DRIVER COGNITIVE WORKLOAD CLASSIFICATION USING PHYSIOLOGICAL RESPONSES

A Thesis

by

DAVID P. WOZNIAK

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Maryam Zahabi
Committee Members,	Thomas Ferris
	Wei Li
Head of Department,	Lewis Ntaimo

May 2023

Major Subject: Industrial Engineering

ABSTRACT

Motor vehicle crashes (MVCs) are a leading cause of death for law enforcement officers (LEOs) in the U.S. LEOs and more specifically novice LEOs (nLEOs) are susceptible to high cognitive workload while driving which can lead to fatal MVCs. To help address this issue, machine learning algorithms (MLAs) can be used for predicting the workload of nLEOs. These MLAs can then be implemented into adaptive in-vehicle technology that will better be able to manage the cognitive workload of LEOs in police operations. A naturalistic ride-along study was conducted with 24 novice nLEOs. Participants performed their normal patrol operations while their physiological responses such as heart rate variation (HRV) and percentage change in pupil size (PCPS) were recorded. After data collection was completed, an MLA was developed and trained based on these data using subjective responses collected from participants and pre-established thresholds for the features extracted from the physiological signals as the ground truth. It was found that the developed MLA could predict cognitive workload with relatively high accuracy given that it was entirely reliant on physiological signals. Future studies should implement the developed MLA into adaptive in-vehicle technology for the prediction of cognitive workload in real-time. Having this technology adapt to the cognitive workload of drivers should reduce cognitive workload of nLEOs and improve road safety.

DEDICATION

This work is dedicated to my mother, father, and brother for their constant support that has allowed me to pursue my educational goals and so much more. Thank you.

ACKNOWLEDGEMENTS

I would like to thank my fellow research team members Junho Park, Jordan Nunn, and Azima Maredia for their contributions to making this project possible.

CONTRIBUTORS AND FUNDING SOURCES

Contributors:

This work was supported by a thesis committee consisting of Dr. Maryam Zahabi and Dr. Thomas Ferris from the Industrial and Systems Engineering Department and Dr. Wei Li from the Landscape Architecture and Urban Planning Department.

All work for this thesis was completed by the student, under the advisement of Dr. Maryam Zahabi.

Funding Sources:

Funding for this study was provided by the National Science Foundation (NSF) (Award Number: 2041889).

TABLE OF CONTENTS

Page

ABSTRACT ii

DEDICATION iii

ACKNOWLEDGEMENTS iv

CONTRIBUTORS AND FUNDING SOURCES v

TABLE OF CONTENTS vi

LIST OF TABLES viii

LIST OF FIGURES ix

1. INTRODUCTION 1

1.1 Law Enforcement Officers 2
1.2 Cognitive Workload Classification 3
1.3 Adaptive Technology 4
1.4 Problem Statement and Research Objectives 5

2. METHOD 6

2.1 Participants 6
2.2 Equipment 6
2.3 Study Procedure 7
2.4 Synchronization Technique 12
2.5 Data Analysis 13
2.6 MLA Development 17

3. RESULTS 19

3.1 Data Screening 19
3.2 MLA Performance 21
3.3 Feature Importance 24

4. DISCUSSION 26

4.1 MLA Selection 26
4.2 Real-time Workload Classification 27
4.3 Technology Applications 28

5. CONCLUSION 31

5.1. Limitations 31
5.2. Future Work 32

REFERENCES 34

LIST OF TABLES

Page

TABLE 1: Accuracy results for most successful seeds of each MLA trained on physiological data. 22

TABLE 2: Precision and ROC AUC results for most successful seeds of each MLA trained on physiological data. 23

TABLE 3: Average training time and testing time for each MLA 24

TABLE 4: Feature Importance for Best MLA Results.....25

LIST OF FIGURES

Page

Figure 1: Naturalistic Observation Equipment Set-up 7
Figure 2: E4 Attachment Procedure 8
Figure 3: Study Set-up 9
Figure 4: Data Movement Chart 12
Figure 5: Weights for Features for Ground Truth 17
Figure 6: Data cleaning process 20
Figure 7: High Cognitive Workload AR HUD 29
Figure 8: Low Cognitive Workload AR HUD 30

1. INTRODUCTION

Motor vehicle crashes (MVCs) are one of the most prevalent causes of death in the U.S. About 46,000 people lost their lives in car crashes and roughly 5.2 million people were seriously injured due to crashes in 2022 alone (NSC, 2022). Furthermore, crash rates for younger, inexperienced drivers are significantly higher in the months immediately following getting their license compared to crash rates several years after getting a license (Curry et al., 2017). Current cognitive performance models (CPMs) used to design technology for vehicles make the critical assumption that the user performing a task is an expert and will never make mistakes or decisions that are not optimal (Kieras & Butler, 1997). This assumption has prevented models from being effectively designed around the cognitive processes of a novice or focusing on how those processes differ from expert cognition. As novice drivers deal with higher cognitive workload (the mental effort an individual exerts to complete a task) compared to experienced drivers, they are also more vulnerable to the risk of MVCs while driving (Moray, 2013). To better aid in the design of technology to reduce cognitive workload, novice performance must be more effectively accounted for.

One way that this issue can be addressed is through adaptive technology, or in-vehicle technology that will adjust its function or presentation in response to the current state of the driver. This could involve displaying less information on a user interface (UI) when the driver has a high cognitive workload or giving the driver warnings to indicate that they are operating in a high-stress condition and making recommendations accordingly. Knowing the cognitive workload of a driver based on their physiological responses is one way to allow adaptive technology to respond to the state of the driver.

1.1 Law Enforcement Officers

Law enforcement officers (LEOs) and more specifically novice LEOs (nLEOs) were selected as the focus of this study because they are the at the highest risk among all drivers to be involved in crashes (Maguire et al., 2002). MVCs are the leading cause of line-of-duty deaths for public safety workers and more specifically LEOs (BLS, 2020). Compared to firefighters and emergency medical services workers, LEOs are involved in a significantly higher number of fatal MVCs overall (BLS, 2019). Additionally, compared to all other occupations, LEOs get into MVCs at a rate of 2.5 times more than the national average (Maguire et al., 2002). Primary reasons for this include the frequent use of in-vehicle technology while driving (Yager et al., 2015), fatigue (Vila & Kenney, 2002), and lack of sufficient training in handling high-demand situations (e.g., pursuit situations, multi-tasking) (Hembroff, Arbuthnott, & Krätzig, 2018).

Use of in-vehicle technology while driving is one of the major causes of LEO crashes. This includes the technology that civilian drivers interact with frequently such as cell phones and global positioning systems (GPS) as well as LEO-specific technology such as mobile computer terminals (MCTs) (a laptop that provides real-time navigation and case information to LEOs) and dispatch radios. The myriad technologies required for LEO patrol activity highlights another issue that increases the MVC rate for novice drivers versus more experienced drivers. nLEOs are an ideal subject for designing driving MLAs due to their naturalistic patrol environment being more likely to induce high workload than a normal driving environment. In prior investigations (Park et al., 2020; Shupsky et al., 2021; Zahabi & Kaber, 2018a, 2018b; Zahabi, Pankok Jr, & Park, 2020) it was found that the mobile computer terminal (MCT) and radio are the most important and frequently used in-vehicle technologies while driving. Note that while LEOs also use this technology when not driving, these studies only considered the portion of the patrolling task when

LEOs were also actively driving their vehicles. Use of these technologies has increased LEOs' distraction and cognitive load while driving (Shahini et al., 2020). In spite of this problem of LEOs and novice drivers being at large risk of MVCs, research on the development of technology to aid them that incorporates their CW or performance has been insufficient.

1.2 Cognitive Workload Classification

To understand how the CW for novice LEOs can be modelled and calculated by a MLA, the differences between novices and experts have to be understood in a cognitive context. A literature review on the gaps in the application of novice cognition to cognitive performance models was conducted, with the results revealing some of the major differences between novice and expert cognition. These differences can be summarized using Wickens' human information processing model (Wickens, 2008). For example, with regards to attentional resources, novices are more likely to be impaired by distractions due to higher attentional resource demands, while experts are less likely to be impaired and can rely on non-visual signals more easily (Regan, Deery, & Triggs, 1998) (Mourant & Rockwell, 1970). With regards to memory, the chunking process for novices is less effective compared to experts, and novices tend to attempt to make decisions before they finish processing all the information they have (Bruer, 1993) (Hutton & Klein, 1999). Most importantly, novices have been found to exhibit higher mental workload than experts when faced with critical decisions like those needed to prevent a MVC. (Ouddiz, Paubel, & Lemercier, 2020). In contrast, experts have better recall than novices, allowing them to more effectively rely on their long-term memory and experiences to make decisions and better manage their overall CW (Horswill & McKenna, 2004). These are some examples that highlight the difference in the cognitive processes of novice and experts in the driving domain that lead to higher risk of high CW for novices.

