR interval, Low frequency component, mid frequency component and high frequency bands.

A significant parameter in determining HRV measures is that of the windows within which the features are derived. As the longer the length of windows for short-term features are > minutes and given our trial lengths are 10 minutes each, it was decided to use ultra-short term features. Kubios software enables time-varying feature analysis for outputting time series of HRV using different metrics. The shortest available time window is 40 seconds, given the length of our event of interest can be as short as two seconds, this window size was chosen as the longer window sizes may average out potential peak responses in the time series.

Mean RR interval

For this measure, the mean R-R value for feature windows of 40 seconds was calculated. A 40 second window was chosen as it allows for capturing of smaller variations in HRV for shorter events in our simulator that may be suppressed through averaging using longer window lengths. Figure 4.5 shows the time-series representation of this procedure for a single participant.

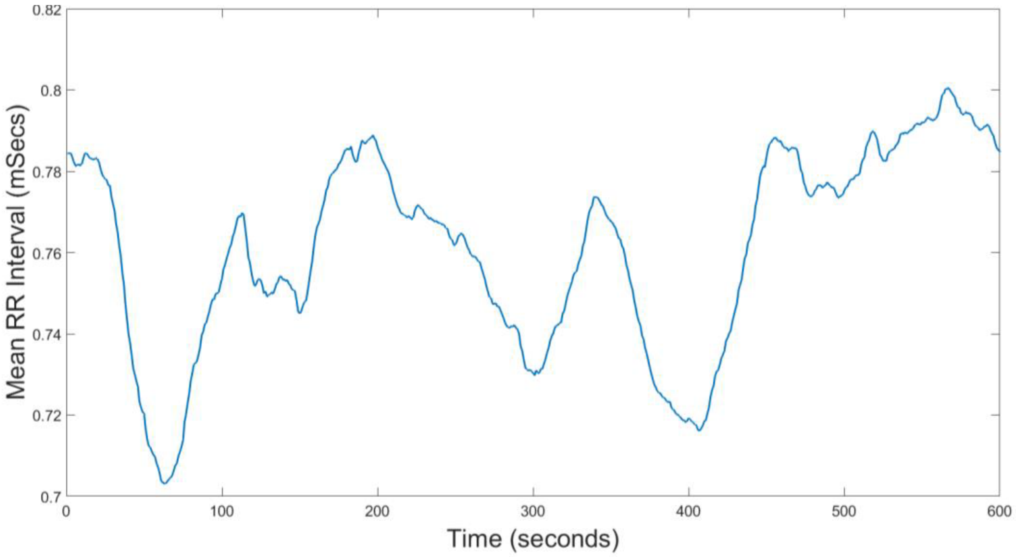


Figure 4.5. Time-Series of Mean RR intervals from single participant over 10 minutes.

Root mean square of successive differences (RMSSD)

The RMSSD is a common HRV metric, calculated using equation 4 over 40 second shifting windows:

$$RMSSD = \sqrt{\frac{1}{N-1}(\sum_{i=1}^{N-1}((|RR|)_{i-1} - (|RR|)_i)^2)} \qquad \text{Eq. 4}$$

In which *N* represents the number of R-R intervals in the window and *R-R* is the length of the R-R interval. The time-series representation of this is shown in Figure 4.6.

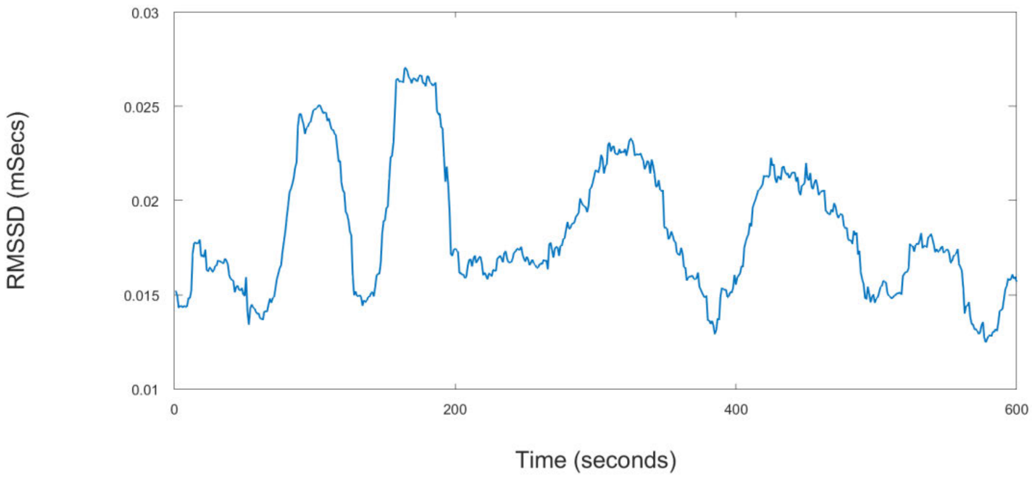


Figure 4.6. Example Time-Series of RMSSD for Single Participant over 10 minutes.

Frequency analysis

For our 40-second windows, the frequency power was determined for the three frequency bands:

1. Low-frequency (LF) 0.02 to 0.06 Hz;
2. Mid-frequency (MF) from 0.07 to 0.14 Hz; and
3. High-frequency (HF) from 0.15 to 0.50 Hz.

An example of this decomposition can be seen Figure 4.7.

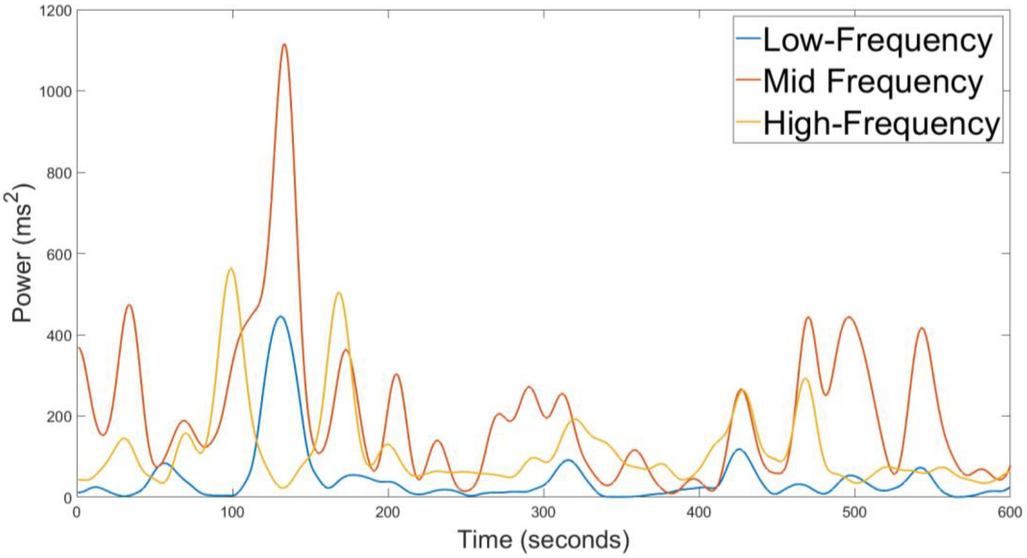


Figure 4.7. Time-Series Representation of Frequency Power at 3 different bands (LF, MF and HF) over a time of 10 minutes.

It should be noted that 40-second windows is exceptionally short for these frequency measures and may affect the validity of their use.

4.5.3. Selection of HRV Measures

From the onset of a stimulus, there is a latency as the input is processed by the participant and the nervous system outputs the signals that are measured by our sensors as a response. This latency varies depending on the contextual circumstances, inter-individual differences in physiology and the psychophysiological signal in question. The precise latency of each signal is difficult to determine in advance given these factors. Resultantly, all the signals gathered from our trials will be assessed for their average latency over our training set. Ideally, the less latency the closer the identified change in psychophysiological signal is to the true time stamp of the onset of the event of interest.

We will perform this analysis of latency between the signals over each measure of HRV and select the signal with the shortest latency. This analysis is performed later in this chapter.

4.5.4. HRV Data Set

The output of the pre-processing procedures outlined in this section will yield five time series for HRV: Mean RR, RMSSD, LF, MF and HF. The measure that demonstrates the smallest latency will be selected. We will have three sets of HRV data, one calibration set and two validation sets from our experiment trials.

4.6. Mouse Data

As discussed in the literature review, there has been some work on deriving the mental state of a user based on their mouse movements whilst performing a task. There are a large number of potential features that can be extracted from mouse data; one example in literature using as many as 64 different features to identify the mood of the user (Zimmermann and Gomez, 1984). Many of these features, such as acceleration, are simply derivatives of more basic features, such as speed, and are therefore highly correlated and omitted.

For this study, two metrics were derived from the mouse data:

1. Mouse speed profile; and
2. A novel measure of mouse gesture normality presented here.

For each participant, mouse position X and Y co-ordinate (in pixels), was recorded, as well as the click state, *MOUSE_DOWN* and *MOUSE_UP*.

4.6.1. Mouse Speed

To derive the mouse speed from the raw data, the following procedure was performed:

The Euclidean distance of each sample of positional data was calculated and divided by the difference in time between each point - formally defined by equation 5.

$$s_i = \frac{\left(\sqrt[2]{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2}\right)}{(t_{i+1} - t_i)} \qquad \text{Eq. 5}$$

In which *s* is the speed, *x* is the x coordinate, *y* is the y coordinate and *t* is the timestamp. Figure 4.8 illustrates the resulting time-series plot.

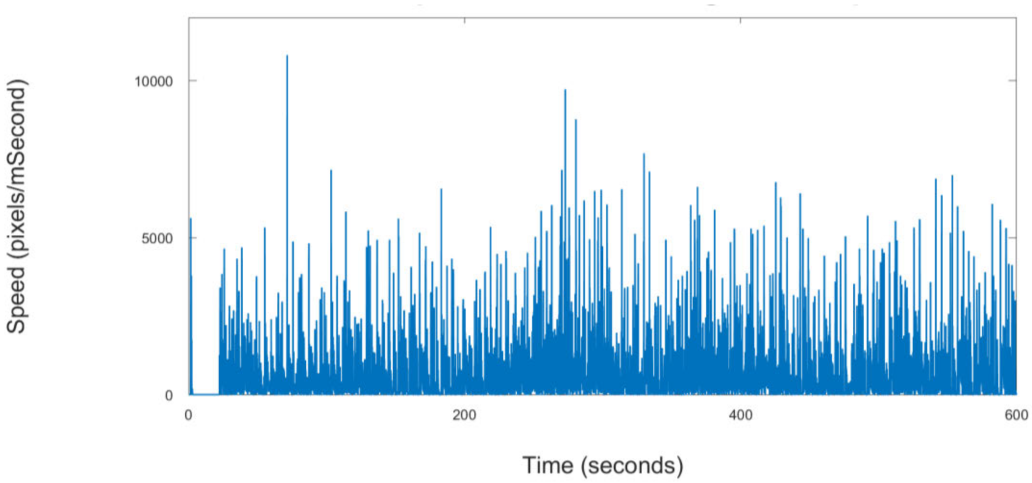


Figure 4.8. Graph of calculated mouse speed in pixels/second for single participant over 10 minutes.

A moving mean average filter is then applied to the speed profile to remove the high-frequency noise (see figure 4.9).

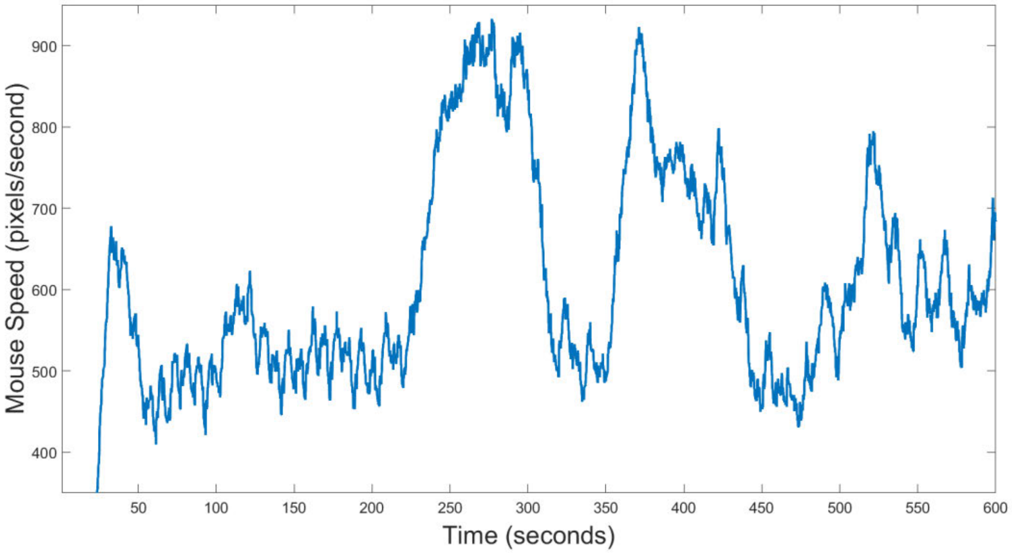


Figure 4.9. Speed Profile across a 10 minute trial with moving average mean filter with 20 seconds window size applied.

The window size of the moving mean averaging filter is one of the parameters tuned in the calibration procedure outlined later in this chapter. The output of this process is a time series representing a speed profile of mouse movement data.

4.6.2. Mouse Gesture Normality

In this section, we present a novel metric to determine the "normality" of mouse movement data. This method models the mouse movement data as a continuous series of 2-D shapes. Our novel method of determining the normality of these shapes works by cataloguing the series of shapes, clustering these shapes, then determining how "normal" each shape is based on each cluster's proportion of the total number of shapes. We then plot this output to a time series to generate a continuous signal of mouse movement "normality".

Delimiting Gestures

To generate a list of gestures from our continuous stream of coordinates, we segment this stream into *sections* or *gestures* by the clicking of the mouse. The gesture begins when the mouse button is pressed and ends when the mouse button is released. This creates a series of delimited gestures, which we can plot on our times-series as seen in Figure 4.10.

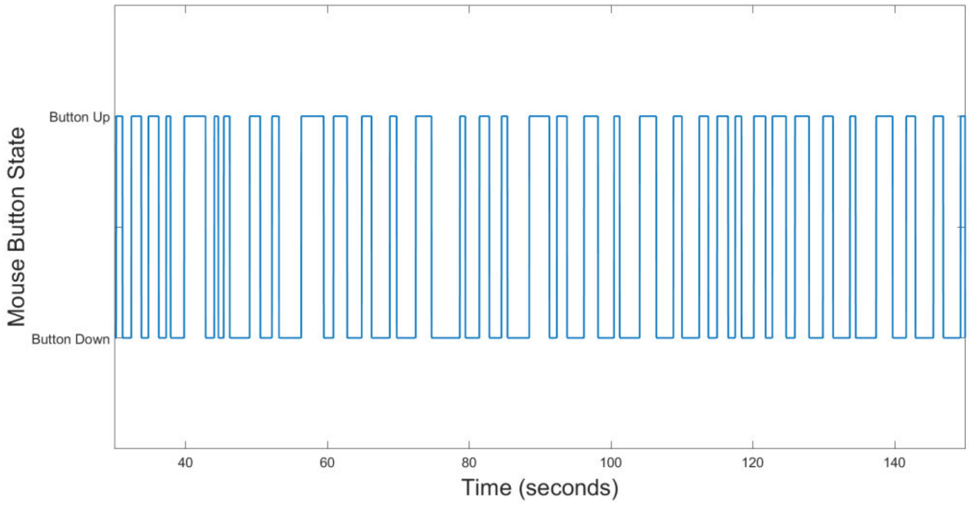


Figure 4.10. Time-series representation of delimited mouse clicks across a 10 minute trial.

Normalising Coordinate Data

As gestures can occur in differing absolute locations on the screen, we need to normalise the gestures so that they can be compared to each other without bias toward their absolute location. To achieve this, we calculate the heading between subsequent coordinates in the list that makes up each gesture – formally defined in equation 6.

$$b_i = \ tan^{-1}\left(\frac{abs(y_{i+1}- y_i)}{abs(x_{i+1}- x_i)}\right) \qquad \text{Eq. 6}$$

In which b is the bearing, y is the y-coordinate and x is the x-coordinate. These bearings provide us with a list of values for a gesture that is resistant to the absolute location on screen. An example of this can be seen in Figure 4.11. Example of two gestures, delimited by mouse clicks (left), being converted to bearing sequences.

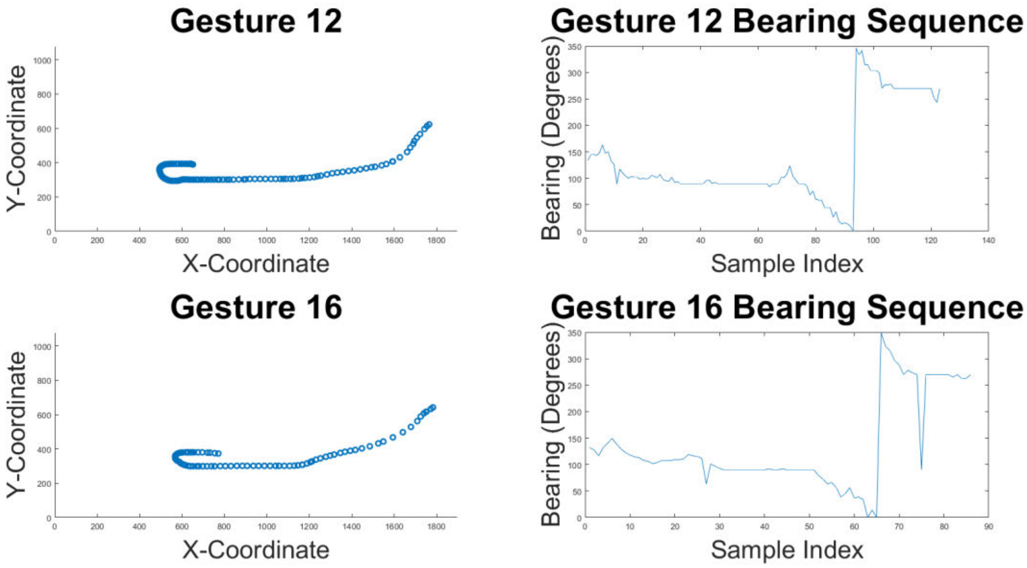


Figure 4.11. Example of two gestures, delimited by mouse clicks (left), being converted to bearing sequences (right).

Comparing Shapes with Dynamic Time Warping

To cluster our bearing sequences into groups, we must first apply a distance metric. As our sequences can vary in length, but may still contain similar shapes, we apply the Dynamic Time

Warping (DTW) method (Sakoe and Chiba, 1978). The DTW algorithm compares two vectors that can be differing lengths by examining a sample-by-sample cost function, to find the optimal global alignment.

The process starts by generating a matrix of distances between all of the points within two of the bearing sequences. Each element of the matrix is calculated using equation 7.

$$D[m,n] = |A_m - B_n| + \min\left(D[m-1,n-1], D[m-1,n], D[m,n-1]\right) \qquad \text{Eq. 7}$$

Where D[m,n] is the distance value at index (m,n), A_m is the value of the first bearing sequence at index m and B_n is the value of the second bearing sequence at index n.

Once the matrix is populated, a path of minimum total value is iteratively calculated from position (m,n) to (0,0), in which the total value is the sum of all matrix elements on the path. This final value is the distance output of the DTW algorithm (see figure 4.12).

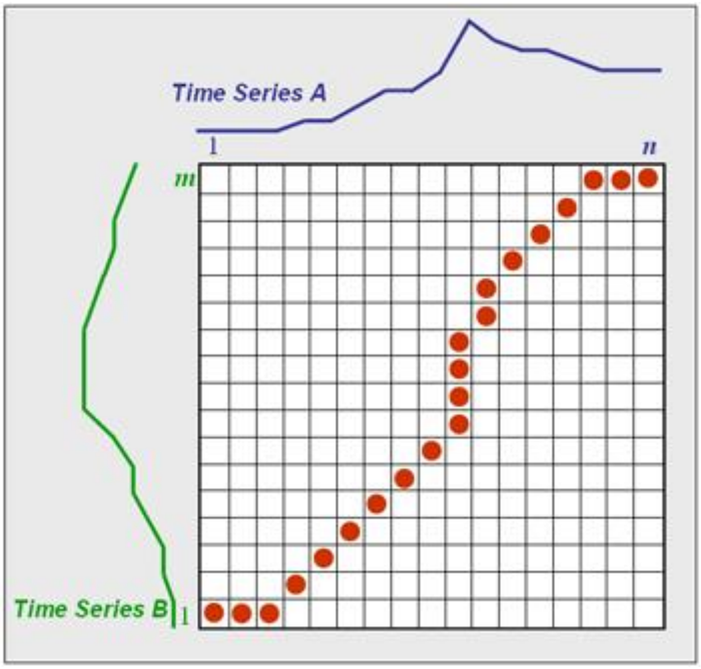


Figure 4.12. Distance Matrix for DTW algorithm. The red dots in the matrix show the optimal alignment between the two time series' A and B.

Clustering with DBScan Algorithm

A distance matrix comparing all bearing sequences with each other is then generated using DTW. This is then clustered into groups of similarity using the DBScan, a well-established clustering algorithm (Daszykowski and Walczak, 2010). The algorithm clusters points in a space based on their density. Figure shows an example of the algorithm’s process. Each point is the centre of a sphere that has a radius of epsilon (input parameter). If a point A in Figure 4.13 was the start point, its sphere contains three further points, which in turn contain two further red points within their radii. C and B are also within the radius of two of the points in the red cluster and are therefore included in the cluster, whereas point N is outside the radius of these points in the cluster and therefore becomes another cluster.

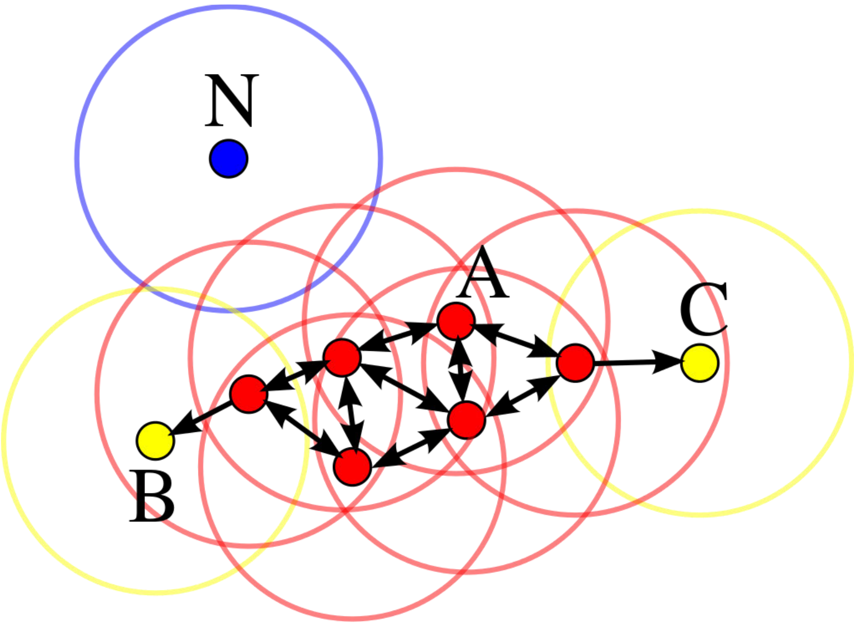


Figure 4.13. Illustration of DBScan Algorithm Output.

Figure 4.13. shows an example of the gestures contained within a single gesture cluster after the DBScan was applied. The red dots represent elements that have been clustered together, with the yellow points demonstrating subsequent elements that will also be classified as red, with the blue dot representing an element that will not be clustered to red given it’s distance from the other elements in the cluster. The epsilon input value of this process forms one of our tuning variables for this measure; the tuning of which will be discussed later in this chapter.

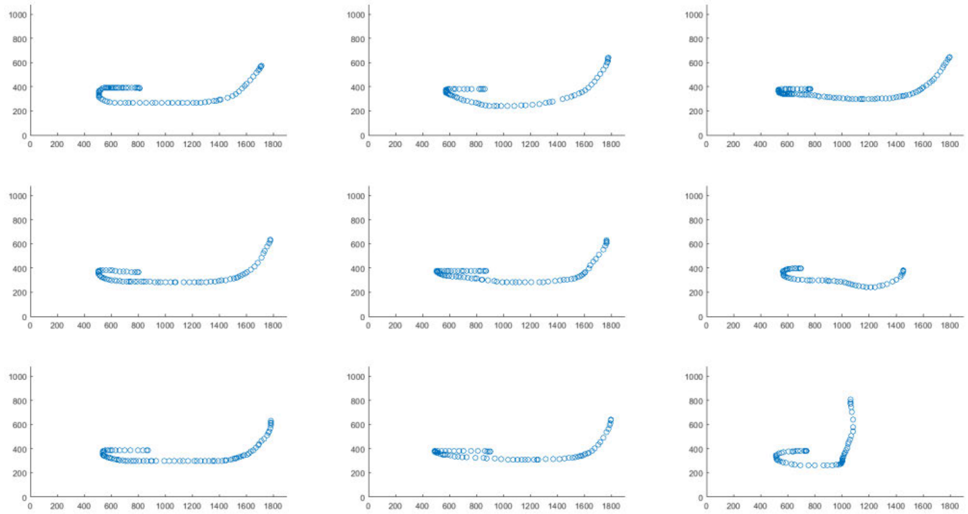


Figure 4.14. Example of gestures that have been clustered together.
Each graph area is a representation of the screen, with points representing X and Y coordinates of a gesture.

Figure 4.14. shows the contents of a single cluster once. The gesture in the bottom right of the figure shows an example of a shape that is different, but within the tolerances of the gesture so as to be included.

Determination of Abnormal Gestures

Our hypothesis for mouse gestures is that the more abnormal gesture shapes will co-occur with the most abnormal events, our Events of Interest (EoI). To determine this abnormality, we sort the clusters by size in descending order, we give the first cluster the value of 1 and each subsequent cluster is labelled in turn. This will allocate a higher value to the more unique gestures. We can then plot the cluster value for each gesture, see Figure 4.15.

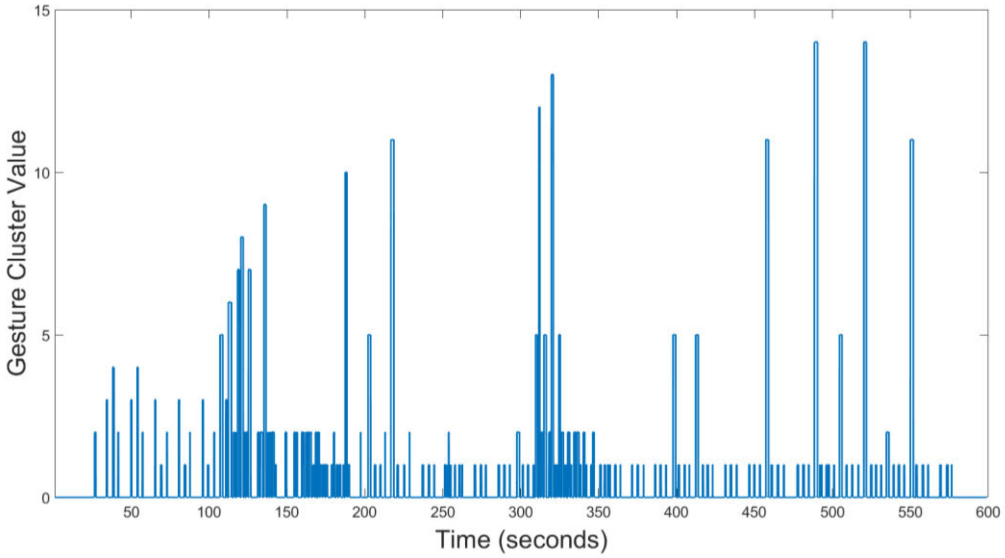


Figure 4.15. Clustered mouse gestures plotted across a 10 minute trial, the x position of the blue bars is the time in which they occurred, the height in y axis represents their cluster number.

From this data, we can apply a moving mean filter to generate a continuous signal that determines the normality of the mouse gestures at a given time. The output of this filter can be seen in Figure 4.16.

