Analysis of Machine Learning Assisted Fatigue Identification in Radiology Readings

by

Matthew Hayes

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved April 2022 by the
Graduate Supervisory Committee:

Troy McDaniel, Chair
Aurel Coza
Hemanth Venkateswara

ARIZONA STATE UNIVERSITY

May 2022

ABSTRACT

Fatigue in radiology is a readily studied area. Machine learning concepts applied to the identification of fatigue are also readily available. However, the intersection between the two areas is not a relative commonality. This study looks to explore the intersection of fatigue in radiology and machine learning concepts by analyzing temporal trends in multivariate time series data. A novel methodological approach using support vector machines to observe temporal trends in time-based aggregations of time series data is proposed. The data used in the study is captured in a real-world, unconstrained radiology setting where gaze and facial metrics are captured from radiologists performing live image reviews. The captured data is formatted into classes whose labels represent a window of time during the radiologist's review. Using the labeled classes, the decision function and accuracy of trained, linear support vector machine models are evaluated to produce a visualization of temporal trends and critical inflection points as well as the contribution of individual features. Consequently, the study finds valid potential justification in the methods suggested. The study offers a prospective use of maximum-margin classification to demarcate the manipulation of an abstract phenomenon such as fatigue on temporal data. Potential applications are envisioned that could improve the workload distribution of the medical act.

ACKNOWLEDGEMENTS

I would like to thank Dr. Bhavik Patel and Dr. Leland Hu of Mayo Clinic for the extremely generous offering of their time. The unique data gathered by the study would not be possible without Dr. Patel's enthusiasm for the prospects of the study. Nor would it have continued if not for Dr. Patel's and Dr. Hu's flexibility and patience in allowing an academic experiment to be conducted within their workplace. I would also like to thank my committee members: Dr. Troy McDaniel, Dr. Aurel Coza, and Dr. Hemanth Venkateswara. I greatly appreciate the time they spent serving as members of my thesis committee. A special thanks to Dr. Coza for giving me his time, insight, and knowledge in his lab as well as in my committee. Finally, a thank you to my family for their beloving support.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Problem Statement

Expertise is not the sole determinant of performance. Environmental, physical, mental, and other factors influence performance efficacy. Intuitively, a child cannot play a sport like basketball at the same performance level as a professional athlete. Although the disparity is primarily due to differences in physical development, similarly severe disparities can be produced by mental factors. Pressure, in its many forms, represents a degree of mental distress. Just as the child cannot compete against the professional athlete, a person executing a task under pressure suffers a performance ailed by burdens they would not have if they were devoid of pressure. With the focal form of performance being fatigue, the main goal of this paper is to explore machine learning data analysis models in the extraction of information pertinent to the timing of the onset of fatigue. In particular, the object of the analysis is the onset of oculomotor and cognitive-emotional fatigue in radiology experts.

Fatigue in Radiology

Quantifying fatigue in itself is not a new subject nor is fatigue in radiology a novel area of research (Hanna et al., 2018; Krupinski et al., 2010; Li et al., 2020; Vosshenrich et al., 2021). Fatigue can be manifested through the growing weakness of muscles in executing physical tasks or as mental tiredness hindering intellectual acuity. Additionally, fatigue in radiology presents a unique form: oculomotor fatigue. Ocular fatigue affects eyesight and is an acute risk during medical image readings (Krupinski et al., 2010). Studies have delved into the objective and subjective measuring of fatigue in

radiology and have found evidence suggesting a decrease in performance as the workday progresses (Krupinski et al., 2010). The effects of a decrease in performance during radiology readings has been shown to include a reduction in diagnostic accuracy coupled with increased oculomotor strain (Krupinski et al., 2010). Even search patterns employed by radiologists during readings become less effective with fatigue, causing an increase in the amount of time and effort to review an image (Hanna et al. 2018). Diagnostic accuracy and review time and effort are essential metrics in an occupation where unsuccessful performance exacts a heavy price. The various effects of fatigue on radiologists' performance are serious when left to fester and pose a clear occupational hazard. The existence of fatigue-born harmful effects in radiology is readily apparent. The nature of such effects is also no stranger to recorded observation. However, the goal of this paper is to analyze the leveraging of machine learning methodologies in the exploration of fatigue in radiology.

Machine Learning Approaches for Fatigue

Statistical methods such as linear regression have been used to quantify fatigue in radiology settings. One such study used the method to analyze the similarities between resident and staff reports across the length of a workday. The study found a negative relationship between staff and resident report similarity over time, suggesting the detrimental effect of fatigue on residents (Vosshenrich et al., 2021). However, heavier applications of statistical machine learning methods can be found in areas such as the construction industry. Li et al. (2020) performed a study where wearable eye-tracking equipment was used to collect gaze data from construction workers to analyze mental fatigue. The study was able to categorize levels of fatigue through a clustering algorithm;

the categories of which were used in support vector machine classification revealing successful classifications of fatigue (Li et al., 2020). Eye tracking is an important avenue for the analysis of fatigue, one which is used in this paper's study, and is represented in other areas outside of construction such as driving (Du et al., 2021). However, while gaze analysis paired with machine learning methods is not uncommon in radiology research, the focus on fatigue apparent in construction and driving based studies is not as widely shared.

Machine Learning in Radiology

Gaze research in the realm of radiology favors perception characterization (Kocejko et al., 2019; Lévêque et al., 2021; Machado et al., 2018; Mall et al., 2018). Using eye-tracking technology, studies aim at understanding how radiologists perceive images (Lévêque et al., 2021; Machado et al., 2018). One such study attempted to identify gaze fixations of radiologists during lung cancer diagnosis reviews. The study was able to collect fixation data from eye-trackers and constructed regions of the lung where radiologists focused their attention (Machado et al., 2018). Another study combined the use of fixation and saccade metrics to analyze search patterns in radiologists performing mammogram reviews. The results suggested a difference in gaze behaviors across radiologists according to their varying degree of experience (Lévêque et al., 2021). While all were expert level radiologists, residents have been used in other research delving into the nature of visual perception. Kocejko et al. (2019) designed an experiment to capture gaze data in residents which could then be processed for assessing skills against those of expert-level and untrained-level image reviewers. The mapping of visual perception data to competency affirmed established metrics used in skill

assessments and suggested an opportunity to monitor skill acquisition through gaze analysis (Kocejko et al., 2019). However, each aforementioned study relegated machine learning tools in favor of general statistical interpretations. Mall et al. (2018) represented a departure from general statistical interpretations of radiologist gaze patterns by adopting deep machine learning approaches. Using a deep convolutional neural network, the study was able to derive accurate models learned according to radiologists' own gaze behaviors as opposed to predetermined metrics (Mall et al., 2018). Utilized in machine learning or general statistical methods, gaze data is an established source of information in radiology research. Yet the interest in fatigue is not as shared as similarly gaze-based research in areas like construction and driving. While visual perception data is congruent across radiology and fatigue experiments another source of information is useful when machine learning methods are being approached.

Study Goals and Hypothesis

The relative novelty of using gaze analysis to tackle fatigue exploration in radiology merits the consideration of tangential studies. Returning to the fatigue analysis in driving research, facial data is another source of information utilized (Du et al., 2021; Khan et al., 2018). Facial features, extracted from simple RGB camera recordings of participants, have been shown to be valid determinants in machine learning models detecting driver fatigue (Du et al., 2021). Additionally, facial expression analysis has been studied as an enabler of driver fatigue recognition (Khan et al., 2018). With the goal of this paper's study borrowing context from multiple different areas, it is important to establish the similarities from where exploration can be directed. In the instance of this research, gaze and facial data are the sources of information collected. Furthermore,

machine learning methods such as support vector machines serve as avenues of analysis. Fatigue acts as the main principle uniting the data and analysis.

The machine learning assisted study of fatigue in radiology presents a certain degree of novelty. Yet, such a study, as previously described, is not uncommon in other areas. Borrowing directly from those studies would result in a review of prediction capabilities: can fatigue of radiologists be modeled from gaze and facial data to provide inferences concerning future levels of fatigue? While such a review has merit and a precedent set by other, non-machine learning based research into radiology fatigue, this paper poses a parallel review. Instead of using machine learning tools such as classifiers for predictive purposes, their use for detecting temporal change is considered. Fatigue exists on the temporal plane; it is related directly to the passing of time. Existing studies have modeled the flowing of fatigue across time, but this research looks to analyze a level deeper than modeling by exploring the nature of temporal change in fatigue. This research assumes the existence of inflection points within the flow of fatigue over time where the degree of the condition changes relatively dramatically. To explore the assumption, the novel use of tools such as classifiers for detection rather than prediction is considered. Further, the setting of this research poses an additional novelty. Gaze and facial data are captured from radiologists, similar to other studies, but in an unconstrained, real-world environment.

Leveraging inspirations from fatigue-related machine learning research to fill a gap in fatigue-related radiology research, the study aims to develop a novel methodology for the analysis of fatigue. The study will accomplish this by implementing a contemporary machine learning technique which allows for the observation of temporal

fatigue characteristics from real-life, real-time radiology visual tasks. Thus, the implementation will explore the hypothesis set by the study:

H1: A methodology exists that uses machine learning classifiers to detect the temporal effects of abstract phenomenon such as fatigue in multivariate space at a degree that allows for the discovery of inflection points demarcating the progress of such phenomenon.

CHAPTER 2

LITERATURE REVIEW

A more general summary of the literature previously introduced can be found in the discussion given by Hooda et al. (2021) on studied machine learning techniques to detect fatigue. Four approaches of detecting fatigue were analyzed: mathematical models, rule based, machine learning, and deep learning. The paper established mathematical models as correlators between different processes affecting fatigue and using features such as sleep and wakefulness duration. The Two Process model was given as an example of a fatigue detector for drivers using features belonging to distinct processes related to sleepiness and wakefulness. The Aircrew Fatigue Evaluation Model is another mathematical model that monitors pilot alertness during flight. Fuzzy logic in the fatigue detecting Fuzzy Inference System algorithm exchanged mathematical models for rule-based analysis. However, the discussion entered the realm of learning by reviewing the varying features studies have used to detect fatigue. Facial features, eye properties, and biological metrics have been used to varying degrees as inputs to different machine learning methodologies. Deriving the state of yawning from face detection was one example given for the use of facial features. Parallelly, the drawing of 63 landmarks from facial behavior to derive eye aspect ratio, nose length ratio, and mouth opening ratio among others was another use of facial features. Capturing the opened or closed state of the eye along with its position and illumination was a discussed use of eye-related features. Biometrics such as skin conductance and electroencephalogram (EEG) signals constituted another category of features experimented with for fatigue detection.

