A Thesis

entitled

Prediction of Pilot Skill Level and Workload for Sliding-Scale Autonomous Systems

by

Sai Kameshwar Rao Nittala

Submitted to the Graduate Faculty as partial fulfillment of the requirements for the Master of Science Degree in Engineering

Dr. Kevin S. Xu, Committee Chair

Dr. Vijay Devabhaktuni, Committee Co-Chair

Dr. Ahmad Y. Javaid, Committee Member

Dr. Scott M. Pappada, Committee Member

Dr. Amanda Bryant-Friedrich, Dean
College of Graduate Studies

The University of Toledo
December 2017

An Abstract of

Prediction of Pilot Skill Level and Workload
for Sliding-Scale Autonomous Systems

by

Sai Kameshwar Rao Nittala

Submitted to the Graduate Faculty as partial fulfillment of the requirements for the Master of Science Degree in Engineering

The University of Toledo
December 2017

There has been tremendous growth in the quality of communication in the human-computer interaction field. Some of the focus areas have included intelligent adaptive interfaces, and multi modality.

An emerging topic in this field of research involves optimal collaboration between humans and machines to achieve a particular goal. One approach to such a goal involves sliding-scale autonomy, in which a machine is designed to dynamically adjust between different levels of autonomy based on a variety of factors, such as the skill level, workload, and behavior of the human operator.

This thesis proposes a system to dynamically predict skill level and workload for pilots on a flight simulator using classification and regression algorithms, respectively. The proposed system uses the pilot's heart rate variability and flight control data. The flight control data includes pilot interactions, such as throttle and aileron, and flight sensor data, such as latitude and longitude.

A user study on fifteen pilots was conducted, each flying the same five predefined routes on a flight simulator. The results indicate that the flight control data alone is sufficient to provide a near perfect classification of a pilot's skill level of either expert or novice. On the other hand, it was found that a combination of flight control and heart rate data produced a more accurate estimate of mental workload and effort.

The findings provide the first step towards a sliding-scale autonomous system for airplane pilots.

To my parents, Siva Bhaskara Rao and Ramani Sailaja for their endless love and support and to the memories of my late grandparents.

Acknowledgments

First and foremost, I would like to start by expressing my sheer gratitude and appreciation for my advisors, Dr. Vijay Devabhaktuni and Dr. Kevin S. Xu for all the guidance, leadership, motivation, patience, and support with which they have provided me over the course of my graduate studies. I would also like to thank Dr. Ahmad Y. Javaid and Dr. Scott M. Pappada for agreeing to serve as members of my thesis committee.

Secondly, I would also like to thank The University of Toledo's Department of Electrical Engineering and Computer Science and Dr. Cyndee Gruden for providing me with financial support in the form of tuition waivers and graduate assistantships.

I especially wish to extend gratitude towards Dr. Alam Mansoor, Dr. Richard Molyet and Dr. Ivie Stein for enriching my graduate experience. I would also like to express my appreciation to the Ohio Federal Research Network (OFRN) and Dr. Ali Reiter for providing the EECS department with the funding for the research portion of my assistantships.

In addition, I want to thank my colleagues Colin Elkin and Ruthwik Junuthula for introducing me to the vast and exciting world of graduate research. I would also like to acknowledge my current and former colleagues of labs NE 2036 and NE 2042 who helped and supported me during the course of my graduate studies.

Most importantly, I would like to thank my parents and friends for their continuous love and support.

Contents

List of Tables

List of Figures

List of Abbreviations

HMI	Human-Machine Interaction
SSA	Sliding-Scale Autonomy
HR	Heart Rate
HRV	Heart Rate Variability
NASA-TLX	NASA Task Load Index
TWA	T-wave Amplitude
LF	Low Frequency
HF	High Frequency
SWAT	Subjective Workload Assessment Technique
RTLX	Raw Task Load Index
FFT	Fast Fourier Transform
MLP	Multi-layered perceptron
SVM	Support Vector Machine
SVR	Support Vector Regression
RBF	Radial Basis Function
kNN	k-Nearest Neighbors
ANN	Artificial Neural Networks
AUC	Area Under the Curve
RMSE	Root-Mean-Squared-Error
R^2	Coefficient of determination
SD	Standard Deviation
EDA	Electrodermal Activity
PPG	Photoplethysmography
ECG	Electrocardiography

Chapter 1

Introduction

1.1 Overview of Autonomous Systems

Most present-day engineering and technological systems possess autonomy to some extent. Autonomy in these areas ideally means self-governing. The autonomy that this kind of system holds is bound by a few parameters, within which it can safely continue to work on its designated task. Such systems do not cause harm to mankind if they work in an ideal manner, and are being controlled otherwise. Human supervision is diligently required for sensitive tasks and in cases of uncertainty.

Autonomous systems have demonstrated that they significantly increase operational capabilities, such as those of armed forces. From [1], autonomous systems are broadly defined into three types: intelligent, scripted and supervised systems.

- Intelligent autonomous systems use an intelligent autonomy technology to instill human intelligence attributes in the back end of the autonomous system elements. This helps the system in decision making, interpretation, and collaboration with other networks and systems.

- Scripted autonomous systems require a preprogrammed script along with well-defined physical models to accomplish the intended mission objective. Such systems have no human interaction after they are deployed.

- Supervised autonomous systems automate the functions of planning, sensing, monitoring, and networking to carry out the activities associated with an autonomous system. This is generally carried out by using the cognitive abilities of human operators, via an interaction setup for decision making, to perceive the meaning of sensor data, diagnose problems, and collaborate with other systems.

Following are some of the important internal processes that make up an autonomous system:

- Planning and Decision: This process is responsible for developing mechanisms for achieving system goals. This area often involves continuous human-machine interaction (HMI) to complete tasks.

- Sensing and Perception: The sensing and perception processes collect and interpret data from sensors and networks. This information is then used to develop a map representation of the goal of the system.

- Monitoring and Diagnosis: These processes are responsible for fault detection and to help prevent data loss, shut down systems, and isolate fault occurrences.

- Networking and Collaboration: The networking and collaboration process collaborates with other autonomous or manned systems in the surroundings with the help of data links and information content.

- Human-System Interface: Humans are required to provide the objectives and control measures at the beginning of the system design. They are needed to interpret the data from the sensors, diagnose problems and authorize the functions of the system.

However, giving more autonomy to systems does not always mean they can perform better on their own. In fact, it can even be counterproductive at times [2, 3].

1.2 Overview of Discrete and Sliding-Scale Autonomy

Technological systems are being built to actively adapt to the ever-changing conditions and requirements of a given objective. The design of such systems brings about the need for adjustable autonomy based on the circumstances and nature of the tasks. Systems with variable autonomy levels are known as discrete autonomy systems [4]. These autonomy levels are resolved by a team of engineers/designers. The robots at the Idaho National Laboratory are an instance of such a system [5].

A high degree of autonomy invariably means that the system can adjust and act on its own accord, without any supervision. On the other hand, a low degree of autonomy would require a high presence of human personnel involved in cognitive tasks for that job and would usually entail a sensitive system.

An important fix in the technological advancement and autonomy sector is trust. Knowing its limitations, the supervisor must be in a position to trust the process and work of a machine. After trust has been initiated, the supervisor is then able to transfer workload in the form of cognitive tasks to the machine [6]. The supervisor need not know the working of the machine if he or she can understand its decisions. Discrete systems contain only predefined autonomy levels, which does not allow for the functions of the system to be modified. This clearly does not provide the operator with the freedom to regulate the system performance.

In an optimal situation, the various levels of autonomy are based on the complexity of the operating environment. Systems are now moving from discretely autonomous to a more sliding-scale approach, providing the end user with much more flexibility to determine the level of autonomy.

Having a provision to contain sliding-scale autonomy (SSA) greatly affects users, as their performance is known to increase over a period. By providing a SSA for

robots, [5] were successfully able to test the advantages of the system over the discrete autonomous system in place. SSA allows the operator to choose the appropriate level of control to achieve the task at hand. For tasks the system can handle semi-autonomously, a high level of control can be used to reduce operator load. This reduction in load allows the operator to control multiple systems at once, effectively multiplying the operator's effectiveness. When the system fails or a task must be done that the system cannot handle, the operator can then use a low level of control to recover from the failure or perform the difficult task. SSA also reduces the bandwidth required for operation of unmanned surface vehicles [7].

In 2008, Scott et al. proposed generic conditions for an effective SSA [8]:

1. The system must be capable of operating at different levels of autonomy.
2. The operator must have controls for each level of autonomy.
3. The operator must be able to select the level of autonomy being used.
4. The frequency in which the lower lever interfaces are used must be low.

Conditions 1 and 4 require the system to have significant autonomous capabilities, while conditions 2 and 3 raise important user interface considerations.

1.3 Overview of Flight Simulation Experiments

Aircraft navigation is a complex time and workload pressured activity affected by individual factors such as level of expertise and age and external factors such as climate.

Simulator experiments have materialized to be an alternate way to implement the practical training of pilots. Simulator settings allow pilots to be put through additional scenarios and exercises that might not be feasible in practical situations and

to also have supervision and an opportunity for repeated practices. These simulators allow pilots to acquire practical scenario skills without having to encounter the same in a real-life setting. Flight simulators allow pilots to use their domain relevant memory in a realistic fashion when compared to typical paper and pencil assessments [9, 10].

Numerous studies have pointed that expert pilots were more likely to possess flexible task management during difficult flight scenarios. There were also varied decision making processes between expert and novice pilots and studies suggest that expert pilots would adapt faster to the ever-changing situations [11]. In a similar setting, expert pilots were found to be adept and knowledgeable when responding to sudden changes in aviation tasks.

Human factors research has continued to signify that an extreme level of mental workload decreases an individual's ability to react to incoming information and further leads to the increased likelihood of human error. The analysis of mental workload has increasingly gained in popularity within the aviation domain [12]. Flight experience has also shown to have some effect in heart rate (HR) responses to the physiological workload of flying in a simulator. HR of an expert was seen to be lower than that of novice pilots [13].

1.4 Overview of Workload

Humans are expected to perform, physically and mentally demanding tasks, especially with the current and future trends in the advancement of technology [14].

Workload is not only established by the nature of the tasks but also the situational environment in which it is performed. Individual behavior and skills are also a factor in determining workload. Tasks may vary from physical actions to cognitive tasks and also depend on the abilities of the individual.

Subjective workload measures attempt to quantify the effort exerted during task

performance, using numerical ratings that do not directly measure either task performance or physiological responses to work. DiDomenico et al. and John Annet [15, 16] suggest that an individual's subjective report of perceptions associated with physical or mental work is generally reflected in the nature of the task, while its demands are reflected on physical and mental resources. Subjective workload assessments are based on an individual's personal feelings and perceptions [17].

Subjective ratings are influenced by the individual's current goals, motives, and plans [16], but rely ultimately on an individual's ability to relate their sensations to some quantitative measure [18]. Individual differences in response to physical and mental demands compound the difficulty in understanding and measuring workload levels.

Operator error prevention and relevant interference would allow for an accurate evaluation of the mental workload in low and high workload scenarios [19, 20, 21, 22, 23].

Given the new demands and expectations placed on individuals during complex task performance, the impact and interaction of physical and mental activity is a vital determinant of overall workload levels.

1.5 Proposed Analysis Technique

The end goal of this research is to advance the current state-of-the-art in monitoring cognitive workload by developing sliding-scale autonomy algorithms based on heart rate variability (HRV) and task-specific measures in order to enhance human-machine teaming and adjustable autonomy.

Thus far, prior work on skill level, mental workload and effort prediction has not taken advantage of data collected from both flight simulators and physiological data.

In this thesis, machine learning algorithms are applied to predict the skill level,

mental workload, and mental and physical effort of the pilots when performing certain tasks in a laboratory setting.

Both flight data and physiological data are collected and used for such predictions. Six different supervised learning classification algorithms are applied to data collected from subjects flying a flight simulator in a laboratory setting, to automatically determine skill level. The usefulness of the HR data collected was also determined for improving skill level prediction.

To determine a mental workload prediction, both flight and physiological data are used in addition to the NASA Task Load Index (NASA-TLX) scores collected to predict the mental workload on pilots flying through different route scenarios. A similar methodology is also followed for the overall effort prediction. The usefulness of the combined data models was also evaluated. The main findings are as follows:

- Using only flight control data, it is possible to obtain near-perfect classification accuracy (0.99 area under the curve (AUC) in the leave-one-subject-out setting) of a pilot's skill level (novice or expert), whereas using only HR data resulted in a weaker classifier (0.66 AUC).

- Standard regression models are unable to predict both mental workload and effort on the 0-20 scale as measured using the NASA TLX self-report tool [14] (R^2 values around 0 for both predictions) in the leave-one-subject-out setting using only flight control data, only HR data, or even a combination of the two data sources.

- A two-stage predictor for the effort and mental workload was proposed, that involves first predicting skill level followed by predicting the effort and mental load using a combination of two regression models, one trained on experts only and one on novices only. Using the two stage predictor on a combination of flight and HR data, moderately accurate estimates of mental workload and effort are

obtained (R^2 values around 0.3 for mental workload and R^2 values around 0.2 for effort).

1.6 Publications and Contributions to Thesis

The major publications and contributions of this research are presented in Table 1.1.

Table 1.1: Publications and contribution to thesis.

Type	**Contribution**
Journal Paper	Pilot Skill Level and Workload Prediction for Sliding-Scale Autonomy
Source Code	A re-usable library with MATLAB and Python source code of the analysis pipeline

1.7 Thesis Organization

The rest of the thesis is organized as follows: Chapter 2 provides an overview of the prior work conducted in the areas of autonomy, workload and NASA-TLX. Chapter 3 describes the data collection procedure and the details of the experimental setup. Chapter 4 elaborates on the proposed methodology, which includes feature construction, and also discusses the machine learning algorithms used in the thesis. It also discusses the proposed analysis followed for the prediction of skill level, mental workload, and effort. Chapter 5 discusses the experimental results from the predictions. Chapter 6 provides a summary of the thesis and suggests a path for future research.

