Improving Driver Drowsiness Detection through Temporal, Contextual, and Hierarchical Modeling

By

Anthony Douglas McDonald

A dissertation submitted in partial fulfillment of the requirements of the degree of

Doctor of Philosophy

(Industrial Engineering)

At the

UNIVERSITY OF WISCONSIN-MADISON

2014

Date of final oral examination: 05/05/14

This dissertation is approved by the following members of the Final Oral Committee:

John D. Lee, Emerson Electric Quality & Productivity Professor, Industrial and Systems Engineering
Bilge Mutlu, Assistant Professor, Computer Science
Douglas Wiegmann, Associate Professor, Industrial and Systems Engineering
Matthew Rizzo, Edgar Reynolds Professor, Neurological Sciences, University of Nebraska Medical Center
Xiaojin Zhu, Associate Professor, Computer Science

Abstract

Drowsiness-related vehicle crashes are a persistent and substantial hazard on today's roadways. Drowsiness mitigation technology promises to reduce these crashes by detecting drowsiness and providing interventions to drivers. Mitigation technology relies on accurate detection algorithms to inspire driver trust and appropriate use of the technology. This dissertation investigates gaps in the current drowsiness detection literature and iteratively develops a series of temporal, contextual, and hierarchical models to address these gaps.

This dissertation uses data collected from a high fidelity driving simulator to predict drowsy-related lane departures. The three studies discussed in this dissertation investigate the effects of dynamic graphical modeling structures, road context integration, and hierarchical context integration on model detection performance. The investigation of dynamic graphical models included Hidden Markov Models, Hidden semi-Markov Models, and Conditional Random Fields. The study of road context integration investigated distributional parameters, Fourier transforms, and Symbolic Aggregate Approximation for generating road context from vehicle speed and acceleration data. The hierarchical context study investigated generation of both road-context and maneuver-level context from speed and acceleration data using Symbolic Aggregate Approximation time-series analysis.

The three studies showed a benefit of including temporal dependencies and maneuver-level context in drowsiness detection algorithms. Including both of these factors significantly reduced false positives generated by the algorithm relative to PERCLOS, a commonly applied algorithm in the drowsiness detection literature, and a steering-based algorithm that did not consider temporal or contextual factors. Maneuver-level context increased detection performance relative to both road type and a hierarchical combination of maneuver-level and road type contexts. State duration modeling undermined model performance and was not effective for drowsiness detection. Together these results provide an improved drowsiness detection model, highlight deficiencies in the current understanding of drowsy driving, and provide benchmarks for future predictive modeling analyses.

Acknowledgements

First and foremost I would like to thank my wife Christina. For all of her sacrifice and understanding she deserves this degree far more than I do. This dissertation is dedicated to her.

I would also like to thank my mother, Elizabeth, who taught me the values of persistence and kindness and what it means to be a good human being, my father, Mark, who created the footsteps I followed and provided a sounding board throughout every step of this process, my brother, Andy, sister-in-law, Stephanie, and my nephew, Luke, who showed me that Madison (and life) was not just for work, as well as my mother and father-in-law, Donna and Mike, who have truly been like a second set of parents for me.

In addition to my familial support, I am very lucky to have had an outstanding PhD advisor in John Lee. John's talent for inspiration and guidance is without equal. I cannot express the profundity of John's impact on my life and work, but it suffices to say that the lessons he taught me will follow me for the rest of my life.

I would also like to thank my committee members: Bilge Mutlu, Doug Wiegmann, Matt Rizzo, and Jerry Zhu who all contributed to this work with their comments and thought provoking questions as well as through their coursework and distinguished bodies of research. I am lucky to have their involvement.

I would certainly not be here without my extended family: Elizabeth Macdonald, Ellen McDonald, Sonya and Walter Baumgartner, Tom, Mia, and Jack Baumgartner, Ryan DeMar, and Mike and Theresa Rapa, all of whom provided significant amounts of support throughout this process and my education in general.

Thank you to the members of the CSL lab (Mai Lee, J.B., Shannon, Mahtab, Pat, Dee, Erin, Vindhya, Rashmi, Maddie, Elease, Kathy…) for your thoughtful comments, support, and friendship. I was very lucky to have you as colleagues and I wish you the best of luck (although you will not need it) on all of your endeavors.

Thank you to the team at the National Advanced Driving Simulator, specifically, Tim Brown, for reviewing papers and supporting this work and Chris Schwarz, who was an integral part of the conception of this dissertation and my understanding of the data.

Thank you to the National Highway Traffic Safety Administration for sponsoring this work.

I would like to also thank Missy Cummings, for hiring football players in her lab, introducing me to Human Factors, and for recommending me to John, and Carl Nehme for his role in my introduction to Human Factors and for showing me the rewards of research.

Finally I would be remiss if I did not acknowledge the love and support of my grandfather, John McDonald, who bought me my first science book, and my great grandparents, Thomas and Mary Renner, who showed me that engineering could be (and is) fun. Although they could not be here to see me finish this dissertation their hand in its creation was undoubtedly significant.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

In 2011, the National Highway Traffic Safety Administration (NHTSA, 2011) reported drowsiness contributes to approximately 83,000 crashes, 37,000 injuries, and 900 deaths each year—accounting for approximately 3% of all traffic-related fatalities. The 100-Car naturalistic driving study found that drowsy driving contributed to 22% to 24% of the crashes and near-crashes observed (Klauer, Dingus, & Neale, 2006). Other crash survey data illustrate that this problem is not unique to American drivers. Maycock (1997) found that drowsiness contributes to 10% of crashes reported by drivers in the United Kingdom. Sagberg (1999) found that drowsiness contributes to approximately 4% of crashes reported by Norwegian drivers and that 10% of male and 4% of female drivers in Norway have fallen asleep behind the wheel. The variance in these estimates reflects the difficulty associated with identifying drowsiness-related crashes. This difficulty is driven by the fact that drowsiness leaves no physical trace and is a subjective experience. These features suggest that the crash statistics and surveys may underestimate the true problem of drowsy driving.

Drowsy driving crashes can be considered as part of a larger class of impairment-related vehicle crashes. This larger class of crashes includes those caused by distraction, alcohol impairment, and both prescription and recreational drug use in addition to drowsiness. This class represents approximately 47% of all fatal crashes (NHTSA, 2012). The primary identifying characteristics of these crashes are that they are preventable and involve a breakdown in the driver's ability to cope with the on-road environment. Despite the fact that the root cause of these impairments differs, the approaches to reducing these crashes are quite similar. These approaches include education (Armstrong, Obst, Banks, & Smith, 2010; Regan, Lee, & Young, 2009; Sagberg, 1999), regulation , and mitigation technology (Balkin, Horrey, Graeber, Czeisler, & Dinges, 2011). Current mitigation technology consists of collecting data from the driver, vehicle, or environment, applying a classification algorithm to these data, and presenting the result of the classification algorithm to the driver (Balkin et al., 2011). The basic mitigation procedure accommodates considerable flexibility in measurement, algorithm, and feedback methods. The interaction between the

driver and the feedback from the mitigation system is a critical failure mode. This success of this type of human-automation interaction strongly depends on the driver's trust in the system and the systems reliability (Lee & See, 2004). The reliability of the system in this case is driven by the robustness of the behavioral measures and the classification algorithm's ability to detect impairment (sensitivity) while limiting false positive classifications (specificity). Thus the detection algorithm is a critical component of the system.

The importance of the detection algorithm for mitigation system success has prompted substantial research on drowsiness detection algorithms. The goals of this research typically center on introducing novel input measures (Dinges & Grace, 1998; Lal, Craig, Boord, Kirkup, & Nguyen, 2003), evaluating the use of machine learning algorithms that have been successful in other domains (Liang, Reyes, & Lee, 2007; Patel, Lal, Kavanagh, & Rossiter, 2011; G. Yang, Lin, & Bhattacharya, 2010; Yeo, Li, Shen, & Wilder-Smith, 2009), or introducing novel pre-processing steps to improve classification (Kutila et al., 2007; Sayed & Eskandarian, 2001). Although synthesis of these studies is difficult due to varying definitions of drowsiness and data sources, no study has reported perfect accuracy and all have non-zero false positive rates. Furthermore although aftermarket drowsiness mitigation systems exist (Halpert, 2012), they have not had a profound effect on the rate of drowsiness-related crashes or the percentage of drivers who engage in drowsy driving (National Sleep Foundation, 2009). The lack of accuracy and impact of current mitigation systems suggest a potential research gap.

Most current algorithms approach drowsiness detection as a supervised machine learning classification problem. Supervised machine learning is a process of training algorithms to detect patterns in data associated with known labels and then use the trained algorithms to assign labels to unseen data (Mitchell, 1997). Although this view has been wildly successful in many domains it has several limitations. Many of the algorithms used in classic supervised machine learning are naïve to the structure of the underlying phenomena they predict and are subject to inductive bias caused by assumptions about the underlying distribution of the data (Kotsiantis, Zaharakis, & Pintelas, 2007). Furthermore, many of these algorithms are designed to perform learning on static problems and perform poorly when applied to

temporal data. Thus approaching drowsiness detection as a supervised machine learning problem could result in poor performance due to a lack of insight into the context of drowsy driving and the temporal profile of drowsy driving.

The significance of context and time in driving can best be understood through theoretical models of driver behavior. The general purpose of these models is to explain driver behavior (and by extension patterns of driving data) in various driving situations (McRuer, Allen, Weir, & Klein, 1977). Three broad paradigms dominate the driver behavioral modeling literature: control theoretic approaches, behavioral adaptation approaches, and cognitive-architectures. Control theoretic approaches view the driver as a servomechanism that provides continuous input to the vehicle to maintain safety and execute maneuvers (McRuer et al., 1977; Weir & McRuer, 1970). Behavioral adaptation models consider driver behavior to be caused by a continuous effort to maintain a constant level of risk or task-difficulty (Fuller, 2005; Wilde, 1988). Cognitive-architectures model the driver as a hierarchical system of control, monitoring, and decision-making, which is driven by the execution of processes (Salvucci, Boer, & Liu, 2001; Salvucci, 2006). Although these perspectives differ, they are united in their conceptualization of driving as a temporal and context (i.e. road type and driving maneuver) dependent process. Furthermore, these models introduce the notion that driving context is hierarchical, i.e. driving behavior is subject to norms that operate on different time scales. This hierarchy, along with general temporal driving context, is a critical determinant of driving behavior and represents a potentially critical omission in current driver drowsiness detection algorithms.

In addition to its significance in driving behavior, time is a central component of drowsiness. Feelings of drowsiness are driven by the circadian rhythm and affected by time spent on vigilance tasks (I. D. Brown, 1994). During driving drowsiness accumulates as the length of the drive increases. Just as with driving behavior, static machine learning algorithms cannot capture these profiles. This is a significant limitation because dependencies within a profile could significantly reduce the variance in drowsiness prediction. For example, once drivers become drowsy they will likely remain in that state. Incorporating these dependencies into drowsiness detection algorithms necessitates a shift from static to

dynamic, temporal methods of supervised machine learning. A subset of these temporal machine learning methods, dynamic probabilistic graphical models (DGM), incorporates explicit models of temporal dependencies and context (Dietterich, 2002; Murphy, 2002). A model using an approach within this subset would fill the research gap produced by the reliance on static machine learning in current drowsiness detection algorithms and provide a structure that facilitates the inclusion of contextual information. However the size of this subset of methods is substantial and the choice of a particular method is data and context dependent, so the choice of DGM model structure drowsiness detection is an open research question.

Dynamic probabilistic graphical models (DGM) facilitate including both temporal and contextual effects into a drowsiness detection algorithm. Like temporal effects, contextual effects are relevant to both general driver behavior and driver drowsiness. For example drowsiness-related crashes frequently occur on straight roads with few environmental stimuli (Lal & Craig, 2001; MacLean, Davies, & Thiele, 2003). Incorporating these conditional likelihoods into a DGM has the potential to significantly increase model detection performance and reduce the rate of false positives; however the process is not straightforward. The initial step of this process requires a data based definition of context. One possible definition of context includes weather and road characteristics. Although this definition is conceptually simple, it relies on external data sources and cannot capture driver's responses to emergent disruptions such as traffic, or behaviors such as lane changes. Incorporating these emergent, contextual factors requires a behavioral based definition of context. Models of driving behavior suggest that this type of context could be derived from speed and acceleration data (Fuller, 2005; Salvucci, 2006). Incorporating these data into a DGM is difficult because the speed and acceleration elements of the driving context cannot be measured instantaneously rather they require aggregation over time. There are a variety of methods to perform this aggregation, or feature generation. As in the case of the graphical model structures, the most appropriate feature generation method is often data and problem specific. The choice of feature generation and the incorporation of these features into the model are open research questions.

Although incorporating context into drowsiness detection algorithms will likely improve detection performance it does not fully capture the intricacies of driver behavior. Driving behavior is fundamentally a hierarchical process consisting of navigation, maneuvers, and vehicle control (McRuer et al., 1977; Michon, 1986, 1989). Navigation of broad road contexts, such as highways and urban arterials, is composed of a sequence of maneuvers that are in turn sequences of vehicle control inputs and the resulting vehicle speeds and accelerations. Adding this contextual hierarchy to a graphical drowsiness detection algorithm will further reduce the variance in drowsiness prediction and will allow the algorithm to locate drowsiness consequence in a specific road and driving context. This increased level of representation facilitates the generation of new theories of driver behavior. Such theories will be beneficial as vehicles become more automated and require transitions in control from automation to drivers. Like the general incorporation of context, hierarchical contextual integration is a difficult process. This process will require development of a new feature generation technique that can capture the unique features of the contextual driving hierarchy.

Drowsiness-related vehicle crashes are a serious public safety concern. These crashes may be reduced by mitigation systems that rely on drowsiness detection algorithms. Currently detection algorithms have been marginally successful, but have not reduced the rate of drowsiness-related crashes. There are three critical gaps in the designs of these current algorithms that can be highlighted through an analysis of driver behavioral models. These three gaps involve a lack of consideration of the temporal, contextual, and hierarchical aspects of drowsy driving behavior. The goal of dissertation is to develop a new algorithm to detect drowsiness addresses these gaps through the use of a dynamic probabilistic graphical modeling framework, and a novel contextual feature generation and integration process. This new algorithm will improve both drowsiness detection performance and provide critical theoretical insights into driving behavior and drowsy driving.

1.1 Research questions

The goal of this dissertation is to develop a drowsiness detection model that considers the temporal, contextual, and hierarchical elements of drowsy driving. This approach is motived by the lack of consideration of these factors in current algorithms and their well-defined relationship with drowsiness and the frequency of drowsiness-related crashes. In pursuit of this goal, this dissertation will answer the following questions:

1. How can dynamic graphical modeling structures provide an accurate and realistic characterization of the temporal nature of drowsy driving?
2. What is the most appropriate method for generating contextual driving features and integrating these features into a dynamic graphical model for real-time drowsiness detection?
3. How can these contextual driving features be integrated into a hierarchy that represents the hierarchical nature of driving behavior and integrates into a dynamic graphical model?

1.2 Practical contributions

The development of a temporal, contextual, and hierarchical drowsiness detection algorithm promises to address several limits of current drowsiness detection algorithms. Furthermore, the inclusion of road context in a hierarchical framework will likely reduce variance in the drowsiness predictions. Given these two improvements, the practical contribution of this dissertation will be an improved drowsiness detection algorithm that can be integrated into both current and future drowsiness mitigation technology. The improved performance of this algorithm relative to previous work will improve the human-automation interaction between drivers and the mitigation system and potentially decrease the frequency of consequences of drowsy driving. In addition the probabilistic graphical structure of this model facilitates the integration of other impairments into its detection capabilities. By identifying and differentiating between these impairments the algorithm has the potential to significantly reduce impairment-related crashes and improve feedback systems.

1.3 Theoretical contributions

The goal of this dissertation is to produce a hierarchical, contextual, and temporal model of drowsy driving behavior. These three characteristics allow such a model to capture drowsy and alert driving behavior under specific environmental conditions. The environment is a central driver and determinate of all human behavior (Gibson, 1966). By incorporating the environment through driving context, the algorithm produced by this dissertation will better capture drowsy driving behavior, and provide a more thorough representation of drowsy driving behavior. Furthermore the structure of this model can be extended to similar driving behavioral scenarios involving distraction, alcohol impairment, or even general evaluations of driver performance across contexts. The assessment of driver-performance across contexts is critical to future automated vehicles that will require technologies capable of assessing the environment and determining an appropriate moment to handover control to a human driver. Therefore, the theoretical contribution of this dissertation is an enhanced theory of driver behavior that provides a link between driving context and driving behavior.

Chapter 2 Motivating a temporal, context-based algorithm for drowsiness detection

Driving impairment is a significant public safety issue that has triggered an extensive body of research on real-time mitigation systems. The goal of these systems is to efficiently diagnose safety critical levels of impairment and then adjust driver support systems, issue alerts, or generate feedback to the driver. The success of these systems is driven by the interplay between the system and the driver, which ultimately depends on the accurate detection of impairment and avoidance of false alarms (Balkin et al., 2011). The accuracy and false alarm rate of a detection algorithm effectively measure the algorithm's ability to identify patterns of data associated with impairment and differentiate between instances where these patterns are caused by context and those where the patterns are caused by impairment. Thus designing an algorithm within the context of drowsy driving is a promising approach to improving detection quality.

This dissertation discusses the development of a temporal, context-based, hierarchical drowsiness detection algorithm. As such, this chapter begins with a general introduction to the consequences of impaired driving, the factors that are known to influence the frequency of drowsiness-related crashes, and current technological interventions aimed at reducing the problem. This section is followed by a review of current drowsiness detection algorithms, which are a central component of current technological interventions. Following this review, current algorithms are compared to theoretical driver and human behavioral models and gaps in the current algorithm literature are defined. The gaps prompt a discussion of temporal machine learning concepts, specifically probabilistic graphical models and time-series feature generation. The chapter concludes by formulating a set of research questions that subsequent chapters will answer through an iterative approach to driver drowsiness detection algorithm design.

2.1 Impaired driving: consequences, influential factors, and safety interventions

Understanding the scope and structure of the problem of drowsy driving is critical to the successful design of a drowsiness detection algorithm. There are three primary sources of data on the frequency and severity of drowsiness-related crashes: crash databases, naturalistic driving data, and surveys. Crash

databases contain accident reports, filed by police officers, which specify the conditions surrounding a crash and contributing causes. A recent analysis of federally maintained data bases conducted by the National Highway Traffic Safety Administration suggested that drowsiness contributes to approximately 83,000 crashes, 37,000 injuries, and 900 deaths each year—accounting for approximately 3% of all traffic-related fatalities (2011). A similar study, which employed a database maintained by the state of North Carolina, found that drowsiness-related crashes typically involved a single vehicle departing is lane and running off the road. Furthermore these crashes often occurred at higher speeds during night-time highway driving, and often involved younger drivers (Pack et al., 1995). These findings support earlier work by Brown (1994) and Knipling and Wang (1994), who found that drowsiness crashes are more frequent among professional drivers. Although these data are valid they likely underestimate the frequency of fatalities and other consequence due to their reliance on police reports and the fact that drowsiness leaves no physical trace on drivers. Naturalistic driving data attempt to resolve these issues by collecting continuous video and kinematic data directly from the vehicle (Neale, Dingus, Klauer, Sudweeks, & Goodman, 2005). The 100-Car naturalistic driving study found that that drowsiness was a factor in 22% to 24% of all crashes and near-crashes, i.e. events triggered by high accelerations or close proximity to other vehicles (Klauer et al., 2006). These data suggest that crash database analyses may underestimate the true problem of drowsiness.

A large body of drowsiness research has used surveys. These surveys supplement naturalistic driving data and crash databases through larger sample sizes and analyses that identify the frequency of engagement in drowsy driving in addition to the frequency of consequences. These data illustrate that the problem of drowsy driving is not unique to American drivers. In a survey of drivers from the United Kingdom, Maycock (1997) found that drivers report drowsiness is a factor in 10% of crashes. A similar Norwegian study found that drowsiness contributes to only 4% of crashes, that 10% of male drivers and 4% of female drivers in Norway have fallen asleep behind the wheel, and that the majority of Norwegian drowsy driving crashes occurred on long straight roadways (Sagberg, 1999). Other studies focused on targeted populations show that nearly half of all long haul truck drivers have fallen asleep behind the

wheel (McCartt, Rohrbaugh, Hammer, & Fuller, 2000). In a summary of previous survey work, MacLean et al. (2003) concluded that 29% to 55% of drivers reported feeling drowsy while driving, 11% to 31% have fallen asleep behind the wheel, and between 4% and 12% of drivers have been involved in a drowsiness-related crash. Together with naturalistic driving data and crash databases, these survey data support the following conclusions:

1. Drowsy driving is a significant contributor to fatal vehicle crashes
2. Crashes attributed to drowsy driving often occur in the evening on long straight highways and are often single car run-off-road crashes.
3. Drowsy driving is more common in professional drivers compared to the general population

These conclusions emphasize the need for novel approaches to drowsy driving mitigation, particularly those that focus on the prevention of drowsy-related lane departures.

2.1.1 The problem structure of drowsiness-related vehicle crashes

Analysis of the consequences of drowsy driving provides a useful method for understanding the severity of the problem; however, it does not explain how drowsy driving crashes occur or what constitutes drowsy driving. The occurrence of a drowsiness-related crash can be viewed as a series latent failures begin with organization factors and unsafe supervision, which facilitate crash pre-conditions, ultimately causing unsafe behavior and resulting in crashes (Reason, 1990). These latent failures have been extensively studied through crash databases, naturalistic and simulated driving studies, and surveys that link latent failures with increases in the frequency of drowsiness-related crashes.

Organizational factors that increase crash rates include, policies that promote shift work, paired driving in a sleeper cab, policies that pressure drivers to arrive on time while following posted speed limits, and lax screening polices for sleep disorders (I. D. Brown, 1994; McCartt et al., 2000; Mitler, Miller, Lipsitz, Walsh, & Wylie, 1997; Pérez-Chada et al., 2005; Philip, 2005). Although many of these factors are unique to professional drivers previous work shows that the pressure to arrive at one's destination influences both professional drivers and the general public (Armstrong et al., 2010; I. D. Brown, 1997; Fletcher, McCulloch, Baulk, & Dawson, 2005). These organizational factors facilitate the

occurrence of crash preconditions that ultimately lead to an increased risk of a drowsiness-related crash. Drowsy driving preconditions are specific to the driver's health and sleep behavior (driver-specific preconditions) or the current drive and driving environment (drive-specific preconditions). Driver-specific preconditions for drowsiness-related crashes include sleep disorders such as Obstructive Sleep Apnea (OSA) or Narcolepsy, reduced sleep on the night before driving, and accumulated sleep debt. Drive-specific preconditions include long periods of driving without stopping, and uneventful highway driving conditions (I. D. Brown, 1994; Dinges, 1995; Lal & Craig, 2001; Maycock, 1997; Williamson et al., 2011). Another commonly reported drive-specific precondition is related to the circadian rhythm. The circadian rhythm is the process that governs the sleep wake cycle of human beings. The process includes two performance "dips," or circadian nadirs, which occur after midnight and after lunch (2 pm to 3 pm). Many studies have related these dips to increases in crash frequency (I. D. Brown, 1994; Dinges, 1995; Pack et al., 1995), however, recent work by Williamson et al. (2011) suggests that these effects may be confounded with the long drive durations and improper sleep schedules.

Organizational factors, when combined with a lack of proper supervision and crash preconditions have a cumulative effect on the risk of a drowsiness-related crash (Dawson, Noy, Härmä, Akerstedt, & Belenky, 2011). For example a sleeper-cab truck driver with a sleeping disorder driving at the end of a four-hour shift is at a much higher risk of a crash compared to the same driver at the beginning of his shift. It is important to note that even with optimal organizational design some crash preconditions can arise and produce drowsiness-related crashes. For example a driver may encounter heavy traffic in a snowstorm that increases the length of the drive. In these cases, as the effect of drowsiness grows stronger it leads to physiological changes that in turn cause performance decrements (unsafe acts), leading to drowsiness-related crashes. Physiological changes include changes in brain activity (specifically increases in delta and theta wave activity), eye closure, and blinking (Lal & Craig, 2001; MacLean et al., 2003; Wierwille, Wreggit, Kirn, Ellsworth, & Fairbanks, 1994). Research on performance-related changes has shown many conflicting results. MacLean et al. (2003) summarize these results and indicate that an increase in lane variability is the most commonly observed behavioral change. Other changes include a

decrease in speed and an increase in the frequency of lane departures. The precise link between drowsiness-related crashes and these physiological and behavioral changes remains an open area of research. The most ambiguous part of this relationship is the point at which drowsiness becomes a risk to the driver (MacLean et al., 2003). This point is not only ambiguous for researchers, but may also be unclear to drivers. This ambiguity, along with the fact that drowsiness is largely a subjective experience, creates a significant obstacle for analyzing drowsiness-related crashes and developing predictive algorithms that rely on ground truth drowsiness data. The organizational factors, unsafe supervision, preconditions, physiological changes, and driving effects are illustrated in Figure 1.

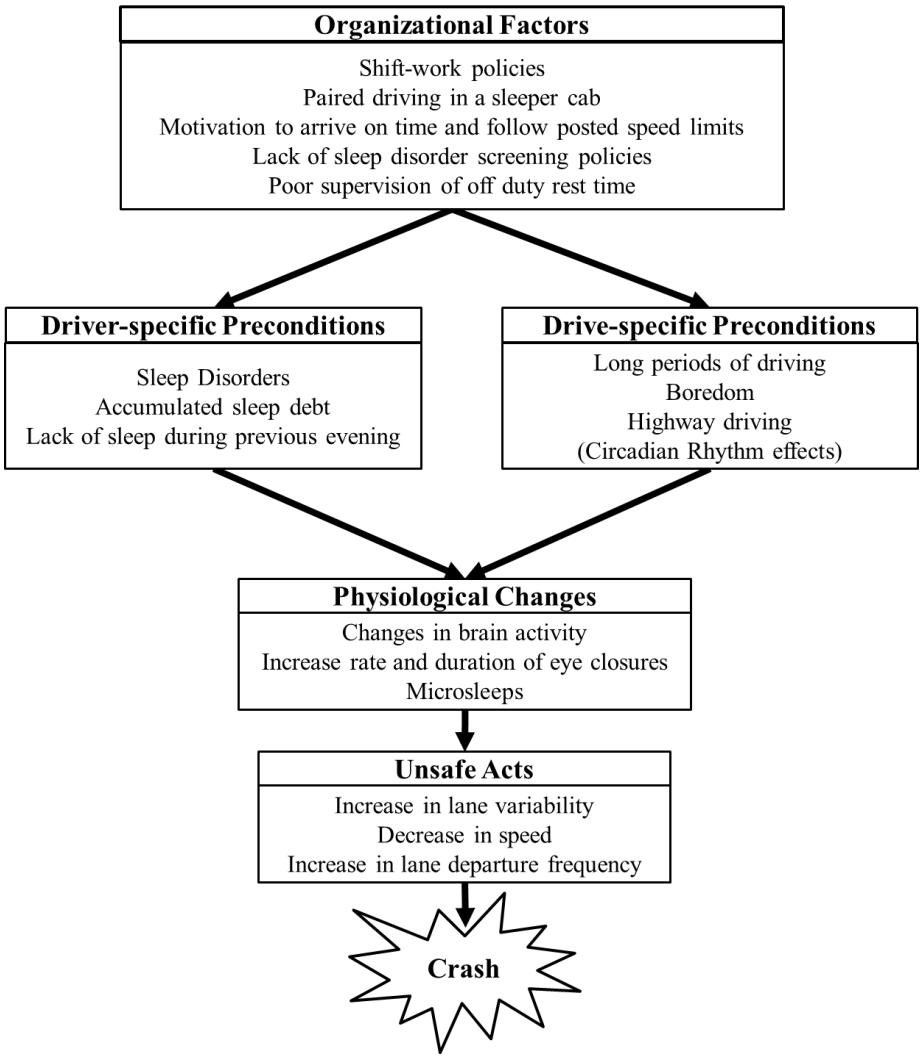

Figure 1 Illustration of the systems structure of drowsy driving crashes.

A systems view of drowsy driving allows one to analyze crash occurrence from a perspective that includes both proximal and distal causes of crashes and highlights behavioral changes that occur prior to crashes. The systems view leads to the following conclusions:

1. Organizational factors that lead to crashes include: poor rest/duty scheduling adherence, shift work, sleeping on the road, and pressure to arrive at one's destination
2. Preconditions for a drowsiness-related crash include sleep disorders, poor sleep during the previous evening, a lack of sleep over an extended period prior to driving, long duration driving, driving in the evening, a lack of stimulating environment, and driving alone
3. As preconditions become more severe they lead to changes in brain activity, increased duration of significant eye closures, increases in lane variability, poor reaction times, and difficulty maintaining ones lane
4. The exact link between physiological changes, performance decrements, and drowsiness-related crashes is unclear, and drivers may not be able to reliably detect when they are dangerously drowsy.

2.1.2 Potential solutions and effectiveness of impaired driving interventions

The multiple levels of latent failures associated with drowsiness-related crashes have prompted a variety of approaches to mitigating crashes. These approaches include restrictions on schedules for professional drivers (Gander et al., 2011), increased education for drivers (Fletcher et al., 2005), laws against drowsy driving (Geist, Merkt, & Altamuro, 2002), and technological approaches (Balkin et al., 2011). The first three types of approaches focus specifically on organizational factors and unsafe supervision, while technological approaches have been developed which address each level of latent failures. Despite this broad set of approaches crash data indicate that the drowsy driving is still a problem. Balkin et al. (2011) suggest that despite their good intentions, approaches that focus on reducing problems due to organizational factors and preconditions are difficult to enforce and thus ineffective. For example a restriction that limits daily driving time does not ensure that drivers spend their off-duty hours sleeping. Armstrong et al. (2010) further suggest that educational approaches do little to diminish the motivations

of drivers to arrive at their destination and are ineffective. Furthermore, drowsiness can arise even in optimal organization conditions, thus these approaches can only affect some types of drowsiness-related crashes. Technological approaches address these issues by providing timely, objective, and individualized feedback (Balkin et al., 2011).

Dinges and Mallis (1998) describe current technological approaches as belonging to one of four categories: readiness-to-perform technologies, mathematical models of fatigue, online driver monitoring technologies, and performance-based monitoring technologies. Readiness-to-perform technologies use pre-drive ocular-motor evaluations and subjective measures to assess a driver's capability of completing a drive safely (Balkin et al., 2011). Mathematical models of fatigue attempt to use known relationships between work hours, sleep, and performance to predict the effects of work patterns on performance (Dawson et al., 2011). In isolation, both of these technologies are limited because they cannot capture drowsiness that emerges during a drive. Their effectiveness might be increased if they are paired with other technologies such as online driver or performance-based monitoring technologies (Dinges, Maislin, Brewster, Krueger, & Carroll, 2005).

Online driver monitoring technologies continuously collect physiological measures from the driver and provide feedback when an algorithm within the system detects patterns in these measures associated with drowsiness. The primary limitation of these technologies is that current methods for collecting the measures required for these technologies are unreliable in many conditions or obtrusive to collect (Balkin et al., 2011; Hartley, Horberry, & Mabbott, 2000). Furthermore because the link between physiological changes and drowsy driving crashes has not been fully defined, the ability of these algorithms to effectively interact with drivers is uncertain. In contrast to online driver monitoring technologies, performance-based monitoring technologies use driver behavioral measures collected from vehicles to assess drowsiness and provide driver feedback. These technologies are advantageous relative to online driver monitoring because they are unobtrusive and can provide feedback that is directly related to performance decrements. The limitation with these technologies is that current methods show high rates of misidentifying drowsiness or false positive rates (Balkin et al., 2011). Furthermore there is little

consensus on the frequency interventions, i.e. how often an algorithm should produce an alert. Given the similarity between online driver and performance-based monitoring approaches, it is natural to unite them under a single term, real-time monitoring approaches.

A critical concern for real-time monitoring approaches is real-world validation. Hartley et al. (2000) reviewed online driver and performance-based monitoring technologies and suggested that while they were promising, few approaches were extensively validated. Dinges et al. (2005) conducted a validation study to evaluate the effect of drowsiness-related feedback on drowsy driving performance from a combination of readiness-for-duty and real-time monitoring approaches. Specifically, they compared driver feedback from a readiness-for-duty actigraphy watch, an eye-closure monitor, the CoPilot (Attention Technologies, Pittsburgh, Pennsylvania), and a lane keeping evaluation technology, SafeTRAC (Applied Perception and AssistWare Technology, Wexford, Pennsylvania). The CoPilot and SafeTRAC monitors produce independent continuous measures of drowsiness based on proprietary algorithms. The study results supported the hypothesis that drivers who received feedback had significantly lower indices of fatigue during nighttime driving, as measured by both the CoPilot and SafeTRAC (Dinges et al., 2005). The primary limitation of this study was that the technologies used were evaluated in combination and so the individual effects of each could not be assessed. Additionally the feedback provided by the systems was limited to visual displays, except when a driver fully departed their lane without signaling. However, the results suggest that real-time monitoring technologies are a promising approach to the problem of driver fatigue.

The lack of extensive evaluation of feedback in the Dinges et al. study is limiting because both operator and performance-based monitoring technologies are forms of automation. Specifically they represent automation that performs information analysis (Parasuraman, Sheridan, & Wickens, 2000). Appropriate use of such automation requires trust and an understanding of system states (Lee & See, 2004). False positives and inappropriate interventions decrease trust and may lead a driver to disengage a system. The false positive rates associated with online driver and performance-based technologies are a

product of their internal algorithm that relates patterns in the data to the driver's state. Thus this algorithm is a critical component of the system success.

Previous solutions aimed at reducing drowsy driving crashes have had some successes yet they have not eliminated the problem. Analysis of the approaches and their components leads to the following conclusions:

1. Methods that attempt to remove on the organizational factors and crash preconditions have limited effectiveness because they cannot address fatigue that emerges during a drive and do not motivate drivers to change their off-work sleep behavior.
2. Real-time technological methods that use driver or performance-based measures to predict drowsiness and provide informative feedback to the driver are a promising direction for future work.
3. The success of these real-time technological methods depends on the relationship between these technologies and drivers.
4. The drowsiness detection algorithm is a critical component of this relationship and thus a critical design component of the system.

2.2 Current drowsiness detection algorithms

Previous work on real-time monitoring technological interventions indicates that the drowsiness detection algorithm is a critical component of their success. A significant amount of research has focused on the development of such algorithms (Sahayadhas, Sundaraj, & Murugappan, 2012; Wang, Yang, Ren, & Zheng, 2006). In all of this work drowsiness detection is framed as a supervised machine learning problem. Supervised machine learning is a process of training predictive models by applying algorithms on sets of labeled data, with the goal of using the trained models to label previous unseen data (Mitchell, 1997). The choice of label (or ground truth), source of input data, and training algorithm are design parameters. In the context of drowsiness detection the supervised machine learning process consists of selecting a ground truth definition of drowsiness, collecting input measures, and training a machine

learning algorithm on the data. Occasionally a data pre-processing step is used to convert input measures to a form that is amenable to the input requirements of machine learning algorithms (eg: Sayed & Eskandarian, 2001). Once the measures of pre-processed data are integrated into the model they are referred to as features. Together the ground truth definition, input measures and features, and choice of machine learning algorithm constitute a drowsiness detection algorithm. Figure 2 shows an example of one common drowsiness detection algorithm, PERCLOS80.

The PERCLOS80 Algorithm			
Input Measure(s)	Feature(s)	Machine Learning Algorithm	Ground Truth Definition of Drowsiness
Eye-Closure	Percentage of a two-minute time window that the Eyes are more than 80% closed	Threshold (or Decision Stump)	Delayed Reaction time

Figure 2 Illustration of the components of the PERCLOS80 algorithm (Dinges & Grace, 1998).

2.2.1 Input measures and feature generation techniques employed in current algorithms

Many different measurement sources have been explored for drowsiness identification including: heart rate (Furman, Baharav, Cahan, & Akselrod, 2008), brain activity (Dinges, Mallis, Maislin, & Powell IV, 1998; Lal et al., 2003), eye closure and tracking (Dinges et al., 1998; Ji, Zhu, & Lan, 2004), lane position (Hanowski, Bowman, Alden, Wierwille, & Carroll, 2008), and steering-wheel angle (Eskandarian & Mortazavi, 2007; Krajewski, Golz, et al., 2009; Krajewski, Sommer, et al., 2009; Sayed & Eskandarian, 2001). Although most previous algorithms focus on a single type of measure, some employ several different measures (Forsman, Vila, Short, Mott, & Van Dongen, 2013; Ji et al., 2004; Tijerina, Gleckler, & Stoltzfus, 1999; Zilberg, Xu, Burton, Karrar, & Lal, 2007). The most commonly applied and theoretically rigorous measures are electroencephalogram (EEG), percent eye-closure over a fixed time window (PERCLOS), and steering-wheel angle (Balkin et al., 2011).

The choice of EEG data as an input measure for drowsiness detection systems grows from the well-established link between spectral EEG patterns and the transition between wakefulness and sleep

(Lal & Craig, 2001). EEG-based algorithms typically involve significant pre-processing. In most cases this preprocessing involves conversion of the original signal to the frequency domain via a Fast Fourier Transform, and then converting the frequency data into input features (Lal et al., 2003; C.-T. Lin et al., 2008; Yeo et al., 2009). Other approaches have generated features through wavelet transforms (Khushaba, Kodagoda, Lal, & Dissanayake, 2011; Wilson & Bracewell, 2000) and independent component analysis (C. Lin et al., 2005). Although it is difficult to evaluate the general effectiveness of these algorithms due to variance in evaluation criteria and ground truth definitions of drowsiness, some have reported error rates of approximately 10% (Lal et al., 2003; Vuckovic, Radivojevic, Chen, & Popovic, 2002). Despite this success, the use of EEG data is highly impractical in a real-world driving setting. Current EEG technology requires electrodes to be attached to a driver's head and the measure may be compromised by electrical impulses from muscles, and other electronic sources inside of the vehicle (Balkin et al., 2011).

PERCLOS measures the percent closure of the eyes averaged over a time window, often as short as one minute (Wierwille et al., 1994). PERCLOS has been incorporated into an algorithm, PERCLOS80 (also referred to as PERCLOS), which predicts drowsiness based on the percentage of time an individual's eyes are more than 80% closed over a 2-minute period. Dinges et al. (1998) demonstrated that the PERCLOS algorithm had over 90% accuracy in detecting degraded performance during a vigilance task, which was more reliable across drivers than EEG, blinks, and head position. PERCLOS has been incorporated into aftermarket devices such as the Co-pilot (R Grace & Stewart, 2001). PERCLOS is so widely accepted that it has even been viewed by some as a ground truth measure of drowsiness (Tijerina et al., 1999; Wierwille et al., 1994). This acceptance suggests that PERCLOS is the paragon of drowsiness detection measures however the measure still has both practical and theoretical limitations. PERCLOS is practically limited because current camera technology required for its measurement is expensive, has not been extensively validated, and may be unreliable when the driver wears sunglasses or under weather conditions that produce high amounts of glare (Balkin et al., 2011). PERCLOS is theoretically limited because it aggregates data over minutes. This is problematic because

microsleep episodes (lasting between 6 and 12 s) that contribute to drowsy-related lane departures may begin and end within a small portion of the two minute time window of PERCLOS calculation (Boyle, Tippin, Paul, & Rizzo, 2008) and might not trigger a PERCLOS-based alarm. Despite its limitations, the overwhelming evidence in favor of PERCLOS suggests that it might be useful for benchmarking new algorithms.

The instrumentation required for PERCLOS and EEG has led researchers to examine Steering-wheel angle, or the deflection of the top of the wheel from the zero point. Steering-wheel angle is typically viewed similarly to EEG in that it requires significant pre-processing and transformation before it becomes a valuable data source. Sayed and Eskandarian (2001) introduced an algorithm based on steering-wheel angle measures that filtered the measures to remove road curvature events, and then discretized into bins customized to drivers. The algorithm ultimately achieved an accuracy of nearly 90% on classifying drivers labeled as sleep deprived or non-sleep deprived. Krajewski et al. (2009) developed a separate algorithm that used a wide range of features derived from steering-wheel angle data. These features characterized the signal in the time domain, frequency domain, and also represented non-linear aspects of the sequence. In all, over 1,200 features were created. The subsequent model achieved 86% accuracy in identifying self-reported sleepiness; however the pre-processing associated with the many features makes real-time application a challenge and creates an opportunity for overfitting. McDonald et al. (2013) presented an approach which used only raw steering wheel angle data. The subsequent model performed comparably PERCLOS in detecting drowsy-related lane departures. Steering-wheel angle methods have achieved comparable performance to both EEG and PERCLOS methods, without the obtrusiveness of EEG or the significant cost and theoretical limitations of PERCLOS. The main limitation of steering-wheel data is that it is highly sensitive to differences in driving activities, such as curve negotiation, and thus detection could be confounded with differences in the driving context (Balkin et al., 2011; Hartley et al., 2000).

2.2.2 Ground truth drowsiness employed in current algorithms

The labels used to identify drowsiness and wakefulness in training and evaluating algorithms represent the ground truth definition of the state being predicted. This ground truth state is critical to the success of the algorithms in real world applications because it influences false positive rates and the type of feedback the algorithm provides to the driver (Balkin et al., 2011). Unfortunately, there is no uniformly agreed upon ground truth measure of drowsiness. This lack of ground truth reflects the lack of a definitive definition of drowsiness itself (Noy et al., 2011) and the ambiguity of the link between subjective feelings of drowsiness and the consequences of drowsiness-related crashes (MacLean et al., 2003). The issue is further complicated because there are large individual differences in drivers' responses to the common ways to induce drowsiness, such as sampling during the circadian nadir, extending periods of wakefulness, and reducing cumulative sleep over a week. These differences even extend to the physiological level, where different measures of brain activity that should define drowsiness sometimes conflict (Lal & Craig, 2001). Furthermore, drowsiness is both a chronic and an episodic phenomenon, meaning that it can affect drivers throughout a drive and occur intermittently during a drive (Richard Grace et al., 1996).

These challenges have led the research community to develop many different labels of drowsiness, including: time of the drive (Sayed & Eskandarian, 2001), time into drive (Zhao, Zhao, Liu, & Zheng, 2012), subjective ratings (Eskandarian & Mortazavi, 2007; Krajewski, Golz, & Sommer, 2009), vigilance tests prior to the drive (Dinges et al., 1998; Forsman et al., 2013; Ji, Lan, & Looney, 2006; Ji et al., 2004), video coding (Lal et al., 2003; Zilberg et al., 2007), and, in simulator studies, the occurrences of crashes (Zhao, Zheng, Zhao, Tu, & Liu, 2011). All of these labels are limited. For example, experimental manipulations such as driving without sleeping the prior evening or driving for a long period does not always generate episodic drowsiness, completing subjective questionnaires may diminish drowsiness, vigilance tests prior to the drive may not capture task-based fatigue, and simulator studies may lead to unnatural driving. Video coding is more robust, but it may be tautological for algorithms such as PERCLOS: the same data are used to define drowsiness and by the algorithm to detect drowsiness.

One promising approach to these limitations is to use a multi-faceted or ensemble approach to defining drowsiness which considers task-performance, subjective ratings, and video coding (McDonald, Lee, Schwarz, et al., 2013).

2.2.3 Machine learning approaches in current algorithms

Many machine learning approaches are available for supervised learning, but only a small subset has been applied to drowsiness and impairment detection. These approaches can be characterized by their training procedure, prediction procedure, and their optimization parameters. Table 1 describes these approaches across the three dimensions. Kotsiantis et al. (2007) provide a concise review of many of these approaches, evaluate their strengths and weaknesses, and conclude that the utility of each approach depends on the data, which suggests that any approach must be confirmed by fitting and evaluating algorithms. These data dependencies and the variety of input/ground truth pairings employed in previous work make objective comparisons between individual approaches difficult. However Table 1 does suggest some concerning themes in previous work. All of the models in Table 1 except for Bayesian Networks (BN) and Dynamic Bayesian Networks (DBN) contain no representation of the problem structure of drowsiness-related crashes illustrated in Figure 1. These algorithms must rely solely on patterns in the data to make predictions. This is particularly concerning for steering-based algorithms that may be highly sensitive to road context. Additionally, all of the models except for Dynamic Bayesian Networks are inherently static and thus do not consider temporal effects of drowsiness that may affect their accuracy.

Table 1 Training procedures, prediction procedures, and previous applications of common machine learning algorithms in drowsiness detection.*Note that these simplifications will be expanded in the section entitled Temporal machine learning approaches and dynamic graphical models.

Machine Learning Approach	*Training Procedure*	*Prediction Procedure*	*Previous Algorithms*
Decision Trees (DT)	Create a tree of conditional (If…Then) statements that partition the training data by ground truth labels. Each partition is selected using information gain (or a similar metric), which measures the reduction of uncertainty in the data after creating a split. After a partition is selected, the data characterized by the partition are evaluated for additional splits. The process stops when the training data are fully partitioned by their labels or when a pre-defined maximum depth of the tree is reached.	Use the values of the features of the unlabeled instance to follow the conditional statements in the tree until reaching a leaf node (the end of a path). Return the label associated with the majority of instances at that node.	(Krajewski & Sommer, 2009)
Neural Networks (NN)	Create a structure of interconnected layers of nodes where each connection has an associated weight and each node represents the application of a sigmoid (or similar) function to the sum of the output of previous nodes multiplied by the weight between the previous node and current node. Adjust the weights using an algorithm known as back-propagation. This process essentially fits a hyperplane that separates the data in a reduced feature space.	Use the weights of the network and the input feature values to calculate the output of the network and then convert the output to a binary label using a threshold.	(Patel et al., 2011; Sandberg, Akerstedt, Anund, Kecklund, & Wahde, 2011; Sayed & Eskandarian, 2001; Vuckovic et al., 2002; Wilson & Bracewell, 2000)
Support Vector Machines (SVM)	Use a linear program to fit a hyperplane to the training data that maximizes the margin between the subsets of training data associated with each ground truth label. The maximization process is subject to a constraint of only allowing a set amount of training instances to fall within the margin (determined by a soft margin parameter). This process can be supplemented with a kernel function, which transforms the data to an alternative feature space.	Use the features of the unlabeled instance to find which side of the margin the instance is on and return the associated label. If the point lies in the margin return the label associated with the side closest to the point.	(Hu & Zheng, 2009; Kutila et al., 2007; Zhao et al., 2012)

Machine Learning Approach	*Training Procedure*	*Prediction Procedure*	*Previous Algorithms*
k-Nearest Neighbor (kNN)	Store the training data in a data structure.	Find the k nearest neighbors from labeled instances to the unlabeled instance using Euclidean distance (or a similar measure) between the feature vectors. Return the label representing the majority of neighbors.	(Khushaba et al., 2011; Krajewski & Sommer, 2009)
Random Forest (RF)	Take a bootstrapped random sample of the training data and the training features and train a decision tree on the sample. Repeat the process many times (e.g. 500 times).	Use the values of the features associated with the unlabeled instance to find predictions from each trained tree, and return the majority vote amongst the trees.	(McDonald, Lee, Schwarz, et al., 2013)
Bayesian Network (BN)	Specify a directed acyclic network structure via an algorithm (e.g. the Chow-Liu algorithm) or domain-knowledge. Use an estimation technique (e.g. maximum likelihood estimation) to calculate the probability of various model states via the dependencies specified by the model structure	Find the probability of each possible label, given the values of the training instance features. Return the label associated with the highest likelihood.	(Ji et al., 2004; J. H. Yang, Tijerina, Pilutti, Coughlin, & Feron, 2009)
Dynamic Bayesian Network (DBN)	Specify a directed acyclic network structure and temporal dependencies between time slices. Estimate the probability of various model states via the dependencies specified by the model structure.*	Find the probability of each possible label given training sequence and return the label associated with the highest likelihood.*	(Ji et al., 2006; G. Yang et al., 2010; J. H. Yang et al., 2009)

2.2.4 Dynamic Bayesian Network drowsiness detection algorithms

A Dynamic Bayesian Network (DBN) is a dynamic probabilistic graphical model. DBN models consist of graph structures—nodes connected by edges—that mimic the dependencies in the underlying problem, and an associated group of probabilities that model the likelihood of model states. The dynamic portion of the model specifies dependencies across discrete time slices. The DBN models warrant further discussion because they may provide a comprehensive solution to drowsiness detection in that they capture both the problem structure (Figure 1) and temporal aspects. The algorithm presented in Ji et al. (2006) combines contextual, facial, ocular, and head-position input to predict drowsiness as defined by

reaction times during a non-driving vigilance task. The contextual information consisted of circadian rhythm, sleep quality, the presence of sleep disorders, and information about one's work environment encoded as binary variables. The parameters for these contextual factors were inferred based on domain knowledge-based estimates. The model's predictions had a correlation coefficient of 0.953 with reaction time delays induced by 25 hours of continuous wakefulness. Yang et al. (2010) extended this work by developing an algorithm that included heart rate, EEG, and eye measures as input in addition to a small subset of contextual factors. As in the earlier work, these contextual factors were binary coded, and consisted of sleep quality and work environment measures. In an unrelated work, Yang et al. (2009) developed a DBN for predicting sleep deprivation using driver performance on a series of stimulus response tasks and force paced driving maneuvers in a driving simulator. Although each of these studies recognizes the importance of time in drowsiness detection, they are all limited in that they do not consider performance-based input, do not include on-road contextual information, and do not employ a rigorous model selection process. Furthermore the consideration of contextual factors that occur at higher levels of the hierarchy of drowsiness-related crashes, such as circadian rhythm may bias these models towards the identification of chronic drowsiness rather than episodic drowsiness.

2.2.5 Potential knowledge gaps in current drowsiness detection algorithms

Current drowsiness detection algorithms can be characterized by their choice of input measure, ground truth definition of drowsiness, and machine learning method. Analysis of each of these components leads to the following conclusions:

1. Steering-wheel angle is advantageous relative to EEG and PERCLOS because it is highly practical and inexpensive.
2. The breadth of support of PERCLOS in previous work suggests it should be used as a benchmark for new algorithms.
3. There is no consensus on a ground truth definition of drowsiness, but ensemble definitions may overcome individual weaknesses of various measures.

4. Most previous algorithms do not consider the temporal or contextual factors associated with drowsiness-related crashes.
5. Methods that employ DBN consider temporal aspects but do not consider the current road context or performance-based measures.
6. No method of drowsiness detection has achieved perfect accuracy and most have error rates greater than 10%.

These conclusions, when combined with the previous section, suggest that previous algorithms are missing critical components of the drowsiness detection problem. The lack of temporal understanding in many algorithms is a critical limitation as time is a central function in drowsiness and drowsy driving consequences. Furthermore the lack of contextual consideration particularly the on-road context is another significant gap.

2.3 Contrasting detection algorithms and driver behavioral models

The significance of the lack of consideration of time and on-road context in current drowsiness detection models can best be understood through an analysis of two types of models that are closely related to the goal of driver drowsiness detection algorithms: theoretical driver behavioral models and behavior identification models. The goal of theoretical driver behavioral models is to explain driver behavior through representations of the fundamental factors that define driving. Despite their broader goals, the environmental aspects of these models provide critical insights into the importance of time and context in driving. Behavioral identification models use behavioral observations, such as driver steering or pictures of an individual, to train algorithms that can identify classes of behavior (e.g., performing a lane change, walking) or intentions of the observed individual. Although these models have different goals and application domains, their approach to time-series human behavioral data provide insights into modeling drowsiness detection.

2.3.1 Theoretical models of driver behavior

There is a breadth of research on theoretical driver behavioral models. Most of this work can be consolidated into three types of models: control theoretic models, behavioral adaptation models, and cognitive architectures. Control theoretic models view the driver as a control element in a continuous system (McRuer & Weir, 1969; Weir & McRuer, 1970). Hence the driver adapts and manipulates dynamic characteristics of the vehicle to satisfy control requirements such as selecting the appropriate tolerances, maintaining the automobile on a particular pathway, reducing path errors within a minimum threshold, and maintaining an established path in the presence of disturbances. Satisfying these control requirements requires input that is both time-sensitive and dependent on road characteristics, such as a curvature (McRuer et al., 1977). The key insight of these models is that time and road geometry are central components to the driving process.

In contrast to control theoretic approaches, behavioral adaptation models view the driver as an entity that seeks to optimize an objective function. Various models view this objective function as a measure of risk or task difficulty (Fuller, 2005; Summala, 1988; Wilde, 1982, 1988, 1994). Hence the driver choses actions and engages in behaviors with the goal of maintain a manageable level of task difficulty. The task difficulty is driven by road characteristics, and the presence of other drivers on the road, and the primary method of modulating task difficulty is through varying the vehicle's speed (Fuller, 2005). Thus behavioral adaptation models expand the notion of road context to include both roadway geometry and interactions with other drivers and develop a link between driver behavior (i.e. speed and acceleration) and context.

Cognitive architectures are comprehensive models that include many of the concepts introduced in control theoretic and behavioral adaptation models (Salvucci et al., 2001; Salvucci, 2006). Cognitive architectures, specifically the ACT-R framework (Anderson & Lebiere, 1998), are hierarchical models which construct complex actions from sets of basic rules. For example in the driving context a complex action would be a lane change, which can be executed using basic rules such as "avoid other vehicles" and "use a turn signal when changing lane." A key component of these models is that they contain a

learning component (Michon, 1989). This component allows drivers to consolidate basic behaviors into repeatable complex actions stored in their memory. This conceptualization of behavior as a repeatable and hierarchical process has significance for the definition of road context. For example execution of complex processes such as a lane change can occur within the context of broader processes such as highway driving. Therefore context is not simply defined by a maneuver itself but also the broader road context.

The three most prominent approaches to theoretical driver modeling suggest that driving is a fundamentally a time-dependent, context-driven activity. The context is hierarchical in nature and depends on a variety of factors including the road geometry, surrounding vehicles, and environment. Furthermore, the context is intimately related to a driver's manipulation of speed. Thus this behavioral hierarchy, along with the time, is a critical determinant of driving behavior and represents a potentially critical omission in current driver drowsiness detection algorithms.

2.3.2 Human behavioral identification models

Human behavioral identification is a broad field that has implications to indexing complex data, adaptive automation, and emerging technology (Brand, Oliver, & Pentland, 1997; Jabon, Bailenson, Pontikakis, Takayama, & Nass, 2011; Turaga, Chellappa, Subrahmanian, & Udrea, 2008). The models produced by this work typically involve a process of collecting labeled behavioral data from various sensors, integrating the data into a set of observations, and training a machine learning algorithm to detect the labels. This process has been applied to a wide variety of domains including in-home activities (van Kasteren, Englebienne, & Krose, 2010), exercise movements (Brand et al., 1997), and driving (Doshi, Morris, & Trivedi, 2011; Kuge, Yamamura, Shimoyama, & Liu, 2000; Oliver & Pentland, 2000). In the driving context behavioral identification models typically focus on recognizing driver activities (e.g. lane changes, passing a lead vehicle) or the state of the driver-vehicle system (e.g. following a lead vehicle, stopped at a traffic signal). This focus contrasts with the goal of drowsiness detection algorithms, which is specifically focused on driver state. Despite the difference in focus, the similarity of process suggests that many of the modeling frameworks used in the driver behavioral detection literature could be applied to drowsiness detection.

The most prominent difference in the behavioral identification and drowsiness detection literature is a focus on time-series modeling frameworks. These frameworks include Dynamic Bayesian Networks (Dagli, Brost, & Breuel, 2003), a variety of Hidden Markov Model structures (Kuge et al., 2000; McCall & Trivedi, 2007; Oliver & Pentland, 2000; Sathyanarayana, Boyraz, & Hansen, 2008; Torkkola, Venkatesan, & Liu, 2005), and Conditional Random Fields (Raksincharoensak et al., 2010). The specific nature of these models will be discussed in the following section; however all of these models are dynamic graphical approaches. The disparity between the frequency of these models in the behavioral detection literature and drowsiness detection algorithm literature further solidifies the need to apply such models in drowsiness detection, particularly because the behavioral detection literature often considers driving (Dagli et al., 2003; Kuge et al., 2000; Oliver & Pentland, 2000; Raksincharoensak et al., 2010; Sathyanarayana et al., 2008; Torkkola et al., 2005).

2.4 Temporal machine learning approaches and dynamic graphical models

Theoretical driver behavioral models and human behavior identification models emphasize the importance of context and temporal modeling in drowsy driving detection algorithms. Incorporating temporal relationships into a drowsiness detection algorithm requires a shift in perspective from static to dynamic machine learning approaches. There are many possible approaches that accommodate this including: sliding windows, recurrent sliding-windows, and probabilistic graphical models (Dietterich, 2002). Sliding windows convert a dynamic machine learning problem to a static problem by binning consecutive data points into fixed windows. They are advantageous because static machine learning algorithms can be used; however, they are limited because they cannot capture relationships between successive windows.

Recurrent sliding windows attempt to address this problem by appending classifications from previous time window into the following feature vectors. Although these methods are reasonable for many applications, they do not align with the goals of this dissertation because they do not allow contextual factors to be integrated into the algorithm. Probabilistic graphical models solve this problem

by explicitly modeling both temporal dependencies and contextual dependencies. Probabilistic graphical models represent a broad range of modeling concepts and structures. The goal of this section is to introduce a set of graphical modeling paradigms and structures that could feasibly be used to detect drowsiness (for a more exhaustive list of graphical modeling structures consult Murphy (2002), Dietterich (2002), and Sutton and McCallum (2006)). Analyses of the human behavioral identification literature and the physiological characteristics of drowsiness direct this study towards three broad classes of models: Hidden Markov Models, Hidden semi-Markov Models, and Conditional Random Fields. The Hidden Markov and Hidden semi-Markov Models represent Dynamic Bayesian Networks and are two of the most common structures explored in human behavioral identification. Conditional Random Fields represent a more recent development in predictive modeling and are also extensively applied in the human behavioral identification literature. This section will review the three structures and the implications of their use for drowsiness detection.

2.4.1 Graphical modeling concepts

Prior to the discussion of probabilistic graphical models, it is useful to review general graph theory and establish terminology. A graph is a collection of nodes connected by edges, or vertices. In a probabilistic framework, nodes represent observed or hidden (random) variables, edges represent dependencies between variables, and the graph itself can be represented by a joint probability distribution. Edges can be directed, representing unidirectional causation, or undirected, representing bidirectional influences. When a directed edge connects two nodes, the receiving node is known as a child and the sending node is known as the parent node, by extension all parents of the parent of the child node are referred to as ancestors. If two nodes are not connected and do not share a common ancestor they are probabilistically independent, meaning that knowledge of the state of one node has no effect on the state of the other. If two nodes are disconnected but share a common ancestor then they are conditionally independent, meaning that once the state of the common ancestor is known, knowledge of the state of one node has no effect on the state of the other (Salamin & Vinciarelli, 2011). The notion of conditional independence is critical to the utility of

graphical models because it reduces the complexity of calculations, such as the likelihood of a particular set of states in a graph, which are central to drowsiness detection.

2.4.2 Dynamic Bayesian Networks (DBN)

Dynamic Bayesian Network models consist of a large class of directed, acyclic, probabilistic graphical modeling structures. Some of the most common DBN structures are assigned more specific names (e.g., Hidden Markov Model, Hidden semi-Markov Model) however they still follow general DBN assumptions. DBNs model temporal processes through a series of discrete time-steps. The models allow for dependencies both within and between each time-step or between multiple successive time-steps. An example of two time steps of a DBN is shown in Figure 3. The figure shows that node B depends on node A from the previous time-step and node O_2 from the current time-step. In the figure, node A and B correspond to hidden states and nodes O_1 and O_2 correspond to observations. In general A and B represent discrete values, such as awake or drowsy, and each observation can be either discrete or continuous. If the observations are discrete then they are related to the hidden states via a table of proportions and if the observations are continuous they can be related to the hidden states via an estimated probability density function.

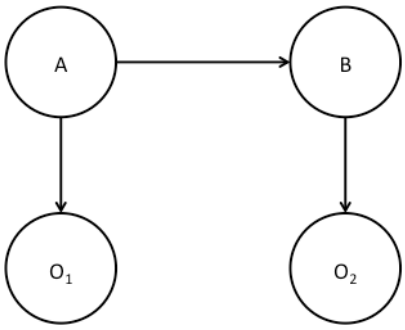

Figure 3 Illustration of the first two time-steps of a simply Dynamic Bayesian Network model. Nodes A and B represent hidden states and the nodes O_1 and O_2 represent observations.

The process of developing and training a DBN consists of two steps: defining the model structure and learning the parameters specified by the structure. Regardless of structure the parameters can be learned through an estimation technique such as Maximum Likelihood Estimation (MLE) if the data are fully labeled. Additionally the parameters can be refined by using a procedure called, the Baum-Welch

algorithm, which is a dynamic programming method based on Expectation Maximization. There are three primary methods for developing a DBN model structure: use an existing structure, learn the structure from the data, or specify the structure using domain knowledge. Developing the structure from the data is an attractive approach; however, it shifts the algorithm design process to a bottom-up data-based approach rather than a top-down theory-based approach. Previous methods of drowsiness detection have used a theory-based approach (Ji et al., 2006). The use of previous structures, such as the Hidden Markov Model, is common in the behavioral detection literature (Oliver & Pentland, 2000), yet it has not been explored in the drowsiness detection literature. Therefore it is worthwhile to review these structures and assess their feasibility for drowsiness detection.

2.4.2.1 Hidden Markov Models

Hidden Markov Models (HMMs) are the most commonly applied and simplest DBN. They have been used extensively in many domains including speech recognition and driving behavioral detection (Kuge et al., 2000; Rabiner, 1989). HMMs model a process in which there is a single random variable that cannot be observed, the hidden state, and a random variable that can be observed, the observation, at each time-step. The hidden state is a discrete value within a fixed set and the observations can be either discrete or continuous. HMMs assume that the hidden state depends on only the previous hidden state and that each individual observation is independent conditioned on the hidden state. The transition probabilities between successive hidden states are defined by a Markov chain (Rabiner & Juang, 1986). An example of three time-steps of an HMM is shown in Figure 4.

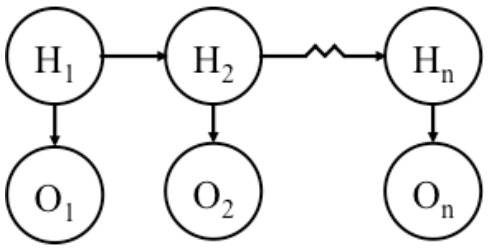

Figure 4 Example of 3 time-steps of an HMM model. Nodes labeled with an "H" represent hidden states and nodes labeled with an "O" represent observations.

The parameters of the HMM can be learned through Maximum Likelihood Estimation and they can be refined through a procedure known as the Baum-Welch algorithm. The Baum-Welch algorithm uses an iterative series of expectations of parameters and maximum likelihood maximizations to find the maximum likelihood estimate for model parameters based on a set of observations. The process consists of maximizing the likelihood of each state at each given time by calculating the likelihood of that state given all possible previous states (known as Forward probabilities) and the likelihood of that state given all possible future states (known as Backward probabilities). The use of Forward and Backward probabilities often leads the Baum-Welch algorithm to be referred to as the Forward-Backward Algorithm. As this process is repeated the model converges on the set of parameters that maximize the model likelihood. After model training, a similar process that uses the Forward probabilities can be used to generate model predictions in real-time (Rabiner, 1989).

HMMs are advantageous because they are supported by an extensive library of software, are easy to implement and have a fixed structure, however they are often a poor model of the underlying situation (Dietterich, 2002). The HMMs inability to explicitly model the duration of the hidden states, reliance on a single hidden state, and strict assumption of independence between the observations may be problematic in many real-world time-series analyses (Brand et al., 1997). However the broad use of HMMs in the human behavioral identification literature along with their relative success in that area suggests that they could be a valuable tool for drowsiness detection.

2.4.2.2 Hidden semi-Markov Models

Hidden semi-Markov Models (HsMM) or Variable-duration HMMs are an extension of basic HMMs that is designed to model the duration of each hidden state. Essentially this extension requires the addition of a layer of nodes to the state space that are used to track the time spent in the current hidden state. This addition is depicted in Figure 5, which shows three time-steps of an HsMM model. Mathematically this change requires the addition of a dependency in the transition probabilities that forces the probability to 0 unless the state duration has finished (Murphy, 2002). The state durations themselves can be either fixed values or based on an estimated probability density function based on the observed state durations in the training data set. As with the HMM, the parameters for the HsMM can be specified with Maximum Likelihood Estimation based on the training data and refined with a similar process to Baum-Welch that takes the state durations into account when calculating Forward and Backward probabilities and maximizing the model likelihood (O'Connell & Højsgaard, 2011). Furthermore the Forward portion of this algorithm can be applied to make predictions in real-time that maximize the likelihood of the given state at that time.

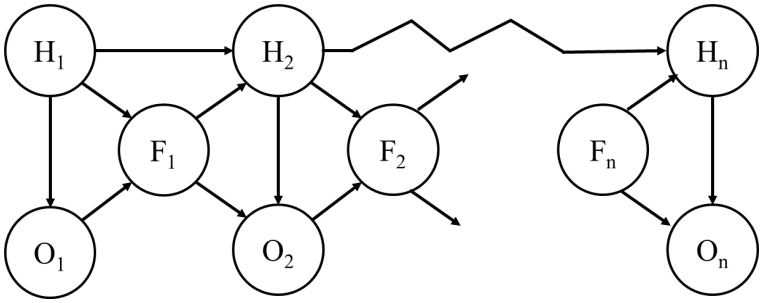

Figure 5 Example of 3 time-steps of a variable duration HMM. Note that the "F" nodes represent transition nodes that are only reached after the duration of the current hidden state.

The advantage of HsMMs is the ability to explicitly model duration of states, which can reduce the classification error however the addition of this node increases the number of parameters in the model, which increases the likelihood of overfitting the model. These extra parameters also increase the run-time of inferring the current hidden state (Murphy, 2002). However some evidence suggests that HsMMs can outperform HMMs in under certain human behavioral conditions (van Kasteren et al., 2010) and so they are valuable to explore for drowsiness detection.

2.4.3 Conditional Random Fields (CRF)

Conditional Random Fields are a class of undirected probabilistic graphical models (Lafferty, McCallum, & Pereira, 2001). An example of a CRF which is analogous to an HMM, the linear chain CRF, is shown in Figure 6. In contrast to DBN models CRFs are discriminative models, which means they do not require one to estimate the joint distribution of the model, only the likelihood of the sequence of labels conditioned on the observations (Dietterich, 2002). Furthermore the graph of the hidden states is undirected. Fundamentally this means that the relationships between an observation and a hidden state specified by a series of functions called potentials (denoted by boxes in Figure 6), which do not need to be probabilities. This means that CRFs do not make any explicit assumptions about the relationship between input data features, which allows the CRF to include highly correlated features that cannot be incorporated into DBNs (Sutton & McCallum, 2006). The ability to include such arbitrary features is critical to success of time-series modeling because it facilitates capturing behavior on a variety of time-scales. For example a drowsiness detection algorithm that used PERCLOS and CRFs could include features based on eye-closure over variable length time windows, which might facilitate detecting both chronic drowsiness and microsleep episodes. In an HMM framework these observations could not be included because they are not independent.

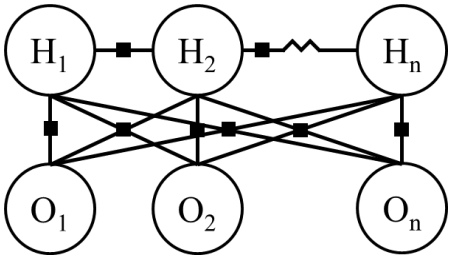

Figure 6 Example of 3 time-steps of a linear chain CRF. Note the undirected arcs between the hidden states and the boxes representing weights on each observation.

CRF models involve a similar training process to HMMs in that they use the Forward-Backward algorithm, although the process is more computationally expensive because the model parameters must be optimized for each training instance (Sutton & McCallum, 2006). Predictions for an entire sequence can be performed using a similar procedure to HMM, however in real-time applications, the Conditional

Random-Field training and evaluation process must be altered such that it optimizes predictions based on the information and weights available at the given time (Barbu, 2009).

CRFs suffer from two principle drawbacks. First, the training of CRFs is very expensive because it requires global adjustments of the potential functions (Dietterich, 2002). This can be minimized through careful state specification; however it may still be an issue for implementation, particularly if the implementation allows for online training which uses individual driver data to customize to the driver. The second drawback is due to the optimization of the hidden states conditional on the observations and the lack of a model of the joint distribution of states and observations. The change in optimization is problematic from a theoretical sense because it inhibits understanding of the causal pathways of the problem; in essence the goal of a CRF is to maximize classification accuracy rather than the explanatory power of the model.

2.4.4 Integrating graphical models and drowsiness

The role of temporal dependencies in both drowsiness and driving suggests time-series approaches may be valuable for drowsiness detection algorithms. Dynamic graphical models provide a direction for this pursuit. Although basic dynamic graphical models have been explored in drowsiness detection work, a thorough analysis of the effect of graphical structure and design on detection performance has not been undertaken. There is a vast array of available graphical modeling structures and paradigms. The goal of this section was to present a subset of these that would be most appropriate for drowsiness detection considering broad problem specifications.

All of these models have strengths and weaknesses relative to drowsiness detection algorithms. HMMs are supported by an extensive library of software and have been extensively validated; however, their pairing of a single state and observation for each time-step might be significantly limiting for drowsiness detection. HsMMs may allow an algorithm to capture microsleep duration and reduce false positive classifications, but they still assume a single observation and strict independence relationships between observations. CRFs provide a structure that is free of many of the assumptions of DBN models yet they might not allow for conclusions to be drawn about general relationships between drowsiness and

observed data patterns and so they limit the theoretical understanding that can be gained through this modeling process.

These structures can also be compared to models structures that are specified by theoretical insight. Although some comparisons exist between these models which suggest generative models ultimately have lower error bound (Ng & Jordan, 2002), the ultimate performance of these any specific context is situation dependent. Therefore the identification of a sufficient structure for drowsiness detection is an empirical question. The question of structure selection depends in this case on the performance of the algorithm, as well as the algorithm's contribution to theoretical understanding of drowsiness detection.

2.5 Feature generation for time-series variables and driving context elicitation

The temporal and hierarchical structure of drowsy driving supports a shift in drowsiness detection algorithms from static machine learning to dynamic graphical models. The observed variables are a critical piece of the graphical model regardless of structure. These observations represent the input variables to the algorithm. In sequential data analyses these observations are often straightforward elements (e.g. a word) and these elements often provide a significant amount of clarity into the part of speech (e.g. "she" is always a pronoun). In time-series analysis the link between measured data and hidden states is less clear. For example individual steering measurements most likely do not provide a significant amount of information on the drowsiness level of the driver. The true relationship between steering and drowsiness is only evident in patterns of steering data that unfold over time (Krajewski & Sommer, 2009; McDonald, Lee, Schwarz, et al., 2013; Sayed & Eskandarian, 2001). Incorporating these patterns into an algorithm directly is problematic because the individual measures are often highly correlated. Furthermore, a feature vector based on such measures can become highly dimensional very quickly. For example a 60-second sample of steering data at 60 Hz would contain 3,600 features. This number of features requires an intermediate feature generation process between the measured data and the dynamic graphical model.

Time-series feature generation is part of a larger body of work known as time-series data mining (Keogh & Kasetty, 2003). Time-series data mining itself grew from database literature on indexing large time-series databases. Many of the time-series feature techniques were originally developed to reduce search times in database operations (Agrawal, Faloutsos, & Swami, 1993). The differences between these methods can be broadly characterized by the way they view time-series. In general the concept of time-series feature generation is to decompose the time-series into a set of components and then retain a small subset of these components that is significantly smaller than the length of the original time-series.

2.5.1 Distributional parameters

The basic approach to time-series feature generation essentially views the time-series as a distribution of samples. The parameters of this distribution can then be used as features to represent the distribution. This method of feature extraction is common in driving research that uses variables such as the standard deviation of lane position over a small time window. This approach is efficient and simple, but the conversion of time series to a distribution is problematic because it loses time information about the signal. For example from a mean, minimum, maximum, and standard deviation perspective a deceleration from 40 to 20 mph and a pattern of acceleration from 20 mph to 40 mph would look quite similar even though they are fundamentally different.

One method of compensating for this limitation is to incorporate non-linear distributional parameters. This approach is particularly relevant for noisy signals, such as steering wheel angle (Das, Zhou, & Lee, 2012). These parameters include the Lyapunov exponent (Wolf, Swift, Swinney, & Vastano, 1985), Sample Entropy (Richman & Moorman, 2000), and correlational measures (Grassberger & Procaccia, 1983). These measures are essentially measures of the regularity of a time-series however they differ in their calculation and quantification of regularity. Lyapunov exponent is calculated by comparing the trajectories of multiple points along the time-series. The Sample Entropy is calculated by computing the negative log of a ratio of counts of similar vectors of fixed size, where similarity is defined by having a distance less than a fixed threshold. Correlation measures are calculated based on a

correlation integral, which measures the mean probability that two states at two points in a time-series are close.

2.5.2 Orthogonal transforms

Orthogonal Transforms represent a time-series as a sum of simple functions such as sine waves (Agrawal et al., 1993). The idea of this perspective is that although all time-series can be decomposed into a set of simple functions that is equivalent to the number of samples in the data, the majority of the information about the time-series is stored in a few of these functions. Consequently the time-series can be reduced by simply decomposing the signal into simple functions, retaining a small subset of these functions and then representing this subset in some reduced feature space (Agrawal et al., 1993; Keogh, Chakrabarti, Pazzani, & Mehrotra, 2001). An example of this process is shown in Figure 7, which shows the Fourier decomposition of a time-series. Fourier Decomposition uses sine waves as the simple decomposition function. In the figure, the top plot shows the original time-series and the bottom three plots show a set of three sine waves that convey the majority of the information in the original time-series. Each of these sine waves can be represented as a phase and magnitude pair, reducing hundreds of data points to six.

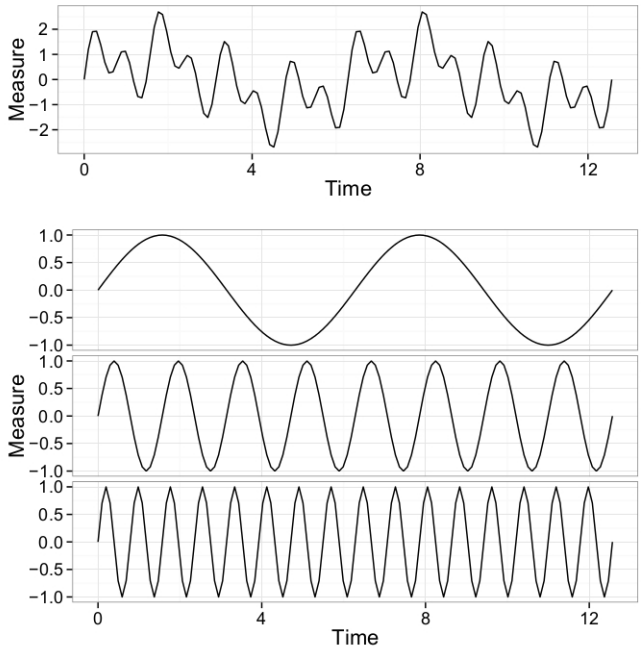

Figure 7 Demonstration of Fourier Decomposition of a signal. The top plot shows the original signal and the bottom three plots show the first three Fourier bases.

Fourier Decomposition (FFT) represents one of the most common orthogonal transforms. Another common transform is the Haar Wavelet transform (DWT), which represents time-series as a sum of square wave functions (Chan & Fu, 1999). Both of these methods are advantageous relative to distributional parameters because they maintain temporal information of the signal however they also suffer from several limitations. FFTs are limited because they assume stationarity and they are not localized in time. Stationarity refers to the fact that the distribution of the time-series, in other words the mean and variance, does not change over time. This assumption causes FFTs to perform poorly in the reduction of time-series like stock prices that have a positive or negative trend in their mean. Time localization refers to the fact that FFT components are fit to an entire signal at once. This broad fitting causes FFTs to struggle when capturing signals such as an impulse function, which have few peaks and are not periodic. The primary disadvantage of DWTs is that they are only defined for time-series of lengths equivalent to integral powers of 2 (i.e. 2^n). Furthermore DWTs are not as well supported as FFT in software such as R. Despite the limitations of FFT, the strict definitions of DWT and the relative

similarity between DWT and Symbolic Aggregate Approximation, described in the following section, leads this dissertation to focus on FFT and Symbolic Aggregate Approximation.

2.5.3 Symbolic Aggregate Approximation

Symbolic Aggregate ApproXimation (SAX) is a feature generation method that converts continuous time-series data to a series of letters via a time-series data reduction technique called Piecewise Aggregate Approximation (PAA). Piecewise Aggregate Approximation represents a time-series as a series of means or constant values. This conversion is accomplished by dividing the time series into equal sized windows, taking the mean of the samples within each window, and then creating a reduced representation of the original time series by simply concatenating the means into a vector (Keogh et al., 2001). SAX is a modification to PAA that converts the output of PAA from a series of constants to a series of letters. This conversion is accomplished by adding additional steps to PAA, which bin the y-axis into quantiles of the normal distribution, assign a letter to each quantile, and then convert the PAA output into letters based on each mean's associated y-axis quantile. This process is shown in Figure 8, which demonstrates the conversion of a continuous speed signal to the word, "dfghhhhii." The primary limitations of SAX are the loss of the true magnitude of the signal and that SAX requires specification of three parameters: the size of the window, the size of the alphabet, and the normalization constant.

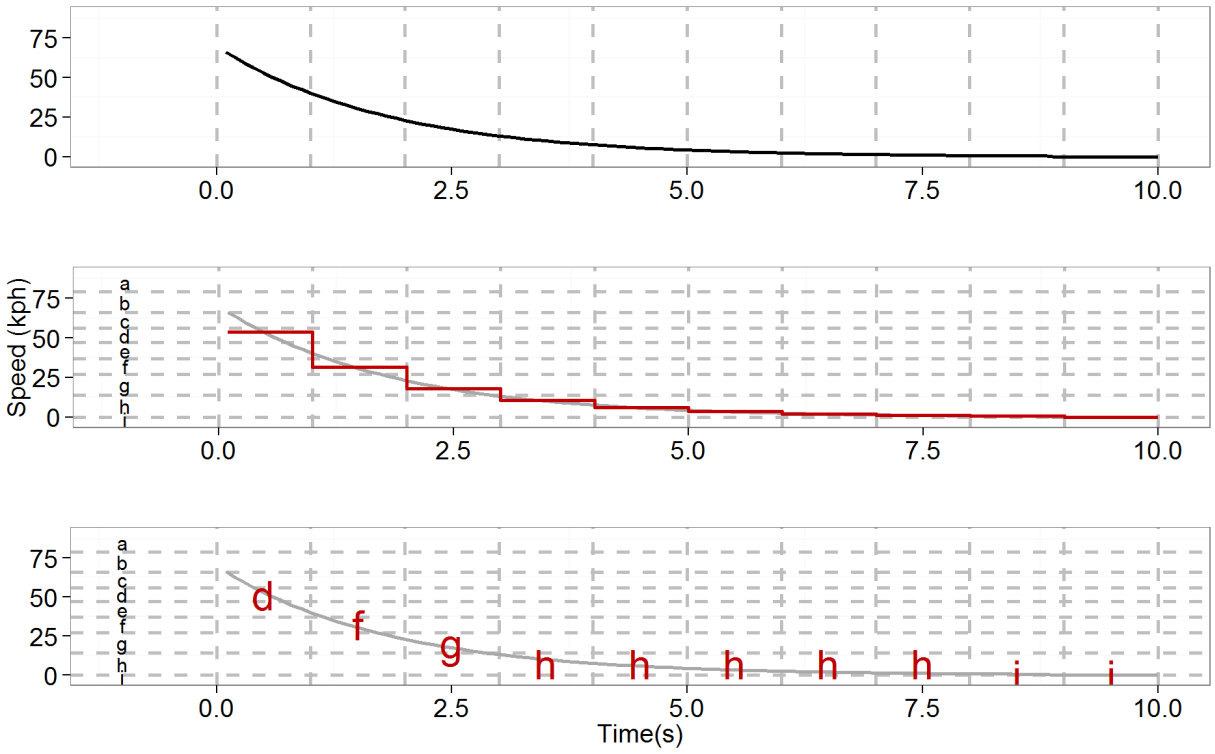

Figure 8 Demonstration of SAX steps for a 10 second sample of data.

2.5.4 Integrating feature generation and drowsiness detection

There are two areas in which the feature generation process is important to drowsiness detection algorithms. The first of these areas is in the integration of performance-related variables, specifically steering wheel angle and pedal input, into the algorithm. The second is the integration of context, through speed and acceleration measures. In both cases the measures represent time-series data and so all of these techniques are applicable. The key difference between the two cases is that the context is part of a larger hierarchical structure. The feature generation technique used to capture context must be generalizable to a hierarchical structure. Like the selection of graphical modeling structure, the choice of feature selection is data dependent (Keogh & Kasetty, 2003), and therefore an empirical process.

Empirical feature selection can be accomplished with one of two broad processes and one of three starting conditions. The two processes are wrapper and filter. Wrapper feature selection evaluates features based on their impact on algorithm classification whereas filter methods evaluate features based on a heuristic, such as information gain, prior to algorithm training. The three starting conditions are forward

selection, backward elimination, and random evaluation. Forward selection methods start with one feature in the model and add features in each iteration as appropriate. Backward elimination methods start with all of the features in the model and iteratively remove features at each step. Random selection methods evaluate random subsets of the features. An empirical feature selection process must select a feature evaluation process, a starting condition, and decide on the types of features to be included in the model.

2.6 Chapter Summary: Insights, gaps in current work, and research questions

Drowsy driving is a complex phenomenon that is currently surrounded by uncertainty. One certainty with drowsy driving is that it frequently results in fatal or injurious run-off road crashes. The frequency of these crashes is related to sleep disorders, time into the drive, sleep hygiene, driving on rural highway environments, and may be related to circadian rhythms. However the true relationship between drowsy driving and these elements is poorly understood. Furthermore very little is known about the relationship between drowsiness and contexts in which crashes rarely occur. This ambiguity when paired with statistics that suggest that the rate of drowsiness-related crashes has not been significantly affected by educational and duty restrictions suggests a need for new drowsiness mitigation solutions.

The most promising approach to preventing drowsiness-related crashes is drowsiness mitigation systems. These systems are a form of automation that collects driving data, and provides feedback to the driver-based on the output of a detection algorithm. The detection algorithm is a critical component of these systems because its reliability is directly tied to driver's trust and use of the mitigation system. Algorithms with high false positive rates are likely to lose driver trust and be misused. Current detection algorithms focus on a wide variety of measures, ground truth definitions of drowsiness, and internal machine learning algorithms. The majority of these algorithms view drowsiness as static phenomenon, and all of these approaches view drowsiness detection as a data pattern detection problem.

Analysis of theoretical driver behavioral models suggests that current drowsiness detection approaches are limited because time and driving context are central components of driving. Context is a complex concept that encapsulates road condition, interaction with other drivers, and maneuvers. It is

hierarchical in nature and unfolds over several time-scales. Theoretical models of driving suggest that speed, and by extension acceleration, depend on and are highly sensitive to context. Therefore these variables may be used to capture driving context in a detection algorithm.

The importance of time for drowsiness detection is also supported by literature on human behavior detection. This literature develops models of time-series human behavior and classifies it into actions or intentions. The most common type of models used in this framework is probabilistic graphical models. The prevalence of these models in the behavior detection literature suggests that they should be adapted for drowsiness detection. Probabilistic graphic model development requires two steps: model structure selection and model parameter learning. The structure selection process is empirical but it can be guided by current structures and theoretically defined structures. The parameter-learning task is dependent on feature selection. Feature selection for drowsy driving variables involves an evaluation of context variables such as speed, and performance variables such as steering wheel angle. All of these data are time-series data and so features representing these data can be derived from them using time-series feature generation methods. Like graphical modeling structures the choice of time-series feature generation method is empirical.

The lack of a thorough understanding of drowsiness, limitations in current algorithms, and the empirical nature of probabilistic graphical model structure and feature selection prompt the following research questions.

1. How can dynamic graphical modeling structures provide an accurate and realistic characterization of the temporal nature of drowsy driving?
2. What is the most appropriate method for generating contextual driving features and integrating these features into a dynamic graphical model in real time?
3. How can these contextual driving features be integrated into a hierarchy that represents the hierarchical nature of driving behavior and integrates into a dynamic graphical model?

The goal of this dissertation is to iteratively answer these three questions over three empirical studies. The remainder of this dissertation will discuss the data used in this work, provide methodologies, results, and

conclusions for the three studies, describe a set of additional analyses, and conclude with general implications of these results for detection algorithms, driving safety research, and the field of Human Factors.

Chapter 3 Driving simulator study of drowsiness

The research questions posed in chapter 2 require empirical investigation with drowsy driving data. This dissertation will employ a dataset collected at the National Advanced Driving Simulator (NADS) located on the campus of the University of Iowa (NADS, 2010). The NADS simulator is a high fidelity driving simulation environment consisting of a 24-foot dome containing a full Chevy Malibu sedan surrounded by 360 degrees of screens. The dome is located on a full motion base that provides 400 meters of lateral and longitudinal travel and 330 degrees of rotation in both directions. During a simulation a driving environment is rendered on the screens surrounding the vehicle. As the driver progresses through this environment the motion base provides real-time haptic feedback. The motion-base structure and the interior of the dome are illustrated in Figure 9 and Figure 10 respectively.

Figure 9 Dome and motion base structure of the NADS II simulator. Reprinted from Brown et al. (2011).

Figure 10 Illustration of a projected driving environment on the inside of the dome. Reprinted from Brown et al. (2011).

The goals of the study were to evaluate algorithms designed to predict alcohol impairment for drowsiness detection, and to develop new algorithms for drowsiness detection. Many of the study design

decisions were influenced by a previous study focused on the development of algorithms to detect alcohol impairment (Lee et al., 2010). However most of these influences do not impact the study relationship to the population of drivers or the realism of the scenario. This chapter will describe this study in detail and provide a basis for algorithm evaluation.

3.1 Study participants

Data were collected from seventy-two subjects as they completed three drives of approximately thirty minutes each. The three drives occurred during the daytime, late evening, and early morning hours respectively. The drivers were healthy men and women from one of three age groups: 21-34, 38-51, 55-68. All participants possessed a valid United States driver's license and were recruited from a participant database maintained by the NADS. Participants were compensated $250 for participating in the study or a prorated amount if they did not complete the study.

Prior to the start of the study, participants were screened for health history, current health status, and status as a morning or evening person. Participants who were pregnant, suffered from chronic disease or a sleep disorder, displayed evidence of substance abuse, and those who took prescription medication that cause or prevent drowsiness were excluded from the study. Additionally participants who strongly tended toward wakefulness in the late evening hours were also excluded from the study. In addition to good health, participants had to have driven a minimum of 10,000 miles per year for the past two years, live within a 30-minute drive of the NADS office, and have sleep patterns in which they slept and woke at approximately the same time every day. On the day of each drive participants were screened for drug use and alcohol consumption, and were removed from the study if they tested positive. Participants sleep was also monitored using an Actigraph and participants who slept less than six hours on the night before the drive were excluded.

3.2 Simulator and data collection

Data were collected continuously throughout each drive from the NADS simulator and a dashboard-mounted eye-tracker (Face Lab™ 5.0, Seeing Machines, Canberra, Australia). The simulator collects a

thorough record of vehicle state (e.g. lane position and speed) and driver inputs (e.g. steering wheel position and accelerator pedal position) originally sampled at 240 Hz. The eye-tracker records a variety of eye measures including gaze and closure, sampled at 60 Hz. In addition to these sensors, EEG data and driver postural data were collected. To balance the data and facilitate data sharing the driving data were down-sampled to 60 Hz prior to analysis.

3.3 Study procedure

The study period consisted of three separate visits to the simulator. The first visit was a screening visit that involved signing of informed consent, drug and alcohol testing, a physical, a training presentation, and an orientation drive in the simulator. The orientation drive lasted approximately 8 minutes and required the participant to successfully complete a left hand turn, and drive on several different types of roads. The drive was guided by a pre-recorded series of audio navigational instructions. After the instructional drive participants were evaluated for symptoms of simulator sickness and removed from the study if they demonstrated significantly high scores for post-drive nausea.

The second and third visits consisted of either a single daytime drive or two drives during the late evening (Early Night condition) and early morning (Late Night) respectively. The order of these days was counterbalanced across participants. These visits were separated by a period of at least three days. During the daytime visit, participants drove themselves to and from the simulator. During the evening visit participants were picked up at their homes after dinner (at approximately 7pm) and driven home after completion of the study. In both visits participants' actigraphy data and Blood Alcohol Composition (BAC) were evaluated for compliance upon their arrival to the simulator. Prior to the start of each drive, participants drowsiness was evaluated with the Stanford Sleepiness Scale (SSS; Hoddes, Zarcone, Smythe, Phillips, & Dement, 1973) and a modified Psychomotor Vigilance Test (PVT; Wilkinson & Houghton, 1982). The SSS is a subjective 1-7 rating scale of drowsiness and the PVT is a visual stimulus haptic response task that measures reaction time. After each drive participants completed the SSS and PVT a second time and also rated their drowsiness during various points in the drive using a

Retrospective Sleepiness Rating. In the daytime session participants were allowed to return home after completion of the post-drive surveys. In the night-time session participants were kept awake between the drives via study personnel intervention and various activities including television and books. Participants were returned home after the completion of the second drive during this session.

The daytime drive began between 9:00 and 12:00, the early night drive began between 11:00 and 1:00, and the late night drive began between 2:00 and 5:00. The late night drive occurred after a minimum of 18 hours of continuous wakefulness. Each drive consisted of three connected segments representing urban, highway, and rural environments. The drive began with an urban segment on a two-lane roadway with posted speed limits between 25 and 45 mph. This segment contained several controlled and uncontrolled intersections along with a series of potential hazards which consisted of other vehicles, motor bikes, and pedestrians entering the roadway. Following the urban segment drivers entered a four-lane divided expressway with a posted speed limit of 70 mph. During this segment drivers followed a lead vehicle and then overtook a series of slower vehicles. After exiting the highway, participants drove on a rural, undivided, two-lane road that culminated in a 300 second period of continuous straight driving. The goal of this drive was to simulate a drive home from an urban parking spot to a rural home location. The drive was approximately 35 minutes long and contained realistic driving surroundings throughout. To balance a potential learning effect three slightly different driving scenarios were used one for each drive. The general structure of these scenarios was the same but the order of turns and other events were reordered. The contents of each drive were categorized into events associated with a specific event name. The event names and descriptions are presented in Table 2.

Table 2 Event descriptions for the simulator study.

Road Type	*Event Name*	*Description*
Urban	Pullout	Pull out of parallel parking space
	Urban	Driving on a narrow urban road with parked cars on both sides
	Green	Navigating through a green light at a controlled intersection
	Yellow	A yellow light dilemma at a controlled intersection
	Left	A left turn at a controlled intersection
	UrbanCurves	Navigating a series of curves on an urban two-lane road with cars parked on both sides
	OnRamp	Navigating a highway entrance ramp
Highway	MergeOn	Merging on to the highway
	Interstate	Driving behind a slow moving vehicle on the highway
	MergingTraffic	Driver approaches an interchange with a vehicle merging about 500 feet ahead of the on-ramp
	InterstateCurves	A series of three curves the driver must negotiate on the interstate with light traffic
	ExitRamp	Navigating a highway exit ramp
	TurnOffRamp	Right turn from the off-ramp onto a rural two-lane undivided road
Rural	Lighted	Driving on a lighted two-lane rural road with a speed limit of 55 mph
	TransToDark	Transition to a segment of the rural road that is unlighted
	Dark	Driving on a rural, two-lane, unlighted 55 mph road with faded lane lines involving some curves
	TransToGravel	Turn slightly to the right onto a gravel road
	Gravel	Driving on the gravel road
	Driveway	Pulling in to a gravel driveway
	GravelRuralExt	Navigating an unlighted gravel rural road that contains a series of curves
	GravelTransToStraight	Transitioning to a straight segment of gravel road.
	PavedTransToRuralStraight	Driving on a transition from a gravel road to a paved rural road
	RuralStraight	Navigating an unlighted paved rural road for 10 minutes
	DarkNoHairpin	Driving in the dark on a straight rural road
	DarkHairpin	Driving in the dark on a hairpin turn on a rural road

3.4 Data and manipulation validation

After the completion of data collection, driving and eye-tracking data were consolidated into a single dataset and the data were verified with a series of automatic checks and visualizations. The data verification step ensured that no data were missing and the size of the dataset was reasonable.

Furthermore each variable was evaluated using three conditions: whether the data lay within a reasonable range, varied in a meaningful manner, and varied continuously.

The data validation was followed by an evaluation of learning effects. Learning effects were evaluated with through analyses of lane deviation, mean speed, and speed deviation during each event in the drive. In general performance did not change across the three drives, 60 of the 75 comparisons revealed no significant changes across the drives. Significant changes in speed accounted for ten of the significantly different comparisons. In each of these comparisons the mean speed increased by at most 4 mph. The other five significant comparisons were related to lane deviation although in four of the five comparisons, lane-keeping performance decreased as experience increased. The only exception to this was lane keeping on a brief segment of the rural road. Thus learning effects did not have a significant impact on the participant performance.

The experimental manipulation was evaluated using cumulative time awake (CTA) measured in minutes, Stanford Sleepiness Scores, PVT reaction times measured in milliseconds, Retrospective Sleepiness Scores (RSS), and the frequency of drowsy-related lane departures. Drowsy-related lane departures were defined through video analyses of all lane departures. During this analysis erroneous lane departures caused by simulator anomalies and visual-manual distraction were removed. Each departure was manually coded using the Observer Rating of Drowsiness (ORD) scale. The ORD scale is a continuous rating between 0 and 100, based on the 60 s preceding each lane departure (Wierwille et al., 1994). In this case the ORD measures were separated into five bins: not drowsy ($ORD < 12$), slightly drowsy ($12 \leq ORD < 37$), moderately drowsy ($37 \leq ORD < 62$), very drowsy ($62 \leq ORD < 90$), and extremely drowsy ($ORD \geq 90$). Departures classified as moderately, very, or extremely drowsy were labeled “drowsy.” The results for CTA, SSS, and PVT are illustrated in Table 3. The results for RSS and drowsy-related lane departures are illustrated in Figure 11 and Figure 12 respectively.

Table 3 Summary statistics for cumulative time awake, measured in minutes, Stanford Sleepiness Scores measured on a 7 point Likert scale, and PVT response times measured in milliseconds. Adapted from (T. Brown et al., 2011).

	Daytime Drive			*Early Night Drive*			*Late Night Drive*		
	Mean	SD	Med	Mean	SD	Med	Mean	SD	Med
CTA	223	73	214	1001	53	995	1230	51	1228
SSS Pre-drive	1.8	0.8	2.0	3.4	1.2	3.0	5.0	1.3	5.0
SSS Post-drive	2.9	1.2	3.0	4.1	1.3	4.0	5.4	1.3	6.0
PVT Pre-drive	371	44	364	397	53	386	430	62	419
PVT Post-drive	394	52	387	412	58	414	460	74	448

The results from Table 3 demonstrate that as expected, all three measures of drowsiness increased from the daytime drive to the early night drive and from the early night drive to the late night drive. Furthermore all three drive conditions demonstrated an increase in both SSS and PVT from pre-drive to post-drive measures, which suggests that the drive induced some level of drowsiness. Both of these findings are confirmed with the RSS and drowsy-related lane departure data. Interestingly drowsy-related lane departures were observed in all three scenarios. This finding seemingly suggests that the inclusion of time on task and circadian measures alone cannot account for all dangerous drowsy driving effects. Furthermore there is a clear connection between the on-road events and the frequency of drowsy-related lane departures. Together these findings validate the approach of integrating temporal and contextual factors into drowsiness detection algorithms designed to detect consequences of drowsy driving.

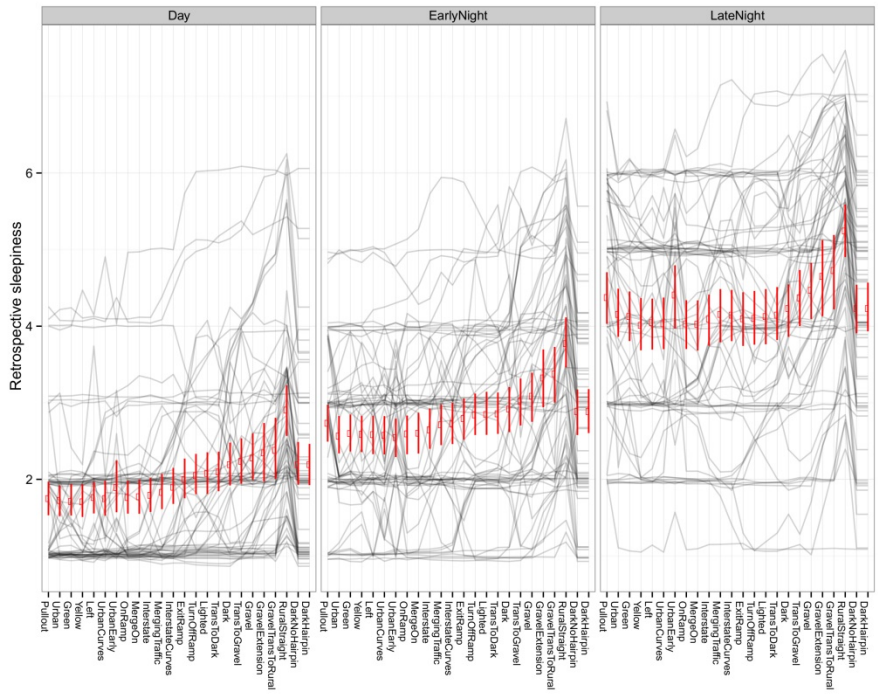

Figure 11 Retrospective sleepiness scores for each driving condition and each driving event. Reprinted from Brown et al. (2011).

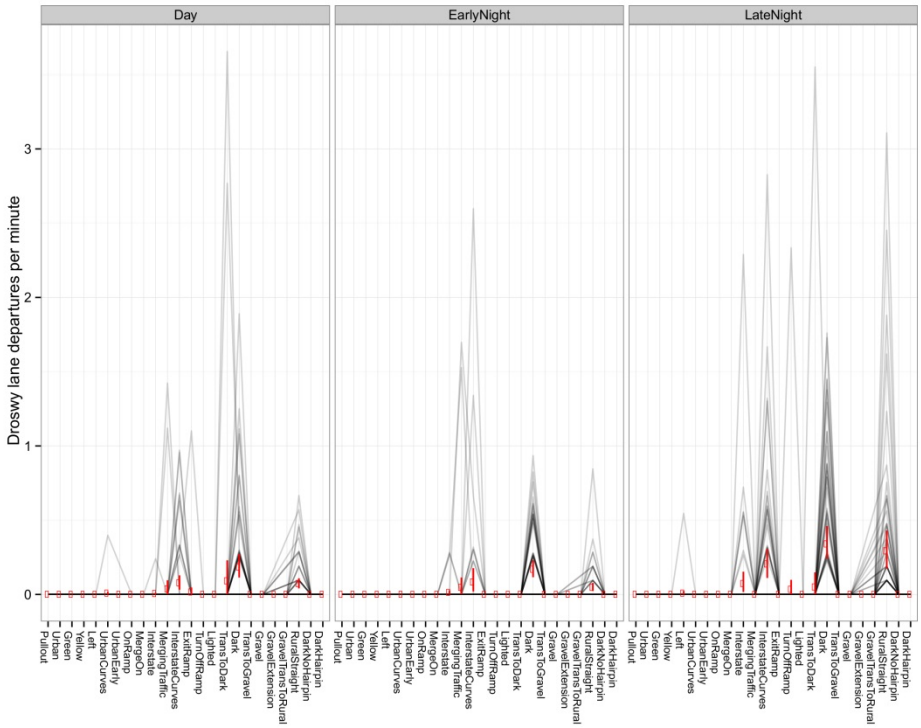

Figure 12 Drowsy-related lane departures per minute for each driving condition and driving event. Reprinted from Brown et al. (2011).

3.5 Data limitations and scope

The manipulation evaluations suggest that the simulator study adequately captured drowsiness over a variety of situations. Despite this adequacy it is important to acknowledge that this dataset has some limitations that in turn have implications for algorithm design and development. The cleanliness of the data collection process essentially removes effects from work schedules, and effects of long-term poor sleep hygiene. Additionally, the inclusion criteria limit the conclusions of this work with respect to professional drivers, individuals with sleep disorders, and those that take prescription medications. Consequently the variables representing these pre-conditions are not included in the algorithm design. Despite these limitations, the data clearly indicate drowsiness inducement in a large sample approximation of the healthy adult population in the United States and therefore are valid for algorithm development.

Chapter 4 Methodology for improving drowsiness detection algorithms

There are four elements of all drowsiness detection algorithms, the ground truth definition of drowsiness, the input measures, the feature generation technique, and the internal machine learning algorithm used to make classifications. The research questions posed by this dissertation empirically address these elements of algorithm design with a particular focus on the machine learning algorithm. To compare algorithms, one must hold some elements of the drowsiness detection algorithm constant. The goal of this chapter is to specify a set of constraints on the algorithm development process that will facilitate comparisons between the models developed in subsequent chapters. Specifically this chapter discusses the partition of data into training and testing datasets, the input measures, and the ground truth definition of drowsiness, the feature generation and selection process, the model evaluation criteria, and also introduces a set of benchmarking comparison models.

4.1 Training and testing data configuration

A central goal of drowsiness detection algorithms and machine learning models in general, is generalizability, the capability to accurately predict unseen data (Mitchell, 1997). One method of analyzing this generalizability is to assess a trained algorithm's ability to predict data that is withheld from the training process. This training and testing also provides a method of measuring overfitting, a condition where the model produces highly accurate predictions due to capturing anomalies in the training data rather than the underlying concept (Mitchell, 1997). The process of partitioning data into training and testing data sets requires a balance in retaining enough data to sufficiently train the algorithm while having a meaningful testing data set. The study data consist of 216 total drives from 72 different participants, representing approximately 160 hours of driving data. The simplest method of partitioning these data into training and testing sets consists of randomly partitioning the individual drives between the training and testing sets based on a threshold. However the goal of this work is to develop a drowsy detection algorithm that can be generally applied to all drivers. In a typical use scenario the model would be tasked with predicting drowsiness in a driver who was not included in the dataset. To simulate this task

the data for these studies were partitioned by driver, with approximately 90% of drivers (65) used in training and 10% (7) drivers used for testing.

Drivers were partitioned into training and testing through a clustering and random selection process. The goal of this selection process was to loosely define types of drowsy drivers and then assign at least one driver of each type to the testing set. Drivers were clustered on a set of demographic variables, driver level aggregation of drowsiness measures, and drive level aggregation of drowsiness measures. The complete set of variables, which includes 42 total variables, is described in Table 4. This set of variables was selected to emphasize the effect of drowsiness on drivers in the partitioning while also taking into account demographic variables that might induce bias in the testing data sample. For example, regardless of the effect of drowsiness across drivers, a model tested on only older female drivers could not be expected to generalize across the population.

Table 4 List of variables used in the driver clustering process.

Type of factor	*Factor*		
Demographics	Age		
	Gender		
	Order of Drives		
Driver level drowsiness	Mean RSS score		
	Max RSS score		
	Min RSS score		
	Mean pre-drive SSS		
	Mean post-drive SSS		
	Mean pre-drive PVT		
	Mean post-drive PVT		
	Mean CTA		
	Total drowsy-related lane departures		
Drive level drowsiness	Day drive mean RSS score	Early night drive mean RSS score	Late night drive mean RSS score
	Day drive max RSS score	Early night max RSS score	Late night max RSS score
	Day drive min RSS score	Early night min RSS score	Late night min RSS score
	Day drive mean pre-drive SSS	Early night mean pre-drive SSS	Late night mean pre-drive SSS
	Day drive mean post-drive SSS	Early night mean post-drive SSS	Late night mean post-drive SSS
	Day drive mean pre-drive PVT	Early night mean pre-drive PVT	Late night mean pre-drive PVT
	Day drive mean post-drive PVT	Early night mean post-drive PVT	Late night mean post-drive PVT
	Day drive mean CTA	Early night mean CTA	Late night mean CTA
	Day drive total drowsy-related lane departures	Early night total drowsy-related lane departures	Late night total drowsy-related lane departures

The dataset was missing at least one of the subjective drowsiness evaluations (RSS, SSS, or PVT) from a total of 10 drivers. The clustering process was performed with Clara clustering, because it can handle missing data (Kaufman & Rousseeuw, 1990). The final number of clusters was selected by comparing the isolation, which a measure of the difference between clusters, between a range of 2 and 15 clusters. Six clusters were found to minimize this value. Table 5 shows the final membership for the six clusters. To reconcile the 6 clusters relative to the 7 drivers required in the testing data for a 90% split, two drivers were included in the final testing set from the first cluster.

Table 5 Cluster membership by drivers for the training and testing data cluster partitioning.

Cluster	1	2	3	4	5	6
Number of drivers	46	7	7	4	5	3

The final testing dataset also contained a representative from every experimental condition, reducing the likelihood of confounding with the trained models. Table 6 shows a demographic summary of the final training and testing datasets. This testing dataset contains 4 female drivers and 3 male drivers, ranging in age from 22 to 57. Five of the drivers began data collection with the early and late night drives and two drivers began with the daytime drive. Three of the clusters are represented by drivers without a drowsy-related lane departure, suggesting that the effects of drowsiness did play a role in the clustering partitions. The testing data were withheld from all feature, parameter, and model selection processes and were only used for evaluation purposes.

Table 6 Demographics and drowsy-related lane departure frequency for the drivers in the test dataset.

Subject	*Age category*	*Age*	*Gender*	*Order*	*Cluster*	*Total Drowsy-related lane departures (DLD)*
A_M1024YF06	Y	22	F	Night First	1	8
A_M1065OF03	O	55	F	Day First	1	25
A_M1153OM03	O	57	M	Day First	2	3
A_M1060OM06	O	57	M	Night First	3	0
A_M1069OF05	O	56	F	Night First	4	0
A_M1036MM06	M	47	M	Night First	5	13
A_M1244MF04	M	50	F	Night First	6	0

4.2 Ground truth definition of drowsiness

The choice of the ground truth definition of drowsiness is critical to the success of a drowsiness mitigation system. The ground truth is essentially equivalent to the maximum amount of information available to communicate to a driver. For example, if an algorithm uses time of day as the ground truth, the algorithm can only warn a driver about the time of day. In this way the choice of ground truth ultimately influences the driver's acceptance and use of the mitigation system (Balkin et al., 2011). This work defines ground truth drowsiness as drowsy-related lane departures. Specifically drowsy-related lane departures identified by the simulator that have a corresponding ORD rating of at least 37 out of 100. These drowsy-related lane departures are different from a routine lane departure, which could be caused

by inattention, correction failures, or other impairments, in that they are preceded by a one minute period of driving in which the driver displayed visible drowsiness indicators such as eye closures, yawning, head nodding, and slumping posture. Drowsy-related lane departures are a sound measure of ground truth because they are generalizable to all drivers, are easy to communicate to drivers, are clearly related to the consequences of drowsiness, and might motivate drivers to pursue action.

The applicability of warning to all drivers regardless of demographic factors is a critical element of an acceptable mitigation system (Dinges & Mallis, 1998). Unlike more general ground truth measures such as cumulative time awake or time on task, that have varied effects across drivers, the definition of a drowsy-related lane departure is constant across drivers. Furthermore the concept of a lane departure is tangible rather than a subjective measure of a complex construct, meaning the occurrence of a lane departure is simpler to communicate to drivers compared to the general malaise of drowsiness. This tangible outcome will facilitate understanding between the driver and the system which in turn will promote trust between drivers and the system (Balkin et al., 2011; Ghazizadeh, Peng, Lee, & Boyle, 2012; Lee & See, 2004). The relationship between the driver and the mitigation system will also likely be supported by the fact that drowsy-related lane departures are directly related to the consequences of drowsiness. Furthermore given that the primary type of crash attributed to drowsiness is the single car road departure (Dinges, 1995; Pack et al., 1995; Sagberg, 1999), and the fact that this type of crash is always preceded by a drowsy-related lane departure, it is reasonable to expect a reduction in crashes with a successful drowsy-related lane departure-based mitigation system. This direct connection to safety critical events subverts the problem of deciding the precise point at which drowsiness becomes dangerous, which is a critical concern with previous algorithms (MacLean et al., 2003).

One limitation of drowsy-related lane departures is that they occur on the order of seconds, whereas the patterns of driver behavior associated with them often occur on the order of minutes. Therefore the detection algorithm developed in this work will use windows of driving behavioral data to predict windows of time surrounding a drowsy-related lane departure. The challenge of this method is that a drowsy-related lane departure can occur at any point within a window of data. In some cases a drowsy

driver may recognize the lane departure and subsequently apply corrective steering or braking to return to their lane. In this scenario the driver is truly only drowsy for the portion of the drive prior to corrective activity. Therefore it is critical to exclude windows that include a correction. Figure 13 shows the steering and brake pedal input surrounding the drowsy-related lane departures observed in the training data. The plots in the figure consist of steering behavior (left) and braking behavior (right) partitioned by the corrective action preceding and following the 10 s on either side of the lane departure. The corrective actions are divided into four categories: none, corrective action after the lane departure, correction before the lane departure, and correction both before and after the lane departure. These four categories represent the four plots on each side of the figure arranged clockwise. Interestingly of the 105 total departures observed 67% (70 departures) did not include any sort of corrective action. The next most common corrective actions were a combination of braking before and after the departure and steering after the departure which occurred in approximately 9% (10) of the cases, and a combination of steering and braking before and after the departure which occurred in nearly 7% (7) of the cases. No other combination of steering and braking corrective activity occurred more than four times in the training data.

These results were incorporated into the ground truth definition of drowsiness by introducing a conditional definition of drowsiness. This conditional definition included all windows containing a drowsy-related lane departure without corrective input and excluded all instances where corrective input was applied immediately following a departure and those where corrective input was provided immediately preceding a departure. Windows with corrective input both before and after the departure were defined as drowsy only when the last sample contained the departure. This inclusion was driven by the hypothesis that the behavior in these windows represents responses to road context rather than corrective input. The goal of this inclusion is to capture drowsy related departures that occur on road geometries that naturally require some type of vehicle input.

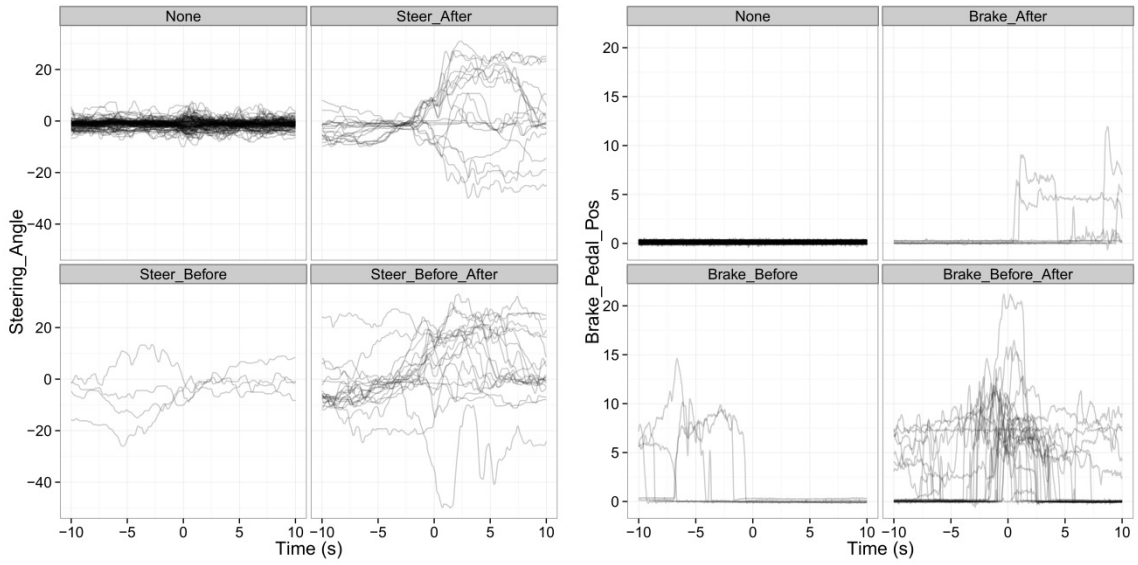

Figure 13 Steering and brake input in the 10 seconds surrounding a drowsy-related lane departure faceted by the corrective action. A braking action is defined as a pedal input of greater than 1 degree and steering action is defined as a steering wheel angle greater than 5 degrees. Note that the plots are centered on the lane departure and the lack of corrective action is significantly more frequent than any type of corrective action.

4.3 Measure selection

The link between driver behavior and drowsiness detection requires collecting driver behavioral measures, converting them into features and using the features in algorithm training. The difference between measures and features then is a processing step, i.e. steering angle is a measure, and mean steering angle over a 60 s window is a feature. There are two types of measures that are critical to the design process in this dissertation: those that capture behavior related to drowsiness detection and those that capture driving context. In both cases there are three general considerations in the measure selection process: reliability, cost, and intrusiveness. The most critical of these is the reliability of the measure. Reliability in this sense is measured in terms of the accuracy and consistency of the measure, the likelihood of missing data, and the relationship between the measure and ground truth. The cost of the measure is also critical, particularly when it is tied to reliability. For example, EEG measures may be reliable but not effective if they require an expensive measurement device. In a more general sense costs might also include battery life, particularly if the mitigation system is implemented on a cellular phone, although this is considered beyond the scope of this work. The intrusiveness of the measure is similar to

the ground truth in that it affects system acceptance. Even the most accurate system will not be used if it is uncomfortable for driver regardless of their life saving potential (consider seat belt adherence). With these general considerations, this work will use steering wheel angle and pedal inputs for drowsiness detection and speed and acceleration measures for driving context.

Steering wheel angle and pedal input are ideal measures for drowsiness detection because they are reliable, inexpensive, and naturally unobtrusive. In this case steering-wheel angle, illustrated in Figure 14, refers to the angle of deflection between the current location of the vertical center of the steering wheel and the vertical. Similarly pedal input (right side of Figure 14) is defined by the angle of the pedals relative to their initial position. Both of these metrics are advantageous relative to previous approaches such as EEG and eye-closure measurements because they do not require expensive sensors, are reliable, and are easy to implement in current vehicles (Balkin et al., 2011). Furthermore steering wheel angle measures have been employed in several successful algorithms in previous work (Krajewski & Sommer, 2009; McDonald, Lee, Schwarz, et al., 2013; Sayed & Eskandarian, 2001) and it is a more direct measure of driver input than driver performance variables, such as standard deviation of lane position, which is known to increase with drowsiness (MacLean et al., 2003). The primary limitation of steering-based algorithms is that they often have high false positive rates (Balkin et al., 2011), which could be an artifact of confounding between periods of inactivity caused by drowsiness, and periods of inactivity caused by the road environment (McDonald, Lee, Schwarz, et al., 2013). Pedal input data should compensate for this, as pedal input is often required in cases where steering input is not.

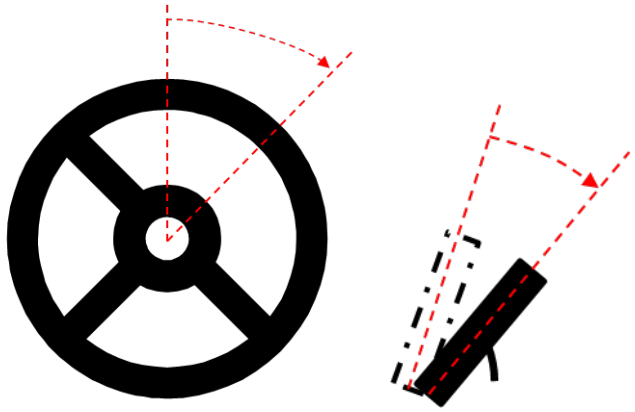

Figure 14 Steering wheel angle and pedal deflection.

Speed and acceleration measures are ideal for context measurement because they capture driving context in real-time in a reliable and cost-effective manner. The alternatives to speed and acceleration include a combination of GIS databases, weather data, and real-time traffic information. This combination introduces a substantial amount of data integration into the model and relies upon external databases that are often poorly maintained. Although connected vehicles promise to enhance the communication of this type of information in the future, these sources cannot capture low level driving maneuvers such as lane changes and turns. Speed and acceleration provide a direct link into these behaviors and the intentions of the driver (Fuller, 2005). Together speed, acceleration, steering, and pedal inputs represent robust, inexpensive, and unobtrusive measures of driver behavior capable of capturing driving context and likely drowsy driving behavior. The next step in the algorithm process is to integrate these measures into an algorithm through converting the measures to features.

4.4 Feature selection

Steering and pedal measures might capture behavior associated with drowsiness however it is difficult to directly incorporate them into a detection algorithm. This difficulty is due to the fact that instantaneous steering and pedal inputs do not disambiguate drowsy drivers from alert drivers, i.e. a drowsy and an awake driver are equally likely to produce any given single measurement. Therefore the measures must be aggregated over time windows. These windows of behavior can be incorporated into a drowsiness detection algorithm by generating features and conducting a feature selection process to find the most

appropriate features. This process necessitates a comparison between window sizes and features. This work evaluated three window sizes: 10 seconds, 30 seconds, and 60 seconds, and four types of features: Distributional parameters, non-linear distributional parameters, SAX based parameters, and Fourier features. The complete set of features is presented in Table 4. The distributional and Fourier features were selected because of the frequency of their use in previous drowsiness detection literature (Das et al., 2012; Jap, Lal, Fischer, & Bekiaris, 2009; Krajewski & Sommer, 2009; Lal et al., 2003). The SAX features were included due to their ability to capture patterns of data succinctly and their growing popularity in naturalistic driving research (McDonald, Lee, Aksan, et al., 2013a, 2013b; McLaurin et al., 2014). The remainder of this section presents the analysis of these features and explains the selection of the final feature set.

Table 7 Features considered in the algorithm development process. Note SAX features were generated only to the maximum window length, i.e. 10 s windows had at most 10 letters.

Feature Type	*Features*
Distributional features	Mean, Minimum, Maximum, Median, Standard Deviation, Skew, Kurtosis
Non-linear distributional features	Lyapunov Exponent, Sample Entropy, Correlation Entropy, Correlation dimension
Fourier Features	Phases and magnitudes of the first 10 components
SAX	SAX features consisting words of length 2, 5, 10, 30, 60, and alphabet sizes consisting of 3, 5, 7, and 9 letters.

4.4.1 SAX features

SAX features differ from the distributional and Fourier features in that they are discrete rather than continuous. Furthermore the SAX variables are in the form of natural language. As with other natural language analyses, the power of SAX features is subject to a tradeoff between the frequency of each observed word and the number of different words observed. As the SAX word length and alphabet size increase, the median frequency of SAX words across the whole data set decreases. The problem with this decrease is that it increases the likelihood of overfitting caused by a random observation of a given word in the training data. Alternatively SAX features generated from small alphabet sizes and word lengths might have so few observed words that they cannot effectively differentiate between drowsy and awake drivers. Figure 15 shows this tradeoff for the SAX features generated for this work. The figure suggests a distinct tradeoff between input settings that have few words and a large median frequency and those that

have many words but a low median frequency. This tradeoff is concerning however it is still worthwhile to evaluate at least some of the SAX features in detail.

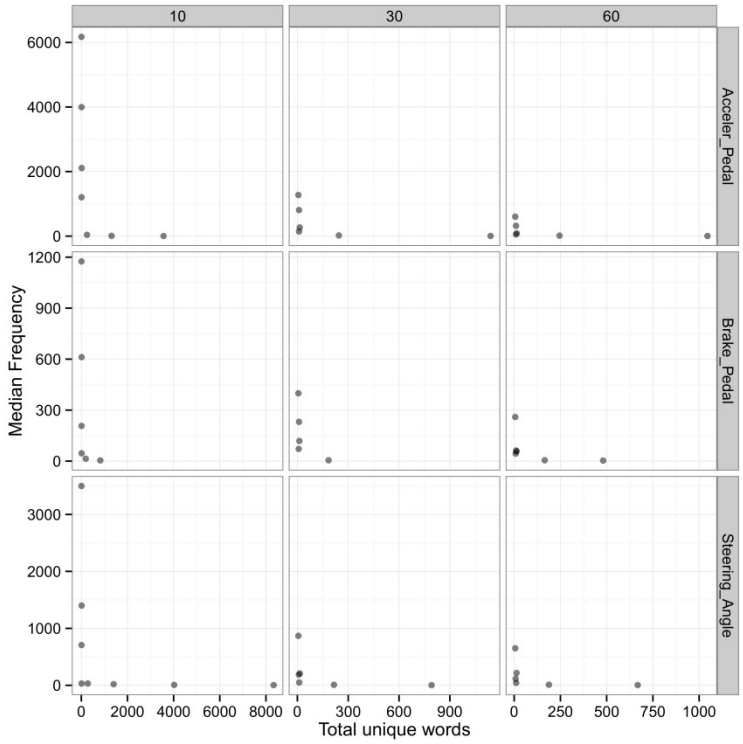

Figure 15 Median frequencies of words by total unique words for SAX features generated by window and measures.

In response to the results presented in Figure 15 all SAX features with a median frequency across the whole training data set of two or less were removed from further consideration. The remaining features were evaluated with a metric called information gain. Information gain is a measure of the value of knowing a certain piece of information, where value is measured by a reduction in uncertainty. For example, knowing if an individual has been up for more than 24 consecutive hours would have a higher information gain in drowsiness detection than knowledge of the same individual's hair color, because being awake for such a long time makes it more likely that the individual is drowsy and hair color likely has no effect. Information gain is on a scale of 0 to 1 where 0 represents a feature of no value and 1 represents a feature that removes all uncertainty from the data. In this case information gain is calculated based on the distributions of windows labeled drowsy and those labeled awake.

Figure 16 shows information gain for the ten SAX features with highest information gain. Interestingly features derived from the 60 s windows represent four of the top six information gain values including the three highest values. None of the features derived from the 10 s windows are represented in the figure. These results suggest that 60 s windows are necessary to capture steering and pedal behavioral patterns associated with drowsiness. The figure also shows that pedal and steering variables are equally represented among the best SAX features. This result validates the consideration of both steering and pedal variables in the algorithm.

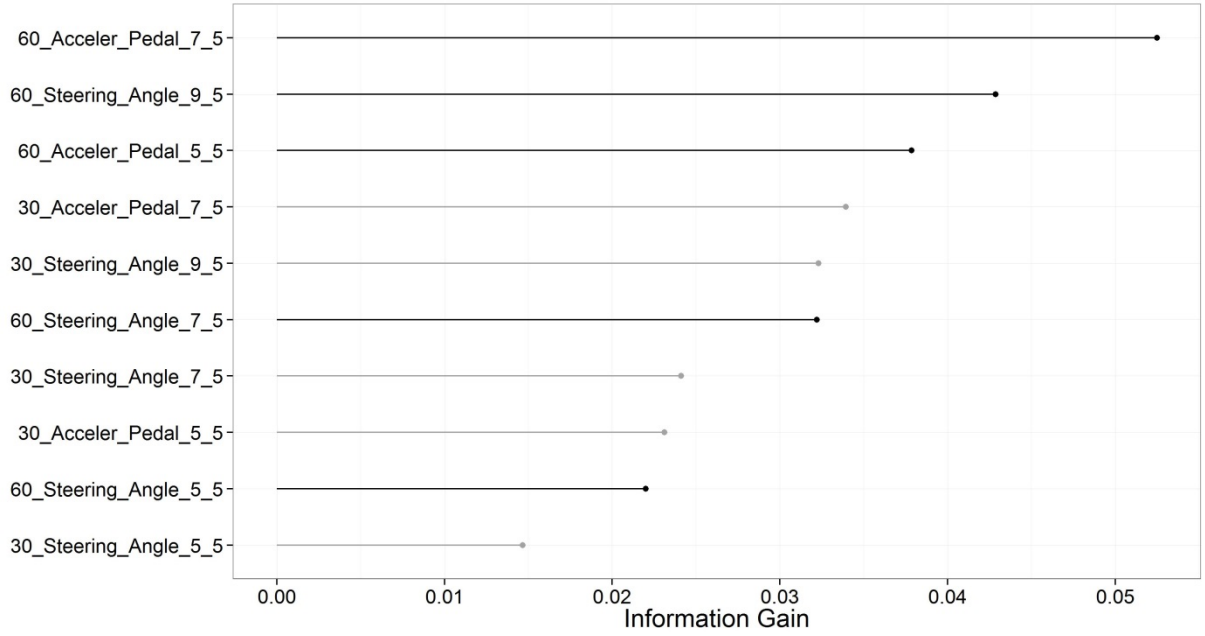

Figure 16 Pin plot of the ten SAX features with the highest information gain. Note the features are presented as: [window size] _ [measure] _ [alphabet size] _ [word length] and the color of the pins indicates the window size.

4.4.2 Continuous features

The distributional, non-linear distributional, and Fourier features are all continuous. Such features require a discretization step to calculate information gain. This discretization step consists of dividing the data into bins defined based on the range of the measure. The process is complicated because an exhaustive optimization of the number and size of bins is computationally infeasible. The simplest method of discretizing the data is to create a binary partition. The location of this partition can be selected by finding the partition that maximizes information gain. This method was used to estimate information gain for the continuous features—shown in Figure 17. As with the SAX features, the features derived from 60 s

windows have the highest information gain, yet all of the features have an order of magnitude lower information gain. This result is not surprising because these calculations allow for only one discrete split in the distribution. In contrast a SAX feature with an alphabet size of 7 and a word length of 5 would have approximately 7^5 "splits." However the generally low information gain with these features suggests that the use of a single continuous feature based on steering and pedal behavior over a time window will not be sufficient to capture drowsiness. This is not surprising particularly given the patterns of steering behavior surrounding drowsy-related lane departures illustrated previously. Many of these patterns had a mean of zero regardless of the type of correction however there was a much larger range of standard deviation, maximum, and minimum values, suggesting that only a combination of mean and one of the other distributional metrics differentiate the states.

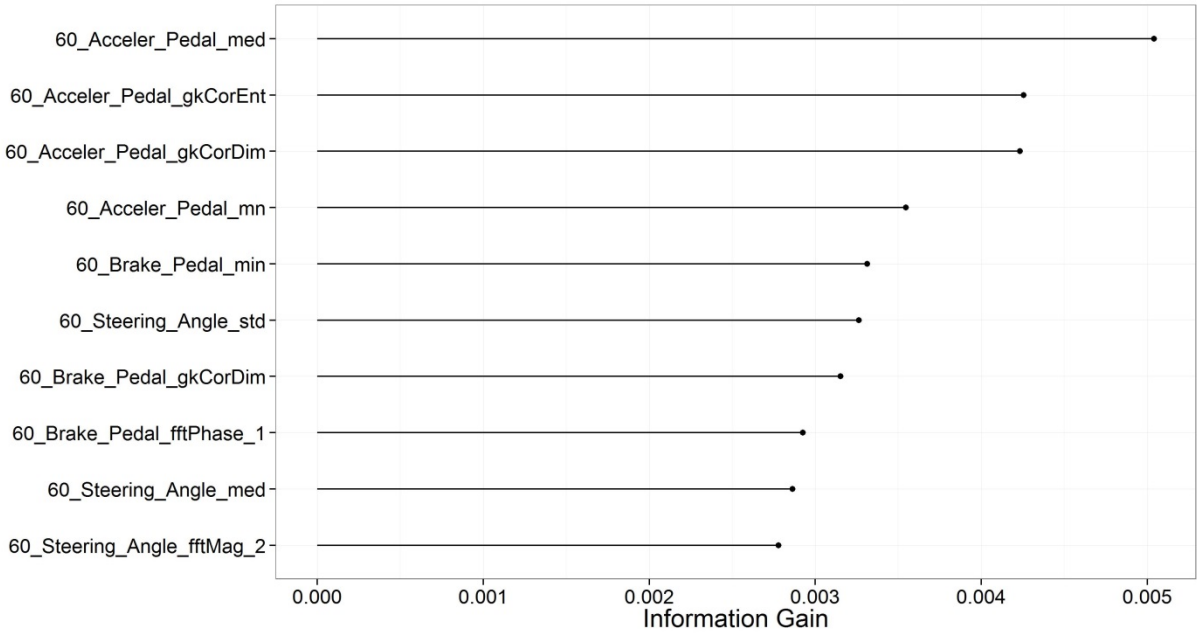

Figure 17 Pin plot of the top ten continuous features according to information gain. Note the feature descriptions are presented as: [window size] _ [measure] _ [feature] and any numbers following this description indicate the component.

4.4.3 Meta-features

One method of analyzing combinations of features as input is to create meta-features. In this case meta-features are defined as features derived from the output of a static machine learning algorithm that uses continuous features as input. The process of creating and selecting meta-features involves fitting a set of machine learning models on sets of continuous features of various window sizes. The models can be

evaluated using a cross validation with the training data. In addition to the basic process, this study used a sampling technique to address the class imbalance between the drowsy and awake data windows in the training data.

4.4.3.1 Addressing class imbalances

The strict definition of drowsiness in this case causes a severe class imbalance between drowsy and awake cases. This imbalance biases machine learning model fitting algorithms to produce null classifiers, or classifiers that predict "awake" for all instances. There are several methods for combatting class imbalance. Primarily these methods revolve around sampling the classes in the training data, i.e. over sampling the drowsy training instances to match the number of awake instances, down-sampling the awake training instances to match the number of drowsy instances, or a hybrid of the two approaches (Kuhn & Johnson, 2013). In this case down-sampling was used to reduce the number of awake instances. This method was selected because the number of drowsy samples required in up-sampling would in turn require that each observed instance be repeated so frequently that subsequent trained models would overfit the training data. The limitation of down-sampling the awake instances is that a very small percentage of the observed awake instances are included in the training set. Some control over this limitation can be gained through the application of domain knowledge. Segments of the current data (and driving data in general) are composed of different contexts and driving environments that do not occur in uniform frequency. For example, a given drive may contain a large proportion of windows of highway driving and only a few windows of left turns. By sampling a dataset organized by these data patterns, one can achieve a sample that replicates the original data while capturing the breadth of included behavior. This concept was employed in this work by first clustering the awake instances and then taking a weighted sample of the cluster membership to produce a down-sampled training data set for the models.

4.4.3.2 Fitting and selecting the model

The next step in the meta-feature generation process is to fit and select an optimal machine learning model. This process can be accomplished by fitting a set of models to the training data. The model parameters can be optimized through an internal cross validation process and the optimized models can be

evaluated with an external cross validation. This study explored seven algorithms: Random Forest (RF), Decision Tree (DT), Naïve Bayes (NB), k-nearest neighbor (KNN), Support Vector Machine (SVM) with a linear kernel, SVM with a radial kernel, and Neural Network (NN). This set of algorithms represents the set of classic machine learning algorithms (Kotsiantis et al., 2007). Each of algorithms was applied separately to the three window sizes and two continuous feature subsets representing features derived from pedal and steering measures respectively. In addition to the separation of features by measurement type, models were fit for subsets of features including all continuous features, and all combinations of Fourier, distributional, and non-linear distributional features. The models were trained and evaluated with the caret package in R (Kuhn et al., 2011).

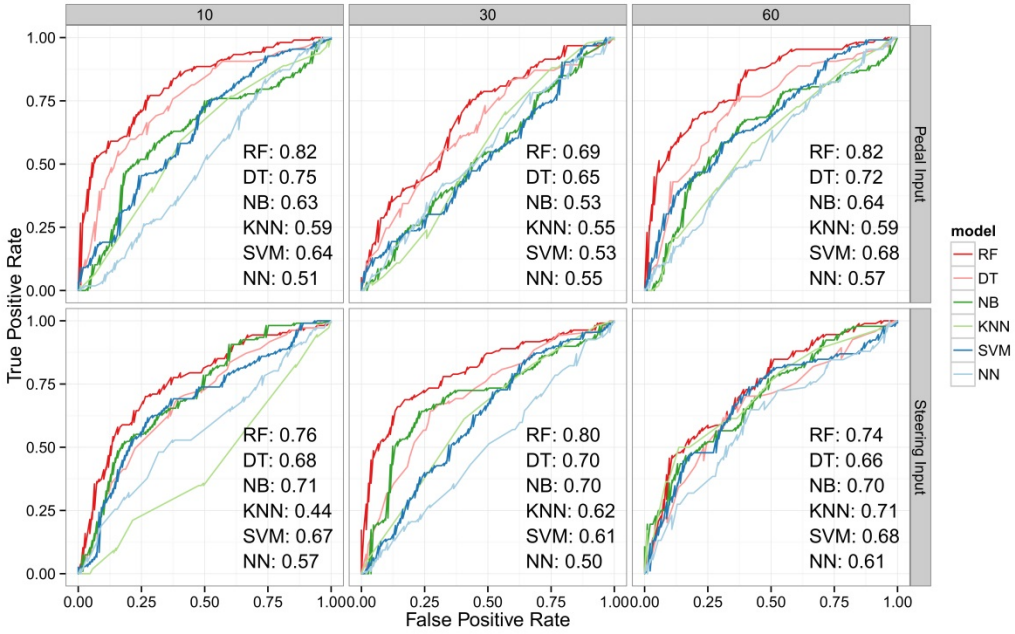

Figure 18 ROC plots of potential models for meta-feature generation for pedal and steering measures and window sizes of 10, 30, and 60 seconds. Note plots are arranged left to right by window size and top to bottom by feature input.

Figure 18 shows Receiver Operating Characteristic (ROC) curves and associated Area Under the Curve (AUC) values for 36 of the fitted models. For simplicity, only the radial kernel SVM models were plotted as they outperformed the linear kernel SVM models in all cases, additionally only the optimal feature subset for each measure type and window size is shown. This subset represents the distributional and non-linear distributional features from the steering measure based models and only the distributional

features from the pedal based models. The individual plots show ROC curves for a particular feature subset and window size paring. Note that an ROC curve plots the true positive rate (TPR) by false positive rate (FPR) across a range of thresholds. The definition of thresholds differs for each classification approach, but they can generally be understood as the minimum evidence required for positive classification. On an ROC plot, a random classifier would follow the line with a slope of one and an intercept of zero. An ideal classifier would follow the edge of the top left quadrant of the graph. The AUC facilitates comparison between two algorithm's ROC curves and a higher AUC indicates a stronger classifier. Figure 18 shows that the Random Forest model outperforms all other models for all window sizes and feature subset combinations, which suggests that the random forest is the optimal algorithm for meta-feature generation in this case.

4.4.3.3 Assessing information gain for meta-features

To generate comparisons between the SAX features, continuous features, and meta-features one must calculate information gain for the meta-features. This process can be accomplished by generating predictions from the fitted random forest models, and then converting those predictions into a binary classification using the thresholds that maximize the AUC. Figure 19 shows a pin plot of information gain values for the six random forests. Notably the 60-second features show a significantly higher information gain than the 30 or 10 second features. This suggests that the 60-second windows should be pursued.

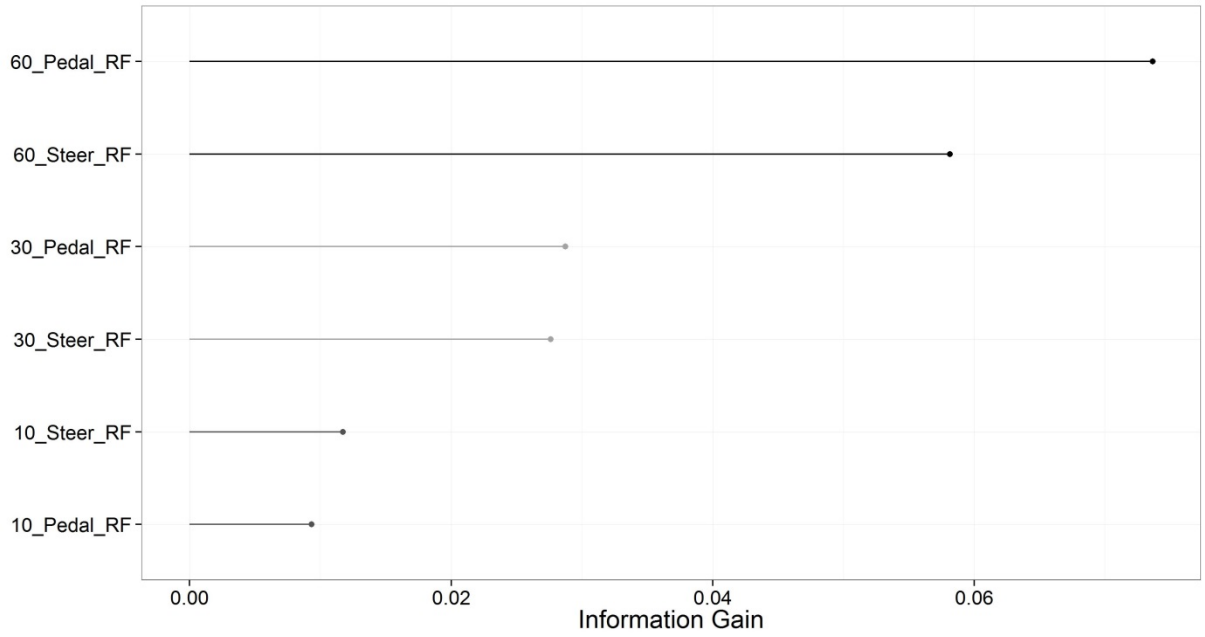

Figure 19 Pin plot of information gain for meta-features. Note the features are presented as: [window size] _ [measure] _ [machine learning algorithm] and the color of the pins is associated with the window size.

4.4.4 Comparing and selecting features

The final feature selection process requires a comparison between SAX, continuous, and meta-features. The order of magnitude difference between the results in Figure 17 and those in Figure 16 and Figure 19 (maximum information gain of 0.005 compared to 0.05 and 0.06 respectively) suggest that simply using the continuous features as observations in a graphical detection algorithm is inadvisable and would produce a significantly worse detection model than the SAX or meta-features. This result is not surprising as the meta-features and SAX features are capable of capturing more complicated patterns in the data. The difference between the meta-features and the SAX features is not as large; however, the meta-features for the 60 s windows have a higher information gain than any of the SAX features. Although it is difficult to quantify, the SAX variables have a higher sensitivity to overfitting than random forest meta-features. The random forest by nature is an ensemble classifier, which has several random sampling based protections against overfitting. Notably the trained model is capable of aligning previously unseen instances with similar known instances contained in the training data and making informed predictions. In contrast the SAX features are fully subject to the number of observations of a given feature and cannot handle unseen features well. These concerns, along with information gain results suggest that the features

used in all subsequent studies should consist of Random Forest models trained on 60 second windows of both steering and pedal data.

4.5 Evaluating the models

The literature on drowsiness detection demonstrates a wide range of evaluation metrics, which impedes effective comparisons between detection algorithms. One of the goals of this dissertation is to unify a set of these various evaluation metrics into a single comprehensive set. To this end, this dissertation explores two types of model evaluation, each supported with a separate series of statistical evaluations. The first type of model evaluation is based on the area under the ROC curve (AUC). The ROC curve plots the true positive rate by the false positive rate of a model for a range of thresholds. The threshold represents the minimum evidence required for a true prediction, and varies in exact definition by algorithm. The AUC of the ROC curve is a robust metric of model evaluation because it is insensitive to the underlying class distribution (Fawcett, 2004). The difference in various model AUC values are statistically evaluated with a bootstrapped significance test from the pROC package in R (Robin et al., 2011). This test estimates significance by calculating a statistic based on the ratio of the difference of two AUC values and the standard deviation of this difference over a set of samples or replicates. This statistic is then compared to the Normal distribution to assess statistical significance. The AUC and the bootstrapped statistical test provide a general comparison between models. In theory, a model with significantly higher AUC represents a performance improvement across nearly all thresholds, and thus is a better model.

The limitation of the AUC is that in practice algorithms require the use of only a single threshold. This threshold represents a design decision that balances the importance of true and false positives. Given this use, it is impractical to simply evaluate and discard models based on AUC values. Thus the AUC analyses will be supplemented with an evaluation at the threshold that produces the maximum sum of sensitivity and specificity on the ROC plot. This evaluation produces a confusion matrix, which gives counts of true and false positive, as well as true and false negatives. In addition it provides specific true and false positive rates. These values can be statistically evaluated using a McNemar's test, which is

recommended in (Dietterich, 1998), or a Fisher's exact test of count data for smaller samples. The goal of these tests is to evaluate whether or not there are practical settings of various models at which one model performs significantly better than another. It is important with these tests to balance the alternatives because in some cases models may produce a higher true positive rate by also allowing more false positives. Therefore all statistical comparisons at fixed thresholds will be calculated such that there is a balance in the alternative value. These tests will demonstrate whether various models can produce the same predictive power with significantly lower cost.

4.6 Benchmarking with previous algorithms

In addition to identifying the ideal structure from the set proposed in this dissertation, it is important to justify the added complexity of these models relative to previous work and the performance of algorithms currently employed in real-world systems. This justification will be established by including comparisons to two baseline algorithms: a steering-based random forest algorithm (McDonald, Lee, Schwarz, et al., 2013), and PERCLOS (Dinges & Grace, 1998) in all analyses in this work. The steering-based random forest represents baseline performance with a sliding window method. This model is ideal for this comparison because it contains many of the same features explored in this work but does not include time or context. Direct comparisons to this model will partially illustrate the effects of added complexity. This model will be referenced as the static steering model, indicating that it is non-temporal and does not include temporal dependencies. PERCLOS represents an industry standard metric that has been widely employed in after-market systems (Richard Grace et al., 1996), and will give a baseline for real-world algorithm performance. Comparisons to PERCLOS also partially illustrate each algorithm's ability to translate to real-world scenarios.

Chapter 5 Evaluating dynamic modeling structures for drowsiness detection

The goal of this study is to empirically assess dynamic graphical modeling structures in drowsiness detection. These models use a combination of observed behavior and previous states to make predictions. The inclusion of previous states in the prediction process should reduce false positives encountered by static prediction algorithms that might be artifacts of the road environment rather than drowsiness. The analysis presented here considers three types of graphical models: Hidden Markov Models (HMM), Hidden semi-Markov Models (HsMM), and Conditional Random Fields (CRF), which differ in their treatment of previous states and observed behaviors. All three of these models view the state of the world as being separated into connected hidden and observable state chains that unfold over time. In the context of drowsiness the hidden states represent drowsy or awake states and the observable states represent features that characterize driver behavior. Hidden Markov Models assume that each drowsiness state is both uniform in duration and representative of the previous state and the current observations. Hidden semi-Markov Models are similar to HMMs except that they remove the assumption of uniform duration of states, which allows states to vary in duration. Conditional Random Fields differ from HMM and HsMM because they allow the observed behavior at each time-step to influence the current hidden state. These structures all have the benefit of using temporal relationships to reduce the false positive rates associated with static detection algorithms however the individual impact of each of their differences and how these differences are affected by different observations is unclear. The remainder of this chapter focuses on quantifying the individual benefits of these modeling structures for detecting drowsy-related lane departures and drowsy driving in general. Following this analysis and a comparison between the models, a single modeling structure will be selected for further study.

5.1 Hidden Markov Models

The structure of HMMs and their ability to consider temporal dependencies suggests that they could produce a dramatic increase in prediction performance relative to a static algorithm. The goal of this analysis is to evaluate this increase in performance relative to the cost of increased model complexity and

multiple sensors. Therefore this analysis compares static algorithms to those using HMM and models containing features derived from a single measure with those including features derived from multiple measures.

5.1.1 Model training

This analysis explored three different HMM models: a model using pedal measure based random forest features as observations (Pedal HMM) , a model using steering measure based random forest features as observations (Steer HMM), and a model that used a combination of steering and pedal measure based random forest meta-features (Combined HMM). The steering random forest contained features representing both linear and non-linear distributional metrics, while the pedal random forest contained only linear distributional metrics. All of the models were trained using the mhsmm package in R (O'Connell & Højsgaard, 2011). The models were continuous density HMMs, meaning that they used the continuous voting output of the random forest models as observations rather than a binary classification. A Normal distribution was fit to the distribution of continuous RF output values associated with drowsy and awake data respectively. In the combined HMM model a multivariate Normal distribution was used to consolidate the pedal and steering random forest output into a single observation. The transition probabilities were initially determined from the transitions observed in the training data and were subsequently refined through an expectation maximization training procedure. In all cases the driver was initially assumed to be awake, which was consistent with the training data. In subsequent time-steps, the algorithm made predictions by converting windows of driver behavior data to votes with the random forest model(s) and then generating a prediction based on the transition probabilities and the densities of the observation distribution.

5.1.2 Model evaluation and discussion

The model fitting results are presented in Figure 20, and Table 8. Figure 20 shows the ROC curves and associated AUC values for the three HMM models, as well as the predictions from the steering and pedal random forests alone, in predicting the test data. The random forest results are presented to clarify the precise benefits of the HMM. Table 8 shows confusion matrices, False Positive Rates (FPR), True

Positive Rates (TPR), Area Under the Curve (AUC) and a 95 % bootstrapped confidence interval for the AUC for each model calculated at the threshold that produced the maximum sum of sensitivity and specificity. The bootstrapped confidence interval is produced by the pROC package in R and derived from repeated resampling comparison between two curves. All tests in this dissertation use 2,000 resamples or replicates.

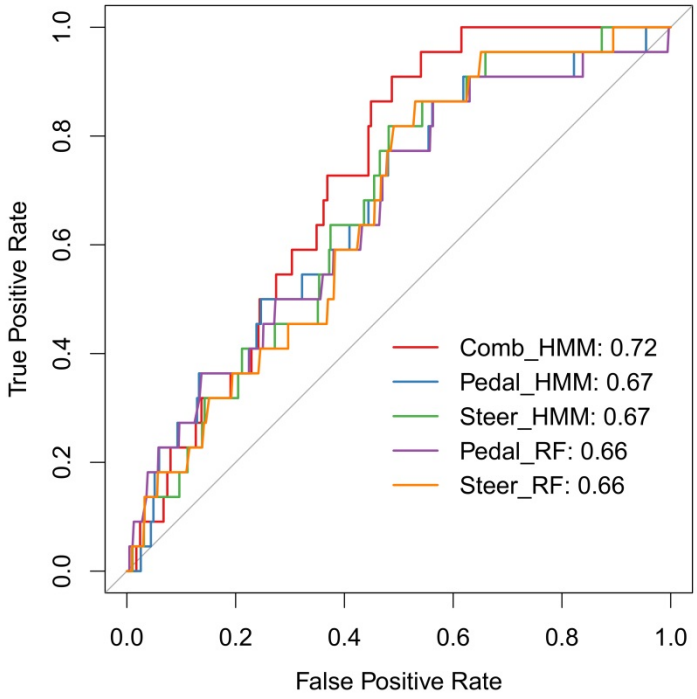

Figure 20 ROC curves for the HMM models and the random forest models predicting the test data.

Figure 20 shows that the HMMs have a higher AUC than the random forest models (Note that the random forest AUC values are not comparable to those calculated during the feature selection process because the AUC values in this figure are based on the test data rather than a cross validation of the training data) in all cases and that the AUC for the combined HMM is highest although the difference is not substantial. Bootstrapped significance tests comparing the combined HMM to the pedal HMM ($D(2000) = 0.99$, $p = 0.32$), steering HMM ($D(2000) = 0.91$, $p = 0.36$), pedal random forest ($D(2000) = 1.00$, $p = 0.31$), and steering random forest ($D(2000) = 0.98$, $p = 0.32$) were not significant, however the bootstrapped confidence intervals in Table 8 suggest all of the models predict significantly better than

random. The FPR and False positive counts in Table 8 show that difference in classification performance is due in most cases to a reduction of false positives between the HMMs and the random forests, with the combined HMM having the lowest false positive rate without a reduction in the identification of True Positives. Fisher's exact test (used due to limited sample sizes) performed on the False Positive Rates showed that the difference in False Positives between the Combined HMM and the Steering angle random forest ($p = 0.001$), Pedal random forest ($p < 0.001$), and Pedal HMM ($p < 0.001$) were all significant.

Table 8 False Positive Rates (FPR), True Positive Rates (TPR), Area Under the Curve (AUC), and bootstrapped 95% AUC confidence intervals for the HMM and random forest meta-feature models. Note the values are calculated at the threshold that produced the maximum sum of sensitivity and specificity for each algorithm.

Model	*True Negatives*	*False Negatives*	*False Positives*	*True Positives*	*FPR*	*TPR*	*AUC*	*Bootstrapped 95 % CI*
Combined HMM	470	3	390	19	0.45	0.86	0.72	(0.65 - 0.80)
Pedal HMM	377	3	483	19	0.56	0.86	0.67	(0.56 - 0.78)
Steering HMM	449	5	411	17	0.48	0.77	0.67	(0.57 - 0.77)
Pedal random forest	376	3	484	19	0.56	0.86	0.66	(0.55 - 0.78)
Steering angle random forest	404	3	456	19	0.53	0.86	0.66	(0.56 - 0.76)

These results are consistent with the hypothesis that the HMM models improve classification performance due to their ability to consider time dependencies, specifically abrupt transitions to drowsy states. The reduction in false positives and stability in sensitivity associated with the HMMs as compared to their random forest counterparts suggests that the HMM structure essentially serves as a filter for erroneous positive classifications without cost to the detection performance of the random forest. It is surprising that the improvement is not dramatic and that it does not generate a statistically significant change in AUC, although the size of the confidence interval suggests that the analysis could benefit from the use of a larger testing dataset. The significant difference in False Positive Rates between the Combined HMM and the random forest models is encouraging and provides evidence to support the hypothesis that under specific settings the Combined HMM outperforms static models. Despite the difficulty created by the narrow performance difference between models, it is somewhat encouraging that

in a real world scenario where a measurement sensor is lost, the performance of an HMM-based algorithm would not suffer.

5.2 Hidden semi-Markov Models

The limitation of HMMs is that they cannot take advantage of state duration. This is limiting from a detection perspective because there is some evidence of durational regularity in drowsiness and alert states (Boyle et al., 2008). This analysis evaluates the impact of this difference and the variance of this impact across measurement sources. Like the HMM analysis, the analysis focuses specifically on comparing HsMM models to static algorithms and the difference between HsMM models using a single measure and those using both steering and pedal input.

5.2.1 Model training

This study explored three HsMM models: a model using pedal measure based random forest features as observations (Pedal HsMM), a model using steering measure based random forest features as observations (Steer HsMM), and a model that used a combination of steering and pedal measure based random forest meta-features (Combined HsMM). The steering and pedal random forest models were identical to those used in the HMM analysis. All of the models were trained using the mhsmm package in R (O'Connell & Højsgaard, 2011). A Normal distribution was fit to the distribution of continuous RF output values associated with drowsy and awake data respectively. In the combined HsMM model a two dimensional Normal distribution was used to consolidate the pedal and steering random forest output into a single observation. The transition probabilities were initially determined from the transitions observed in the training data. Similarly the state durations were modeled using Gamma distributions fit to the observed state durations in the training data. Both of these parameters were optimized through an expectation maximization process. In all cases the driver was initially assumed to be awake, which was consistent with the training data.

5.2.2 Model Evaluation and discussion

The HsMM model results are presented in Figure 21, which shows ROC curves for the three HsMM models and the two static random forest models, and Table 9, which presents confusion matrix count

values, FPR, TPR, AUC, and bootstrapped confidence intervals for the fitted models. The figure indicates that the detection performance of the HsMM models is substantially worse than the simple random forest models and by extension the HMM models from the previous chapter. Furthermore the bootstrapped AUC confidence intervals for the HsMM models contain 0.5, which suggests the models performance is similar to a random classifier. The results in Table 9 suggest that this decrease is largely due to a substantial increase in False Positives in the HsMM models.

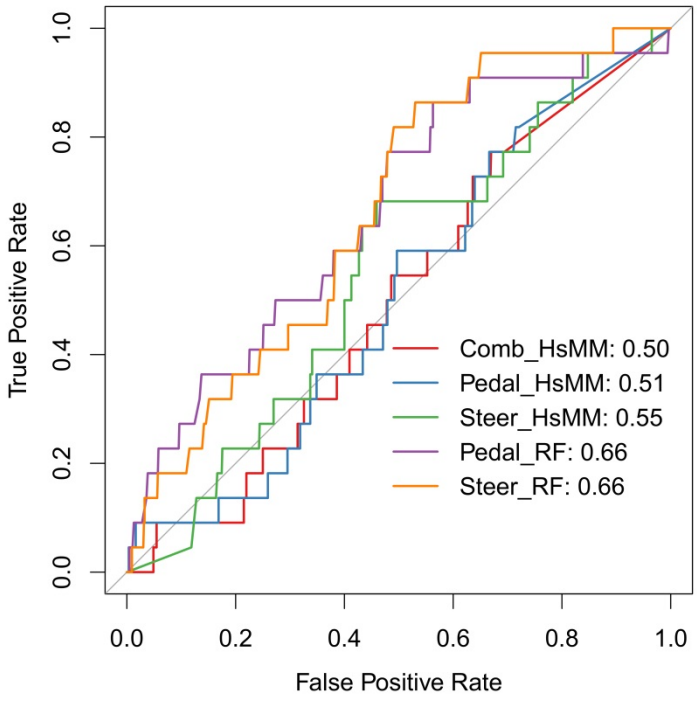

Figure 21 ROC curves for the Hidden semi-Markov Models.

Table 9 Confusion matrix counts, FPR, TPR, AUC, and bootstrapped 95 % Confidence Interval values for the HsMM models and the static random forest models.

Model	*True Negatives*	*False Negatives*	*False Positives*	*True Positives*	*FPR*	*TPR*	*AUC*	*Bootstrapped 95 % CI*
Combined HsMM	284	5	576	17	0.67	0.77	0.50	(0.39 - 0.61)
Pedal HsMM	287	5	573	17	0.67	0.77	0.51	(0.40 - 0.62)
Steering HsMM	265	5	595	17	0.69	0.77	0.55	(0.44 - 0.66)
Pedal random forest	376	3	484	19	0.56	0.86	0.66	(0.55 - 0.78)
Steering angle random forest	404	3	456	19	0.53	0.86	0.66	(0.56 - 0.76)

The substantial decrease in performance associated with the added complexity of the HsMM is surprising however a more thorough consideration of the definition of drowsiness and state organization provides valuable context and insight. Drowsiness in this case is defined as a window of driving data where a lane departure occurred, the driver was rated to be drowsy, and no correction was applied and all windows not defined as drowsy were considered awake. The mechanics of the HsMM are such that the

model must transition states after the duration of a given state is completed. Meaning that in the two hidden state case the model must transition to awake after being drowsy. When the duration of the awake states observed in the training data has a large variance, this will bias the model to a higher frequency of false positives. Furthermore previous studies such as (van Kasteren et al., 2010), demonstrate a benefit of HsMMs where observations associated with two states were similar but the duration of the two states were different in some type of predictable manner. For example, the arm muscle activation involved in window cleaning versus tooth brushing. In this case the observations from the random forests are quite different between states and inspection of the durations of these states suggests that they are quite irregular. These results strongly suggest that HsMM models are not advantageous for detecting drowsy-related lane departures. They also provide evidence at a more fundamental level that the transitions to various levels of drowsiness are regular, however, the length of time at each level is not, and therefore durations do not provide a benefit for drowsiness detection.

5.3 Conditional Random Fields

Conditional Random Fields are a potentially advantageous predictive modeling framework relative to HMM and HsMM because they do not require strict assumptions about the relationship between subsequent observations. The removal of this assumption allows CRF models to consider the entire sequence of observations in their classification. In theory, this broader consideration should significantly improve classification. The goal of this analysis is to evaluate this theory relative to static models.

5.3.1 Model Fitting

This analysis investigated three Conditional Random Field models trained with the General Conditional Random Field Toolbox for MATLAB (Murphy, 2006). The analysis was constrained in that the CRF toolbox does not support continuous observations, and development of a more expansive toolbox was beyond the scope of this work. The toolbox constraints were accommodated by converting the continuous random forest voting output to a binary label of drowsiness using a threshold which produced the maximum sum of sensitivity and specificity. This binary conversion is somewhat limiting however it

maintains the general trend of predictions from the random forests. Following this conversion the model parameters were optimized using an internal procedure within the CRF toolbox that is similar to expectation maximization. Predictions for the models were also generated within the CRF toolbox. In this case features were generated relative to predictions across the entire sequence of testing observations however in a real world application this type of prediction would be impossible (i.e., using future unseen instances to predict the current state). Although there are methods to accommodate CRF predictions in real time (Barbu, 2009), they require a generalization over unseen portions of the sequence and would almost certainly underperform the results presented here.

5.3.2 Model Evaluation and discussion

Figure 22 and Table 10 show the results for the Conditional Random Field models. The CRF models show a wider range of performance than the other two graphical modeling methods. Notably the steering CRF predicts about the same as a random classifier and the pedal and combined CRFs predict significantly better than random. Bootstrapped significance test between the three models show that the differences between the combined CRF model and steering CRF ($D(2000) = 8.62$, $p < 0.001$), and the pedal CRF and steering CRF ($D(2000) = 4.13$, $p < 0.001$) are significant, however the difference between the combined CRF and the pedal CRF is not significant at the $p = 0.05$ level ($D(2000) = 0.67$, $p = 0.50$). It is interesting that the combined CRF is able to compensate for the poor performance of the steering predictions in this case whereas the steering CRF is not. The FPR and TPR values in Table 10 show that at the threshold of highest sensitivity and specificity, the CRF models can reduce false positives, although this comes at the expense of model sensitivity. This result seems somewhat odd in that it seemingly disagrees with the AUC results. However the result may simply be an artifact of the use of binary observations which essentially reduces the degrees of freedom of the state predictions and in turn the number of threshold inflection points on the ROC curve. A model with continuous observations can have an infinite variety of observations and therefore infinite levels of output. In contrast this model has a limit on the different types of observation at each time-step corresponding to all combinations of binary output

from the random forest input (two in the single measure case, four in the combined case). This limits the potential output and by extension the available thresholds on the ROC curve.

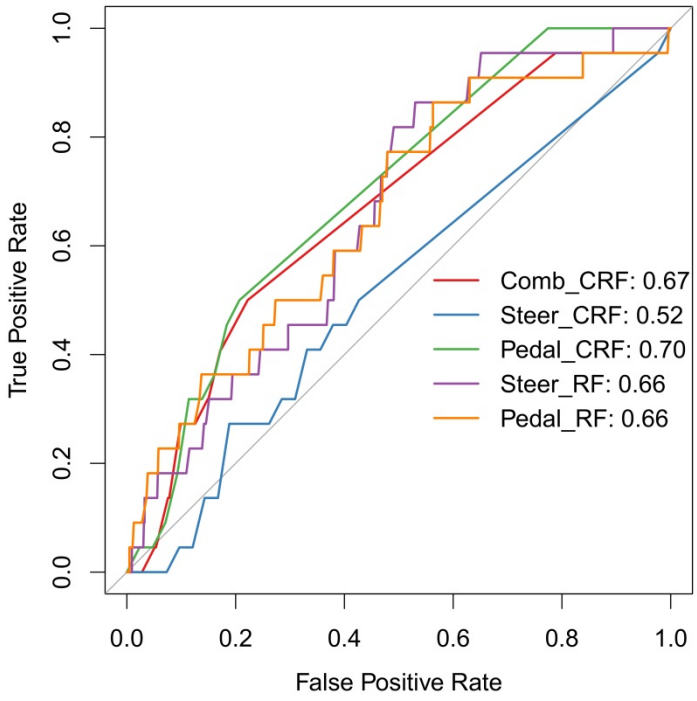

Figure 22 ROC curve for the CRF model compared to the random forest meta-features.

Table 10 Confusion matrix, TPR, FPR, AUC, and bootstrapped confidence interval for the Conditional Random Field Model.

Model	*True Negatives*	*False Negatives*	*False Positives*	*True Positives*	*FPR*	*TPR*	*AUC*	*Bootstrapped 95% CI*
Combined CRF	669	11	191	11	0.22	0.50	0.67	(0.56 - 0.77)
Steering angle CRF	698	16	162	6	0.19	0.27	0.52	(0.40 - 0.63)
Pedal CRF	178	11	682	11	0.79	0.50	0.70	(0.61 - 0.78)
Pedal random forest	376	3	484	19	0.56	0.86	0.66	(0.55 - 0.78)
Steering angle random forest	404	3	456	19	0.53	0.86	0.66	(0.56 - 0.76)

Together these results suggest that Conditional Random Fields could be a valid approach to drowsiness detection, particularly in a post-hoc detection scenario. Although it is difficult to fully quantify the performance reduction associated with a transition to real-time predictions, the performance would almost certainly be poorer. It is interesting that the CRF models do not provide a more significant

increase in performance given that they use many more parameters. This lack of benefit relative to additional complexity suggests that CRF models may be less advantageous than simple static models that have a similar level of performance with fewer parameters.

5.4 Comparing modeling structures and benchmarks

The goal of this study is ultimately to select a graphical model for further analysis with contextual data and to evaluate the chosen model relative to PERCLOS, a proven real-world detection algorithm, and a random forest steering algorithm, which provides a benchmark for evaluating additional complexity. Figure 23 shows ROC curves and smoothed ROC curves representing the three graphical modeling structures explored in this study and the two benchmarking algorithms. The smoothed ROC curves estimate classifier performance on a larger testing dataset. The individual models represent the highest performing model by AUC from each of the three graphical modeling explorations: the combined HMM model, the steering HsMM model, and the pedal CRF model. The smoothed ROC curves on the right side of Figure 23 show that the HMM model outperforms all of the other models and that this improvement reflects both a lower false positive rate and a higher sensitivity across a large range of thresholds. In contrast, the HsMM model performs more poorly than both benchmark algorithms and the CRF model has relatively similar performance despite being the most complex model, as measured by number of parameters. Bootstrapped significance tests of these models show that the difference between the combined HMM and the HsMM models is significant ($D(2000) = 2.56$, $p = 0.01$), however the difference between the combined HMM and the other models is not significant at the $p = 0.05$ level.

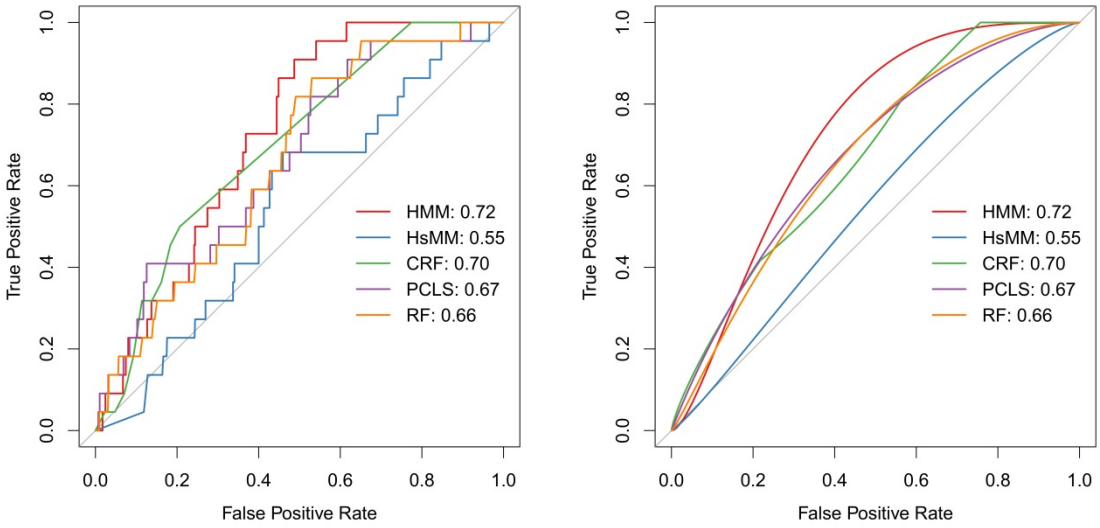

Figure 23 ROC (left) and smoothed ROC (right) curves for the HMM, HsMM, CRF, PERCLOS, and Random Forest models.

Although the AUC test results were not significant beyond the HsMM comparison, evaluations at the true positive rate associated with the maximum performance of the HMM model did show significant differences in performance. Fisher's exact tests of the false positive rate of the HMM and PERCLOS ($p < 0.001$), the steering random forest ($p < 0.001$), and HsMM ($p < 0.001$) suggested that the HMM model could predict the same frequency of true positives at a significantly lower rate of true positives for at least one threshold. Together these results suggest a preference for the HMM model.

5.4.1 Assessing model limitations and successes

The ROC curves in Figure 23 suggest that the HMM model reduces false positives and generally improves classification relative to the static random forest model and the other graphical models evaluated in this study. However the ROC space is limited in that it does not link the improvements directly to the driving scenario. The loss of this link makes it difficult to evaluate the hypothesis that the HMM model reduces false positives caused by driver behavior driven by context rather that a state of drowsiness. Figure 24 makes this link by showing false positives by simulator event for the combined HMM model and the static random forest model. In the figure they grey bars are false positives removed by the transition to HMM and the black bars are remaining false positives. In both cases the false

positives are calculated based on the threshold that produces the maximum sum of sensitivity and specificity, which corresponds to the same sensitivity across both models.

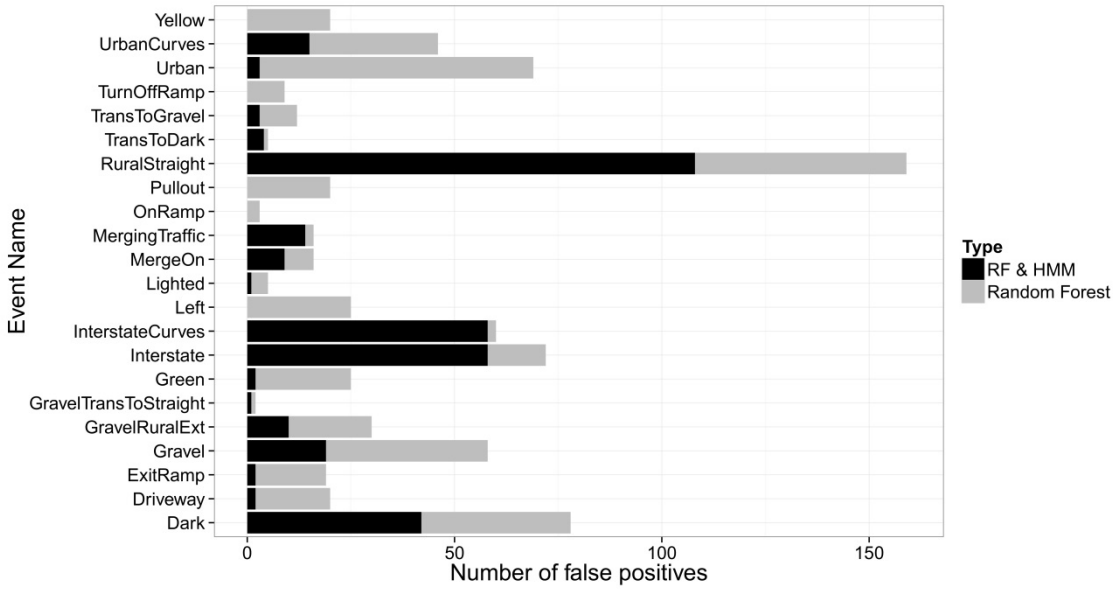

Figure 24 False positives by simulator event removed by applying the HMM relative to the static random forest model (grey) and those retained by both models (black). Note False Positives were identified based on a threshold that maximized the sum of sensitivity and specificity.

Figure 24 shows that the HMM eliminates all of the false positives for the Yellow (a yellow light dilemma), the Turn off Ramp (turn off of a highway on-ramp), the Pullout (pulling out of a driveway), the On Ramp, and the Left Turn events. In addition the Fisher's exact tests showed that the HMM significantly reduced false positives associated with the Urban Curves ($p < 0.001$), Urban ($p < 0.001$), Transition to Gravel (a segment of transition between a rural road and a gravel drive; $p = 0.03$), Green (passing through a green light; $p < 0.001$), Gravel Rural Extension ($p < 0.01$), Gravel ($p < 0.001$), Driveway ($p < 0.001$), and Dark ($p < 0.01$) events. This is promising because these events almost exclusively occur after other events of high demand. These results suggest that the model effectively captures the fact that awake drivers tend to stay awake. The number of persistent false positives in several events such as the Merging Traffic and Interstate Curves, which are not likely to exhibit a drowsy-related lane departure, suggests that the model could be improved by the addition of contextual information regarding the frequency of false positives in various contexts. This result further establishes the need to include driving context in detection algorithms.

5.5 Discussion and theoretical implications

The specific goal of this study was to evaluate dynamic graphical modeling structures for detecting drowsy-related lane departures and demonstrate that such models provide improved classification performance relative to static models and previous benchmarks. Despite this narrow focus, the study provides several important broader insights into drowsiness detection, the phenomenon of drowsiness, and predictive modeling. This section describes these insights and provides recommendations for future work.

Perhaps the most important result of this study, from a drowsiness detection standpoint, is the increase in performance generated by the transition from a static algorithm to the HMMs. Although previous work advocates the use of dynamic models (Ji et al., 2006; G. Yang et al., 2010; J. H. Yang et al., 2009), few explorations have directly compared and quantified the effect of dynamic models relative to static models with a robust dataset. The increase in AUC on an unseen testing data set achieved by the HMM in this study suggests that HMM models should be considered in future algorithm developments, particularly using driver behavior to detect drowsiness. In addition to the combined HMM results, it is encouraging to note that detection performance does not significantly degrade when a measurement source is removed. This result is promising given the potential use of cruise control or sensor failures in real-world scenarios. In contrast to the HMM, the HsMM and CRF results are not as encouraging. Detection problems with larger state spaces and more regular transition durations may benefit from the use of HsMM and post-hoc analyses with larger observation sets might be needed to realize the benefits of CRF. For example, in the multiple impairment detection case, HsMM might differentiate alcohol from distraction because distracted driving occurs in shorter time frames. CRFs might be used as a post-hoc drive analysis tool that provides feedback to drivers. Even with these applications, this study found no evidence to suggest that HsMM models could successfully predict drowsy-related lane departures in real-time, and minimal evidence that CRFs could improve on previous algorithms that require fewer parameters.

From a more general drowsiness perspective, the results of this study confirm the notion that temporal state dependencies exist in drowsiness (i.e. an awake driver will likely stay awake in the near future and a drowsy driver will likely stay drowsy). The detection performance of the HMM model not only supports this state transition concept, but also demonstrates that the regularity of these state transitions can be harnessed to predict drowsiness. Interestingly the HsMM results suggest that despite the regularity of transitions between states, the duration of these transitions is random. More generally, this suggests and confirms the existence of acute drowsiness events. If time on task effects were the sole cause of drowsy-related lane departures, HsMMs would perform well by predicting a pattern of long-duration awake states followed by a drowsiness state. The poor performance shows that this strategy is not viable. Despite these results, it would be beneficial to re-evaluate HsMM with a more robust ground truth measurement that tracks all levels of the gradual transition to a drowsy-related lane departure.

In the context of predictive modeling, this study provides a contribution to time-series anomaly detection and graphical modeling analysis for human behavioral identification. Although all of the results and conclusions here are data dependent, they may serve as a benchmark for similar analyses in the future. The most prominent result in the predictive modeling context is that HMMs can be used to essentially filter noise from a static machine learning model in time-series inference problems. Furthermore, the HMM component may be able to overcome false positives that might occur as a result of treating a time-series inference problem as a static classification task. Alternatively CRF models might be used in a similar manner, although the performance may lag behind HMMs. Finally, the HsMM results here suggest that the use of HsMM models for binary inference in anomaly detection is not advisable.

Overall the results from this study are promising in that they confirm the primary hypothesis: considering temporal dependencies in a drowsiness detection algorithm improves performance. However the best dynamic graphical model observed here, the combined HMM algorithm, produces an unacceptably high false positive rate at the threshold that produces the maximum sum of sensitivity and specificity. This might reflect the model's failure to consider the context in which drowsiness occurs and so it cannot differentiate driver behavior driven by context and driver behavior driven by drowsiness.

Subsequent analysis should analyze these effects and focus on the integration of driving context into detection algorithms.

Chapter 6 Analyzing contextual driving features for drowsiness detection

The previous study demonstrated that Hidden Markov Models are an effective tool for improving the drowsy related lane departures relative to PERCLOS and a static random forest steering algorithm. Further analyses suggest that a reduction of false positives causes this improvement. Despite this improvement, the Hidden Markov algorithm did not remove all of the false positives from the static steering algorithm. A substantial portion of the remaining false positives occurred during simulator events that were unlikely to induce a drowsy-related lane departure. Integrating such contexts into a detection algorithm should reduce the false positives observed from the HMM model alone and improve classification.

Although the simulator data presented in this work clearly define this context with labeled simulator events, the definition of on-road context is a substantially more difficult problem in the real world. Furthermore, while some studies suggest that on-road context may be a reflected in speed and acceleration behavior (Fuller, 2005), theory on driver behavior provides very little guidance on converting such a reflection into a compact yet meaningful form that reflects areas of the road where a drowsiness incident is likely to occur. The central challenge of the conversion is finding a time-series feature generation method that reduces speed and acceleration data while maintaining its central characteristics. The previous study suggests that to perform optimally this conversion method must also convert data into a form that can be integrated into dynamical graphical models such as an HMM. The goal of this study is to examine contextual feature generation methods and the integration of their resulting features into a graphical detection model. The study will focus on three methods of contextual feature generation: Symbolic Aggregate Approximation, Distributional characteristics, and Discrete Fourier Transforms, applied to three contextual data sources: speed, lateral acceleration, and longitudinal acceleration all integrated into the Hidden Markov Model from the previous chapter. The subsequent sections describe the feature generation and selection process for each feature generation method, illustrate their integration

into the HMM, and then compare the prediction results of the subsequent models relative to the previous study and the benchmarking models.

6.1 Symbolic Aggregate Approximation

Symbolic Aggregate Approximation is a feature generation method that converts continuous univariate time-series data into discrete words. The resulting output significantly reduces the size and complexity of the original data without a substantial loss of information. Furthermore the discrete nature of the words facilitates integration into a dynamic graphical modeling framework because observed word frequencies are essentially maximum likelihood estimates, i.e. frequency tables can simply be added into the model without further manipulation. The primary limitation of SAX is that it requires several input parameters: the word length, alphabet size, and normalization procedure. Previous work suggests that a global normalization of speed and acceleration behavior relative to a large sample of driving data is effective producing meaningful output for driving data (McDonald, Lee, Aksan, et al., 2013b), however there is very little guidance in the literature regarding the alphabet size and word length settings, particularly for drowsy driving detection features. Similarly, there is very little support for particular graphical modeling structures for integrating SAX features into a graphical drowsiness detection algorithm. Therefore this study focused on conducting an optimization process over a wide range of SAX input settings and integration structures to find the optimal settings.

6.1.1 Feature creation

The SAX features used in this study were generated through a custom function written in R. SAX was applied separately to 60 s windows of speed, lateral acceleration, and longitudinal acceleration data. In each application data were normalized by a global mean and standard deviation calculated from a uniform sample of 5,000 measurements across all drivers and conditions. Sample repetitions suggested that these mean values were relatively stable and did not change substantially with resampling. These global means and standard deviations are presented in Table 11.

Table 11 Means and standard deviations used in the normalization step of SAX.

Variable	*Mean*	*Standard Deviation*
Speed	46.31 mph	16.48 mph
Lateral Acceleration	-0.058 m/s^2	1.556 m/s^2
Longitudinal Acceleration	-0.003 m/s^2	1.129 m/s^2

After the data were normalized, they were converted into letters by dividing the data into a set of windows, taking the mean of the windows, and then assigning the means to a set of bins on the y-axis defined by quantiles of the normal distribution. The number of windows represented the word length and the number of quantiles represents the alphabet size. Given the results of the feature generation process, three word lengths were explored: one, two, and three, these word lengths were selected to balance the tradeoff between word frequency and the number of rarely observed words. Initially SAX features were calculated for word lengths of five however these words were found to have many rare observations and were removed from subsequent analyses. In addition the varied word length alphabet sizes of 3, five, seven, and nine letters were explored. This range effectively covers the range suggested by Lin et al. (2007) and the alphabet size used in McDonald et. al (2013b). The resulting application of SAX produced a total of thirty-six features, one for each combination of variable, word length, and alphabet size.

6.1.2 Model structures

In addition to the range of SAX inputs explored, this work also investigated four different modeling structures for integrating SAX into an HMM model. The four structures correspond to: a structure that considers only speed features, a structure that includes speed and lateral acceleration features, a structure that considers speed and longitudinal acceleration, and a structure that includes speed, lateral acceleration, and longitudinal acceleration independently. Two time steps of each of these model structures are illustrated in Figure 25. For the speed only structure and the independent structure, individual frequency tables were created for the variables. In contrast the combined speed and acceleration structures used a combined frequency table of speeds and the given acceleration. The goals of these structures were to establish the importance (or lack thereof) of the SAX acceleration words in the definition of context, and examine the impact of considering speed and acceleration as independent or dependent measures.

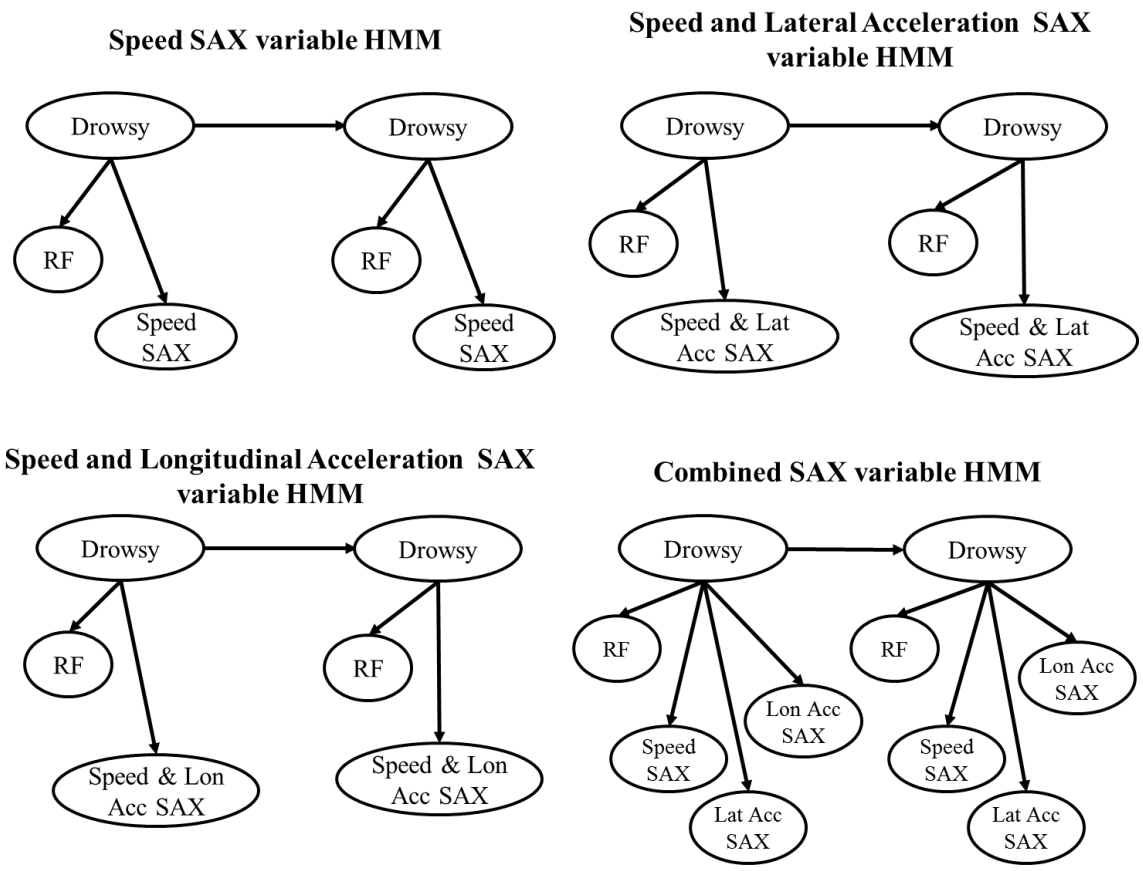

Figure 25 Modeling structures explored in the analysis of SAX features. Note that each quadrant displays 2 time steps and that the structures were designed to demonstrate the benefit of including acceleration in the definition of context and the impact of considering the SAX features independently.

6.1.3 SAX model evaluation and selection

Each structure was fit for all combinations of the SAX word length and alphabet size, and the resulting models were evaluated using the test data to determine the optimal structure and SAX settings in the space. In all 48 models were evaluated. Figure 26 shows the AUC results and associated bootstrapped confidence intervals for the SAX evaluation. The figure shows the AUC for each SAX input setting in a single plot arranged by word length (columns) and alphabet size (rows), with the model structure on the x-axis. The highest AUC value, with a word length of three, alphabet size of three, and the combined speed and longitudinal acceleration structure is highlighted in red.

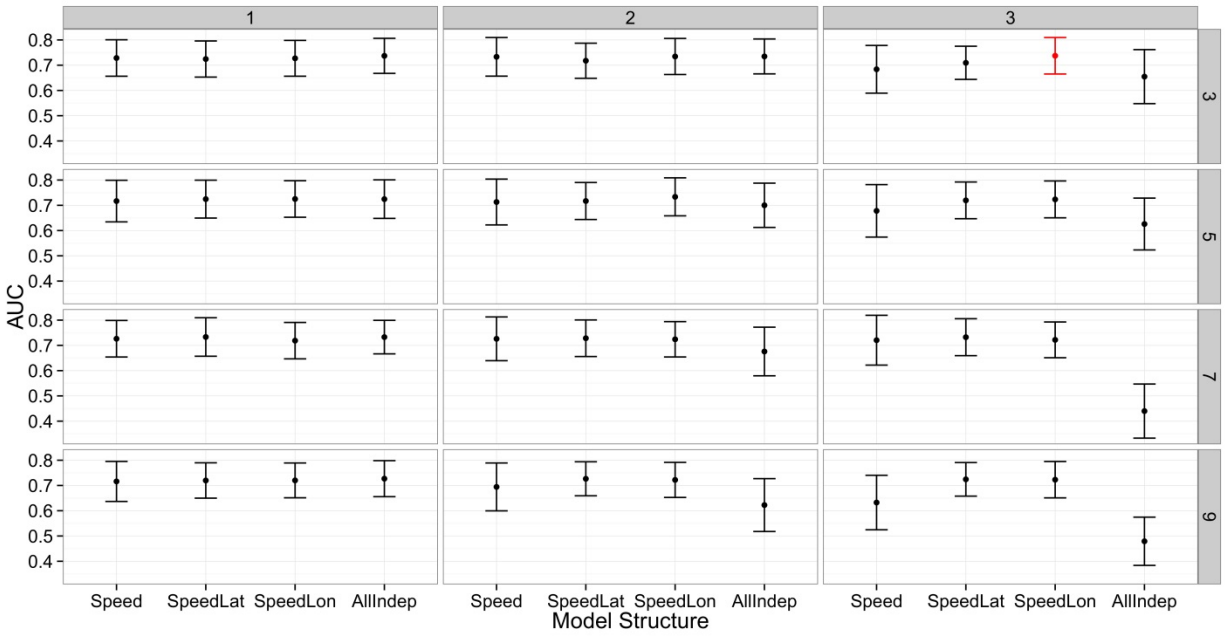

Figure 26 Area under the curve results with 95% bootstrapped confidence intervals for the SAX models arranged by word length (columns), alphabet size (rows), and model structure (x-axis labels). Note that the x-axis labels from left to right correspond to models that consider only speed, those that contain speed and lateral acceleration, those that contain speed and longitudinal acceleration, and models that consider all three features independently. The model with the highest AUC (with a word length of 3, alphabet size of 3, and the SpeedLon structure) has been highlighted in red.

The results suggest that the models in general are relatively insensitive to both word length and alphabet size, however the performance sharply drops off for the speed only structure and the all independent variable structure as the word length and alphabet size increase. It seems that smaller alphabet sizes are more robust to changes in both word length and model structure. The lack of extensive variability in the results might be due to the fixed structure of the drive however the results do seem to confirm the importance of a combination of acceleration and speed in the definition of context and advocate for smaller alphabet sizes. Although there is not overwhelming evidence that the combined speed and longitudinal structure with an alphabet size and word length of three is the optimal model, the results here suggest that it should be retained for further analysis with the knowledge that a decrease in word length or a change in structure would not significantly affect performance.

6.1.4 Model results and discussion

The previous section illustrated the effect of SAX features and modeling structures but it did not consider the performance of SAX relative to the final model from the previous study. A comparison of these two

models ROC curves and their prediction statistics at the threshold of maximum sensitivity and specificity are shown in Figure 27 and Table 12 respectively.

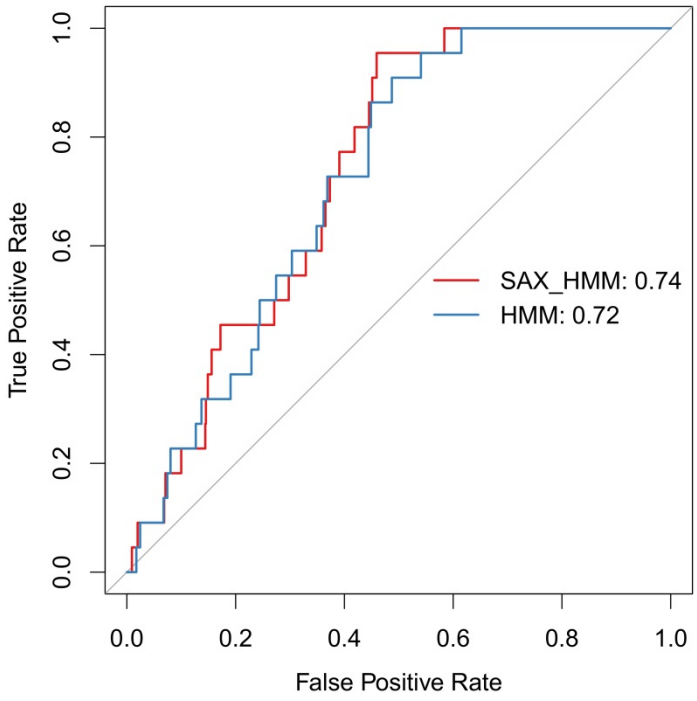

Figure 27 ROC curves and AUC values for the SAX HMM model, the combined HMM from the first study, and a static random forest model.

Table 12 Confusion matrix, TPR, FPR, AUC, and bootstrapped confidence interval for the SAX model and the Combined HMM model from the first study.

Model	*True Negatives*	*False Negatives*	*False Positives*	*True Positives*	*FPR*	*TPR*	*AUC*	*Bootstrapped 95% CI*
SAX HMM	472	2	388	20	0.45	0.91	0.74	(0.67 - 0.81)
Combined HMM	441	2	419	20	0.49	0.91	0.72	(0.65 - 0.80)

The ROC plot suggests that the SAX HMM produces only a modest improvement, 0.02 AUC, over the previous HMM. This difference is not significant according to a bootstrapped significance test, $D(2000) = 0.93$, $p = 0.35$). However, a close inspection of the ROC curve reveals several regions of the curve where the SAX HMM curve is substantially to the left of the original HMM curve. These regions are important because they highlight areas where SAX reduces false positives without a cost to identification of

drowsy-related lane departures. A Fisher's Exact Test at the threshold displayed in Table 12 shows that the reduction of false positives is not significant at the $p = 0.05$ level ($p = 0.1$). The results are encouraging, but suggest a need for an alternative contextual definition or another method of reducing false positives.

6.2 Distributional measures

Distributional measures are important to investigate because they are extensively used in the transportation analysis literature, particularly analyses of driver performance. The features are advantageous because they are simple and fast to calculate and they provide substantial data reduction for time-series data. Their limitation is their calculation involves a loss of temporal patterns across the signal, meaning that the original signal cannot be reconstructed from distributional measures unlike SAX or Fourier Transforms. This could be problematic because the link between speed and acceleration and context might be related to patterns of data rather than a single number. Another concern with distributional measures is that they are continuous. When these features are integrated into a graphical model they must be included as a distribution with the same dimensionality as the number of features. As this number of features increases, the likelihood of observing any particular point in the distribution tends to decrease and by extension the observations related to this distribution have a smaller effect in classification. In light of the popularity of distributional features and their limitations this study sought to examine their application to the definition of context and integration into drowsiness detection algorithms.

6.2.1 Feature selection

A total of 11 distributional features including: Mean, Minimum, Maximum, Median, Standard Deviation, Skew, Kurtosis, Lyapunov Exponent, Sample Entropy, Correlation Entropy, and Correlation dimension were considered in this analysis. Each feature was calculated separately for speed, lateral acceleration, and longitudinal acceleration, creating 33 contextual distributional features. This number of features could be prohibitive to classification so a smaller feature subset was created by applying a feature selection method based on the Kulback-Leibler (KL) divergence, which can be thought of as a continuous version

of information gain. The method consisted of calculating the KL divergence between the distributions of each feature associated with drowsy instances and awake instances in the training data set and then selecting a subset of features that displayed the highest KL divergence. Figure 28 shows a plot of KL divergence for each of the variables in rank order. The KL divergence method retained four features: mean speed, maximum speed, minimum speed, and median speed. The figure shows that after these values there was a sharp decrease in KL divergence between the two distributions that suggested additional features would not benefit the analysis.

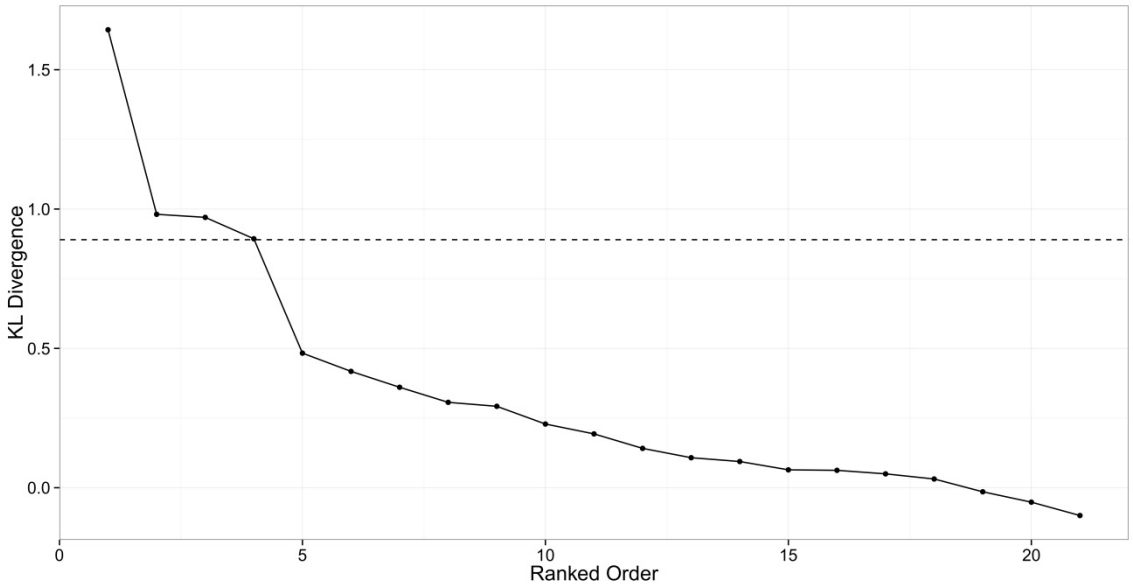

Figure 28 KL divergence for the distributional features. Each point on the plot represents a feature and the dotted line indicates the cutoff threshold used to select retained features.

6.2.2 Model structures

Two separate modeling structures were explored in this evaluation: a structure that integrated distributional features and the subset of distributional features identified in the previous section as multinomial distributions independent of the random forest meta-feature input, and a structure that integrated the features directly into the random forest prior to integration into the HMM. This later structure used the entire feature set because the design of the random forest allows for a natural feature filtering and therefore is robust to the number of distributional features. The goal of evaluating these structures is to identify the benefit of considering distributional context independently of the pedal and

steering measures as compared to considering context and drowsiness meta-features independently. Two time steps of each of these structures are illustrated in Figure 29.

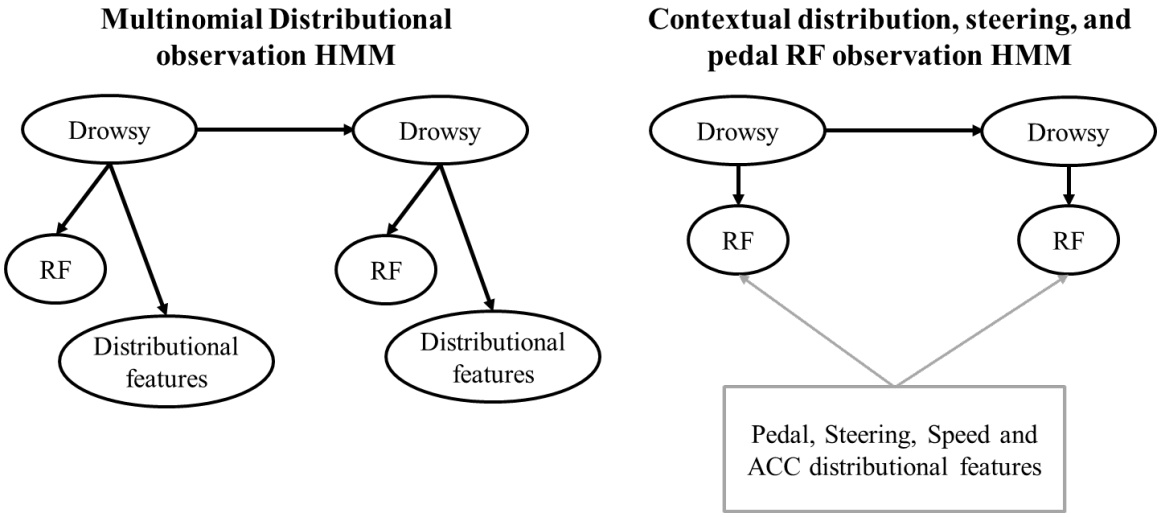

Figure 29 Modeling structures explored in the analysis of distributional features. Note that each image displays 2 time steps. The grey box on the right side of the figure represents features integrated into the RF training that are consistent across time.

6.2.3 Model results and evaluation

The combination of the two feature sets and the two structures created a total of 3 different distributional feature contextual models. Note that the distributional random forest structure was not fit to the subset of features identified by KL divergence, because the random forest essentially employs an internal feature selection process that would make use of a subset redundant. Figure 30 and Table 13 show the results of these model applications along with the combined HMM model from the previous chapter.

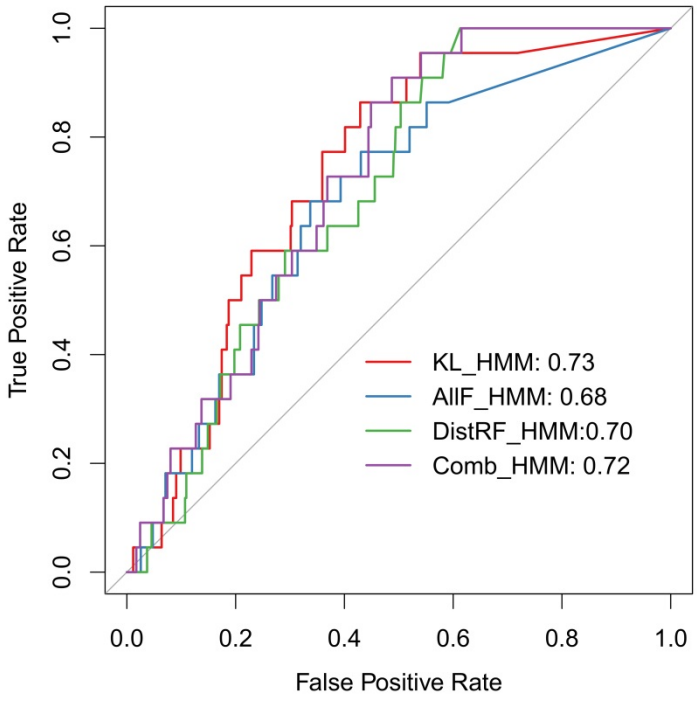

Figure 30 ROC curves for the four HMM models fit with distributional parameters and the combined HMM model from study 1. The labels in order correspond to an HMM model using features selected by KL distance as a multinomial observation (KL_HMM), an HMM model containing all of the distributional features as a multinomial observation (AllF_HMM), a model that integrated the distributional contextual features into the HMM (DistRF_HMM), and the combined HMM model from the previous study (Comb_HMM).

Table 13 Confusion matrix, TPR, FPR, AUC, and bootstrapped confidence interval for the distributional context models and the Combined HMM model from the first study.

Model	*True Negatives*	*False Negatives*	*False Positives*	*True Positives*	*FPR*	*TPR*	*AUC*	*Bootstrapped 95% CI*
ALL F HMM	570	7	290	15	0.34	0.68	0.68	(0.58 - 0.78)
Dist. RF HMM	337	1	523	21	0.61	0.95	0.70	(0.62 - 0.78)
KL HMM	491	3	369	19	0.43	0.86	0.73	(0.65 - 0.82)
Combined HMM	441	2	419	20	0.49	0.91	0.72	(0.65 - 0.80)

The ROC curve results suggest that the KL divergence feature subset and the independent context model structure (KL HMM) perform best. Bootstrapped significance tests comparing each of the models suggested that none of the differences were significant. Although the difference between the KL HMM and the combined HMM from the previous chapter is not significant there are several regions where the addition of context by distributional features seems to benefit classification performance. Like the SAX

HMM model, the improvement from the KL HMM is a result of a reduction of false positives in several regions of the ROC curve. Although the difference is not statistically significant, the model containing all the features in a multinomial distribution shows a substantial decrease in AUC (0.05). This decrement is interesting because it seems to confirm the idea that an increase in dimensionality of an observation can undermine model performance. Similarly, the structure that integrates the distributional features into the random forest meta-feature generation step also performs slightly worse that the KL divergence subset. This result suggests that considering contextual features independently of those associated with drowsiness is more beneficial than combining the information in a single observation.

6.3 Discrete Fourier Transform measures

Discrete Fourier Transform (FFT) features form somewhat of a middle ground between distributional and SAX features. FFTs decompose a sample of univariate time-series data into a sum of a set of sine waves of increasing frequency. After this conversion the original time-series can be reduced by the assumption that the majority of information about the time-series shape is concentrated in the first few components of the sum (Agrawal et al., 1993). A magnitude and phase pair can describe each of these components in the frequency domain. This conversion is similar to SAX in that it maintains information about patterns within the signal and that the signal can be reconstructed from the reduced components. It is similar to the distributional metrics in that it is continuous and does not require input values other than the number of components to retain. FFTs are valuable to explore for drowsiness detection given this position and the fact that they are commonly used as features in other drowsiness detection work (Lal et al., 2003; C.-T. Lin et al., 2008; Yeo et al., 2009). Despite this broad use, the application of Fourier features in drowsiness detection has been almost exclusively associated with EEG data. Although it seems like a feasible alternative there is very little evidence in the current literature supporting the use of Fourier features for defining road context from speed and acceleration variables.

6.3.1 Feature selection

The calculation of Fourier features requires specifying the number of Fourier components to retain. Generally the goal of this selection is to ensure reasonable reconstruction of the original signal. The definition of reasonable in this case is subjective. In this work 10 Fourier components were retained for each window, this resulted in 2 Fourier features for each window (corresponding to the phase and magnitude of the component). Like the other two feature generation methods, Fourier features were calculated for speed, lateral acceleration, and longitudinal acceleration separately and their calculation was based on each window rather than across the entire drive. This combination produced a total of 60 Fourier features for each window. This number of features creates severe dimensionality problems and requires a feature reduction method or an alternative modeling structure. In this case the KL divergence method discussed previously was applied to reduce the feature set. As with the distributional features, four features were retained. Figure 31 shows the KL divergence for each of the Fourier features along with the threshold used to retain features. In this case there is not a clear cutoff point, so the threshold was selected in order to balance the contributions of the features with dimensionality. Furthermore an exploratory analysis suggested that models retaining four features outperformed those with three, five, or six features. Interestingly these features consisted of only lateral acceleration phases and magnitudes and included the second component phase and magnitude, the third component magnitude, and the fifth component magnitude.

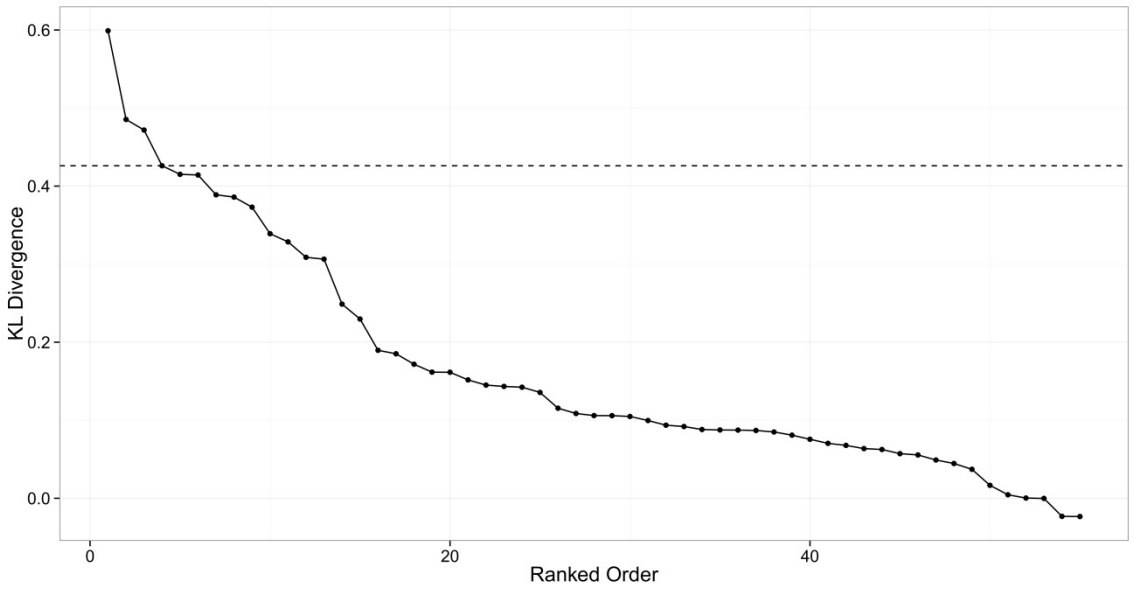

Figure 31 KL divergence for the Fourier features. Each point on the plot represents a feature and the dotted line indicates the cutoff threshold used to select retained features.

6.3.2 Model structures

A similar pair of modeling structures to the distributional features was explored here. The structures consisted of one structure that used the KL divergence features in a multinomial distribution and considered these features independently of steering and pedal features, and one structure that integrated the Fourier features directly into the random forest model along with the pedal and steering features. As with the distributional features the goal of these two structures was to isolate the importance of considering context independently versus considering context together with drowsiness indicators.

6.3.3 Model results and evaluation

The combination of model structures and feature generation methods consisted of fitting two models with Fourier contextual features, one where a feature subset selected by KL divergence was integrated into an HMM model through a multinomial distribution that was independent of the pedal and steering random forest (KL_HMM), and one model where the Fourier feature set was included in a random forest model with pedal and steering features prior to integration with the HMM (FFTRF_HMM). Figure 32 and Table 14 show the ROC curves and prediction statistics on the held aside test set for these two models along with the combined HMM (Comb_HMM) model from the previous chapter.

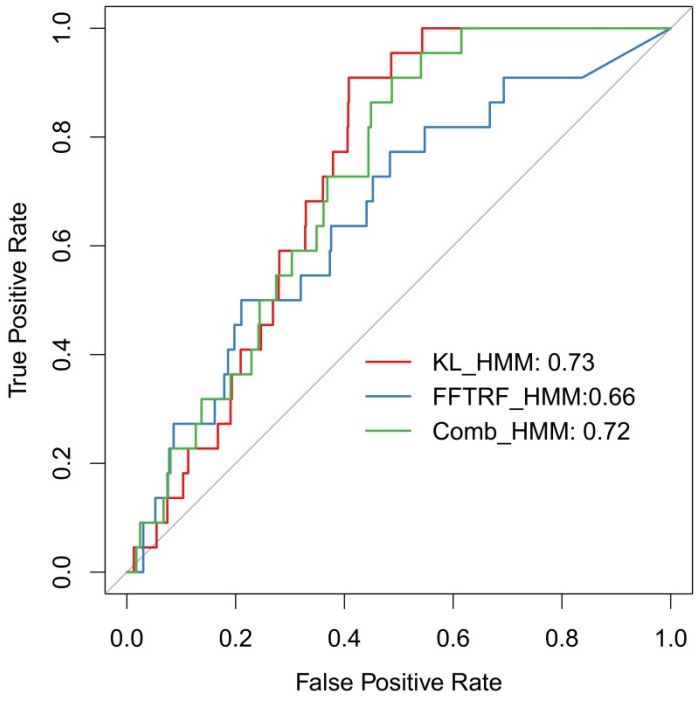

Figure 32 ROC curves for the four HMM models fit with Fourier contextual features and the combined HMM model from study 1. The labels in order correspond to an HMM model using features selected by KL distance as a multinomial observation (KL_HMM), a model that integrated the distributional contextual features into the HMM (FFTRF_HMM).

Table 14 Confusion matrix, TPR, FPR, AUC, and bootstrapped confidence interval for the FFT feature context models and the Combined HMM model from the first study.

Model	*True Negatives*	*False Negatives*	*False Positives*	*True Positives*	*FPR*	*TPR*	*AUC*	*Bootstrapped 95% CI*
FFT RF HMM	151	2	709	20	0.82	0.91	0.70	(0.62 - 0.78)
KL HMM	509	2	351	20	0.41	0.91	0.73	(0.67 - 0.80)
Combined HMM	441	2	419	20	0.49	0.91	0.72	(0.65 - 0.80)

In terms of AUC value, the KL divergence feature subset HMM outperforms both the Combined HMM (by 0.01 AUC) and the random forest HMM with Fourier, Pedal, and Steering features (by 0.03). Bootstrapped significance tests show that this difference is not significant, but the ROC curves clearly show several regions where KL HMM has either a higher sensitivity or a lower false positive rate at the same false positive rate or sensitivity respectively. This difference is also highlight at the threshold for maximum sensitivity and specificity in Table 14, which shows that the KL HMM can detect the same number of drowsiness instances while reducing the false positives rate. As with the Distributional feature

case, it is interesting that the model that integrates contextual features into the random forest rather than considering them separately performs worse than the KL HMM. Together these results might suggest that the value of speed and acceleration data in drowsiness detection relates to their ability to characterize the relationship between on-road environments rather their ability to identify drowsiness itself.

6.4 Comparing contextual feature generation methods

This study was motivated by the fact that the temporal relationships exploited by the HMM drowsiness detection algorithm did not eliminate false positive classifications in driving contexts where a drowsy-related lane departure was highly unlikely. The goal of this study was to explore feature generation methods that could capture context and integrate it into HMM model predictions and evaluate the hypothesis that such an integration would increase detection performance relative to an HMM model ignorant of context, as well as the two benchmarking algorithms. This impact was evaluated with ROC and smooth ROC curves, shown in Figure 33, from the three best performing models associated with each feature generation method were plotted alongside of the curves for HMM from the previous chapter, PERCLOS, and the static random forest benchmarking model. The left plot suggests that there is not a substantial difference between the three feature selection methods although the AUC results show that the SAX model has the highest AUC. Bootstrapped significance tests show that none of the differences in AUC are significant. Furthermore there is not a substantial difference between any of the contextual models and the HMM model although the AUC does indicate a slight improvement. The estimates from the smooth ROC curves clarify this difference and do suggest that the SAX and FFT models would produce better detection performance over a large portion of the ROC curve. Encouragingly these models are expected to substantially outperform the static random forest model and PERCLOS for almost the entire range of thresholds.

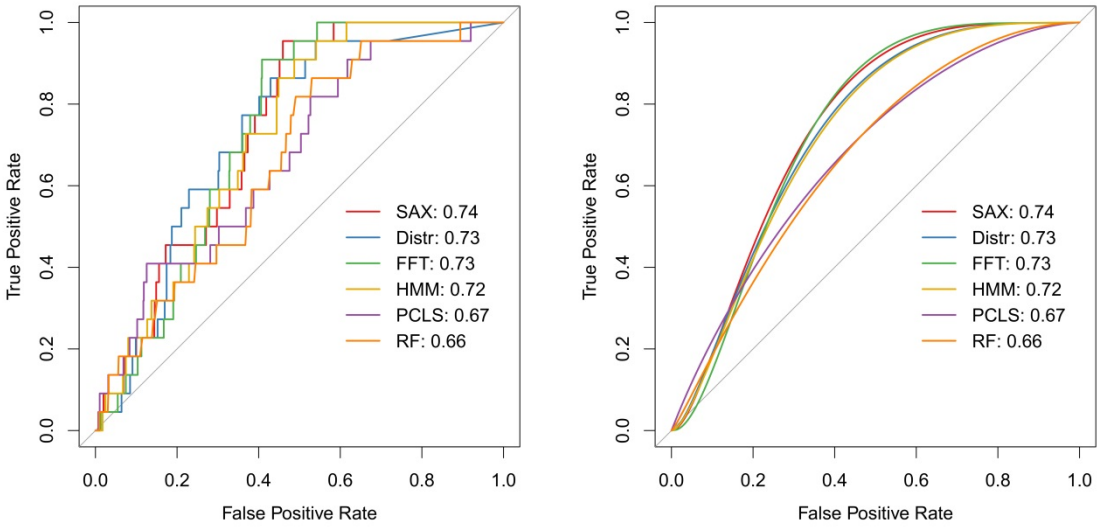

Figure 33 ROC (left) and smoothed ROC (right) curves for the SAX, Distributional, and Fourier context inclusive models beside the HMM model from the first study, PERCLOS, and Random Forest models.

6.4.1 Evaluating success and model limitations

Despite the fact that the classification improvement between the contextual models and the HMM model without context is not significant, it is still beneficial to analyze the improvement and to identify its source. The intuition behind including context suggested that it would reduce the number of false positives compared to the HMM without context. Therefore it is beneficial to revisit the analysis of false positives from the previous chapter with the best performing contextual inclusive model, the SAX HMM, and the best performing model from the previous chapter, the combined HMM or simply HMM. Figure 34 shows a histogram of false positives by simulator event for the two models. The counts are arranged by color with the black bars corresponding to false positives present in both the HMM without context and the SAX HMM and the grey bars corresponding to false positives eliminated by the addition of context (i.e. not present in the SAX HMM predictions). The false positives were calculated at the threshold that maximized the sum of sensitivity and specificity for the HMM model, and a threshold that produced the same sensitivity from the SAX HMM. This later threshold was selected because a comparison at the maximum sum of sensitivity and specificity for both models would confound the sensitivity by allowing the SAX HMM model to predict both fewer true positives but also fewer false positives.

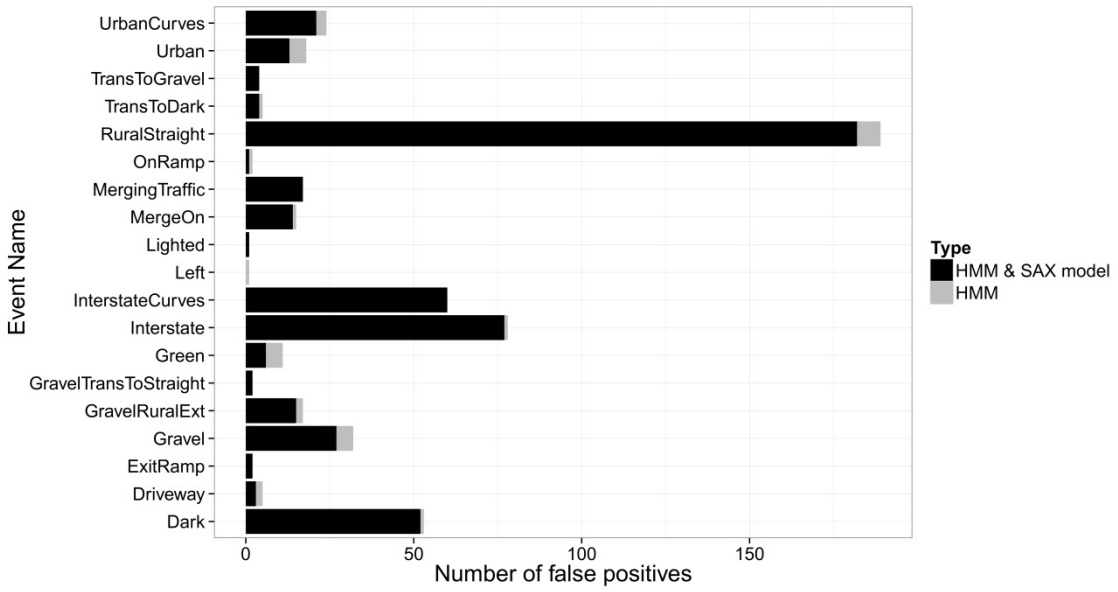

Figure 34 Histogram of false positives removed by the SAX model (grey) relative to the HMM model from the first study and those that remain in the prediction (black). Note that the False positives were identified based on a threshold associated with the highest sum of sensitivity and specificity.

The results in Figure 34 are encouraging in several respects, namely decrease in false positives associated with the Urban Curves, Urban, and Merge On events. Unfortunately none of these changes is significant according to a Fisher's Exact Test. These reductions are consistent with the hypothesis that the SAX contextual features are effective at identifying events where a drowsy-related lane departure is highly unlikely.

Beyond the encouraging results, the figure also shows the models primary limitation, which is illustrated specifically in the large amount of false positives associated with the Rural Straight, Interstate curves, Interstate, and Dark events. These events are characterized by long periods of mostly straight driving with very little input required from the driver. This limitation can be understood through the SAX input to the model, i.e. the word frequency across drowsy and awake states in the training data. Figure 35 shows a histogram of SAX speed word frequencies by event for the three letter alphabet and three letter word length input to the SAX HMM algorithm. This alphabet size means that "a" corresponds to higher speeds, and "c" corresponds to slower speeds. The figure shows that the SAX context effectively partitions higher consistent speed events that are associated with drowsy-related lane departures from more variable speed contexts that are not likely to result in a drowsy-related lane departure. Furthermore

there seems to be a distinct signature associated with each of the events in the figure. Thus SAX effectively captures on-road context although this context itself does not provide significant clarity for drowsiness. The limited benefit of adding SAX context to the algorithm suggests a need for a deeper level of context that can capture driving-maneuvers.

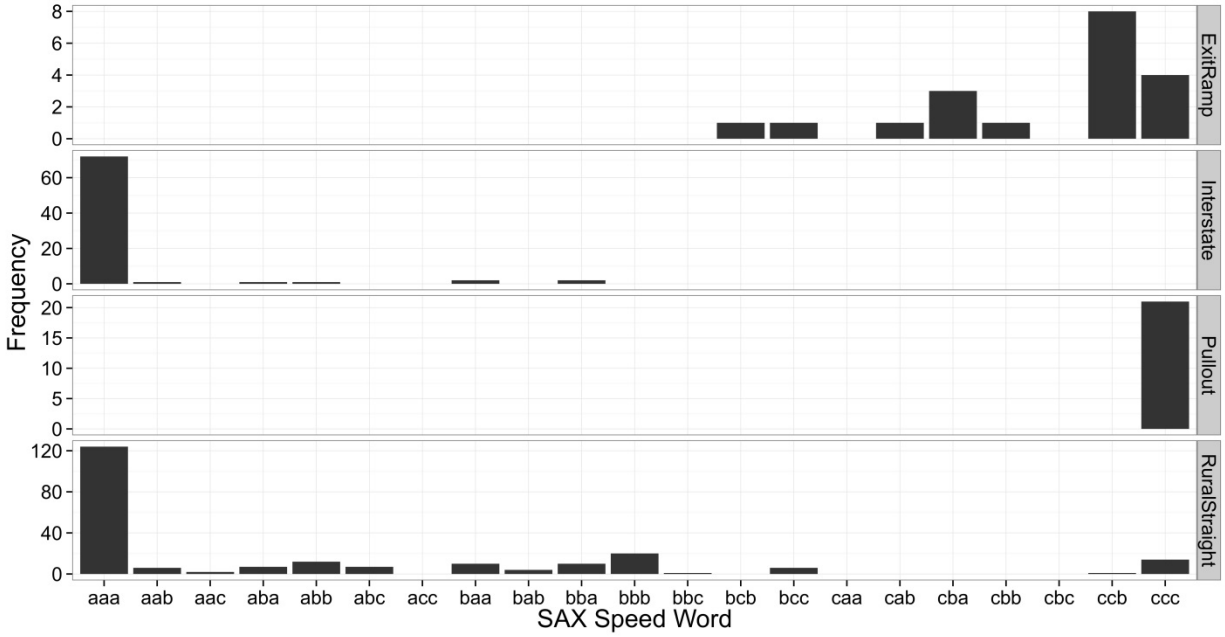

Figure 35 SAX speed word frequencies by event for a subset of the driving events.

6.5 Discussion and theoretical implications

The goals of this study were to evaluate the effect of integrating context into a dynamic graphical drowsiness detection algorithm and explore various time-series methods for generating driving contextual features. Although the results do not suggest that such integration substantially improves detection performance, there is some evidence that such integration reduces false positives relative to a dynamic graphical model that does not consider context. Beyond the simple classification improvement this study, like the work in the previous chapter, provides insights and contributions to the drowsiness detection literature, the field of driving safety, and the general predictive model literature.

The primary lesson from this work relative to the drowsiness detection literature is the combination of defining context by speed and acceleration and integrating this context into a graphical

detection algorithm may reduce false positives relative to a model that does not consider context. There is also some evidence that the feature generation method does not have a substantial impact, although projections of detection predictions across a large sample suggest that feature generation methods that retain the shape of the signal, i.e. SAX and FFT, will outperform distributional feature based methods. Furthermore structures that consider this context independently from drowsiness indicators seem to outperform those that do not. The results also suggest that more substantial performance gains would be produced by a contextual definition that both partitions context at the macro (road type) level, as well as the micro (maneuver) level.

The performance limitations of the model related to this context also have implications in the wider field of driving safety research. At the theoretical level, the false positive frequency results suggest that associating drowsy driving crashes with types of roadways is not sufficient to describe the true nature of the impact of drowsiness on driving. In essence, inside of the "driving on a rural highway" context is an additional layer of unexplained variance that might be used to enhance both the definition of a drowsiness-related crash, but also provide insights to the process of drowsiness in general. At a practical level this study contributes to the growing body of work on the value of SAX for transportation analysis (McDonald, Lee, Aksan, et al., 2013a, 2013b; McLaurin et al., 2014). The specific benefit of SAX in this case is its ability to capture context in a concise form that can be easily translated into many other statistical modeling analyses. More generally, the performance of SAX and FFT relative to distributional measures suggests that future analyses would benefit from exploring a wider range of feature generation and dependent measures.

The importance and value of SAX observed here might also be extended to the broader predictive modeling literature. It seems that SAX is a viable alternative for time series feature generation that might benefit a wider range of applications. Unfortunately this analysis did not provide significant insight into the input parameter selection however other studies might be able to identify benefits with a similar approach to the one employed here but applied to a larger dataset. Although SAX slightly outperformed the other two feature generation methods, the analysis suggested that this difference was not substantial

and there is very little evidence to suggest that it is generally superior to the other two feature generation methods.

Overall, the results observed here are encouraging in that they seem to both give a clear indication of the value of road context in detecting drowsy-related lane departures, but also provide a seemingly clear path to algorithm improvement. The goal of this path should be to essentially further partition the contextual definitions to access the micro or maneuver level speed and acceleration behavior. Such an analysis should either demonstrate the limits of this contextual definition method or illustrate the maneuver level context associated with drowsiness and drowsy-related lane departures.

Chapter 7 Analyzing Contextual hierarchies for drowsiness detection

The previous studies show that dynamic graphical models improve drowsy-related lane departure detection relative to static models and that adding context to dynamic graphical models reduces false positives relative to dynamic graphical models without context. However all of the previous models demonstrate an unacceptable false positive rate at the threshold that detects all drowsy instances. The majority of these false positives are centered on events where the road type (or macro context) suggests that drowsy-related lane departure is likely. The reason for this limitation is likely due to the model's failure to capture micro level context or driving maneuvers within each road context. Practically these maneuvers refer to actions such as lane changes or responses to other vehicles, which occur at a different frequency when a driver is drowsy (MacLean et al., 2003). Integrating this micro level context into the detection algorithm, should improve the differentiation between drivers who are simply driving on a road segment that may induce drowsiness and those that are drowsy driving on a similar segment.

Identification and integration of these maneuvers from real-world data is a difficult problem. However the results from the previous study and driver theory suggest that speed and acceleration behavior might provide the data necessary to identify driving context. One example of this process is shown in Figure 36. The top plot in the figure depicts a full drive of speed data with a segment with similar macro-context highlighted in yellow. The bottom plot shows the data from the yellow segment normalized to the segment mean. Inside of that lower segment, micro-level contextual events are highlighted. This figure encapsulates the contextual knowledge built into the algorithms discussed in the previous chapter (the yellow segment of the top plot), and the goal of this study, to capture the purple, green, and pink shaded regions and integrate them into a detection algorithm.

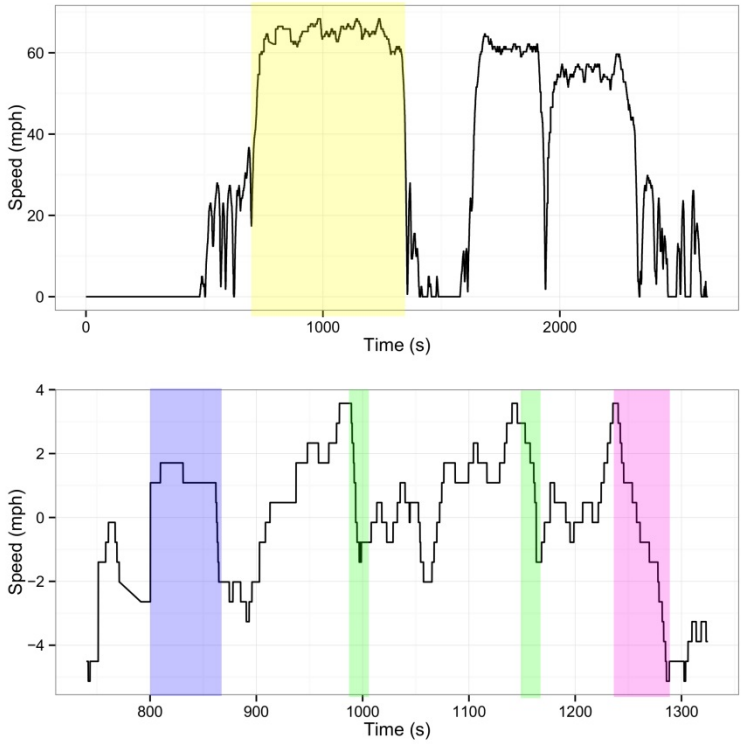

Figure 36 Speed data for a single drive (top plot) and a driving segment (bottom plot) with highlights corresponding to contexts that could be relevant to drowsiness detection.

Accomplishing this identification and integration requires the application of a time-series feature generation method capable of capturing sequences or segments of data. The success of SAX in the previous study suggests that it would be a valid tool for this analysis because it can capture such segments in a concise form and produced the highest AUC value observed in the previous study. The pursuit of SAX and the definition of micro-context introduces several additional challenges namely, the proper window size for calculating the micro context windows, the micro-level SAX input settings, and the model structure. The goal of this study is to explore the parameter space surrounding these challenges and ultimately develop a detection algorithm that exceeds the performance of the previous models.

7.1 Model settings and exploration space

Given the absence of extensive literature regarding the definition of maneuvers in real time, it is difficult to precisely define window sizes and SAX input for this analysis. The goal of this exploration is to provide a starting point for these parameters by selecting from a wide range guided by the analyses in

the previous chapters. The window sizes explored in this study: 5 s, 10 s, and 30 s, attempt to identify a range of maneuvers, while maintaining enough relevant information to classify behavior. Similarly the SAX input parameters explored in this study reflect knowledge gained during the feature selection analysis and the previous chapter. As in the previous chapter SAX was applied to speed, lateral acceleration, and longitudinal acceleration although for this study the data were normed to the mean of the window rather than a global measure. Initially two word lengths: three, five and four alphabet sizes: three, five, seven, and, nine were explored. However preliminary analyses suggested that the five-letter words did not repeat across drives, meaning that the five-letter words created a strong likelihood of model overfitting. Accordingly this setting was removed from future analyses. Combining the window sizes and alphabet size input to SAX produced 12 different settings. Each of these settings was applied to the speed and the accelerations producing a set of three words for each micro window. This definition of features at the 5 s, 10 s, or 30 s level creates an additional challenge because the time windows do not naturally align with the 60 s windows used in previous studies to predict drowsiness and define context. Thus these features had to be aggregated across a set of smaller windows into features that represented the full 60 s. This process was accomplished by counting the frequency of each word from the smaller windows and then adding these frequency counts to a feature vector that was used in the random forest meta-feature generation portion of model training. This frequency count strategy is common in other symbolic analyses (Leslie, Eskin, & Noble, 2002) and was employed by McDonald et al. (2013a). Figure 37 illustrates the SAX micro-feature creation and model integration.

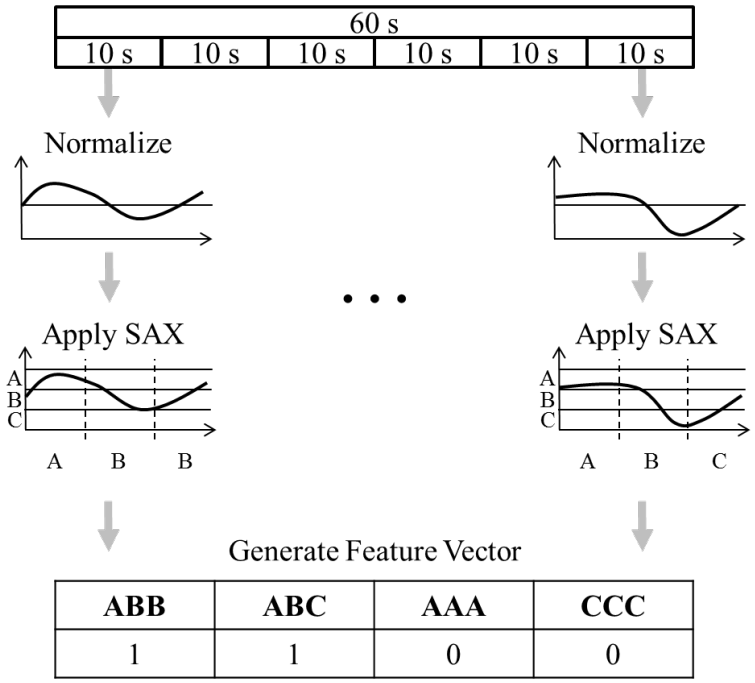

Figure 37 Demonstration of the SAX maneuver-level feature generation process.

In addition to the exploration of SAX and window size settings, this study explored four different model structures—depicted in Figure 38, which were all based on the HMM models from the previous studies. The structures differed in their inclusion of road context, as defined by the 60s globally normalized SAX features from the previous study, and their treatment of the maneuver level context. This maneuver context was treated in two ways. The first method involved including the maneuver context in the random forest meta-feature training with the steering and pedal features. This structure mimics the structure utilized for the distributional parameters and Fourier features in the previous study and is depicted in the top two diagrams in Figure 38. The second method of maneuver treatment consisted of training three separate random forest meta-features and incorporating them into the HMM model as a multivariate observation. These structures align with the Combined HMM model from the first study. Although the difference between these structures is subtle, it is an important comparison because it tests the value of the maneuver level context when considered independently of the pedal and steering features versus dependently. The road context inclusion and exclusion allows one to test the value of road context

itself and the hypothesis that with road and maneuver context combined, one can fully partition situations where a drowsy-related lane departure is likely.

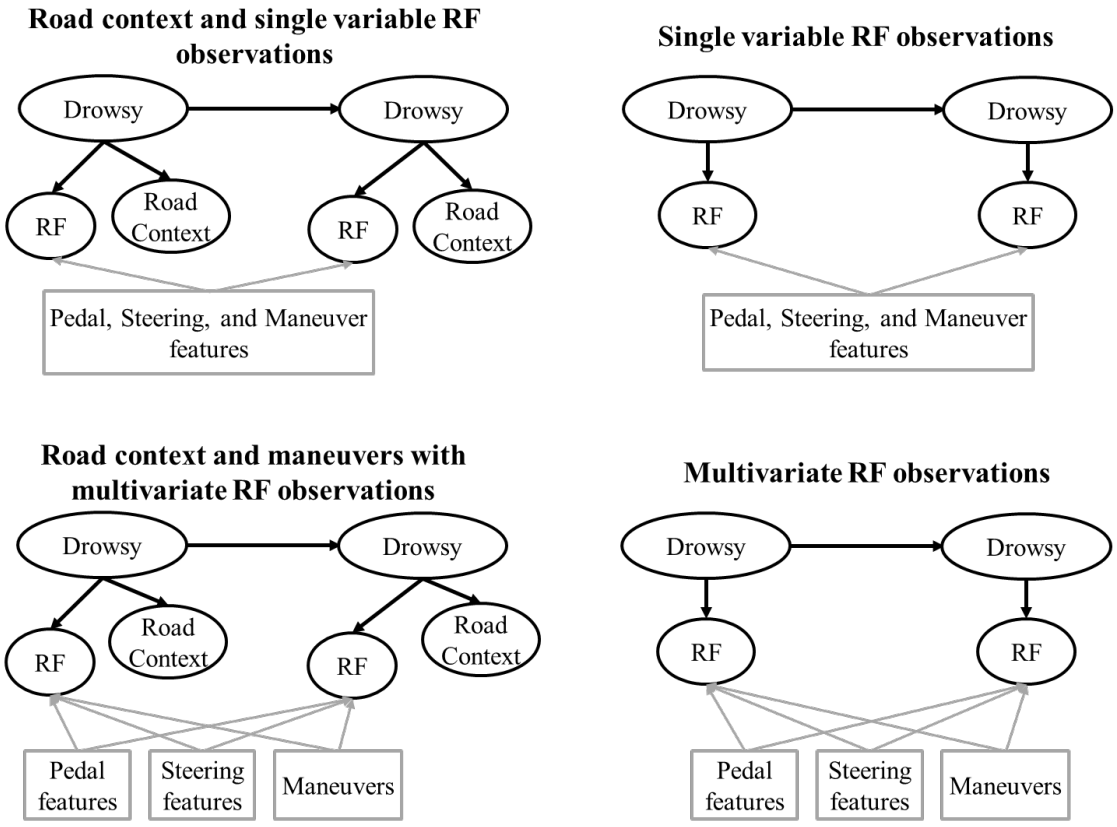

Figure 38 Model structures examined in this study. Note the grey boxes indicate features used to train the random forests and multiple grey boxes correspond to multiple random forest models.

7.2 Model fitting results

The 4 structures and 12 SAX and window input settings produced a total of 48 different models. These models were tested on the held aside test set and the AUC results with bootstrapped 95% confidence intervals are shown in Figure 39. The figure shows a plot for each structure and window size pairing and four points corresponding to the various alphabet sizes investigated in the study. The model with the highest AUC: Multivariate random forest with no high level road context, with a window size of 10 s, and an alphabet size of 3 (MVRF), is highlighted in red in the figure.

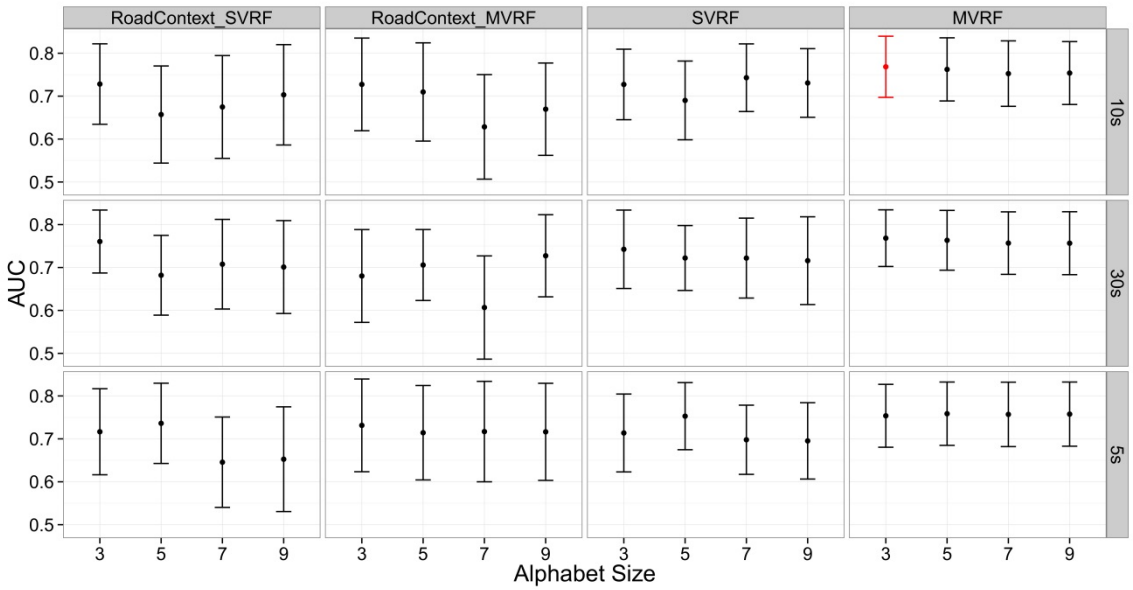

Figure 39 AUC results with 95% bootstrapped confidence intervals for the 48 models analyzed in this study tested with the held-aside test data, each plot corresponds to a model structure and window size. The structure abbreviations correspond to: Single RF with road context, Multiple RF with road context, single variable RF, and multiple variable RF. The model with the highest AUC, MVRF structure with a window size of 10 s and an alphabet size of 3, is highlighted in red.

These results suggest that the combination of macro context and low level maneuver level context actually decreases the detection performance relative to a model with only maneuver context. Furthermore the combination increases the variance of model prediction. This difference is surprising because it almost directly contradicts the hypothesis that the combination of contexts would work together to increase drowsiness detection performance. One should note that the size of this difference is not significant based on the bootstrapped test, but the lack of improvement relative to additional model parameters still suggests that a model without high-level context should be preferred. Beyond the model structure effect there are no clear trends across the models except that in several cases the larger alphabet sizes, seven and nine, show a decrease in AUC. This effect is particularly noticeable in the Road context and single random forest (RoadContext_SVRF) model with five-second windows.

7.3 Comparing contextual models to HMMs and benchmarks

This study was motivated by the fact that the temporal relationships and the contextual features introduced in the SAX HMM drowsiness detection algorithm did not eliminate false positive classifications in driving contexts where a drowsy-related lane departure was likely. The goal of this study

was to explore contextual driving features at two different time scales representing high-level road context and low-level maneuver context, with the hypothesis that the combinations of these two levels of features would allow the algorithm to identify and further separate instances where the driver was simply driving on a road context that could induce drowsiness from those where the driver was actually drowsy. The previous analysis did not provide support for this hypothesis rather it suggested that the lower level alone may still improve drowsiness detection. This improvement was analyzed further by plotting ROC and smooth ROC curves from the model with the highest AUC from the previous analysis, MVRF with an alphabet size of 3 and a window size of 10 s (maneuver-level context HMM), alongside of the curves for the SAX HMM, and HMM from the previous chapters, PERCLOS, and the static random forest benchmark model. Figure 40 shows these plots.

The results of the ROC plots in the left figure show that the maneuver-level context (MLC) HMM does not provide a benefit in the lower regions of the curve, where the FPR is less than 0.2, but provides a substantial benefit for much of the rest of the curve, particularly over the PERCLOS (PCLS) and Steering Random Forest (RF) algorithms. The smooth ROC curve plot accentuates these differences, but also suggests that the maneuver level context and SAX HMM would perform quite similarly over a larger data set. The AUC values for each curve relative to the maneuver-level context model were statistically evaluated with a bootstrapped significance test with 2,000 replicates. The bootstrapped test results suggested that there was no significant differences in AUC between the MLC HMM algorithm and any of the other algorithms at the $p = 0.05$ level however the difference between the MLC HMM and the Steering random forest algorithms was significant at the $p = 0.1$ level, ($D(2000) = 1.75$, $p = 0.08$).

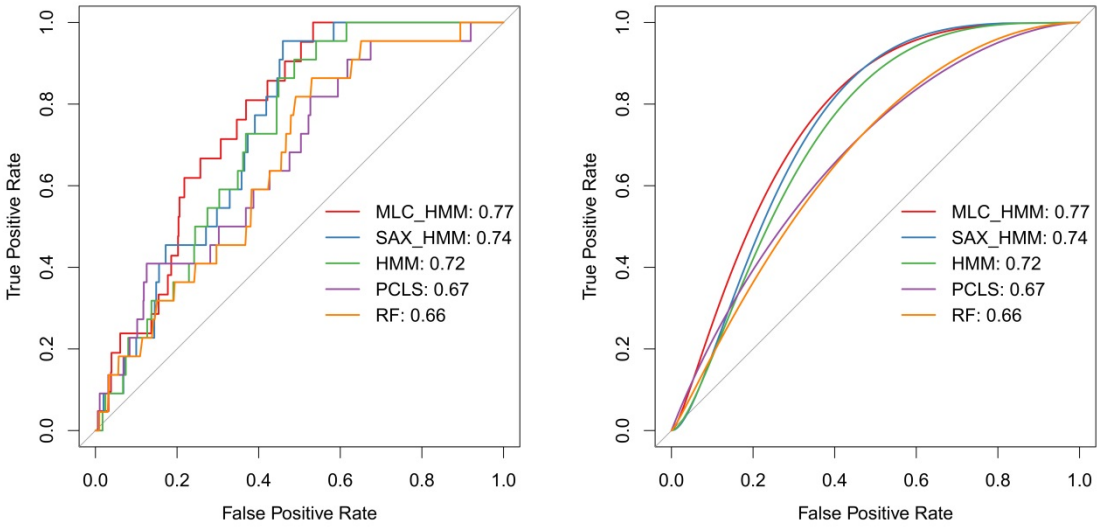

Figure 40 ROC (left) and smoothed ROC (right) curves for the Maneuver level context (MLC) HMM model beside the SAX HMM model from the second study, HMM model from the first study, PERCLOS, and Random Forest model.

The most interesting result in these plots is that the maneuver-level context HMM has the lowest false positive rate of all the algorithms for detecting all of the drowsy-related lane departures. Table 15 shows this result along with a full confusion matrix and bootstrapped 95 % confidence intervals for all of the models displayed in Figure 40. The differences in false positive rates between the maneuver-level context HMM and the other models were statistically evaluated using McNemar's test (Dietterich, 1998). The results showed that the difference in false positives was significant in all cases: SAX HMM ($M(1) = 17.88$, $p = 0.002$), HMM ($M(1) = 9.57$, $p < 0.001$) , PERCLOS ($M(1) = 110.67$, $p = 0.001$), Steering angle random forest ($M(1) = 124.47$, $p < 0.001$).

Table 15 Confusion matrix, TPR, FPR, AUC, and bootstrapped confidence intervals for the Maneuver level context (MLC) model, the SAX HMM algorithm, and HMM algorithm from the previous studies, and the PERCLOS and Steering random forest benchmarking algorithms.

Model	*True Negatives*	*False Negatives*	*False Positives*	*True Positives*	*FPR*	*TPR*	*AUC*	*Bootstrapped 95% CI*
MLC HMM	392	0	448	21	0.53	1	0.77	(0.70 - 0.84)
SAX HMM	350	0	490	21	0.58	1	0.74	(0.67 - 0.81)
HMM	324	0	516	21	0.61	1	0.72	(0.65 - 0.80)
PCLS	68	0	772	21	0.92	1	0.67	(0.56 - 0.78)
RF	89	0	751	21	0.89	1	0.66	(0.56 - 0.76)

7.4 Analyzing false positives and model limitations

The ROC and confusion matrix results show that adding maneuver-level context into the model reduces false positives relative to models that only include high level context, models that only consider time, and static models. The results demonstrate the value of including maneuver-level context in a detection algorithm however they do not address the specific benefits of the inclusion relative to road contexts. Figure 41 shows a histogram of false positives, at the threshold that correctly predicts all drowsy cases, by event for the MLC HMM model and the SAX HMM model from the previous study. False positives present in both models are shown with black bars and false positives present in the SAX HMM model, but not the MLC HMM are colored grey.

The results shown in the figure are encouraging in that they show the MLC HMM model further reduces false positives associated with the Rural Straight event as well as the Interstate and Dark events. A Fisher's exact test, employed due to the small sample size, showed that these differences were significant at the $p = 0.05$ level although the difference in the Rural Straight event was significant at the $p = 0.1$ level, $p = 0.07$. However beyond these reductions, the results are somewhat disappointing. The number of remaining false positives in the Rural Straight, Interstate, Interstate Curves, and Dark events suggests the need for further model additions and perhaps compromises on the mitigation side of the drowsiness mitigation system.

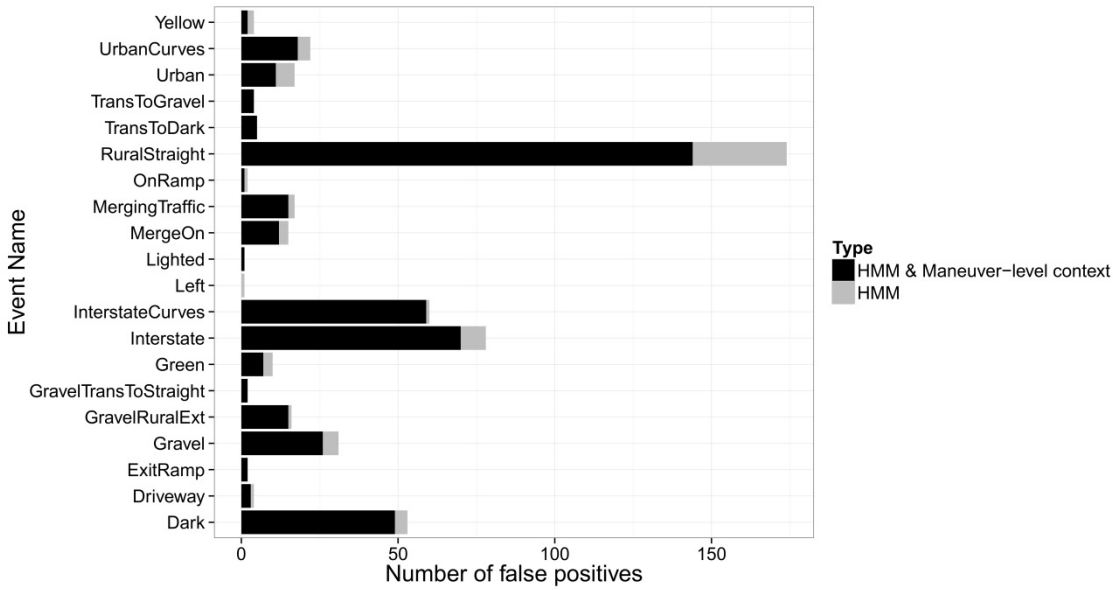

Figure 41 Histogram of false positives removed by the MLC HMM algorithm (grey) relative to the HMM algorithm from the first study and remaining false positives (black). Note that the False positives were identified based on a threshold required to detect all drowsy-related lane departures.

7.5 Discussion and theoretical implications

The goal of this study was to evaluate the effects integrating macro-level road context and micro-level maneuver context into an HMM-based drowsy detection algorithm. The results show that an algorithm containing only maneuver-level context produced better detection performance than algorithms with the combination of contexts and those with road context alone. Although this difference was not significant based on AUC, the maneuver-level context model did have a significantly lower false positive rate at the threshold that detected all of the drowsy instances. