1.0 Literature review

## 1.1 Background

Cognitive load is an increasingly studied topic which encapsulates a broad spectrum of fields, including health, psychology (Torkamani-Azar, Lee & Bednarik, 2022; Joseph and Murugesh, 2020) and defense applications (Tazeem, Sajjanhar, Lee and Jia, 2021; Murthy & Biswas, 2022). One such method of discerning potential levels of cognitive workload is through the use of eye tracking (Larsson, 2016; Murthy & Biswas, 2022). Previously, much of this has been tracked manually by researchers, but with the rise of computer vision and machine learning has come the potential of automated end to end methods of categorization (Tazeem et al., 2021).

The following is a comprehensive analysis of recent literature. Firstly, search methodology will be provided, followed by key features, proposed datasets and machine learning models, and papers that have used these datasets and models to allow for a baseline, and finally the proposed direction for filling a gap in the literature and associated research questions.

## 1.2 Literature search methods

A comprehensive literature review was conducted to review recent literature on eye movement and cognitive load. Further to this, state of the art machine learning methods employed in this area were examined to determine popular and effective methods, and identify potential gaps in recent literature.

The literature review was conducted through the use of Google Scholar, SCOPUS and IEEE Explore. Literature was filtered to articles from 2019 and beyond to ensure that only the most current literature was obtained. A total of 21 papers were reviewed for this literature review. Several were excluded as they were not relevant enough to the research topic.

The following search terms were included: eye movement cognitive load machine learning, eye tracking data set, eye movement data set, machine learning cognitive load

## 1.3 Features

A number of features have been found to be beneficial in machine learning models used to predict cognitive load in recent literature. Pupil size, blinks and fixation have all been found to be strong predictors ((Shojaeizadeh, Djamasbi, Paffenroth & Trapp, 2019; Murugesh (2020), Jerčić, Sennersten & Lindley, 2020; Ramakrishan, Balasingam & Biondi, 2021; Bafna, Hansen and Baekgaard, 2020; (Appel, Sevcenko, Wortha, Tsarava, Moeller, Ninaus, Kasneci, & Gerjets, P, 2019; Babu, JeevithaShree, Prabhakar, Saluja, Pashilkar, & Biswas, 2019).

### 1.3.1 Pupil Size

Shojaeizadeh, Djamasbi, Paffenroth & Trapp (2019) sought to predict task demand through eye movement data. One of the key findings from their study was the importance of pupil size data in predicting task demand, with their model built using saccade-to-fixation pupil dilation and pupil dilation variance ratio, with an accuracy of 79%, using a random forest model.

This is not the only study examining this relationship, and further researchers have corroborated this, including Joseph and Murugesh (2020), Jerčić, Sennersten and Lindley (2020) and Ramakrishan, Balasingam and Biondi (2021).

Ramakrishan et al. (2021) simulated 4 different experiments to test the relationship between multiple eye movement measures including pupil size and cognitive load. These tasks were unmanned vehicle operation, memory recall tasks, delayed memory recall tasks, and finally simulated driving. They found that in all but one task, pupil size increased, on average as the task difficulty increased. The one task where this was not the case was due to a confounding stimulus of light. In summary, they found that while pupil size was a strong indicator of cognitive load, that pupil size alone is not enough due to the affects of other stimuli, such as light. As this research will be using a publicly available dataset, this is not something that can be controlled for as it is limited to datasets that are available. It may, however be an important direction for future research.

### 1.3.2 Blink Frequency

Another physical indicator that has been linked to cognitive load is blinking; in particular, blink frequency. Multiple researchers have provided evidence to show this link, including Shojaeizadeh et al. (2019), and Bafna, Hansen and Baekgaard (2020)**.**

Bafna, et al. (2020) sought to measure the link between cognitive load and blink rate during a typing task. They had 18 participants memorize simple and difficult sentences over a period of four days. Using blink rate, pupil diameter frequency and interval, typing speed and error rate as features in their prediction they found that as the task performance increased, typing performance lowered; and while blink frequency, duration and interval increased. This further provides support for this concept that blink frequency can be a useful predictor when determining cognitive load.

### 1.3.3 Fixation

The final feature that will be discussed in this literature review are fixations. This is another physical indicator that has been linked to cognitive load. Fixation is the point that a gaze is directed to. The duration is the length of time that a person’s gaze if fixated at a specific point and the count is the number of times a fixation occurs within a task.

This feature is another eye movement feature that has been identified as integral to predicting cognitive load in recent literature (Appel, Sevcenko, Wortha, Tsarava, Moeller, Ninaus, Kasneci, & Gerjets, P, 2019; Babu, JeevithaShree, Prabhakar, Saluja, Pashilkar, & Biswas, 2019; ).

Liu, Li, Yeh, & Chien (2022) narrowed in on fixation as a predictor for cognitive load. They obtained fixation data and level of cognitive load using a modified video game. They changed the number of tiles and words presented in this game to simulate low and high levels of cognitive load. The measure of cognitive load was done with the NASA TLX score. This is the same measure used by by Ktistakis, Skaramagkas, Manousos, Tachos, Tripoliti, Fotiadis and Tsiknakis (2022), which will be examined later.