Moreover, the thesis includes an appendix that contains an elaborate list of the features extracted and results from the IMWUT paper.

Chapter 2

Literature Review

2.1 Review of Autonomy

The advent of autonomous systems has led to tremendous amounts of increase in production and service of machinery. This improvement in productivity is achieved by increasing the number of systems in each person's control. The processing capability of humans is limited, which limits the amount of attention to be devoted to each system. This can be overcome by adding intelligence to the systems, thereby allowing them to operate in autonomous and semi-autonomous modes.

Autonomy can be used to determine the effectiveness of the human-machine team. Due to the multidisciplinary nature of the human-machine interaction, autonomy has been conceptualized in a disparate way. Autonomy has been applied in varying degrees to a wide variety of sectors such as health care nursing tasks [24], domestic assistance [25], search and rescue [26], and education [27]. Due to the wide range of service applications, human-machine interaction is often necessary, and systems of varying autonomy levels are expected to interact with humans having limited or no formal training [28].

The earliest categorization for the various modes of automation was proposed by Sheridan et al. [29]. They deduced a ten-point scale and categorized a higher

level of automation as representing increased autonomy and lower levels as decreased autonomy. The scale is shown in Table 2.1. This taxonomy specified what information was communicated to the human as well as allocation of function split between the human and automation. However, the scale used for taxonomy was restricted to a particular set of discernible points along the continuum of automation and was applied mainly to the output functions of decision making and action selection. One of the disadvantages of the scale was the lack of detailed specification of input functions which was related to the way in which information was acquired.

Table 2.1: Decision making automation levels (Sheridan & Verplank).

Level of Automation	Description
1.	The computer offers no assistance; the human must make all decisions and actions
2.	The computer offers a complete set of decision/action alternatives, or
3.	Narrows the selection down to a few, or
4.	Suggests one alternative, and
5.	Executes that suggestion if the human approves, or
6.	Allows the human a restricted veto time before automatic execution
7.	Executes automatically, then necessarily informs the human, and
8.	Informs the human only if asked, or
9.	Informs the human only if it, the computer, decides to
10.	The computer decides everything, acts autonomously, ignores the human

This 10 point scale was created on a general notion. In 1999, Endsley et al. in [30] proposed a revised taxonomy with greater specificity on input functions. The automation levels have been described with the help of the Endsley and Kaber model [30]. They defined autonomy taxonomy into four generic models which include:

- monitoring - scanning displays
- generating - formulating options or strategies to meet goals
- selecting - deciding upon an option or strategy

- implementing - acting out a chosen option

Parasuraman et al. in [31] suggested that stages of automation be classified as input and output functions, which can be automated to differing degrees along a continuum of fully manual to fully automated. The stages of automation included the following:

(1) Information acquisition

(2) Information analysis

(3) Decision and action selection

(4) Action implementation.

The idea behind Parasuraman and colleagues' model was to provide an objective basis for making the choice to examine the extent required for the automation of the task. The authors proposed a method to evaluate the consequences of the human operator as well as the automation. This method was used to identify potential design issues and provide a process to determine the appropriate levels or ranges of automation.

In the information acquisition stage, automation was used for supporting the processes related to sensing and registering input data. This stage of automation was used for supporting the human sensory and perceptual processes, such as assisting humans with monitoring environmental factors. In this stage, the automation included systems that scan and observe the environment, such as radar, infrared, or goggles. Automation was also used to organize sensory informations such as an automated air traffic control system that prioritizes aircraft for handling at higher levels of information acquisition.

Automation that performed tasks similar to human cognitive function, such as working memory, was categorized under the information analysis stage. Automation in this stage was also used to provide integration of multiple input values, make predictions, or summarize data to the user.

In the decision selection stage, automation was chosen from decision alternatives. Automation in this stage would help the medical doctors by providing recommendation for diagnosis or help with the navigational routes for aircraft in order to avoid inclement weather.

In the last stage, the action implementation stage, automation was used to execute the chosen action. As it was the end stage, the automation may complete all, or subparts, of the task. The implementation stage was used to support an autopilot function in an aircraft or to use automatic stapler in a photocopy machine.

The research conducted by Huang and colleagues supports the viewpoint stating that higher autonomy requires less HMI [32, 33, 34]. Their research was mainly focused on developing a framework for autonomy and metrics used to measure robot autonomy. Although this framework was used primarily within military applications, the general framework has been cited more generally as a basis for human-machine interactive autonomy [35].

In [32], a negative linear correlation between frequency of interaction and autonomy was developed, and it was suggested that as the level of autonomy increased, the interaction frequency decreased. Their model included constructs such as human intervention, operator workload, operator skill level, and the operator-to-robot ratio. Similarly, other researchers have also proposed that higher robot autonomy requires less HMI [35]. Autonomy has been described as the amount of time that a person can neglect the system, and neglect time refers to the measure of how the systems task effectiveness declines over time when the system is neglected by the user [36].

There is a striking contradiction between the idea that higher autonomy reduces the frequency of interaction and the traditional concept proposed by HMI researchers that higher robot autonomy enables more sophisticated interaction [28, 37, 38]. Goodrich et al. concluded that it would be harder to achieve autonomous robots that engage in peer-to-peer collaboration with humans without social interac-

tion [38].

The role of the system and human in the HMI distinguishes two conceptualizations of autonomy [39, 40, 41, 42]. Intervention and interaction most likely defines the role of a human being in connection to the system autonomy. On the other hand, intervention could also be interpreted as a specific type of interaction, as suggested in [32], which refers to the frequency of human control. However, performance problems (e.g., [43, 44]) can be caused by a system acting autonomously without intervention and reflects the human out-of-the-loop phenomenon in automation. The frequency of intervention may be more applicable when the role of the user or human being is to operate the robot (e.g. tele-operation or monitoring). On the other hand, the sophistication of the interaction might be more applicable when the role of the human is that of a bystander (e.g. social partner, coworker, or supervisor). Intervention and interaction should be considered simultaneously when determining the level of autonomy among the robots. This can be achieved by evaluating the amount and level of interaction that is required.

Ideally, higher autonomy levels are desirable, but this requires effort, and the design of such systems must be reliable. The relation between human intervention and failure rate is almost linear in nature. In terms of cost effect, the ideal solution would be to lower the failure rate to a point at which it is cheaper for a human to intervene rather than to undergo system failure. To use this solution, the human must occasionally have access to low levels of control to recover from system failures while performing most operations at a high level of control to reduce operator load.

Bradshaw et al., in [45], described adjustable autonomy as "the system being governed at a sweet spot between convenience and comfort." Altering the level of autonomy in certain ways would allow the researcher to get mixed-initiative interaction. [45] used an interesting vacuum cleaner analogy to explain to the researchers, the concept of mixed-initiative interaction. He stated that the most manual machine

was a simple vacuum cleaner operated by a human arm. Apart from the motor, all its actions were supervised by the human. The opposite, a fully-autonomous vacuum cleaner, turned itself on, vacuumed until it decided that it had finished and then retreated back to its storage place to recharge. This vacuum cleaner required no action or initiation from the user. Such a type of mixed-initiative interaction can be achieved at varying levels.

In addition to dealing with communication delays, adjustable autonomy has also been applied to problems in which human workload and safety are considerations. The concept has been applied in both software [46, 47] and hardware agents [48]. Although promising, challenges in creating systems that effectively employ adjustable autonomy include issues in mixed initiatives [48, 49], intervention, responsibility, and trust [50]. Researchers from aviation and other human-factors areas provide meaningful insights into the application of adjustable autonomy in the HMI domain [51].

For many of the applications in which adjustable autonomy and mixed initiatives are appropriate, it is desirable to allow the human to interact with the system as naturally as possible. This led to research in advanced interfaces, such as gesture recognition [52, 53], emotive computing [54], natural language-based interfaces, and virtual reality-based displays [55]. Additionally, this also led to research in systems learning from human operators [56] and research in designing intelligent interface agents [57].

The key element in mixed initiative systems is the on-running dialogue between human and machine, in which both parties share responsibility for mission safety and success. This work was well characterized by [58], who emphasized a system centered view to HMI. Related concepts are also present in some approaches to share control [59] as well as in situation-adaptive autonomy in aviation automation [51].

SSA allows the operator to choose the appropriate level of control to achieve the task at hand. For tasks that the system can handle semi-autonomously, a high level

of control can be used to reduce operator load. This reduction in load allows the operator to control multiple systems at once, effectively multiplying the operators effectiveness. When the system fails or a task must be done that the system cannot handle, the operator can use a low level of control to recover from the failure or perform the difficult task. SSA also reduces the bandwidth required for operating the systems.

2.2 Review of Workload

Over the past few years, the high intensity work-life of aircraft pilots and air traffic control operators has led attention to the area of workload. Increase in task demands are highly correlated to the subject's capabilities, which may lead to errors in human factor issues, which become critical for safety. In 1986, Gopher et al. in [60] provided a state of the art review of workload and its definitions. In 2003, [61] presented a current review of the workload measurement methods and suggested a few professional recommendations on various techniques for use in simulations involving humans. Castor et al. in [62] provided an assessment process to help choose among the different measures depending on the phenomenon under study.

Task complexity is directly related to mental demand and increases in the processing stages for a task requirement. Both mental demand and task complexity depend on the goals set for task performance. Task difficulty is related to the processing effort that is required by the individual and is dependent on context, capacity, strategy, and state of the allocation of resources.

In 1993, [63] pointed out that workload is not only task-specific but also person-specific. Workload was further defined as the specified amount of information processing for a task performance. It was elaborated that workload was dependent on the individual and the interaction between task structure and operator, the same tasks

did not result in an equal amount of workload for all individuals.

O'Donnell et al. in [64] provided an alternate definition of workload. They defined workload as the portion of the operator's capacity required to perform a task. Years later, [65, 66] stated that mental workload was dependent on the demands in relation to the amount of resources the operator can allocate and is a relative concept.

To reduce the workload on operators, there have been numerous advancements and developments in equipment design. Collision avoidance systems, driver impairment navigation systems, and traffic information systems have been designed to help drivers but this has resulted in an overload of information processing [67]. Scheduling is done to prevent overload, but the operator's limitations must be considered while doing so. High road-environment demands include having to merge in heavy traffic [68, 69], while the effects of alcohol, monotony and fatigue have shown to increase workload by a reduction in capacity [70].

Table 2.2: Factors affecting workload.

Driver State Affecting Factors	Driver Trait Factors	Environmental Factors
monotony	experience	road environment demands
fatigue	age	traffic demands
sedative drugs & alcohol	strategy	vehicle ergonomics

A list of factors affecting driver workload is explained in Table 2.2. The table displays driver state, trait, and environmental factors that influences workload. Factors may either increase or decrease mental workload.

O'Donnel et al. in [64] defined primary-task performance as a measure of comprehensive effectiveness of man-machine interaction. They also stated that it was required to involve task performance and workload measures to draw conclusions about human-machine interaction and additionally learning more about the operator's strategy.

One of the requirements for psycho-physiological measures is to accurately study

and predict mental workload as well as the reference data that establishes the subjects' unstressed background state. In an ideal scenario, this would act as a baseline for the evaluation of an individual specific experimental setting.

Recent studies have involved an operator's physiological measures to study the overall workload during tasks. The advantages with these measures are that they can be collected continuously, and they are relatively unobtrusive due to miniaturization. Another advantage is that these measures do not require a response from the operator. Pupil diameter was found to be sensitive to global activation [71]. Similarly, the evoked cortical brain potential was sensitive to particular stages of information processing [72].

In 1992, [73] claimed that the main determinant in HR response in experience pilots in the absence of physical effort is mental workload. However, in settings such as laboratory experiments or automobile driving, the workload levels are lower than pilot workload levels [74]. The HR is influenced by not only physical effort but also emotional factors [75], such as high responsibility or the fear of failing a test [76]. Other factors affecting cardiac activity are speech and high G-forces [74]. Sedative drugs and time-on-task results in fatigue which leads to a decrease in average HR [77], while low amounts of alcohol are reported to increase HR [78].

In 1963, [79] suggested that HRV in the time domain can also be used to measure mental load. HRV decrease is more sensitive to increases in workload than HR increase, despite there being several insensitivity reports for HR and HRV [80]. One of the causes for finding no effect of mental load on HRV is due to the global nature of the measure and its sensitivity to physical load. [75] showed that an increase in physical load decreased HRV and increased HR, on the other hand, an increase in mental load was followed by a reduced HRV and no effect on HR. Fatigue is reported to increase HRV [77] while hand low amounts of alcohol decrease HRV [81].

Mascord et al. however, report an increase in HRV due to low amounts of alcohol

and suggest that this phenomenon may be attributed to alcohol-induced fluctuations in the autonomic control of HR [78]. A decrease in power in the mid frequency and high frequency bands has been shown to be associated with mental effort and task demands [76, 82, 83, 84].

[85, 86] concluded that spectral measures are primarily sensitive to task-rest differences and do not moderate increases in difficulty within a task. Per [85], only large differences, such as the transition from single to dual task or automatic vs. controlled processing, can trigger observable differences on spectral measures. In the higher workload regions in which performances are affected and overload emerges, the sensitivity of the measure is nonlinear to increases in the workload [87].

Table 2.3: Alternate naming of heart rate measures.

Variable/Frequency band	Abbreviation	Alternative name, i = inverse (related)
Heart rate	HR	Inter-beat-interval (IBI)[i], Heart Period (HP) [i]
Heart rate variability	HRV	Sinus Arrhythmia, Variation coefficient (Modulation index)
T-wave	TWA	T-wave Amplitude
Low frequency band	LF	Temperature band, Slow-wave component
High frequency band	HF	Respiratory Sinus Arrhythmia, V-component, Respiration band

In Table 2.3, alternative naming of HR measures and HRV-frequency bands are listed.

