Predicting Mental Workload during Semi-Autonomous Driving using Physiological and Driving Performance Measures

Justin S. Lee[2], Nade Liang[1], Gaojian Huang[1], and Brandon Pitts[1]
[1]School of Industrial Engineering, Purdue University
[2]School of Electrical and Computer Engineering, Purdue University

ABSTRACT

Vehicle automation is developing at a rapid rate worldwide. Some SAE Level-1 automated driving systems (ADS), or semi-autonomous driving systems, are becoming widely available. Examples include Adaptive Cruise Control (ACC) and Lane Keeping System (LKS). There are new challenges to road safety when using these types of ADS as continuous driver input is still needed. Some of these challenges are associated with suboptimal driver mental workload (MWL). Predicting a driver's MWL can help mitigate these challenges and improve road safety. The goal of this study is to investigate the capabilities of a selection of supervised Machine Learning (ML) classification techniques to produce driver behavior-driven prediction models of driver's MWL. Subjective measurements of MWL using the NASA Task Load Index were used as comparison to determine the effectiveness of the ML classification techniques. Data used in this study was collected on a medium fidelity driving simulator. Driver behavior was monitored using a combination of physiological and driving performance measures. Findings show that the random forest classification models performed the best in predicting MWL compared to other models such as decision tree or support vector machine. Heart rate and heart rate variability data, followed by eye tracking data, were the most significant features for the classification models. The result of this study provides key insights into the importance of measuring heart rate and eye tracking data when designing future systems that predict MWL of drivers in real-time.

KEYWORDS

Human Factors, Mental Workload, Semi-Autonomous Driving, Eye Tracking, Machine Learning, Random Forest

1 Introduction

There are six different levels of automation, defined by the SAE's definition of levels of automation (SAE International, 2021). At SAE Level 0, there is no automation and is referred to as manual driving. When analyzing driving performance at SAE Level 0, there are usually two modes of control that are analyzed, longitudinal and lateral control. As the level of automation increases by one, the modes of control available to the driver decreases by one. At SAE Level 1, one dimension of control is automated. However, continuous driver input is still needed which can lead to road safety concerns. At SAE Level 2, both longitudinal and lateral control are automated. At this level, the main concern for a driver's performance is their takeover performance. A takeover event is where manual intervention is required by the driver due to limitations in the automation system or failure. At SAE Level 3, the driver can do secondary tasks during automation, but a takeover event may still occur unexpectedly. Thus, driver input is still required at these levels. SAE Level 4 does not require any takeovers within the range of capabilities of the automation system. However, once the vehicle exits the range of capabilities, the driver must resume control. SAE Level 5 is full automation with no driver intervention required.

A driver's mental workload has been found to be a function of the situational complexity, and it negatively impacts driving performance (Paxion et al., 2014). Tran et al. (2017) states that both a high and too low mental workload degrades driving performance. Du et al. (2020) mentions that driver's cognitive and emotional states influence driver's takeover performance. Degradation in driver performance can lead to road safety concerns. Determining a driver's mental workload may be useful when designing alert systems for takeover requests, designing automation systems to account for driver behavior during silent failures, and real-time prediction of a driver's performance. This will help reduce traffic accidents and errors while drivers are using semi-autonomous and autonomous vehicles that require takeovers.

Mental workload is commonly measured subjectively using NASA Task Load Index (NASA-TLX) scales (Hart & Staveland, 1988) when measuring mental workload during different driving scenarios. NASA TLX scales are subjective measurements of a person's mental state where subjects are asked to rate various metrics on a scale including mental demand, physical demand, temporal demand, performance, effort, and frustration level. These studies often involve different levels of automation, from SAE Level 0, 1, and 2. For example, Chen et al. (2019) measured drivers' mental workload at SAE Level 0, 1, and 2 using a driving simulator. In other studies, at higher levels of automation, secondary tasks that are unrelated to driving (non-driving-related tasks, NDRTs) and its impacts are explored including its impacts on driving performance.

It has been shown by past research (Foy & Chapman, 2018) that mental workload is reflected in various driver physiological measurements (such as eye movements) and driver behavior. Specifically, galvanic skin response has been found to be an effective indicator for mental workload level (Foy & Chapman, 2018) as well as gaze behavior (Du et al., 2020) for predicting mental workload in driving. HRV was found to be a good indicator for mental workload (Mulder, 1992).

Predicting a driver's mental workload based on physiological factors and driving performance metrics at different level of automation has not been thoroughly explored in the literature. There already exists research that discusses which physiological factors are more important for predicting driver performance and mental workload (Du et al., 2020; Marquart et al., 2015). However, research into developing a supervised machine learning model that predicts drivers' mental workload during automated driving is limited.

Several studies exist that develop models to predict mental workload or driving performance using classification models. One study used classifiers to predict differences between normal and elevated cognitive workload during manual driving induced by a secondary n-back task (Solovey et al., 2014). Another study used classifiers to predict different types of driver distraction during manual driving using driving behavior and physiological measures (McDonald et al., 2020). A different study used eye tracking measures to predict perceived workload during robotic surgical skills training using a naive bayes classifier (Wu et al., 2020). A fourth study used classifiers to predict takeover performance of drivers at different mental workloads induced by a secondary task using physiological data, driving behavior, and external driving conditions (Du et al., 2020). The classifiers used in these four studies include decision tree, logistic regression, random forest, multilayer perceptrons, naive bayes, SVM, and kNN. The accuracy of the classification models ranged from around 0.5 to 0.9. These past studies influenced the design and model selection of this study.

To address the research gap in predicting driver mental workload at a semi-autonomous level, the goal of this study is to: develop models to predict driver mental workload using physiological and driving performance measures and to determine which factors are the most important for mental workload prediction.

2 Methods

2.1 Data Used

The data came from a study whose goal was to determine the effects of secondary tasks on driving performance at automation level 1. A medium fidelity, fixed-base driving simulator (miniSim) was used. Adaptive cruise control (ACC) or Lane Keeping Systems (LKS) were available to participants. Participants were required to keep constant headway when using the LKS system, maintain the vehicle in the center of the lane when using ACC, or keep constant headway and maintain the vehicle in the center of the lane when driving manually. While participants were driving, they were tasked with an auditory N-back test. They were either 0-back (no task), 1-back, or 3-back. There was a total of nine conditions that were a combination of the usage of different assisted driving systems and the N-back tests. This was done to vary overall task complexity and mental workload. Each condition was done in two consecutive blocks with a short break in between each block. The study utilized a within-subject design, and the nine conditions were randomly ordered. This is detailed in Figure 1.