They used one way ANOVA to analyse the relationship between fixations and cognitive load. Their findings were that in high cognitive load tasks, participants had longer fixation durations and less fixations. This study is important to note as the same cognitive measure is used by Ktistakis et al. (2022) on the proposed COLET dataset. It is therefore possible that similar results are obtained when using machine learning models on this dataset.

|  |  |
| --- | --- |
| **Feature** | **Reason for inclusion** |
| Pupil Size | Pupil size has been linked to differing levels of cognitive load in several studies (Shojaeizadeh, Djamasbi, Paffenroth & Trapp, 2019; Murugesh (2020), Jerčić, Sennersten & Lindley, 2020; Ramakrishan, Balasingam & Biondi, 2021  ) |
| Blink rate | Higher blink rate can indicate higher levels of cognitive load (Shojaeizadeh et al., 2019; Bafna, Hansen and Baekgaard, 2020) |
| Fixation duration | Longer durations of fixations indicates higher cognitive load (Appel, Sevcenko, Wortha, Tsarava, Moeller, Ninaus, Kasneci, & Gerjets, P, 2019; Babu, JeevithaShree, Prabhakar, Saluja, Pashilkar, & Biswas, 2019) |

*Table 1. Summary of eye movement measures.*

## 1.4 Data Sets

This section will firstly be prefaced by noting that there are very limited, appropriate, open data sources available that could be used in this study. Of all the studies examined, only 5 contained open data sources. At least one contained a link to repositories that were no longer supported, and one contained the note that it would be supplied upon request, however due to time restrictions this is not feasible to obtain. As a result, 2 open-source data sources were included in this research and are described below. He et al. (2022) used a specific dataset available to physiotherapists, however this is not available to use for this research. Bozkir et al. (2019) used a private dataset they compiled in their previous work using a virtual reality driving simulation. This was done specifically to examine the effect of a virtual or augmented reality setup.

### 1.4.1 Colet

The Colet dataset was created by Ktistakis et al. (2022). The authors sought to create a publicly available dataset that combined both objective and subjective measures of cognitive load. They noted these levels of cognitive load as low, medium and high.

In order to create this dataset, they recruited 28 participants and had them perform 4 activities. These activities consisted of visual search activities of variable complexity. Their performance was measured in 2 different ways. The first was using the NASA-TLX. This is a measure that incorporates multiple measures, including mental demand, physical demand, frustration, and effort (NASA, 2020). Multiple objective measures were included and are explained in detail below in table 2.

Further to this, performance measures were also included. This was included in the form of number of mistakes and time taken to complete. This is reported as both an inverse efficiency score, reaction time and percentage of correct errors.

Ktistakis et al. (2022) applied several machine learning methods to this dataset to predict cognitive load. The models applied were Gaussian Naïve Bayes, Random Forest, Support Vector Machine, Ensemble Gradient Boosting, K-Nearest Neighbours, Bernoulli Naïve Bayes, Logistic Regression and Decision Trees.

Hyperparameters were tuned through a random grid search iterated 1,000 times and a test-train split of 20-80 was applied. Ktistakis et al. (2022) reported both accuracy and F-score and applied a k-fold cross validation of 5. The results obtained are detailed in table 4

|  |  |
| --- | --- |
| **Key measure** | **Author Measures** |
| Fixation | Frequency, duration |
| Saccade | Frequency, duration, velocity |
| Blink | Frequency, duration, |
| Pupil | Diameter |

*Table 2. Key cognitive load measures and measures obtained by authors*

## 1.4.2 A Machine Learning Approach for Detecting Cognitive Interference Based on Eye-Tracking Data

Rizzo, Ermini, Zanca, Bernabini and Rossi (2022) provide a publicly available dataset. They sought to determine cognitive load through the use of cognitive interference in the form of a Stroop test. They tracked eye movement data across four tasks with varying levels of cognitive interference.

The data consists of 64 subjects that were tasked with performing 2 actions, reading and naming. They varied the level of cognitive load by either presenting interference or not presenting interference. This interference presented itself as a more difficult reading task. For example, in the reading without interference task participants would read the world ‘blue’ in black text, while the same word would be presented with incorrect spelling and an opposite colour, for example ‘blu’ in red text. Multiple eye movements were recorded, including number of fixations, length of fixation, and saccades. Unfortunately no blink data, nor any pupil data was recorded.

|  |  |
| --- | --- |
| **Key measure** | **Author Measures** |
| Fixation | Number of fixations, average fixation length, maximum fixation length, horizontal and vertical regressions |
| Saccade | Frequency, duration, amplitude, angle, maximum angle, average angle, minimum distance, maximum distance |

*Table 3. Measures used by Rizzo et al. in relation to the key*

## 1.5 Previous Machine Learning Models

The previous machine learning models used by authors to predict cognitive load will be reviewed. Table 4 (below) shows accuracies and F1 measures (where reported) obtained by the authors on the datasets reviewed in section 1.4 or their own datasets. Accuracy is the proportion of correct predictions, and the F1 score is the harmonic mean of the precision and recall, combining precision and recall into a single metric (Google, 2022). Deep learning model performance have been noted in table 4, however this research will limit its scope to non-deep learning models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy Obtained** | **F1 Score** | **Author** | **Dataset** |
| KNN | 97.2% (low, medium, high) | Not reported | He et al. (2022) | eyePhysioset (not publicly available) |
|  | 78.82% (high, low) | 77.29% (high, low) | E. Bozkir, D. Geisler and E. Kasneci (2019) | Private dataset compiled by the author |
| Decision Tree | 73.4% (high, low) | 71.65% (high, low) | E. Bozkir, D. Geisler and E. Kasneci (2019) | Private dataset compiled by the author |
|  | 86.8% (high, low) | FILL THIS IN | Joseph, A.W., Vaiz, J.S. and Murugesh, R., (2021) | Private dataset compiled by the author |
|  | 74% (low/medium, high) | 73% (low/medium, high) | Ktistakis et al. (2022) | COLET Ktistakis et al. (2022) |
| SVM | 93.3% (low, medium, high) | Not reported | He et al. (2022) | eyePhysioset (not publicly available) |
| 80.7% (high, low) | 80.98 (high, low) | E. Bozkir, D. Geisler and E. Kasneci (2019) | Private dataset compiled by the author |
| 67% (reading with vs without interference. Fixation, saccade, normalization) | 68% (reading with vs without interference. Fixation, saccade, normalization) | Rizzo, Ermini, Zanca, Bernabini and Rossi (2022) | A Machine Learning Approach for Detecting Cognitive Interference Based on Eye-Tracking Data |
| 72% (Naming with vs without interference, Fixation, saccade, normalization) | 68% (Naming with vs without interference, Fixation, saccade, normalization) | Rizzo, Ermini, Zanca, Bernabini and Rossi (2022) | A Machine Learning Approach for Detecting Cognitive Interference Based on Eye-Tracking Data |
|  | 69% (low, medium) | 69% (low, medium) | Ktistakis et al. (2022) | COLET Ktistakis et al. (2022) |
| Gaussian Naïve Bayes | 59% (low, medium high), 88% (low, high) | 59% (low, medium high), 86% (low, high) | Ktistakis et al. (2022) | COLET Ktistakis et al. (2022) |
| Logistic Regression | 51% (low, medium, high) | 50% (low, medium, high) | Ktistakis et al. (2022) | COLET Ktistakis et al. (2022) |
| 68% (Naming with vs without interference, Fixation, saccade, normalization) | 69% (Naming with vs without interference, Fixation, saccade, normalization) | Rizzo, Ermini, Zanca, Bernabini and Rossi (2022) | A Machine Learning Approach for Detecting Cognitive Interference Based on Eye-Tracking Data |
|  | 68% ( reading with vs without interference), Fixation, saccade, normalization) | 68% (reading with vs without interference), Fixation, saccade, normalization) | Rizzo, Ermini, Zanca, Bernabini and Rossi (2022) | A Machine Learning Approach for Detecting Cognitive Interference Based on Eye-Tracking Data |
| Gradient Boosting Ensemble Model | 58% (low, high) | 57% (low, high) | Ktistakis et al. (2022) | COLET Ktistakis et al. (2022) |
| Random Forest | 84% (low, high) | 84% (low, high) | Ktistakis et al. (2022) | COLET Ktistakis et al. (2022) |
| 97.8 (low, medium, high) | Not reported | He et al. (2022) | eyePhysioset |
| 74.36 (low, high) | 72.99% (low, high) | E. Bozkir, D. Geisler and E. Kasneci (2019) | Private dataset compiled by the author |
|  | 62% (reading with vs without interference. Fixation, saccade, normalization) | 67% (reading with vs without interference. Fixation, saccade, normalization) | Rizzo, Ermini, Zanca, Bernabini and Rossi (2022) | A Machine Learning Approach for Detecting Cognitive Interference Based on Eye-Tracking Data |
|  | 62% (naming with vs without interference. Fixation, saccade, normalization) | 59% (naming with vs without interference. Fixation, saccade, normalization) | Rizzo, Ermini, Zanca, Bernabini and Rossi (2022) | A Machine Learning Approach for Detecting Cognitive Interference Based on Eye-Tracking Data |
| Artificial Neural Net (ANN) | 78% (naming with vs without interference. Fixation, saccade, normalization) | 70 % (naming with vs without interference. Fixation, saccade, normalization) | Rizzo, Ermini, Zanca, Bernabini and Rossi (2022) | A Machine Learning Approach for Detecting Cognitive Interference Based on Eye-Tracking Data |
| 68% (reading with vs without interference. Fixation, saccade, normalization) | 68% (reading with vs without interference. Fixation, saccade, normalization) | Rizzo, Ermini, Zanca, Bernabini and Rossi (2022) | A Machine Learning Approach for Detecting Cognitive Interference Based on Eye-Tracking Data |
| feedforward neural network | 91.8% | Not reported | He et al. (2022) | eyePhysioset (not publicly available) |
| recurrent neural network | 94.4% | Not reported | He et al. (2022) | eyePhysioset (not publicly available) |

*Table 4. Summary of accuracies obtained in papers using KNN, Decision Tree and SVM machine learning models.*

### 1.5.1 KNN

K-Nearest-Neighbours (KNN) is a popular, simple, high performing machine learning algorithm that has been used previously by researchers to predict cognitive load from eye movement data (He et al., 2022; Bozkir, Geisler, Kasneci, 2019; Ktistakis, Skaramagkas, Manousos., Tachos, Tripoliti, Fotiadis, & Tsiknakis, 2022). KNN seeks to classify a data point based on surrounding data points, or ‘neighbours’.

Previously, one of the main drawbacks of this approach was the amount of computational power needed, however processing power has continued to increase, and this has become less of an issue over time (Cunningham, Delany, 2022). A further issue noted with KNN is that it will struggle with non-relevant features (Osisanwo, Akinsola, Awodele, Hinmikaiye, Olakanmi & Akinjobi, 2017). As noted prior, the proposed models will be built on features identified through initial feature analysis, so this will be a non-issue.

Further to this, KNN is more appropriate than other models that have been applied in previous research such as Gaussian Naïve bayes due to its discriminative nature, meaning that it seeks to predict labels (Varghese, 2018); in this case, high and low cognitive load

### 1.5.2 SVM

Support vector machine classifiers seek to develop a hyperplane between values, enabling it to assign data points to classes (Noble, 2006). It has been extensively used as a classifier across many fields, including cognitive load given eye movement data (Bozkir, et al., 2019; Rizzo et al. 2022; Ktistakis et al. 2022).

SVM models have several advantages and disadvantages. The first advantage is that SVM is a model that works well for classification, however it is a non-probabilistic model (Sui, He, Vilsen, Meng, Teodorescu & Stroe, 2021), meaning that it is not able to consider random variation. This is a potential downside to the single use of the model.

SVM works well when there are two distinct classes, however it can underperform when there are more features than data points (Scikit-learn, 2022). This is a non-issue with the proposed datasets as neither of these have more features than data points, and further to this, the goal of the research is to classify binary classes, high and low cognitive load. Due to its high performance on similar datasets, difference to KNN and decision trees, it will be used as one of the proposed models in the final, stacked, ensemble model.

### 1.5.3 Decision Trees

Decision trees are models that classify through multiple nodes. Each node will seek to classify a data point based on several input features (Kingsford & Salzberg, 2008). From these multiple nodes, the final classification is made.

One of the strengths of decision trees is that they can handle both categorical and continuous data and can classify quite quickly, in a simple and interpretable way (Kingsford & Salzberg, 2008). It is a strong model for prediction and, unlike KNN or SVM, it bases its classification on multiple nodes. It does not require previous assumptions of the distribution of the data, and handles collinearity efficiently (Varghese, 2018). It is, however less robust and accurate than a decision tree, which is an ensemble model consisting of multiple decision trees.

As with the proposed KNN and SVM models, it has been used previously to great effect, is an appropriate classifier, and differs from the previously reviewed models. Due to this, it will be included in the final, stacked, ensemble model.

### 1.5.4 Gaussian Naïve Bayes

Gaussian Naïve Bayes (GNB) classifiers work on the probabilistic theory of Bayes, and classifies data points accordingly. It works under the assumption of a gaussian distribution on an attribute given the class label. It has been used in multiple previous classification problems including text classification, document classification (Jahromi & Taheri, 2017), and was used by Ktsikatis et al. (2022) in classifying high, medium and low levels of cognitive load given eye movement data.

One of the key strengths of GNB is that it can still classify accurately even when independence assumptions are violated (Al-Aidaroos, Bakar & Othman, 2010). A further strength is its flexibility and ability to work well with large datasets (Akkaya & Colakoglu, 2019). The main weakness of this model is the requirement for large datasets to get the best results (Akkaya & Colakogly, 2019).

Due to its high accuracy on the Colet dataset, ability to work even when independence assumptions are violated, and flexibility, it will be used as one of the benchmark models and used within the final, stacked model.

### 1.5.5 Ensemble Models

Two main ensemble models were discovered during the literature search. These are random forests and gradient boosted regression trees. These are usually chosen for their low overfitting, low computational expense and speed (Appel, 2021). There is a lack of literature examining other ensemble methods, including stacked modelling which is a method to address single, weak learning models.

The main advantage of using a stacked ensemble model is the ability to strengthen a model’s predictive power by combining single, weak learning models.

## 1.6 Conclusion

Based on recent literature, there is limited literature examining the effectiveness of ensemble models outside Random Forest and Gradient Boosted models in predicting levels of cognitive load using eye movement data. in the literature for ensemble machine learning models beyond gradient boosted regression trees and random forests.

Several machine learning models will be used, and a final, ensemble model will be used. The specific models that will be used are Support Vector Machine (SVM, Gaussian Naïve Bayes (GNB) and Decision Trees. Each of these models has been used in recent research to high levels of accuracy with a summary shown below in table 4. Please note that Ktistakis et al. (2022) sought to classify multiple classes. The classes predicted are noted next to the accuracies obtained in the table 4.

It is therefore expected that by using multiple models together, that this research will improve upon the existing models on the datasets presented by Ktistakis et al. (2019) and Rizzo et al. (2022). There is limited current literature examining ensemble models beyond gradient boosted models and regression trees when seeking to classify cognitive load based on eye movement. This research will also seek to further research in the field by investigating this gap. Further to this, by employing 3 different categorization methods; a Gaussian probabilistic approach with GNB; hyperplanes with SVM and a probabilistic approach with CART then the resulting stacked ensemble model will have a broader approach to categorization than a single, weaker model.

This research will seek to answer 3 main research questions. The first research question is, ‘which features in the available datasets are relevant for detecting cognitive load?’, the second is ‘can eye movement features be used to determine high and low cognitive load through the use of machine learning?’ and the second question is ‘will a stacked ensemble model outperform the previously presented models on the previously presented datasets?’.

This literature review has presented several important measures in predicting cognitive load through eye movement data, including fixations, pupil size and blinks. It has presented multiple machine learning models that have been used to great effect previously, and will serve as baseline models, as well as 2 data sets that will be used for analysis. It has presented the strengths and weaknesses of these models, as well as an argument for using ensemble models. The next part of this research will introduce the research design.

Research Design

# Data Pre-processing

## Colet

The COLET dataset was provided as a matlab file. This was converted to excel files to analyse the data more easily in python.

Multiple features were extracted based on the previous literature review. The count of blinks was obtained by counting the number of data points within the blink data, grouped by participant by task, indicating the number of blinks per participant per task. The average pupil diameter was obtained by taking the average value of the diameter grouped by participant by task. The count of fixations and the count of saccades were obtained by using the velocity identification threshold. This was used by the Colet authors (Ktistakis et al., 2022). It works under the assumption that saccades and fixations have different velocities. Further to this, consecutive fixations are collapsed into groups and counted once, and the difference in time between observations must be above a specific threshold for it to be considered a fixation. These counts are then grouped together by participant and by task. This is summarised below in table 1.

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Velocity Threshold** | **Time Threshold** | **Group** |
| Saccade | Must be greater than 0.45 | N/A | N/A |
| Fixation | Must be less than 0.45 | Must be greater than 0.55 between movements | Consecutive Fixations are grouped together and count as 1 towards the final |

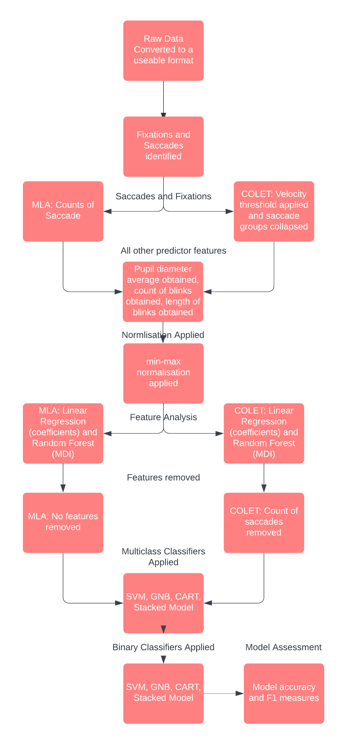
*Table 1. Summary of thresholds applied to identify fixations and saccades in the COLET dataset*

The final Colet dataset consists of the number of blinks per participant per task, average blink length per participant per task, the average score identified through the NASA TLX survey per participant per task, the average pupil diameter per participant per task, the number of fixations per participant per task and the count of saccades per participant per task.

|  |  |
| --- | --- |
| **Feature** | **Aggregate applied** |
| Blinks | Count |
| Blink Duration | Average |
| Pupil Size | Average |
| Fixations | Count |
| Saccades | Count |
| Cognitive load | N/A |

*Table 2. Summary of features for the COLET dataset*

Any rows with missing values were dropped as these cannot be used in the machine learning analysis. Values were also converted to integer values to allow for analysis. The values of the NASA TLX recordings were converted from their raw values to 1, 2, or 3 meaning low, medium and high cognitive load. These values were identified from the author’s paper where they noted 50-100 is high, 30-49 is medium, and 0-29 are considered low values.

Finally, the feature data was scaled to values between 0 and 1 using a min-max scaler. A summary of the entire process is shown in figure 1 (below). 

*Figure 1. Summary of modelling process.*

A Machine Learning Approach for Detecting Cognitive Interference Based on Eye-Tracking Data (MLA)

A Machine Learning Approach for Detecting Cognitive Interference Based on Eye-Tracking Data (referred to as MLA for the remainder of this report). The second dataset was provided as 1996 excel files, however could not be opened in python due to errors. These were therefore converted to CSV files for ease of analysis. These files were obtained from Rizzo, Ermini, Zanca, Bernabini and Rossi (2022).

Multiple files were included, but not all were used in the analysis. Interest area was not used as the purpose of this research was not to identify if the eye movement data itself was correct, but to use the data to identify cognitive load. Fixation files were used, gaze files were used and saccades files were used.

Minimal pre-processing was required for this data. Pupil size was recorded for both left and right eyes sequentially. These values were combined into one eye record. Average diameter was taken from these values.

The count of fixations were taken from the ‘current\_fix\_pupil’ field by taking the count of this, and grouping by the participant, resulting in a count of fixations by participant by task. Similarly, the count of saccades was obtained by taking the count of values for ‘current\_sac\_duration’, resulting in the count of saccades by participant by task. The count of blinks were taken by counting the number of values for the ‘current\_sac\_blink\_end’ field, resulting in the count of blinks by participant and task.

As with the COLET dataset, any rows with missing data were dropped as these missing values could not be used in the machine learning analysis.

In this case, the target feature for cognitive load are the tasks, as these are what was manipulated by the authors. The target variables were the tasks (1 to 4), and for the binary condition, the interference tasks were combined to form one target, and those without interference were combined to form the second target.

|  |  |
| --- | --- |
| **Feature** | **Aggregate applied** |
| Blinks | Count |
| Blink Duration | Average |
| Pupil Size | Average |
| Fixations | Count |
| Saccades | Count |

*Table 3. MLA Features*

# Feature Analysis

## COLET

The COLET dataset features were analysed in 2 different ways. The first way was through a linear regression model. The feature coefficients were used as a way to examine feature importance. The second method for feature analysis was to examine the mean decrease in impurity using a random forest model. The results are summarised below in figures 2 and 3. As the count of saccades had the lowest coefficient score and acted as a detractor in the linear regression analysis, and did not contribute significantly in the random forest mean decrease in impurity analysis, the feature was removed from the dataset.

*Figure 2. COLET feature analysis linear regression*

*Figure 3. COLET feature analysis regression tree*

## MLA

The same feature analyses were conducted on the MLA dataset. The results of these are summarised below in figures 4 and 5. It was decided that no features would be removed following this analysis. The reasoning was that while the count of saccades was the largest detractor in the linear regression analysis, it was also the largest contributor in the regression tree analysis.

*Figure 4. Linear regression feature analysis of the MLA features.*

*Figure 5. Regression tree feature analysis of the MLA features.*

# Hyperparameter Tuning

## Support Vector Classifier

Prior to model building the model hyperparamters were tuned. A 10 cross fold validation, random grid search method was used.

The optimal hyperparamters found for both datasets were as follows:

COLET: 'C': 10, 'gamma': 1, 'kernel': 'rbf'

MLA: 'C': 100, 'gamma': 1, 'kernel': 'sigmoid'

## Gaussian Naïve Bayes (GNB)

The same approach was used to tune the hyperparameters for the Gaussian Naïve Bayes classifiers.

The optimal hyperparamters found for both datasets were as follows:

COLET: var\_smoothing = 0.533669923120631

MLA: var\_smoothing = 2.848035868435805e-09

## Classification and Regression Tree (CART)

The same approach was used to tune the hyperparameters for the CART models.

The optimal hyperparamters found for both datasets were as follows:

COLET: 'max\_depth': 2, 'max\_features': 0.6, 'min\_samples\_leaf': 0.08

MLA: 'max\_depth': 3, 'max\_features': 0.2, 'min\_samples\_leaf': 0.04

# Models

Finally, with the hyperparameters tuned, the models were created and used to predict their targets. The first set of models predicted the targets as provided by the authors. For COLET, this was predicting low, medium and high cognitive load. This was identified through the mean NASA TLX score ranges. As noted by Ktistakas et al. (2022), this was scores from 0-29 as low, 30 to 49 as medium, and 50 to 100 as high. These 3 levels were used for the multiclass classifiers, and the high and low scores were merged into 1, and are considered ‘abnormal’ ranges, while the medium score acted as the second target.

For the MLA dataset, the authors did not measure levels of cognitive load, but did manipulate the levels of load through their activities. These activities were split into 4. Reading without interference, reading with interference, naming without interference and naming with interference. For the multiclass classifiers, all 4 were target variables, while for the binary classifiers, the interference conditions were grouped together, while the no interference were grouped together.

The models used were a support vector classifier (SVC), classification and regression tree (CART) and a Gaussian naïve bayes (GNB) model. Finally, these classifiers were stacked together into a stacked, ensemble model.

# Results

The results of each classifier across both the binary and multiclass classifiers are noted below in table 4. The accuracy was greater across all classifiers than the F1 score, indicating lower precision and recall. The stacked classifiers had difficulty identifying abnormal cognitive load levels, which results in the lower-than-expected accuracy obtained.

Overall, the binary classifiers outperformed the multiclass classifiers. The models did not perform as well as those in the presented papers, and the stacked models did not perform as well as expected. The likely reasons for these results will be examined below in the discussion.

For both datasets classifiers were able to identify cognitive load at greater than 50% accuracy (except for the stacked binary classifier on the MLA dataset).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Classifier** | **Dataset** | **Accuracy (Multiclass)** | **Accuracy (Binary)** | **F1 (Macro average - Multiclass)** | **F1 (Macro average - Binary)** |
| SVC | COLET | 0.482 | 0.724 | 0.344 | 0.579 |
| CART | COLET | 0.482 | 0.621 | 0.456 | 0.505 |
| GNB | COLET | 0.482 | 0.655 | 0.222 | 0.396 |
| Stacked | COLET | 0.448 | 0.655 | 0.311 | 0.396 |
| SVC | MLA | 0.370 | 0.629 | 0.344 | 0.617 |
| CART | MLA | 0.407 | 0.621 | 0.328 | 0.444 |
| GNB | MLA | 0.222 | 0.512 | 0.155 | 0.493 |
| Stacked | MLA | 0.222 | 0.482 | 0.217 | 0.475 |

*Table 4. Summary of classifier results.*

# Discussion

Following the model results, this paper will discuss the reason for differences in analysis when compared to the authors and limitations before discussing the responses to the previously noted research questions.

There are several possible reasons for the difference between the results obtained in this paper compared to previous literature. While processing the COLET dataset it was noted that there were large amounts of missing and incorrect data, particularly in blink data. Further to this, there was a large amount of missing blink data. Records with missing data was removed, resulting in a smaller amount of remaining data. This low sample size will impact the model prediction as has been shown in previous literature (for example, Vabalas, Gowen, Poliakoff & Casson, 2019).

Further to this, the method of identifying saccades and fixations in the COLET dataset was unclear. It was interpreted as noted in table 1. If this was interpreted incorrectly this will result in further difference in analysis for the COLET dataset.

Another reason for difference is the difference in hyperparameters. It is unclear exactly which hyperparameters were used by Rizzo et al. (2022) and Ktistakis et al. (2022). Finally, due to the random selection in the test-train split, there will be a difference in the data selected for testing and training data in both datasets, which will further result in differences.

It is difficult to compare the results obtained when compared to those presented by Rizzo et al. (2022). In order to limit the scope of the study, the binary targets identified were created by combining both interference conditions and by combining both non-interference conditions. This differs from what Rizzo et al. (2022) performed for their analysis.

While the performance of the stacked modelling was not as high as expected, the values do not differ significantly from the individual models. As the stacked model is comprised of each of these single models, it is not unsurprising that it does not significantly outperform the individual models.

The following research questions were posed at the beginning of this paper. “which features in the available datasets are relevant for detecting cognitive load?”, the second is “can eye movement features be used to determine high and low cognitive load through the use of machine learning?” and the second question is “will a stacked ensemble model outperform the previously presented models on the previously presented datasets?”.

Through feature analysis it was identified that blink frequency, blink durations, fixation frequency, and pupil diameter were all identified as relevant features for the COLET dataset. The MLA dataset is more difficult to answer. All features were found to be detractors in a linear regression analysis, while the greatest detractor in this analysis was found to be the strongest predictor in the regression tree analysis. It was therefore determined that all features would be used in the analysis.

As noted previously, both datasets classifiers were able to identify cognitive load at greater than 50% accuracy (except for the stacked binary classifier on the MLA dataset). This indicates that yes, based on the data examined here, eye movement data can be used to predict cognitive load.

Finally, comparing the models to those presented in previous research, they underperform when compared to those presented by Ktistakis et al. (2022) on the COLET dataset, and also underperform compared to those presented by Rizzo et al. (2022) on the MLA dataset. This underperformance includes the results of the stacked models (both binary and multiclass). While they do not underperform compared to the single models by a wide margin, they still perform worse than expected.

# Threats to validity

Several threats to validity are identified. Neither studies indicate from where the authors recruited their participants are from. This may be an threat to external validity as it limits the applicability of the research in a real world setting.

An internal threat to validity is the identification of saccades and fixations in the COLET dataset. As noted previously, this was interpreted as shown in table 1. This was identified through the methods noted by Ktistakis et al. (2022). If this is interpreted differently it will change the count of saccades and fixations, and thus will change the feature values. Another internal threat to validity is the ample size. As it is quite small this will impact the model performance.

# Future Directions

Following this research there are several ideas for future directions. One such direction would be employing the use of deep learning. Multiple researchers have used deep learning to predict cognitive load from eye movement data, including Rizzo et al. (2022).

Further to this, the inclusion of additional features is suggested for future directions. Due to time limitations a limited number of features were included, however using the same features as the authors including minimum, maximum saccade and fixation durations, minimum, maximum and average velocity, amongst many others.

Finally, a further suggestion would be to obtain data separately rather than relying on existing data. This will require additional work including ethical considerations, equipment hire and participant recruitment, but will allow a closer supervision of data collection.

# Conclusion

In conclusion, this paper has posed and attempted to answer 3 research questions using multiple machine learning models and multiple datasets. The research questions posed were ‘which features in the available datasets are relevant for detecting cognitive load?’ the second is ‘can eye movement features be used to determine high and low cognitive load through the use of machine learning?’ and the second question is ‘will a stacked ensemble model outperform the previously presented models on the previously presented datasets?’. It was found that using the datasets provided by Rizzo et al. (2022) and Ktistakis et al. (2022) that almost all features as noted in tables 2 and 3 were identified as useful in detecting cognitive load. It was also found that cognitive load could be detected through the use of machine learning models, but that the stacked models did not outperform the individual models.

# References

Akkaya, B. and Çolakoğlu, N., 2019. Comparison of Multi-class Classification Algorithms on Early Diagnosis of Heart Diseases.

Al-Aidaroos, K.M., Bakar, A.A. and Othman, Z., 2010, March. Naive Bayes variants in classification learning. In *2010 international conference on information retrieval & knowledge management (CAMP)* (pp. 276-281). IEEE.

Appel, T. 2021. *Cross-participant and cross-task classification of cognitive load based on eye tracking* (Doctoral dissertation, Eberhard Karls Universität Tübingen).

Appel, T., Sevcenko, N., Wortha, F., Tsarava, K., Moeller, K., Ninaus, M., Kasneci, E. and Gerjets, P. 2019 Predicting cognitive load in an emergency simulation based on behavioral and physiological measures. In *2019 International Conference on Multimodal Interaction* (pp. 154-163).

Babu, M.D., JeevithaShree, D.V., Prabhakar, G., Saluja, K.P.S., Pashilkar, A. and Biswas, P. 2019. Estimating pilots’ cognitive load from ocular parameters through simulation and in-flight studies. *Journal of Eye Movement Research*, *12*(3).

Bafna, T., Hansen, J.P.P. and Baekgaard, P. 2020. Cognitive load during eye-typing. *ACM symposium on eye tracking research and applications* (pp. 1-8).

Bozkir, E. Geisler, D. and Kasneci E. 2019. Person Independent, Privacy Preserving, and Real Time Assessment of Cognitive Load using Eye Tracking in a Virtual Reality Setup *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 1834-1837, doi: 10.1109/VR.2019.8797758.

Cunningham, P. and Delany, S.J. 2021. K-nearest neighbour classifiers-a tutorial. *ACM Computing Surveys (CSUR)*, *54*(6), pp.1-25.

Google, 2022. Evaluate the performance of processors, accessed 02/12/2022. Available: <https://cloud.google.com/document-ai/docs/workbench/evaluate>

He, D., Wang, Z., Khalil, E. B., Donmez, B., Qiao, G., & Kumar, S. 2022. Classification of Driver Cognitive Load: Exploring the Benefits of Fusing Eye-Tracking and Physiological Measures. *Transportation Research Record*, *2676*(10), 670–681. <https://doi.org/10.1177/03611981221090937>

Jahromi, A.H. and Taheri, M., 2017, October. A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features. In *2017 Artificial intelligence and signal processing conference (AISP)* (pp. 209-212). IEEE.

Jerčić, P., Sennersten, C. & Lindley, C., 2020. Modeling cognitive load and physiological arousal through pupil diameter and heart rate. *Multimedia Tools and Applications*, *79*(5), pp.3145-3159.

Joseph, A.W. and Murugesh, R., 2020. Potential eye tracking metrics and indicators to measure cognitive load in human-computer interaction research*. J. Sci. Res,* 64(1), pp.168-175.

Joseph, A.W., Vaiz, J.S. and Murugesh, R. 2021. Modeling Cognitive Load in Mobile Human Computer Interaction Using Eye Tracking Metrics. In *International Conference on Applied Human Factors and Ergonomics* (pp. 99-106). Springer, Cham.

Kingsford, C. and Salzberg, S.L., 2008. What are decision trees?. *Nature biotechnology*, *26*(9), pp.1011-1013.

Ktistakis, E., Skaramagkas, V., Manousos, D., Tachos, N.S., Tripoliti, E., Fotiadis, D.I. and Tsiknakis, M., 2022. COLET: A dataset for COgnitive workLoad estimation based on eye-tracking. *Computer Methods and Programs in Biomedicine*, *224*, p.106989.

Liu, J.C., Li, K.A., Yeh, S.L. and Chien, S.Y., 2022. Assessing Perceptual Load and Cognitive Load by Fixation-Related Information of Eye Movements. *Sensors*, *22*(3), p.1187.

NASA, 2020. ‘NASA TLX: Task Load Index’ Available online: <https://humansystems.arc.nasa.gov/groups/tlx/>

Noble, W.S., 2006. What is a support vector machine?. *Nature biotechnology*, *24*(12), pp.1565-1567.

Osisanwo, F.Y., Akinsola, J.E.T., Awodele, O., Hinmikaiye, J.O., Olakanmi, O. and Akinjobi, J., 2017. Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, *48*(3), pp.128-138.

Pomplun, M. and Sunkara, S., 2019. Pupil dilation as an indicator of cognitive workload in human-computer interaction. *Human-Centered Computing* (pp. 542-546). CRC Press.

Ramakrishnan, P., Balasingam, B. and Biondi, F., 2021. Cognitive load estimation for adaptive human–machine system automation. *Learning control* (pp. 35-58). Elsevier.

Rizzo, A., Ermini, S., Zanca, D., Bernabini, D. and Rossi, A., 2022. A Machine Learning Approach for Detecting Cognitive Interference Based on Eye-Tracking Data. *Frontiers in Human Neuroscience*, *16*.

Scikit-Learn, 2022. *Support Vector Machines,* Accessed 22/11/2022.

Sui, X., He, S., Vilsen, S.B., Meng, J., Teodorescu, R. and Stroe, D.I., 2021. A review of non-probabilistic machine learning-based state of health estimation techniques for Lithium-ion battery. *Applied Energy*, *300*, p.117346.

Tazeem, H., Sajjanhar, A., Lee, G. and Jia, D. (2021) *Random Forest for Event Classification of Eye Movements: Towards Effective Cognitive Workload Estimation*

Vabalas, A., Gowen, E., Poliakoff, E. and Casson, A.J., 2019. Machine learning algorithm validation with a limited sample size. *PloS one*, *14*(11), p.e0224365.

Varghese, D. (2018). *Comparative Study on Classic Machine learning Algorithms,* Accessed 23/11/2022