Wilson et al. in [17] suggested that HR provided an index of overall workload. Also, spectral analysis of HRV is more useful as an index of mental workload than the time series features.

A restriction in the use of HR measures is that, due to the idiosyncratic nature of the measure, operators are usually required to serve as their own control in workload assessment [88, 89]. [90] recommended that no corrective action be taken in cases in which the verbalization duration is short or in which speech is relatively infrequent.

Another important factor influencing HRV is physical load. Finally, age may affect the use of HR measures as restriction of subjects to specific age groups may be required if HRV is the primary workload measure. HRV may decrease with increasing age due to, amongst others, a decrease in blood vessel flexibility [91].

An alternate and easier method to calculate workload has been self-assessment methods. These methods involve rating demands based on numerical or graphical scales. Few subjective techniques use scales that are categorical such as the modified Cooper-Harper scale. Other techniques are open ended, such as a standard reference task which would work as an anchor in relation to other subject rated tasks. [92] subdivided rating scale methods into single-dimensional ratings such as overall workload, hierarchical ratings such as the Bedford Scales, and multidimensional ratings such as Subjective Workload Assessment Technique (SWAT) and NASA-TLX. The relevance and usage of the NASA-TLX scales are elaborated in the following section.

2.3 Review of NASA-TLX

The NASA-TLX tool is one of the most widely used instruments for measuring subjective workload. This tool provides an overall index of mental workload as well as the relative contributions of six sub-scales: mental demands, physical demands, and temporal demands, effort, frustration, and perceived performance.

The psychometric characteristics of the NASA-TLX are well documented, validated and were used initially by the Human Performance Group at the NASA Ames Research Laboratory as a tool for subjective evaluation of individuals workload in flight simulation [93, 94], air traffic control studies [95], automated and manual control [96], and vigilance tasks [97]. More recently, it has been used in a variety of tasks outside of the aeronautical field including the medical domain [98, 99, 100], for assessment of workload perception and clinical fields [101].

A principal reason for the popularity of the NASA-TLX among researchers is its ease of implementation. The multidimensionality of the NASA-TLX allows for a more detailed analysis of the workload source relative to other techniques that are based primarily on rankings of mental workload sources. The NASA-TLX is also very portable and can be used in operational experiments.

[102] in 2004 compared three subjective workload instruments (SWAT, NASA-TLX, and the Workload Profile) and showed that all three instruments had high correlations (between 0.73 and 0.79) with one measure of performance but also showed that NASA-TLX had a higher correlation than the two other instruments with a second measure of performance. Concurrent validity of the NASA-TLX was found to be higher than the concurrent validity of the other two workload instruments.

Some researchers have used modified versions of the original NASA-TLX. The use of an unweighted or raw TLX (RTLX) is the most common, as high correlations have been shown between the weighted and unweighted scores [103, 104]. Cao et al., modified the NASA-TLX sub-scales in their study of vehicle navigation systems [105].

Studies have explored the relationship between NASA-TLX ratings and other performance factors, such as fatigue [106], stress [107], trust [108], experience [109], and situational awareness [110]. Other NASA-TLX studies have involved measures of physiological (e.g. cardiovascular, muscular, and skin-related or brain-related) function thought to index different aspects of workload [111].

Research has shown the NASA-TLX to be favored most by subjects, when compared with other subjective workload assessment techniques (e.g. SWAT, the Cooper-Harper scale), and to be highly correlated with other measures of workload [112, 113].

In a review in 2006, [14] estimated that the NASA-TLX has been used in more than 300 studies, mainly in air traffic control and civilian or military aviation. The scores have also been used in a health care setting [114]. The literature indicates the advantages of decision making and flight control of experts in a wide range of various

tasks [115].

Chapter 3

Data Collection

To achieve the project objectives, our colleagues at the Wright State Research Institute (WSRI) and Perduco utilized a protocol that was reviewed and approved by the Wright State University Institutional Review Board (SC# 6315). Under this protocol, both expert and novice pilots were recruited from a local community. Novice pilots were defined as individuals who had less than 40 hours of flight time, while expert pilots were defined as having greater than 40 hours of flight time. The data collection procedure is discussed in the following subsections.

A total of 15 pilots were considered for the experiments which included 12 novices and 3 experts.

Table 3.1: Flight route description.

Flight Route	Duration (minutes)	Environment	Difficulty
1	10	Daytime/Clear	Easy
2	15	Nighttime/Clear	Hard
3	15	Daytime/Clear	Easy
4	15	Daytime/Cloudy	Hard
5	10	Daytime/Clear	Easy

3.1 Data Collection Apparatus

An X-plane flight simulator [116] was used to measure flight metrics such as location, plane orientation, and speed. Each subject was provided 10 minutes or less to familiarize himself to the X-plane flight simulator less than one week prior to testing. Each subject was asked to fly a Cessna aircraft over five different routes with varying levels of difficulty, as described in Table 3.1.

Each flight was completed either by successfully reaching all the waypoints and landing the aircraft or by crashing. The first and fifth flights were identical for direct comparison of learning effects. The corresponding waypoints are shown in Figure 3-1. Heart rate measurements were collected from an ear-clip. The collected data was synchronized and outputted in a time aligned format.

The data was collected at an operating frequency of 1 Hz. A NASA-TLX survey was administered after each flight as a subjective measure of workload, and the subjects were also provided 5 minutes of rest.

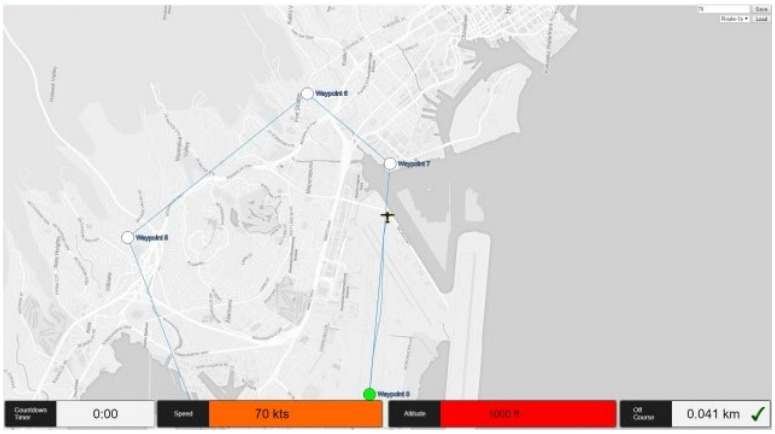

Figure 3-1: Waypoint goals.

The X-plane interface display provided the subjects with a first person view of the cockpit. Figure 3-2 better illustrates the first person view. The cockpit display provided visualization of the current speed, altitude, and other metrics. While flying,

the subject also had vision of an interface, that provided a map with requirements including flight path with waypoints and target altitude and speed to be met at each waypoint. The interface displayed target altitude, target speed, and distance from path at the bottom of the screen. The rest of this interface was occupied by a map of the flight area with a picture of the plane's current position in X-plane, as well as a drawn path and waypoints that need to be hit.

Figure 3-2: First person view of the cockpit.

The X-plane simulator allowed the subjects to control the plane with a yoke controller and foot pedals. The yoke controller allowed for input for a ground-bound wheel brake, a throttle, an air flap control for in-air braking. The spatial orientation of the plane was controlled using a yoke. The foot pedals slid backward and forward to control the ground-bound turning direction, while acting as a toe brake when pressed down. The yoke controller and foot pedals are shown in Figure 3-3.

3.1.1 Flight Data

The collected flight data included pilot inputs, such as aileron, rudder, elevator, heading, and throttle as well as flight sensor data, namely latitude, longitude, altitude,

Figure 3-3: Yoke controller and foot-pedals setup.

and speed. The collected flight data for one of the subjects is shown in Figure 3-4.

3.1.2 Self-Report Data

The NASA-TLX provides a workload assessment on six different scales: mental demand, physical demand, temporal demand, performance, frustration, and effort. Subjects were given the survey after each flight. The measures for novices and experts over each route are shown in Figure 3-5. A particular interest is the difference between the mean mental workload, and the mean effort (physical and mental) across experts and novices, which was utilized in our proposed two stage approach (discussed in chapter 4). For the experiments conducted, the mental workload and effort are believed to be the best measure for the experimental tasks, and therefore, this thesis focuses on these predictions.

The detailed information of the subjects is shown in Table 3.2.

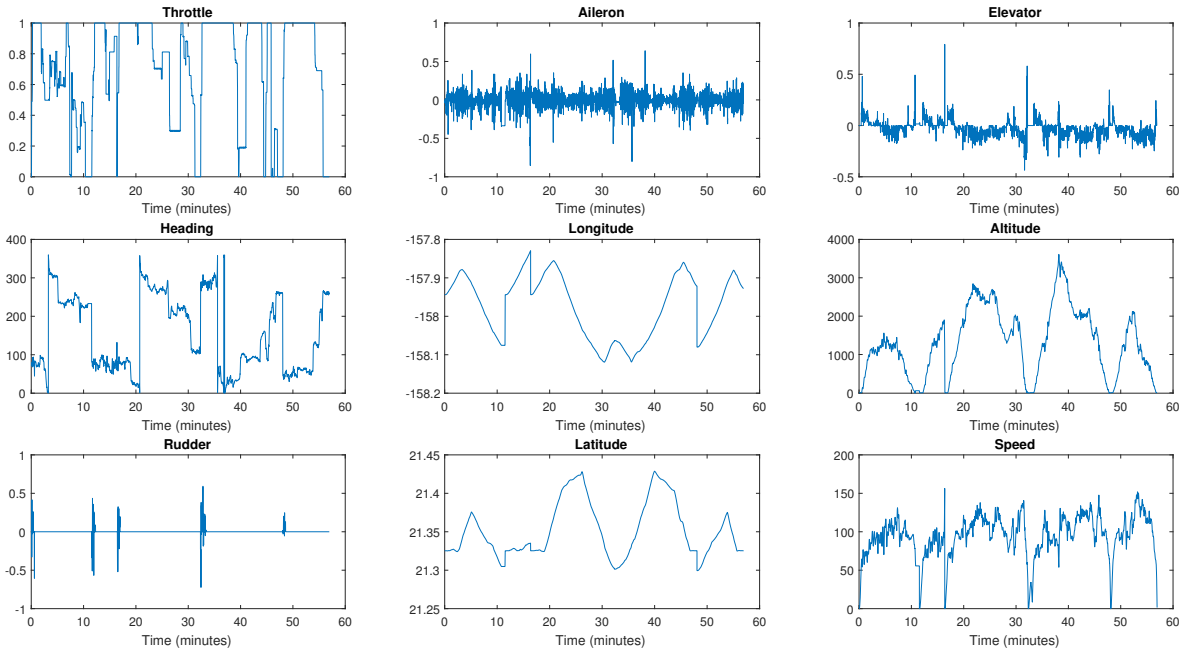

Figure 3-4: Raw flight input data for a single pilot over an entire session (five routes).

3.2 Experimental Procedure

Test subjects participated in the experiment over the course of two days. On the first day, the participant was given informed consent and a demographic survey to complete. Once the documentation was completed, the subject received an acclimation sheet detailing information about the experiment. While the subject read the sheet, the experimenter flew a basic route and answered any questions, ensuring to disclose necessary information about how each input device will affect the behavior of the plane. The user interface interpretation was also addressed. This included the behavior of waypoints and metric data. After observing a flight, the subject flew the same basic flight to become acclimated with the equipment.

On the second day, the data collection took place. The experiment was completed

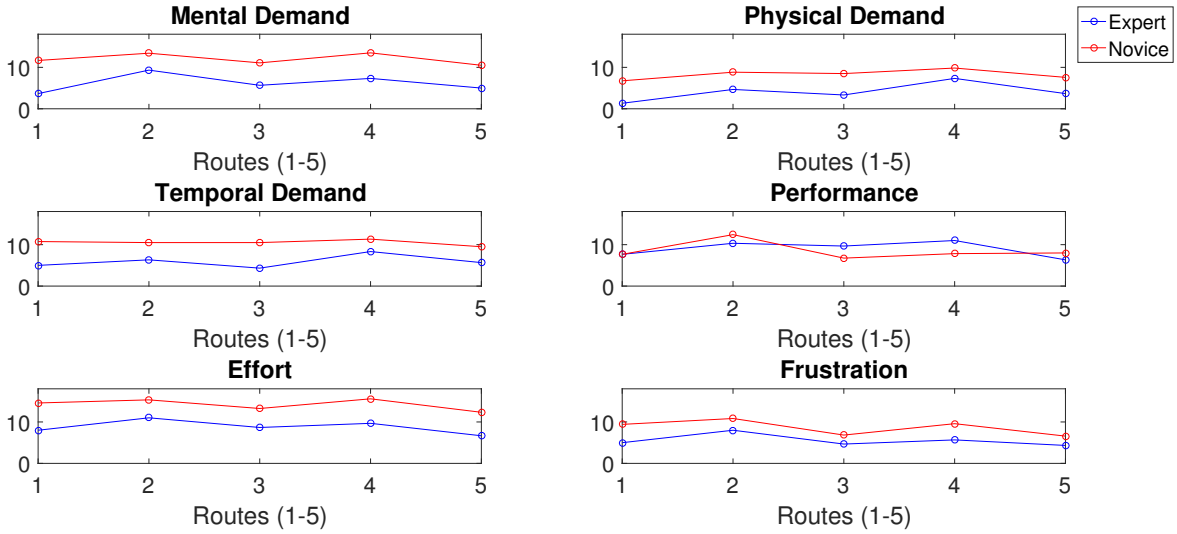

Figure 3-5: Mean NASA-TLX measures for novices and experts across each route.

Table 3.2: Information about the fifteen subjects.