Coinciding with the explanation of common features was the appraisal of the machine learning models that consumed them (Hooda et al., 2021).

AdaBoost, Support Vector Machine (SVM), and Gradient Boosting Decision Tree (GBDT) were among the machine learning models investigated. AdaBoost was used with eye data to select features and train a binary classifier capable of detecting eye states: opened or shut. SVM was used across studies designing models to detect signs of fatigue from biological metrics and facial features. Skin conductance was used within an SVM model with relatively high accuracy in the detection of fatigue levels while components of facial behavior reached similarly high levels of accuracy when used as features in another SVM model. A study interested in the characteristics of EEG signals leveraged the GDBT method to detect levels of fatigue in drivers when coupled with features drawn from such signals. However, while the paper offered clear examples of machine learning techniques being applied to the detection of fatigue, they were compared to similarly minded deep learning approaches (Hooda et al., 2021).

Autoencoders were the predominant topic in the paper's discussion of fatigue detection through deep learning. The studies presented were interested in biometric data such as EEG and electro-oculogram (EOG) signals. Layered autoencoders were then used to feed outputs into regression or classifier models that could label data as existing in a state of fatigue or not (Hooda et al., 2021). The summary of machine learning developments in fatigue detection provided by Hooda et al. (2021) serves as context for the objective of this study. The priorly introduced examples of fatigue identifiers in the fields of driving and construction supplement those reviewed here (Du et al., 2021; Khan et al., 2018; Li et al., 2021). However, the study does not aim to develop a state-of-the-art

machine learning model to pair with those previously discussed, but to observe the characteristics of an established model in the area of fatigue. Further, the lack of a common intersection between fatigue and machine learning in radiology research presents an opportunity to make such an observation in a novel environment. Radiology research has been introduced as having various studies comment on the nature of fatigue in the profession as well as those that derived expertise metrics with machine learning techniques (Krupinski et al., 2010; Mall et al., 2018). The study hopes to assuage the gap present in radiology research by putting forth the remarked observations against the hypothesis of the study.

CHAPTER 3

EXPERIMENT

Experimental Procedure

With institutional review board approval, the data gathering took place at the Mayo Clinic Hospital, Phoenix, Arizona. A setup consisting of a lab computer and monitor, camera, and a Tobii-30 eye tracker was placed within a radiology reading room at a designated workstation. The workstation was designated by the volunteering radiologist of the particular day. The radiologist performed their shift, reading from the hospital setup while the cameras of the lab study recorded gaze and facial features onto the lab computer. The physical lab equipment was setup in the morning before the radiologist began their first readings of the day. Minimal interference from the researchers occurred until the completion of the recording session. Recordings were taken in fifty-minute intervals until the radiologist completed the self-allotted time for the study. Gaps existed between fifty-minute intervals as some recording sessions had members of the study start a new fifty-minute recording after the completion of preceding recordings. One recording session was done with continuous fifty-minute recordings ignoring the need for the manual starting of successive recordings. No restrictions were placed upon the radiologists during the recording session. Freedom of movement, time spent reading, and interruptions were left to the discretion of the radiologist to achieve the least amount of simulation in their readings.

Four recording sessions took place with two attending radiologists partaking in two sessions each. The set of sessions for one radiologist lasted four and seven hours while the set of sessions for the other radiologist lasted six hours each. Along with not

limiting the actions of the radiologists, the study did not impose an image selection to be read from. Thus, the images being read were not predetermined by the study and were genuine, real-time cases. With the sensitive nature of the images and readings, the study did not use the camera to capture data outside face recordings. Additionally, the gaze captured by the eye tracker was overlayed and calibrated to a screen on the lab monitor mimicking the dimensions of the hospital imaging monitor without capturing the contents of the hospital monitor. The data recording was stored and processed with a third-party software application, iMotions.

Data Acquisition

Upon completion of a recording session, the recordings of the camera and the gaze tracker were available for processing within the iMotions application. The software natively supports facial recognition and expression analysis from camera recordings as well as the measurement of gaze metrics from eye tracking data. However, the study extracted a subset of the metrics available from the software's facial and gaze analysis. The chosen metrics, deemed applicable due to interesting exploration or their previously observed effects on fatigue, fell into the two categories of their respective pieces of equipment.

Expression metrics: anger, contempt, disgust, fear, joy, sadness, surprise, engagement, and valence. The expression metrics were drawn from iMotions' facial recognition analysis, using the Affectiva Affdex toolkit, of the camera's recordings. The emotion variables are represented as intensity measurements where the larger the value of an emotion, the more intense the facial expression displaying that emotion and oppositely for the lower a value is. The variables themselves are drawn from the calculated

correlation between certain facial landmarks. The study pursued the use of facial expressions as they have been shown as determinants of fatigue in areas such as driving (Kaplan et al., 2015). Engagement is represented as a value between 0 and 100 where the higher the value, the more expressive the expression. More specifically, engagement measures the weighted sum of several facial movements. While not explicitly observed in other fatigue related studies, engagement is an interesting variable for the study as it serves as a parallel to the expression metrics. As the emotion variables themselves are drawn from facial landmarks, engagement being a higher-level composite value of facial landmarks can corroborate for or against observations drawn from the emotions. Valence is represented as a value between -100 and 100 where the more negative the value, the more negative the connotation of facial features while the more positive the value, the more positive the connotation. Again, valence is calculated by certain facial movements that contribute either adversely or positively towards the metric. As with engagement, valence is uncommon in similar fatigue studies but has the potential to serve as supporting evidence.

Gaze metrics: left and right pupil diameter, area covered, gaze velocity, gaze acceleration, fixation duration, saccade amplitude, saccade peak velocity, and saccade peak acceleration. The gaze metrics were drawn from iMotions' gaze analysis of the eye tracker recordings. The pupil diameters were calculated from the Tobii eye tracking software. Pupil diameter has been observed as a substantial factor in fatigue detection (Li et al., 2020; Yamada & Kobayashi, 2018). Area covered was calculated by multiplying the standard deviation of interpolated x and y gaze coordinates given by internal iMotions gaze analysis. Studies of fatigue in radiology have observed a decrease in search pattern

efficacy as fatigue progresses (Hanna et al., 2018). The decrease in search efficacy suggests an increase in the observed area of an image, a suggestion the study found merit in and a cause for exploration. Gaze velocity and acceleration are provided by iMotions' internal gaze analysis. Gaze velocity has been used in proposed fatigue detection models while gaze acceleration is more uncommon but was selected for its use to correlate with gaze velocity (Li F. et al., 2019). Fixation duration was also calculated by iMotions' internal gaze analysis. Fixations are represented as clusters of similarly positioned gaze points by iMotions, the duration consequently being the amount of time, in milliseconds, the gaze is within a cluster's area. Fixation duration as an oculomotor metric has been investigated for its use in fatigue detection (Naeeri et al., 2021; Yamada & Kobayashi, 2018). Fixations have been observed in radiology-related fatigue studies where the number of fixations during a reading increased the more fatigued the radiologist (Hanna et al., 2018). The duration of fixations serves as an interesting metric to be observed in comparison with recorded effects of similar metrics such as the number of fixations. Saccade amplitude, peak velocity, and peak acceleration are all drawn from the internal iMotions gaze analysis. Saccades are represented, in iMotions, as rapid minute movements of the gaze between fixation points. Saccade amplitude is the total distance of the movement between fixations. All three saccade metrics are heavily researched in the area of fatigue and have been shown to suffer acute effects from fatigue (Di et al., 2012; Li F. et al., 2019; Naeeri et al., 2021; Yamada & Kobayashi, 2018).

Data Preprocessing

Before analysis, the raw data from iMotions, with the selected metrics, was exported into a comma separated file format. The exported data was represented as a

multivariate time series where the selected metrics from both the camera and eye tracker were aligned with timestamps ordered by iMotions. Given a timestamp in milliseconds, the data would contain the recorded measurements of the aforementioned metrics if any existed at that time. Due to the unrestrained nature of the recording sessions, sparsity and noise are apparent in the data. However, Wang et al. (2012) studied the clustering of multivariate time series data through the use of feature extraction. Other studies have also reviewed the potential for feature-based representations of multivariate time series data (Fulcher, 2017). This study took inspiration from the approaches and transformed the exported data to allow for the extraction of global features to potentially assuage the nature of the data.

Particularly, the data was arranged into buckets. The buckets would be an interval of time, defined by the study, that covered from the start of one recording session until its end. Although the recordings for a session were done in fifty-minute periods and contained gaps from the end of one period to the start of the next, the buckets were ordered sequentially regardless of gap time. For instance, a bucket time of fifteen-minutes would result in sixteen buckets from a four-hour recording session containing four fifty-minute recordings (three full-length fifteen-minute buckets and one five-minute long bucket for each fifty-minute recording). The buckets, ordered sequentially starting from one, represented classes. In the fifteen-minute bucket example, the first class represents the first active fifteen minutes of the recording session, the second class the second fifteen minutes, and so on. After the data was arranged into buckets, global features from the multivariate time series were drawn at another time-based granularity. Allowing another example, a bucket time of fifteen-minutes with a feature length of three minutes

would result in five feature vectors belonging to that bucket. Specifically, the feature length of three minutes denotes the aggregation of datapoints into global features such as mean anger, mean saccade amplitude, and saccade peak velocity variance among others. The bucketing and feature extraction of the data allowed feature vectors and classes to be created from the original multivariate time series. Additionally, several z-score standardization techniques were used on the data, the results of which will be discussed in a later section.

CHAPTER 4

ANALYSIS METHODS

The goal of the data analysis was to explore the possibility of using machine learning classifiers in the detection of temporal change with regards to fatigue. To that end, the study experimented with the use of support vector machine (SVM) classifiers. SVMs are an optimal margin classifier where a decision function is learned from training data that computes a relation between datapoints and a separating hyperplane between different classes of data. The separating hyperplane maximizes the margin, the smallest distance between the hyperplane and the closest datapoints (Boser et al., 1992). In the two-class, linearly separable example illustrated in Figure 1, SVM finds the optimal decision boundary that separates the data space into two areas, one for each class.