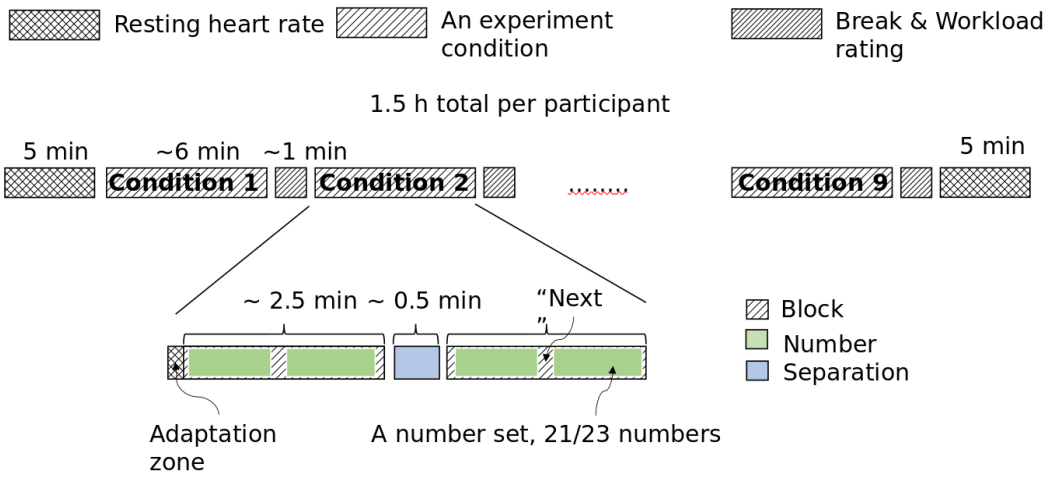

Figure 1. Condition and Block layout (Liang, 2019).

After each condition, participants were asked to evaluate their mental workload from driving and the secondary tasks using NASA-TLX scales. The weighted overall score was computed. During the experiment, participants' heart rate, eye tracking metrics, and driving performance were measured using an ECG heart rate sensor and data from the driving sim. The heart rate sensor used was the Polar H10 HR monitor (Tarvainen et al., 2018). The eye tracking measurements were recorded using FOVIO FX3 (Seeing Machines Inc. Canberra, Australia).

There was a total of twenty-five participants. With nine different conditions, there are a total of 225 samples across the nine conditions. However, data for certain participants were missing, and data from participant one were removed so only 210 of the 225 samples across the nine conditions were used.

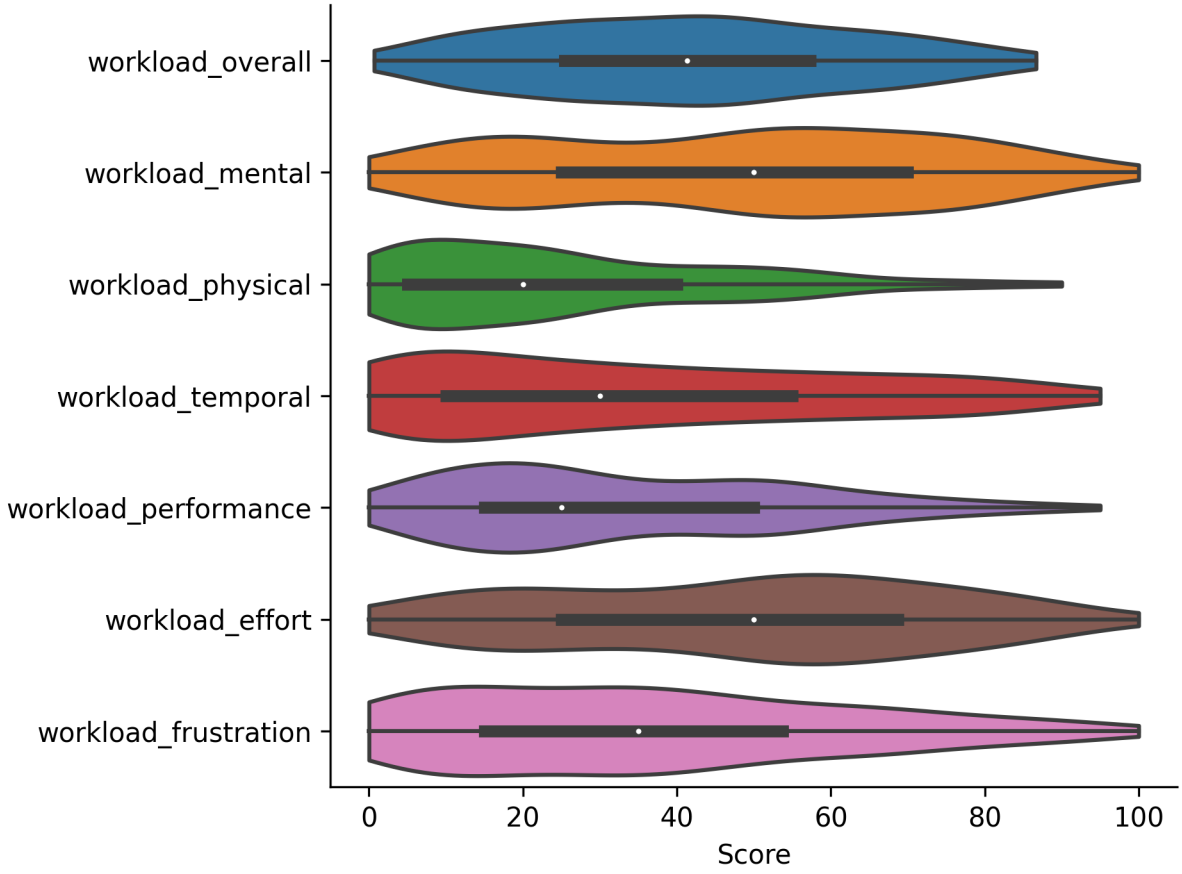

Figure 2. Violin plots of NASA-TLX scores and subscales

The distribution for some of the subscales do not appear to be normal (Figure 2). As such, different methods for determining High or Low workload were created to see if this would have a significant impact. These methods are outlined later.

2.2 Classification Models

The classification models used in this study are as follows:

- Logistic Regression (Logistic)
- Support Vector Machine with Radial Basis Function Kernel (SMV_rbf)
- Support Vector Machine with Linear Function Kernel (SMV_lin)
- Gaussian Naive Bayes
- Decision Tree
- Random Forest

The computer models were implemented using the scikit-learn package in python (Pedregosa et al., 2011). These models were selected due to their use for similar applications in other studies (Du et al., 2020; McDonald et al., 2020).