Subjects' Information				
Subject ID	Flight hours	Gender	Age	Dominant hand
1	0	M	30	Right
2	1	M	32	Right
3	0	M	29	Right
4	1200	M	43	Left
5	0	M	22	Right
6	0	M	24	Right
7	3000	M	51	Right
8	0	M	30	Ambidextrous
9	0	M	41	Right
10	0	M	29	Right
11	0	M	24	Right
12	0	M	22	Right
13	0	M	24	Right
14	0	M	32	Right
15	2200	M	40	Right

over the course of two hours in a series of five flight scenarios as described in Table 3.1. The order of these flights was the same for every subject, and each flight scenario took approximately 20 minutes to complete. The first and the last flight scenario were the same in order to measure training effects. The second flight took place at night to measure the effects of black hole syndrome. The third flight was similar to routes one and five. It took place at dusk with no other factors affecting the flight. The goal was to separate flights two and four with an easier simulation to limit the influence of human factors with respect to the reported task load. Flight four was a turbulent flight to test for task load and control inputs in a more physically demanding environment. During the experiment, subjects were told to fly as close to the goal altitude, speed, and path as possible.

Chapter 4

Methods and Experiment Setup

In this chapter, we elaborate on the methodology used for the prediction and also discuss the experiment setup.

A third-order median filter was used as a preprocessing technique to remove noise from the collected data. Taking into account the different route scenarios, the subjects completed each route over the range of 5 - 19 minutes.

One-minute time windows were considered to be a good match for the feature construction of the flight data. After an extensive study of the physiological data analysis [117], the HR data was split into four-minute time windows, keeping in mind the frequency domain requirements for feature extraction.

4.1 Feature Extraction

In the areas of machine learning and pattern recognition, feature extraction is the process of deriving features from the initially acquired data. It is mainly used to extract hidden information from the data that may not be learned directly from the raw data [118]. This helps to save computational complexity and is known to increase the performance of the classifier. Basic feature extraction techniques usually involve frequency, time, and statistical based features. The most common set of these features are used here.

Flight Data A concoction of statistical, frequency, and time domain features were extracted. The following features were calculated for each of the flight data signals at every time window:

- Mean, maximum, minimum, and standard-deviation of the original signal.
- Mean, maximum, minimum, and standard-deviation of the derivative of the original signal.
- Mean, maximum, minimum, and standard-deviation of the double-derivative of the original signal.
- Mean, maximum, minimum, and standard-deviation of the Haar wavelet function and the corresponding detail coefficients at levels 1, 2, and 3 of the decomposed structure.

Wavelet transforms are used here since it is one of the most desired time frequency transformations in present day research [119]. Haar wavelet was used for prediction [120], due to low computing requirements, simplicity and its orthogonal properties making Haar transforms one of the most widely used wavelets in the signal processing sector [121].

Heart Rate Data HR data was segmented into four-minute time windows owing to the needs of the frequency components for the calculation of power estimates. A fast Fourier transform (FFT) was conducted [117] to extract frequency components. The high frequency (HF) and low frequency (LF) components and their intervals are taken into consideration. The total LF power, HF power, and autonomic balance denoted as the ratio of LF to HF were the three frequency domain features extracted [122]. The frequency ranges for the LF and HF components were taken from [123]:

- LF: 0.04 Hz - 0.15 Hz

- HF: 0.15 Hz - 0.40 Hz

The following time domain features were also extracted:

- Mean, maximum, minimum, and standard-deviation of the original signal.
- Mean, maximum, minimum, and standard-deviation of the derivative of the original signal.
- Mean, maximum, minimum, and standard-deviation of the double-derivative of the original signal.

4.2 Data Normalization

Data normalization was performed to bring all the variables into proportion with respect to one another. This improves model behavior and lowers the bias in the classifier learning. Normalization also ensures that the network is not ill conditioned. The coefficients of the classifier reflect in the contribution towards the model. Several algorithms such as multi layered perceptron (MLP) and support vector machines (SVM) have shown faster convergence results on normalized data [124, 125]. In certain fields of statistics, normalization is done in terms of scaling to detect anomalies.

In the experiments, after the feature extraction was completed, the feature matrix was normalized to the ranges [-1,1]. This is computed using the following formula:

$$X_{new} = \frac{X_{old} - X_{new}}{max(X_{new}) - min(X_{new})}, \tag{4.1}$$

where X_{new} is the new normalized value, min(X_{new}) equals -1, max(X_{new}) equals +1 and X_{old} is the original value of the element.

4.3 Feature Selection

Feature selection was performed on both the sets of constructed data features. Feature selection in machine learning is the process of choosing a subset of relevant features which contribute most to the predictions. Feature selection also helps to remove redundant features and improves the performance of the model. By using feature selection, the analysis has a faster computational time as the dimensionality of the feature space is lowered.

By removing the redundant features, feature selection helps reduce the over-fit of a model which enhances generalization while noise components are removed. Over-fitting usually arises when a model is overly complex and/or when a machine learning algorithm models the training data too well. In such cases, the model not only learns the relevant data but also the noise that has a negative impact on the model performance. This in turn leads to the substandard predictive performance of the model and behaves poorly on the testing data. A lot of techniques have been proposed to avoid over-fit (e.g. normalization, grid search, early stopping, cross-validation) and these are generally carried out by either testing the model's performance on a left-out data set or by penalizing the parameters involved in the machine learning algorithm.

In the experiments conducted, feature selection using the L1-penalized logistic regression was examined, which automatically performs feature selection. This is discussed in the next section.

4.4 Grid Search

Grid search is an exhaustive search algorithm that searches for the best set of parameters within a manually specified range in the parameter space and is found to be reliable in low dimensional spaces. Grid search in SVMs are generally evaluated by a performance metric which is an evaluation on a left-out data set [126]. The

grid search algorithm trains the classifier with all possible outcomes of the parameter range and outputs the set of values with which the highest score was achieved in the validation set [126].

The initial choice of the learning parameters for a model is an important step in obtaining good training models [126]. The complexity of grid search increases exponentially as the number of parameters to be tuned increases and is not the most preferred method when it comes to larger and complicated models and datasets [127].

For support vector regression (SVR), the radial basis function (RBF) kernel which was found to be accurate in [128] was used. Both linear and RBF kernels were modeled for the SVM classifier. The penalty term was tuned for the SVM classification. The inverse of regularization strength parameter was tuned for both the penalized logistic regression models. For the k-nearest neighbors (kNN) classifier, the number of neighbors parameter was tuned. The number of estimators parameter was tuned for the random forest classifier and random forest regression and the multiplier term for the LASSO regression was adjusted so as to retain the lowest mean squared error. Both the penalty term and the kernel coefficient terms for the SVR are tuned using the grid-search.

4.5 Machine Learning Algorithms

Machine learning is broadly defined as the ability of a machine to learn without having to provide explicit instructions. It finds applications in many a field ranging from pattern recognition to pharmaceuticals and economics. It comprises of algorithms which learn data and build models based on a few predefined parameters. These algorithms are generally used in decision making, pattern recognition and classification. The algorithms most widely used in literature are artificial neural networks (ANN), SVMs, and clustering algorithms.

4.5.1 Supervised Classification Algorithms

In supervised learning, classification is the problem of identifying a category from a list of categories, to which an observation belongs. This is determined on the basis of a training set of data with known class membership. Algorithms that implement classification are known as classifier.

In this thesis, six widely used classification algorithms were employed for skill level prediction: SVMs with both linear and RBF kernels, kNN classifiers, random forests, and logistic regression with both the penalized models (l1 and l2). Area under the receiver operating characteristic curve (AUC) is chosen as a performance metric.

4.5.1.1 Support Vector Machines

Support Vector Machines are one of the most extensively used supervised learning algorithms, first explicitly developed by Vapnik [129] in 1995. In this method, data is mapped to higher dimensions through nonlinear mapping for the simplification of distinguishing patterns [130]. The term support vectors are those data points that determine the largest difference of separation amongst two groups. SVMs are widely used for image classification and in the biological sector. In recent times, the performance of an SVM is often being considered as a benchmark for categorization and classification tasks as well as a basis for comparison to other machine learning techniques.

4.5.1.2 Random Forest Classification

Random forests are a robust classification and regression based ensemble method. The ensemble of the individual trees formed from the bootstrap samples is known as a random forest. It is highly suitable for real time implementation [131]. Random forests are used in fields ranging from gene selection [132] to predicting customer retention [133] and ecology [134].

4.5.1.3 Logistic Regression

Logistic regression is a widely used algorithm in machine learning for classification. It was developed by David Cox in 1958 [135]. The logistic model is used to estimate the probability of a response based on predictor variables (features).

L1 regularized logistic regression requires solving a convex optimization problem. In particular, it is often used for feature selection and for avoidance of over-fitting, and has been shown to have good generalization performance in the presence of many irrelevant features [136, 137].

4.5.1.4 K-Nearest Neighbor Classification

In the field of pattern recognition, the kNN algorithm is a non-parametric method used for both classification and regression [138]. For kNN classification, the input has k-nearest training instances, and the output is a class membership dependent on the majority vote of the test instance's neighbors. This method is amongst the most simple machine learning algorithms and is an instance based learning approach. The algorithm is usually sensitive to the local data structure [139].

4.5.2 Supervised Regression Algorithms

Supervised regression algorithms are a set of statistical models used to estimate relationships between variables. These algorithms are widely used in forecasting stock prices and understanding gene networks. The algorithms makes predictions from data by learning the relationship among the features and the observed responses. Regression in some cases, refers to specifically the estimation of continuous response (dependent) variables, as opposed to the discrete response variables used in classification [140].

The features extracted from the flight and HR data were also used to determine

the amount of mental workload and effort put in by the pilots. For the prediction of mental workload and effort, LASSO regression, SVR, and random forest regression were employed. Data from the NASA-TLX served as the ground truth. The root mean-squared-error (RMSE) and the coefficient of determination (R^2) score were chosen as the performance metrics.

4.5.2.1 LASSO Regression

In the area of machine learning, LASSO regression is a regression analysis method that performs both feature selection and regularization to enhance the prediction accuracy and to interpret the statistical model produced. It was introduced in 1996 by Robert Tibshirani [141]. This regression was initially designed for least squares models' [142]. LASSO's ability to perform subset selection relies on a form of constraint and has a variety of interpretations including in terms of geometry, Bayesian statistics, and convex analysis [143].

4.5.2.2 Support Vector Regression

SVMs possess a lot of different characteristics: absence of local minima, usage of kernels, number of support vectors and sparsity of the solutions. Similar to classification, the SVR contains all the main features that characterize the maximum margin algorithm. A non-linear function is learned by the machine by mapping into high dimensional feature space. SVR attempts to minimize the generalization error bound so as to achieve generalized performance.

4.5.2.3 Random Forest Regression

Random forest regression was first proposed by Brieman et al [144]. It is a flexible and robust regression method used for modeling the input-output functional relationship. In the random forest, each tree acts on its own accord and the final prediction

is made by considering the mean of the individual tree outputs. It also handles high dimensional data effectively [131].

4.6 Experiment Setup

The predictions were carried out using the machine learning libraries [145] in Python.

Two validation methods were chosen for sample prediction accuracy:

- Leave-a-subject-out validation (15 subjects)
- Leave-a-route-out validation (5 routes)

The leave-a-subject-out validation method involved holding all data for one subject as testing and training the model on data from all other subjects. This process was repeated over all possible subjects involved, which is done to evaluate the ability of the model to generalize the performance on a new pilot data.

As the name suggests, the leave-a-route-out validation method holds data for one route and trains on data from the other routes. All subjects are considered in this case. This method examines variability across the five pre-defined routes. This was done to evaluate the model performance to a new-route assignment for which no training data exists.

Both methods are a standard in most user and driver based studies. These methods are preferred over other cross-validation methods such as 10-fold cross validation, due to the dependence of time windows for the same subject (pilot).

There were some equipment malfunctions in the HR data collection for a few subjects over certain routes. Out of the 75 total routes (5 flown by each of the 15 pilots), 68 routes have both flight control and HR data. Results on these 68 routes are reported.

To examine the importance of the different types of data collected in the experiments, each of the machine learning algorithms were tested on two different feature sets: flight data only, containing a total of 252 features constructed from the flight control data variables; and HR data only, containing a total of 15 features constructed from the HR data.

4.6.1 Skill Level Prediction

The ability of a system to modify its functionality based on the skill level of the operator is of utmost importance in the design of a sliding-scale autonomous system. The skill level of the operator is likely to affect the optimal level of autonomy.

To combat the class imbalance problem (12 novices and 3 experts), the sampling designs were modified [146]. The predefined class weights were adjusted in a way that the weights are inversely proportional to the corresponding class frequencies. This method was followed for both the l1 and l2 penalized logistic regression and the SVMs.

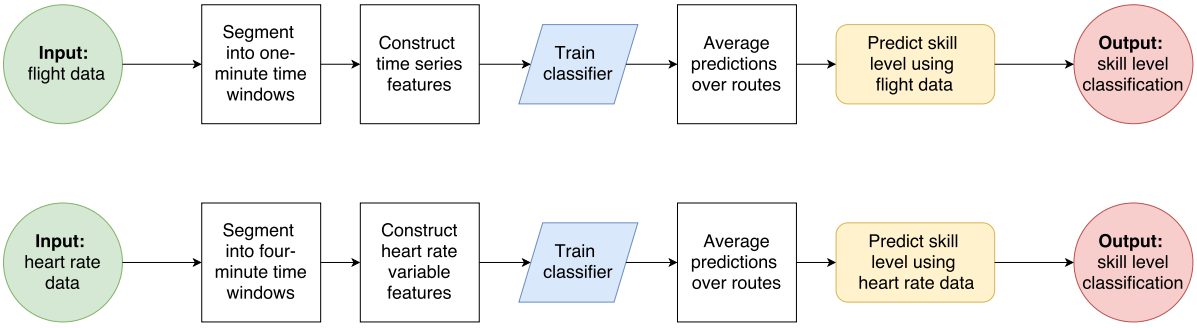

Figure 4-1: Data analytics pipeline for pilot skill level prediction using flight inputs and heart rate.

The pipeline of the analysis followed for skill level prediction is shown in Figure 4-1.