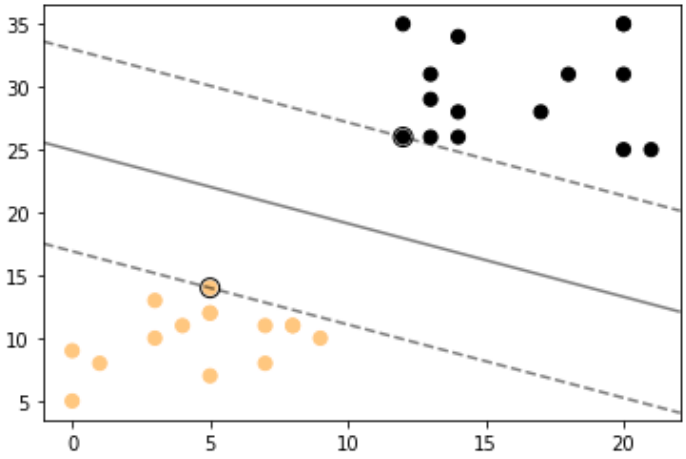

Figure 1. Visualization of SVM hyperplane construction.

The solid line represents the decision boundary or hyperplane while the dashed lines represent the support vectors. The support vectors contain the datapoints from which the optimal hyperplane can be found as the distance between the support vectors and the hyperplane is the maximized margin (Cortes & Vapnik, 1995). Although an SVM classifier looks to maximize the margin, a solution can also be gained through the

equivalent problem of minimizing the norm of the feature weights, w (Boser et al., 1992). However, complexity is added to the SVM objective when non-linearly separable or non-separable data is introduced. Non-separable data refers to the situation where datapoints from one class exist in the opposite class's area. In a linear case as well as others, such a situation can be solved by introducing a slack variable, ζ, that allows datapoints in opposing areas to remain 'valid' so long as they do not exceed the distance from their correct areas allowed to them by the slack variable (Cortes & Vapnik, 1995). The method is known as soft-margin SVM, where any datapoint exceeding the slack distance is considered an error and the tuning of such error acceptance is done through a hyperparameter, *C.* A large *C* has the SVM search for a margin that misclassifies the smallest amount of datapoints possible whereas a small *C* has the SVM search for a maximum margin regardless of the number of misclassifications incurred (Pedregosa et al., 2011). However, non-separability can occur in a case of non-linear separability as well. Non-linearly separable data refers to the inability to construct a linear separating hyperplane between data classes. In such cases, the kernel trick technique is adopted within the SVM objective function. The kernel trick maps datapoints that are non-linearly separable in their native dimension, to higher dimensions where they can potentially become linearly separable (Cortes & Vapnik, 1995). Thus, the margin classification of SVMs paired with their documented capabilities in complex dataspaces offered an interesting potential use for this study.

For implementation, the Python library Scikit-learn and its SVM derivation was used. Named support vector classifier in the library, Scikit-learn provides the following formulation of an SVM classifier (Pedregosa et al., 2011):

$$\begin{aligned} &\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta_i \\ \text{subject to} \quad &y_i(w^T \phi(x_i) + b) \geq 1 - \zeta_i, \\ &\zeta_i \geq 0, i = 1, \ldots, n \end{aligned} \tag{1}$$

Equation 1 represents the feature weight vector minimization problem. However, the impact of the slack variable and error acceptance can be observed in the second term of the minimization function as well as the function bounds. Although Equation 1 is given by the library, an equivalent formulation is used in the library's implementation that allows for the kernel trick (Pedregosa et al., 2011):

$$\begin{aligned} &\min_{\alpha} \frac{1}{2} \alpha^T Q \alpha - e^T \alpha \\ \text{subject to} \quad &y^T \alpha = 0 \\ &0 \leq \alpha_i \leq C, i = 1, \ldots, n \end{aligned} \tag{2}$$

Equation 2 removes the feature weight vector for the dual coefficients, α, and the matrix, *Q*. *Q* contains the kernel *K* which in turn contains the kernel function φ. The kernel function is responsible for the mapping of feature vectors, datapoints, into higher dimensions. The library offers several kernel functions to choose from, but the study primarily focused on the linear kernel with some testing with the radial basis function kernel.

Using the SVM implementation provided by the Scikit-learn library, the study conducted three major experiments on the pre-processed recording data. First, the study looked to observe the temporal patterns of the bucketed data. Given a time granularity for the classes of data, described as bucketing in the data pre-processing section, the data would be used to train an SVM for every unique one-versus-one class scenario. The SVM parameters were kept constant through each scenario, those mainly being the linear kernel type and an error acceptance value, *C*, of 1. After a one-versus-one SVM was

trained, the decision function and accuracy of the fitted SVM were evaluated against the training data. In particular, the trained SVMs were not used with the goal of creating generalizable classifiers, but rather with the goal of observing the attributes of the fitted model.

The decision function is given in the following equation by the Scikit-learn library:

$$\sum_{i \in SV} y_i \alpha_i K(x_i, x) + b \tag{3}$$

The decision functions of the trained SVMs were evaluated against the training data to achieve a distance-based perspective of the margin constructed by the models. To corroborate the perspective observed from the decision function, the accuracy of the trained SVMs was also evaluated. The accuracy allowed a similarity-based perspective of the constructed margin. Altogether, the accuracy and decision function of the trained SVMs formed the basis of the experimental temporal analysis methodology.

The second major experiment maintained the one-versus-one scenarios of the first experiment but also adopted a leave-one-out approach. The one-versus-one scenarios of the second experiment only concerned the first class against every other class but were repeated for every removal of a single feature. The feature vectors used to train the SVM models in the first experiment contained *n* features. In the second experiment, one feature would be removed at a time and a full set of SVMs were trained with *n-1* sized feature vectors for each class one-versus-*x* for all *x* not one cases until all features had been removed once. The decision functions and accuracies of the resulting SVMs were again evaluated and aggregated according to what feature was missing.

The third experiment continued with the results of the second and calculated the difference between the decision function results and accuracies found by the general class one-versus-rest SVMs and each leave-one SVM. The differences were then ranked in descending order allowing an interpretation of the effects different features had on the temporal pattern of the data. With the third experiment completed, the results were analyzed and constituted a proof of concept for the experimental methods outlined in this paper.

CHAPTER 5

RESULTS

Experiment One: Temporal Trends

The results of experiment one, the observation of temporal patterns, are shown in the following figures:

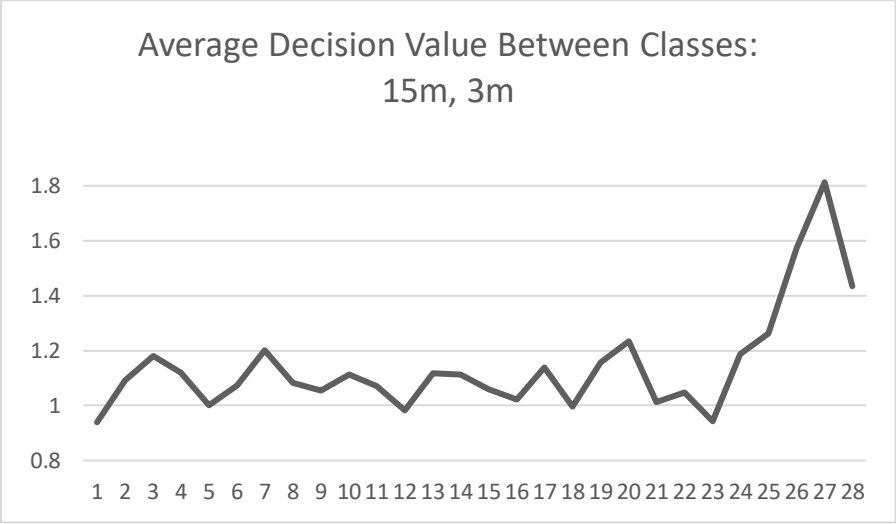

Figure 2. Progression of average decision values.

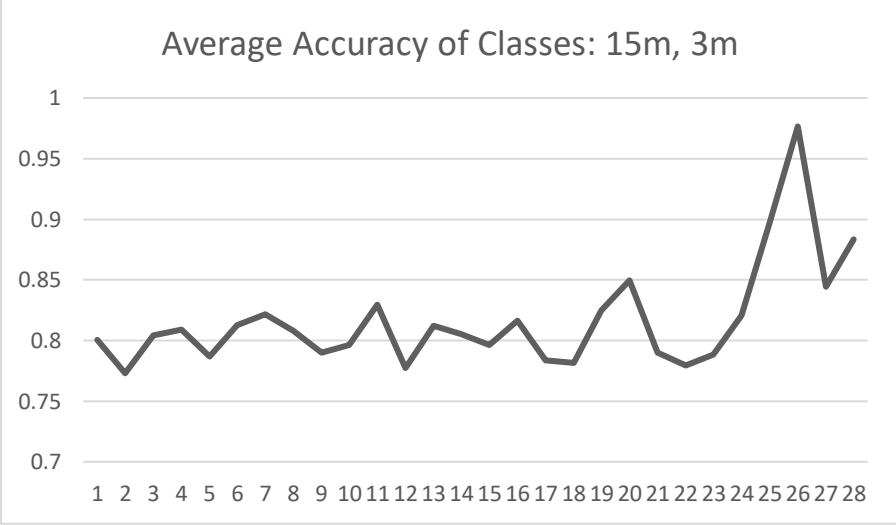

Figure 3. Progression of average accuracy.

Figures 2 and 3 show the class-based average of their respective metrics for a bucketing interval of fifteen minutes and a feature length of three minutes. The figures were derived from the following heatmaps respectively:

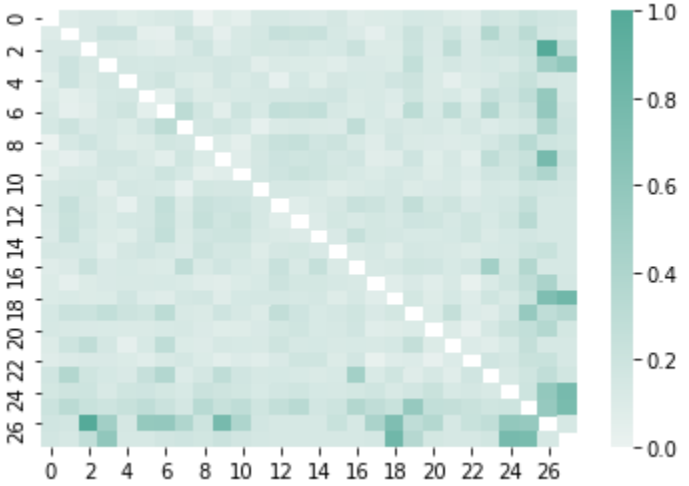

Figure 4. Heatmap of individual average decision values.