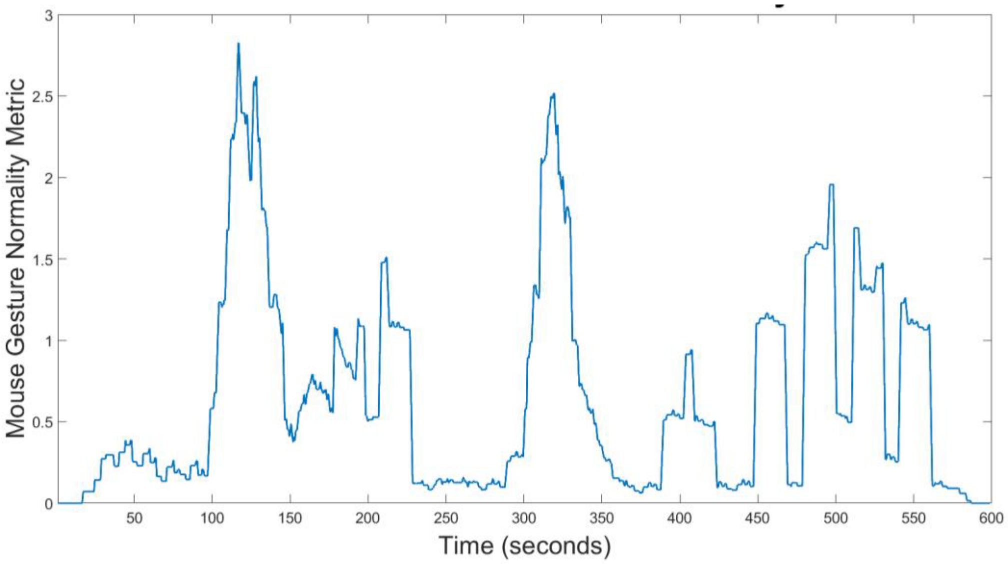


Figure 4.16. Output of mouse gesture normality metric across a 10 minute trial, the y axis representing the "abnormality" of the mouse gestures, the higher the value, the more abnormal.

This novel mouse normality metric has three parameters:

1. The DBscan algorithm epsilon value;
2. The moving mean average window size; and

3. The minimum peak prominence.

These values can be tuned using a calibration data set such as the one gathered from our experiment trials.

4.6.3. Mouse Data Set

The processes outlined in this section outline transforms our raw mouse positional data into two time series: speed profile and shape normality profile. The mouse gesture normality metric is a novel metric to this work. The output of this process is six data sets for each participant, a speed and gesture normality profile for each of the three trials.

4.7. Ground Truth

The final data stream gathered from the simulator is the video recording of the operators using the simulator. This video shows the exact moments at which the events occurred to each participant. The manual annotation of this data highlights an example of the complex and time-consuming difficulties of manual data labelling procedures in real world applications. To assess the accuracy of our analysis of the times at which the events of interest occurred, each participant's video will be examined and the times recorded manually.

To add some perspective to this as a motivation for this research, each participant has 30 minutes of video footage, on average; it takes 2.5 times the length of the video to determine all the events from the footage. As the creators of the simulation space, we define ourselves as experts in this specific application, therefore this average time represents the fastest possible time. In this experiment, we included 30 participants (this will be discussed in a later chapter); this results in 37.5 hours of manual review to determine the correct classifications for this study, highlighting a perfect scenario in which an automated data labelling system could be of use.

For the manual data labelling, there could also be an element of deviation between the onset of different participants events of interest. To minimize the variations in manual definitions of the different event times, a definition of each event is outlined below. These definitions were used to keep the manually located timestamps as consistent as possible.

4.7.1. Sudden Traffic Surge

This event is determined to have started from the moment the first aircraft appears in the surge. As this is programmed into the simulator for the first test set, this time is known prior to the participant using the simulator. The time at which the surge is deemed to have ended is also preprogramed; when the surge ends, normal traffic resumes, thus the moment when the first aircraft appears from the normal traffic section is the time at which the surge has ended. For the second test set, the time at which the traffic surge is deemed to have begun is achieved by identifying when a specific aircraft enters the screen, the type and trajectory of which is unique to the starting of the traffic surge event. Equally, normal traffic resuming is heralded by an aircraft of specific type at a unique trajectory, the presence of this aircraft determines the end of the traffic surge event for the second testing set.

4.7.2. Cloud Cover

The cloud appears at a specific time that is preprogramed for the first test set. As the cloud moves at a constant speed across the screen, it always leaves the field of view 30 seconds after it appears. Thus, the time frame for cloud cover is understood before the participant uses the simulator. For the second test set, the moment when the cloud can first be seen entering the screen is determined to be the start, the end of the event is set at 30 seconds after this point.

4.7.3. Speeding aircraft

The speeding aircraft appear at a time preprogramed into the simulator for the first test set; the appearance of the aircraft at this time marks the start of the speeding aircraft event (which will be the same for all participants in the first test set). The speeding aircraft event is deemed to have ended when the participant has landed the aircraft. This time had to be derived manually by inspecting the screen recording of the individual’s simulator run. The speeding aircrafts appear at specific locations unique to this event type. Thus, for the second test set, when these aircrafts were spotted on the screen recording, the event is deemed to have started, the end of the event is determined in the same fashion as the first test set.

4.7.4. Unresponsive aircraft

Determining the times in which this event occurred was more complex. This is due to the aircraft being identical to all other aircrafts in normal traffic. The unresponsive aircraft appears at locations unique to this event, and were spotted manually using this knowledge. For the participant however, it is only recognisable as unresponsive once they attempt to move it with the mouse. Even then, the user may either not notice that their click had no effect, or simply believe that it was human error on their part. This creates a complication in discovering the ground truth as it is not possible to determine exactly when the event starts. It was decided that for consistency, the moment when the participant first clicks the unresponsive aircraft will be the moment that this event started and the moment the aircraft leaves the screen will be the end of the event.

4.7.5. Note on Ground Truth Labels

As we are assessing elements of cognition in this work, it should be noted that the elements of event start time variance pose a non-trivial problem to creating an accurate ground truth.

When using psychophysiological sensor equipment to assess different elements of cognition, the start time of the onset of the stimulus being investigated is often in a highly controlled environment. In some studies, the participant is explicitly warned of the onset of a stimulus with either some text appearing on the screen or an audible tone. The nature of *when* an event started is not often discussed in literature surround practical applications of psychophysiological cognitive load measurements. For example, with the use of the Task Evoked Pupillary Response (TEPR) and other event-related potentials, the start time is a critical element of the analysis as the baseline for readings is often defined as the preceding seconds up to this point. In an example when an arithmetic question is audibly posed to a participant, this time is often labelled as "question being read" with analysis occurring from the moment the question has finished being asked, but this doesn't not account for individuals who are already processing the information cognitively as the question is being read verses those who may begin this process a second or so after the end of the question. This notion is being mentioned here as a limitation of this study is that the start times of the events is ultimately a decision made by a fallible human; though we have attempted to mitigate this by standardising definitions of event start times in advance.

4.7.6. Ground Truth Data Set

The labels generated through the process outlined in this section form our last data stream processed from the raw data. This data is in three sections, one calibration set and two validation sets, the calibration set will be used to optimise the classification analysis.

4.8. Determining the Times of Events of Interest from Time Series

With our data streams from our psychophysiological signals pre-processed, we will now analyse the data to derive the times in which our *Events of Interest* (EoI) occurred. Our methods for determining points at which events occurred is based on the hypothesis that an event will stimulate a response in the participant’s physiological response.

Previous work in this area pays specific attention to characterizing types of cognitive load within trials. Each of the trials are designed to have subjectively differing mental workload differences. There is little literature on a continuous task with multiple event types within a single trial. Many of the previous works discussed in the literature review focus on methods such as the *Task-Evoked Pupillary Response* (TEPR) (Beatty and Lucero-Wagoner, 2000; Mosaly, Mazur and Marks, 2017; van der Wel and van Steenbergen, 2018). The difficulty with applying this marker in a situational context is that, in order to calculate it, the amplitudes of the pupil diameter increase need to be compared against a baseline pupil diameter, the proper discovery of which has wide ranging approaches. Mathôt notes that the method is sometimes applied inappropriately for the context (Mathôt *et al.*, 2018). The issue with the practical application of TEPR as a recall classifier is that it specifically requires knowledge of the exact moment the stimulus was provided in order to characterise the response. In this work, the identification of the time in which the stimulus occurred is the output and cannot be provided as an input.

Equally, with methods such as HRV, they are demonstrated to have a significant response to increased cognitive load but either comparing a rest scenario to a specific scenario that was previously known to have contained a task of higher difficulty. When longer stimulus scenarios are presented such as those in the study by Fallahi et al, the labelled data often encompasses a significantly broader definition such as “higher traffic density” that the operator experienced for significant lengths of time, these methods do not identify specific individual events that occur within the scenario.

These measures specifically require either a highly controlled lab environment, which limits its practical deployment or a well-labelled dataset, which requires manual labelling anyway which defeats the objective of this work. This labelled data set allows the methods to perform statistical analysis on predetermined segments of time to assess if there is significant change in the signal when compared to, for example, a pre-determined rest period. In some cases, a large number of repeated trials occur and results averaged to assess significance. These methods are not applicable in this research as they all rely on a pre-labelled data set.

In real world environments, no knowledge can be assumed for these purposes, for example, a rest period could be defined as a moment taken for a personal thought or when the operator's attention drifts. As a result, the baseline of their cognitive load is unknown and cannot be determined without expressly defining the environment that represents rest and labelling it accordingly.

In this research, we have a hypothesis that an event of interest will create a response in the cognitive load from the operator. We therefore model the time series as a continuous series of peaks that we assume are stimulated from the simulation environment. Resultantly, we present a novel method for assess cognitive load signals in time series, the method locates all peaks in the time series and then filters the peaks by tuning the parameters that characterise the peaks. This method does not require any knowledge of resting baseline or context of the previous elements of the simulation environment. The characteristics of the peaks that represent real events of interest are determined by tuning the peak finding procedure using the training data.

4.8.1. Peak Finding Procedure

We define a peak in the data as a point P_t in which $P_{t-1} < P_t > P_{t+1}$. We then calculate the *prominence* of each peak in the time series. We extend a horizontal line from the peak in both directions. This line terminates when it either crosses another point in the signal or reaches the end of the time series, in both directions. The data points under each of these extended lines, left and right, form two sub sections. The minima are found of these two subsections. The greater value of these two minima forms the reference value for the prominence; the difference between the peak value and this reference is our peak prominence. This process is illustrated in Figure 4.17.

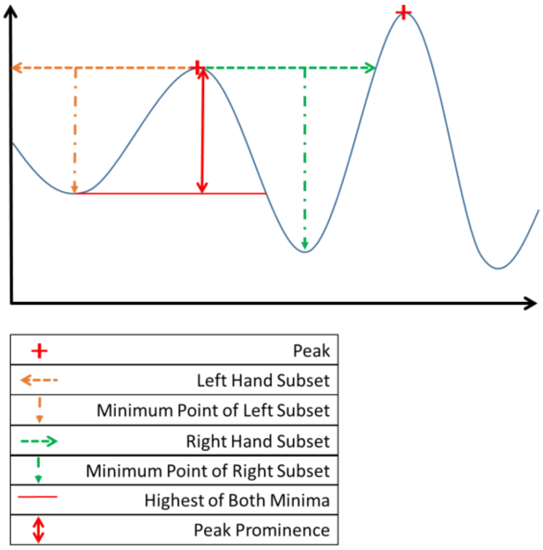


Figure 4.17. Illustration of Peak Prominence Procedure

A visual inspection of the pupil diameter data shows that certain areas show a significant increase in pupil diameter. Figure 4.18. shows an example of a single participant’s pupil data.

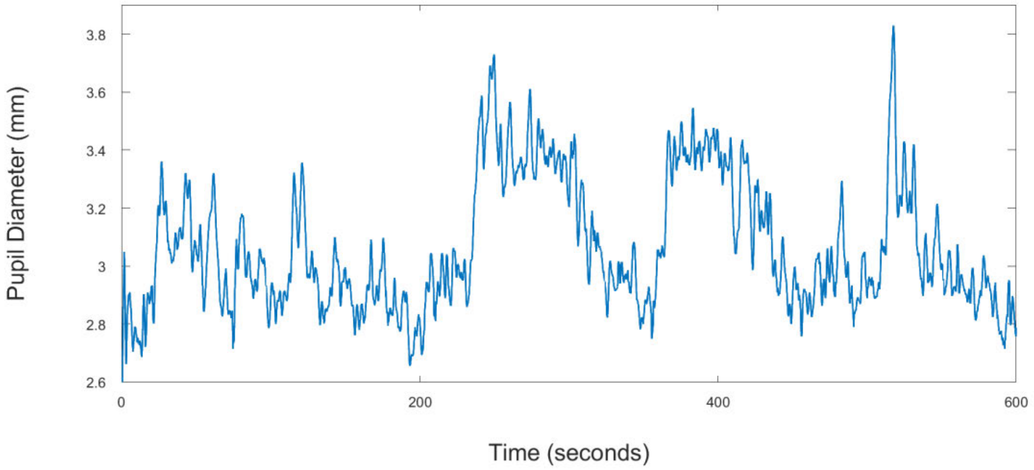


Figure 4.18. Sample from test set of a single participant's pupil diameter data showing absolute pupil diameter over a time period of 10 minutes.

If we overlay the times in which Events of Interest (EoI) occurred, we can clearly see a relationship between these periods of increased pupil diameter and the EoI (see figure 4.19.).

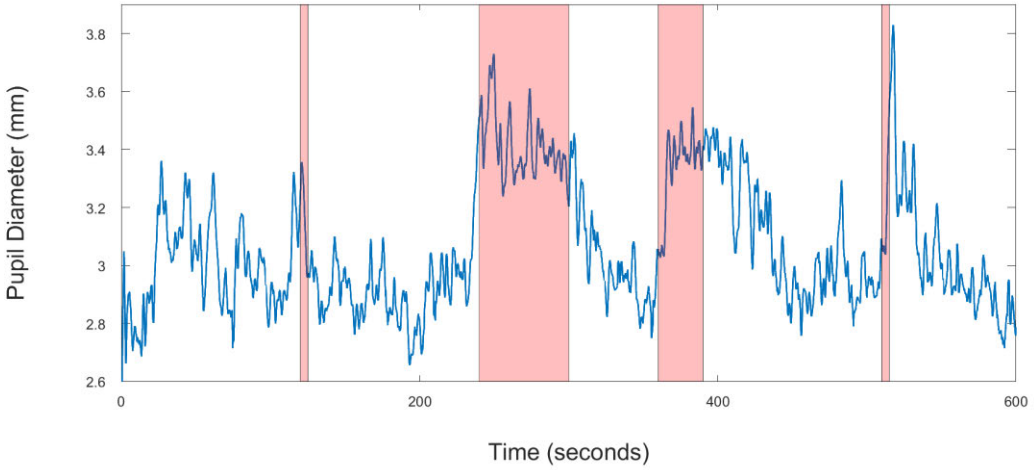


Figure 4-19. A sample of a single participant's pre-processed pupil data over a 10 minute period. The red highlighted areas show times in which Events of Interest were manually verified to have occurred.

We can apply the algorithm to the entire time series signal to find all the peaks within the signal, then we can filter out peaks of certain prominence.

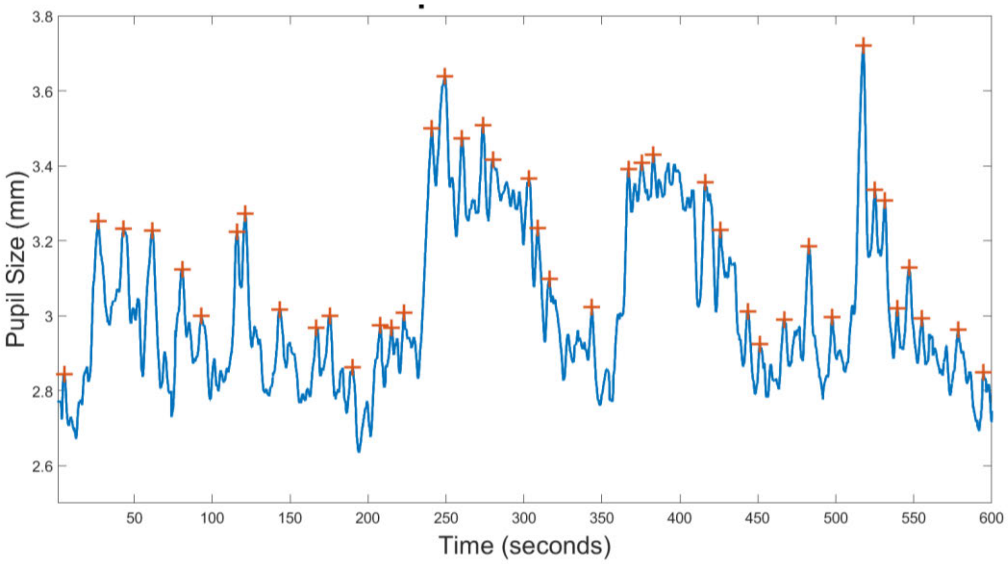


Figure 4.20. Peak Finding procedure result with all peaks below 0.1 prominence filtered out.

Figure 4.20 shows the pre-processed pupil data with the peak finding procedure applied, filtering out all peaks beneath a prominence of 0.1. Given the higher frequency components of the signal, we can see a larger number of peaks located. Our hypothesis is that events in the simulator will stimulate a peak response from the pupil; we must now differentiate peak responses from normal events and those from our abnormal events of interest, which we expect to be greater. We can filter out smaller peaks in two ways:

1. To adjust the data with a moving average filter in order to smooth the peaks from the signal, or;
2. To adjust the minimum peak prominence input of the procedure.

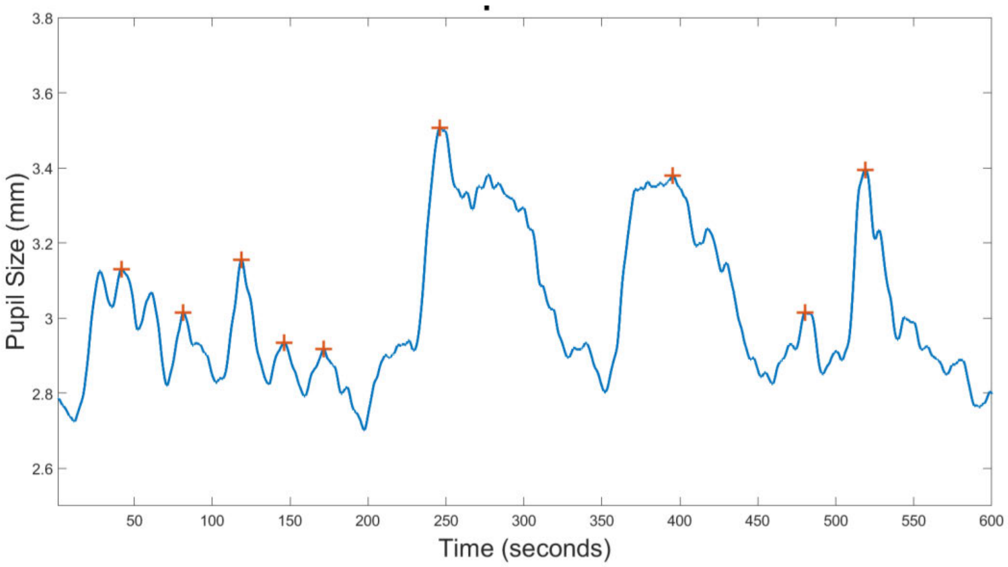


Figure 4.21. Peak finding result when moving mean filter of window size of 10 seconds applied

Figure 4.21.shows the effects of a 10-second moving mean window filter on the number of peaks detected by the peak finding procedure compared to Figure 4.20.

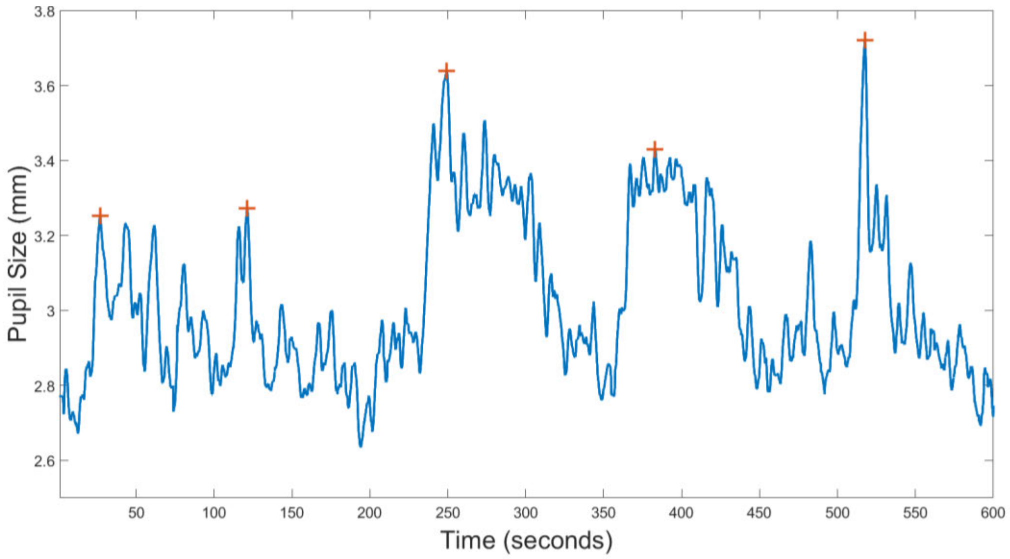


Figure 4.22. Results of peak find procedure of unfiltered data with minimum peak prominence set at 0.5.

Figure 4.22 shows the results of the peak finding procedure on unfiltered data when the minimum prominence has been increased to 0.5. We can see that the effect of both techniques reduces the number of peaks located, identifying those that represent the highest-amplitude peak responses.

4.8.2. Tuning Metrics with Calibration Data

In this section we discuss the calibrating of the parameters of the methods outlined in the preceding sections. We will examine each type of data in turn, optimising the parameters for best classification precision and recall of events of interest. This optimisation will be performed across each individual, to then be tested on the validation set.

Classification Recall and Precision

Given that our methods cannot generate true negative results, we assess the performance of our classifier using recall and precision. Recall is the proportion of all the events that were correctly classified – formally defined in equation 8.

$$Recall = \frac{Total\ True\ Positives}{Total\ True\ Positives + Total\ False\ Negatives} \qquad \text{Eq. 8}$$

Precision is the proportion of classifications that were correct – formally defined in equation 9.

$$Precision = \frac{Total\ True\ Positives}{Total\ True\ Positives + Total\ False\ Positives} \qquad \text{Eq. 9}$$

Both recall and precision are important factors in determining the accuracy of the system. As a result, we combine the measures using a weighted average. We use the F1 score for the calibration data for each data type (pupils, HRV, Mouse speed and mouse gesture). The F1 score is the weighted average of the recall and precision results – formally defined in equation 10.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad \text{Eq. 10}$$

Pupils

The pupil data method has two parameters: the size of the moving average window and the minimum peak prominence. We calculate the F1 score for all combinations of these inputs; we then select the optimum values for our parameters before repeating this process to obtain window size and minimum peak prominence for all participants (see figure. 4.23).

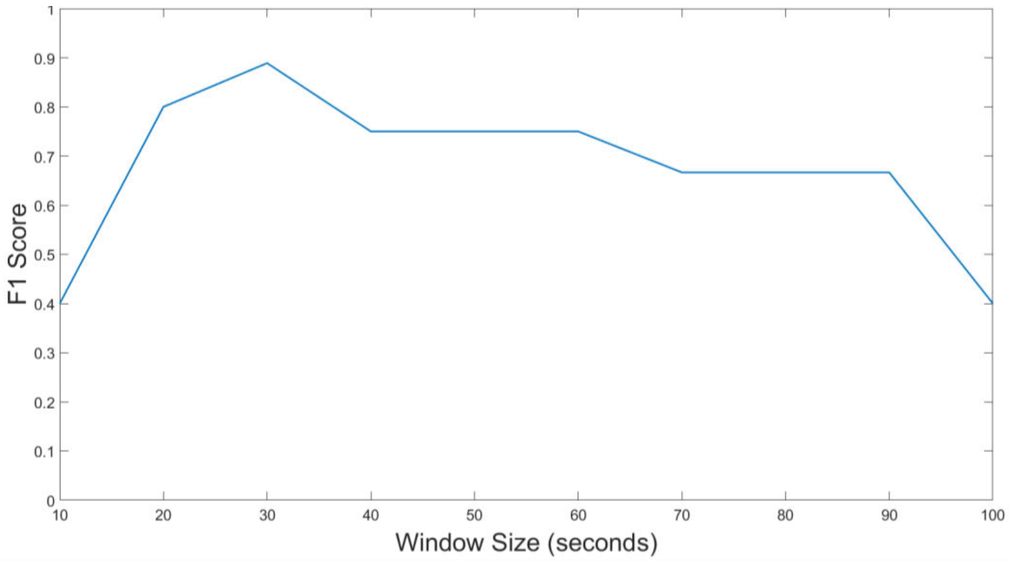


Figure 4.23. Example of optimising pupil classifier F1 Score. The F1 score changes depending on the window size in this example; the window size of 30 seconds was selected as it yields the greatest F1 score of 0.9.

HRV

We optimise our HRV values in the same fashion as the pupil data. The input parameter for the HRV data is the minimum peak prominence. We assess this value across each participant to find the optimum F1 score. We also take an average of this score across all participant for each of our HRV metrics to assess the optimum measure to use.

Table 4.1. Average F1 Scores Across Different HRV Metrics

HRV Measure	Average F1 Score
HRV - HF	0.421
HRV - MF	0.500
HRV - LF	0.444
HRV - Mean RR	0.725
HRV - RMSSD	0.807

We see from the results in Table 4.1. that the mean RR and RMSSD (Root mean square of successive differences) metrics outperform the frequency measures by approximately 20%. The highest performing metric is the RMSSD, this metric is one of the most frequently cited time domain metrics in literature and our results confirm its validity as a basis of a method to analyse cognitive load signals.

Mouse

Our mouse speed metric has two input parameters:

1. The mean average window size; and
2. The minimum peak prominence.

These are optimised in the same fashion as the pupil data, iterating over combinations of both parameters and a maximum F1 score chosen.

The procedure was repeated for the mouse gesture metric over the three input parameters: DBscan epsilon, mean average window size and minimum peak prominence.

4.8.3. Summary

In section 0 we presented our method for determining the times at which event of interest occurred in a control room scenario using psychophysiological signals. We incorporate the understanding of physiological response peaks and the limitations of current methods for this research scope. We showed how we will calibrate the parameters from each of our data streams to achieve optimal classification accuracy using the F1-score. For the pupil, heart and mouse speed data, we tune the size of the window for the moving average filter and the minimum peak prominence. We also tune these parameters for the mouse gesture method, with the addition of the DBScan epsilon value. By properly calibrating these parameters for each of our participants using their calibration data set, we optimise the methods for testing on our validation data sets.

4.9. Determining Correct Classifications

The output of our methods is a time at which a peak of sufficient magnitude was determined for that particular data type. To determine if this timestamp was correctly classified as an event of interest, the timestamp output was compared with the ground truth.

The issue with this approach is that it does not account for latency in response for shorter duration events. Latency periods vary depending on the measure used as well as variations occurring person-to-person, based on the moment at which the individual was aware the event was occurring. For pupil diameter, an average latency for a pupil response can be about 2 seconds (Hoeks and Levelt, 1993), whereas heart related measures seem to have widely varying recorded latency periods of up to 10 seconds. Intuitively, we expect there to be a greater latency with the mouse data as this response has the added delay of cognition + motor action response. To determine the length of the latency, or delay, window for each channel of data, we examine the time between the onset of an event and a response (if any) in each channel. The purpose of calculating this window size is to get a value that best reflects the natural latency in each measure; this was performed before any optimisation was performed.

4.9.1. Latency Response per Channel

Here we assess the time between the event starting in ground truth and the timestamp method output in each channel for each participant in our training set. For each participant we have four events in the 10-minute training data. We place a 20-second window starting at the onset of an event in ground truth and test to see if the classified data stream responded, we then calculate the time taken to respond for each participant for each event. If no response is detected within the 20-second window, we do not record any latency (false-negative).

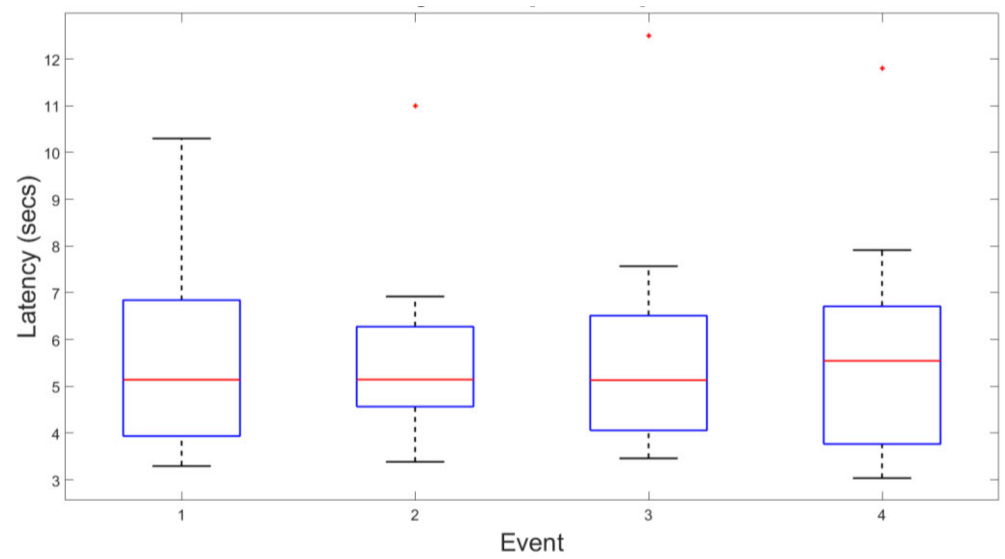


Figure 4.24. Latency Boxplot of Pupil Responses over the four events in training data.