As previously stated, both HR and flight data features are extracted duly con-

sidering the respective time windows and normalized to the ranges of [-1,1]. AUC was chosen as the performance metric. The supervised classification algorithms were tuned for specific parameters using a grid search where parameters with the highest AUC were retained for classification. All of the above methods involved a binary classification output, in which the classes were differentiated as a novice or expert.

Classification was performed using two different analysis patterns to investigate the advantages of a SSA. The coarse analysis detected the average skill predictions-per-route for each pilot, while the fine analysis detected the skill predictions across each of the time windows in a particular route. Since the prediction of skill level across each window is likely to be noisy, more focus was put on the results from the coarse analysis.

4.6.2 Mental Workload and Effort Prediction

The features extracted from the flight and HR data are used to determine the amount of mental workload on the pilots. A similar process was carried out to measure the amount of effort put in by the pilots. The RMSE and R^2 score were chosen as the performance metrics. A grid-search was performed on the parameters for all supervised regression algorithms, and the parameters with the least mean-squared-error were retained for prediction. Two different analysis approaches were conducted:

- Single-Stage Approach
- Two-Stage Approach

4.6.2.1 Single-Stage Approach

The single-stage analysis pipeline for mental workload and effort prediction for the single-stage prediction is illustrated in Figure 4-2. As seen in the figure, the pipeline was similar to the classification analysis only to be replaced by the regression

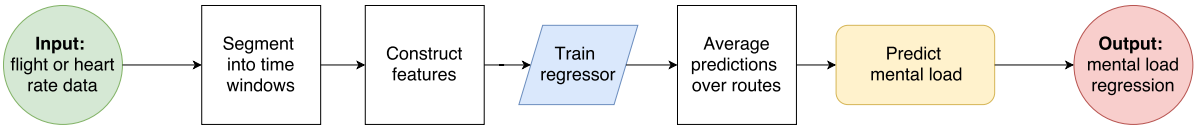

Figure 4-2: Data analytics pipeline for single-stage pilot mental workload and effort prediction using either flight inputs or heart rate data.

models and corresponding mental workload and effort prediction. The two data sources (flight and HR) were also combined to learn the importance of data fusion. A convex combination of the individual model predictions was performed. This convex combination model was optimized to retain the least RMSE using the same leave-a-subject-out and leave-a-route-out validation methods as for skill level prediction. The convex combination was chosen over training a single model on an enlarged feature set due to the differences in the lengths of time windows for flight control and HR data.

4.6.2.2 Two-Stage Approach

From the results discussed in Chapter 6, it was seen that the single-stage approach was unable to accurately predict mental workload.

In an effort to improve the prediction, a two-stage regression approach was proposed. The analysis pipeline for the two-stage regression is shown in Figure 4-3. In this approach, two individual expert and novices regression models were trained, each with both expert and novice pilot data keeping in mind the ground truth. To improve the prediction model, the predicted probabilities attained from the skill classification (logistic regression) were also incorporated in the analysis. The predicted class probabilities were appropriately multiplied with the predictions from the regression models and summed to attain a final prediction.

This method was considered citing a real-time setting where an arbitrary pilot's

workload is to be determined. A convex combination of the flight and HR data prediction was used to determine the final mental workload and effort prediction.

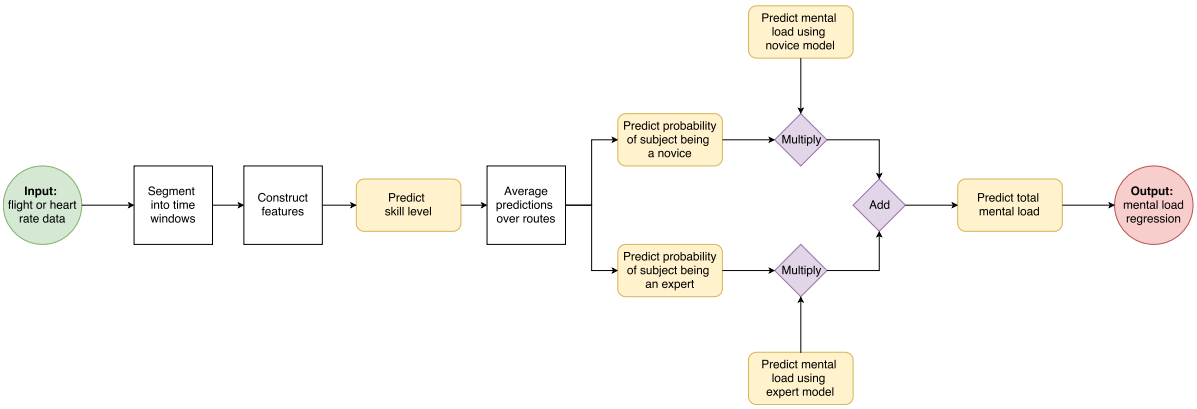

Figure 4-3: Data analytics pipeline for proposed two-stage pilot mental workload and effort prediction by first predicting skill level then combining two different mental load and effort regression models. The pipeline can be applied to either flight inputs or heart rate data.

Chapter 5

Results and Discussion

5.1 Skill Level Prediction

The results of the coarse analysis for skill level prediction are shown in Tables 5.1 and 5.2, respectively.

Table 5.1: Skill level prediction AUC using leave-a-subject-out CV.

Leave-a-Subject-out Validation		
Algorithm	Flight only	HR only
Logistic Regression- L1 Penalty	0.95	0.67
SVM-RBF kernel	0.99	0.66
Logistic Regression	0.99	0.67
SVM-Linear kernel	0.88	0.64
Random Forest	0.95	0.46
kNN	0.99	0.49

It is seen that, by using only the flight data-derived features (flight only), the classification model was successfully able to predict the skill level of the subjects, and a near-perfect AUC was attained in each setting. This indicated that the model was able to reliably predict the skill level of a pilot based on the way the pilot controls flight. On the other hand, using only HR data resulted in an inferior classifier when compared to the flight data.

To understand the importance of the different features constructed from the flight

Table 5.2: Skill level prediction AUC using leave-a-route-out CV.

Leave-a-Route-out Validation		
Algorithm	Flight only	HR only
Logistic Regression- L1 Penalty	1	0.77
SVM-RBF kernel	1	0.86
Logistic Regression	1	0.82
SVM-Linear kernel	0.94	0.83
Random Forest	1	0.84
kNN	1	0.81

control data, features with non-zero weights in the L1-penalized logistic regression were examined. The mean numbers of features over both the validation methods for flight control data are shown in Table 5.3. From the number of features selected, it is apparent that the two most important variables are the speed and heading. It is likely that the experts' prior flight experience allowed them to fly through the routes both faster and more on-course compared to the novices. All the 15 HR data features were deemed important by the L1-penalized logistic regression.

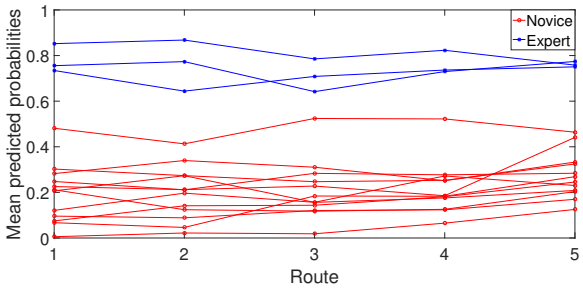

Figure 5-1: Mean predicted probabilities per route using logistic regression in the leave-a-route-out validation.

The mean predicted probabilities per route (coarse analysis) over all subjects in the leave-a-route-out validation setting with logistic regression is shown in Figure 5-1. The coarse analysis in the leave-a-subject-out validation is illustrated in Figure 5-2. From Figure 5-1, it can be seen that separating experts from novices is extremely straightforward in the leave-a-route-out setting. It is more impressive that separating

Table 5.3: Importance of flight data variables: number of features selected (calculated by the 'L1' penalty); Mean (SD).

Flight data variables	Leave-a-Subject-out Validation	Leave-a-Route-out Validation
Throttle	20.53 (1.36)	20.40 (2.19)
Aileron	19.93 (1.62)	19.20 (2.16)
Elevator	16.60 (1.45)	16.60 (3.78)
Heading	24.26 (1.22)	23.20 (0.83)
Longitude	15.46 (1.30)	15.20 (2.77)
Altitude	20.26 (1.22)	19.40 (2.60)
Rudder	14.66 (1.67)	13.60 (2.07)
Latitude	14.13 (1.35)	13.40 (2.70)
Speed	26.40 (1.18)	26.40 (1.52)

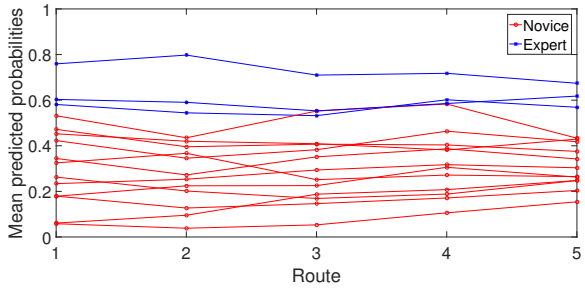

Figure 5-2: Mean predicted probabilities per route using logistic regression in the leave-a-subject-out validation.

experts from novices can be done quite accurately even in the leave-a-subject-out setting as shown in Figure 5-2. Notice that there are only 2 pilots, 1 novice and 1 expert, for which the algorithm makes incorrect predictions on any route.

The fine analysis was conducted to study the skill prediction at each time window. The theory that, as time progresses, the results from the fine analysis would match that from the coarse analysis is supported by Figures 5-3 and 5-4. This aligned with the experimental findings. The results from the fine analysis involving the individual time windows over the leave-a-subject-out validation and leave-a-route-out validation are shown in Figures 5-3 and 5-4, respectively. The prediction AUCs are over each of the first 5 minutes for each route. Notice from Figure 5-4 that in the leave-a-route-out setting it is even possible to accurately predict skill level on a minute-by-minute basis

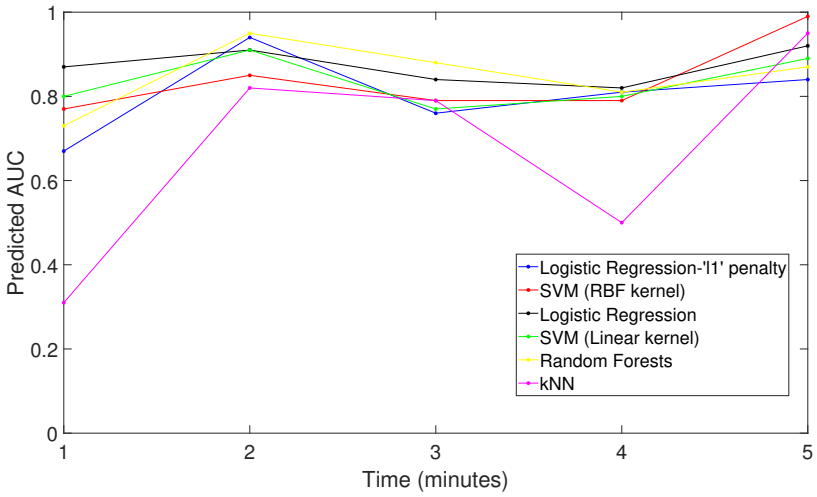

Figure 5-3: Predicted probabilities over individual time windows for all classifiers in the leave-a-subject-out validation.

without any averaging or smoothing over time. This indicated that the proposed model is robust to the introduction of new routes. On the other hand, from Figure 5-3, the prediction AUCs on a minute-by-minute basis are lower than in the coarse analysis setting. This is also a reasonable result, as it should certainly be difficult to predict whether a pilot is an expert or novice from only observing one minute of flight from that pilot. After averaging over 10-15 minutes of flight as done in the coarse analysis, a near-perfect prediction of skill level was obtained.

5.2 Mental Workload Prediction

The results of the mental workload prediction from the single-stage approach are tabulated in Tables 5.4 and 5.5, respectively. The key observation is that the individual regression models were unable to predict the mental workload. A combination of the two data sources did not help in the prediction either. Better results were achieved in the leave-a-route-out setting, where the variation due to differences in self-report scales of the pilots are no longer present. In this setting, an acceptable

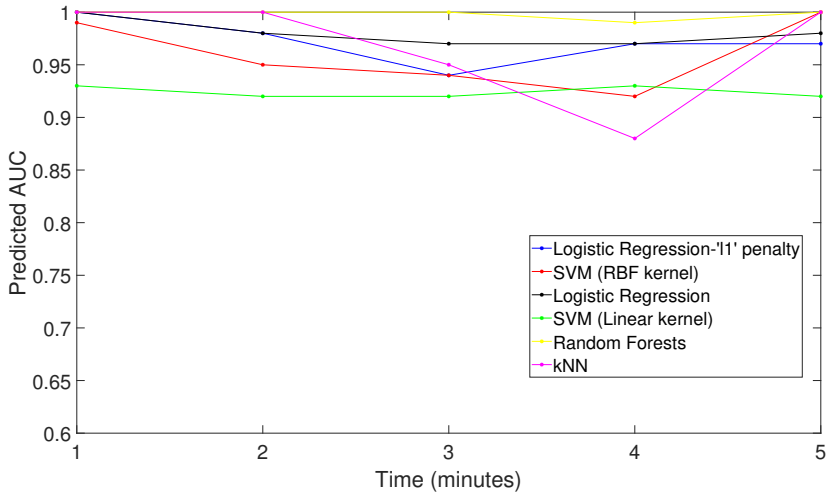

Figure 5-4: Predicted probabilities over individual time windows for all classifiers in the leave-a-route-out validation.

R^2 score of 0.53 was obtained. It is notable that the kernel-based SVR model did a better job in predicting the mental workload in this setting when compared to the linear LASSO model and the ensemble tree model.

Table 5.4: Mental workload prediction using leave-a-subject-out CV in the single-stage approach; RMSE (R^2).