Because of these differences, technology, CPMs, and MLAs that target novices specifically have rarely been able to capture all the idiosyncrasies between novices and experts to effectively model or predict their mental workload in various driving situations. While there have been plenty of attempts to model driver cognitive workload using MLAs in the past, these approaches relied on physiological variables that would be cumbersome to implement for naturalistic driving tasks or rely entirely on driving simulator data to develop their MLAs (Islam et al., 2020; Son, Oh, & Park, 2013). The approach taken here is novel in that it relies on physiological variables captured while the nLEO is performing their normal patrol duties. These physiological variables were collected using non-intrusive devices and included heartrate variability (HRV), percentage change in pupil size (PCPS), blink rate (BR) and galvanic skin activity (GSR). These physiological variables have all been validated as effective indicators of workload, and when combined under a single algorithm can be used to effectively determine an individual's cognitive workload at a given moment (McDonald, Ferris, & Wiener, 2019; Singh, Conjeti, & Banerjee, 2013; Zahabi et al., 2022). Because of the way this algorithm functions, it is able to circumvent many of the issues facing current MLAs with implementation into in-vehicle technology and is able to be effectively incorporated into adaptive systems that can help reduce crash rates for vulnerable driving populations.

1.3 Adaptive Technology

Adaptive technology refers to technology that will change its presentation or function in response to some change in the environment. In this case, it refers to in-vehicle technology that detects the driver's workload and responds accordingly, generally by adjusting the salience of information that is presented to the driver. As mentioned in section 1.2, the salience of information is one of the main factors contributing to how novices determine what information to focus on regardless of

importance, and in driving situations even very brief lapses in attention can lead to fatal MVCs. While performance on tasks has been shown to be most effective when an optimal level of arousal is maintained under the Yerkes-Dodson law (Yerkes & Dodson, 1908), the goal of this technology is to reduce the impact on the driving task as much as possible. Because of this, not all of the traditional rules of maximizing efficiency apply, and the changes made are more focused on maintaining optimal levels of cognitive workload for the driving task. Adaptive technology therefore has the dual goal of maximizing driver safety and making information as accessible as possible for the driver when their CW allows for it. This is why a MLA that can monitor the CW of a driver is essential to ensuring that the technology can respond quickly and effectively in the crucial seconds that a driver has to make decisions to avoid potential MVCs.

1.4 Problem Statement and Research Objectives

Novice LEOs are at a significantly higher risk of getting into MVCs compared to other occupations and the general population. Taking advantage of adaptive technology that accounts for the CW of the driver might help reduce these crash rates. This technology could also be generalizable to novice drivers that struggle with higher CW than expert drivers as well. Therefore, it is necessary to develop and integrate MLAs that can detect and predict CW for nLEOs in real time and provide that information to adaptive technology.

The objective of this study was to develop a MLA that could predict the cognitive workload of nLEOs using features that could be measured in real time while the patrol task is being performed.

2. METHOD

2.1 Participants

Twenty-four (24) participants were recruited (age: M = 30.76, SD = 5.07; gender: 6 females, 18 males). In order to qualify for this study, participants needed to have normal or corrected-to-normal vision without glasses (glasses prevented the eye tracking device from working properly), have less than 5 years of primary patrol experience (Hillerbrand, 1989), and have more than 1.5 years of driving experience. On average, participants had 3.04 years of primary patrol experience with standard deviation of 2.39. The required driving experience was included to mitigate the potential effects of having the participants be novices at both driving and primary patrol activity, as the intent of this study was to observe how secondary tasks could affect the CW of nLEOs while they are performing their duties in the vehicle. From this pool of participants, four participants were excluded from the final count due to failing to collect enough useful data (reasons such as equipment failures or participants choosing to end the study early due to needs associated with their jobs). All participants read and signed the provided informed consent form before participating in the study. As the study took place during the participants' normal working hours, they were not compensated for their time. The study protocol was approved by Texas A&M Institutional Review Board (IRB2021-0757D).

2.2 Equipment

To measure all the physiological responses necessary for this experiment, several different types of measurement equipment were employed. The Empatica E4 (Empatica) device was used to measure the HRV and GSR data from the participant. To measure the pupillometry data, the Pupil Labs Eye Tracking glasses (Pupil Labs) were used. These devices are validated for use in measuring these physiological measures and were synchronized before data were collected by

plugging the E4 watch into the laptop used to run the eye tracking software (Fuhl et al., 2016; Schuurmans et al., 2020). The ride-alongs were also recorded using a dash camera attached behind the front seats of the police vehicle. Figure 1 illustrates the details on how the naturalistic observation study set-up looked in the police vehicles. Note that the charging port used to power the dash camera came from the ports in the police vehicle itself.

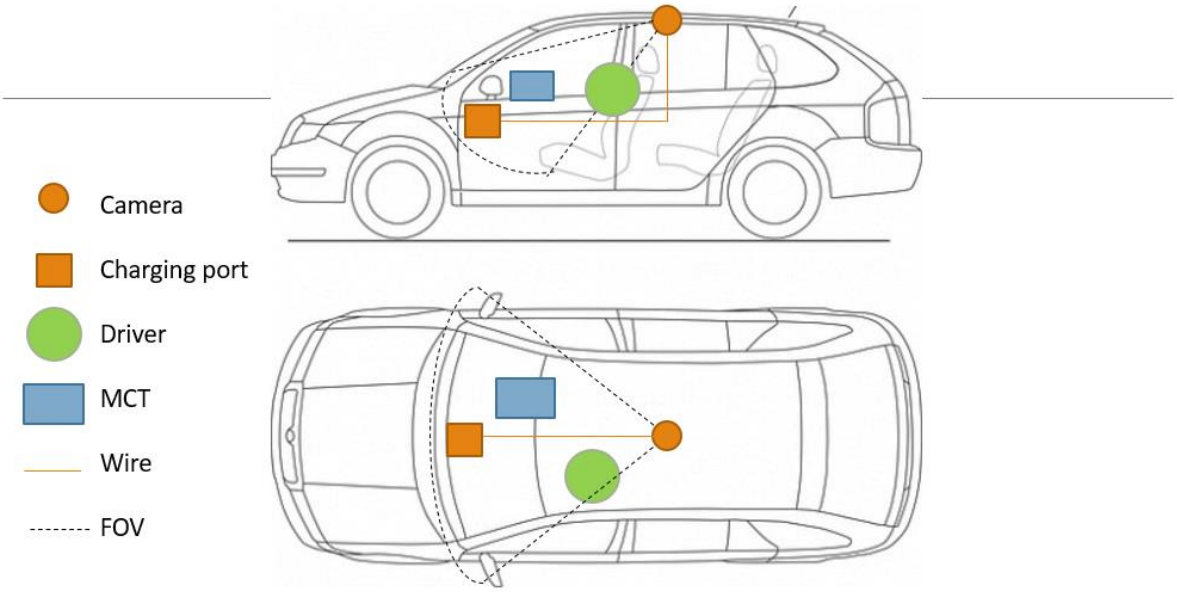

Figure 1: Naturalistic Observation Equipment Set-up

2.3 Study Procedure

Upon arriving at the police station with all the required equipment, the researcher presented the participant with an informed consent form that explained the details of the study and potential risks and benefits. Once this was signed, the Empatica E4 was attached to the wrist of the participant and activated to give the device time to calibrate while other set-up procedures were completed. Figure 2 shows how the E4 was attached to the participants' wrists.

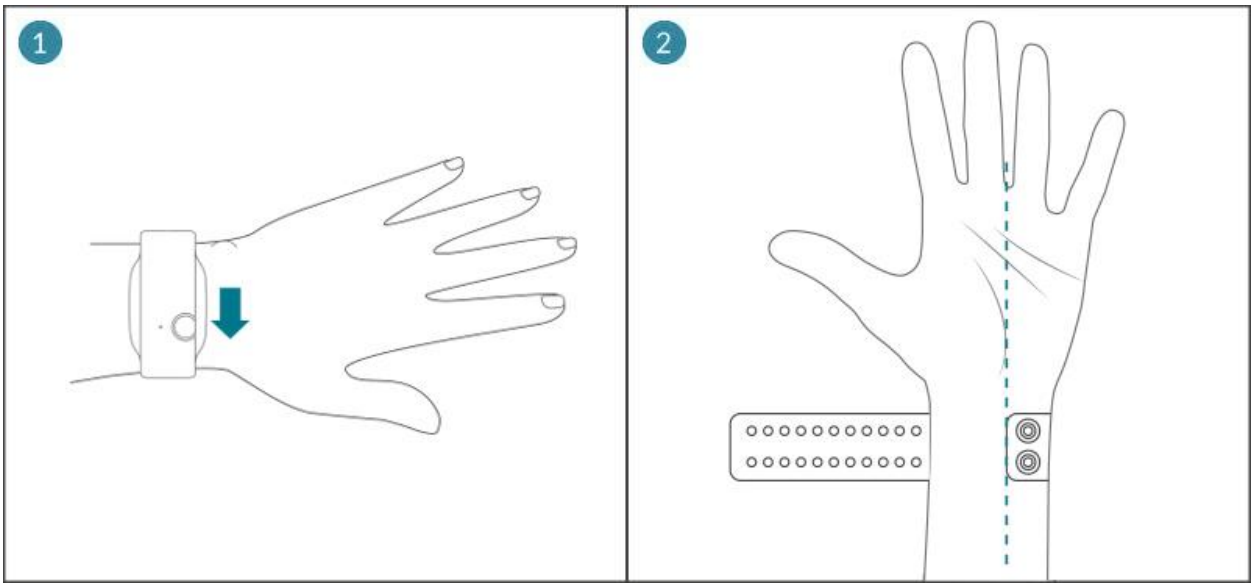

Figure 2: E4 Attachment Procedure (Empatica, 2020)