We see from the boxplot in Figure 4.24. that the average pupil response recorded is approximately 5 seconds, with the fastest responses being approximately 3 seconds.

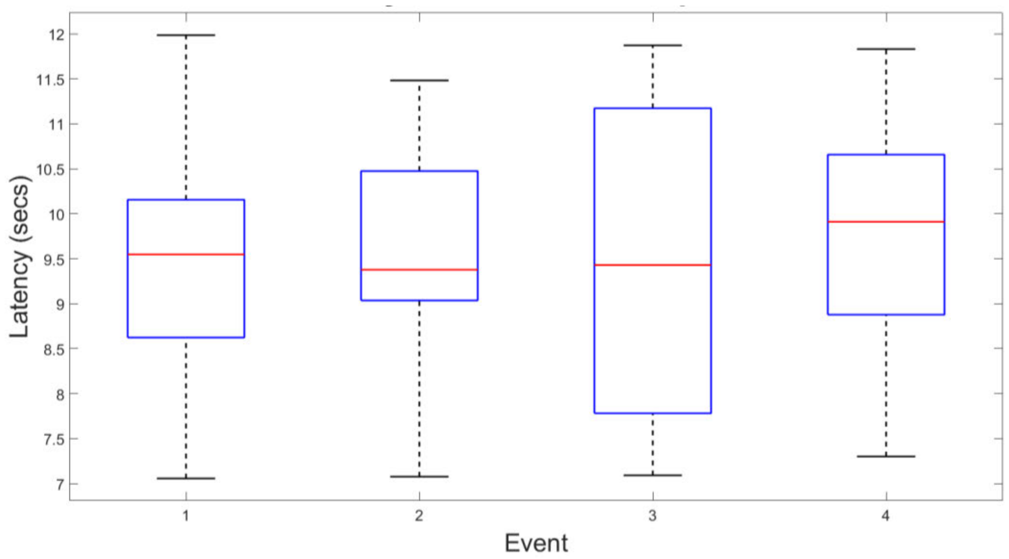


Figure 4.25. Latency Boxplot of HRV responses using the mean R-R measure.

Using the mean RR measure we note a delay of approximately 9.5 seconds for a response, as expected, longer than the pupil response (see figure 4.25).

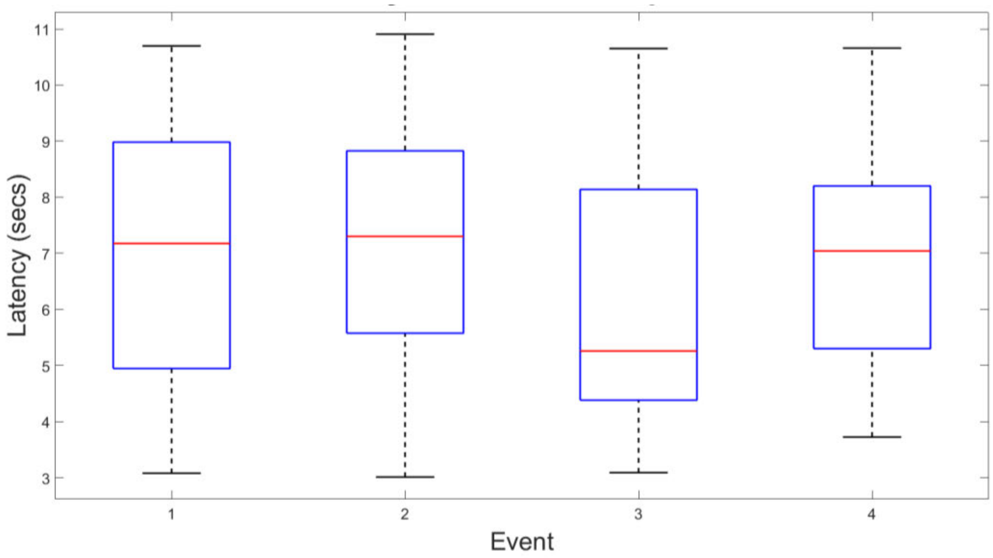


Figure 2.26. Latency Boxplot of HRV responses using the RMSSD measure.

We see in Figure 2.26. that response times of the root mean square of successive differences respond on average, faster than their mean RR equivalents.

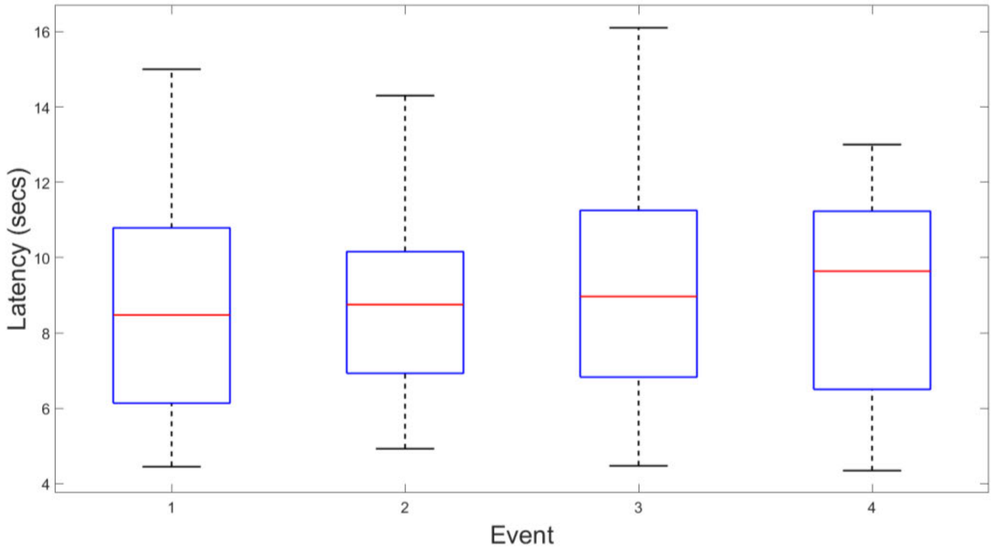


Figure 4.27. Latency Boxplot of HRV responses examining the Low-Frequency component measure

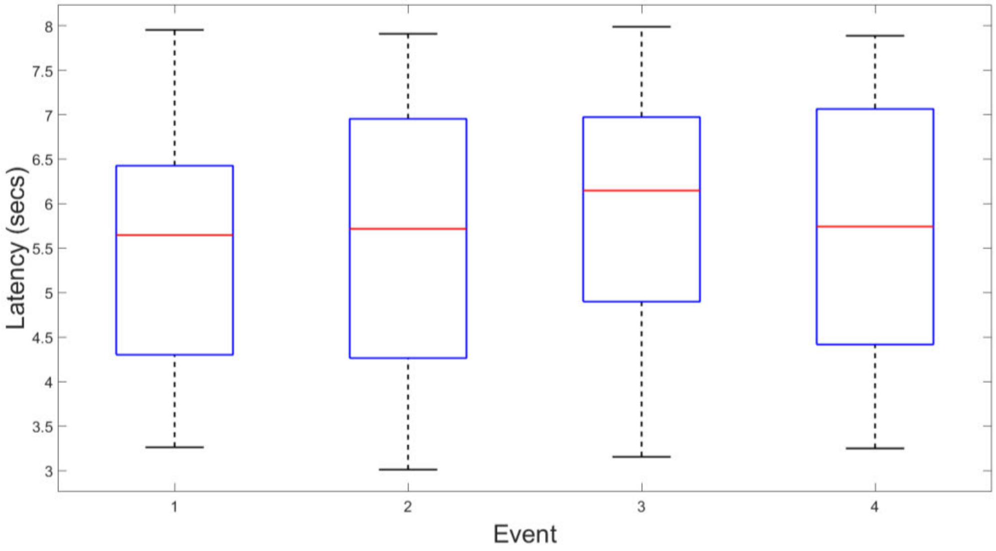


Figure 4.28. Latency Boxplot of HRV responses examining the Mid-Frequency component measure.

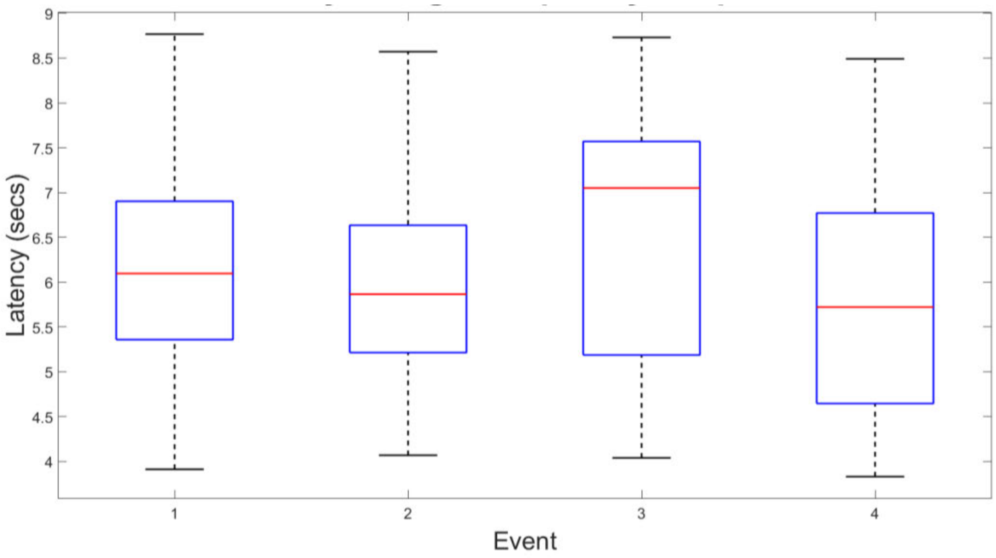


Figure 4.29. Latency Boxplot of HRV responses examining the High-Frequency component measure.

From examining the latency in our frequency measures, we see an increased average response time to that of the mean RR measure and in the mid-high frequency, we see response times comparable to that of the pupil measures See figures 4.27-4.29).

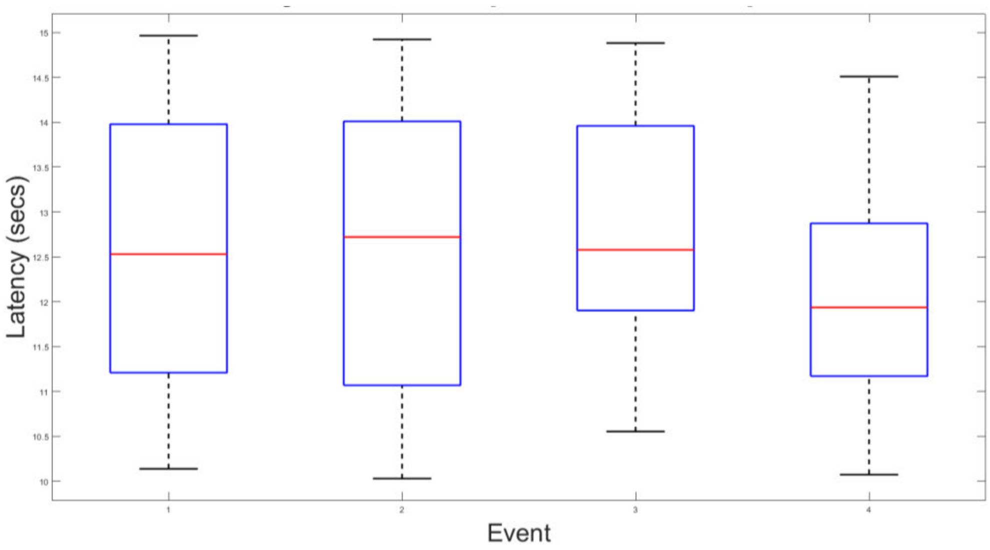


Figure 4.30. Latency Boxplot of mouse speed metric responses.

The higher response times shown in Figure 4.30. Figure 4.31 confirms our hypothesis that mouse measures would have a longer latency than physiological measures.

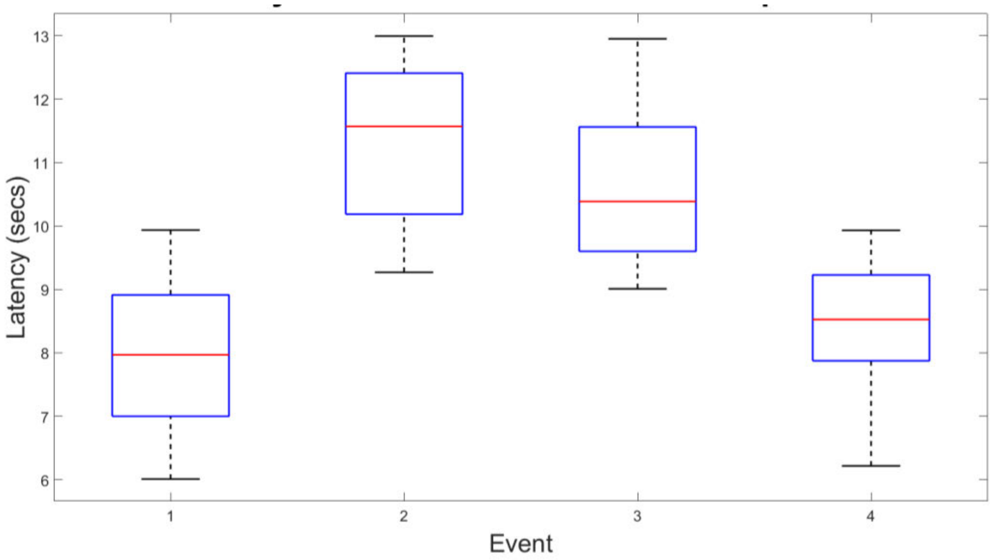


Figure 4.31. Latency Boxplot of mouse gesture metric responses.

The response speed for the mouse gesture metric has the most variation between the events, most notable, between events 2&3 and 1&4. This is intuitive as events 1&4 are the short-term speeding aircraft events, which required a faster response.

4.9.2. Latency Summary

Each measure has some latency; a window needs to be selected for each measure that accounts for this. Naturally, the larger the window, the higher number of correct classifications. If the window becomes too large however, we run the risk of incorrectly labelling false positives as true positives. Establishing this temporal segmentation is not a trivial matter, "common sense" backed with established theory is employed here to establish this window. This is discussed in work by Scheirer in which they attempt to define these windows, but concede that a fixed window, though necessary, does not factor in the many variables at work during an applied study involving psychophysiological measures (Scheirer *et al.*, 2002). Here, we establish our windows through an examination of literature, which provides a wide range of latency values and also through an examination of latency values in our training data. These values are not precise, but do agree with values reported in the literature. As briefly discussed in section 0, the ground truth is established as objectively as possible, but factors outside of the realm of reasonable objective precision are at play. We will discuss these factors more in the Results and Discussion chapter.

Table 4.2. Maximum Response Times of all metrics derived from raw data.

	Maximum Response Time (seconds)				
Response Measure	**Event 1**	**Event 2**	**Event 3**	**Event 4**	**Maximum Response**
Pupil	10.25	6.85	7.55	7.94	10.25
HRV - Mean RR	11.99	11.49	11.8	11.76	11.99
HRV - RMSSD	10.81	10.92	10.74	10.75	10.92
HRV - LF	15	14.3	16.1	13	16.10
HRV - MF	7.95	7.91	7.99	7.89	7.99
HRV - HF	8.77	8.57	8.73	8.49	8.77
Mouse - Speed	10.2	14.91	14.51	-	14.91
Mouse - Gesture	9.94	12.99	12.95	9.93	12.99

In Table 4.2. we see the maximum values from each boxplot shown previously. Table 4.3. shows the chosen latency windows for each measure for clarity.

Table 4.3. Response Latency Windows Chosen for each data type.

Response Measure	Latency Window (Seconds)
Pupil	10
HRV - Mean RR	12
HRV - RMSSD	11
HRV - LF	16
HRV - MF	8
HRV - HF	9
Mouse - Speed	15
Mouse - Gesture	13

A graphical illustration showing this classification validation process using latency windows can be seen in Figure 4.32.

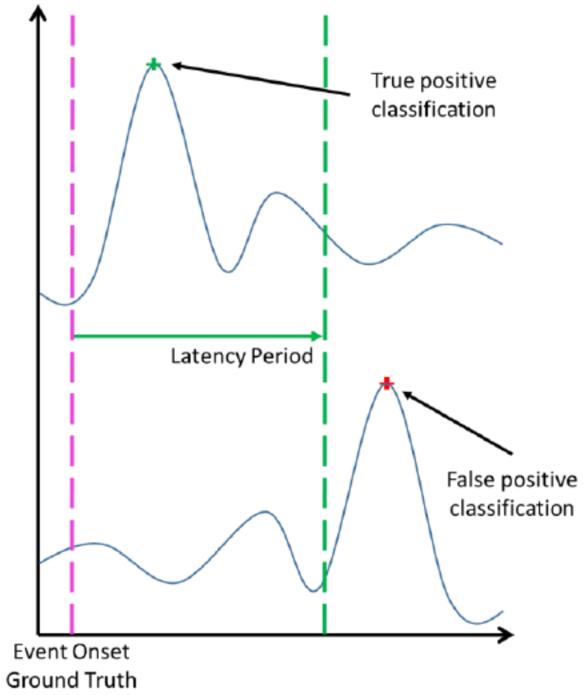


Figure 4.32. Graphic illustration of latency window.

This method defines how we determine if an event is correctly classified or not. The purple vertical line denotes the time at which the event was determined to have started – the determination of this time is explain in section 0. For each data type (pupils, heart, mouse) for each participant, we cycle through every event in the ground truth for that participant, for each event we search for a timestamp - generated by the EoI detection method - in a window starting at the point when the event was manually deemed to have occurred + the latency window for that particular data type. If the timestamp that the particular data method deemed to be an event of interest occurs within this window, it is deemed a true positive classification, if outside, it is deemed a false positive.

4.10. Summary

In this chapter, we outline our methods for transforming the raw data collected from our experiment trials into time series and determining the times at which events of interest occurred within. The methods designed and described in this chapter will allow us to answer our research questions. As for each research question, there in an input of raw data and an output of time stamps.

For each data type, we have determined a method to automatically retrieve time stamps at which events of interest are predicted to have occurred. We start by using standard methods to pre-process each of the psychophysiological signals. We also present a novel method to create a time series of mouse movement data that represents how “normal” the shapes being generated by the mouse are.

We then present a novel method to analyse time series data for significant peaks of operator response. This method circumvents issues with some of the methods in literature that require highly controlled environments or pre-labelled datasets in order to characterise signals.

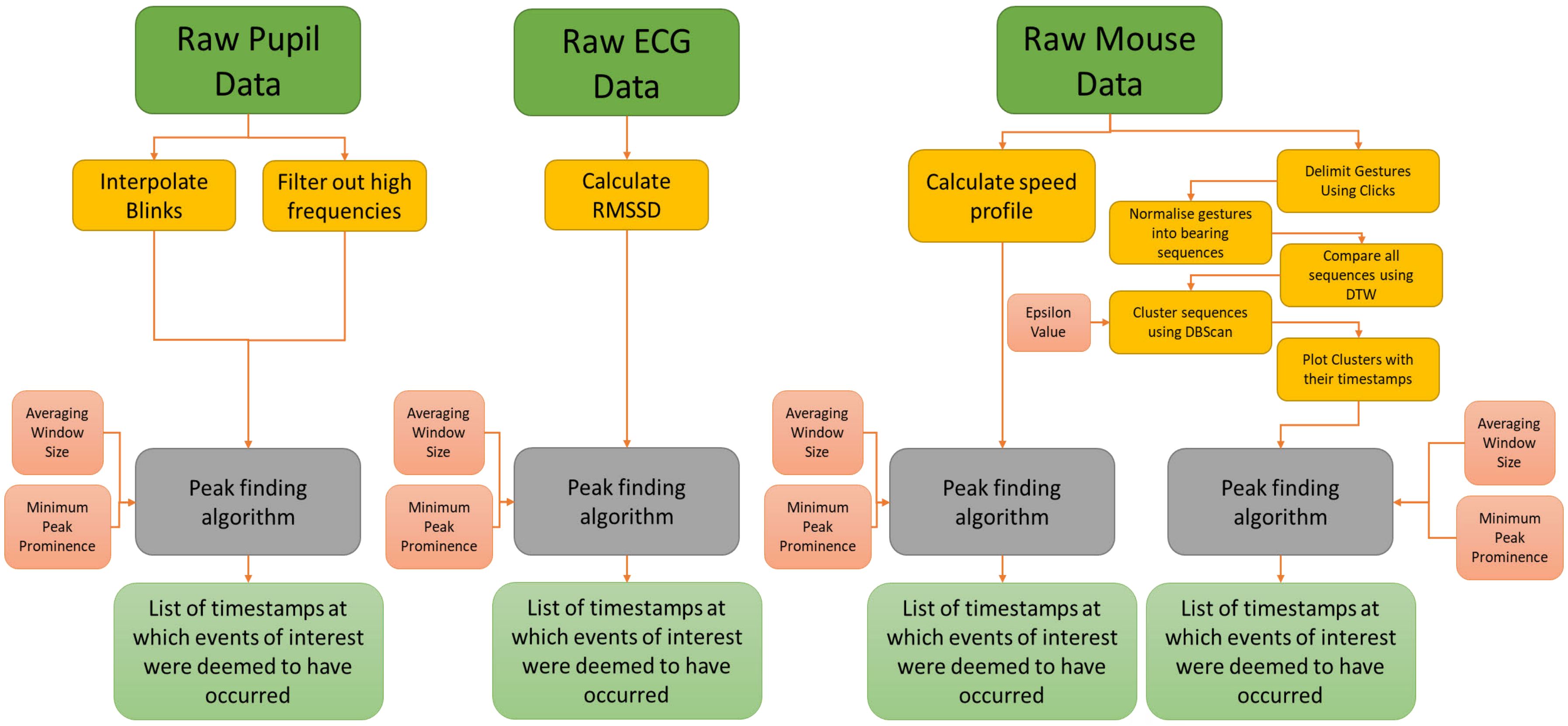

Raw Pupil Data
Interpolate Blinks
Filter out high frequencies
Averaging Window Size
Minimum Peak Prominence
Peak finding algorithm
List of timestamps at which events of interest were deemed to have occurred
Raw ECG Data
Calculate RMSSD
Averaging Window Size
Minimum Peak Prominence
Peak finding algorithm
List of timestamps at which events of interest were deemed to have occurred
Raw Mouse Data
Calculate speed profile
Delimit Gestures Using Clicks
Normalise gestures into bearing sequences
Compare all sequences using DTW
Epsilon Value
Cluster sequences using DBScan
Plot Clusters with their timestamps
Averaging Window Size
Minimum Peak Prominence
Peak finding algorithm
List of timestamps at which events of interest were deemed to have occurred
Averaging Window Size
Minimum Peak Prominence
Peak finding algorithm
List of timestamps at which events of interest were deemed to have occurred

Figure 4.33. Process diagram detailing the data process' to transform the raw data types into timestamps in which events of interest were deemed to have occurred. The red boxes show the parameters for the peak find algorithm that are calibrated to each participant using the calibration data set.

Figure 4.33. shows the process of transforming the raw data into timestamps using the methods described in this chapter. The red boxes denote the parameters for each data types that are calibrated to each participant using the calibration data set.

We also perform analysis on all of our classified data streams to determine the effect of latency on each channel. We discover that pupils have the fastest response time to the onset of an event of interest followed by the RMSSD of the ECG signal, thus we select the RMSSD as our chosen metric of HRV for this study. We also determine the latency for the mouse movement data. This latency study allows us to define custom windows, which are critical in the determination of whether an automatically generated time stamp agrees with the manually determined time stamps. The comparison of which will determine the answer to our research questions pertaining to the validity of this methodology of automatically data labelling.

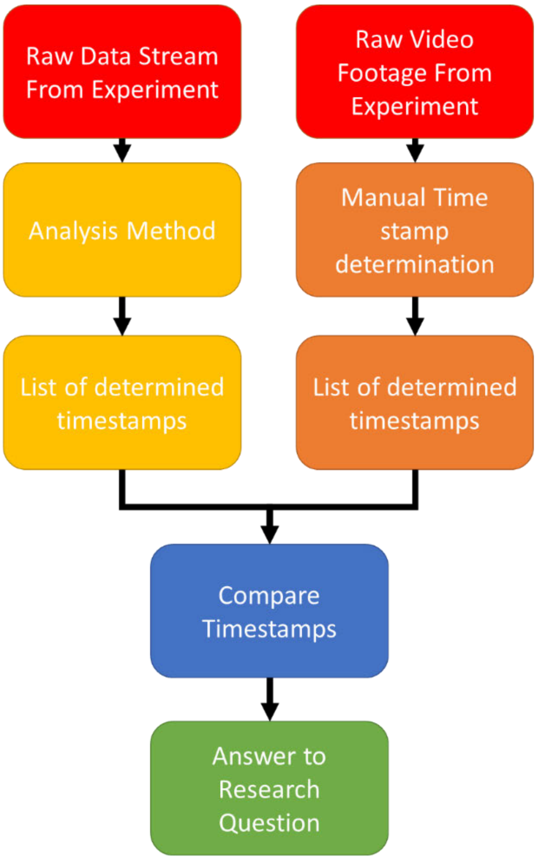

Raw Data Stream From Experiment
Raw Video Footage From Experiment
Analysis Method
Manual Time stamp determination
List of determined timestamps
List of determined timestamps
Compare Timestamps
Answer to Research Question

Figure 3.34. Flowchart demonstrating the transformation of raw data to time stamps.

In Figure 3.34.we show the transformation of raw data into time stamps. This is performed automatically using the methods outlined in this chapter. To assess the answer to our research questions we must also have a list of time stamps that represent the ground truth. This chapter also details how the ground truth is obtained. For the purposes of this research, the ground truth was obtained manually by annotating the data using definitions of each of the events of interest. These timestamps take a considerable amount of time and effort to obtain and, as discussed, may still not be totally accurate, but for our purposes, are considered to be the benchmark standard by which we will compare our automatically determined time stamps. We will assess the precision and recall capabilities of our system in the next chapter.

5. Results and Discussion

5.1. Introduction

In the previous chapters, we have generated a set of research questions by assessing the state of literature in this field. We then developed a methodology to generate data that would enable us to answer our research questions. We then created a set of analysis methods to transform the data gathered from the experimental trials into time stamps. We also developed a methodology for manually acquiring ground truth time stamps from video footage of our trials and gathered this data to compare our automatically generated timestamps against.

In this chapter, we will assess the accuracy of our automatic labelling system. We will assess this by comparing the automatically generated timestamps against the ground truth. This accuracy will be determined by way of precision and recall. Recall representing the proportion of potential correct timestamps found and precision representing the proportion of events that were *correctly* identified.

We first outline the resulting data from the experimental trials, detailing the nature of the data acquired and the participants used in the study.

We then present the results of the analysis. The results are structured by answering each of our research questions separately. For each question, we examine the recall and precision results over several sub sets of data. We first examine the results gathered for the first validation set. We then assess results for the second validation set. We then assess the results when averaged over the population for the first validation set. We then assess the precision and recall results on a per event basis to determine if different event types were more or less likely to be classified by our analysis method. We then summarise the results in the discussion.

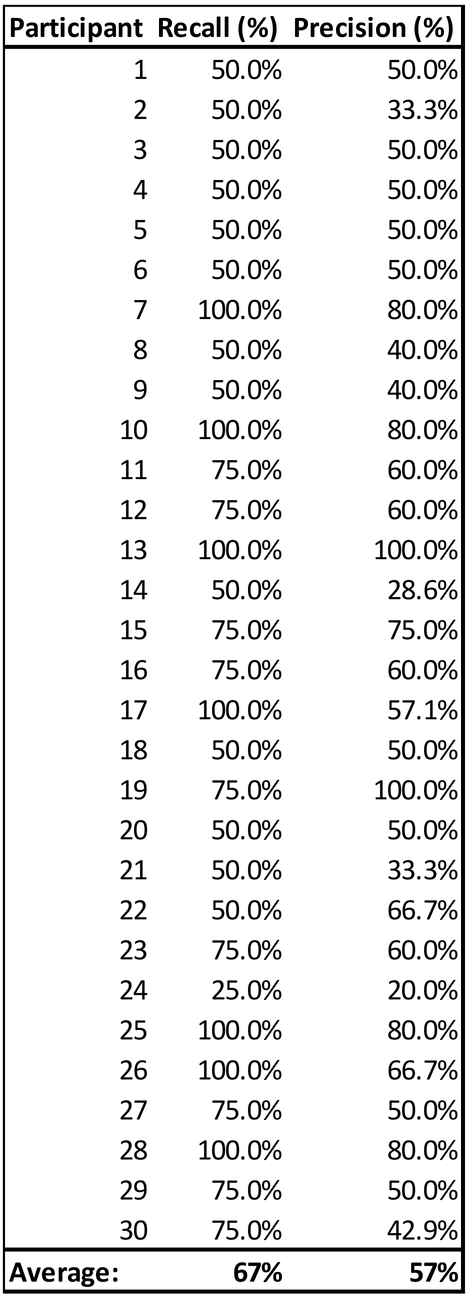
5.2. Research Question 1: *Can pupil dilation data be used to identify the times at which events of interest occur in an operator scenario?*

Our first research question focuses on the viability of pupillometry data as an automated annotation tool for unlabelled datasets. Our results were tuned on a training set of data then tested on 2 different sets of validating data. We will present the results of each participant in terms of *recall* and *precision*. Given that our time series has no mechanism for defining *true negative* classification results, a standard confusion matrix will yield little value.