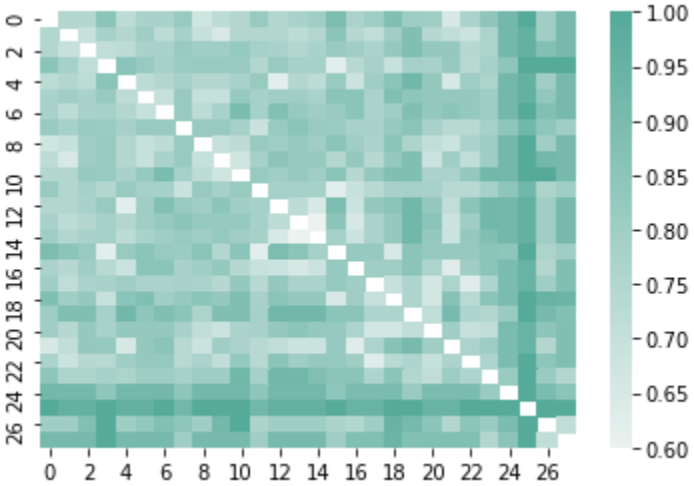

Figure 5. Heatmap of individual accuracies.

Figures 4 and 5 display the results of individual SVMs for their respective metrics; Figure 4 contains averaged SVM decision value results while Figure 5 contains the accuracy of SVMs. Along with decision value and accuracy metrics, the two being

the primary metrics of the experiment, a distance metric was also observed in experiment one. The figures for the distance metric are provided below.

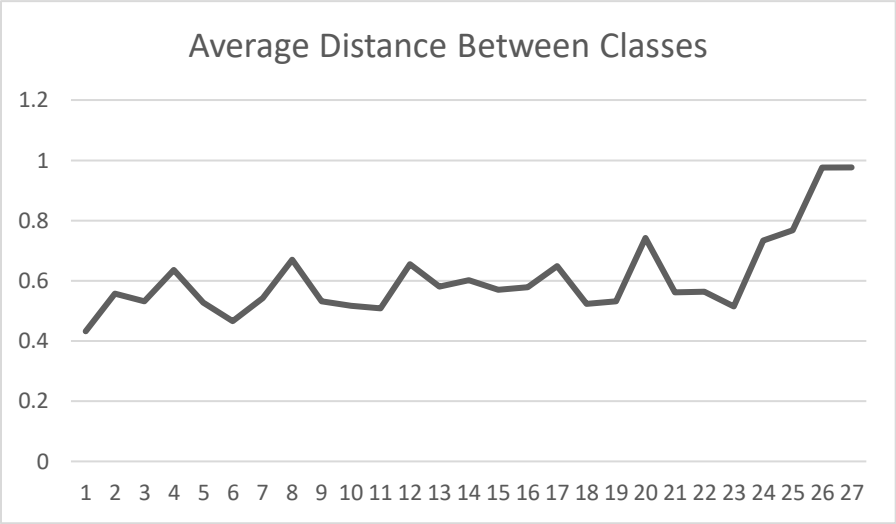

Figure 6. Progression of average distance.

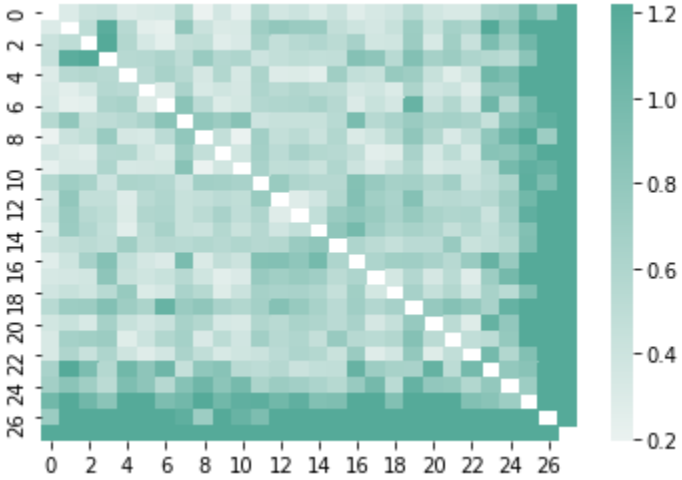

Figure 7. Heatmap of individual distances.

Figures 6 and 7 show the distance between training data and their respective hyperplane. The distance metric was drawn from the following equation (Boser et al., 1992):

$$distance(x) = \frac{|D(x)|}{\|w\|} \tag{4}$$

Additionally, the data used to produce the preceding figures had any missing values for the various sensor readings replaced with zero and standardized using the z-score measurement. The standardization took place before any bucketing into time classes or aggregation into feature vectors. Explaining Figure 3, the point for class 1 represents the averaged accuracy of linear SVMs trained on class one versus all other classes independently and scored against the training data. Figure 2 follows a similar pattern but is instead the average of the absolute average decision function results for each SVM trained in the one-versus-one scenarios. Different standardization techniques using z-score were also tested during the first experiment. One technique waited to standardize the data until the data aggregation step, meaning the data was only standardized within the feature vectors themselves. Additionally, the technique did not replace any missing values in the original dataset with zeroes until feature vector standardization. Another technique followed the similar pattern of standardizing feature vectors independently but allowed the emplacement of zeroes within the overall data before standardization. The effects of these techniques compared to the first are shown in the following figure.

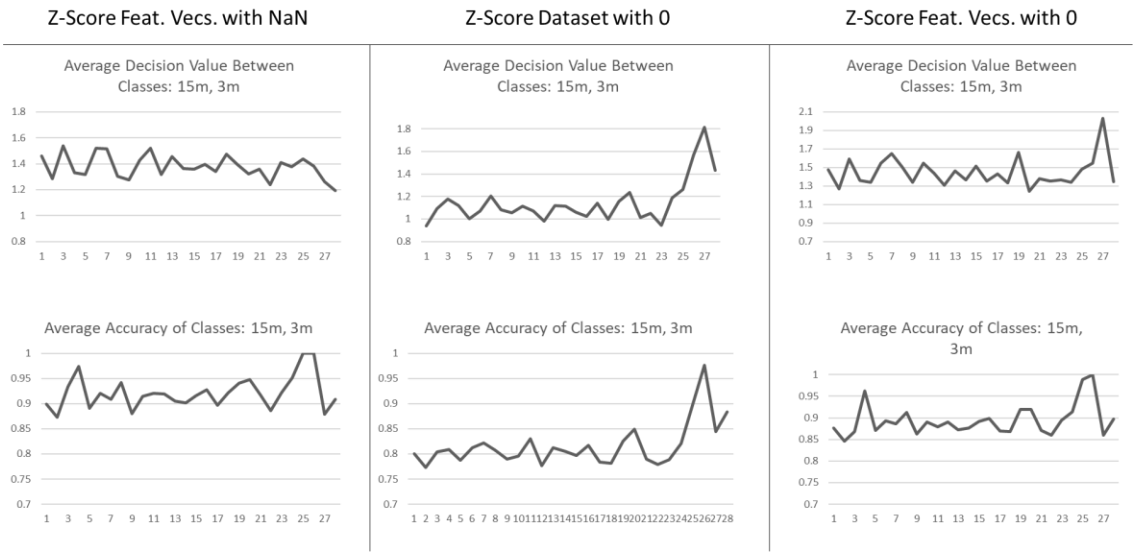

Figure 8. Comparison of standardization techniques.

Figure 8 highlights the effects of the different standardization techniques, but the study favored the use of the first technique, middle in the figure, and all ensuing results are shown with its adoption. However, different time granularities for classes were also tested. One granularity smaller and one larger than fifteen minutes were considered, the results follow:

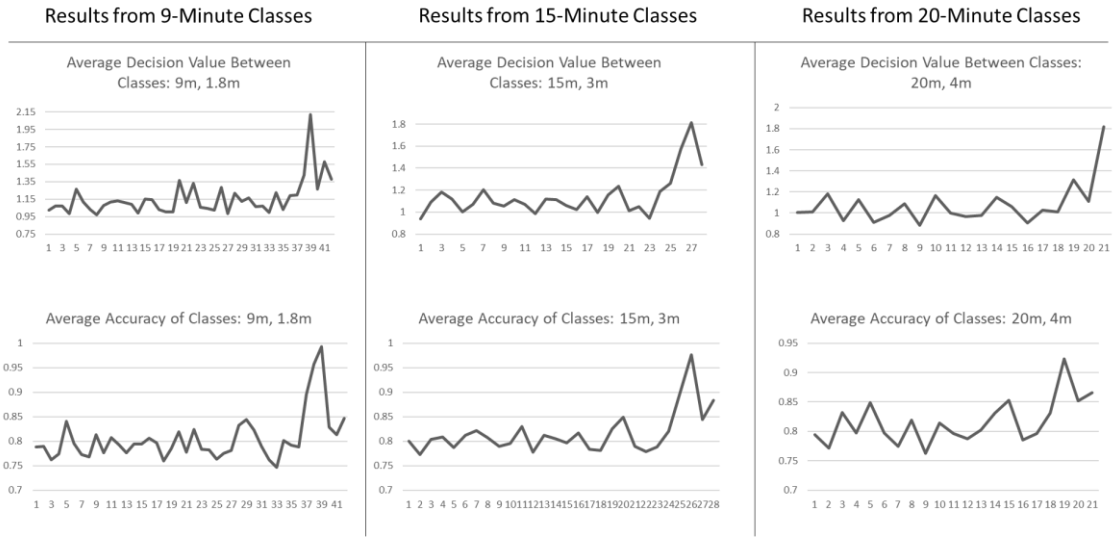

Figure 9. Comparison of time classes.

Just as with the first standardization technique adoption, the study favored the use of fifteen-minute classes going into the subsequent experiments. The second and third experiments have their results combined as the third is no more than the ranking of the second's outputs. Echoing earlier sentiments, the following graphs show results of data standardized using the first technique mentioned and grouped into fifteen-minute classes with three-minute feature vector lengths.