2.3 Data Preprocessing and Flow

The data was preprocessed and labeled. Binary classification was done using the NASA-TLX individual and overall scores. The exact target variables are mental, physical, temporal, performance, effort, and frustration individual scales as well as the overall NASA-TLX scores.

Each target was a number ranging from zero to one hundred and were labeled either Low or High mental workload for classification purposes. Different schemes for determining cutoffs for each label were done. This was done to see if the distribution of workload subscales would affect model performance when splitting between High or Low workload when using the 50-50 split. This is outlined in Table 1.

Table 1. Different Split Methods Used

Name	**Description**	
50-50 split	Scores below fifty were determined Low and scores above fifty were determined High	Low < 50 High > 50
Mean split	Scores below the mean for its category were determined Low and scores above the mean were determined High	Low < Score Mean High > Score Mean
50^{th} percentile split	Scores below the 50^{th} percentile for its category were determined Low and scores above the 50^{th} were determined High	Low < 50^{th} percentile High > 50^{th} percentile

The features were normalized, and repeated k-fold cross validation was done with 10 splits and 3 repeats for each model.

The entire feature set that was used is listed in detail in Table 2.

Table 2. List of Features Used

Eye Tracking		
Name	**Description**	**Relevance**
Pupil dilation/diameter	Diameter of pupil	It was found that larger dilations correlate to increased mental workloads (Marquart et al., 2015).
	Measurements: • Mean • Standard Deviation	
Gaze Entropy	Randomness of gaze locations	It was found that "gaze entropy increased as perceived workload increased" (Wu et al., 2020).
	Measurements: • Negative sum of element product of p multiplied by the log base 2 of p, where p is the normalized histogram counts of the gaze locations	
Fixation Count	Total number of fixations	Fixations are commonly used in the literature as a measure of mental workload. It was found that fixation duration increased with increasing mental demand (Marquart et al., 2015; Wu et al., 2020) while others found the opposite (Foy & Chapman, 2018).
	Measurements: • Count	
Fixation duration	Dwell time of each fixation	
	Measurements: • Mean • Standard Deviation	
Saccade length	Linear distance between consecutive fixation points in a saccade	Measures of saccades have been used in past studies relating eye fixations and driver's mental workload (Marquart et al., 2015).
	Measurements: • Mean • Standard Deviation	
Blink duration	Length of time eye is not visible for a consecutive number of eye measurements	Found to decrease with increasing mental workload (Marquart et al., 2015).
	Measurements: • Mean • Standard Deviation	
Blink rate	Count of blinks per minute	Mixed results according to Marquart et al., 2015. Found to be a good indicator for predicting mental workload in manual driving (Tran et al., 2017).
	Measurements: • Mean • Standard Deviation	
Blink latency	Total time elapsed between each recorded blink	Found to increase with increasing mental workload (Marquart et al., 2015).
	Measurements: • Mean • Standard Deviation	

Heart Rate		
Name	**Description**	**Relevance**
HRV (time and frequency domain)	Heart rate variability. R-R interval is the time interval between consecutive ECG R-waves	HRV measures, including metrics on the R-R interval, and HR were found to be significant in classification models when predicting driver takeover performance at SAE level 3 (Du et al., 2020). HRV and HR was also used to measure mental workload in a simulated driving environment (Mehler et al., 2009). HR was used to classify driver distraction in classification models (McDonald et al., 2020). Frequency measures including LF and HF have been found to decrease under higher mental workload (Mehler et al., 2011).
	Measurements: • Standard Deviation • SDNN (Standard deviation of normal-to-normal R-R intervals) • RMSSD (Root mean square of successive R-R interval differences) • NNxx (Count of successive R-R intervals that differ by more than x seconds) • pNNxx (Relative count of successive R-R intervals that differ by more than x seconds) • TINN (Triangular interpolation of normal-to-normal intervals) • HRV Triangular index • VLF (Very low frequency) • LF (Low frequency) • HF (High frequency) • LF/HF (ratio of LF and HF)	
HR	Heart rate	
	Measurements: • Mean • Minimum • Maximum • Mean R-R interval	
Nonlinear parameters	Nonlinear analysis and metrics of HRV	Poincare was found to be a good indicator for different mental effort task loads (Mukherjee et al., 2011).
	Measurements: • Poincare plots • ApEn (approximate entropy) • SampEn (sample entropy) • DFA (Detrended fluctuated analysis) using software from (Tarvainen et al., 2018).	

Driving Metrics		
Name	**Description**	**Relevance**
Headway	Bumper to Bumper Distance between driving car and front car, as defined by SAE International, 2018.	These metrics are commonly used when assessing driver's takeover performance at automation level 3 (McDonald et al., 2020). These metrics are also relevant for assessing driving performance at automation level 1 for the different conditions and usages of the available ADS to drivers.
	Measurements: • Mean • Standard Deviation	
Mean Time to Collision	Time till driving car collides with front car at current speed and accelerations, as defined by SAE International, 2018.	
	Measurements: • Mean • Standard Deviation	
Lane Deviation	Perpendicular distance of driving car's center from the center line of the lane	
	Measurements: • Mean • Standard Deviation	
Speed	Speed of the driving vehicle	
	Measurements: • Mean • Standard Deviation	

Each feature was evaluated for its importance using mutual information. Then, various computer models using the top k most relevant features were trained and evaluated. Each model used the entire feature set or a subset of the feature set including subsets such as physiological measurements only and driving metrics only. This was to see if models' performance changed significantly if only certain features were used.

2.4 Model Evaluation

The classification models were evaluated using their accuracy scores. The AUC – ROC score and F1 scores were also computed. Model fitting time was not taken into consideration.

Figure 3 summarizes data processing, model training and evaluation.