While the E4 calibrated itself, the participant was asked to fill out a demographic questionnaire. This was to ensure that the participant met the requirements for the study (see section 2.1) and to ensure that the participant did not participate in demanding physical activity while the E4 calibrated itself. If the participant did not remain in a resting state during the calibration process, then the E4 would not be able to reliably calibrate itself to the normal physiological responses of the participant. The researcher used this time to set up the eye tracking glasses and the dash camera in the participant's police vehicle as shown in Figure 1. Apriltags were attached to surfaces that participants frequently looked at (MCT, radio, windshield, etc.) to improve tracking quality. Once this was completed, the participant was asked to put on the eye tracking device and a calibration procedure was executed where the participant was asked to look at each of the four apriltags placed on the windshield without blinking as directed by the researcher. Following this step, a baseline pupil diameter was collected by running the eye tracking software for two minutes (Zahabi et al., 2021) while the participant remained in the vehicle without doing anything.

Once the calibration was completed, the study was initiated. A unique synchronization technique explained in section 2.3 was performed to ensure that data from the E4, dash camera, and the eye tracking glasses could be synchronized after the data collection. The participant was then instructed to perform their normal patrol duties while wearing the eye tracking glasses and Empatica E4. The researcher remained in the passenger seat of the vehicle to monitor the equipment and ensure that data collection proceeded smoothly. The researcher did not initiate interactions with the participant to ensure that the patrol was as naturalistic as possible. Figure 3 illustrates the set-up for the experiment, with a participant in the driver's seat on the left and a researcher on the right. Note the myriad technologies that the participant has to interact with during their patrols and the apriltags that can be seen on the MCT for tracking eye movements.

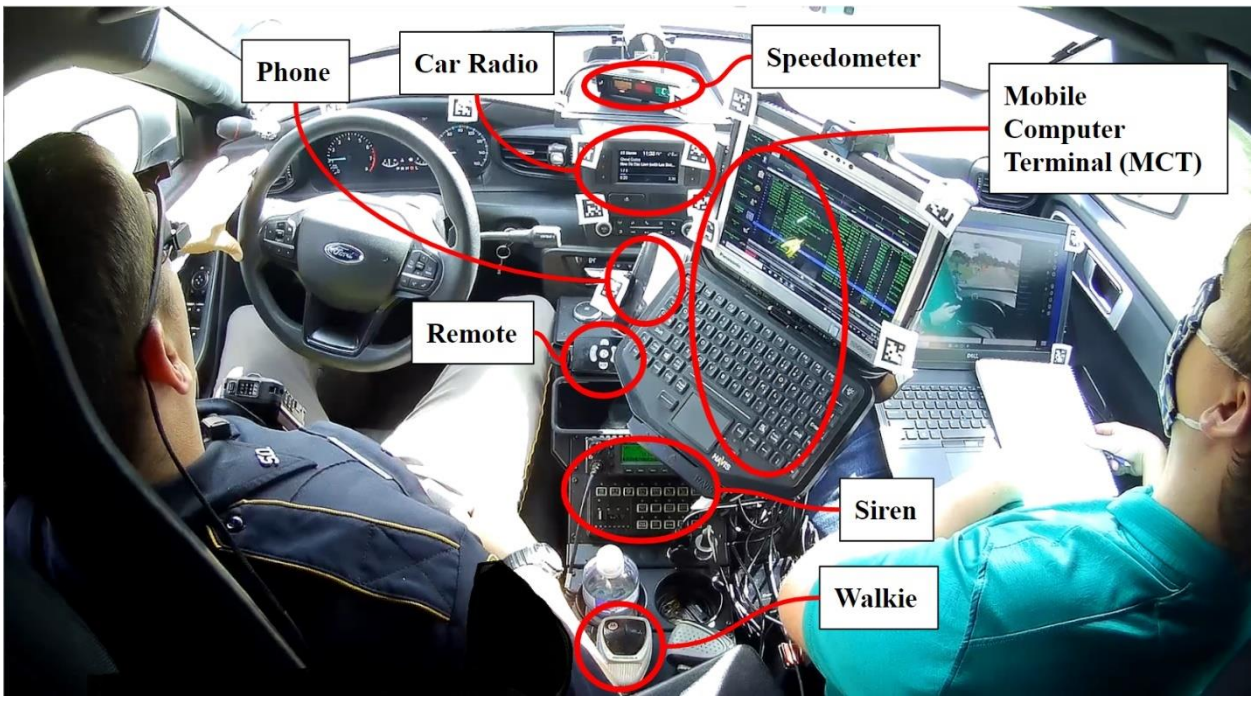

Figure 3: Study Set-up

Once the experiment began, the researcher was responsible for taking note of important events that happened during data collection. This included equipment failures, changes in the weather that could affect the quality of eye tracking data collection, and any event that could potentially affect the driving situation, such as the initiation of a police chase. In the event of equipment failure, the researcher was required to stop the experiment and address the issue with the piece of equipment that failed when the vehicle had stopped. Once the issue was resolved, the experiment could be resumed once all necessary calibration and synchronization steps were taken. Weather issues that caused significant lapses in eye tracking quality were documented and accounted for in post-processing by adjusting confidence thresholds or removing these sections of data entirely. In the event of severe weather, a timestamp could be taken on the E4 to indicate the pausing of data collection until conditions allowed for data of higher quality to be collected.

The most important role of the researcher was to verify and maintain the integrity of the data collection equipment in the event of emergency situations like police chases. While data would still be collected during emergency maneuvers when possible, the high speed and sudden changes in direction required a prioritization of the safety of the study equipment over the quality of data. The times of the start and stops of these instances were recorded and examined separately from the rest of the data to determine whether the raw data would be useful for MLA development. As this was a naturalistic observation study, the role of the researcher before the data analysis step was primarily administrative, with the goal of making the impact of the measuring equipment on the participants as low as possible. This method allowed for data collection that more accurately represented real-world changes in novice CW compared to previous studies that employed driving simulators for data collection.

Data collection continued until at least three hours of data were collected or the participant chose to stop the experiment for any reason. When participants were required to stop and exit their vehicle to do their police duties, data collection was paused, and the participant removed their eye tracking glasses (but not the Empatica E4). After the participant finished any required tasks for their job and was ready to drive again, the synchronization technique discussed below in section 2.4 was repeated and the naturalistic observation resumed.

Once the study was concluded, the participant returned to their police station and the equipment used for the observation was removed. A Driver Activity Load Index (DALI) questionnaire was given to participants to evaluate their CW during the driving part of their patrol. Once this was filled out, the participant was given a copy of the informed consent form for their records and thanked for their participation. Figure 4 outlines how data were collected and moved from the devices used for data collection and transformed into a useable format. Note the Empatica E4 and Pupil Labs eye tracking glasses both rely on the same laptop for gathering and synchronizing data streams.

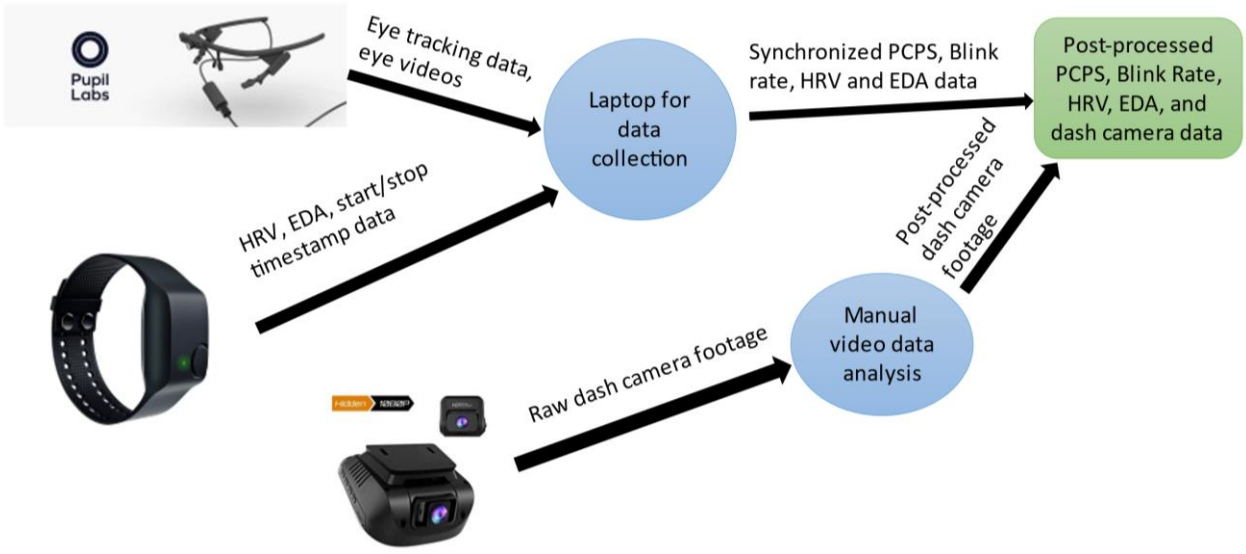

Figure 4: Data Movement Chart

2.4 Synchronization Technique

The first step in the synchronization process was ensuring that the internal clock of the Empatica E4 was synchronized with the internal clock of the computer used to run the eye tracking software. This could be accomplished by simply plugging the E4 into the computer in question to charge. Once this was done and the observation was ready to be conducted, the dash camera and eye tracking software were turned on to record and a timestamp was taken with the E4 by pressing the button on the face shown in Figure 2. Doing this caused a red LED to flash on the E4 for three seconds. This procedure was done in view of both the dash camera and the world camera of the eye tracking software. A file within the E4's data storage was used to hold the timestamp that occurred each time the button was pressed. In post-processing, the data collected by all devices before this timestamp could be discarded to ensure that all data were synchronized.