Leave-a-Subject-out Validation			
Algorithm	Flight only	HR only	All features
SVR	4.25 (-0.02)	4.30 (-0.04)	4.19 (0)
LASSO	4.28 (-0.03)	4.40 (-0.09)	4.22 (0)
Random Forests	4.31 (-0.05)	4.41 (-0.10)	3.59 (0.26)

The two-stage approach was considerably better when compared to the previous approach in the leave-a-subject-out setting, as seen in Table 5.6. Using the skill level prediction and the individual novice and expert regression models, the two-stage approach was able to more accurately predict the mental workload when combining both the flight and HR data, resulting in R^2 values around 0.3, compared to 0 for the single-stage predictor, which is seen from Tables 5.6 and 5.7 respectively. The two-stage predictor provided a slight improvement in the leave-a-route-out setting

Table 5.5: Mental workload prediction using leave-a-route-out CV in the single-stage approach; RMSE (R^2).

Leave-a-Route-out Validation			
Algorithm	Flight only	HR only	All features
SVR	2.86 (0.54)	3.96 (0.12)	2.86 (0.54)
LASSO	3.95 (0.12)	4.06 (0.07)	3.87 (0.15)
Random Forests	2.92 (0.52)	4.08 (0.07)	2.92 (0.52)

also (R^2 values around 0.58).

Using a weighted combination of skill level probabilities and individual regression models, this approach was considered as it reflects a real-world scenario in which an out of the data-set sample/subject would be evaluated provided the skill level is unknown. The convex combination of the HR and flight data was able to better predict the mental workload when compared to the individual data set models. This is seen in both approaches.

For the above mentioned approaches, flight control data only analysis with all routes in consideration was also conducted to include the 7 additional routes without HR data. This did not yield any significant change in the results.

Table 5.6: Mental workload prediction using leave-a-subject-out CV in the two-stage approach; RMSE (R^2).

Leave-a-Subject-out Validation			
Algorithm	Flight only	HR only	All features
SVR	4.09 (0.05)	4.15 (0.03)	3.50 (0.30)
LASSO	4.18 (0.01)	4.27 (0.01)	3.65 (0.23)
Random Forests	4.19 (0.01)	4.59 (-0.19)	3.68 (0.23)

Table 5.7: Mental workload prediction using leave-a-route-out CV in the two-stage approach; RMSE (R^2).

Leave-a-Route-out Validation			
Algorithm	Flight only	HR only	All features
SVR	2.71 (0.59)	3.45 (0.33)	2.70 (0.59)
LASSO	3.40 (0.35)	3.73 (0.21)	3.34 (0.37)
Random Forests	2.78 (0.57)	3.94 (0.12)	2.78 (0.57)

5.3 Effort Prediction

A similar process to that of the mental workload prediction was followed for effort prediction. The results from the single stage approach for the total effort prediction are shown in Tables 5.8 and 5.9, respectively. As observed for the mental workload prediction, the single stage regression models were not able to predict the total effort, and better results were achieved in the leave-a-route-out validation.

The combined model, using both the flight and HR data derived features, added little value to the prediction. R^2 scores in the ranges of 0.2-0.3 were achieved in the leave-a-route-out validation. The kernel based SVR model did a a better job in predicting the effort when compared to the other methods.

Table 5.8: Effort prediction using leave-a-subject-out CV in the single-stage approach; RMSE (R^2).

Leave-a-Subject-out Validation			
Algorithm	Flight only	HR only	All features
SVR	3.86 (-0.03)	3.83 (-0.01)	3.60 (0.11)
LASSO	4 (-0.10)	3.82 (-0.01)	3.63 (0.09)
Random Forests	4.23 (-0.23)	4.09 (-0.15)	3.59 (0.11)

The two-stage approach for effort prediction performed comparatively better than the single-stage approach, as seen in Tables 5.10 and 5.11, respectively. The leave-a-route-out validation setting performed the effort prediction better than the leave-a-subject-out validation as was seen for the mental workload prediction.

The convex combination of both data sources resulted in R^2 values of around 0.4

Table 5.9: Effort prediction using leave-a-route-out CV in the single-stage approach; RMSE (R^2).

Leave-a-Route-out Validation			
Algorithm	Flight only	HR only	All features
SVR	3.00 (0.38)	3.78 (0.01)	2.99 (0.39)
LASSO	3.56 (0.13)	3.70 (0.06)	3.48 (0.17)
Random Forests	3.05 (0.36)	3.87 (-0.03)	3.01 (0.38)

in the leave-a-route-out setting. The convex combination of data proved to provide better results when compared to the individual data models and the same was seen in both validation approaches.

Table 5.10: Effort prediction using leave-a-subject-out CV in the two-stage approach; RMSE (R^2).

Leave-a-Subject-out Validation			
Algorithm	Flight only	HR only	All features
SVR	3.65 (0.08)	3.65 (0.08)	3.16 (0.31)
LASSO	4.02 (-0.01)	3.94 (-0.07)	3.52 (0.14)
Random Forests	4.25 (-0.24)	4.19 (-0.21)	3.39 (0.20)

Table 5.11: Effort prediction using leave-a-route-out CV in the two-stage approach; RMSE (R^2).

Leave-a-Route-out Validation			
Algorithm	Flight only	HR only	All features
SVR	2.92 (0.41)	3.31 (0.25)	2.87 (0.43)
LASSO	3.28 (0.26)	3.47 (0.17)	3.24 (0.28)
Random Forests	3.02 (0.37)	3.73 (0.04)	3.00 (0.38)

Chapter 6

Conclusive Remarks

The end goal of this research was to investigate multiple machine learning algorithms and concepts, and to incorporate these methods to develop a novel pipeline analysis for the prediction of pilot skill level and workload (mental workload and effort). Hence, analysis approaches for skill level, mental workload, and effort prediction were presented in Chapter 4.

After a brief literature review focusing on autonomy and workload in Chapter 2, Chapter 3 elaborated on the data collection and experimental set-up procedure followed by Wright State Research Institute and Perduco.

We found that by using flight and HR data, it is possible to predict the skill level of the subjects. On the other hand, it was found that standard regression models were unable to predict workload accurately. This was seen for all data sources (flight only, HR only, and both).

When comparing the multiple data sources, it was found that the HR data added little value to the prediction. This does not imply that the HR data provided little information, but rather could simply mean that the features extracted were not ideal for the above mentioned predictions. The prediction results from the HR data analysis could also call for better physiological data signals to be collected, such as electrodermal activity (EDA), photoplethysmography (PPG) data, and electrocardiography

(ECG) data. Another issue with the HR data results could be attributed to the low sampling rate of the HR data collection equipment. Researchers have suggested that HRV analysis be executed with data that has a sampling rate of more than 250 Hz [147].

For skill level prediction, using just the flight data derived features, a near perfect prediction accuracy was achieved, and the pipeline suggested in Figure 4-1 could be used in a real-time setting to help aid the research. The collection of the various flight control input signals proved beneficial to determine the differences in an expert and novice's aviation abilities. From the flight control input data importances, we also learned that the heading and speed variables were considered to be most important in the model learning for skill level prediction.

Mental workload and effort prediction were considered as these variables were found to be the most important parameters for pilot prediction in the current experimental settings.

For both mental workload and effort prediction, the two-stage approach incorporating the individual regression models and the probability estimates from the skill level prediction proved to be a more comprehensive approach, when compared to the single stage model, which was quite straightforward. For both approaches, the convex combination of the flight and HR data produced the better results among all comparisons. This helped us to ascertain the importance of data fusion in predictive modeling and, more importantly, workload prediction.

An additional issue with the HR data is the presence of motion artifacts that may lead to unwanted solutions. We manually adjusted all discrepancies in the data, but this approach would not be ideal in a real-time setting.

The SVR model with the RBF kernel worked out to be the better choice in mental workload and effort predictions in both validation settings. For classification, all algorithms employed produced competitive accuracies.

In the experiments, no artifact detection methods were used. Employing an artifact detection algorithm would have likely increased the prediction accuracy for each algorithm, but its effect on the combined model is unclear. Overall, however, the progress made by our approach will be quite beneficial to the advancements of human performance, autonomy, and human-computer interaction as a whole.

6.1 Future Work

Skill and workload prediction is one of the first steps in building a human-machine team with the ultimate goal of modifying the level of task assignment based on the workload feedback determined as the experiment progresses.

One of the areas of potential future work could be focused on feature extraction. The most common statistical based values were used to derive features. A wider variety of features especially for the HR data could be computed in order to have a better effect on the predictions. As mentioned in the previous section, workload measures could be better predicted by collecting additional physiological data signals such as EDA, PPG and ECG and performing appropriate feature construction.

These approaches could also be modified to a more complex learning scenario by using semi-supervised or active learning methods. This would greatly help the experiments to be evaluated in a real-time setting thereby providing feedback to the subjects on-the-fly. Methods such as deep neural networks can also be used, as they have the capability to automatically learn features from the raw data.

References

[1] National Research Council. *Autonomous Vehicles in Support of Naval Operations*. The National Academies Press, Washington, DC, 2005.

[2] M. L. Cummings, Carl E. Nehme, Jacob Crandall, and Paul Mitchell. *Predicting Operator Capacity for Supervisory Control of Multiple UAVs*, pages 11–37. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[3] Michael A. Goodrich, Dan R. Olsen, Jacob W. Cr, and Thomas J. Palmer. Experiments in adjustable autonomy. pages 1624–1629, 2001.

[4] D. J. Bruemmer, D. D. Dudenhoeffer, and J. Marble. Dynamic autonomy for urban search and rescue. In *2002 AAAI Mobile Robot Workshop, Edmonton, Canada*, 2002.

[5] Munjal Desai. *Sliding scale autonomy and trust in human-robot interaction*. Ms thesis, University of Massachusetts Lowell, 2007.

[6] H. A. Yanco and J. Drury. "where am i?" acquiring situation awareness using a remote robot platform. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, volume 3, pages 2835 – 2840 vol.3, Oct 2004.

[7] Kuntao Cui, Zuochang Yang, and Wenli Sun. The collaborative autonomy and control framework for unmanned surface vehicle. In *Frontier of Computer Science and Technology (FCST), 2015 Ninth International Conference on*, pages 242–247. IEEE, 2015.

[8] Scott Lenser and Chris Jones. Practical problems in sliding scale autonomy: a case study, 2008.

[9] Douglas Wiegmann, Juliana Goh, and David O'Hare. The role of situation assessment and flight experience in pilots' decisions to continue visual flight rules flight into adverse weather. 44:189–97, 02 2002.

[10] Angela T Schriver, Daniel Morrow, Christopher Wickens, and Donald Talleur. Expertise differences in attentional strategies related to pilot decision making. 50:864–78, 01 2009.

[11] Mark Wiggins and David O'Hare. Expertise in aeronautical weather-related decision making: A cross-sectional analysis of general aviation pilots. 1:305–320, 12 1995.

[12] Barry H Kantowitz and Patricia A Casper. Human workload in aviation. 1988.

[13] Caroline Dussault, Jean-Claude Jouanin, Matthieu Philippe, and Charles Yannick Guezennec. Eeg and ecg changes during simulator operation reflect mental workload and vigilance. *Aviation, Space, and Environmental Medicine*, 76(4):344–351, 2005.

[14] Sandra G. Hart and Lowell E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in Psychology*, 52:139 – 183, 1988. Human Mental Workload.

[15] Angela DiDomenico and Maury A. Nussbaum. Interactive effects of physical and mental workload on subjective workload assessment. *International Journal of Industrial Ergonomics*, 38(11):977 – 983, 2008.

[16] John Annett. Subjective rating scales: science or art? *Ergonomics*, 45(14):966–987, 2002. PMID: 12569049.

[17] Glenn F Wilson and F Thomas Eggemeier. Psychophysiological assessment of workload in multi-task environments. *Multiple-task performance*, 329360, 1991.

[18] Bruce J Noble and Robert J Robertson. *Perceived exertion*. Human Kinetics Publishers, 1996.

[19] Pete Hancock and Raja Parasuraman. Human factors and safety in the design of intelligent vehicle-highway systems (ivhs). 23:181–198, 12 1992.

[20] P.A. Hancock and W.B. Verwey. Fatigue, workload and adaptive driver systems. *Accident Analysis and Prevention*, 29(4):495 – 506, 1997. Fatigue and Transport.

[21] Najmedin Meshkati, Peter Hancock, Mansour Rahimi, and Suzanne M. Dawes. Techniques in mental workload assessment. 01 1995.

[22] Raja Parasuraman, James Christensen, and Scott Grafton. Neuroergonomics: The brain in action and at work. 59:1–3, 08 2011.

[23] Raja Parasuraman and Glenn F. Wilson. Putting the brain to work: Neuroergonomics past, present, and future. *Human Factors*, 50(3):468–474, 2008. PMID: 18689055.

[24] Carolyn Ells, Matthew R Hunt, and Jane Chambers-Evans. Relational autonomy as an essential component of patient-centered care. *IJFAB: International Journal of Feminist Approaches to Bioethics*, 4(2):79–101, 2011.

[25] David P Ellerman. *Helping people help themselves: autonomy-compatible assistance*. The World Bank, 2000.

[26] David J Bruemmer, Donald D Dudenhoeffer, and Julie L Marble. Dynamic-autonomy for urban search and rescue. In *AAAI mobile robot competition*, pages 33–37, 2002.

[27] Walter Berka, Jan De Groof, and Hilde Penneman. *Autonomy in education*, volume 3. Springer Science & Business Media, 2000.

[28] Sebastian Thrun. Toward a framework for human-robot interaction. *Human–Computer Interaction*, 19(1-2):9–24, 2004.

[29] Thomas B Sheridan and William L Verplank. Human and computer control of undersea teleoperators. Technical report, MASSACHUSETTS INST OF TECH CAMBRIDGE MAN-MACHINE SYSTEMS LAB, 1978.

[30] Mica R Endsley. Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics*, 42(3):462 – 492, 1999.

[31] Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, 30(3):286–297, 2000.