Experiment Two & Three: Feature Impacts on Temporal Trends

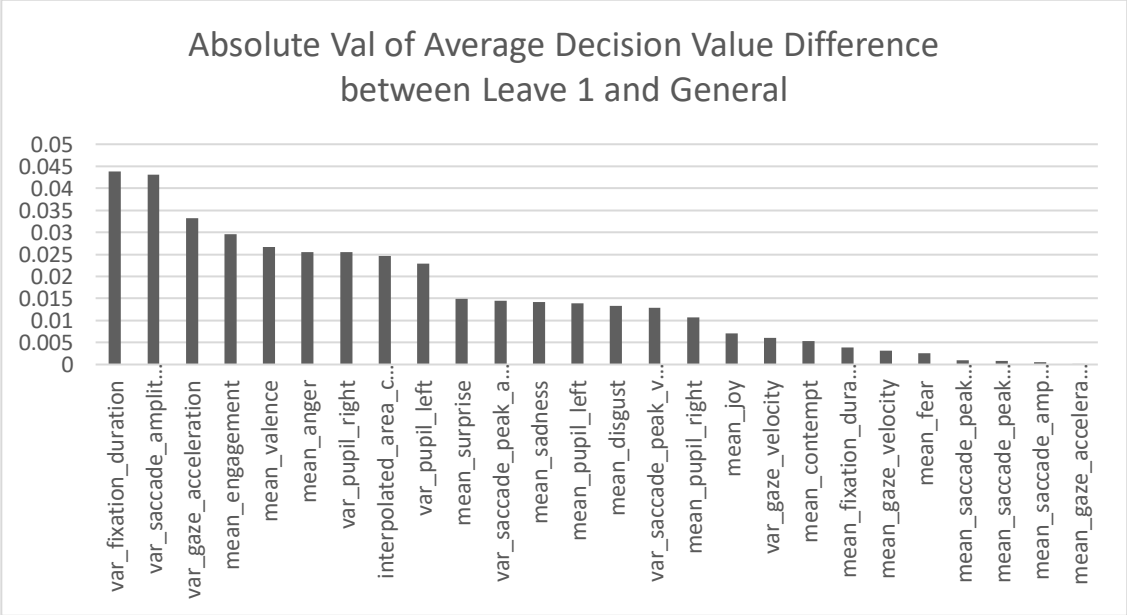

Figure 10. Individual feature impacts on average decision values.

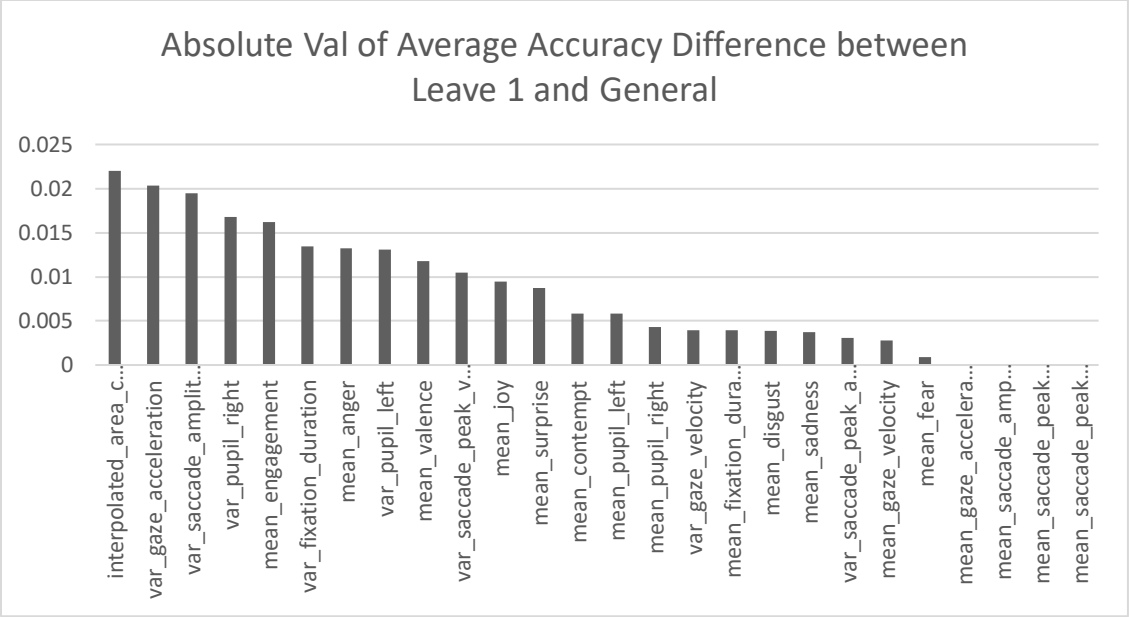

Figure 11. Individual feature impacts on average accuracy.

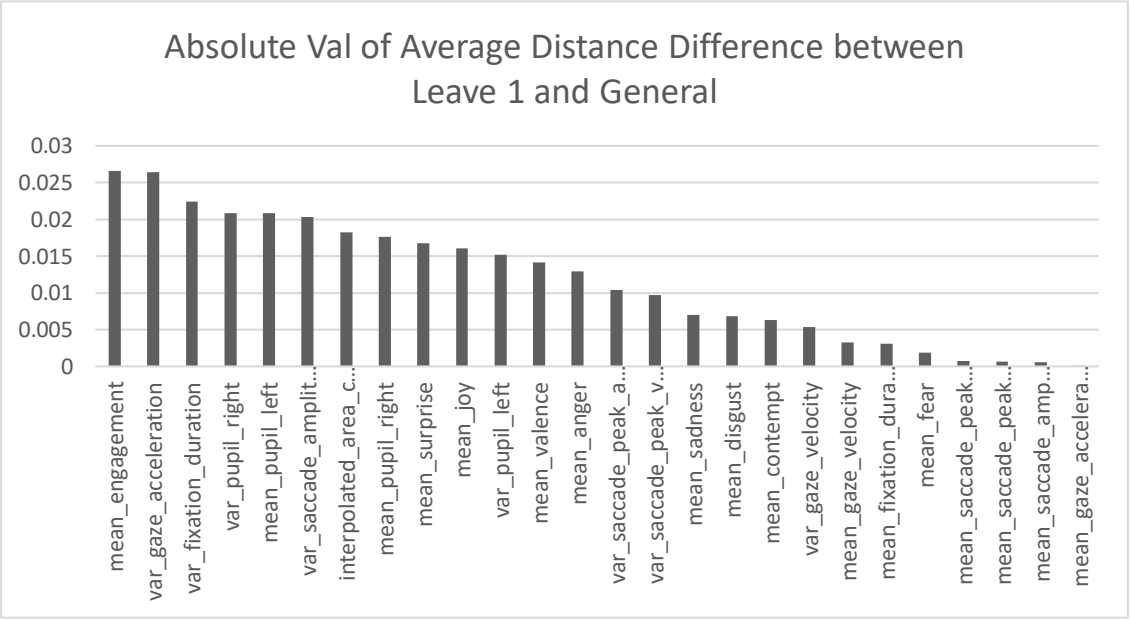

Figure 12. Individual feature impacts on average distance.

Figures 10, 11, and 12 show the rankings of the differences between the metrics given by the overall one-versus-rest SVM and the values given by the one-versus-rest SVMs missing their respective features. For instance, Figure 10 shows the removal of the variance of fixation duration feature had the largest effect on the average decision values

produced by trained SVMs. Parallelly, Figure 11 shows the removal of the interpolated area covered feature had the largest effect on the average accuracy produced by trained SVMs. The results illustrate the effects of SVM classifiers on observing trends in temporal data. Discussing the interpretation of the results will hopefully offer insight into the usefulness of the experimental methods used in this study.

CHAPTER 6

DISCUSSION

Interpretation Context

Fatigue research in radiology is not uncommon nor is the use of machine learning techniques in fatigue detection a rarity. However, the data used in this analysis represents a degree of novelty as the studies previously discussed have used data gathered in relatively controlled environments. To the best of the author's knowledge this is the first-time full shift gaze analysis and emotional data were recorded in live radiology readings. Furthermore, the use of an SVM classifier to observe temporal trends below the abstraction of fatigue and its occurrence in radiology readings is another relative novelty. The experiments done in this study showcase a potential procedure to be used in such an area. The results of which, as to be discussed here, carry interpretations that suggest a legitimacy of the novel perspectives of classifiers offered by this research.

Given the data gathered by the study, could an inflection point be found where the data could be split into pre-fatigue and post-fatigue areas? Could the data help investigate the validity of H1? For such an investigation a means of deciphering differences in the data at-hand was required. However, the data was a multivariate time series, any difference had to be made with regards to time. Linear SVM classifiers were proposed as an avenue to decipher such temporal differences. SVMs look to create an optimal margin between different classes of data; it follows that the margin and other attributes of a trained SVM model could be viewed as differentiators. Specifically, if a margin can be formed between two classes that is smaller or larger than a margin between two other classes, then an interpretation of a difference between the classes exists. However, before

attempting to gain any interpretation from the margin or other attributes of an SVM, the data collected by the study had to be pre-processed to allow for such an interpretation to have meaning in a temporal setting.

The multivariate time series' of the study were grouped into classes representing a time granularity. Each recording session consisted of multiple fifty-minute recordings from which the multivariate time series' were extracted. Any time granularity chosen to represent classes for the SVM thus acted as a filter for the temporal trends of the data. A granularity too high would effectively filter out the details of any temporal trend while a granularity too small would filter in an exhaustive amount of details. After testing several granularities, fifteen minutes were found to offer the relatively best perspective of temporal trends. Within each fifteen-minute interval, the data was aggregated into composite metrics at another time-based granularity. Again, other values for the length of the composite metric were tested before settling on a three-minute interval. The composite metrics chosen were as follows: the mean of anger, contempt, disgust, fear, joy, sadness, surprise, engagement, valence, left pupil diameter, right pupil diameter, gaze velocity, gaze acceleration, fixation duration, saccade amplitude, saccade peak velocity, and saccade peak acceleration, the variance of left pupil diameter, right pupil diameter, gaze velocity, gaze acceleration, fixation duration, saccade amplitude, saccade peak velocity, and saccade peak acceleration and lastly the interpolated area covered. The composite metrics were taken for every three-minute interval within a fifteen-minute class effectively creating five feature vectors for each instance of a fifteen-minute class. With data from four recording sessions, the longest of which containing seven fifty-minute recordings, the classes ranged from label 1 to 28 (a fifty-minute recording

contains three full-length fifteen-minute classes followed by one class constituting the final five minutes). To observe the difference between classes, an SVM was trained for every unique combination of classes. From these trained SVMs the results shown in Figures 2 through 7 were obtained.

Temporal Trend Interpretations

Figure 2 shows the trend of average decision values across the different classes which themselves represent the passage of time. To illustrate a finer granularity of the trend in Figure 2, the heatmap in Figure 4 is provided. Figure 4 shows the average decision values for each case of unique class combinations. Figure 2 is no more than the column-wise averaging of the values in Figure 4. Each cell in Figure 4 is the averaged absolute sum of the decision function in Equation 3 computed against the training data used to train the particular SVM. The decision values in both Figure 2 and 4 can be interpreted as a form of distance from the hyperplane constructed by a linear SVM. An average decision value closer to zero can be seen as the training data existing with a higher closeness to their fitted hyperplane than training data whose averaged decision value is larger. Additionally, Equation 4 can used to relate decision values more closely to a distance measurement in a linear SVM.