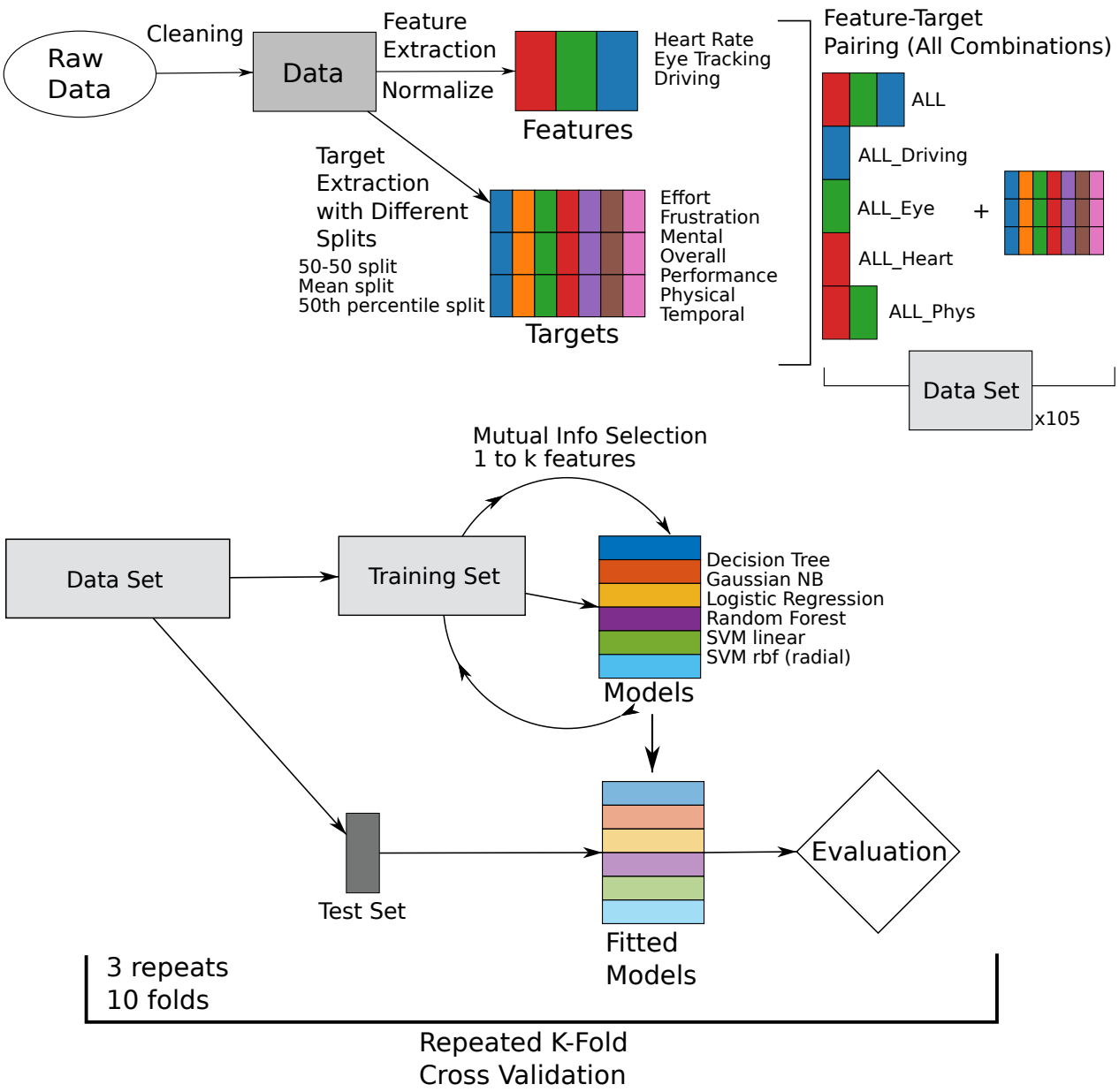

Figure 3. Data flow, Preprocessing, Model Training and Evaluation

2.5 Exploration

As exploration, models were trained using features extracted from block two (Figure 1) only. The features were still assigned to the targets with the same methodology described previously. This was done to see if observation window may have an effect on model performance. Smaller observation window sizes may also be more practical for real-time mental workload prediction. Additionally, the overall workload score collected at the end of each conditions may not reflect fluctuations in mental workload at earlier parts of the condition, mainly in block one (Figure 1).

3 Results

3.1 Model Performance

Table 3. Model performance using all features predicting overall workload across different splits (Block One & Block Two Data Included)

50-50 split			
Model	**Mean Accuracy**	**Mean ROC AUC**	**Mean F1 score**
Decision_Tree	0.673	0.641	0.520
Gaussian_NB	0.673	0.690	0.396
Logistic_Regression	0.663	0.699	0.190
Random_Forest	0.746	0.795	0.584
SVM_linear	0.663	0.698	0.456
SVM_rbf	0.716	0.745	0.414
Mean split			
Model	**Mean Accuracy**	**Mean ROC AUC**	**Mean F1 score**
Decision_Tree	0.627	0.632	0.623
Gaussian_NB	0.629	0.723	0.650
Logistic_Regression	0.671	0.738	0.669
Random_Forest	0.724	0.797	0.722
SVM_linear	0.719	0.770	0.721
SVM_rbf	0.697	0.760	0.705
50^{th} percentile split			
Model	**Mean Accuracy**	**Mean ROC AUC**	**Mean F1 score**
Decision_Tree	0.633	0.634	0.621
Gaussian_NB	0.627	0.722	0.658
Logistic_Regression	0.671	0.738	0.669
Random_Forest	0.732	0.796	0.729
SVM_linear	0.719	0.770	0.721
SVM_rbf	0.697	0.760	0.705

The models for the different splits all performed comparably to one another with a mean accuracy ranging from about 0.63 to about 0.73 for the overall workload target using all features. Table 3 shows that the random forest classifier performed the best across all splits. Note that the mean accuracy of the best model after mutual information feature selection was done is shown only. The different splitting methods for target classification did not have a significant impact on model performance.

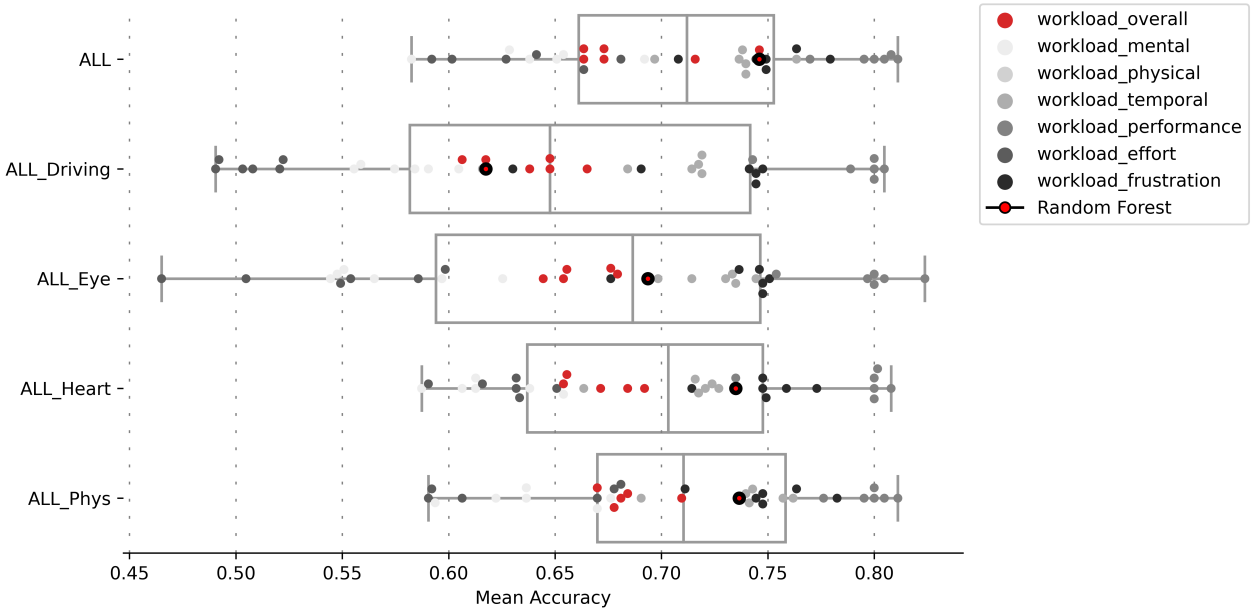

Figure 4. Accuracy of all models across all feature subsets and target variables for the 50-50 split.

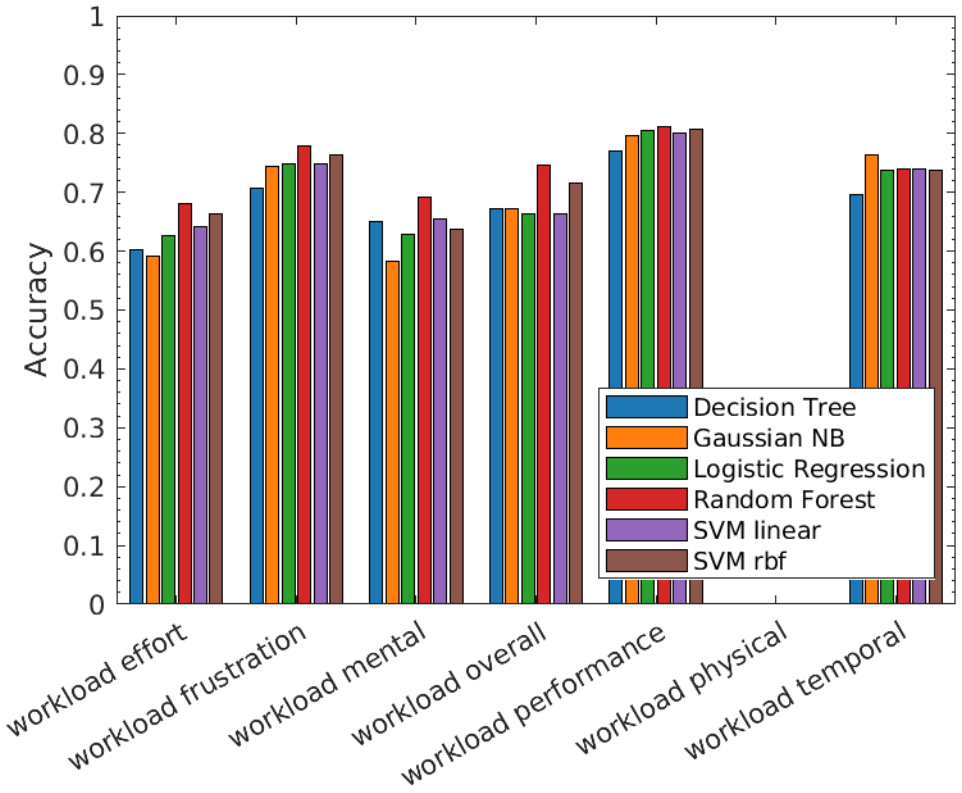

Figure 5. Accuracy *of models trained with all features across all target variables for the 50-50 split*

The accuracy for models whose target variable was workload physical with the 50-50 split were not computed (Figure 5). The data set had very few samples that had workload physical classified as High using the 50-50 split because of the distribution of workload physical scores. As such, the accuracy of these models was not meaningful as during training, some splits would have no samples with High workload physical in the test fold. The accuracy for frustration, performance, and temporal were higher than accuracy for other targets.

The swarm plots (Figure 4) show the mean accuracy of all models for all target variables across all feature subsets for the 50-50 split. Note that the Random Forest classifier trained with the overall workload target has been highlighted by a black outline. Models with all physiological data performed better than models with only eye tracking or only heart rate data. Models trained with all physiological data performed similarly to models trained with all features. Models trained with only driving behavior performed poorly or worse compared to other models. The 50-50 split has a distinct clustering of models trained with the same target variable whereas the other splits have a less defined clustering. Models whose target was overall workload generally fell within the interquartile range and were close to the mean.

Model performance across different target variables were not significantly different with random forest still performing marginally better (Figure 5).

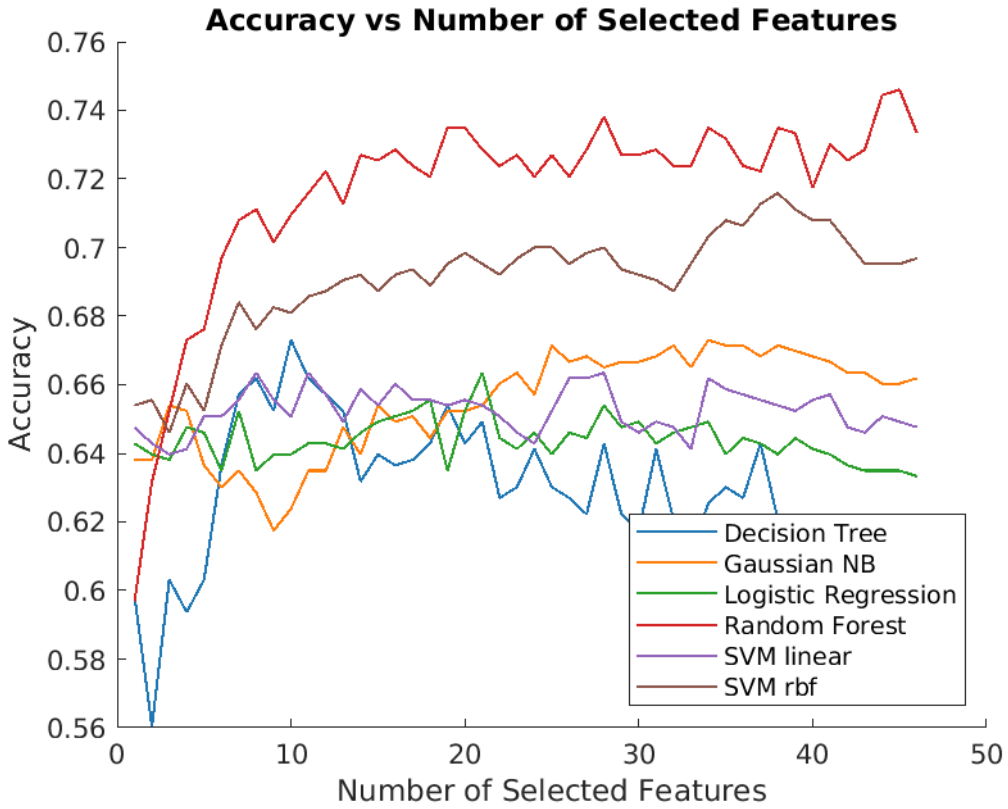

Figure 6. Accuracy of models as the number features used changes. The models shown were trained with all features for the overall workload target using the 50-50 split.