[32] Hui-Min Huang. Autonomy levels for unmanned systems (alfus) framework volume i: Terminology version 2.0. *Special Publication (NIST SP)-1011*, 2004.

[33] Hui-Min Huang, Kerry Pavek, Mark Ragon, Jeffry Jones, Elena Messina, and James Albus. Characterizing unmanned system autonomy: Contextual autonomous capability and level of autonomy analyses. In *Unmanned Systems Technology IX*, volume 6561, page 65611N. International Society for Optics and Photonics, 2007.

[34] Hui-Min Huang, Kerry Pavek, Brian Novak, James Albus, and E Messin. A framework for autonomy levels for unmanned systems (alfus). *Proceedings of the AUVSI's Unmanned Systems North America*, pages 849–863, 2005.

[35] Holly A Yanco and Jill Drury. Classifying human-robot interaction: an updated taxonomy. In *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, volume 3, pages 2841–2846. IEEE, 2004.

[36] Michael A Goodrich and Dan R Olsen. Seven principles of efficient human robot interaction. In *Systems, Man and Cybernetics, 2003. IEEE International Conference on*, volume 4, pages 3942–3948. IEEE, 2003.

[37] David Feil-Seifer, Kristine Skinner, and Maja J Matarić. Benchmarks for evaluating socially assistive robotics. *Interaction Studies*, 8(3):423–439, 2007.

[38] Michael A Goodrich and Alan C Schultz. Human-robot interaction: a survey. *Foundations and trends in human-computer interaction*, 1(3):203–275, 2007.

[39] Matthew Johnson, Jeffrey M Bradshaw, Paul J Feltovich, Catholijn Jonker, Maarten Sierhuis, and Birna van Riemsdijk. Toward coactivity. In *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*, pages 101–102. IEEE Press, 2010.

[40] Robin R Murphy and Debra Schreckenghost. Survey of metrics for human-robot interaction. In *Human-Robot Interaction (HRI), 2013 8th ACM/IEEE International Conference on*, pages 197–198. IEEE, 2013.

[41] Jean Scholtz. Theory and evaluation of human robot interactions. In *System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on*, pages 10–pp. IEEE, 2003.

[42] Jean C Scholtz. Human-robot interactions: Creating synergistic cyber forces. In *Multi-Robot Systems: From Swarms to Intelligent Automata*, pages 177–184. Springer, 2002.

[43] Mica R Endsley. *Designing for situation awareness: An approach to user-centered design*. CRC press, 2016.

[44] Mica R Endsley and Esin O Kiris. The out-of-the-loop performance problem and level of control in automation. *Human factors*, 37(2):381–394, 1995.

[45] Jeffrey M Bradshaw, Paul J Feltovich, Hyuckchul Jung, Shriniwas Kulkarni, William Taysom, and Andrzej Uszok. Dimensions of adjustable autonomy and mixed-initiative interaction. In *International Workshop on Computational Autonomy*, pages 17–39. Springer, 2003.

[46] Martha E Pollack, Ioannis Tsamardinos, and John F Horty. Adjustable autonomy for a plan management agent. In *AAAI Spring Symposium on Agents with Adjustable Autonomy*, pages 22–24, 1999.

[47] Paul Scerri, David Pynadath, and Milind Tambe. Adjustable autonomy in real-world multi-agent environments. In *Proceedings of the fifth international conference on Autonomous agents*, pages 300–307. ACM, 2001.

[48] Dennis Perzanowski, Alan C Schultz, William Adams, and Elaine Marsh. Goal tracking in a natural language interface: Towards achieving adjustable autonomy. In *Computational Intelligence in Robotics and Automation, 1999. CIRA'99. Proceedings. 1999 IEEE International Symposium on*, pages 208–213. IEEE, 1999.

[49] George Ferguson, James F Allen, Bradford W Miller, et al. Trains-95: Towards a mixed-initiative planning assistant. In *AIPS*, pages 70–77, 1996.

[50] Toshiyuki Inagaki. Trust, autonomy, and authority in human-machine systems: Situation-adaptive coordination for systems safety. *Proc. Cognitive Systems Engineering for Process Control 96'*, pages 176–183, 1996.

[51] Toshiyuki Inagaki. Situation-adaptive responsibility allocation for human-centered automation. *Transactions of the Society of Instrument and Control Engineers*, 31(3):292–298, 1995.

[52] David Kortenkamp, Eric Huber, R Peter Bonasso, et al. Recognizing and interpreting gestures on a mobile robot. In *Proceedings of the National Conference on Artificial Intelligence*, pages 915–921, 1996.

[53] Richard M Voyles and Pradeep K Khosla. Tactile gestures for human/robot interaction. In *Intelligent Robots and Systems 95.'Human Robot Interaction and Cooperative Robots', Proceedings. 1995 IEEE/RSJ International Conference on*, volume 3, pages 7–13. IEEE, 1995.

[54] Cynthia Breazeal et al. A motivational system for regulating human-robot interaction. In *Aaai/iaai*, pages 54–61, 1998.

[55] F Steele, Geb Thomas, and Theodore Blackmon. An operator interface for a robot-mounted, 3d camera system: Project pioneer. In *Virtual Reality, 1999. Proceedings., IEEE*, pages 126–132. IEEE, 1999.

[56] Richard M Voyles and Pradeep K Khosla. A multi-agent system for programming robots by human demonstration. *Integrated Computer-Aided Engineering*, 8(1):59–67, 2001.

[57] Robin R Murphy and Erika Rogers. Cooperative assistance for remote robot supervision. *Presence: Teleoperators & Virtual Environments*, 5(2):224–240, 1996.

[58] Terrence Fong, Charles Thorpe, and Charles Baur. A safeguarded teleoperation controller. In *IEEE International Conference on Advanced Robotics (ICAR)*, number LSRO2-CONF-2001-002, 2001.

[59] Thomas Rofer and Axel Lankenau. Ensuring safe obstacle avoidance in a shared-control system. In *Emerging Technologies and Factory Automation, 1999. Proceedings. ETFA'99. 1999 7th IEEE International Conference on*, volume 2, pages 1405–1414. IEEE, 1999.

[60] Daniel Gopher and Emanuel Donchin. Workload: An examination of the concept. 1986.

[61] Erick Farmer and Adam Brownson. Review of workload measurement, analysis and interpretation methods. *European Organisation for the Safety of Air Navigation*, 33:1–33, 2003.

[62] M Castor, E Hanson, E Svensson, S Nählinder, P LeBlaye, I MacLeod, N Wright, J Alfredson, L Ågren, P Berggren, et al. Garteur handbook of mental workload measurement. *GARTEUR, Group for Aeronautical Research and Technology in Europe, Flight Mechanics Action Group FM AG13*, 164, 2003.

[63] William B Rouse, Sharon L Edwards, and John M Hammer. Modeling the dynamics of mental workload and human performance in complex systems. *IEEE transactions on systems, man, and cybernetics*, 23(6):1662–1671, 1993.

[64] Robert D O'Donnell and F Thomas Eggemeier. Workload assessment methodology. 1986.

[65] TF Meijman and G Mulder. Arbeidspsychologische aspecten van werkbelasting.(workpsychological aspects of work-demand) in pjd drenth, h. thierry & ch. j. de wolf (red.). *Nieuw handboek A&O psychologie*, 1992.

[66] F Zijlstra and Th Meijman. Het meten van mentale inspanning met behulp van een subjectieve methode (measurement of mental effort with a subjective method). *Mentale belasting en werkstress. Een arbeidspsychologische benadering*, pages 42–61, 1989.

[67] Wim B Verwey. Adaptable driver-car interfacing and mental workload: a review of the literature. Technical report, INSTITUTE FOR PERCEPTION RVO-TNO SOESTERBERG (NETHERLANDS), 1990.

[68] Walter Schneider, Sue T Dumais, and Richard M Shiffrin. Automatic/control processing and attention. Technical report, ILLINOIS UNIV CHAMPAIGN HUMAN ATTENTION RESEARCH LAB, 1982.

[69] Barry H Kantowitz. Heavy vehicle driver workload assessment: Lessons from aviation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 36, pages 1113–1117. SAGE Publications Sage CA: Los Angeles, CA, 1992.

[70] Walter W Wierwille and F Thomas Eggemeier. Recommendations for mental workload measurement in a test and evaluation environment. *Human Factors*, 35(2):263–281, 1993.

[71] Margaret M Bradley, Laura Miccoli, Miguel A Escrig, and Peter J Lang. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, 45(4):602–607, 2008.

[72] Shravani Sur and VK Sinha. Event-related potential: An overview. *Industrial psychiatry journal*, 18(1):70, 2009.

[73] Alan H Roscoe. Assessing pilot workload. why measure heart rate, hrv and respiration? *Biological psychology*, 34(2):259–287, 1992.

[74] Glenn F Wilson. Applied use of cardiac and respiration measures: Practical considerations and precautions. *Biological Psychology*, 34(2):163–178, 1992.

[75] Dhong H Lee and Kyung S Park. Multivariate analysis of mental and physical load components in sinus arrhythmia scores. *Ergonomics*, 33(1):35–47, 1990.

[76] PGAM Jorna. Heart rate and workload variations in actual and simulated flight. *Ergonomics*, 36(9):1043–1054, 1993.

[77] David J Mascord and Richard A Heath. Behavioral and physiological indices of fatigue in a visual tracking task. *Journal of safety research*, 23(1):19–25, 1992.

[78] David J Mascord, Jeannie Walls, and Graham A Starmer. 20 fatigue and alcohol: interactive effects on human performance in driving-related tasks. *Fatigue and driving: Driver impairment, driver fatigue, and driving simulation*, page 189, 1995.

[79] JWH Kalsbeek and JH Ettema. Scored regularity of the heart rate pattern and the measurement of perceptual or mental load. *Ergonomics*, 6(3):306–307, 1963.

[80] Walter W Wierwille, Mansour Rahimi, and John G Casali. Evaluation of 16 measures of mental workload using a simulated flight task emphasizing mediational activity. *Human Factors*, 27(5):489–502, 1985.

[81] J Gonzalez Gonzalez, A Mendez Llorens, A Mendez Novoa, and JJ Cordero Valeriano. Effect of acute alcohol ingestion on short-term heart rate fluctuations. *Journal of studies on Alcohol*, 53(1):86–90, 1992.

[82] Yasuko Itoh, Yoshio Hayashi, Ippei Tsukui, and Susumu Saito. The ergonomic evaluation of eye movement and mental workload in aircraft pilots. *Ergonomics*, 33(6):719–732, 1990.

[83] JA Veltman and AWK Gaillard. Indices of mental workload in a complex task environment. *Neuropsychobiology*, 28(1-2):72–75, 1993.

[84] Richard W Backs and Kimberle A Seljos. Metabolic and cardiorespiratory

measures of mental effort: the effects of level of difficulty in a working memory task. *International Journal of psychophysiology*, 16(1):57–68, 1994.

[85] Peter GAM Jorna. Spectral analysis of heart rate and psychological state: A review of its validity as a workload index. *Biological psychology*, 34(2):237–257, 1992.

[86] Fred GWC Paas, Jeroen JG Van Merriënboer, and Jos J Adam. Measurement of cognitive load in instructional research. *Perceptual and motor skills*, 79(1):419–430, 1994.

[87] Jans Aasman, Gijsbertus Mulder, and Lambertus JM Mulder. Operator effort and the measurement of heart-rate variability. *Human factors*, 29(2):161–170, 1987.

[88] Lambertus Johannes Maria Mulder. *Assessment of cardiovascular reactivity by means of spectral analysis*. PhD thesis, Rijksuniversiteit, 1988.

[89] Erik J Sirevaag, Arthur F Kramer, CHRISTOPHER D WICKENS MARK REISWEBER, DAVID L STRAYER, and JAMES F GRENELL. Assessment of pilot performance and mental workload in rotary wing aircraft. *Ergonomics*, 36(9):1121–1140, 1993.

[90] Stephen W Porges and Evan A Byrne. Research methods for measurement of heart rate and respiration. *Biological psychology*, 34(2):93–130, 1992.

[91] Gijsbertus Mulder. The heart of mental effort. *Unpublished doctoral dissertation, University of Groningen, Groningen, The Netherlands*, 1980.

[92] SG Hart and C Wickens. Workload assessment and prediction. manprint, an approach to systems integration, 257-296, 1990.

[93] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in psychology*, 52:139–183, 1988.

[94] Thomas E Nygren. Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload. *Human Factors*, 33(1):17–33, 1991.

[95] Ulla Metzger and Raja Parasuraman. Automation in future air traffic management: Effects of decision aid reliability on controller performance and mental workload. *Human Factors*, 47(1):35–49, 2005. PMID: 15960085.

[96] Mica R Endsley. Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics*, 42(3):462–492, 1999.

[97] James L Szalma, Joel S Warm, Gerald Matthews, William N Dember, Ernest M Weiler, Ashley Meier, and F Thomas Eggemeier. Effects of sensory modality and task duration on performance, workload, and stress in sustained attention. *Human factors*, 46(2):219–233, 2004.

[98] Matthew B Weinger, Alison G Vredenburgh, Cynthia Mills Schumann, Alex Macario, Kevin J Williams, Michael J Kalsher, Brian Smith, Phuong C Truong, and Ann Kim. Quantitative description of the workload associated with airway management procedures. *Journal of Clinical Anesthesia*, 12(4):273 – 282, 2000.

[99] Matthew B Weinger, Swapna B Reddy, and Jason M Slagle. Multiple measures of anesthesia workload during teaching and nonteaching cases. *Anesthesia & Analgesia*, 98(5):1419–1425, 2004.

[100] Scott Levin, Daniel France, Robin Hemphill, Ian Jones, Kong Chen, Dorsey Rickard, Renee Makowski, and Dominik Aronsky. Tracking workload in the emergency department. 48:526–39, 02 2006.

[101] Yuliya Yurko, Mark Scerbo, Ajita S Prabhu, Christina E Acker, and Dimitrios Stefanidis. Higher mental workload is associated with poorer laparoscopic performance as measured by the nasa-tlx tool. 5:267–71, 10 2010.

[102] Susana Valdehita, Eva Ramiro, Jess Garca, and Jos M. Puente. Evaluation of subjective mental workload: a comparison of swat, nasa-tlx, and workload profile methods. 53:61 – 86, 01 2004.

[103] Alvah C Bittner Jr, James C Byers, Susan G Hill, Allen L Zaklad, and Richard E Christ. Generic workload ratings of a mobile air defense system (los-fh). In *Proceedings of the Human Factors Society Annual Meeting*, volume 33, pages 1476–1480. SAGE Publications Sage CA: Los Angeles, CA, 1989.

[104] William Moroney, D.W. Biers, F.T. Eggemeier, and J.A. Mitchell. A comparison of two scoring procedures with the nasa task load index in a simulated flight task. pages 734 – 740 vol.2, 06 1992.

[105] Alex Cao, Keshav K. Chintamani, Abhilash K. Pandya, and R. Darin Ellis. Nasa tlx: Software for assessing subjective mental workload. *Behavior Research Methods*, 41(1):113–117, Feb 2009.

[106] Stuart D. Baulk, Katie J. Kandelaars, Nicole Lamond, Gregory D. Roach, Drew Dawson, and Adam Fletcher. Does variation in workload affect fatigue in a regular 12-hour shift system? *Sleep and Biological Rhythms*, 5(1):74–77, Jul 2007.

[107] Sean Reilley, Anthony F. Grasha, Gerald Matthews, and John Schafer. Automatic-controlled information processing and error detection in a simulated pharmacy-verification task. *Perceptual and Motor Skills*, 97(1):151–174, 2003. PMID: 14604036.

[108] Carl W. Turner, Jennifer A. Safar, and Karthik Ramaswamy. The effects of use on acceptance and trust in voice authentication technology. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(5):718–722, 2006.

[109] Adams Greenwood-Ericksen, Tal Oron-Gilad, James L. Szalma, Shawn Stafford, and Peter A. Hancock. Workload and performance: A field evaluation in a police shooting range. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 48(16):1953–1957, 2004.

[110] David B Kaber, Emrah Onal, and Mica R Endsley. Design of automation for telerobots and the effect on performance, operator situation awareness, and subjective workload. *Human Factors and Ergonomics in Manufacturing*, 10(4):409–430, 2000.

[111] Shinji Miyake. Multivariate workload evaluation combining physiological and subjective measures. *International Journal of Psychophysiology*, 40(3):233 – 238, 2001. Psychophysiology in.

[112] Vernoi Battiste and Michael Bortolussi. Transport pilot workload: A comparison of two subjective techniques. *Proceedings of the Human Factors Society Annual Meeting*, 32(2):150–154, 1988.

[113] Susan G. Hill, Helene P. Iavecchia, Alvah C. Bittner, Jr., James C. Byers, Allen L. Zaklad, and Richard E. Christ. Comparison of four subjective workload rating scales. *Hum. Factors*, 34(4):429–439, August 1992.

[114] Gloria Young, Lyubov Zavelina, and Vallire Hooper. Assessment of workload using nasa task load index in perianesthesia nursing. *Journal of PeriAnesthesia Nursing*, 23(2):102 – 110, 2008.

[115] Quinn Kennedy, Joy L Taylor, Gordon Reade, and Jerome A Yesavage. Age

and expertise effects in aviation decision making and flight control in a flight simulator. *Aviation, space, and environmental medicine*, 81(5):489–497, 2010.

[116] Austin Meyer. X-plane. *Laminar Research*, 2007.

[117] L A Lipsitz, J Mietus, G B Moody, and A L Goldberger. Spectral characteristics of heart rate variability before and during postural tilt. relations to aging and risk of syncope. *Circulation*, 81(6):1803 – 1810, 1990.

[118] Media Anugerah Ayu, Siti Aisyah Ismail, Ahmad Faridi Abdul Matin, and Teddy Mantoro. A comparison study of classifier algorithms for mobile-phone's accelerometer based activity recognition. *Procedia Engineering*, 41:224 – 229, 2012.

[119] I. Daubechies. The wavelet transform, time-frequency localization and signal analysis. *IEEE Transactions on Information Theory*, 36(5):961 – 1005, Sep 1990.

[120] T. Zikov, S. Bibian, G. A. Dumont, M. Huzmezan, and C. R. Ries. A wavelet based de-noising technique for ocular artifact correction of the electroencephalogram. In *Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society] [Engineering in Medicine and Biology*, volume 1, pages 98 – 105 vol.1, 2002.

[121] Radomir S. Stankovi and Bogdan J. Falkowski. The haar wavelet transform: its status and achievements. *Computers & Electrical Engineering*, 29(1):25 – 44, 2003.

[122] A. Natarajan, K. S. Xu, and B. Eriksson. Detecting divisions of the autonomic nervous system using wearables. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5761 – 5764, Aug 2016.

[123] J. B. Bolkhovsky, C. G. Scully, and K. H. Chon. Statistical analysis of heart rate and heart rate variability monitoring through the use of smart phone cameras. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1610 – 1613, Aug 2012.

[124] P. Juszczak, D. M. J. Tax, and R. P. W. Duin. Feature scaling in support vector data description.

[125] Tim Salimans and Diederik P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 901 – 909. Curran Associates, Inc., 2016.

[126] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. A practical guide to support vector classification. 2003.

[127] Ralf Herbrich. *Learning Kernel Classifiers: Theory and Algorithms*. MIT Press, Cambridge, MA, USA, 2001.

[128] Sara Taylor, Natasha Jaques, Weixuan Chen, Szymon Fedor, Akane Sano, and Rosalind Picard. Automatic identification of artifacts in electrodermal activity data. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, pages 1934–1937. IEEE, 2015.

[129] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273 – 297, September 1995.

[130] Cornelis Joost Van Rijsbergen. Information retrieval. 1979.

[131] Vladimir Svetnik, Andy Liaw, Christopher Tong, J Christopher Culberson, Robert P Sheridan, and Bradley P Feuston. Random forest: a classification

and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43(6):1947–1958, 2003.

[132] Ramón Díaz-Uriarte and Sara Alvarez De Andres. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1):3, 2006.

[133] Bart Larivière and Dirk Van den Poel. Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications*, 29(2):472–484, 2005.

[134] D Richard Cutler, Thomas C Edwards, Karen H Beard, Adele Cutler, Kyle T Hess, Jacob Gibson, and Joshua J Lawler. Random forests for classification in ecology. *Ecology*, 88(11):2783–2792, 2007.

[135] David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 215–242, 1958.

[136] Joshua Goodman et al. Exponential priors for maximum entropy models. In *HLT-NAACL*, pages 305–312, 2004.

[137] Andrew Y Ng. Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, page 78. ACM, 2004.

[138] Naomi S Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[139] Steve Russell, Lisa A Meadows, and Roslin R Russell. *Microarray technology in practice*. Academic Press, 2008.

[140] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[141] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[142] Leo Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384, 1995.

[143] Robert Tibshirani. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.

[144] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[145] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[146] Gary King and Langche Zeng. Logistic regression in rare events data. *Political analysis*, 9(2):137–163, 2001.

[147] M. Z. Poh, D. J. McDuff, and R. W. Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Transactions on Biomedical Engineering*, 58(1):7 – 11, Jan 2011.

Appendix A

Detailed feature description and Additional Results

A.1 The detailed description of flight data features

Table A.1: Description of the flight data constructed features

Feature Index	Statistical Features	Flight data signal
1	mean	Signal
2	maximum	
3	minimum	
4	standard deviation	
5	mean	First derivative of the signal
6	maximum	
7	minimum	
8	standard deviation	
9	mean	Second derivative of the signal
10	maximum	
11	minimum	
12	standard deviation	

Continued on next page

Table A.1 – continued from previous page

Feature Index	Statistical Features	Data/signal
13	mean	Haar wavelet transform of the signal
14	maximum	
15	minimum	
16	standard deviation	
17	mean	Haar wavelet detail coefficients at level 1 (signal)
18	maximum	
19	minimum	
20	standard deviation	
21	mean	Haar wavelet detail coefficients at level 2 (signal)
22	maximum	
23	minimum	
24	standard deviation	
25	mean	Haar wavelet detail coefficients at level 3 (signal)
26	maximum	
27	minimum	
28	standard deviation	

The following flight data signals were used to derive the above mentioned features- throttle, aileron, elevator, heading, longitude, altitude, rudder, latitude, speed.

A.2 The detailed description of HR data features

Table A.2: Description of the HR data constructed features

Feature Index	Statistical Features	Data/signal
1	mean	Heart rate
2	maximum	
3	minimum	
4	standard deviation	
5	mean	First derivative of the heart rate
6	maximum	
7	minimum	
8	standard deviation	
9	mean	Second derivative of the heart rate
10	maximum	
11	minimum	
12	standard deviation	
13	mean	total low frequency power (LFP)
14	maximum	total high frequency power (HFP)
15	minimum	autonomic balance (ratio of LFP to HFP)

A.3 Results from the ACM-IMWUT paper

The results from the mental workload prediction and the corresponding plots of the skill prediction were conducted on the L1 penalized logistic regression.

Table A.3: Skill level prediction AUC using leave-one-subject-out and leave-one-route-out CV.

Algorithm	Leave a Subject Out		Leave a Route Out	
	Flight only	HR only	Flight only	HR only
Logistic regression (L1 penalty)	0.95	0.67	1	0.86
SVM (RBF kernel)	0.99	0.66	1	0.77

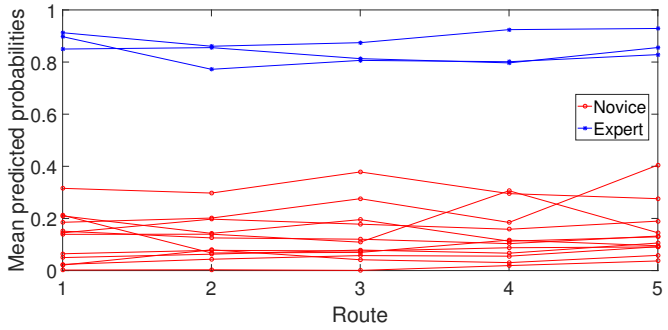

Figure A-1: Mean predicted probabilities per route using logistic regression (L1 penalty) in the leave-a-route-out validation.

Table A.4: Mental workload prediction using leave-one-subject-out and leave-one-route-out CV in the single-stage approach; RMSE (R^2).

	Leave a Subject Out			Leave a Route Out		
Algorithm	Flight only	HR only	All features	Flight only	HR only	All features
SVR	4.25 (-0.01)	4.30 (-0.04)	4.23 (-0.01)	2.85 (0.53)	3.96 (0.12)	2.85 (0.53)
LASSO	4.28 (-0.03)	4.40 (-0.09)	4.28 (-0.03)	3.94 (0.12)	4.06 (0.07)	3.87 (0.15)

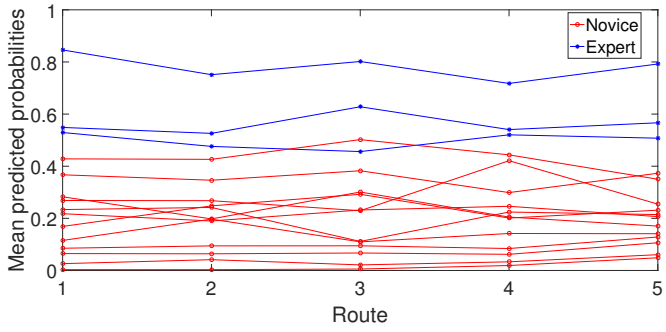

Figure A-2: Mean predicted probabilities per route using logistic regression (L1 penalty) in the leave-a-subject-out validation.

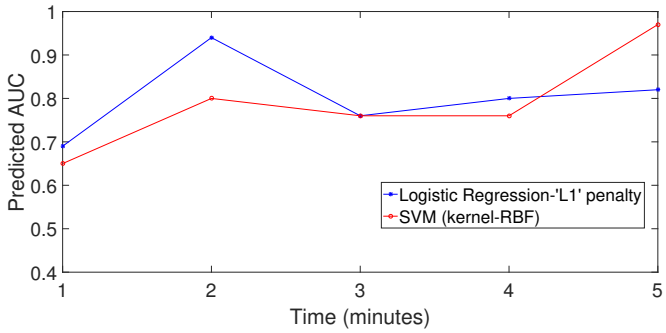

Figure A-3: Predicted AUC over individual time windows for all classifiers in the leave-a-subject-out validation.

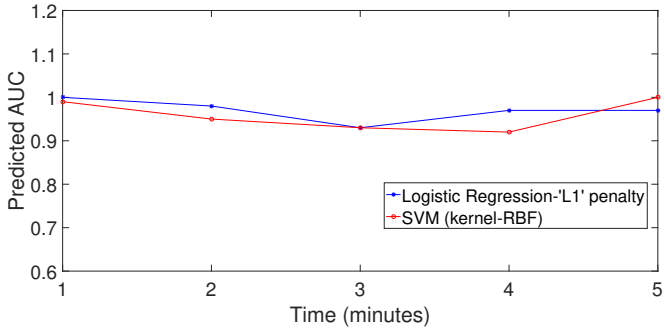

Figure A-4: Predicted AUC over individual time windows for all classifiers in the leave-a-route-out validation.

Table A.5: Mental workload prediction using leave-one-subject-out and leave-one-route-out CV in the two-stage approach; RMSE (R^2).

	Leave a Subject Out			Leave a Route Out		
Algorithm	Flight only	HR only	All features	Flight only	HR only	All features
SVR	3.99 (0.12)	4.15 (0.03)	3.48 (0.31)	2.71 (0.59)	3.53 (0.30)	2.71 (0.59)
LASSO	4.08 (0.06)	4.31 (-0.05)	3.59 (0.27)	3.38 (0.35)	3.80 (0.19)	3.35 (0.37)