When the participant had to stop the observation to conduct police activities, a similar procedure was executed. Before turning off any devices, another timestamp was taken with the E4 to signal that no data beyond that timestamp was to be collected for the observation. Then, the eye tracking glasses could be removed, and the participant could freely leave the vehicle to conduct their business. Upon returning, another timestamp was taken with the E4 once all devices were reactivated to indicate that data from beyond that timestamp was to be used for the experiment. This procedure repeated itself until data collection was complete, upon which a final timestamp could be taken from the E4 to ensure that no further data was considered and the end time for all devices was synchronized. This synchronization procedure was accurate down to the millisecond and UNIX time was used as the metric for synchronization of all data streams. More information about the synchronization approach can be found from Wozniak et al. (2022)

2.5 Data Analysis

Once data collection was completed, the data were post-processed and the synchronized data streams were combined in order to create usable lines of data for training the MLA. The data were combined by using the synchronization method explained in section 2.4 to create a data frame for each participant. Data points that fell within periods where the police vehicle was stopped or occurred before or after the start and end timestamps respectively were removed. Rows of data were found in five-minute intervals starting from where the observation began. This interval was chosen because it is the standard interval used for collecting root mean squared standard deviation (RMSSD) data (Electrophysiology, 1996). Additionally, this interval increased the resistance of data rows to sudden spikes from some features like PCPS that may artificially indicate high workload due to minor data collection errors. In the event that a break in the data occurred before

a full five-minute interval elapsed, the partial interval was examined to determine if it contained enough data to be useful. If this was the case, it was included in the final data set.

From the raw physiological data, several features were extracted that were validated measures of cognitive workload. The baseline data collected before and after the observation period for each participant was used to calculate PCPS from the averaged pupil diameter over time for each participant. RMSSD and the low frequency/high frequency (LF/HF) ratio for each participant were calculated using HRV data. From the GSR data statistics, the skin conductance level (SCL) and skin conductance response (SCR) in the form of the SCLm, SCLc, SCRh, SCRa, and SCRr were extracted. Blink rate was calculated within each five-minute interval using the number of blinks recorded during that period by the Pupil Labs eye tracking software. The PCPS was also calculated using Pupil Labs data and the baseline pupil diameters recorded before and after the data collection period. Each of these values were then put together in a table to create a set of five-minute intervals for each participant.

Due to errors encountered in data collection as a result of naturalistic observations, not every five-minute data interval was able to be used. In the example of one participant, data collection failed completely due to the participant adjusting the Empatica E4 such that it was unable to accurately record data while the participant was away from the vehicle and out of sight of the researcher. For this and other instances where one of the four raw data streams could not be collected, the five-minute intervals associated with that set of data were discarded. In the case missing only one or two raw data streams for only some intervals within a participant, the missing values were approximated using a decision tree. This decision tree looked at all of the values for the missing data value from the other five-minute intervals and made an estimate of the missing value using

that participant's other data. This process was only used for data values within participants due to the natural differences that exist between participants with regards to average physiological values.

The final step in data analysis before MLA development could begin was to establish ground truth workload values for each of the five-minute intervals. Because it would be unreasonable to assign DALI values taken for the overall ride-along as the CW for each interval, these ratings were weighted against the physiological features themselves to establish a ground truth. The classification of CW was divided into two groups, high and low. This number of groups was chosen based on using fuzzy logic in MATLAB to separate the collected physiological data into 2 CW groups, 3 CW groups, and 5 CW groups to see which separation resulted in the most even split of data (based on pre-established thresholds for each physiological feature). It was found that 2 CW groups resulted in the most even split of the data, with 40.1% of data rows being classified as low CW. This was also considered to be reasonable due to the naturally high CW that was expected to be experienced by nLEOs during their patrol task.

Each feature, including DALI ratings, was assigned to be either a high impact, medium impact, or low impact feature for establishing ground truth CW. High impact variables included DALI, RMSSD, SCLm and SCLc due to their resistance to environmental factors, high number of validating studies, and ability to detect minute changes in workload (Cinaz et al., 2013; Fallahi et al., 2016; Mehler, Reimer, & Coughlin, 2010; Mehler, Reimer, & Wang, 2011; Pauzié, 2008a, 2008b; Reimer & Mehler, 2011; Shimomura et al., 2008; Zakerian et al., 2018). These features had a weight of 0.125 for determining the ground truth workload, with the thresholds for these physiological variables were established by previous studies (Abhishekh et al., 2013; Abusharha, 2017; Arthur, 1990; De Waard & Brookhuis, 1996; Pfleging et al., 2016; Zahabi et al., 2021). Medium impact variables include the LF/HF ratio, SCRh, SCRa, and SCRr, due to lower resilience

to environmental factors and high correlation to other physiological measures (Cinaz et al., 2013; Fallahi et al., 2016; Hsu et al., 2015; Novak, Mihelj, & Munih, 2011; Rodriguez Paras, 2015; Verwey & Veltman, 1996). These features had a weight of 0.075 each in determining ground truth workload. Finally, PCPS, BR, and avgGSR were assigned as low impact features due to the nature of data collection impeding the quality of eye tracking data and the noise factor associated with raw avgGSR values (Cardona & Quevedo, 2014; Faure, Lobjois, & Benguigui, 2016; Iqbal et al., 2005; Johns, Sibi, & Ju, 2014; Kahng & Mantik, 2002; Kosch et al., 2019; Pfleging et al., 2016; Stern, Boyer, & Schroeder, 1994). These features were assigned an importance weight of 0.0667 each. The specific weights chosen for each group were selected to keep the weight gap between feature groups relatively low while still maintaining a significant difference between the high impact and low impact features. While there is no literature exploring how these features should be prioritized over each other for machine learning or specific weighting guidelines, the weights selected here were inferred based on the above literature and thorough examination of the collected data. Figure 5 illustrates the breakdown in feature weight assignment between all of the features. Once these weights were applied to all of the five-minute intervals, ground truths could be established and an MLA could be developed.

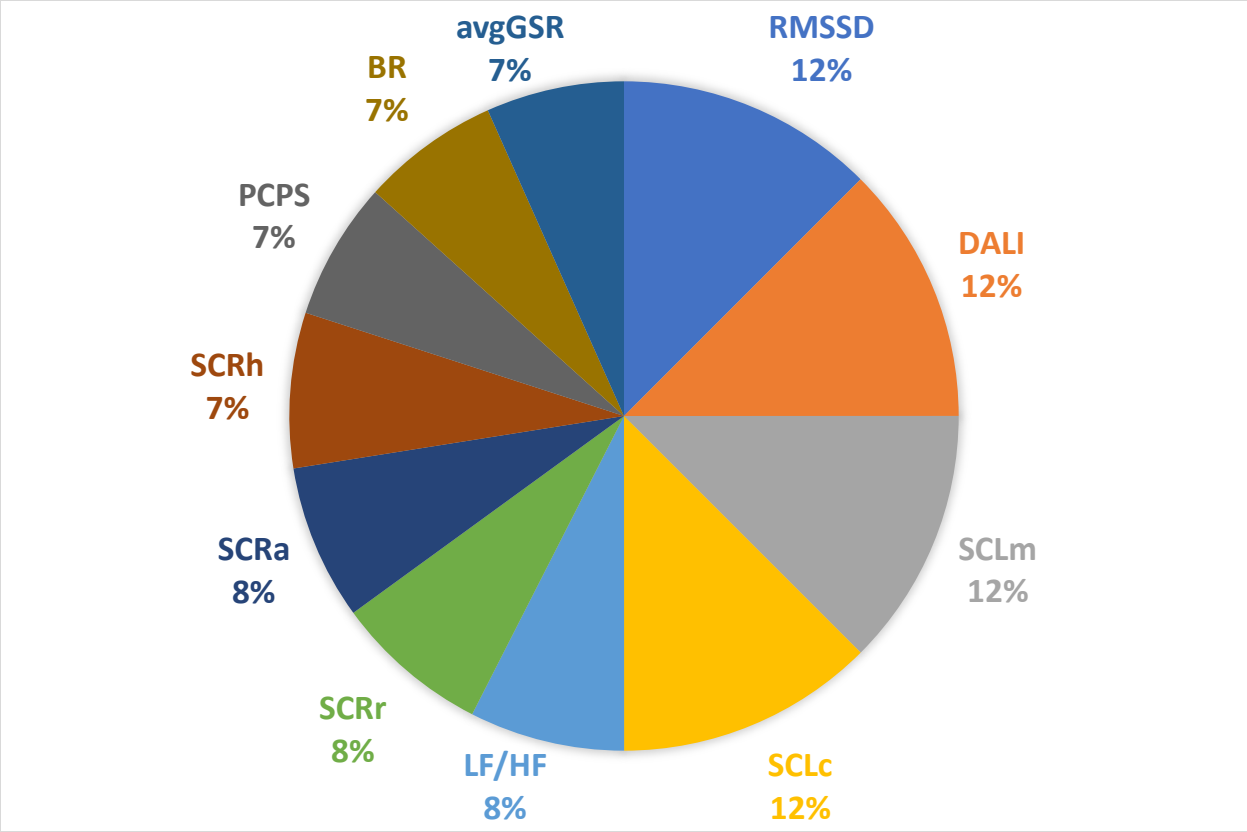

Figure 5: Weights for Features for Ground Truth

2.6 MLA Development

Development of an MLA to predict novice cognitive workload was completed first by developing MLAs that were found to be prevalent in the prediction of CW in the driving domain, which included decision trees, random forests, naïve bayes algorithms, and support vector machines (McDonald, Ferris, & Wiener, 2019). The quality of each of these MLAs was assessed across multiple seeds on the metrics of receiver operating characteristic (ROC) area under the curve (AUC), precision, and accuracy. The classification dataset used for validation and testing consisted of a randomly selected set of 20% of the overall collected data, with the remaining 80% of the data being used to train the model. All of the extracted features were considered in the final classifications made by each model, though the feature importance varied from model to model.

The algorithm that performed the best was then evaluated based on existing criteria for effective MLA performance to determine whether it was successfully able to predict the cognitive workload of nLEOs in the driving domain.

3. RESULTS

3.1 Data Screening

After grouping all of the raw data into 5-minute intervals, a total of 769 rows of data were initially created. Of those rows, only 328 had all metrics filled in with no missing values. Before missing data could be filled in, outliers in the data had to be removed. This was done by finding for each column all rows that had a value more than two standard deviations larger or smaller than the mean for that column with the value in question removed. This led to the removal of 75 rows of data, with an additional 66 rows of data being removed from participants that did not complete the experiment in a way that allowed the raw data to be salvageable. This includes issues such as but not limited to tampering with the E4 and failing to comply with instructions to ensure that data was synchronized. Once these rows were removed, the next step in the filtering process was to remove rows for which there were no values to extrapolate average values within the participant for that column. This caused the removal of one participant's data entirely due to the E4 failing to record GSR values for that participant and the subsequent removal of 22 other rows from different participants due to similar reasons. As 229 rows of data needed some form of extrapolation in order to fill in missing values, a decision tree algorithm was created to predict the missing values for each missing data entry for these rows. To prevent individual differences from confounding the predictions for these missing values, the algorithm only considered values within participants when filling in data. Figure 6 below illustrates this data cleaning process visually.

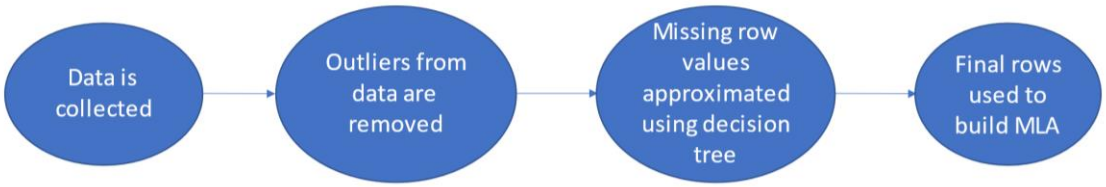

Figure 6: Data cleaning process

From the original 24 recruited participants, the data for 4 participants were excluded due to errors in data collection that occurred as a result of the study taking place in a naturalistic setting. This left 20 participants' worth of raw data to be grouped into 5-minute sections. After removing outliers and filling in missing values, a total of 557 rows of data were captured. Once ground truths had been assigned to each data row, there were 228 rows with the high workload classification and 329 rows with the low workload classification, meaning that approximately 59% of the rows were classified as being low workload. Given the nature of naturalistic observation, this can be considered a successful split of the data.

Not all participants contributed the same number of rows of data. Some participants contributed around 10 data rows while others contributed over 30. These differences were once again due to differences in both the length of data collection and the quality of the data collected due to the limitations of naturalistic observations. It is also important to note that the participant number was

used only for filling in missing data entries and establishing the ground truth for all data rows. All participant data were combined for the final MLA training to ensure that as much data could be used for training the algorithm as possible and to accommodate for a large range of potential high and low CW data rows.

3.2 MLA Performance

Based on previous studies on effective MLAs in the driving domain that relied on physiological variables, the following MLAs were selected to be trained by the collected data: decision trees (DT), random forests (RF), naïve bayes algorithms, and support vector machines (McDonald, Ferris, & Wiener, 2019). Table 1 summarizes the performance of these algorithms on the basis of their accuracy in classifying a randomly selected set of 20% of the total dataset with the rest being used to train the algorithm. For the MLAs besides naïve bayes, hyperparameters were selected and tuned using 10-fold cross validation sets from the training data repeated three times, with the best set of hyperparameters in terms of test data accuracy being selected. The support vector machine MLAs are split into radial and polynomial kernels indicated by SVMr and SVMp respectively. These two kernels were selected based on their use in a previous study evaluating MLAs in the driving domain using physiological variables only (McDonald, Ferris, & Wiener, 2019). Additionally, the no information rate (NIR) refers to the rate of success at guessing the classification of a row of data with no other information available. Tuned hyperparameters for each seed include minimum n value, cost complexity, and tree depth.

A trained MLA was considered successful if it had a higher accuracy than the NIR with at least 95% confidence. The results displayed the average of 5 seeds that performed the best from a total of 50 seed tests for each MLA in order to showcase their most effective performance. If an MLA had fewer than 5 seeds perform better than the NIR within a 95% confidence interval (CI), then

only seeds that met this condition were considered when averaging results. Overall, success rates were 42% for RF, 34% for SVMr, 8% for naïve bayes, 6% for SVMp, and 4% for DT. This does not mean that the other tested seeds were necessarily worse than the NIR (the vast majority of seeds performed better than the NIR on average), but it means they failed to meet the 95% benchmark used to define success in this context. From these results it was found that the RF model performed the best both in terms of high accuracy and consistent performance when compared to the NIR across multiple seeds.

TABLE 1

Accuracy results for most successful seeds of each MLA trained on physiological data.

Metric	Accuracy (%)	NIR (%)	95% CI (%)
RF	73.21	59.24	(64, 81.1)
SVMr	67.7	56.25	(58.22, 76.20)
SVMp	68.62	57.4	(59.2, 77.02)
DT	62.5	52.23	(52.86, 71.45)
NB	71.4	53.57	(57.81, 82.69)

* RF = Random Forest, SVMr = Support Vector Machine radial, SVMp = Support Vector Machine Polynomial, DT = Decision Tree, NB = Naïve Bayes, NIR = No information rate.

Additional metrics that were evaluated to determine the best MLA include the AUC and precision. AUC is a measure of model performance at any given threshold that evaluates the predictive ability of learning algorithms (Huang & Ling, 2005) while precision refers to the degree of difference between various samples. For both metrics, high values indicate a more effective model. It was

found that the RF model performed the best on average for AUC while NB performed the best on precision when looking at the best performing seeds overall. The results of the AUC and precision comparisons are shown in Table 2. Note that no AUC was calculated for the naïve bayes MLA because no hyperparameters were manipulated.

TABLE 2

Precision and AUC results for most successful seeds of each MLA trained on physiological data.

Metric	AUC (%)	Precision (%)
RF	79.48	73.16
SVMr	69.8	68.12
SVMp	72.4	71.99
DT	64.84	66.33
NB	N/A	76.03

* RF = Random Forest, SVMr = Support Vector Machine radial, SVMp = Support Vector Machine Polynomial, DT = Decision Tree, NB = Naïve Bayes, ROC AUC = Receiver Operator Characteristics Area Under the Curve.

Another factor to keep in mind when evaluating the MLA performance was training time and test time. These terms refer to the amount of time on average it took to train and run test data through the MLAs respectively. Test time is often used as a method of computational cost, and in the case of this MLA a small test time is a good indication that the MLA will be easier to implement in real-time. Once again, the RF MLA outperformed the other MLAs in test time with an average test time roughly 0.06 seconds faster than the second fastest MLA. On the other hand, RF performed the worst by far in training time, taking nearly five minutes longer than the next slowest MLA to finish training. Though NB had a much faster training time than the other MLAs, this performance

is not enough to overlook the poor results in the other considered metrics. Another reason that a small testing time is critical has to do with the speed at which data were collected. GSR data were collected at 4 Hz, but HRV and pupillometry data were collected much faster, at 64 Hz and around 130 Hz, respectively. To account for these frequent value updates, the developed MLA needed to have a testing time able to at least match these data intake rates combined, and only the NB and RF MLAs were able to accomplish this. Table 3 below displays the average training time and testing time for each MLA. Note that training time is in minutes and testing time is in seconds.

TABLE 3

Average training time and testing time for each MLA

Machine Learning Algorithm	Training Time (minutes)	Test Time (seconds)
RF	6.97	0.02
SVMr	1.48	0.28
SVMp	2.34	0.31
DT	2.20	0.15
NB	0.62	0.08

* RF = Random Forest, SVMr = Support Vector Machine radial, SVMp = Support Vector Machine Polynomial, DT = Decision Tree, NB = Naïve Bayes

3.3 Feature Importance

In addition to developing the MLA for future incorporation to in-vehicle technologies, the importance of each of the features for all of the different algorithms was considered. This information is summarized in Table 4 and contains the feature importance for the best results from

each of the tested MLAs. Note that while the different algorithms had their own individual scales for identifying which features were important, they have been adjusted to a weighting factor that sums to 1 for more convenient comparisons here. This means that the features with the highest value between all the others in a column is considered the most important for classification by the MLA in that column. Note that for SVMp and SVMr some feature importance ratings were negative, indicating that those features were not useful in predicting CW, so the scaling in the table below is adjusted accordingly. For every MLA except for NB, SCRr was found to be the most important feature, with features such as avgEDA and SCLc being considered less important.

TABLE 4

Feature Importance for Best MLA Results

Model	**RF**	**DT**	**SVMp**	**SVMr**	**NB**
SCRr	0.21	0.41	3.90	3.90	0.26
LF/HF	0.11	0.16	0.54	0.54	0.0052
SCLm	0.10	0.15	0.11	0.11	8.86E-06
Blink Rate	0.092	0.080	-0.49	-0.49	0.41
SCRa	0.073	0.044	-3.63	-3.63	0.0045
SCRh	0.060	0.042	-0.90	-0.90	0.18
RMSSD	0.067	0.039	0.63	0.63	0.0041
SCLc	0.12	0.030	0.053	0.053	0.00097
avgGSR	0.067	0.025	0.59	0.59	0.10
PCPS	0.11	0.024	0.20	0.20	0.022

4. DISCUSSION

4.1 MLA Selection

Out of all of the MLAs tested, only the RF MLA consistently performed better than the NIR rate in terms of accuracy while meeting the precision and AUC guidelines found for creating effective MLAs, which include precision ratings on average of at least 0.7 as well as AUC values of around 0.85 (Lee, Lessler, & Stuart, 2010; Pencina et al., 2008). Specific values for good accuracy are not standardized, so 0.7 was used as a general benchmark in line with the precision recommendation. These values are general guidelines, as the effectiveness of a MLA is primarily determined by its ability to learn as it obtains more data, meaning that future data collection should be able to improve this MLA to validate its effectiveness in predicting cognitive workload (El Naqa & Murphy, 2015). Of the MLAs, the DT algorithm performed the worst, failing to perform significantly better than the NIR rate roughly 96% of the time. Naïve bayes and SVMp algorithms performed poorly as well. As naïve bayes assumes full independence of features and it was reasonable to assume that at least some of the features in this dataset were related to each other, this result was expected for naïve bayes (Lewis, 1998). However, it was anticipated that DT and SVMp would perform much better than they actually did. One possible reason for this discrepancy would be the use of data collected in a naturalistic setting rather than in a laboratory or in a simulator study. This could also be due to the stringent requirements placed on tested MLAs in order to be considered successful in terms of accuracy with regard to the NIR. The incorporation of other variables into future iterations of both these MLAs and the most successful MLA, RF, in future studies could yield more satisfactory results.

To validate the selection of the RF algorithm, several factors were considered, most notably the accuracy, precision, and AUC on average of all the seeds tested. As the RF algorithm performed

better than its other competitors, it was the final MLA selected. Note that while specific seeds might have had other MLAs perform better than the RF, the RF algorithm provided the most consistent results. Another point of note is that RF algorithms are popular in the driving domain and have been used in several studies in the past, making incorporation to other technologies and comparisons easier to make (Das & Khilar, 2019; Ferreira et al., 2017; Rahman, Saleem, & Iyer, 2019). These studies investigated several different avenues of application such as using phones or in-vehicle GPS for the implementation of adaptive technology, and the advantages of a MLA like this that can collect and use data for training and prediction while driving would be invaluable to the development of these technologies.

4.2 Real-time Workload Classification

The primary reason that physiological variables alone were considered in the naturalistic observation ride-along study design was because one of the end goals for the developed MLA was to be able to classify workload in real-time. The applications of this classification would be to implement them into technology that can use the current workload of the driver to adjust the in-vehicle technology to accommodate them. Given that novice drivers are more prone to high workload and this higher workload can lead to more potentially fatal mistakes, understanding when these risks might happen using MLAs is critical. To offset the potential issues in accuracy of the MLA, new samples need to consistently be taken to have the workload update as frequently as possible. Individual differences also need to be accounted for by having the MLA be trained specifically with data collected for an individual participant and supplemented by the already collected data. To test this, a real-time algorithm was developed in python and tested in a lab setting to see if data could be recorded and run through the developed MLA in real-time. This was proven to be the case and the developed MLA was able to predict cognitive load in real-time with

data collected without having to stop data collection. This finding is crucial when considered in tandem with the test times for the developed MLAs.

In addition to being the most accurate on average, the RF algorithm was by far the fastest in calculating the output when fed the same amount of test data as the other MLAs. Though the difference of a few hundredths of a second might seem insignificant, the longer it takes a MLA to output results in real-time, the higher the risk that the output it provides will be too late to be useful. If output calculation is unable to keep up with the rate at while physiological signals are processed, then this delay can continue to build until the output is many seconds, possibly even minutes behind the original reading in the worst case scenario. To prevent this, a MLA that can calculate output as quickly as possible is essential, giving RF more credence as an optimal choice among the tested MLAs. Note that for the purposes of this study, the final products included the developed MLA and the identification of the common scenarios experienced by nLEOs while on patrol, but future studies should be able to use the proposed model to create new technology that can improve the safety of nLEOs.

4.3 Technology Applications

Adaptive technology has the most potential to take full advantage of the real-time workload classification MLA developed from this study. An example of this can be found in the information that police officers receive about a suspect on their MCT screens. During ride-alongs, the LEOs interacted with their in-vehicle technologies in four complex scenarios. These scenarios refer to secondary tasks that the LEO completes while driving that are likely to induce high CW. The scenarios included talking on the phone, searching for or inputting directions onto a GPS device, communicating on a walkie-talkie, and searching for information on the MCT that requires physical contact with the MCT. For the last task in particular, LEOs are frequently confronted with

overwhelming amounts of text-based information that obscure important information and make it difficult to determine the nature of the situation they are driving to while focusing on the road. At the same time, this information might be important and more experienced officers would have fewer problems processing this information. To address this issue, adaptive technology based on workload can be implemented. One example of what this could look like is illustrated in Figures 7 and 8. These images show a prototype using an augmented reality heads-up display (AR HUD) that would automatically adjust its appearance based on the CW of LEOs. If the LEO is experiencing high workload, then a simpler version of the information they need for their current patrol task will be displayed similar to Figure 7. Note that the red pictures indicate the violations.

Figure 7: High Cognitive Workload AR HUD

Conversely, when the LEO's CW is low, more information can be made available to them without increasing the risk of a MVC. In this case, a display similar to Figure 8 could be shown. This is just a single example of how adaptive technology could effectively be integrated with real-time CW classification from the developed MLA to improve LEO safety and reduce MVCs. Future studies should validate this proposed MLA with additional user-testing and then implement it in real-world scenarios to evaluate the effectiveness of this technology in reducing cognitive workload. To that end, the feature importance findings of this study can be useful. Given the difficulty of implementing a multitude of physiological measures into in-vehicle technology,

prioritizing GSR in accordance with the findings on feature importance may allow for more effective implementation than trying to accomplish everything at once.

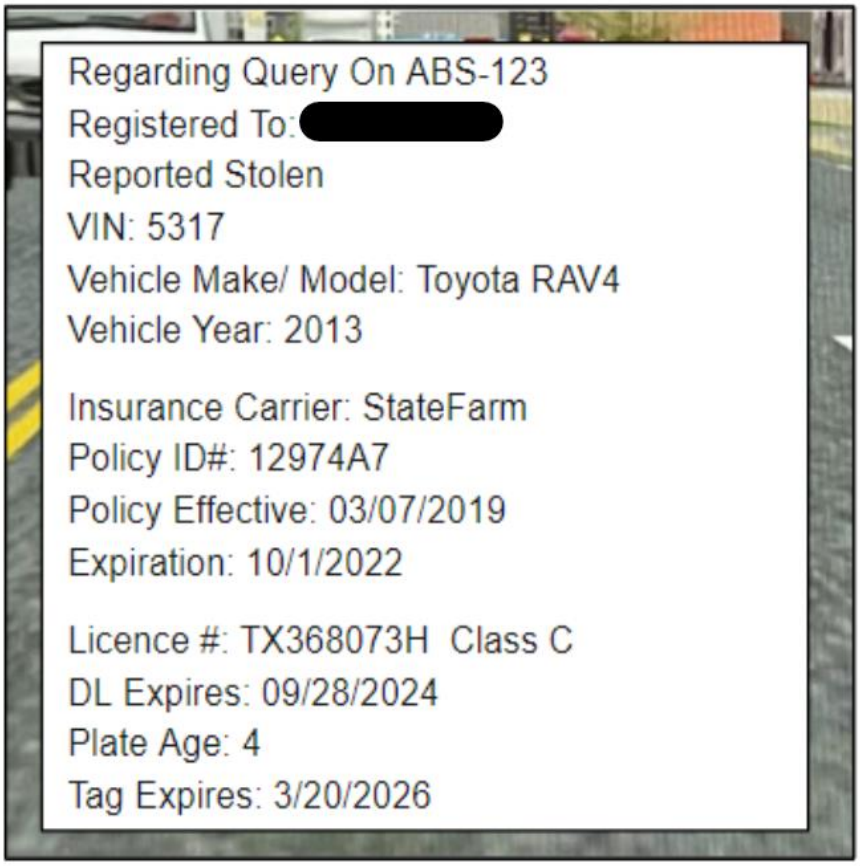

Figure 8: Low Cognitive Workload AR HUD

5. CONCLUSION

In this ride-along study, an MLA for classifying the workload of nLEOs was developed. This MLA used physiological variables in the form of HRV, GSR, BR, PCPS, and extracted features from these responses to classify cognitive workload. One of the key contributions of this MLA over other MLAs in the domain is its ability to be implemented without impeding the normal patrol duties of a LEO. Because of this, the results can be used to develop technology that can predict the workload of users in real-time and adapt the functions of a vehicle accordingly, either to emphasize or de-emphasize secondary tasks. Common scenarios that nLEOs contend with during their normal patrol duties were also identified and can serve as a starting point for the design of more realistic simulator studies. These scenarios included complex scenarios shown to induce high CW such as inputting directions on a GPS device and searching for information on a MCT that requires physical contact with the MCT. Future studies should seek to implement this MLA into adaptive technology in both real world and simulated settings to evaluate how effective it is at predicting CW for both low-risk and high-risk scenarios respectively. Through this, the MLA can be refined with the incorporation of performance variables gathered from existing in-vehicle technology to improve MLA accuracy and precision, providing faster, more accurate readings. Incorporating the developed MLA with adaptive technology can help nLEOs to better manage their tasks in the vehicle and can improve their safety in police operations.

5.1. Limitations

There were several limitations of this study that should be addressed when conducting further experiments to develop MLAs for this domain. First, the nature of naturalistic observation caused significant amounts of data to be lost over the course of the experiment. Mitigation techniques that were found to be effective to maximize the amount of data collected included only performing

data collection in the daytime, targeting cloudy days to avoid the amount of eye data lost to sunlight glare, and ensuring that all devices are properly secured to avoid data loss that goes undetected until the experiment is over. Because subjective workload evaluations could only be collected at the end of the observation period, they might not have accurately represented the entire ride-along with regards to the cognitive workload. The development of a method for establishing ground truth workload was implemented to help circumvent this limitation. Using physiological features only for classification allowed for real-time algorithms to potentially make use of this MLA, but it also does not account for subjective measures and task performance that were also shown to be effective predictors of cognitive workload. This sort of limitation could be overcome by implementing technology that is able to track user performance and already exists in the form of technology like lane-keep assist and cruise control.

5.2. Future Work

Future experiments need to continue testing the developed MLA with different scenarios to see how effective it is at predicting CW in scenarios other than the ones identified in this study. This could be done with driving simulator studies or further naturalistic observation studies. An advantage of non-naturalistic observation studies is the potential to incorporate other measures of CW such as task performance thanks to the experiment being performed in a controlled environment. One goal of the developed MLA is to eventually incorporate the performance of the LEO on the task they are currently performing to more accurately evaluate their CW. The results of these future experiments should be a more robust and adaptive MLA that can take advantage of the driving performance of LEOs to adjust how adaptive in-vehicle technology interacts with the driver. Ideally, wearable devices that are less intrusive without sacrificing effectiveness should be employed. While the Pupil Labs eye tracking device did not impede the patrol task of LEOs

significantly, implementing a wireless version of the glasses or one that functions as sunglasses that most officers wear while on duty would improve the quality of implemented MLAs while reducing any induced cognitive load by the system on LEOs. Development and implementation of this technology would greatly improve the quality of data collection and real-time MLA implementation overall.

REFERENCES

Abhishekh, H. A., Nisarga, P., Kisan, R., Meghana, A., Chandran, S., Raju, T., & Sathyaprabha, T. N. (2013). Influence of age and gender on autonomic regulation of heart. *Journal of clinical monitoring and computing, 27*(3), 259-264.

Abusharha, A. A. (2017). Changes in blink rate and ocular symptoms during different reading tasks. *Clinical optometry, 9*, 133.

Arthur, E. (1990). Physiological Metrics of Mental Workload: A Review of Recent Progress.

BLS. (2019). Table A-6. Fatal occupational injuries resulting from transportation incidents and homicides by occupation, all United States, 2015. 2016. Retrieved from https://www.bls.gov/iif/oshwc/cfoi/cftb0300.xlsx

BLS. (2020). Fatal Occupational Injuries to Emergency Responders. Retrieved from https://www.bls.gov/iif/oshwc/cfoi/er_fact_sheet.htm

Bruer, J. T. (1993). The mind's journey from novice to expert. *American Educator, 17*(2), 6-15.

Cardona, G., & Quevedo, N. (2014). Blinking and driving: the influence of saccades and cognitive workload. *Current eye research, 39*(3), 239-244.

Cinaz, B., Arnrich, B., La Marca, R., & Tröster, G. (2013). Monitoring of mental workload levels during an everyday life office-work scenario. *Personal and ubiquitous computing, 17*(2), 229-239.

Curry, A. E., Metzger, K. B., Williams, A. F., & Tefft, B. C. (2017). Comparison of older and younger novice driver crash rates: Informing the need for extended Graduated Driver Licensing restrictions. *Accident Analysis & Prevention, 108*, 66-73.

Das, R., & Khilar, P. M. (2019). *Driver behaviour profiling in VANETs: comparison of ensemble machine learning techniques.* Paper presented at the 2019 IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP).

De Waard, D., & Brookhuis, K. (1996). The measurement of drivers' mental workload.

El Naqa, I., & Murphy, M. J. (2015). What is machine learning? In *machine learning in radiation oncology* (pp. 3-11): Springer.

Electrophysiology, T. F. o. t. E. S. o. C. t. N. A. S. o. P. (1996). Heart rate variability: standards of measurement, physiological interpretation, and clinical use. *Circulation, 93*(5), 1043-1065.

Empatica. (2020). Wear your E4 wristband. Retrieved from https://support.empatica.com/hc/en-us/articles/206374015-Wear-your-E4-wristband-

Fallahi, M., Motamedzade, M., Heidarimoghadam, R., Soltanian, A. R., & Miyake, S. (2016). Effects of mental workload on physiological and subjective responses during traffic density monitoring: A field study. *Applied ergonomics, 52*, 95-103.

Faure, V., Lobjois, R., & Benguigui, N. (2016). The effects of driving environment complexity and dual tasking on drivers' mental workload and eye blink behavior. *Transportation research part F: traffic psychology and behaviour, 40*, 78-90.

Ferreira, J., Carvalho, E., Ferreira, B. V., de Souza, C., Suhara, Y., Pentland, A., & Pessin, G. (2017). Driver behavior profiling: An investigation with different smartphone sensors and machine learning. *PLoS one, 12*(4), e0174959.

Fuhl, W., Tonsen, M., Bulling, A., & Kasneci, E. (2016). Pupil detection for head-mounted eye tracking in the wild: an evaluation of the state of the art. *Machine Vision and Applications, 27*(8), 1275-1288.

Hembroff, C. C., Arbuthnott, K. D., & Krätzig, G. P. (2018). Emergency response driver training: Dual-task decrements of dispatch communication. *Transportation research part F: traffic psychology and behaviour, 59*, 222-235.

Hillerbrand, E. (1989). Cognitive differences between experts and novices: Implications for group supervision. *Journal of Counseling & Development, 67*(5), 293-296.

Horswill, M. S., & McKenna, F. P. (2004). Drivers' hazard perception ability: Situation awareness on the road. *A cognitive approach to situation awareness: Theory and application*, 155-175.

Hsu, B.-W., Wang, M.-J. J., Chen, C.-Y., & Chen, F. (2015). Effective indices for monitoring mental workload while performing multiple tasks. *Perceptual and motor skills, 121*(1), 94-117.

Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering, 17*(3), 299-310.

Hutton, R. J., & Klein, G. (1999). Expert decision making. *Systems Engineering: The Journal of The International Council on Systems Engineering, 2*(1), 32-45.

Iqbal, S. T., Adamczyk, P. D., Zheng, X. S., & Bailey, B. P. (2005). *Towards an index of opportunity: understanding changes in mental workload during task execution.* Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems.

Islam, M. R., Barua, S., Ahmed, M. U., Begum, S., Aricò, P., Borghini, G., & Di Flumeri, G. (2020). A novel mutual information based feature set for drivers' mental workload evaluation using machine learning. *Brain Sciences, 10*(8), 551.

Johns, M., Sibi, S., & Ju, W. (2014). *Effect of cognitive load in autonomous vehicles on driver performance during transfer of control.* Paper presented at the Adjunct Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications.

Kahng, A. B., & Mantik, S. (2002). *Measurement of inherent noise in EDA tools.* Paper presented at the Proceedings International Symposium on Quality Electronic Design.

Kieras, D. E., & Butler, K. A. (1997). Task Analysis and the Design of Functionality. *The computer science and engineering handbook, 23*, 1401-1423.

Kosch, T., Karolus, J., Ha, H., & Schmidt, A. (2019). *Your skin resists: exploring electrodermal activity as workload indicator during manual assembly.* Paper presented at the Proceedings of the ACM SIGCHI Symposium on Engineering Interactive Computing Systems.

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in medicine, 29*(3), 337-346.

Lewis, D. D. (1998). *Naive (Bayes) at forty: The independence assumption in information retrieval.* Paper presented at the European conference on machine learning.

Maguire, B. J., Hunting, K. L., Smith, G. S., & Levick, N. R. (2002). Occupational fatalities in emergency medical services: a hidden crisis. *Annals of emergency medicine, 40*(6), 625-632.

McDonald, A. D., Ferris, T. K., & Wiener, T. A. (2019). Classification of Driver Distraction: A Comprehensive Analysis of Feature Generation, Machine Learning, and Input Measures. *Human Factors, 0*(0), 0018720819856454. doi:10.1177/0018720819856454

Mehler, B., Reimer, B., & Coughlin, J. F. (2010). *Physiological reactivity to graded levels of cognitive workload across three age groups: An on-road evaluation.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Mehler, B., Reimer, B., & Wang, Y. (2011). *A comparison of heart rate and heart rate variability indices in distinguishing single-task driving and driving under secondary cognitive workload.* Paper presented at the Driving Assesment Conference.

Moray, N. (2013). *Mental workload: Its theory and measurement* (Vol. 8): Springer Science & Business Media.

Mourant, R. R., & Rockwell, T. H. (1970). Mapping eye-movement patterns to the visual scene in driving: An exploratory study. *Human factors, 12*(1), 81-87.

Novak, D., Mihelj, M., & Munih, M. (2011). Psychophysiological responses to different levels of cognitive and physical workload in haptic interaction. *Robotica, 29*(3), 367-374.

NSC. (2022). Preliminary Semiannual Estimates. Retrieved from https://injuryfacts.nsc.org/motor-vehicle/overview/preliminary-estimates/

Ouddiz, S., Paubel, P.-V., & Lemercier, C. (2020). How do novice and expert drivers prepare for takeover when they are drivengers of a level 3 autonomous vehicle? Investigation of their visual behaviour. *Le travail humain, 83*(4), 361-378.

Park, J., McKenzie, J., Shahini, F., & Zahabi, M. (2020). *Application of Cognitive Performance Modeling for Usability Evaluation of Emergency Medical Services In-Vehicle Technology.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Pauzié, A. (2008a). *Evaluating driver mental workload using the driving activity load index (DALI).* Paper presented at the proc. of European conference on human interface design for intelligent transport systems.

Pauzié, A. (2008b). A method to assess the driver mental workload: The driving activity load index (DALI). *IET Intelligent Transport Systems, 2*(4), 315-322.

Pencina, M. J., D'Agostino Sr, R. B., D'Agostino Jr, R. B., & Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in medicine, 27*(2), 157-172.

Pfleging, B., Fekety, D. K., Schmidt, A., & Kun, A. L. (2016). *A model relating pupil diameter to mental workload and lighting conditions.* Paper presented at the Proceedings of the 2016 CHI conference on human factors in computing systems.

Rahman, A. A., Saleem, W., & Iyer, V. V. (2019). *Driving behavior profiling and prediction in KSA using smart phone sensors and MLAs.* Paper presented at the 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT).

Regan, M. A., Deery, H. A., & Triggs, T. J. (1998). *Training for attentional control in novice car drivers: A simulator study*. Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Reimer, B., & Mehler, B. (2011). The impact of cognitive workload on physiological arousal in young adult drivers: a field study and simulation validation. *Ergonomics, 54*(10), 932-942.

Rodriguez Paras, C. (2015). *Exploring physiological measures for prediction and identification of the redline of cognitive workload.*

Schuurmans, A. A., de Looff, P., Nijhof, K. S., Rosada, C., Scholte, R. H., Popma, A., & Otten, R. (2020). Validity of the Empatica E4 wristband to measure heart rate variability (HRV) parameters: A comparison to electrocardiography (ECG). *Journal of medical systems, 44*(11), 1-11.

Shahini, F., Zahabi, M., Patranella, B., & Mohammed Abdul Razak, A. (2020). *Police Officer Interactions with In-vehicle Technologies: An On-Road Investigation.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Shimomura, Y., Yoda, T., Sugiura, K., Horiguchi, A., Iwanaga, K., & Katsuura, T. (2008). Use of frequency domain analysis of skin conductance for evaluation of mental workload. *Journal of physiological anthropology, 27*(4), 173-177.

Shupsky, T., Lyman, A., He, J., & Zahabi, M. (2021). Effects of mobile computer terminal configuration and level of driving control on police officers' performance and workload. *Human factors, 63*(6), 1106-1120.

Singh, R. R., Conjeti, S., & Banerjee, R. (2013). A comparative evaluation of neural network classifiers for stress level analysis of automotive drivers using physiological signals. *Biomedical Signal Processing and Control, 8*(6), 740-754.

Son, J., Oh, H., & Park, M. (2013). Identification of driver cognitive workload using support vector machines with driving performance, physiology and eye movement in a driving simulator. *International Journal of Precision Engineering and Manufacturing, 14*(8), 1321-1327.

Stern, J. A., Boyer, D., & Schroeder, D. (1994). Blink rate: a possible measure of fatigue. *Human factors, 36*(2), 285-297.

Verwey, W. B., & Veltman, H. A. (1996). Detecting short periods of elevated workload: A comparison of nine workload assessment techniques. *Journal of Experimental Psychology: Applied, 2*(3), 270.

Vila, B., & Kenney, D. J. (2002). The prevalence and potential consequences of police fatigue. *NIJ Journal, 248*, 17-21.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human factors, 50*(3), 449-455.

Wozniak, D., Park, J., Nunn, J., Maredia, A., & Zahabi, M. (2022). *Measuring Cognitive Workload of Novice Law Enforcement Officers in a Naturalistic Driving Study*. Paper presented at the HFES 66th International Annual Meeting, Atlanta, GA.

Yager, C., Dinakar, S., Sanagaram, M., & Ferris, T. K. (2015). Emergency vehicle operator on-board device distractions. *Texas A&M Transportation Institute Technical Report, 2015*, 1-50.

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Punishment: Issues and experiments*, 27-41.

Zahabi, M., & Kaber, D. (2018a). Effect of police mobile computer terminal interface design on officer driving distraction. *Applied ergonomics, 67*, 26-38.

Zahabi, M., & Kaber, D. (2018b). Identification of task demands and usability issues in police use of mobile computing terminals. *Applied ergonomics, 66*, 161-171. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0003687017301953?via%3Dihub

Zahabi, M., Nasr, V., Mohammed Abdul Razak, A., Patranella, B., McCanless, L., & Maredia, A. (2021). Effect of secondary tasks on police officer cognitive workload and performance under normal and pursuit driving situations. *Human factors*, 00187208211010956.

Zahabi, M., Pankok Jr, C., & Park, J. (2020). Human factors in police mobile computer terminals: A systematic review and survey of recent literature, guideline formulation, and future research directions. *Applied ergonomics, 84*, 103041.

Zahabi, M., Shahini, F., Yin, W., & Zhang, X. (2022). Physical and cognitive demands associated with police in-vehicle technology use: an on-road case study. *Ergonomics, 65*(1), 91-104.

Zakerian, S. A., Zia, G., Nasl Seraji, G., Azam, K., & Mortezapour, A. (2018). Reliability and validity of the driver activity load index for assessing mental workload among drivers in production companies. *Journal of Occupational Hygiene Engineering Volume, 5*(2), 65-71.