Equation 4 contains the decision function D and the feature weight vector w, both of which are provided by the Scikit-learn library. The decision function in Equation 4 can be drawn from Equation 3 while the feature weight vector is a provided attribute of the library's SVM models. A heatmap using Equation 4 is given in Figure 7. Figure 7 shows the average distance of training samples from their hyperplanes using Equation 4. However, any value above twice the median has been masked with the highest degree of

heat. Following the same procedure for Figure 2, the graph in Figure 6 can be constructed to highlight the flow of average distances across the classes.

Figure 6 shows the progression of the column-wise averages from Figure 7, class 28 is left out as all distances of class 28 were above the double median threshold. Figures 6 and 7 illustrate the distance between training samples and their hyperplanes allowing an interpretation of the differences between time classes. The higher a classes average distance to the hyperplane, the further spread out the data for each class is from each other. The higher spread, the greater degree of difference between classes. The progression of this difference can be used to shed light on the flow of more abstract variables such as fatigue. Both Figures 2 and 6 show the progression of class difference using different metrics, decision values and distance respectively, yet a structure is common between the two. Further, Figures 4 and 7 display a more concise supporting view of the structure. Fluctuations exist across the sequence of classes but a positive trend culminating in a large, rapid growth can be observed. Remembering the time-based nature of the classes, it can be assumed that the structure in the figures relates to a temporal pattern. As time moves forward, it can be postulated that the data is undergoing an increasingly differentiating phenomenon. The figures would suggest this phenomenon affects the data in such a way that an inflection may exist that can allude to the separation of the data into areas containing different degrees of differentiation. To corroborate the structure and interpretation of Figures 2, 4, 6, and 7, the study evaluated the accuracy of the trained models as seen in Figure 5.

As with Figures 2 and 6, Figure 3 was constructed from the column-wise averaging of the values from Figure 5. However, in the case of viewing the flow of

accuracies in Figures 3 and 5, a different interpretation of relationships between classes can be reviewed. The accuracy evaluated is the amount of correctly predicted feature vectors given by feeding the training data into the trained SVM model. A higher accuracy denotes a decision function and hyperplane that can better separate the training data than a lower accuracy. In the case of observing temporal trends, a high accuracy is seen as two classes being dissimilar from each other whereas a low accuracy demonstrates similarity, or the inability of the SVM to construct a perfectly separating hyperplane given the allowed slack for the soft margin. In both cases, as with decision values and distance, accuracy can be used to evaluate the differentiation between time classes.

Observing Figures 4 and 7 against Figure 5 reveals the similar, generally positive structure of the data. While distance and decision values grew as time progressed, accuracy increased as well. The outer edges of the heatmap grow in severity compared to the growth preceding them. The increase in distance and decision values demarcate data-hyperplane separation, accuracy highlights the intermixing of data classes. The increase of accuracy suggests an increase in dissimilarity between later time classes. Yet despite the nuances in each perspective, both form a similar opinion of temporal trends. Therefore, in the case of this study, accuracy and distance contain a level of corresponding evidence for SVM-based temporal differentiation. The results from experiment one clearly offer a structure of data variables over time that suggests an abstract phenomenon. Thus, the evidence from the results can be used to support the claim of H1 in the existence of a classifier-based methodology that can detect temporal effects. However, to better view the effects of individual data variables on the observed

structure and support H1's claim of the methodology being applicable in multivariate space, the results from experiments two and three were gathered.

Feature Impact Interpretations

Figures 10 and 11 show the effects of different variables on their respective trends. However, only the unique combinations of class one with other classes were considered. Thus, the results represent the impact of variables on the relationships between the first fifteen-minute interval and every other interval. Figure 10 focuses on the difference of decision value trends. The figure suggests the variance in fixation duration and in saccade amplitude account for the two largest discrepancies in class one's relationships with other classes. Particularly, the absence of the variables results in a trend of decision values that differs the most from the trend with their inclusion. To view the rankings according to the accuracy trend, Figure 11 is provided. Figure 11 instead proposes that the interpolated area covered and variance in gaze acceleration led to the two largest discrepancies in class one's relationships with other classes. However, despite the opposing suggestions of the top two variables, an interestingly similar observation can be made across the two figures. Each figure places a higher number of gaze related metrics towards higher degrees of impact than facial related metrics. Some facial metrics such as mean anger and engagement are given similar levels of impact but the general favoring of gaze over facial variables remain. Viewing Figure 12 for the rankings according to distance trends also supports the observation.

The interpretation of this observation leads to a belief that gaze features are impacted by the abstract phenomenon from experiment one more heavily than facial features. However, this impact itself is miniscule as the differences shown in Figures 10,

11, and 12 are themselves tiny. Nonetheless, the relationship between the individual effects of different variables provides an interesting set of observations compounded with those from experiment one. The abstract phenomenon is manipulating the data across the flow of time with a positive trend. However, positive in the case of the metrics reviewed in experiment one corresponds to a greater difference or isolation of classes. The increase in isolation matches a hypothetical progression of fatigue; as time progresses, fatigue becomes more severe. A stronger involvement of fatigue should then impact the overall state of an individual adversely. Specifically, the state of a person with maximum fatigue should be extremely differentiable from the state exhibiting average fatigue. The severity of fatigue can be a possible explanation of the increase in the isolation of time classes observed in experiment one; a positive trend should exist as fatigue isolates the state of an individual further away from preceding states. However, fatigue itself is extremely hard to define; more confidently, the structure of results from experiment one suggests a manipulation by some abstract phenomenon that may be fatigue related. No claim is made that fatigue is the direct cause of the manipulation, but the claim is made that an increasingly differentiating, temporal manipulation can be observed using SVM classifiers.

The results shown and experiments discussed have been garnered from the complete concatenation of all data gathered. However, another application of the experiments was done on the data separated into their respective subjects. Two subjects participated in the data gathering and the results of their individual analyses can be found in the appendix. Regardless, the study focused on analyzing the full set of data and offers the individual subject results as an aside.

Methodology Interpretations

Returning to the question of finding an inflection point where the data can be split into pre-fatigue and post-fatigue areas with the results from the experiments evokes confidence. Confidence not in the concrete solution to the problem, but confidence in the methodologies of the study in serving as a steppingstone towards a solution. The results clearly show a structure of data, in multiple evaluation metrics, that follows a trend one would believe is manipulated by fatigue. Fatigue should increase over time and adversely affect the state of an individual. The results from experiment one do show a separation of the state of variables as time progresses. The results from experiment two and three offer insight as to the nature of those variables on their shared state across time. However, despite these promising observations, this study does not claim to have defined fatigue or the exact location of an inflection point that defines a state change of fatigue. The study offers the observations as justification for the use of SVM classifiers in the analysis of temporal trends and inflection points. Such a use, as seen on the data from the study, has potential in the realm of fatigue. The results produced by the study vindicate the methodology outlined, granting firm supporting evidence for H1. Further, the experiments were done without any direct coupling to the radiology setting outside of the data gathered therefore allowing the methodologies of this study to be applied to fatigue analysis in other areas.

CHAPTER 7

CONCLUDING REMARKS

Conclusion

The study set out to analyze the complex trend of fatigue against the claims of H1 a level lower than similar studies. Machine learning fatigue identifiers have been studied that can detect a state of fatigue. The state of fatigue in these studies consist of metrics from both gaze and facial features. However, the detection is static with regard to time; the state of an individual is either marked as fatigued or not fatigued. Delving a level lower than a binary classification of fatigue led to this study contemplating the identification of fatigue dynamically with time. The main question drawn from such contemplation was the location of an inflection point in the progression of time that could demarcate pre-fatigue and post-fatigue areas. To explore the question, a novel arrangement of data was proposed that could be used in a novel application of a contemporary machine learning classifier, SVM. Instead of viewing feature vectors as determinants of a static state of fatigue, feature vectors were viewed as malleable constructs existing in different frames of time. Using the nature of SVM classifiers to construct optimal margins, the manipulation of these time-anchored constructs could be observed. The observations of which could be extrapolated to the manipulation by an abstract phenomenon such as fatigue. The results of the study suggest such a manipulation on both the structure of the frames of time as well as the features within them. The study is confident with the use of SVM classifiers to evaluate temporal trends in a multivariate space and is hopeful the procedure outlined in the paper offers a potential boon in the further study of fatigue identification in radiology and elsewhere.

Future Work

The ability to use an SVM classifier to differentiate between temporal areas is a prospective method to decipher fatigue. Although machine learning methods have been used as identifiers of fatigue, the use of SVMs in this study are more suited for the analysis of temporal trends in the search for an inflection point. Although this study did not aim to directly find such an inflection point, the methods described in the experiments can be used as inspiration in the definite search of an inflection point. Upon finding the inflection point, a classifier or other machine learning models could be constructed and use the knowledge of inflection to help monitor the temporal progression of fatigue. Additionally, as the study gathered data in a real-world, unconstrained setting, the experimental procedure outlined can serve as an extendable baseline for conducting studies in similar settings.

Limitations

Several limitations exist throughout the results produced by the study. The data itself is noisy and sparse. Sparsity arises from the unconstrained nature of the study where recordings may be ongoing, but the subjects may leave the recording area or orient themselves away from the cameras. Additionally, noise is apparent in the data from its processed nature. The data is derived from third party software, iMotions and Tobii, whose accuracy cannot be completely guaranteed by the study. The facial expression and some gaze variables are produced by iMotions calculations and are not validated by the study. Further, Tobii requires gaze calibrations of subjects which again cannot have their accuracy guaranteed by the study. Gaze calibrations were taken at the beginning of

recording sessions and were not repeated for the rest of the session leading to another source of noise.

Parallelly, the data was also limited. Only two recording sessions from two radiologists each were gathered totaling to about twenty-three hours of recording. The data is then limited in its number of subjects and the amount of data obtained from each. Thus, the reliability of the results leaves much room for improvement. The results are not generalizable and better serve as a proof of concept for future work.