5.2.1. First Validation Set

The first validation set contains data taken from the second half of the first experiment. This data represents 10 minutes of operator work. The 10-minute section contained four EoI that were to be correctly classified from pupil data. Though a useful measure, *number of correct classifications* does also include false positives, making recall and precision more valuable measures of the quality of our classifier. Table 5.1. contains the results of the classification recall and precision for all participants in validation set one.

Table 5.1 Table of Recall and Precision Results For Pupil Data Classifier on Validation Set 1.



Participant	Recall (%)	Precision (%)
1	50.0%	50.0%
2	50.0%	33.3%
3	50.0%	50.0%
4	50.0%	50.0%
5	50.0%	50.0%
6	50.0%	50.0%
7	100.0%	80.0%
8	50.0%	40.0%
9	50.0%	40.0%
10	100.0%	80.0%
11	75.0%	60.0%
12	75.0%	60.0%
13	100.0%	100.0%
14	50.0%	28.6%
15	75.0%	75.0%
16	75.0%	60.0%
17	100.0%	57.1%
18	50.0%	50.0%
19	75.0%	100.0%
20	50.0%	50.0%
21	50.0%	33.3%
22	50.0%	66.7%
23	75.0%	60.0%
24	25.0%	20.0%
25	100.0%	80.0%
26	100.0%	66.7%
27	75.0%	50.0%
28	100.0%	80.0%
29	75.0%	50.0%
30	75.0%	42.9%
Average:	**67%**	**57%**

We can see from these results an average recall of ~67%. We also see that 16 of the 30 participants recall was at least 75%; with seven of those demonstrating perfect recall. We also see and average precision of 57% across all participants, with 14 participants scoring above 50% precision.

We can make two key observations at this stage:

1. All participants correctly classified at least one EoI.
2. There is a subset of particular individuals that scored higher than average recall and precision results.

We will now examine a few representative samples from these results.

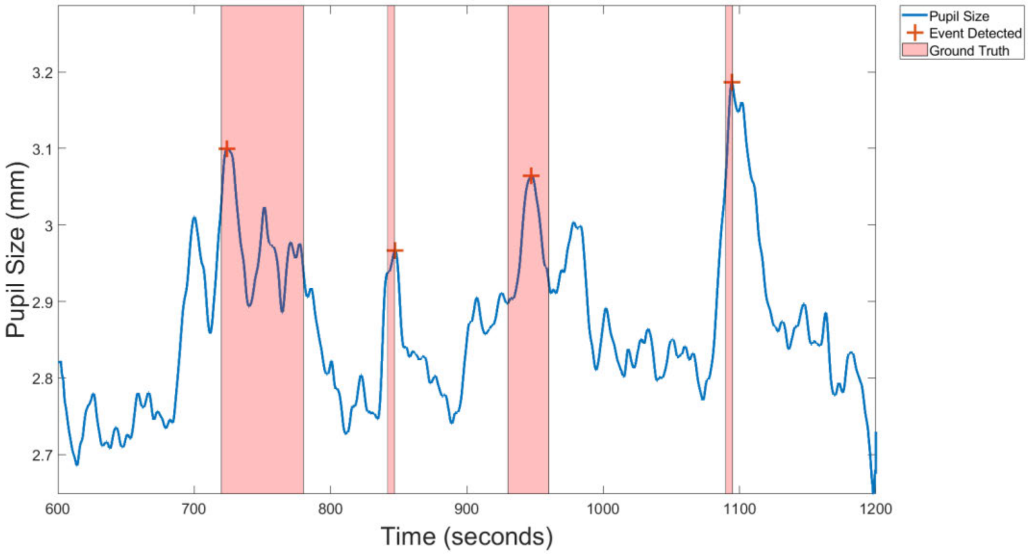


Figure 5.1. Results From Participant 13 - an example of perfect recall and precision.

Figure 5.1 is an example of a participant that was classified with perfect recall and classification. We note that for long-form events (traffic surge and cloud cover – on the figure: 1st and 3rd shaded areas), the pupil response is longer in terms of increased amplitude than those of the short term events (2nd and 4th). The long-form events also appear to have a preceding amplitude increase before the events actually started. A number of potential reasons for this can be considered:

- as a natural side-effect of the averaging filter that was first applied to the data.

- the participant responding to event that were not pre-programmed into the simulator, for example accidentally creating a situation in which two aircraft nearly collide, requiring intervention and increasing cognitive load and therefore pupil diameter. We will discuss in depth these potential scenarios later in the chapter.

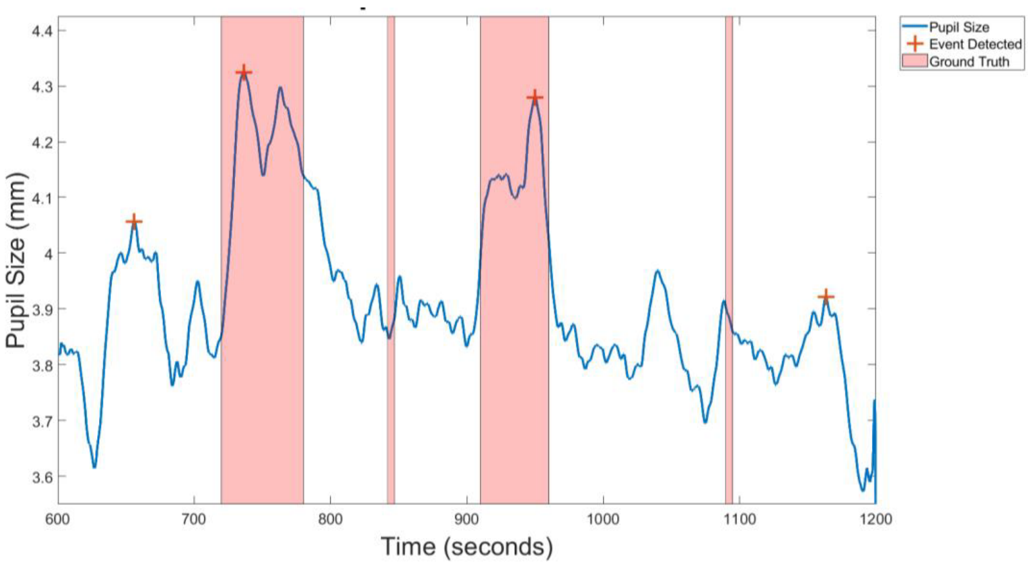


Figure 5.2. Results From Participant 4 - an example of 50% recall and precision.

In Figure 5.2. we show an example of a participant achieving 50% recall and 50% precision. This particular individual shares these results with six other participants. We see from the results that the long-form events are correctly classified, we can also visually confirm a significant amplitude increase during these events. Neither of the shorter term events were correctly classified; the first speeding event (2nd highlighted section), seems to show no significant pupil response whereas the second speeding event (4th highlighted section) does show a response that was misclassified by our algorithm (false negative). The two false positives here could be as a result of other events not pre-programmed into the simulator – we will discuss this further in the chapter.

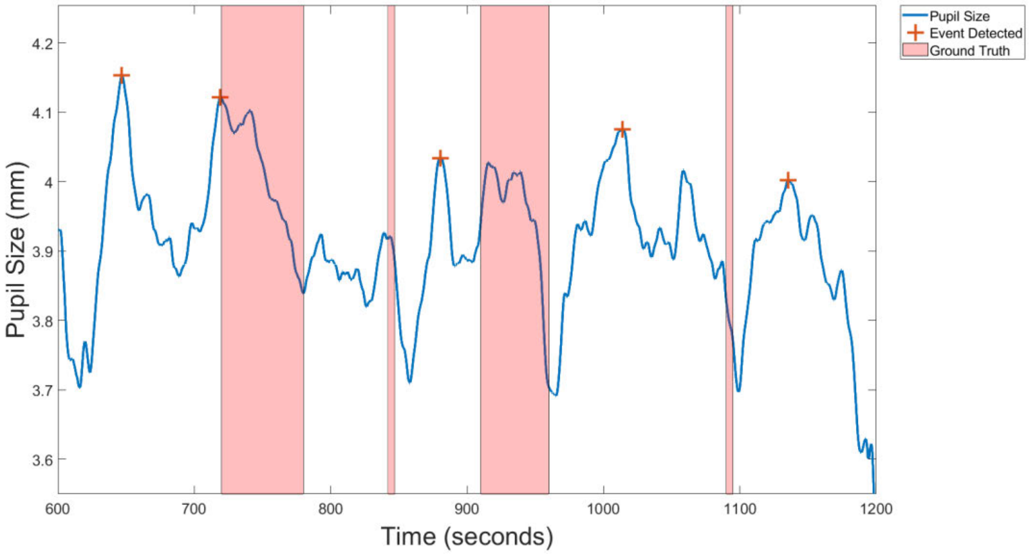


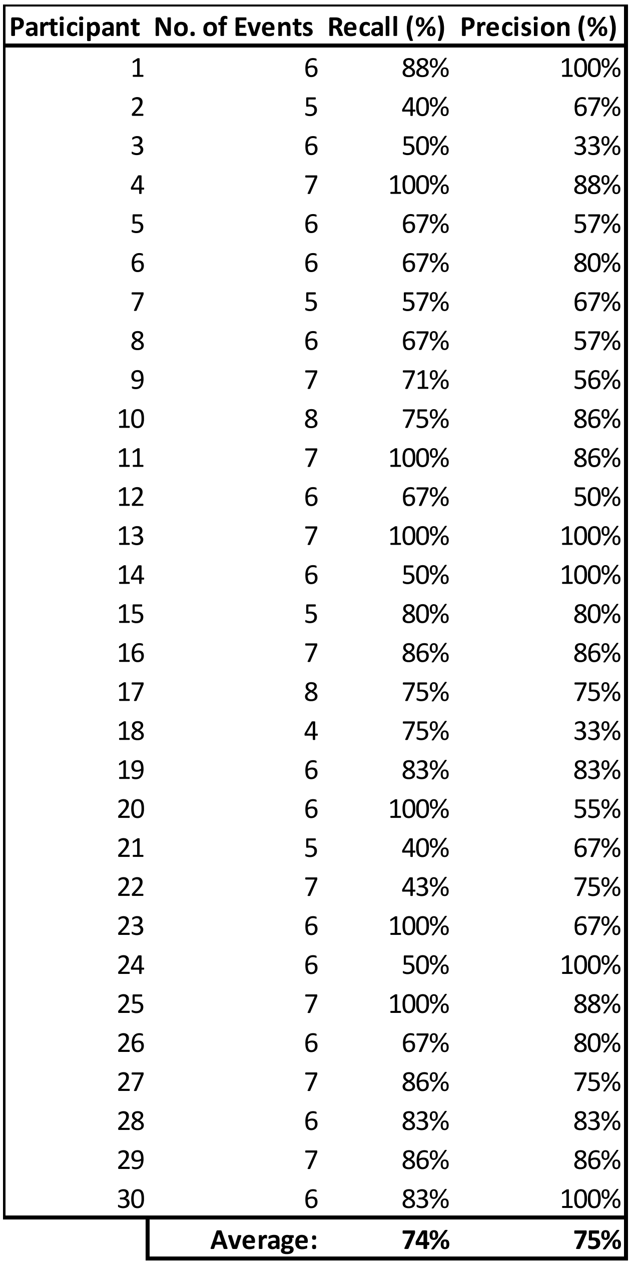
Figure 5.3. Results From Participant 24 - an example of 25% recall and 20% precision.

Figure 5.3. shows an example of a participant with poor recall and precision: 25% and 20% respectively; by these, measures, the worst performing of all participants in this validation set. Four of the five events detected by our method are false-positive, with one correct classification. It should be noted here that the event correctly classified was the long-form traffic surge; it appears so far in the investigation, that these events are more likely to be detected correctly than the short form events which we will discuss this later in the chapter. The four false positives are well outside of our latency windows, seemingly showing no relationship to the ground truth. Our reasoning for these events is again, down to potential events outside our pre-programmed EoI (discussed later). The presence of such a poor performing example shows a limitation of the methods used; being that if this particular *operator's* data was used to classify a data set, it would have no practical use.

5.2.2. Second Validation Set

The second validation data poses a slightly different set of EoI, in that each participant faces the same set up of the simulator, but actions by the participant affect the timing of certain events. Namely, the event *non-responsive aircraft* has a flexible timing between participants, as the definition of the ground truth is based on when it was deemed the participant *noticed* the non-responsive aircraft, rather than when the aircraft appeared as it is impossible to determine if it is responsive until the participant attempts to “contact” it, details of this labelling procedure were laid out in section 0. Our 10-minute section of data contains between four and eight events depending on the manner in which the participant played the simulator.

Table 5.2. Results of Classifier on Pupil Data From Second Validation Set.



Participant	No. of Events	Recall (%)	Precision (%)
1	6	88%	100%
2	5	40%	67%
3	6	50%	33%
4	7	100%	88%
5	6	67%	57%
6	6	67%	80%
7	5	57%	67%
8	6	67%	57%
9	7	71%	56%
10	8	75%	86%
11	7	100%	86%
12	6	67%	50%
13	7	100%	100%
14	6	50%	100%
15	5	80%	80%
16	7	86%	86%
17	8	75%	75%
18	4	75%	33%
19	6	83%	83%
20	6	100%	55%
21	5	40%	67%
22	7	43%	75%
23	6	100%	67%
24	6	50%	100%
25	7	100%	88%
26	6	67%	80%
27	7	86%	75%
28	6	83%	83%
29	7	86%	86%
30	6	83%	100%
	Average:	**74%**	**75%**

For example, if a user never interacts with a non-contact aircraft, it will not be classed as an event in ground truth, therefore reducing the total number of events experienced. In table 5.2. we see the classification results of the pupil data from the second validation set. The second set shows an average recall of 74%, a 7% improvement on the first validation set. We also note a more significant 18% improvement in precision, from 57% to 75%. For this dataset, the pupil classification shows 90% of participants classifying at least 50% of the EoI.

As we are currently assessing the results of individuals, it is also notable that the results of participant 13 had perfect recall and precision for both dataset 1 and 2. The results from participant can be seen in Figure 5.4.

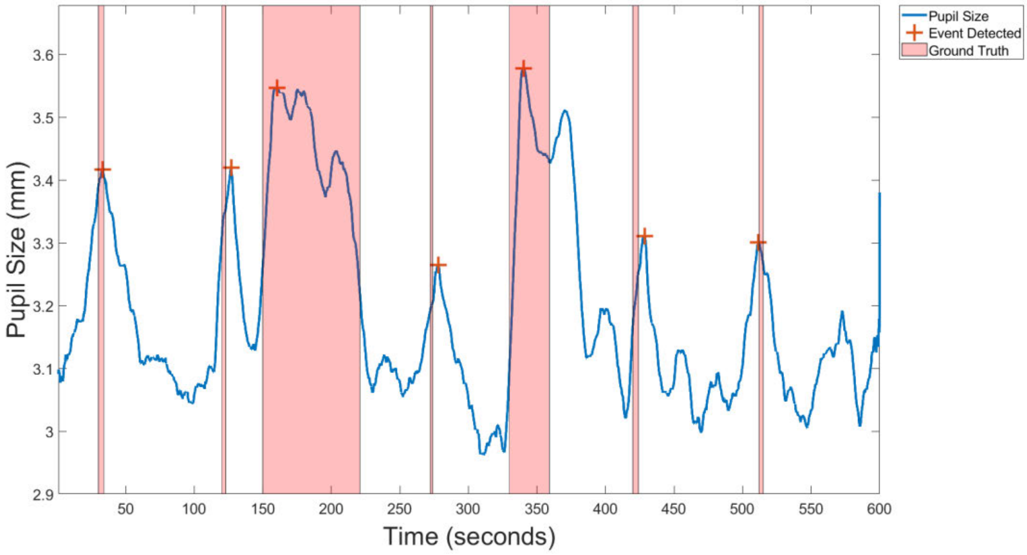


Figure 5.4. Classified pupil data for participant 13 from validation set 2.

Given the varying precision and recall of the various participants, this provides some evidence to suggest that the method is more effective with certain individuals than others. There are innumerable inter-individual factors that can affect the psychophysiological response of different people. It requires a certain combination of factors to select an operator that will be best suited to this application, an area for further work. It could be a simple case of the participant being more proficient at the task than others, assuming the only taxing element of the task for them being the EoI, whereas others could be applying more cognitive resources to the *normal* aspects of the simulation. As with the first validation set, we will also examine a representative choice of participants, an average example and a poor example.

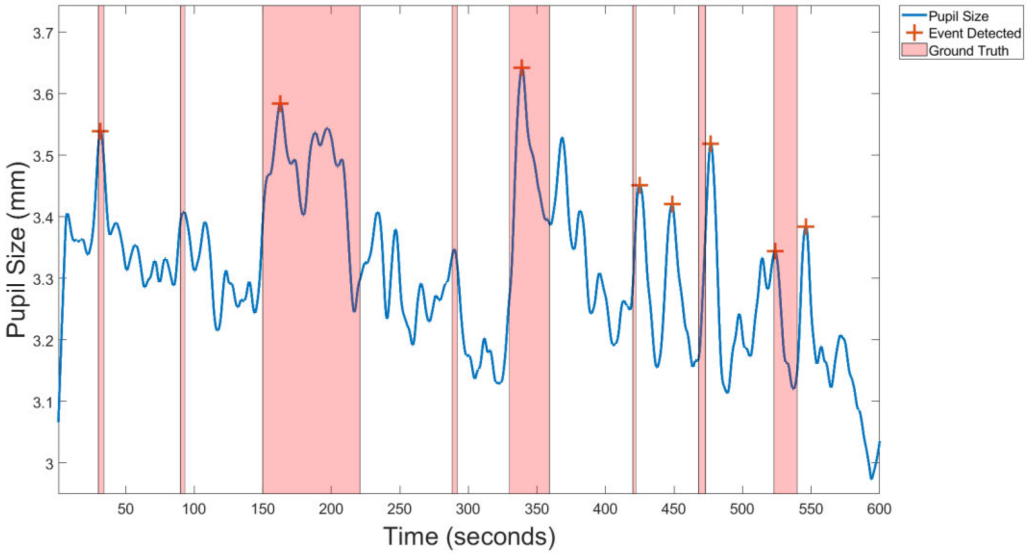


Figure 5.5. Classified pupil data from participant 17.

In Figure 5.5. we see data from participant 17, with recall and precision of 75%. Here, as in the first data set, we see significant and sustained pupil dilation for the long-form events of traffic surge and cloud cover (red ground truth bars 3 & 5), which are correctly classified. We also note that the short-form speeding aircraft and non-contact aircraft (red ground truth bars 2 & 4), are false-negatives, yet we do see a pupil response for these events. A lower threshold of peak prominence will have correctly classified these events, however, the threshold is set from the training data. We also note a false-positive classification at ~450 seconds; assuming response is correlated with events in the data, the event was not annotated as one of our ground truth EoI's, we will discuss this further in the chapter.

The challenge posed by the ground truth of the non-contact aircraft becomes apparent towards the end of this data; the complex event appears to have stimulated two peaks of response, one peak for the moment when the participant realised the aircraft did not respond to first contact and a second for when they realised it would not respond to a second attempt. This complex response yielded two peaks, but only one event, in a practical application, it would be recognised that both these peaks referred to the same event, yet this method marks it as a misclassification, reducing the measured precision.

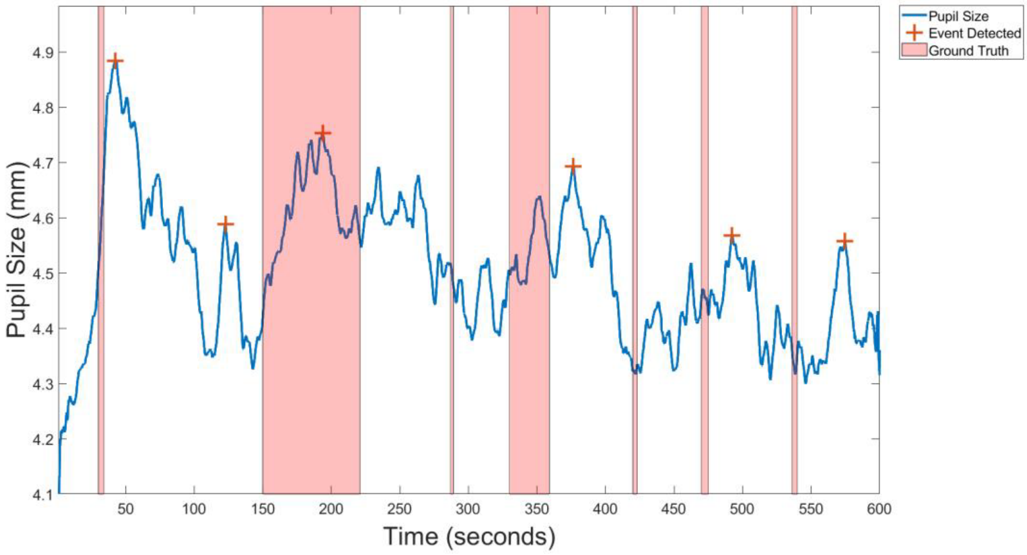


Figure 5.6. Classified pupil data for participant 7 from validation set 2.

The data in Figure 5.6. is from participant 7, with recall of 57% and precision of 67%. We note that, although only the first was correctly classified, both of our long-form events demonstrate a sustained pupil dilation increase. On multiple occasions for this participant we note a significant latency between event ground truth and peaks in pupil dilation; again multiple factors mean that it is not possible to state that these peaks are related to preceding events.

5.2.3. Results as a Population

In this section we will examine the pupil results as a population, averaging over all participants. In an applied example, it is unlikely that population-level results will have any practical use, short of some very specific examples: these being situations in which multiple operators were working on the same incoming data or in control room situations that deal with highly regular temporal events. For example, air traffic control has a schedule that includes times of high traffic at regular intervals.

Typically, psychophysiological data is examined over populations over multiple trials, this is mainly for characterising the responses of human outputs and referring them back to a psychological process from a controlled stimulus. In practise, as we have seen from our exploratory discussion of the pupil data in the previous sections, on an individual basis, there is potential for individuals to not respond to particular events in the same fashion as other individuals.

Given in our first validation dataset, the events of interest all occur at the same time across all participants, we examine the average pupil dilation data over all participants. We use the same analysis methods as applied to the individual participants, using the training data to optimise the parameters for the classifier.

The pupil data is first normalised, a baseline pupil value was established for each participant by taking an average of the measurements within the first 15 seconds (in which no stimulus is presented, this is our *resting* pupil diameter), this baseline was then subtracted from all values for each participant. An average was then obtained by taking the mean value for each sample of data across the 30 participants. The result of which can be seen in Figure 5.7.

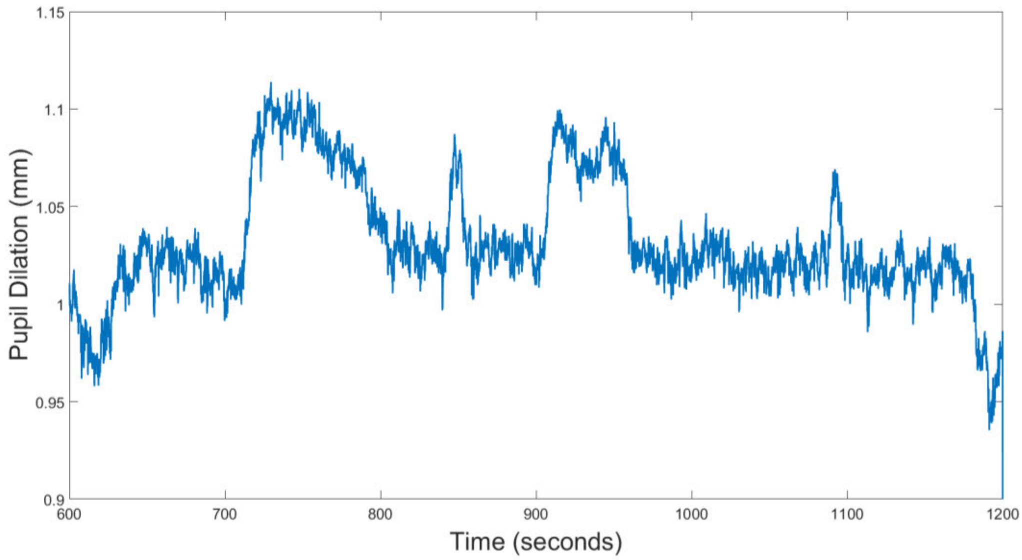


Figure 5.7. Averaged Pupil Data Across 30 participants.

We then classify the data using our pupil data method, the results of which can be seen in Figure 5.8.

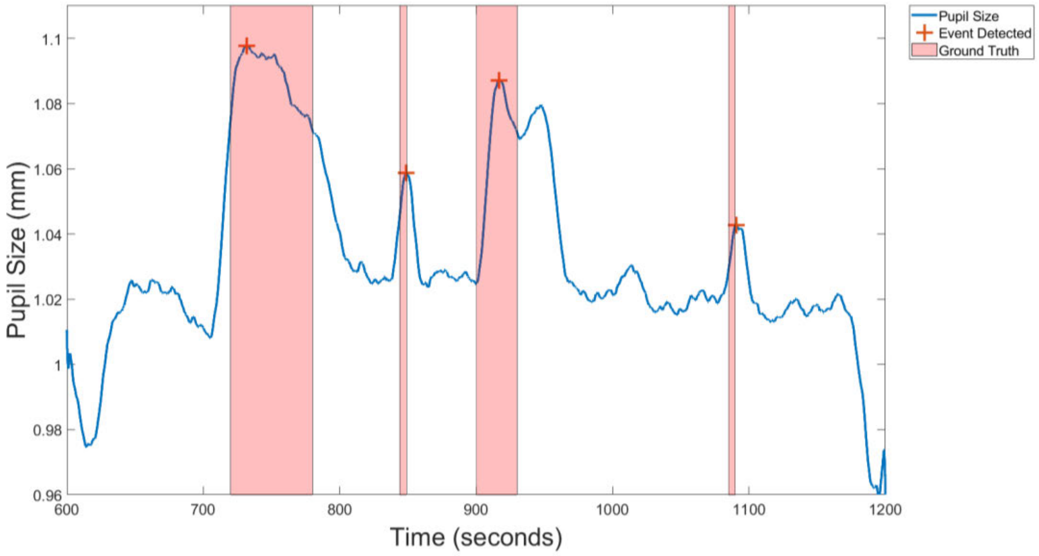


Figure 5.8. Classified pupil data across population of 30 participants.

The results for the population were recall and precision of 100%. The graphical representation in Figure 5.8. shows clear responses for each of our events of interest. It can be said that averaging the results across the participants improves the classification recall and precision.

5.2.4. Event Type Analysis

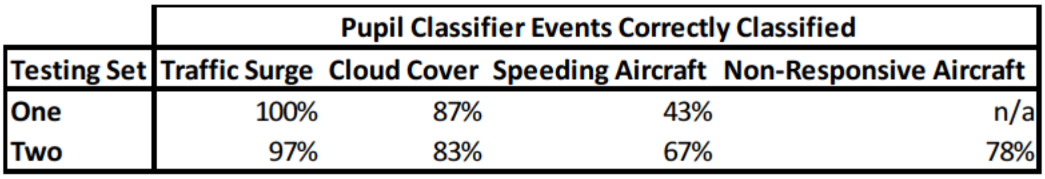
We have discussed the average classification results for the pupil classifier, we will now discuss the results of the classifier by event type. As the types of event vary in nature and in length, we will examine if the classifier is more sensitive to certain types of event.

In our two validation data sets we have four different types of event:

1. Traffic surge;
2. Cloud cover;
3. Speeding aircraft; and
4. Non-responsive aircraft.

We can split these events into two main types: long-form and short-form. The traffic surge and cloud over are our long-form events which last 1 minute and 45 seconds on average respectively. Speeding aircraft and non-responsive aircraft events are our short-form event, lasting on average 6-12 seconds each. The nature of each type of event is also different, as outlined in the methodology chapter. Table 5.3. show the results broken down by event type.

Table 5.3. Classification results for pupil data broken down by event type.



	Pupil Classifier Events Correctly Classified			
Testing Set	Traffic Surge	Cloud Cover	Speeding Aircraft	Non-Responsive Aircraft
One	100%	87%	43%	n/a
Two	97%	83%	67%	78%

Our results show that across both validation sets that the classifier performs better on the longer-from events than the short.

Traffic Surge Event

The traffic surge event, which represents a longer period of increased difficulty, has the best classification performance – given this event is most likely to stimulate a surge in cognitive load for a sustained period, this result is to be expected. We see that the classification performance carries over to the more complex second validation set, showing the robustness of the pupil classifier for this type of event.

We noted in our exploratory examples and the population average results the significant response peaks associated with the traffic surge event. This ubiquitous response was expected in the first validation set; given the structure of the event sequence ensured the event was followed and preceded by periods of normality. The results from the second validation set show that this response can be expected even when the event can precede or follow other events of interest by a comparatively narrow margin that may have affected the nature of the pupil response, given its latency and the more complex nature of the second validation set.

Cloud Cover Event

The cloud cover event has the second best classification performance for the pupil data. This event, like the traffic surge, represents a cognitive load stimulus sustained over a longer period of time compared to the short-form events. This period of time is however, slightly shorter than that of the traffic surge (~15 seconds shorter). This shortened time may be responsible for the slight (~13.5%) drop in classification performance when compared to the traffic surge event, signifying that there is a relationship between the length of the EoI and the likelihood of classification.

The nature of the event is also different to that of the traffic surge. The traffic surge immediately increases the difficulty of the task and sustains that change of difficulty for the duration of the event. The cloud cover event however, provides an obstacle that requires the participant to change their strategy. The cognitive load increase for this task results mainly from the realisation that the strategy requires changing, then implementing the change; once this has been done, the difficulty of the event tails off.

This of course also depends on the ability of the participant to recognise and achieve this, which accounts for the inter-individual differences in the classification performance, which will, always pose an obstacle for psychophysiological measures to classify events. We also see a slight drop in classification performance between validation sets 1 and 2, which is attributable to the more complex and less regular nature of the second validation set.

Speeding Aircraft Event

The speeding aircraft is the first of our two short-form events. Lasting on average ~10 seconds, it is expected to quickly spike an increase in the user's cognitive load as they attempt to handle the event before an aircraft collision occurs. This event, though short, does break the normal behaviour of the simulator space. It is expected that the participant is to quickly comprehend the event, and make inputs to mitigate any problems that could occur. We note a significant drop in classification performance for the speeding aircraft event to 43%. There are notable individual examples that responded very clearly to the speeding event as we saw in the exploratory discussion, who achieved 100% recall and precision.

The event being short in this fashion may lead to it not being classified for the following reasons; the response stimulated from the event is not significant enough to be separated from noise in the signal, in which case, the optimisation from the training data will have increased the classification threshold above these responses to improve the precision of the results rather than including all responses of this size. Another reason may be that the event was simply too short to stimulate a response at all, which we noted in a few examples. The reason for this could be down to the cognition process of the individual or that the participant did not deem the event to be different enough from the normal procedure to yield a response.

We do see however, in the population average, that both speeding aircraft events are correctly classified and visually show significant responses above noise. This favours our first explanation of the poor performance by demonstrating that, on average, participants did respond with sufficient amplitude to differentiate the average to a clear response, just not enough on an individual basis.

Non-Responsive Aircraft

The non-responsive aircraft is the second short-form event, the event represents a significantly more complex cognitive event that is difficult to handle and also difficult to manually identify after the fact from video analysis. The event manifests as an aircraft entering the screen that visually looks no different from any other aircraft object. The only difference is that it cannot be controlled by the participant, a fact they are unaware of until they attempt to move it. This unyielding aircraft continues to move across the screen at the same speed and heading until it exits on the other side of the screen. The participant must attempt to move the aircraft and realise that it will not respond to their input, at which point they must make provisions to ensure the aircraft does not collide with other aircrafts currently on screen.

The moment at which this realisation takes place is expected to be the moment at which the cognitive load of the participant begins to rise, culminating when the participant has made necessary operational changes to compensate for the unresponsive aircraft. This increase in event complexity may be the reason for its higher classification performance (78%). The time taken to formulate a solution to the problem posed by the event may be longer, this may provide the sustained period of cognitive load increase required to consistently increase the response of the pupil beyond that of noise.

5.2.5. Summary

We have presented results in this subsection to address our research question: can pupil dilation data be used to identify EoI in an operator scenario?

The general performance of our pupil data classifier shows a recall of 67% with a precision of 57% for our first dataset. The training data used to develop the pupil classifier came from the first half of the 20 minute first experiment. We expect, therefore that this will yield greater recall and precision results for the 1st validation set, and less so in the 2nd validation set. However, the more complex second validation set, that used training data from the first experiment, outperformed the results for the first validation set in both recall and precision, significantly so in precision; yielding an 18% increase in correct event detection.

Once averaged, the results from the first validation set are very clear; demonstrating 100% recall and precision. Showing that the average pupil responses across a population are clearly capable of identifying events of interest in complex control datasets. As discussed before however, in practise, a population experiencing an identical set of events in the temporal domain is unlikely. Across the averages of the participants however, there does exist a set of individuals whose data yielded perfect or very high precision and recall and there also existed those that pulled the average further down. Future work perhaps exists in the identification and experimentation with those individuals whose pupil responses appear to be sensitive to abnormal events in complex data, then using their results as a training set. We note that these high-performing individuals exist across both datasets when regarding pupil data classification.

On an individual event basis, we note that there is a skew toward correct classification of longer-form events, the reasons for which we discussed above. We see an average correct classification rate of 92% for long-form events compared to 63% average for short-form events. This potentially rules out this method for identifying events of a short-form nature, depending on the acceptable false-positive and false-negative rates for a particular application. Further work on this could involve using more classification features to better identify the short term events, using the breadth of the response as well as the prominence to delineate between short and long-form event types.

In conclusion, the answer to our research question is that pupil data classification shows promise as a classifier for retrieving events of interest from complex control task datasets. Specifically, when certain individuals are used as the operator and long-form events are the events of interest, it can be said that according to these findings, pupil data can be used as a classifier to retrieve these events reliably.

5.3. Research Question 2: *Can HRV data be used to identify the times at which events of interest occur in an operator scenario?*

As discuss in the previous section, we will now examine the results of classifying the HRV data on both our validation sets. We will first examine the classification results from the first validation set, examine a cross-section of individual participant classification performance and then examine the second validation set in similar fashion. We will then examine the performance of the classifier with respect to different event types. Finally, we will assess the data over the population rather than individuals.

5.3.1. First Validation Set

Our first validation set results are presented in Table 5.4., with average recall and precision of 64% and 53%, these are very similar to the results of the pupil classifier.

Table 5.4. Classification Results For RMSSD HRV Metric For Validation Set 1.

RMSSD

Participant	Recall (%)	Precision (%)
1	75%	100%
2	50%	50%
3	75%	60%
4	25%	50%
5	50%	50%
6	100%	80%
7	50%	40%
8	50%	33%
9	75%	60%
10	75%	50%
11	75%	43%
12	75%	33%
13	100%	67%
14	50%	50%
15	75%	75%
16	75%	38%
17	75%	38%
18	50%	50%
19	50%	67%
20	75%	100%
21	0%	0%
22	75%	50%
23	75%	50%
24	50%	50%
25	100%	57%
26	100%	67%
27	50%	50%
28	50%	50%
29	75%	60%
30	25%	20%
Average:	**64%**	**53%**

We will now examine a graphical result of an individual results:

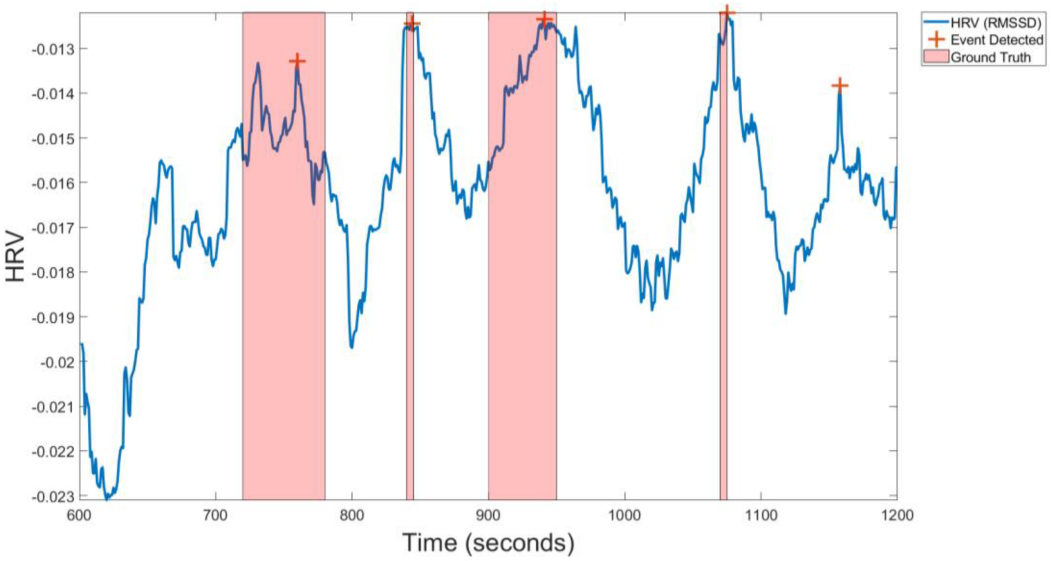


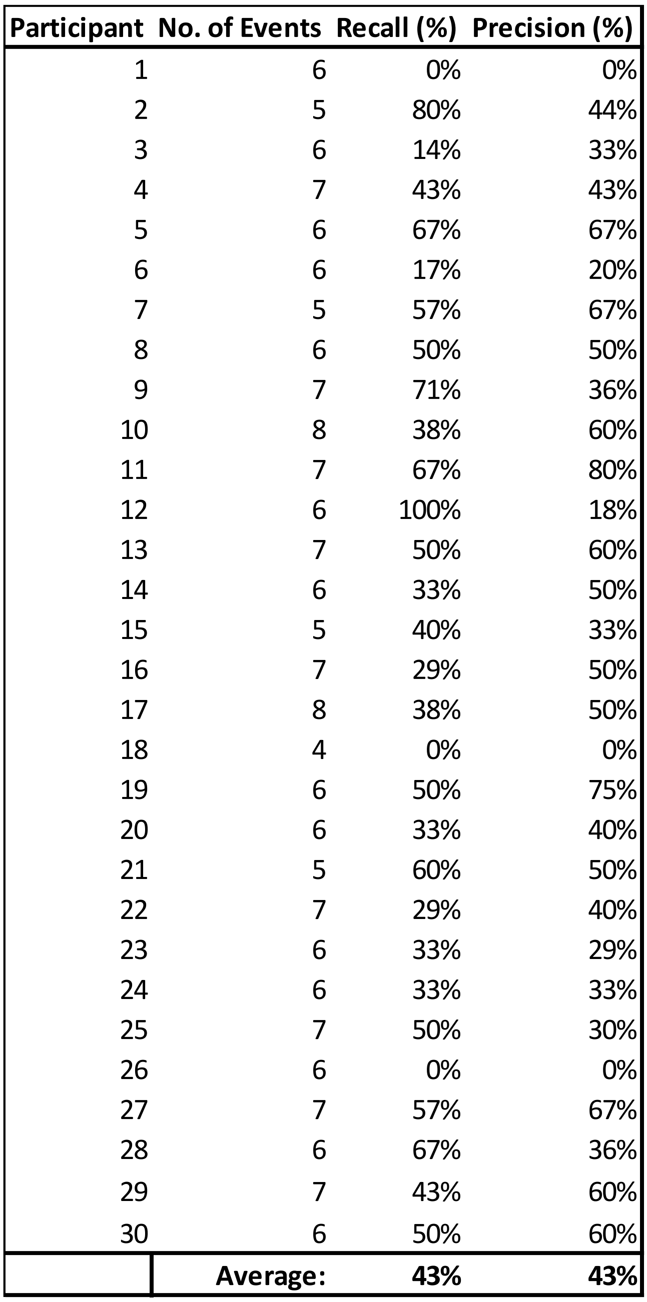
Figure 5.9. Classified HRV data from participant 6 showing perfect recall and 80% precision.

As with the pupil data, we note in Figure 5.9. that the long-form (1st and 3rd red bars) events stimulate a wider peak response than the short-form (2nd an 3rd red bars) events. We also note a false positive result at toward the end of the data – the reason for which may have been stimulated by a non-programmed event.

5.3.2. Second Validation Set

As discuss previously, the second validation set contains a greater number of EoI and also events that are more complex to identify. Given the work of the literature, we hypothesise that the slow reactive rate of HRV measures would be less precise when a denser set of events are presented in the same time frame. The classification results from the second validation set are presented in Table 5.5.

Table 5.5. Classification Results from HRV data for data set 2.



Participant	No. of Events	Recall (%)	Precision (%)
1	6	0%	0%
2	5	80%	44%
3	6	14%	33%
4	7	43%	43%
5	6	67%	67%
6	6	17%	20%
7	5	57%	67%
8	6	50%	50%
9	7	71%	36%
10	8	38%	60%
11	7	67%	80%
12	6	100%	18%
13	7	50%	60%
14	6	33%	50%
15	5	40%	33%
16	7	29%	50%
17	8	38%	50%
18	4	0%	0%
19	6	50%	75%
20	6	33%	40%
21	5	60%	50%
22	7	29%	40%
23	6	33%	29%
24	6	33%	33%
25	7	50%	30%
26	6	0%	0%
27	7	57%	67%
28	6	67%	36%
29	7	43%	60%
30	6	50%	60%
	Average:	**43%**	**43%**

As we hypothesised, both the recall and precision of the results are less than the first data set.

This is most likely due to the increased number of events in the same time frame as dataset 1.

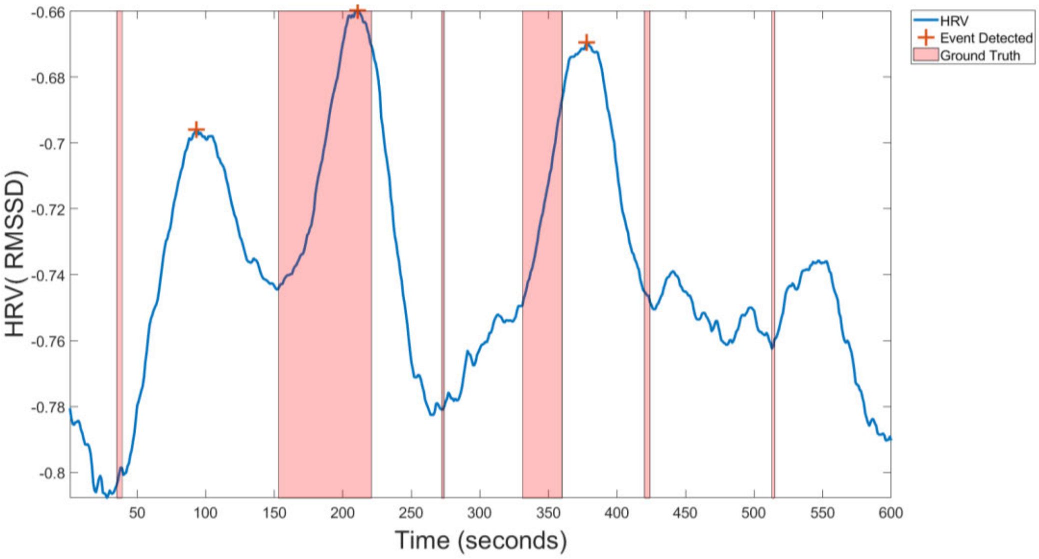


Figure 5.10. Classified HRV data for participant 8 from the second validation set.

We see in Figure 5.10. an example of a participants data, the 2 long-form events (red bars 2 & 4) show clear responses in the HRV data, yet the short-form events appear to have no significant response in the HRV data.

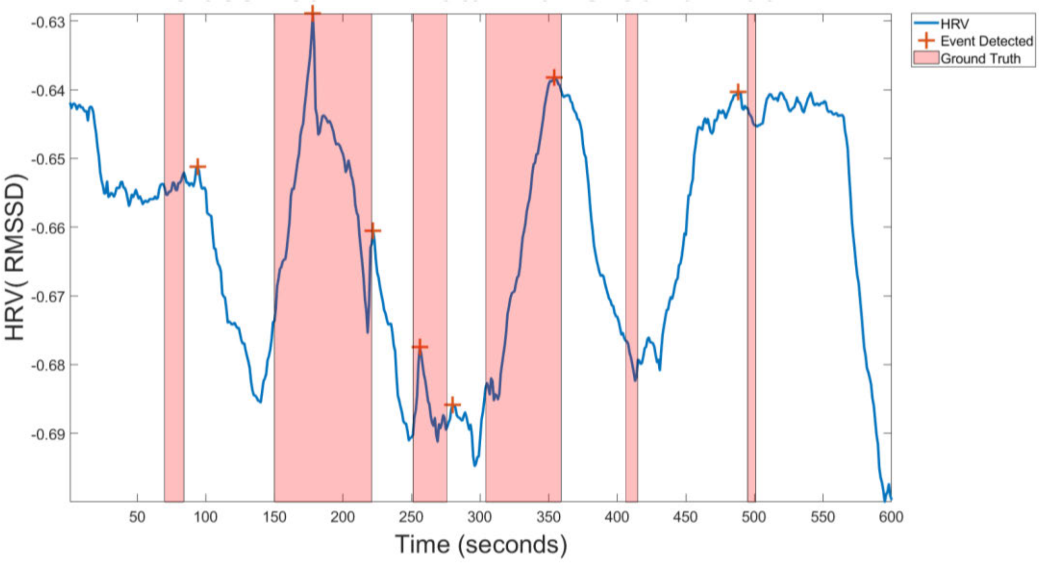


Figure 5.11. Classified HRV data for participant 11 from the second validation set.

In the example result shown in Figure 5.11 however, we see a correctly classified response to a short-term (non-responsive aircraft) event (3rd red bar). This particular example however, shows that the participant took longer to recognise that this was a non-contact aircraft, thus increasing the length of the ground truth of the event. When compared against the length of the same event in Figure 5.10., we see that the event is longer in the second example, perhaps this increased length of event is enough to trigger a response that can be classified.

This does appear to favour a conclusion that this measure is simply not appropriate for identifying short term events consistently which we shall discuss in the next section. It should also be noted that the two long-form events in Figure 5.11. trigger a significant response and are correctly classified. We will examine the sensitivity of this metric to event types in a further section.

5.3.3. Results as a Population

As done for the pupil data, we will now examine the classification results when the HRV data is averaged over the population. The same averaging procedure is applied; the results of which can be seen in Figure 5.12.

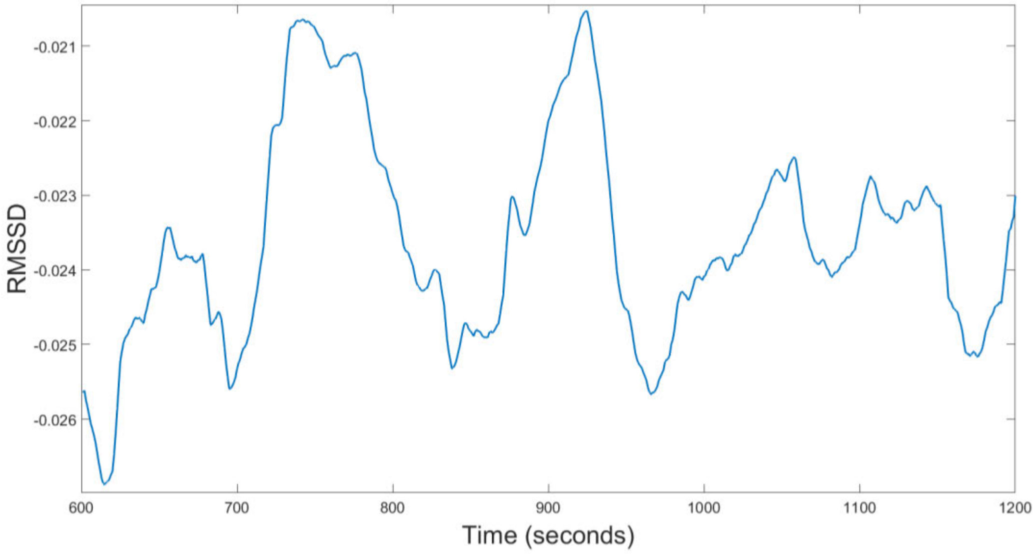


Figure 5.12. Averaged HRV Data over population of 30 participants for validation set 1.

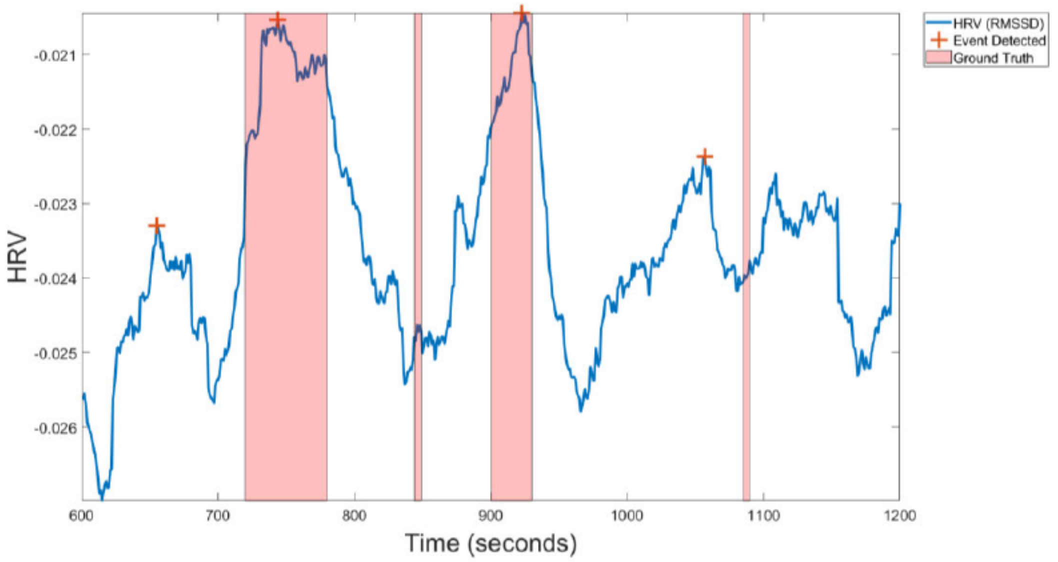


Figure 5.13. Classified HRV data across population of 30 participants for validation set 1.

The classification results of the averaged HRV data can be seen in Figure 3.13. The average of the population appears to have the same characteristic large responses to the long-form events (red bars 1 & 3) as the individual cases we examined. This being said, the classification yielded only 50% recall and precision, with 2 very distinctive false-positives.

5.3.4. Event Type Analysis

As done for our pupil classifier, we will now examine the classifier's performance with regards to each event type. The results of which can be seen in Table 5.6.

Table 5.6. HRV Classifier results across individual event types for validation set 1 & 2.

	HRV Classifier			
Testing Set	**Traffic Surge**	**Cloud Cover**	**Speeding Aircraft**	**Non-Responsive Aircraft**
One	87%	90%	43%	n/a
Two	77%	73%	18%	28%

We see in these results, as we did in the pupil results, that a skew exists in favour of correctly classifying the long-form events. In 5.3.4. we discuss the specifics of the events in regards to our predictions regarding their potential cognitive response.

In these HRV results we see a clear relationship between classifier performance and long-form events. HRV typically is rarely used to examine stimuli that last less than 5 minutes, it is therefore intuitive that we see a drop off of classification performance for our short-form events (being roughly 10 seconds). We note that the performance is also decreased in the second set, given the second data sets increased complexity, this is to be expected.

Though the relationship of better performance for longer events holds in the second validation set, there is a more significant loss in classification performance for speeding aircrafts. This is most likely due to the denser event sequence, with HRV response requiring a longer recovery time between events; this makes a distinctive peak a short period of time less likely if a new event of interest starts soon after another.

5.4.5. Summary

We have presented results in this subsection to address our research question: can ECG data be used to identify EoI in an operator scenario? Our classification results from validation set 1 and 2 showed that the HRV classifier was capable of recall and precision of 64% and 53% respectively for the first validation set and 43% recall and precision for the second set. The HRV measure certainly shows less precision as a classifier than pupil dilation. In literature, the use of HRV has typically been on events lasting ~5mins and not for short term events such as the ones presented here.

Due to the slower response times of HRV, the measure may not be so responsive to short term events. This is confirmed when our data is averaged across the population of participants. We note that the long-form events yield a clear and precise response in HRV, yet the lack of precision in the smaller events is problematic for the averaged results as it produces more false-positives, whereas pupil response average increased the recall and precision to 100%. The results at this stage show the HRV measure as a useful metric for recalling the long-form events.

Given this, the results from the event-by-event analysis confirm the value of the measure at classifying long-form events, showing an average recall of 82% for long-form events and 30% recall for short-form events. These show the potential for HRV as a classifier for the long form events but provides strong evidence that the metric has no practical value when classifying short-term events as the low recall percentage is made less useful by the poor precision scores.

5.4. Research Question 3: *Can mouse movement data be used to identify the times at which events of interest occur in an operator scenario?*

In this section we present and discuss the results of the mouse metric classifiers on the first and second validation data sets. We will first examine the results across all participants for the first data set, examine some graphical examples and discuss the output. We will then examine the results for the second validation data set. The results for the first data set at population level shall then be presented and discussed. The results for the mouse metrics classifications sensitivity to certain events will then be discussed. Finally, we shall discuss the constellation of results we have obtained in order to answer our research question.

5.4.1. First Validation Set

We present the classification recall and precision results for all participants across validation set 1 in Table 5.7.

Table 5.7. Classification results across all participant for first validation set.

	Mouse Speed Metric		Mouse Gesture Metric	
Participant	**Recall (%)**	**Precision (%)**	**Recall (%)**	**Precision (%)**
1	50%	67%	50%	67%
2	75%	100%	100%	100%
3	100%	50%	75%	75%
4	75%	75%	50%	67%
5	75%	75%	50%	100%
6	75%	60%	25%	50%
7	100%	100%	50%	50%
8	100%	67%	50%	100%
9	75%	50%	50%	67%
10	75%	100%	75%	60%
11	75%	38%	50%	100%
12	75%	100%	50%	100%
13	75%	100%	25%	33%
14	50%	100%	25%	50%
15	75%	60%	25%	20%
16	25%	13%	25%	33%
17	100%	67%	50%	67%
18	75%	75%	50%	50%
19	100%	80%	50%	100%
20	75%	100%	50%	67%
21	100%	80%	50%	67%
22	75%	75%	50%	100%
23	50%	50%	50%	67%
24	75%	75%	50%	100%
25	75%	75%	75%	100%
26	75%	100%	50%	67%
27	100%	100%	50%	67%
28	50%	100%	50%	50%
29	75%	75%	50%	100%
30	50%	50%	50%	100%
Average:	**75%**	**75%**	**50%**	**72%**

Our hypothesis for the mouse speed metric is that EoI will stimulate faster input responses from the operator. We see an average recall and precision of 76% and 75% respectively for the mouse speed metric. We see that 80% of the participant's classifiers scored at least 75% recall, with two individuals scoring perfect recall and precision.

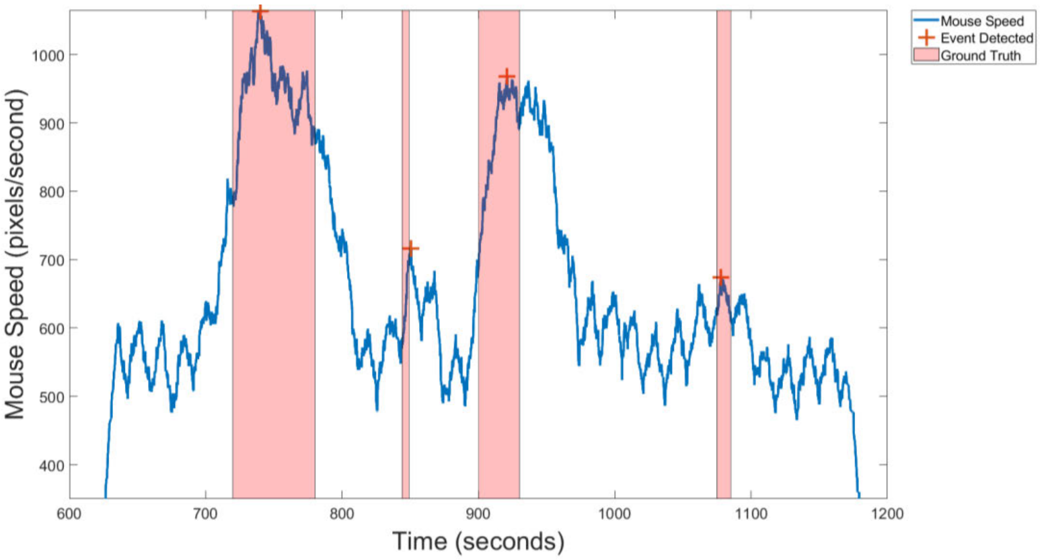


Figure 5.14. Classified mouse speed metric from participant 7 for validation dataset 1 showing perfect recall and precision

We see in Figure 5.14 the graphical display of the results from participant 7; showing perfect recall and precision. We also note here the significant response from the two long-form events (red bars 2 & 4), which, intuitively, agrees with the hypothesis in that a sustained period of abnormality would stimulate an increased period of faster mouse activity. This example represents the ideal result that fits our hypothesis.

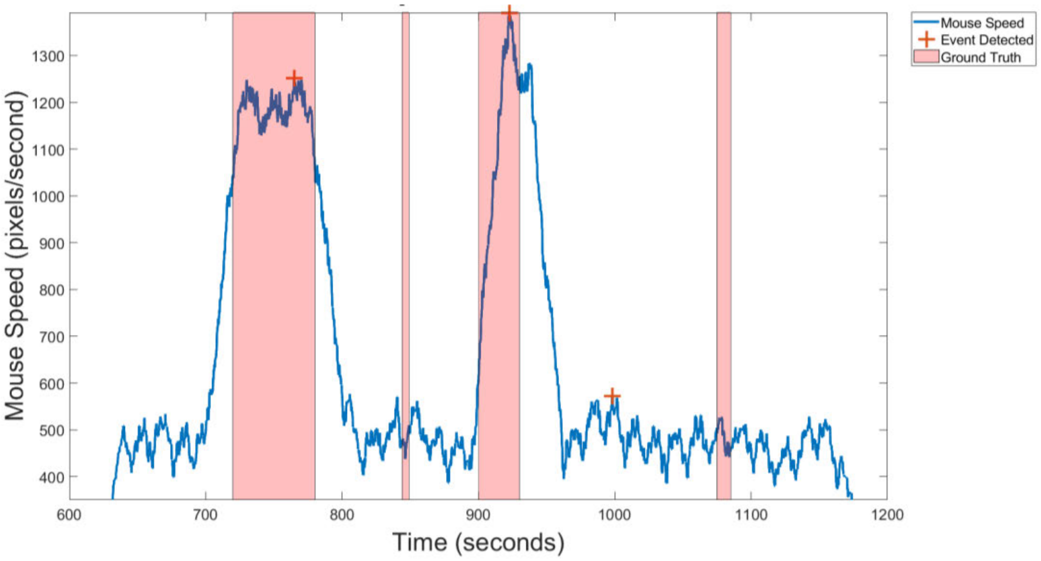


Figure 5.15. Classified mouse speed metric from participant 18 for validation dataset 1 showing recall and precision of 75%.

Figure 5.15. shows an example participant that scored 75% recall and precision. Here we note the same significant, correctly classified responses to the long-form events (red bars 2 & 4). The remaining events seem indistinguishable from noise however, with a slightly more prominent showing our false positive result. This example demonstrates that this metric appears more sensitive than long-form EoI – we shall discuss this later within the chapter.

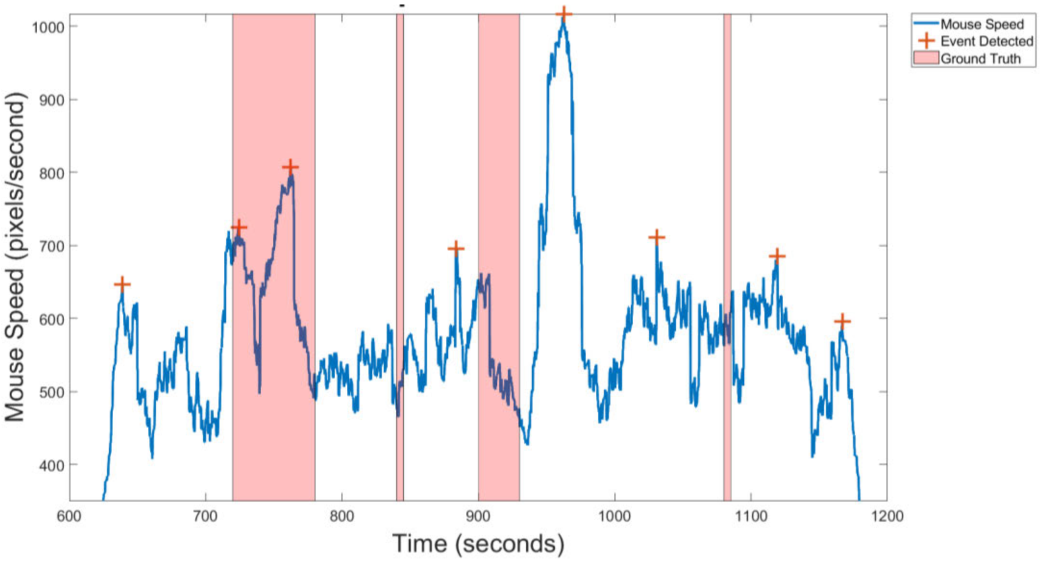


Figure 5.16. Classified mouse speed metric from participant 16 for validation dataset 1 showing recall and precision of 25% and 13% respectively.

Figure 5.16. shows a poor performing participant's data. Though we see two distinct periods of high mouse speed, only the first co-occurs with our long-form EoI with the second significant peak lagging significantly behind the stimulus EoI to be a false-positive. Five other false positives also have no relationship with other EoI.

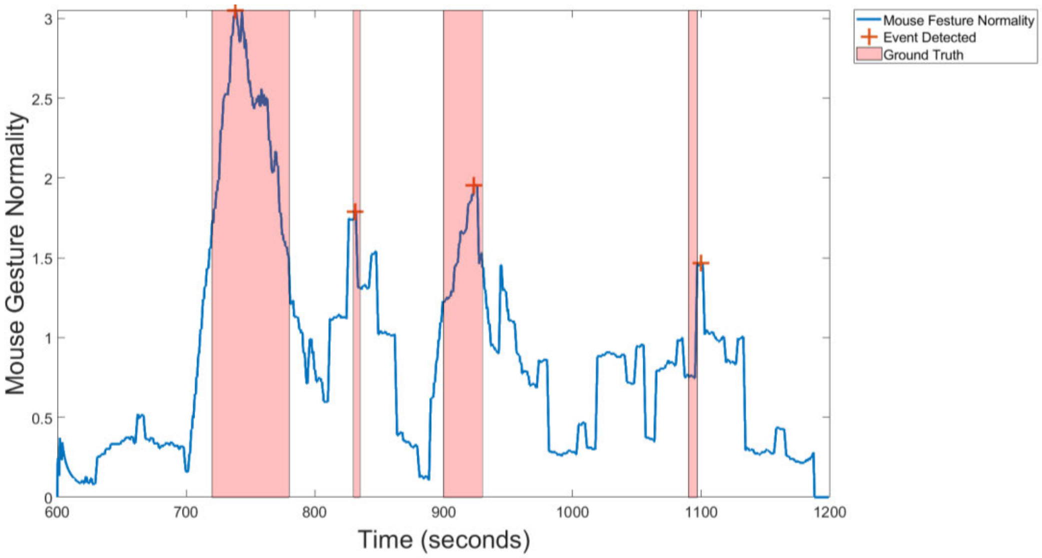


Figure 5.17. Classified mouse gesture normality metric for participant 2 showing perfect recall and precision.

In Figure 5.17. we see an example of perfect recall and precision for dataset one. This example represents the upper envelope of the performance of this metric when paired with an appropriate individual. We note in this metric also, that there is a relationship between the size of the response and the length of the EoI, with the long-form events (red bars 1 & 3) showing the widest response peaks. Though the only participant's data to show perfect recall and precision, it does show the potential for the measure to be an effective tool in finding EoI.

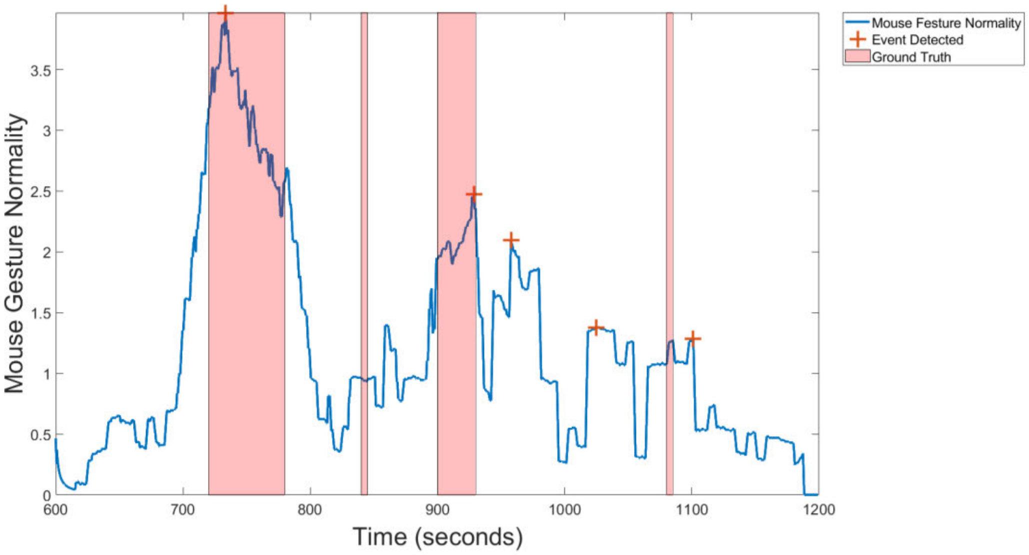


Figure 5.18. Classified mouse gesture normality metric for participant 18 showing 50 % recall and precision.

In Figure 5.18. we see an example of mouse gesture data that had 50% recall and precision. We note the two most significant responses co-occur with our 2 long form events (red bars 1 & 3), but in this example we see potential peaks relating to the short-form events (red bars 2 & 4). However, the first is both a false negative, sitting underneath the threshold obtained in the training set and also lagging behind the stimulus. The second short-form event appears to be in a peak of abnormal mouse gesture data, but the zenith lagging too far behind for a positive classification, but showing that with potentially a greater training set, that this participant's data fits with our hypothesis as well and our ideal example in Figure 5.17.

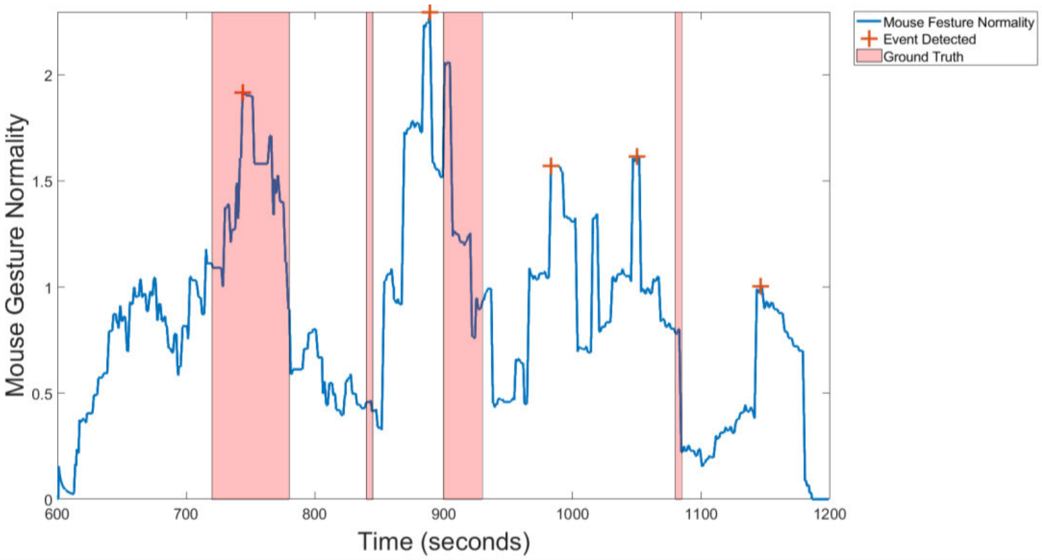


Figure 5.19. Classified mouse gesture normality metric for participant 16 showing 25 % recall and 33% precision.

In Figure 5.19 we see one of the poorer examples of the gesture normality metric in terms of classification performance, scoring 25% recall and 33% precision. We see again, that even in this poor example, two significant peaks near our long-form events. Though in this case the second large peak precedes the event, leading to a false-positive, the reason for this is unknown but will be discussed further in the chapter. The other 3 false-positives equally appear to have no relationship to the pre-programmed EoI.

5.2.2. Second Validation Set

Our second validation set contains more EoI in the ten-minute period. It also contains a differing variety and order of events. This increased complexity and density will provide a challenge for our mouse gesture normality metric. We present the results of the classifier in table 5.8.

Table 5.8. Classification results for mouse metrics across all participants in second validation set.

		Mouse Speed Metric		Mouse Gesture Metric	
Participant	**No. of Events**	**Recall (%)**	**Precision (%)**	**Recall (%)**	**Precision (%)**
1	6	67%	100%	33%	100%
2	5	40%	67%	60%	30%
3	7	71%	83%	43%	50%
4	7	71%	71%	71%	100%
5	6	83%	63%	50%	60%
6	5	60%	75%	60%	33%
7	7	71%	71%	29%	100%
8	6	100%	75%	50%	60%
9	7	43%	60%	43%	50%
10	8	75%	67%	13%	100%
11	6	100%	60%	33%	100%
12	6	83%	71%	33%	50%
13	6	67%	67%	83%	83%
14	6	83%	83%	33%	67%
15	5	60%	100%	60%	33%
16	7	86%	55%	43%	60%
17	8	88%	88%	50%	57%
18	4	50%	33%	75%	33%
19	6	67%	44%	33%	50%
20	6	83%	71%	17%	33%
21	5	60%	60%	40%	67%
22	7	43%	75%	29%	67%
23	7	57%	100%	67%	80%
24	6	17%	25%	17%	100%
25	6	83%	63%	50%	75%
26	7	71%	56%	29%	67%
27	7	86%	100%	43%	50%
28	6	83%	83%	33%	40%
29	7	71%	83%	43%	38%
30	6	67%	44%	33%	50%
	Average:	**70%**	**70%**	**43%**	**63%**

We can see from the results in Table 5.8 that there is a drop in performance for our mouse metrics when applied to the more complex second validation set. The speed metric shows 70% recall and precision compared to the 75% recorded for the first validation set and the gesture metric shows recall and precision of 43% and 63% respectively. Considering the added complexities involved in the second validation set, this could be anticipated; we will examine a cross-section of individuals and discuss the outcomes further.

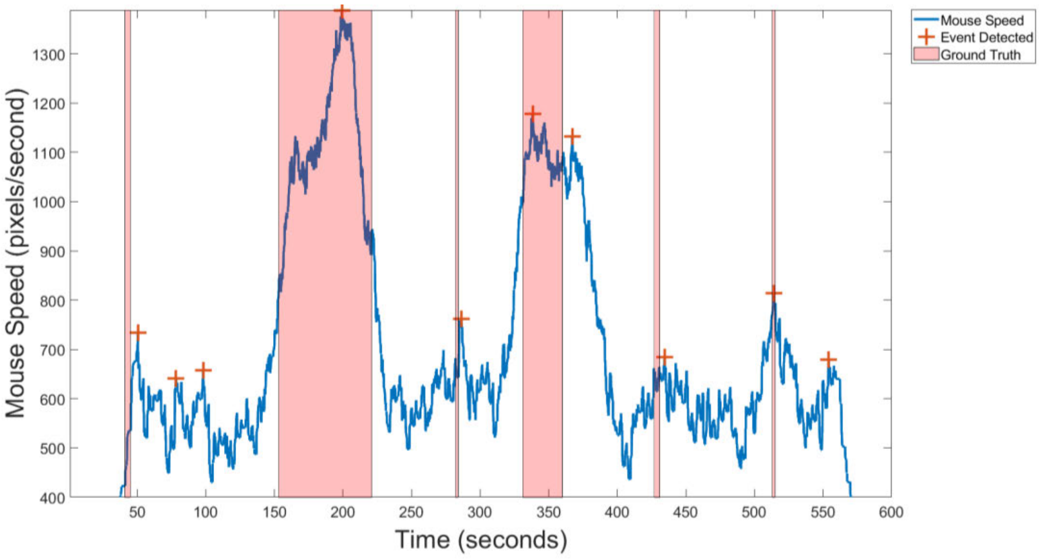


Figure 5.20. Classified mouse speed data from participant 8 for the second validation set showing perfect recall and 75% precision.

In the example in Figure 5.20. is from participant 8, with perfect recall and precision of 75%. This is our ideal participant from this metric in this data set. We see the characteristic large responses in speed for the long-form events (red bars 2 & 4). We also see all four of the short-form events correctly classified. If these peaks do relate to abnormal simulator situations, the false-positives are not abnormal events that were pre-programed into the simulator, again, we will discuss these later in the chapter.

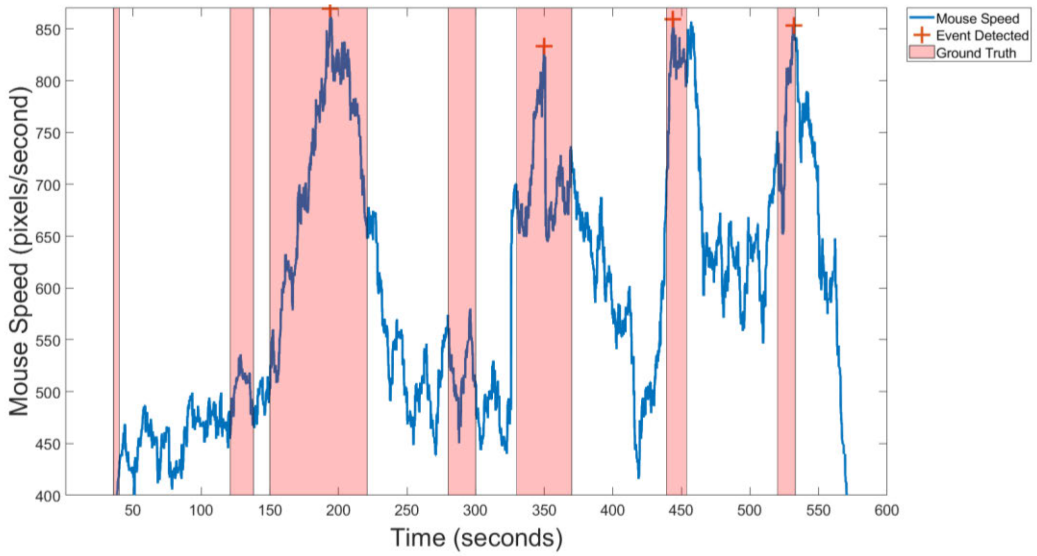


Figure 5.21. Classified mouse speed data from participant 23 for the second validation set, demonstrating 57% recall and perfect precision.

The example in Figure 5.21. shows a participant that had a greater number of EoI than the example in Figure 5.20, achieving recall of 57% and perfect precision. We note here the same characteristic larger responses for the long-form events (red bars 3 & 5). We also note for two of the short-form event (red bars 2 & 4) there is some noticeable response that would be classified with a lower threshold, an example of where potentially more training data would improve the results for this individual. This data demonstrates that there is potential for this metric to perform better than the recall results show.

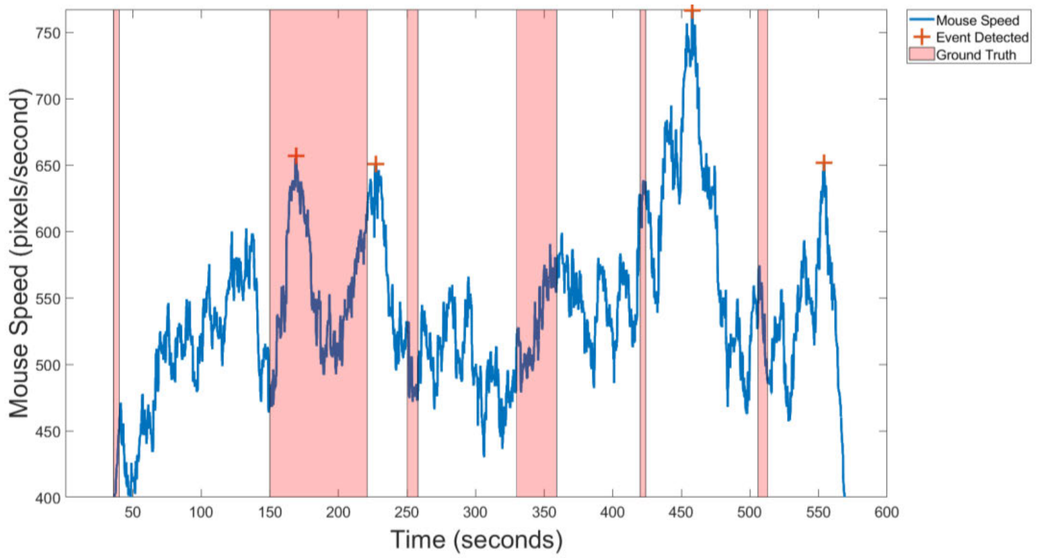


Figure 5.22. Classified mouse speed data from participant 24 for the second validation set, demonstrating 17% recall and perfect precision.

The example in Figure 5.22. is the poorest performing individual for this metric with recall of 57% and precision of 25%. This interesting example demonstrates again that the training data provided an inadequate threshold for responses; we can see responses for events 4, 5 and 6 that were false-negatives. We also note a lack of significant peaks around our long-form events. In this particular example, the non-contact aircraft (red bar 5) appears to have caused a significant and delayed response for the participant, a potential explanation may be that, in dealing with the non-contact aircraft, a significant queue of incoming aircraft built up and then had to be subsequently dealt with very quickly, stimulating the significant and delay response in this data. This raises potential further work on different methods for assessing methods for correct classifications, we shall discuss this in further sections.

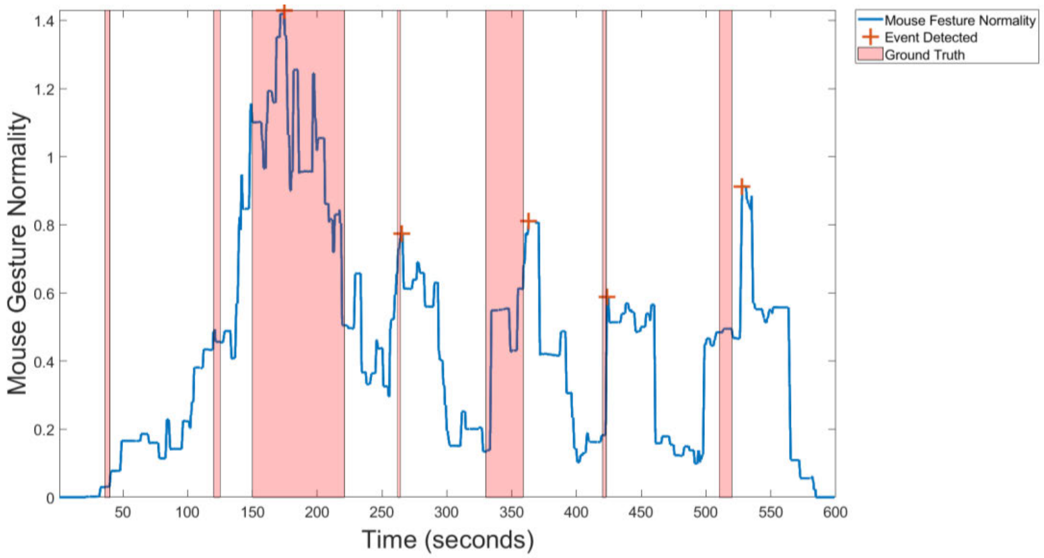


Figure 5.23. Classified mouse gesture data for participant 4 from validation set 2.

Our overall results for the mouse gesture normality metric showed a 7% decrease in recall and 11% reduction in precision than our first validation set. We see an ideal example from the second validation set in Figure 5.23, showing recall of 71% and perfect precision. We note the typical large response for the first long-form event (3rd red bar) but no such response for the second long-form event (5th red bar). We do otherwise see clear responses that co-occur with our EoI, with the exception of the first two events.

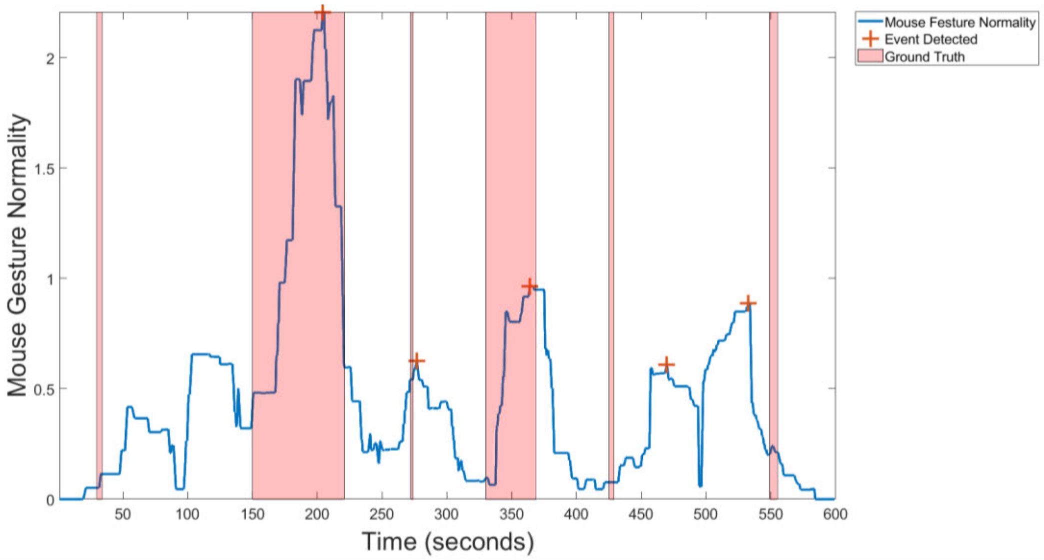


Figure 5.24. Classified mouse gesture data for participant 5 from validation set 2.

Figure 5.24. shows the data from participant 5, which scored a recall of 50% and precision of 60%. We note in this average example a significant response for the first long-form event (red bar 2), the second long-form event does yield a large response but not to the same extent as the first. We see no further relationship between the two false positives and the EoI.

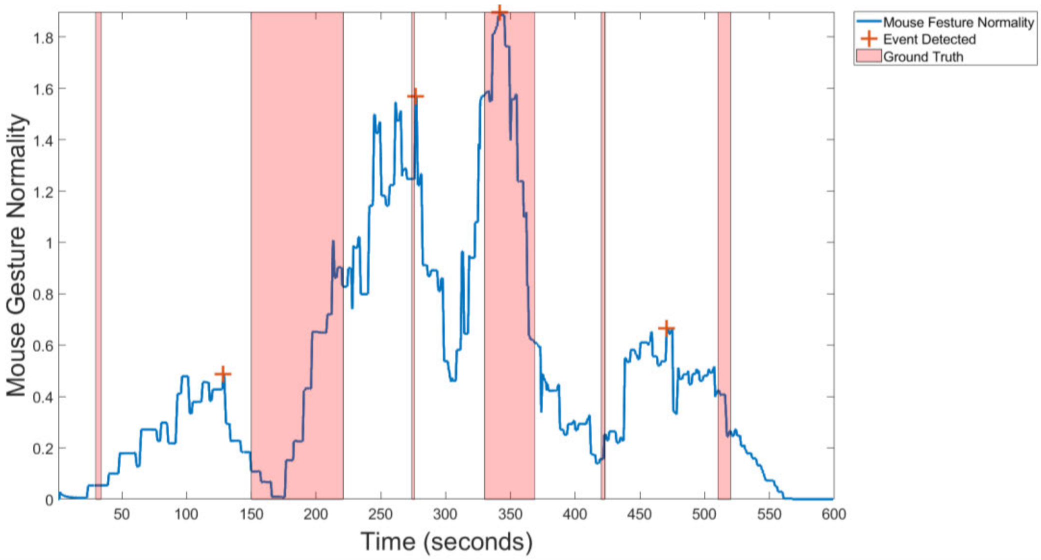


Figure 5.25. Classified mouse gesture data for participant 19 from validation set 2.

Figure 5.25. shows a poor performing example participant, with recall of 33% and precision of 50%. We note a significant response for the second long-form event (4th red bar). The remaining events show no relationship to the EoI save the short-form event (3rd red bar); this appears to be the summit of a large response that is lagging behind the first long-form event.

5.4.3. Results as a population

As with the pupil and HRV data, we will now examine the data as a population, first for the mouse speed metric, then for the gesture normality metric. For the speed metric, the data is per-sample averaged over the 30 participants (see Figure 5.26).

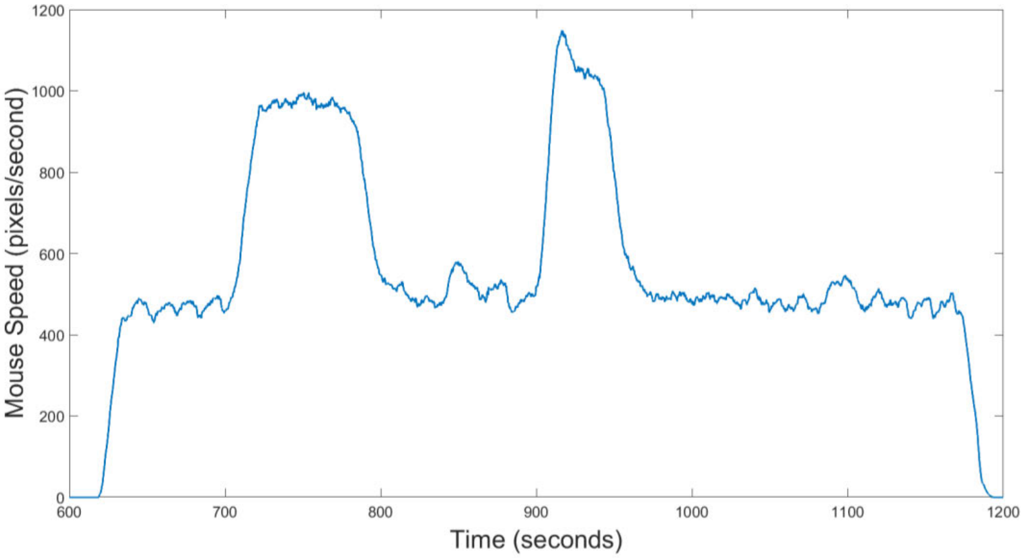


Figure 5.26. Mouse speed data averaged over 30 participants for validation set 1.

We can see the characteristic large response peaks are pronounced in the averaged data.

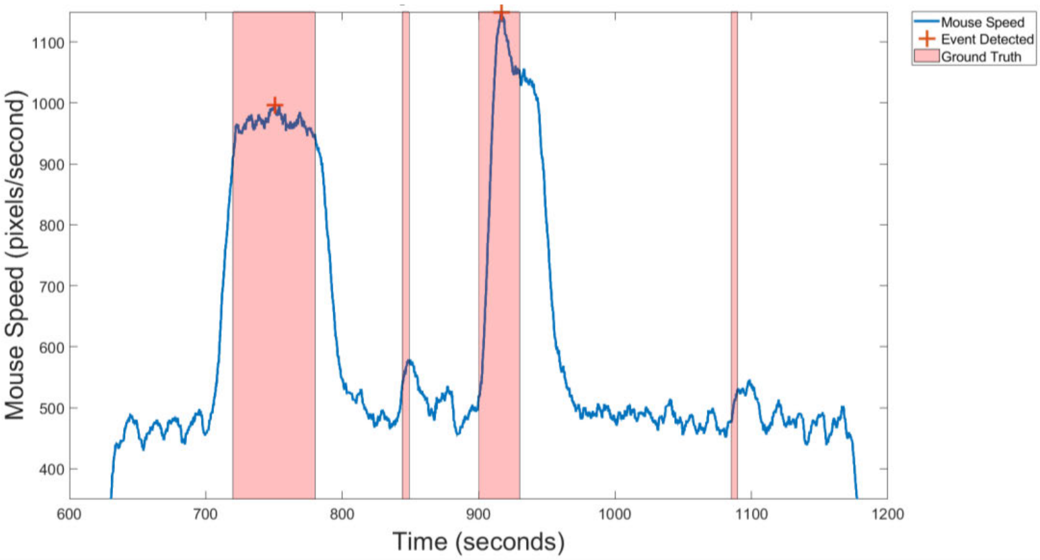


Figure 5.27. Classified Mouse Speed data across averaged data from 30 participants for validation set 1.

Figure 5.27. shows the classification results for the averaged data. Though we do see small peaks occurring at the short-form events, they are false-negatives and the averaged data yields a recall and precision of 50%. The metric at a population level is clearly more sensitive to the long-form events.

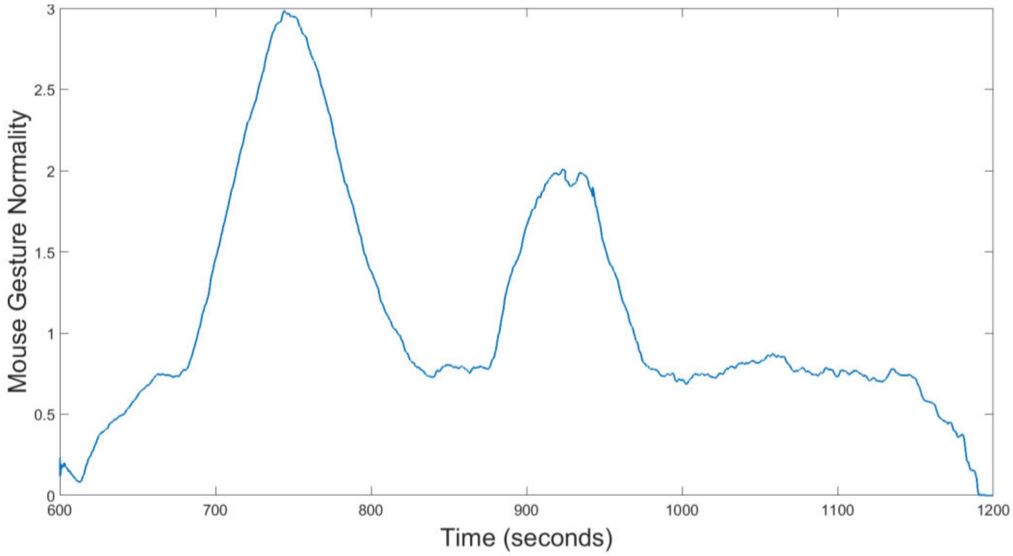


Figure 5.28. Averaged mouse gesture data over the population of 30 participants for validation data set 1.

In Figure 5.28 we see the mouse gesture data averaged over the 30 participants using the same, per sample mean approach.

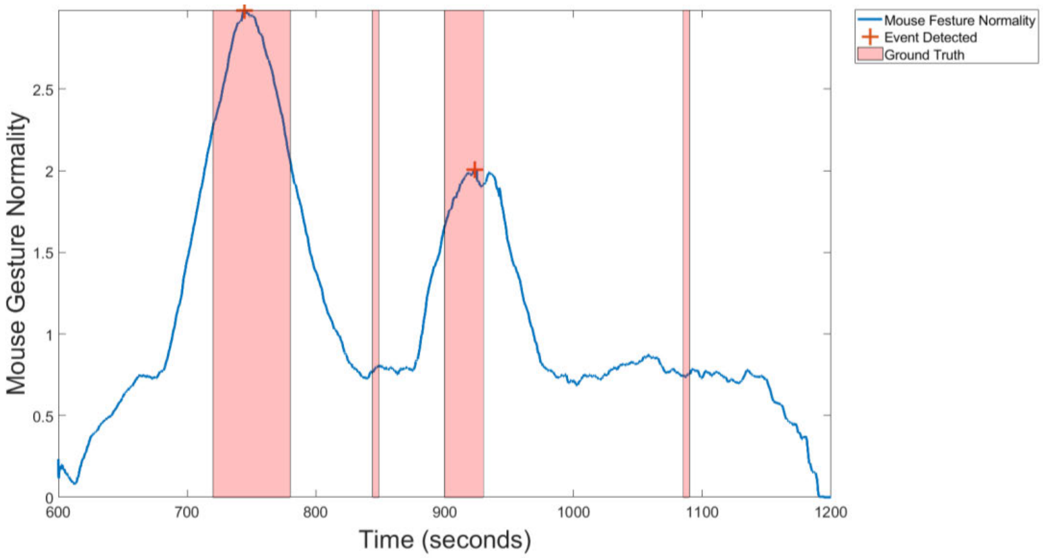


Figure 5.29. Classified mouse gesture data from population average over 30 participants from validation data set 1.

The results in Figure 5.29 show the mouse gesture normality metric yields recall and precision of 50% for the averaged population. The results here are very similar to the mouse speed metric average, showing that, on average, mouse-based metrics appear more sensitive to the long-form EoI.

5.4.4 Event Type Analysis

As with the pupil and HRV metrics, we will assess the individual event performance of our mouse metric classifiers. The results of which can be seen in Table 5.9.

Table 5.9. Mouse speed and mouse gesture classifier recall results across individual event types for validation set 1 & 2

Testing Set	Traffic Surge	Cloud Cover	Speeding Aircraft	Non-Responsive Aircraft	Mouse Speed Classifier
One	97%	97%	55%	n/a	
Two	100%	87%	42%	69%	
Testing Set	Traffic Surge	Cloud Cover	Speeding Aircraft	Non-Responsive Aircraft	Mouse Gesture Classifier
One	100%	83%	8%	n/a	
Two	67%	67%	20%	39%	

We note that, as with the other metrics, the mouse speed and gesture metrics have a far higher average recall percentage for long-form events.

Traffic Surge

We see that the recall percentage for this event is near 100% for the mouse speed metric. This fits intuitively with the nature of the event, as the constant influx of aircraft at a faster rate will require, on average, faster movements.

The recall percentage for the mouse gesture metric is also very high for the first validation set (100%), yet falls to 67% in the second validation set. One reason for this may be due to the greater number of events of interest and the more complex events in the second validation set, leading to a broader range of differing mouse gestures, some of which would not have been seen in the training data. It is likely that a more appropriate training set is required to improve the recall of this metric.

Cloud Cover

The recall for the speed metric for validation set one is 97% for the cloud cover event, this recall falls slightly for the second validation set, in which 87% is achieved. This result is also expected, as the long-form nature of the cloud cover event requires the participant to input mouse actions quickly to mitigate the obscuring effects of the cloud.

The mouse gesture recall for the first validation set is 83% falling to 67% for the second validation set. Again, for this long-form events, the higher recall results are expected, due to the sustained nature of abnormal gestures required to mitigate the cloud interference that occurs only once throughout the ten-minute simulator run, making the set of gestures associated with it relatively abnormal by comparison to those of normal procedure. Given the second validation set has a larger number of event of interest, it is reasonable to hypothesize that a greater number of abnormal gestures will be required, thus slightly reduce the extent to which the gestures during the cloud cover event in the second validation set are abnormal, slightly reducing the recall percentage.

Speeding Aircraft

We observe a reduction in recall percentage for the short-form events in the mouse speed metric. The operator is required to react quickly to a speeding aircraft to bring it under control, resulting in a peak on our mouse speed metric. The lower recall percentage compared to our long-form events may be explained by the short-term nature of this response; the increase of speed to control the speeding aircraft may only be for a fraction of a second, not enough to register a prominent enough peak. For those correctly classified, the knock-on effect of dealing with this abnormal aircraft may have resulted in the participant increasing the average speed of input to compensate for time lost dealing with the errant craft.

The action required to handle a speeding aircraft in an ideal fashion, needs only a single gesture. This single gesture would have to be sufficiently different in shape to normal operating gestures in order to be classified correctly. This may explain the weak performance of the gesture classifier of 8% recall for the first validation set and 20% for the second.

Non-Responsive Aircraft

The slight improvement in performance of the mouse speed classifier may be as a result of the complex nature of the non-responsive aircraft event. The period between the participant realising the aircraft cannot be controlled and inputting mitigating actions to compensate can be relatively frenetic, which is translated in this case into increased average speed. The lower recall percentage by comparison to the long-form events may be due this period of increased speed lagging too far behind the event to be correctly classified within the allowed latency period.

The same reasoning can be applied to the mouse gesture metric which has a recall percentage of 39%. The abnormal gestures required to handle the knock-on effects of the non-responsive aircraft may have produced a peak that lags behind the allowed latency window.

5.4.5. Summary

We have presented results in this subsection to address our research question: can mouse movement data be used to identify EoI in an operator scenario?

The mouse speed metric we have used for a classifier in this work demonstrates results comparable to our pupil data, outperforming the average recall and precision pupil classifier results for the first validation set and returning classification recall and precision of near 100% for the long-form events in the first validation set. These results carry over to our more complex second validation set, returning average recall and classification of 100% and 87% respectively. This demonstrates the value of this metric in isolating long-form events of interest. We also see recall results averaging 55% across our short-form events. The metric appears less sensitive to these shorter term events; further work here may involve utilising a broader range of mouse metrics to identify a classifier that is more sensitive to these events.

We also present a novel mouse gesture normality metric that returns recall of 50% and precision of 72% for our first dataset and recall of 43% and precision of 63% for our second set. The metric is clearly more sensitive to the long-form events, yielding an average recall of 79% for long-form events and 22% for short form. The metric is widely applicable to different scenarios due to the ubiquitous use of a mouse in control situations.

It should be mentioned that we expect greater classification results for this particular scenario given the intricately understood parameters of the EoI. Increase in mouse speed is expected for all of our EoI, we equally expect abnormal mouse gestures for our EoI. The robustness of these measures across differing contexts requires further examination.

5.5. Discussion Summary

This chapter has presented the classification results for three data streams: pupil diameter, ECG and mouse movements. The classifiers developed for each were applied to the validation data from two experiments. The recall, precision and F1 Score of each of the measures was evaluated, a summary of these averages can be seen in Figure 5.30.

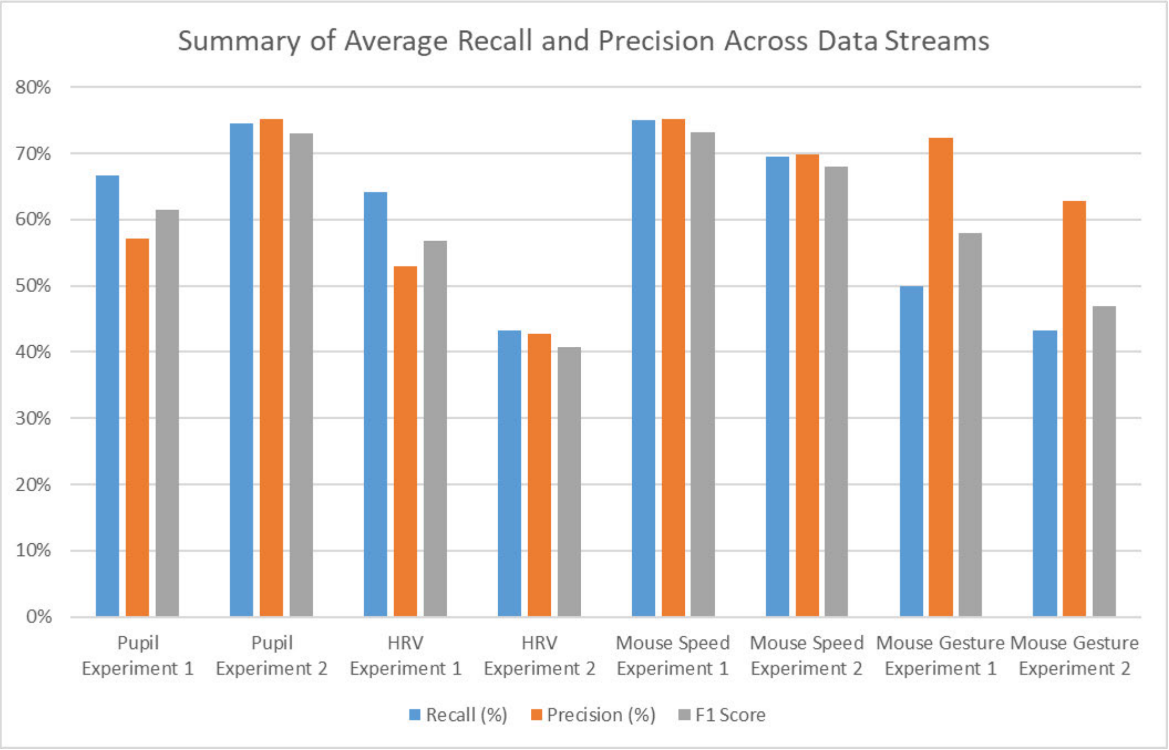


Figure 5.30. Bar chart showing average percentage recall and precision across all 30 participants for each classifier for experiments 1 & 2.

For the physiological measures, pupil diameter demonstrates better performance than HRV. For the first experiment, the pupil classifier demonstrated a 4.65% higher average F1 score, but for the second, more complex experiment, there is a drop off in HRV performance and slight improvement in pupil performance – leading to the pupils having a F1 score 32.29% greater than HRV. As discussed above, this result demonstrates that a higher density of short-form events reduces the value of HRV as a classifier of events of interest. The mouse speed classifier demonstrates the highest average F1 score over experiments 1 and 2. Though the mouse gesture classifier does report lower average F1 scores than the mouse speed measure, it does also report higher precision compared to its recall percentage; meaning that though it is less likely to classify all events, the events it does classify are more likely to be correct.

Each measure was also assessed on an event-by event basis. The performance of each classifier was clearly more robust when recalling the long-form events (traffic surge and cloud cover), as these events create a situation that requires a longer period of increased cognitive load. It is expected that the classifier will be more sensitive to these events as a sustained event is more likely to create an increase the average response from each measure, increasing the likelihood it will meet the parametric requirements for a correct classification. This type of event is also more similar in characteristic to those found in literature that have been demonstrated to generate significant responses in cognitive load.

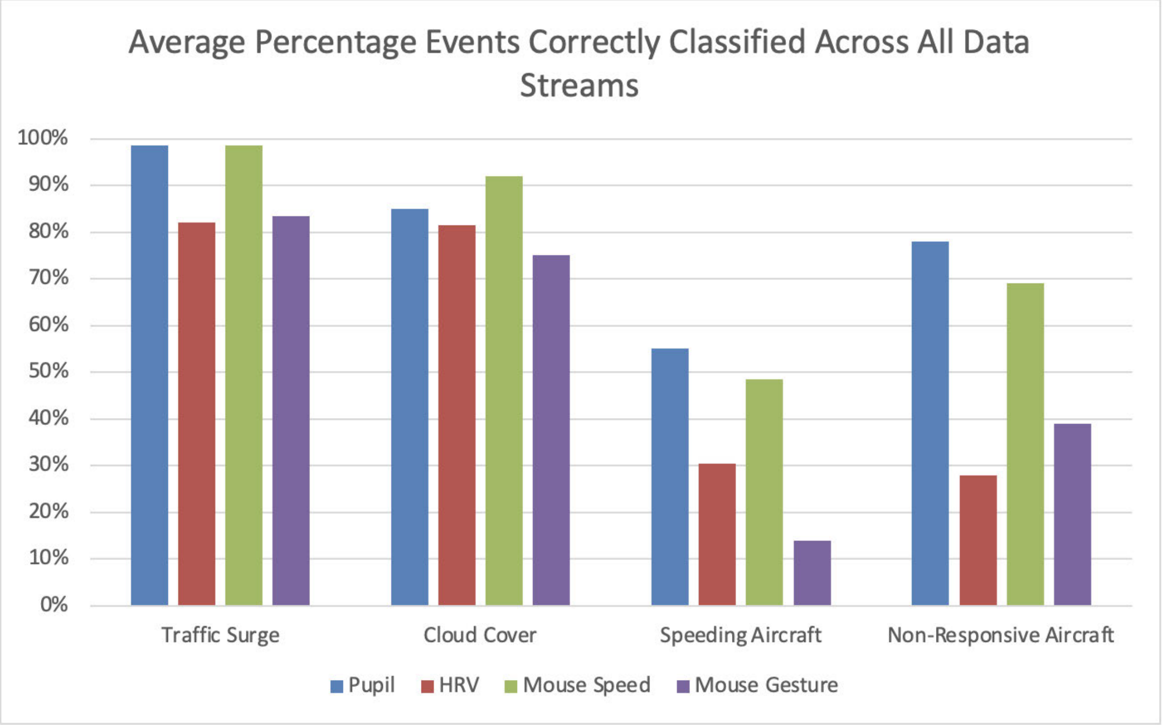


Figure 5.31. Percentage of correct classifications across each event.

It can be seen in Figure 5.31. that pupil diameter and mouse speed are the highest performing classifiers across all events. The speeding aircraft event was the event that all measures were least sensitive to. This can be explained by a very short duration of the event that will reduce the classification performance for the opposite reason of that of the long-form events. Another reason for the poor performance for this event for the mouse metrics is that the event can be handled with relatively few changes to the mouse input. For example, if the participant can click and drag a path from the speeding aircraft immediately, this represents both a very short increase in average speed that will not generate a large enough response in that measure, it also may represent a single very ‘normal’ gesture, which will not increase the gesture normality data stream significantly enough to be classified.

Overall, these results demonstrate that physiological indices can be used to identify the times in which long-form events occurred from complex datasets. With average classification between 81.5% and 98.5% for the long form events. The results also demonstrate that the use of mouse speed as a classifier of long form events yields equally high performance of 92% - 98.5%. Finally, the use of mouse gesture normality as a classifier of events of interest demonstrates less value for long-form events than the speed based metric (average 16% less correct classifications), but the metric may be more widely applicable over differing contexts given its design, though examination of this needs further work.

6. Conclusion

6.1. Summary

The research objective for this EngD project was: *To determine if the times at which events of interest occurred within complex control room scenarios can be retrieved without manual intervention from the operator.*

We present in this thesis a body of work to meet this aim, which we have achieved.

The research was inspired by the nature of the data labelling problem. The data labelling problem inhibits many different applications of machine learning, but arguably this problem is more acute in control room scenarios where obtaining labelled data is made more difficult again. With a combination of a very small pool of available experts to label data and the inherent complexity of the data they operate within, standard manual labelling methods are very time intensive and expensive to use in this context. Labelling the data in real time also poses a significant challenge as the inherent nature of the domains of control room operators are inherently high-stress, high-stakes environments where the attention of the operator is critical.

Our motivation for the research comes from the pairing of this problem to that of cognitive load. Cognitive overload is often cited as the main performance issue suffered by control room operators. Cognitive load increases when more mental resources are applied to a certain task. Cognitive load is most commonly measures subjectively through the posing of structured questionnaires such as the NASA Task Load Index, this allows researchers to characterise the difficulty of certain tasks with more dimensionality. Cognitive load can also be measured objectively by monitoring human psychophysiological signals. Cognitive stimuli can influence the autonomic nervous system to unconsciously change outputs that can be measured such as pupil diameter and heart rate.

We attempt to determine if these two fields: objective cognitive load and the control room data labelling problem, can be linked to create an automatic data labelling system by passively monitoring an operators psychophysiological output. We hypothesise that the events we wish to label will yield a change in psychophysiological output that differs from the baseline cognitive load of the regular tasks of the operator. We would then determine when events of interest occur through analysis of these signals post hoc, resultantly, further honing our research objective:

To determine if the times at which events of interest occurred within complex control room scenarios can be retrieved without manual intervention from the operator by analysing their psychophysiological signals.

We first assessed literature to determine the state of the art in this field. Through this, we discover that a wide variety of psychophysiological signals have been applied in control room scenarios to characterise task difficulty and operator performance limits. We also discover there is a gap in the literature for using these signals to determine the *time* at which events of interest occur. Often, the analysis techniques used on these signals requires precise knowledge of the time of an event in order to perform the characterisation.

We then develop a methodology to achieve our research objective by designing a piece of research. We select two well-documented signals to measure; ECG and pupil diameter. We also decide to use mouse input metrics given the inherent simplicity of integrating this data in this context. We then generate 3 research questions:

Can pupil dilation data be used to identify the times at which events of interest occur in an operator scenario?

Can HRV data be used to identify the times at which events of interest occur in an operator scenario?

Can mouse movement data be used to identify the times at which events of interest occur in an operator scenario?

We then develop a simulation environment that will enable us to test our hypothesis in a control room scenario. The environment contains 4 events of interest that are manually labelled to determine the recall and precision of our analysis.

We then pre-process the data gathered from the experiments, creating definitions for ground truth labelling the video feed as a benchmark. We then select an HRV metric based of an accuracy and latency study, determining that the significant difference in latencies of each measure needs to be carefully considered when determining correct classifications.

We then develop a novel time series analysis of the psychophysiological signals, modelling the series as peak responses as filtering the responses by tuning parameters of peak prominences and data smoothing. We also develop a novel technique for determining a time series of "normality" for mouse movements.

6.2. Main Findings

Of the two physiological measures, the pupil diameter measure is the best performing metric at determining times of events of interest; yielding an average F1 score of 67% across both the simpler first experiment and the more complex 2nd experiment. Upon inspection of the inter-event classification performance, we note that the pupil diameter correctly recalled 91.5% of the long form events as an average across all participants. We also demonstrate that the value of the pupil diameter metric as a correct classifier for the short-form events is less than that of the long-form, correctly recalling the speeding aircraft and non-contact aircraft events at 55% and 78% respectively. These results conclude that pupil diameter can successfully be used to locate times of events of interest, with robust classification performance for long-form events and with significantly less value as a classifier for short-form events.

As with the pupil diameter, the HRV measure demonstrates higher recall of the long-form events; showing an average correct classification of 82% for long for events compared to an average 29.5% for short form. These results conclude that HRV is not capable of consistently locating times of short-form events of interest but can be used as a tool to locate these times for long form events.

The mouse speed measure demonstrates an average F1-score of 70.5% across the two experiments - the highest of all the measures. As with the other measures, we note a drop-off in classification performance between the long and short-form events (from an average of 95% to 58.5%). We can conclude that the mouse speed measure is a suitable classifier to determine long –form events from complex data (assuming that the complex data required a mouse as an input). As for the short-form events, we can conclude that the measure is similarly as sensitive to these events as the pupil measure, though with an average correct recall of 58.5%, we cannot conclude that it is a robust measure.

The mouse gesture normality measure presented here demonstrates an average F1 score of 52.2% - from this average across all participants we can conclude that it is not a robust or consistent measure for determining times of events of interest from complex data. Upon inspection of the event-by-event results however, we can demonstrate that the measure has more value at determining short term events, with an average correct recall of 79%. This demonstrates the value of the measure during events that require a more sustained period of abnormality over those events that require a very short period of abnormal gestures.

In context of the existing literature in this field, we present a totally novel set of results. To compare the results obtained here we must reframe the work in terms of the events definition. In literature, the events are pre-partitioned and the human data streams gathered within each of these partitions and compared against either other partitions or some benchmark. The vast majority of studies determined that changing the difficulty of the task led to a significant change in psychophysiological signals (depending on both the type of signal and how it was processed). This leading conclusion led to the investigation in this research to apply the same logic in reverse; determining the *times* of the events of changing difficulty based on when the psychophysiological changed significantly (again, depending on the type of signal and how it was processed). The key difference being the context in which the event could occur, by comparing a known section against some previously known baseline would be inappropriate in a long-term scenario as it cannot distinguish against “normal” activity as “normal” activity has to be clearly defined across the length of the trial. If we compare our results against those of (Pedrotti, Benedetto, *et al.*, 2014) that attempt to classify pupil diameter as a measure of cognitive load against 4 increasing categories of difficulty in a simulated driving task; they achieve an average classification accuracy of 83%. On average, for the same psychophysiological measure, we achieve an average F1-score of 67%. This is significantly less accurate than our comparison study, this could be due to several factors that are important to note; firstly, that the context of the signal measurement comparison can affect the nature of the accuracy results (as they have clearly defined time sections of increased difficulty) with pauses in between these time periods. Also to consider are that we employ a greater sample size (30 participants vs their 16) and their domain context of driving is also a different scenario to our control room context.

In conclusion, we determine that psychophysiological signals can be used to identify the times at which events of interest occur. The caveats of this approach become clear when event-type analysis is performed. Given the structure of increasing task difficulty in literature is often to maintain this for a sustained period of time, this agrees with our findings that the measures perform significantly better at detecting long-form events. Detection of short-term events is certainly possible with pupil diameter and would be an interesting case for further research.

We also conclude there is no concrete evidence to suggest HRV is capable of detecting short-term events. This is to be expected given the nature of windowing the data for HRV procedures; window lengths of 40 seconds, considered ultra-short form, are likely to average out any smaller responses such as the short-term events here that last ~2 seconds.

6.3. Limitations of the Work

We assume that the mental state of the participants was affected entirely by the simulator they were interacting with. It is not known precisely what was stimulating the participant at any given time. Controls were established in the experimental set up to mitigate potential external stimuli, but internal thought processes and their potential effect on the mental state of the participant could not be controlled.

The nature of the roles of control room operators means that for any given dataset, the times at which events were detected based on the physiological and mouse input metrics would depend on the individual, not an average across a population. As seen in the results here, that individual could yield perfect precision and recall, or very low precision and recall, it would not be possible to confirm the validity of the results without a manual inspection of the data to establish a ground truth, which would negate the purpose of using the signals acquired from the operator.

As mentioned in section 0, the characterization of the participants was not as thorough as was possible. This missing information may have been able to shed light on the nature of significantly higher performing individuals.

6.4. Contributions of This Research

This research is the first study to research retrieval of the *times* at which events of interest occurred in a control room scenario. We determine that pupil diameter and HRV can be used in this way to determine these events if the events in question last for a time of approximately 30 seconds. This is also the first study to determine if these signals are capable of identifying short-form events lasting only 2 seconds or so. We determine that of the signals tested, only pupil diameter demonstrates that potential to recall such events, a valuable contribution to the field as often events of this nature and length are dismissed due to their difficult to detect nature. We also present a novel mouse movement metric that can produce a time series of the normality of the gestures created by the mouse in the context of the timespan recorded. This method has many potential applications beyond this research in human factors and other activity-recognition and assessment fields.

6.4. Suggestions for Further Work

The work presented in this thesis demonstrates the potential for significant value in automated human-data-driven labelling systems. Though the results show that this concept is possible, there is also significant potential for further work to extend the concept and explore new research territories.

As mentioned in the methods chapter, a more thorough and robust process to find participants for the trials could be undertaken to examine some of the behaviours in the data that were not controlled in this work. Firstly, a larger sample size would improve the conclusions from the results presented here. Secondly, though high-level exclusionary factors were implemented here, there exist innumerable, myriad factors that could be used to either exclude, select or observe from the participants. We first mentioned in the results and discussion chapter, that there exist a subset of high-performing individuals of whom the methods here provided perfect or near perfect recall and precision. An area for further work would be to investigate if these individuals have characteristics that identify them as suitable candidates for physiological signal-based event detection. Characteristics such as profession and competence with computer interfaces may lead to further findings that isolate particular participant profiles that are better suited to these methods.

Further work would also involve applying the methods here to real-world scenarios. By deploying the psychophysiological signal capturing technologies to a range of differing domains, such as those mentioned in the literature review, assessments could be made to determine if particular domains are more or less suitable for abnormal event detection through psychophysiological assessment. For example, are the stresses on the cognitive function of a pilot too myriad and ranging to enable detection of events that are different from the normal functioning of their duties compared to the driver of a car? Research could also determine whether the skill or experience of the pilot/driver affected the methods presented here.

Finally, the methods to analyse the data streams in this research treated each stream separately to determine the sensitivity of each to events of differing types. Further work here could expand these methods to include techniques that combine the various signals together to create ensemble classifiers. The potential of these techniques is ranging, from either simply increasing the specificity and accuracy of the system to potentially creating classifiers that are capable of distinguishing between different types of events.

References

Aasman, J., Mulder, G. and Mulder, L. J. M. (1987) 'Operator effort and the measurement of heart-rate variability', *Human Factors*, 29(2), pp. 161–170. doi: 10.1177/001872088702900204.

Van Acker, B. B. *et al.* (2020) 'Mobile pupillometry in manual assembly: A pilot study exploring the wearability and external validity of a renowned mental workload lab measure', *International Journal of Industrial Ergonomics*, 75(August 2019). doi: 10.1016/j.ergon.2019.102891.

Alnaes, D. *et al.* (2014) 'Pupil size signals mental effort deployed during multiple object tracking and predicts brain activity in the dorsal attention network and the locus coeruleus', *Journal of Vision*, 14(4), pp. 1–1. doi: 10.1167/14.4.1.

Anderson, E. W. *et al.* (2011) 'A user study of visualization effectiveness using EEG and cognitive load', *Computer Graphics Forum*, 30(3), pp. 791–800. doi: 10.1111/j.1467-8659.2011.01928.x.

Aricò, P. *et al.* (2016) 'Adaptive Automation Triggered by EEG-Based Mental Workload Index: A Passive Brain-Computer Interface Application in Realistic Air Traffic Control Environment', *Frontiers in Human Neuroscience*, 10(October), p. 539. doi: 10.3389/fnhum.2016.00539.

Beatty, J. and Lucero-Wagoner, B. (2000) 'The pupillary system', *Handbook of psychophysiology (2nd ed.).*, pp. 142–162. Available at: http://prx.library.gatech.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2000-03927-005&site=ehost-live.

Berka, C. *et al.* (2007) 'EEG Correlates of Task Engagement and Mental\nWorkload in Vigilance, Learning, and Memory Tasks', *Aviation, Space, and Environmental Medicine*, 78(5), pp. B231–B244. Available at: http://www.b-alert.com/augcog/ASEM_unofficial_final_DO_NOT_DISTRIBUTE_EXTERNALLY.pdf.

Bernhardt, K. A. *et al.* (2019) 'The effects of dynamic workload and experience on commercially available EEG cognitive state metrics in a high-fidelity air traffic control environment', *Applied Ergonomics*, 77(January), pp. 83–91. doi:

10.1016/j.apergo.2019.01.008.

Bhavsar, P., Srinivasan, B. and Srinivasan, R. (2016) ‘Pupillometry Based Real-Time Monitoring of Operator’s Cognitive Workload to Prevent Human Error during Abnormal Situations’, *Industrial and Engineering Chemistry Research*, 55(12), pp. 3372–3382. doi: 10.1021/acs.iecr.5b03685.

Brouwer, A. M. *et al.* (2012) ‘Estimating workload using EEG spectral power and ERPs in the n-back task’, *Journal of Neural Engineering*, 9(4), p. 045008. doi: 10.1088/1741-2560/9/4/045008.

Castaldo, R., Melillo, P. and Pecchia, L. (2015) ‘Acute mental stress assessment via short term HRV analysis in healthy adults: A systematic review’, *IFMBE Proceedings*. Elsevier Ltd, 45, pp. 1–4. doi: 10.1007/978-3-319-11128-5_1.

Chen, F. (2011) ‘Neuro-ergonomic Research for Online Assessment of Cognitive Workload’, (September), pp. 568–571. Available at: http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA551175.

Cohen, S., Janicki-Deverts, D. and Miller, G. E. (2007) ‘Psychological stress and disease’, *Journal of the American Medical Association*, 298(14), pp. 1685–1687. doi: 10.1001/jama.298.14.1685.

Coyne, J. and Sibley, C. (2016) ‘Investigating the use of two low cost Eye tracking systems for detecting pupillary response to changes in mental workload’, *Proceedings of the Human Factors and Ergonomics Society*, pp. 37–41. doi: 10.1177/1541931213601009.

Das, R. *et al.* (2014) ‘Cognitive Load measurement - A comparative study using Low cost Commercial EEG devices’, *3rd Internaltional Conference on Advances in Computing, Communications & Informatics*, pp. 1188–1194.

Daszykowski, M. and Walczak, B. (2010) ‘Density-Based Clustering Methods’, *Comprehensive Chemometrics*, 2(34), pp. 635–654. doi: 10.1016/B978-044452701-1.00067-3.

Denison, R. N., Parker, J. A. and Carrasco, M. (2019) ‘Modeling pupil responses to rapid sequential events’, *bioRxiv*. doi: 10.1101/655902.

Ding, Y. *et al.* (2020) ‘Measurement and identification of mental workload during simulated computer tasks with multimodal methods and machine learning’, *Ergonomics*. Taylor & Francis, 63(7), pp. 896–908. doi: 10.1080/00140139.2020.1759699.

Egeth, H. and Kahneman, D. (1975) *Attention and Effort*, *The American Journal of Psychology*. doi: 10.2307/1421603.

Fallahi, M. *et al.* (2016) 'Effects of mental workload on physiological and subjective responses during traffic density monitoring: A field study', *Applied Ergonomics*. Elsevier Ltd, 52, pp. 95–103. doi: 10.1016/j.apergo.2015.07.009.

Ferdous, N. (2014) 'Using Pupil Size as an Indicator for Task Difficulty in Data Visualization', *Proceedings of the IEEE Visualization Conference*, pp. 1–2.

Ferrari, M. and Quaresima, V. (2012) 'A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application', *NeuroImage*. Elsevier Inc., 63(2), pp. 921–935. doi: 10.1016/j.neuroimage.2012.03.049.

Fishburn, F. A. *et al.* (2014) 'Sensitivity of fNIRS to cognitive state and load', *Frontiers in Human Neuroscience*, 8(February), p. 76. doi: 10.3389/fnhum.2014.00076.

Van Gerven, P. W. M. *et al.* (2004) 'Memory load and the cognitive pupillary response in aging', *Psychophysiology*, 41(2), pp. 167–174. doi: 10.1111/j.1469-8986.2003.00148.x.

Gevins, A. *et al.* (2007) 'Monitoring Working Memory Load during Computer-Based Tasks with EEG Pattern Recognition Methods', *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 40(1), pp. 79–91. doi: 10.1518/001872098779480578.

Glenn F. Wilson (2002) 'An Analysis of Mental Workload in Pilots During Flight Using Multiple Psychophysiological Measures', *The International Journal of Aviation Psychology*, 12(1), pp. 18–32. doi: 10.1207/S15327108IJAP1201.

Haapalainen, E. *et al.* (2010) 'Psycho-physiological measures for assessing cognitive load', p. 301. doi: 10.1145/1864349.1864395.

Handford, M. (1987) *Where's Waldo?* Little, Brown Boston.

Harper, R. P. and Cooper, G. E. (1986) 'Handling qualities and pilot evaluation', *Journal of Guidance, Control, and Dynamics*, 9(5), pp. 515–529. doi: 10.2514/3.20142.

Hart, S. G. and Staveland, L. E. (1988) 'Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research', *Advances in Psychology*, 52(C), pp. 139–183. doi: 10.1016/S0166-4115(08)62386-9.

Hayes, T. R. and Petrov, A. A. (2016) 'Mapping and correcting the influence of gaze position on pupil size measurements', *Behavior Research Methods*, 48(2), pp. 510–527. doi: 10.3758/s13428-015-0588-x.

Heine, T. *et al.* (2017) 'Electrocardiographic features for the measurement of drivers' mental workload', *Applied Ergonomics*. Elsevier Ltd, 61, pp. 31–43. doi: 10.1016/j.apergo.2016.12.015.

Hess, E. H. and Polt, J. M. (1964) ‘Pupil size in relation to mental activity during simple problem-solving’, *Science,* 143(3611), pp. 1190–1192. doi: 10.1126/science.143.3611.1190.

Hidalgo-Muñoz, A. R. *et al.* (2018) ‘Cardiovascular correlates of emotional state, cognitive workload and time-on-task effect during a realistic flight simulation’, *International Journal of Psychophysiology*. Elsevier, 128(November 2017), pp. 62–69. doi: 10.1016/j.ijpsycho.2018.04.002.

Hill, S. G. *et al.* (1992) ‘Comparison of four subjective workload rating scales’, *Human Factors,* 34(4), pp. 429–439. doi: 10.1177/001872089203400405.

Hoeks, B. and Levelt, W. J. M. (1993) ‘Pupillary dilation as a measure of attention: a quantitative system analysis’, *Behavior Research Methods, Instruments, & Computers,* 25(1), pp. 16–26. doi: 10.3758/BF03204445.

Hogervorst, M. A., Brouwer, A. M. and van Erp, J. B. F. (2014) ‘Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload’, *Frontiers in Neuroscience,* 8(OCT). doi: 10.3389/fnins.2014.00322.

Hossain, G. and Yeasin, M. (2014) ‘Understanding effects of cognitive load from pupillary responses using hilbert analytic phase’, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops,* pp. 381–386. doi: 10.1109/CVPRW.2014.62.

Itakura, F. and Umezaki, T. (2005) ‘Distance measure for speech recognition based on the smoothed group delay spectrum’, in *Acoustics, Speech, and Signal Processing, IEEE International conference on ICASSP ’87.,* pp. 1257–1260. doi: 10.1109/icassp.1987.1169476.

Jainta, S. and Baccino, T. (2010) ‘Analyzing the pupil response due to increased cognitive demand: An independent component analysis study’, *International Journal of Psychophysiology*. Elsevier B.V., 77(1), pp. 1–7. doi: 10.1016/j.ijpsycho.2010.03.008.

Jimenez-Molina, A., Retamal, C. and Lira, H. (2018) ‘Using psychophysiological sensors to assess mental workload during web browsing’, *Sensors (Switzerland),* 18(2), pp. 1–26. doi: 10.3390/s18020458.

Kahneman, D. and Beatty, J. (1966) ‘Pupil Diamtere and Load on Memory’, *Science,* pp. 1583–1585. Available at: http://science.sciencemag.org/.

Kaklauskas, A. (2015) ‘Web-based biometric computer mouse advisory system to analyze a user’s emotions and work productivity’, *Intelligent Systems Reference Library,* 81, pp. 137–173. doi: 10.1007/978-3-319-13659-2_5.

Kiefer, P. *et al.* (2016) ‘Measuring cognitive load for map tasks through pupil diameter’, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* 9927 LNCS, pp. 323–337. doi: 10.1007/978-3-319-45738-3_21.

Kin, H. and Epps, J. (2016) ‘Pupillary transient responses to within-task’, *Computer Methods and Programs in Biomedicine*, 137, pp. 47–63. doi: 10.1016/j.cmpb.2016.08.017.

Klimesch, W. (1999) ‘EEG alpha and theta oscillations reflect cognitive and memory performance: a rKlimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. Brain Research Reviews, 29(2-3), 169–195. doi:10.1016/S016’, *Brain Research Reviews*, 29(2–3), pp. 169–195. doi: 10.1016/S0165-0173(98)00056-3.

Klingner, J., Kumar, R. and Hanrahan, P. (2008) ‘Measuring the task-evoked pupillary response with a remote eye tracker’, *Proceedings of the 2008 symposium on Eye tracking research & applications - ETRA ’08,* 1(212), p. 69. doi: 10.1145/1344471.1344489.

Ko, A. (2013) ‘Kolakowska13’, *Proceedings of IEEE International Conference on Multimedia and Expo*, pp. 548–555.

Krejtz, K. *et al.* (2018) ‘Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze’, *PLoS ONE*, 13(9), pp. 1–23. doi: 10.1371/journal.pone.0203629.

Lecoutre, L. *et al.* (2015) ‘Evaluating EEG measures as a workload assessment in an operational video game setup’, *PhyCS 2015 - 2nd International Conference on Physiological Computing Systems, Proceedings*, (February), pp. 112–117.

Lees, F. (2016) *Lees’ Loss Prevention in the Process Industries*. 3rd edn, *Lees’ Loss Prevention in the Process Industries*. 3rd edn. Butterworth-Heinemann. doi: 10.1016/c2009-0-24104-3.

Lisi, M., Bonato, M. and Zorzi, M. (2015) ‘Pupil dilation reveals top-down attentional load during spatial monitoring’, *Biological Psychology*. Elsevier B.V., 112, pp. 39–45. doi: 10.1016/j.biopsycho.2015.10.002.

Luque-Casado, A. *et al.* (2016) ‘Heart rate variability and cognitive processing: The autonomic response to task demands’, *Biological Psychology*. Elsevier B.V., 113, pp. 83–90. doi: 10.1016/j.biopsycho.2015.11.013.

Mallick, R. *et al.* (2017) ‘The use of eye metrics to index cognitive workload in video games’, *Proceedings of the 2nd Workshop on Eye Tracking and Visualization, ETVIS 2016*. IEEE, 21005, pp. 60–64. doi: 10.1109/ETVIS.2016.7851168.

Mandrick, K. *et al.* (2016) 'Neural and psychophysiological correlates of human performance under stress and high mental workload', *Biological Psychology*. Elsevier B.V., 121, pp. 62–73. doi: 10.1016/j.biopsycho.2016.10.002.

Mansikka, H., Virtanen, K., *et al.* (2016) 'Fighter pilots' heart rate, heart rate variation and performance during an instrument flight rules proficiency test', *Applied Ergonomics*. Elsevier Ltd, 56, pp. 213–219. doi: 10.1016/j.apergo.2016.04.006.

Mansikka, H., Simola, P., *et al.* (2016) 'Fighter pilots' heart rate, heart rate variation and performance during instrument approaches', *Ergonomics*. Taylor & Francis, 59(10), pp. 1344–1352. doi: 10.1080/00140139.2015.1136699.

Mansikka, H., Virtanen, K. and Harris, D. (2019) 'Comparison of NASA-TLX scale, modified Cooper–Harper scale and mean inter-beat interval as measures of pilot mental workload during simulated flight tasks', *Ergonomics*. Taylor & Francis, 62(2), pp. 246–254. doi: 10.1080/00140139.2018.1471159.

Marinescu, A. C. *et al.* (2018) 'Physiological Parameter Response to Variation of Mental Workload', *Human Factors*, 60(1), pp. 31–56. doi: 10.1177/0018720817733101.

Marquart, G. and de Winter, J. (2015) 'Workload assessment for mental arithmetic tasks using the task-evoked pupillary response', *PeerJ Computer Science*, 1, p. e16. doi: 10.7717/peerj-cs.16.

Marshall, S. P. (2002) 'The Index of Cognitive Activity: measuring cognitive workload', *Proceedings of the IEEE 7th Conference on Human Factors and Power Plants*, pp. 5–9. doi: 10.1109/HFPP.2002.1042860.

Mathôt, S. *et al.* (2018) 'Safe and sensible preprocessing and baseline correction of pupil-size data', *Behavior Research Methods*. Behavior Research Methods, 50(1), pp. 94–106. doi: 10.3758/s13428-017-1007-2.

Matthews, G. *et al.* (2017) 'Metrics for individual differences in EEG response to cognitive workload: Optimizing performance prediction', *Personality and Individual Differences*. Elsevier Ltd, 118, pp. 22–28. doi: 10.1016/j.paid.2017.03.002.

Mosaly, P. R., Mazur, L. M. and Marks, L. B. (2017) 'Quantification of baseline pupillary response and task-evoked pupillary response during constant and incremental task load', *Ergonomics*. Taylor & Francis, 60(10), pp. 1369–1375. doi: 10.1080/00140139.2017.1288930.

Murphy, P. R. *et al.* (2014) 'Pupil diameter covaries with BOLD activity in human locus coeruleus', *Human Brain Mapping*, 35(8), pp. 4140–4154. doi: 10.1002/hbm.22466.

Narins, P. M. (1995) 'Evolution of Human Walking', *Scientific American*, 273(2), pp. 78–83.

Nickel, P. and Nachreiner, F. (2004) ‘Sensitivity and Diagnosticity of the 0.1-Hz Component of Heart Rate Variability as an Indicator of Mental Workload’, *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 45(4), pp. 575–590. doi: 10.1518/hfes.45.4.575.27094.

Nourbakhsh, N. *et al.* (2012) ‘Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks’, *Proceedings of the 24th Conference on Australian Computer-Human Interaction OzCHI ’12*, pp. 420–423. doi: 10.1145/2414536.2414602.

Palinko, O. *et al.* (2010) ‘Estimating cognitive load using remote eye tracking in a driving simulator’, *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications - ETRA ’10*, p. 141. doi: 10.1145/1743666.1743701.

Pan, J. and Tompkins, W. J. (2007) ‘A Real-Time QRS Detection Algorithm’, *IEEE Transactions on Biomedical Engineering*, BME-32(3), pp. 230–236. doi: 10.1109/tbme.1985.325532.

Pape, A., Wiegmann, D. and Shappell, S. (2001) ‘Air traffic control (ATC) related accidents and incidents: A human factors analysis’, *Proceedings of the 11th …*, (February 2016), pp. 1–4. Available at: http://www.aviation.illinois.edu/avimain/papers/research/pub_pdfs/isap/papewiegmshappavpsy01.pdf.

Pecchia, L. *et al.* (2018) ‘Are ultra-short heart rate variability features good surrogates of short-term ones? State-of-the-art review and recommendations’, *Healthcare Technology Letters*, 5(3), pp. 94–100. doi: 10.1049/htl.2017.0090.

Pedrotti, M., Mirzaei, M. A., *et al.* (2014) ‘Automatic Stress Classification With Pupil Diameter Analysis’, *International Journal of Human-Computer Interaction*, 30(3), pp. 220–236. doi: 10.1080/10447318.2013.848320.

Pedrotti, M., Benedetto, S., *et al.* (2014) ‘Automatic Stress Classification With Pupil Diameter Analysis’, *International Journal of Human-Computer Interaction*, 30(3), pp. 220–236. doi: 10.1080/10447318.2013.848320.

Pereira, T. *et al.* (2017) ‘Heart rate variability metrics for fine-grained stress level assessment’, *Computer Methods and Programs in Biomedicine*, 148, pp. 71–80. doi: 10.1016/j.cmpb.2017.06.018.

Peysakhovich, V. *et al.* (2015) ‘Frequency analysis of a task-evoked pupillary response: Luminance-independent measure of mental effort’, *International Journal of Psychophysiology*. Elsevier B.V., 97(1), pp. 30–37. doi:

10.1016/j.ijpsycho.2015.04.019.

Peysakhovich, V., Dehais, F. and Causse, M. (2015) ‘Pupil Diameter as a Measure of Cognitive Load during Auditory-visual Interference in a Simple Piloting Task’, *Procedia Manufacturing*, 3(Ahfe), pp. 5199–5205. doi: 10.1016/j.promfg.2015.07.583.

Peysakhovich, V., Vachon, F. and Dehais, F. (2017) ‘The impact of luminance on tonic and phasic pupillary responses to sustained cognitive load’, *International Journal of Psychophysiology*. Elsevier B.V., 112, pp. 40–45. doi: 10.1016/j.ijpsycho.2016.12.003.

Porter, G., Troscianko, T. and Gilchrist, I. D. (2007) ‘Effort during visual search and counting: Insights from pupillometry’, *Quarterly Journal of Experimental Psychology*, 60(2), pp. 211–229. doi: 10.1080/17470210600673818.

Privitera, C. M. *et al.* (2010) ‘Pupil dilation during visual target detection’, *Journal of Vision*, 10(10), pp. 3–3. doi: 10.1167/10.10.3.

Puma, S. *et al.* (2018) ‘Using theta and alpha band power to assess cognitive workload in multitasking environments’, *International Journal of Psychophysiology*. Elsevier, 123(October 2017), pp. 111–120. doi: 10.1016/j.ijpsycho.2017.10.004.

Reid, G. B. and Nygren, T. E. (1988) ‘The Subjective Workload Assessment Technique: A Scaling Procedure for Measuring Mental Workload’, *Advances in Psychology*. North-Holland, 52(C), pp. 185–218. doi: 10.1016/S0166-4115(08)62387-0.

Ren, P. *et al.* (2014) ‘Off-line and on-line stress detection through processing of the pupil diameter signal’, *Annals of Biomedical Engineering*, 42(1), pp. 162–176. doi: 10.1007/s10439-013-0880-9.

Rhodes, B. J. *et al.* (2007) ‘SeeCoast: Automated port scene understanding facilitated by normalcy learning’, *Proceedings - IEEE Military Communications Conference MILCOM*. doi: 10.1109/MILCOM.2006.302306.

Rozado, D. and Dunser, A. (2015) ‘Combining EEG with Pupillometry to Improve Cognitive Workload Detection’, *Computer*, 48(10), pp. 18–25. doi: 10.1109/MC.2015.314.

Ryu, K. and Myung, R. (2005) ‘Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic’, *International Journal of Industrial Ergonomics*, 35(11), pp. 991–1009. doi: 10.1016/j.ergon.2005.04.005.

Sakoe, H. and Chiba, S. (1978) ‘Dynamic Programming Algorithm Optimization for Spoken Word.pdf’, *IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING*, ASSP-26(1).

Scharinger, C., Kammerer, Y. and Gerjets, P. (2015) ‘Pupil dilation and EEG alpha frequency band power reveal load on executive functions for link-selection processes during text reading', *PLoS ONE*, 10(6), p. e0130608. doi: 10.1371/journal.pone.0130608.

Scheirer, J. *et al.* (2002) 'Frustrating the user on purpose: A step toward building an affective computer', *Interacting with Computers*, 14(2), pp. 93–118. doi: 10.1016/S0953-5438(01)00059-5.

Schuller, B., Lang, M. and Rigoll, G. (2002) 'Multimodal emotion recognition in audiovisual communication', in *Proceedings - 2002 IEEE International Conference on Multimedia and Expo, ICME 2002*, pp. 745–748. doi: 10.1109/ICME.2002.1035889.

Science, E. (2017) *Monitoring Heart Rate Variability (HRV) is so much more valuable than just monitoring heart rate.* Available at: http://www.myithlete.com/what-is-hrv/ (Accessed: 12 January 2018).

Shaffer, F. and Ginsberg, J. P. (2017) 'An Overview of Heart Rate Variability Metrics and Norms', *Frontiers in Public Health*, 5(September), pp. 1–17. doi: 10.3389/fpubh.2017.00258.

Shakouri, M. *et al.* (2018) 'Analysis of the sensitivity of heart rate variability and subjective workload measures in a driving simulator: The case of highway work zones', *International Journal of Industrial Ergonomics*. Elsevier B.V, 66, pp. 136–145. doi: 10.1016/j.ergon.2018.02.015.

Sharma, N. and Gedeon, T. (2012) 'Objective measures, sensors and computational techniques for stress recognition and classification: A survey', *Computer Methods and Programs in Biomedicine*. Elsevier Ireland Ltd, 108(3), pp. 1287–1301. doi: 10.1016/j.cmpb.2012.07.003.

Shi, Y. *et al.* (2007) 'Galvanic skin response (GSR) as an index of cognitive load', *CHI '07 extended abstracts on Human factors in computing systems - CHI '07*, p. 2651. doi: 10.1145/1240866.1241057.

Speier, C., Valacich, J. S. and Vessey, I. (2007) 'The Influence of Task Interruption on Individual Decision Making: An Information Overload Perspective', *Decision Sciences*, 30(2), pp. 337–360. doi: 10.1111/j.1540-5915.1999.tb01613.x.

Sun, F.-T. *et al.* (2012) 'Activity-Aware Mental Stress Detection Using Physiological Sensors', *Mobile Computing, Applications, and Services*, 76, pp. 282–301. doi: 10.1007/978-3-642-29336-8_16.

Tao, D. *et al.* (2019) 'A systematic review of physiological measures of mental workload', *International Journal of Environmental Research and Public Health*, 16(15), pp. 1–23. doi: 10.3390/ijerph16152716.

Tarvainen, M. P. *et al.* (2014) ‘Kubios HRV - Heart rate variability analysis software’, *Computer Methods and Programs in Biomedicine*. Elsevier Ireland Ltd, 113(1), pp. 210–220. doi: 10.1016/j.cmpb.2013.07.024.

Tattersall, A. J. and Hockey, G. R. J. (2006) ‘Level of Operator Control and Changes in Heart Rate Variability during Simulated Flight Maintenance’, *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(4), pp. 682–698. doi: 10.1518/001872095778995517.

Taub, M. *et al.* (2017) ‘Using multi-channel data with multi-level modeling to assess in-game performance during gameplay with CRYSTAL ISLAND’, *Computers in Human Behavior*. Elsevier Ltd, 76, pp. 641–655. doi: 10.1016/j.chb.2017.01.038.

Tjolleng, A. *et al.* (2017) ‘Classification of a Driver’s cognitive workload levels using artificial neural network on ECG signals’, *Applied Ergonomics*, 59, pp. 326–332. doi: 10.1016/j.apergo.2016.09.013.

Trejo, L. J. *et al.* (2007) ‘EEG-Based Estimation of Mental Fatigue: Convergent Evidence for a Three-State Model’, *Foundations of Augmented Cognition*, pp. 201–211. doi: 10.1007/978-3-540-73216-7_23.

Truschzinski, M. *et al.* (2018) ‘Emotional and cognitive influences in air traffic controller tasks: An investigation using a virtual environment?’, *Applied Ergonomics*, 69(January), pp. 1–9. doi: 10.1016/j.apergo.2017.12.019.

Velichkovsky, B. M. *et al.* (2000) ‘Visual fixations and level of attentional processing’, *Proceedings of the symposium on Eye tracking research & applications - ETRA ’00*, pp. 79–85. doi: 10.1145/355017.355029.

Vidulich, M. A. (1987) ‘Absolute Magnitude Estimation and Relative Judgement Approaches to Subjective Workload Assessment’, in *Tsang, Pamela S.*

Vlastos, D. D. *et al.* (2020) ‘Can a Low-Cost Eye Tracker Assess the Impact of a Valent Stimulus? A Study Replicating the Visual Backward Masking Paradigm’, *Interacting with Computers*, 32(2), pp. 132–141. doi: 10.1093/iwc/iwaa010.

Wahn, B. *et al.* (2016) ‘Pupil sizes scale with attentional load and task experience in a multiple object tracking task’, *PLoS ONE*, 11(12), pp. 9–22. doi: 10.1371/journal.pone.0168087.

van der Wel, P. and van Steenbergen, H. (2018) ‘Pupil dilation as an index of effort in cognitive control tasks: A review’, *Psychonomic Bulletin and Review*. Psychonomic Bulletin & Review, pp. 2005–2015. doi: 10.3758/s13423-018-1432-y.

Wong, H. K. and Epps, J. (2016) ‘Pupillary transient responses to within-task cognitive load variation’, *Computer Methods and Programs in Biomedicine*. Elsevier Ireland Ltd, 137, pp. 47–63. doi: 10.1016/j.cmpb.2016.08.017.

Woods, D. D. *et al.* (2012) 'Can We Ever Escape from Data Overload? A Cognitive Systems Diagnosis', *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, pp. 174–178. doi: 10.1177/154193129904300310.

Yan, S., Wei, Y. and Tran, C. C. (2019) 'Evaluation and prediction mental workload in user interface of maritime operations using eye response', *International Journal of Industrial Ergonomics*. Elsevier, 71(145), pp. 117–127. doi: 10.1016/j.ergon.2019.03.002.

Zhai, J. and Barreto, A. (2006) 'Stress detection in computer users based on digital signal processing of noninvasive physiological variables', *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*, (May), pp. 1355–1358. doi: 10.1109/IEMBS.2006.259421.

Zhang, H. *et al.* (2014) 'Detection of variations in cognitive workload using multi-modality physiological sensors and a large margin unbiased regression machine', *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2014*, 2014, pp. 2985–2988. doi: 10.1109/EMBC.2014.6944250.

Zimmermann, P. G. and Gomez, P. (1984) 'Extending usability : putting affect into the user-experience', *Medical Education*, (1).