Models generally performed better as the number of features used increased (Figure 6). However, after a certain point, there was only marginal improvement or sometimes a decrease in performance.

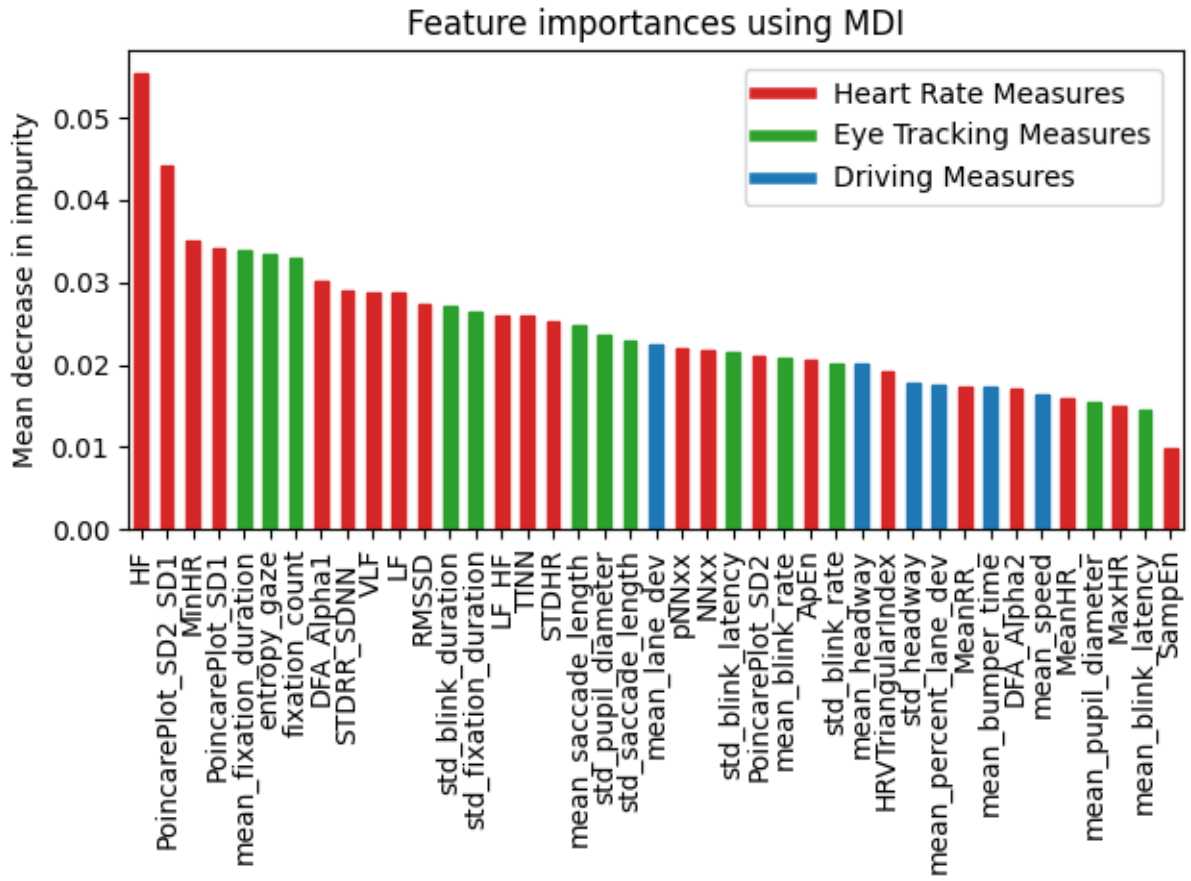

Figure 7. Feature Importance using mean decrease in impurity (MDI) for the best random forest classifier.

The feature importances of the best performing random forest classifier were computed (Figure 7). Heart rate measures and eye tracking measures were found to be the most important. The best performing random forest classifier used a total of forty-two features. However, performance between a random forest classifier using twenty features and the best performing random forest classifier was not significant (Figure 6).

3.2 Selected Features

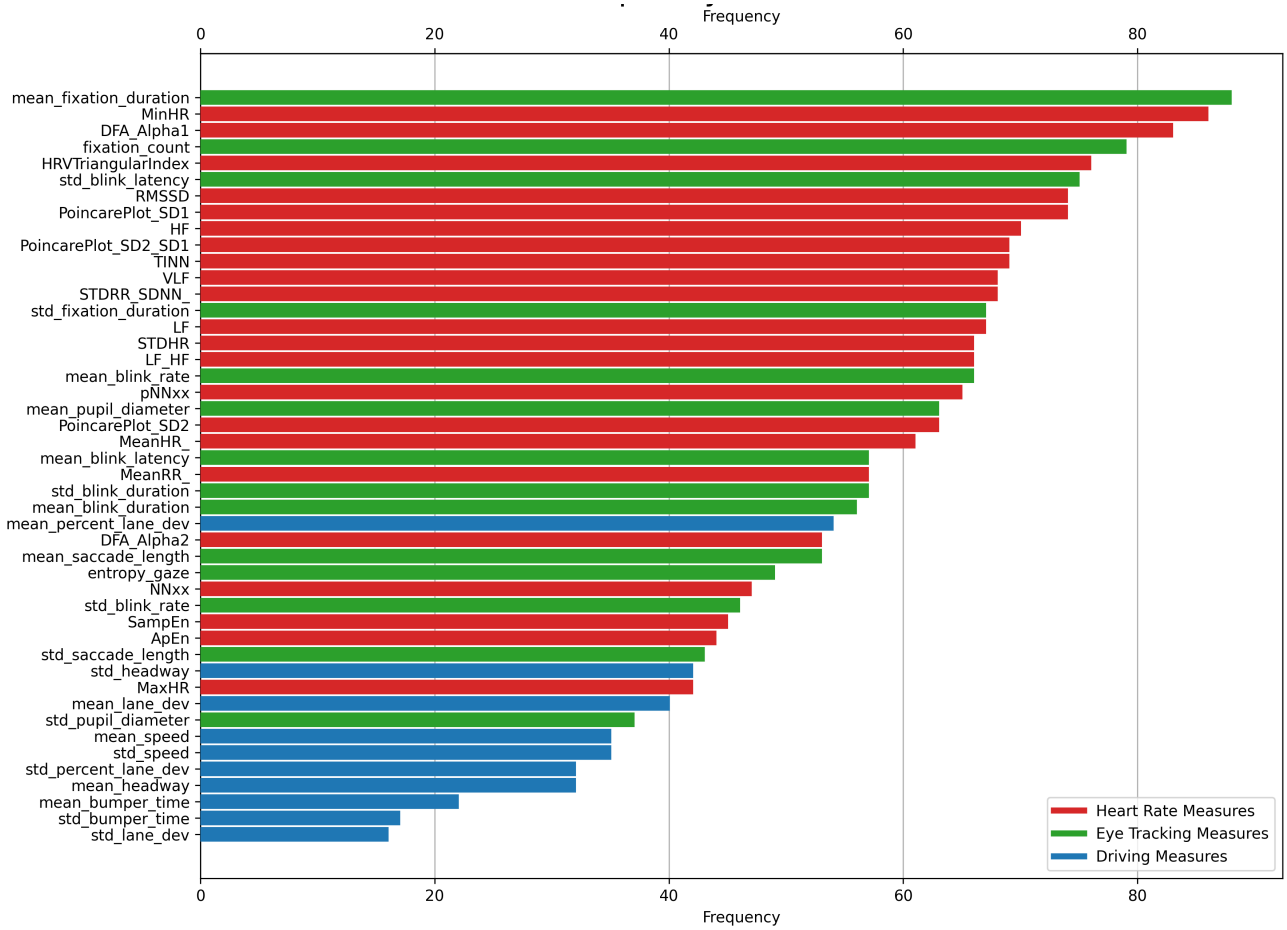

Figure 8. The most frequently used features across all models for the 50-50 split.

Models generally selected to use, based on mutual information selection, heart rate measures and eye tracking measures more than driving behavior measures (Figure 8).

3.3 Exploration

Table 4. Model performance using all features predicting overall workload across different splits (Block Two data only).

50-50 split			
Model	**Mean Accuracy**	**Mean ROC AUC**	**Mean F1 Score**
Decision_Tree	0.732	0.705	0.605
Gaussian_NB	0.617	0.708	0.567
Logistic_Regression	0.667	0.711	0.292
Random_Forest	0.770	0.809	0.626
SVM_linear	0.737	0.757	0.593
SVM_rbf	0.700	0.754	0.399
Mean Split			
Model	**Mean Accuracy**	**Mean ROC AUC**	**Mean F1 Score**
Decision_Tree	0.635	0.634	0.622
Gaussian_NB	0.648	0.715	0.667
Logistic_Regression	0.652	0.730	0.658
Random_Forest	0.724	0.790	0.720
SVM_linear	0.681	0.741	0.682
SVM_rbf	0.716	0.776	0.732
50th percentile split			
Model	**Mean Accuracy**	**Mean ROC AUC**	**Mean F1 Score**
Decision_Tree	0.643	0.640	0.625
Gaussian_NB	0.648	0.715	0.668
Logistic_Regression	0.657	0.730	0.658
Random_Forest	0.721	0.784	0.715
SVM_linear	0.681	0.741	0.678
SVM_rbf	0.717	0.776	0.734

There was no significant improvement in model accuracy for any of the splits if observation window size was constrained to block two (Figure 1) (Table 4). Random forest improved from an accuracy around 0.75 (Table 3) to 0.77 (Table 4).

4 Discussion

The objective of this study was to see how effectively certain classification techniques predict mental workload in the driving environment and to determine which factors are the most important for mental workload prediction. In general, the random forest classifier performed the best with an accuracy of 0.75. It was found that heart rate measures and eye tracking measures were the most important, with driving performance measures being the least important in mental workload prediction.

4.1 Split Methods and Target Selection

The distribution of scores for certain subscales may have affected model performance. However, model performance across the three splitting methods were not significantly different (Table 3). Since the mean split and 50^{th} percentile split require computing the mean and 50^{th} percentile across a large sample while the 50-50 split can be done with a smaller sample size, the 50-50 split is more practical. Further analysis was thus done using the 50-50 split.

Using overall workload as the target variable was comparable to using a subscale in regard to model accuracy (Figure 5). However, the distribution of overall workload in the data set was more similar to a normal distribution than some of the subscales (Figure 2). Additionally, the overall score reflects the various factors of mental workload. As such, it is recommended to use the overall workload NASA-TLX score.

4.2 Model Performance

Models trained with all physiological features only performed similarly with models trained with all features with a mean accuracy of 0.72 compared with a mean accuracy of 0.73. Models trained with only eye tracking features or only heart rate features also performed similarly but slightly worse as well. However, this was not the case for models trained with only driving features. The features most commonly selected were also from either heart rate measures or eye tracking measures. This is similar to a previous study (Solovey et al., 2014) which found that models trained with only heart rate data or only physiological data performed comparably to models that were trained with all features.

The random forest classifier was found to perform the best at predicting mental workload. This was consistent across all the different split methods. This is similar to what Du et al. (2020) found when predicting driver takeover performance. The random forest classifier may have performed the best because they aggregate the results of many decision trees, which reduces overfitting (Du et al., 2020). Logistic regression classifiers generally performed the worst.

The results from the exploration section show no significant improvement in overall model performance (Table 4). There was marginal improvement in the random forest classifier. This may indicate that if there were details in mental workload in block 1 not reflected in the final overall mental workload score, they were not as significant on model performance.

4.3 Important Features

Heart rate measures have been commonly used in the past to effectively measure mental workload (Mulder, 1992). This aligns with this study's results that heart rate measures were good indicators

for mental workload. Many of the most selected heart rate features were HRV measures with some being nonlinear measurements. The medium frequency band has been found to be sensitive to mental workload (Charles & Nixon, 2019) which is in the range of the LF band used in this study. However, the HF band was selected more often than the LF band by models. One caveat about HRV measures it that it is not a good indicator for mental workload sensitivity (Paxion et al., 2014). However, this study only differentiates between low and high mental workload which may be a reason HRV measures were still selected more often.

Eye tracking measures were also found to be important indicators for mental workload prediction. However, they were overall found to be less important, and models trained with only eye tracking features performed slightly worse than models trained with only heart rate features. This could be due to the study design of the data set used. Participants were tasked with either keeping constant headway or keeping the driving vehicle within the lane. This requires participants to have their visual attention on a certain spot like the lead vehicle or the lane markers. Eye tracking measures may reflect this behavior as well as mental workload. This reflects the discussion that eye tracking measures may reflect visual workload and visual tasks (Marquart et al., 2015).

Driving measures were not found to be important indicators for mental workload prediction. This is a similar finding to Solovey et al. (2014), which found that models performed significantly worse when predicting mental workload using only driving behavior. One reason is that driving/situational complexity was not taken into account. It was found that scenario type was one of the top sixteen important features when predicting driver takeover performance (Du et al., 2020). It was found that increased situational complexity generally increased mental workload and decreased driving performance; additionally, drivers in a manual setting were found to use compensatory mechanisms to maintain good driving performance even if mental workload increased (Paxion et al., 2014). A similar effect can be happening in this study where despite increased mental workload, drivers were able to compensate and still maintain proper headway or keep within the designated lane. Situational complexity was not significantly varied.

4.4 Limitations and Future Work

Several limitations should be taken into consideration for future work. First, the measurement of features were done across both blocks one and two or just block two in the exploration. This means the observation windows were over a span of several minutes. Details and fluctuations in features throughout each condition may have been lost and affect model performance. Future work should explore the possibility of using shorter observation windows for more granular prediction of mental workload. Second, only a limited number of participants were included in this study. The relatively small sample size may affect the robustness of the classification models. Future work should include more participants and demographic diversity to allow more generalizable results.

5 Conclusions

This study aimed to predict mental workload of drivers in a semi-automated setting using machine learning models. It was found that the random forest classifier performed the best with an accuracy around 0.75. In addition, this study identified that the most selected features for mental workload prediction were heart rate measures and eye tracking measures. Models that used only physiological measures performed comparably to models that used all features available while models that used only driving performance measures performed significantly worse than models that used all features available. The result of this study can be used to guide the design of future in-vehicle mental workload prediction systems. A better understanding of drivers' real-time mental workload will improve the development of an adaptive human-machine interface bettering driving performance and road safety.

Acknowledgments

I would like to thank the SURF program at Purdue University for providing support for this research and the tools to conduct research.

I would like to thank my SURF Graduate Advisor, Jie Zhu, for his support in navigating the SURF program and conducting research.

References

Charles, R. L., & Nixon, J. (2019). Measuring mental workload using physiological measures: A systematic review. *Applied Ergonomics*, *74*(September 2016), 221–232. https://doi.org/10.1016/j.apergo.2018.08.028

Chen, W., Sawaragi, T., & Horiguchi, Y. (2019). Measurement of Driver's Mental Workload in Partial Autonomous Driving. *IFAC-PapersOnLine*, *52*(19), 347–352. https://doi.org/10.1016/j.ifacol.2019.12.083

Du, N., Zhou, F., Pulver, E. M., Tilbury, D. M., Robert, L. P., Pradhan, A. K., & Yang, X. J. (2020). Predicting driver takeover performance in conditionally automated driving. *Accident Analysis and Prevention*, *148*. https://doi.org/10.1016/j.aap.2020.105748

Foy, H. J., & Chapman, P. (2018). Mental workload is reflected in driver behaviour, physiology, eye movements and prefrontal cortex activation. *Applied Ergonomics*, *73*(June 2017), 90–99. https://doi.org/10.1016/j.apergo.2018.06.006

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

Liang, N. (2019). *Assessing the Effects of Cognitive Secondary Tasks and Automation Type on Changes in Heart Rate: Implications for the Potential Use of Nanotechnology*. *August*.

Marquart, G., Cabrall, C., & de Winter, J. (2015). Review of Eye-related Measures of Drivers' Mental Workload. *Procedia Manufacturing*, *3*, 2854–2861. https://doi.org/10.1016/j.promfg.2015.07.783

McDonald, A. D., Ferris, T. K., & Wiener, T. A. (2020). Classification of Driver Distraction: A Comprehensive Analysis of Feature Generation, Machine Learning, and Input Measures. *Human Factors*, *62*(6), 1019–1035. https://doi.org/10.1177/0018720819856454

Mehler, B., Reimer, B., Coughlin, J. F., & Dusek, J. A. (2009). Impact of Incremental Increases in Cognitive Workload on Physiological Arousal and Performance in Young Adult Drivers. *Transportation Research Record*, *2138*(1), 6–12. https://doi.org/10.3141/2138-02

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*. 590–597. https://doi.org/10.17077/drivingassessment.1451

Mukherjee, S., Yadav, R., Yung, I., Zajdel, D. P., & Oken, B. S. (2011). Sensitivity to mental effort and test–retest reliability of heart rate variability measures in healthy seniors. *Clinical Neurophysiology*, *122*(10), 2059–2066. https://doi.org/https://doi.org/10.1016/j.clinph.2011.02.032

Mulder, L. J. M. (1992). Measurement and analysis methods of heart rate and respiration for use in applied environments. *Biological Psychology*, *34*(2–3), 205–236. https://doi.org/10.1016/0301-0511(92)90016-N

Paxion, J., Galy, E., & Berthelon, C. (2014). Mental workload and driving. In *Frontiers in Psychology* (Vol. 5, Issue DEC). Frontiers Research Foundation. https://doi.org/10.3389/fpsyg.2014.01344

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research*, *12*, 2825–2830.

SAE International. (2018). Surface Vehicle. In *SAE International* (Vol. 4970, Issue 724).

SAE International. (2021). *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*.

Solovey, E. T., Zec, M., Perez, E. A. G., Reimer, B., & Mehler, B. (2014). Classifying driver workload using physiological and driving performance data: Two field studies. *Conference on Human Factors in Computing Systems - Proceedings*, 4057–4066. https://doi.org/10.1145/2556288.2557068

Tarvainen, M. P., Lipponen, J., Niskanen, J., & Ranta-aho, P. O. (2018). USER ’ S GUIDE HRV Standard. *Kubios Oy*, *173*(3), 5–172.

Tran, C. C., Yan, S., Habiyaremye, J. L., & Wei, Y. (2017). Predicting driver’s work performance in driving simulator based on physiological indices. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10688 LNCS*, 150–162. https://doi.org/10.1007/978-3-319-72038-8_12

Wu, C., Cha, J., Sulek, J., Zhou, T., Sundaram, C. P., Wachs, J., & Yu, D. (2020). Eye-Tracking Metrics Predict Perceived Workload in Robotic Surgical Skills Training. *Human Factors*, *62*(8), 1365–1386. https://doi.org/10.1177/0018720819874544