REFERENCES

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A Training Algorithm for Optimal Margin Classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 144–152. Pittsburgh, Pennsylvania, USA. doi:10.1145/130385.130401

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. doi:10.1007/BF00994018

Di Stasi, Renner, R., Catena, A., Cañas, J. J., Velichkovsky, B. M., & Pannasch, S. (2012). Towards a driver fatigue test based on the saccadic main sequence: A partial validation by subjective report data. *Transportation Research. Part C, Emerging Technologies*, *21*(1), 122–133. https://doi.org/10.1016/j.trc.2011.07.002

Du, G., Li, T., Li, C., Liu, P. X., & Li, D. (2021). Vision-Based Fatigue Driving Recognition Method Integrating Heart Rate and Facial Features. *IEEE Transactions on Intelligent Transportation Systems*, *22*(5), 3089–3100. doi:10.1109/TITS.2020.2979527

Fulcher, B. D. (2017). *Feature-based time-series analysis*. doi:10.48550/ARXIV.1709.08055

Hanna, T. N., Zygmont, M. E., Peterson, R., Theriot, D., Shekhani, H., Johnson, J. O., & Krupinski, E. A. (2018). The Effects of Fatigue From Overnight Shifts on Radiology Search Patterns and Diagnostic Performance. *Journal of the American College of Radiology : JACR*, *15*(12), 1709–1716. https://doi.org/10.1016/j.jacr.2017.12.019

Hooda, Joshi, V., & Shah, M. (2021). A comprehensive review of approaches to detect fatigue using machine learning techniques. *Chronic Diseases and Translational Medicine*. https://doi.org/10.1016/j.cdtm.2021.07.002

Kaplan, Guvensan, M. A., Yavuz, A. G., & Karalurt, Y. (2015). Driver Behavior Analysis for Safe Driving: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, *16*(6), 3017–3032. https://doi.org/10.1109/TITS.2015.2462084

Khan, Hussain, S., Xiaoming, S., & Yang, S. (2018). An Effective Framework for Driver Fatigue Recognition Based on Intelligent Facial Expressions Analysis. *IEEE Access*, *6*, 67459–67468. https://doi.org/10.1109/ACCESS.2018.2878601

Kocejko, Gorycki, T., Polinski, A., Bujnowski, A., Kaczmarek, M., Ruminski, J., Nowakowski, A., & Wtorek, J. (2019). Using Eye-tracking to get information on the skills acquisition by the radiology residents. *International Conference on Human System Interaction, HSI, 2019-*, 54–59. https://doi.org/10.1109/HSI47298.2019.8942597

Krupinski, E. A., Berbaum, K. S., Caldwell, R. T., Schartz, K. M., & Kim, J. (2010). Long radiology workdays reduce detection and accommodation accuracy. *Journal of the American College of Radiology : JACR*, *7*(9), 698–704. https://doi.org/10.1016/j.jacr.2010.03.004

Lévêque, Young, P., & Liu, H. (2021). Studying the gaze patterns of expert radiologists in screening mammography: A case study with breast test Wales. *European Signal Processing Conference, 2021-*, 1249–1253. https://doi.org/10.23919/Eusipco47968.2020.9287678

Li, F., Chen, C.-H., Xu, G., Khoo, L. P., & Liu, Y. (2019). Proactive mental fatigue detection of traffic control operators using bagged trees and gaze-bin analysis. *Advanced Engineering Informatics*, *42*, 100987. doi:10.1016/j.aei.2019.100987

Li, J., Li, H., Umer, W., Wang, H., Xing, X., Zhao, S., & Hou, J. (2020). Identification and classification of construction equipment operators' mental fatigue using wearable eye-tracking technology. *Automation in Construction*, *109*, 103000. doi:10.1016/j.autcon.2019.103000

Machado, M., Aresta, G., Leitão, P., Carvalho, A. S., Rodrigues, M., Ramos, I., … Campilho, A. (2018). Radiologists' Gaze Characterization During Lung Nodule Search in Thoracic CT. *2018 International Conference on Graphics and Interaction (ICGI)*, 1–7. doi:10.1109/ITCGI.2018.8602697

Mall, Brennan, P. C., & Mello-Thoms, C. (2018). Modeling visual search behavior of breast radiologists using a deep convolution neural network. *Journal of Medical Imaging (Bellingham, Wash.)*, *5*(3), 035502–035502. https://doi.org/10.1117/1.JMI.5.3.035502

Naeeri, S., Kang, Z., Mandal, S., & Kim, K. (2021). Multimodal Analysis of Eye Movements and Fatigue in a Simulated Glass Cockpit Environment. *Aerospace*, *8*(10). doi:10.3390/aerospace8100283

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. Retrieved from https://scikit-learn.org/stable/about.html#citing-scikit-learn

Vosshenrich, J., Brantner, P., Cyriac, J., Boll, D. T., Merkle, E. M., & Heye, T. (2021). Quantifying Radiology Resident Fatigue: Analysis of Preliminary Reports. *Radiology*, *298*(3), 632–639. doi:10.1148/radiol.2021203486

Xiaozhe Wang, Wirth, A., & Liang Wang. (2007). Structure-Based Statistical Features and Multivariate Time Series Clustering. *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, 351–360. https://doi.org/10.1109/ICDM.2007.103

Yamada, & Kobayashi, M. (2018). Detecting mental fatigue from eye-tracking data gathered while watching video: Evaluation in younger and older adults. *Artificial Intelligence in Medicine*, *91*, 39–48. https://doi.org/10.1016/j.artmed.2018.06.005

APPENDIX A

INDIVIDUAL SUBJECT RESULTS

Subject 1:

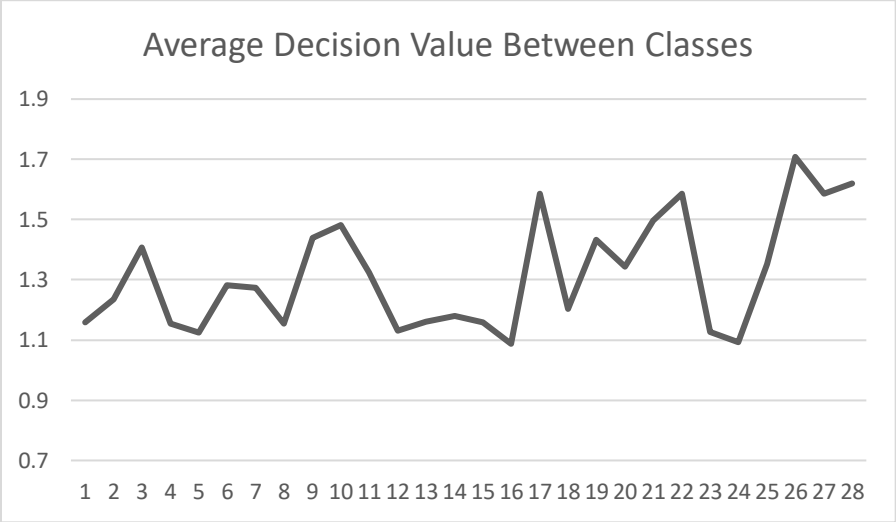

Figure A1. Progression of average decision values for subject 1.

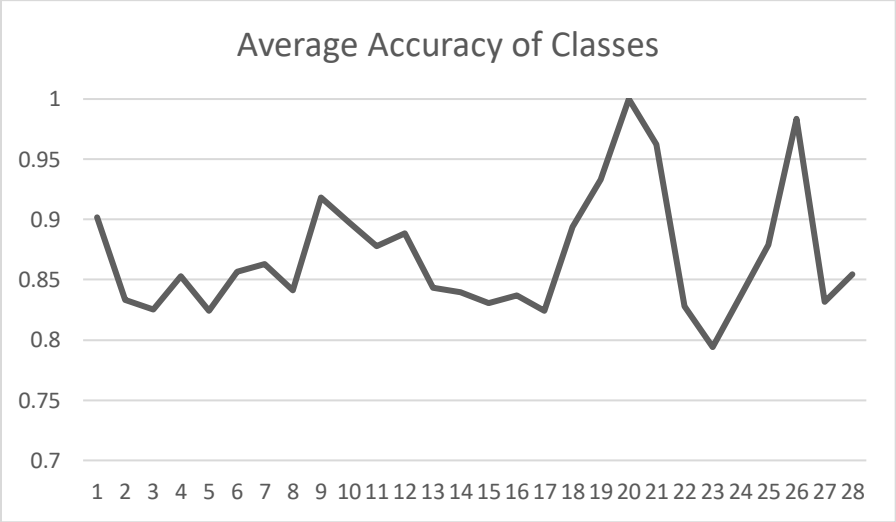

Figure A2. Progression of average accuracy for subject 1.

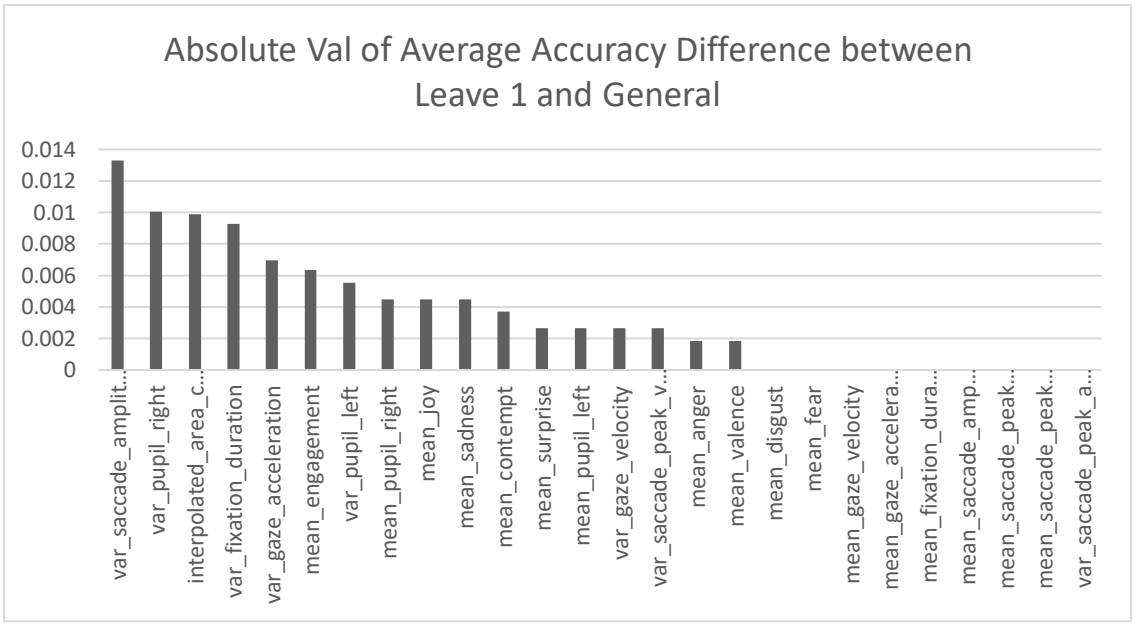

Figure A3. Individual feature impacts on average decision values for subject 1.

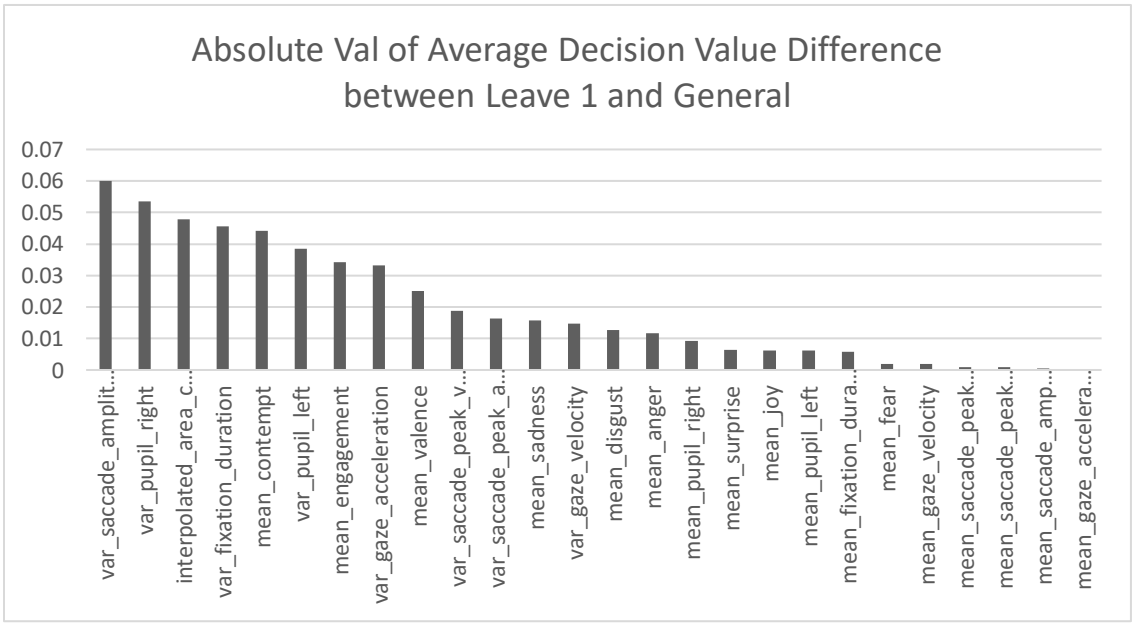

Figure A4. Individual feature impacts on average accuracy for subject 1.

Subject 2:

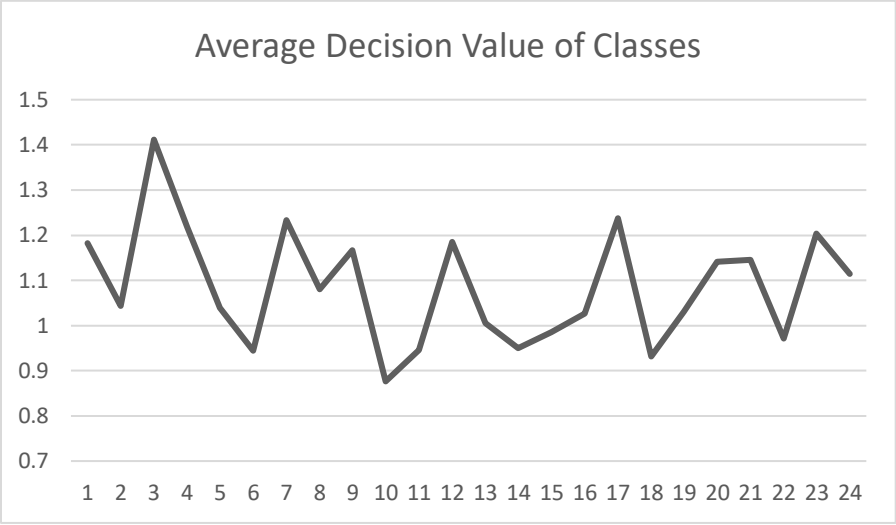

Figure A5. Progression of average decision values for subject 2.

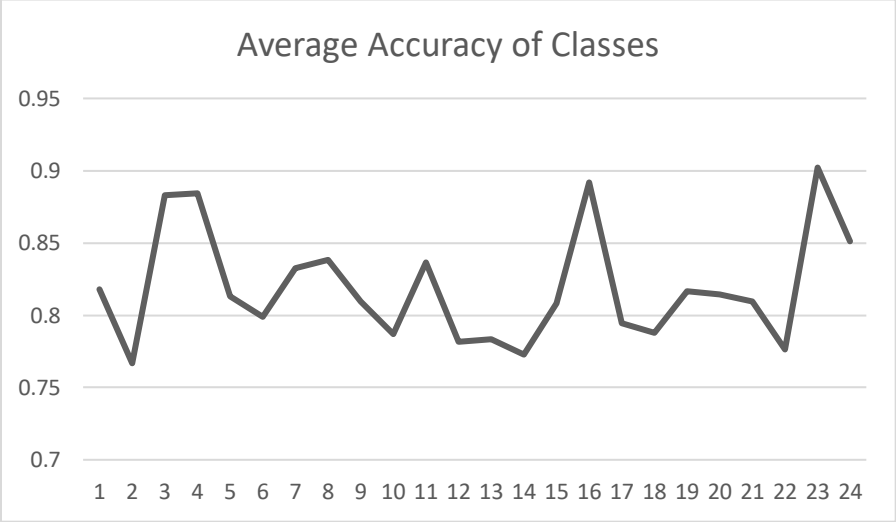

Figure A6. Progression of average accuracy for subject 2.

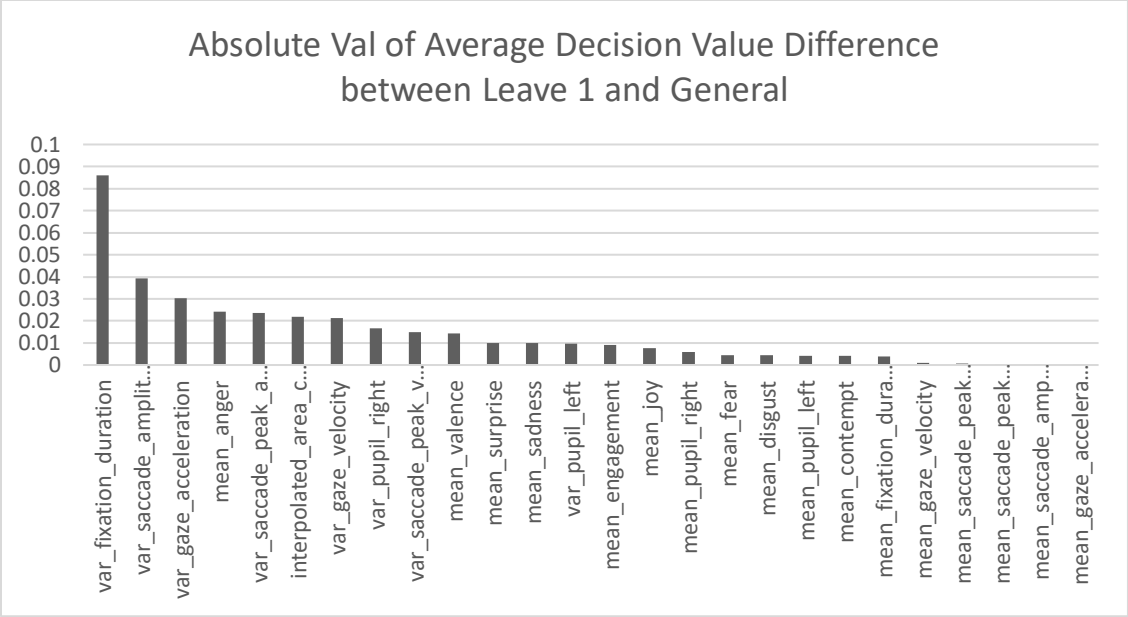

Figure A7. Individual feature impacts on average decision values for subject 2.

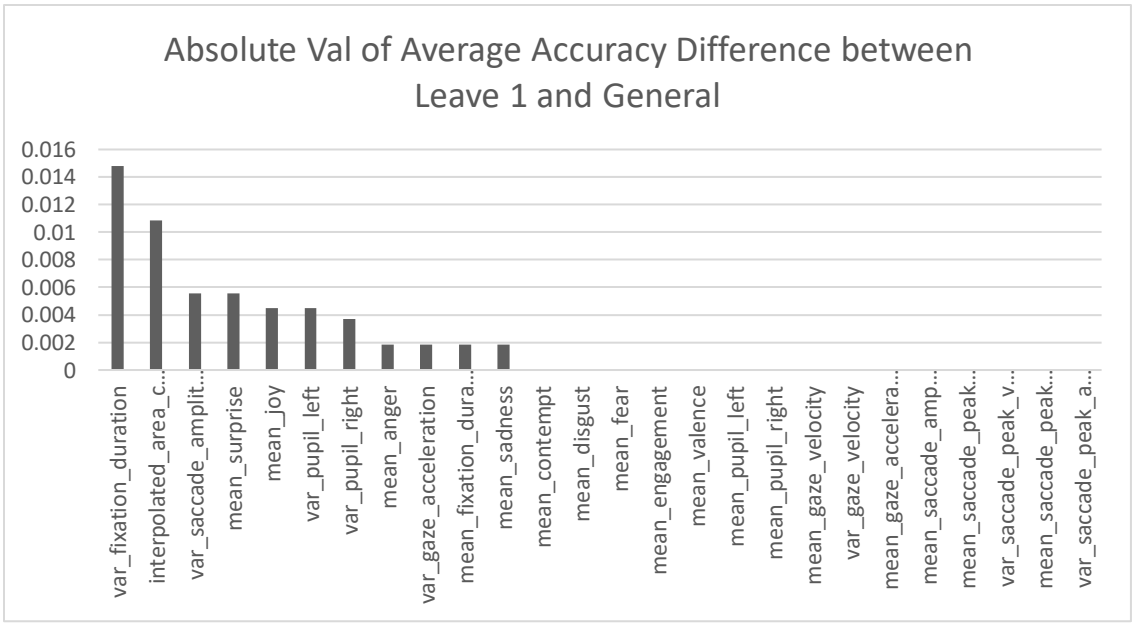

Figure A8. Individual feature impacts on average accuracy for subject 2.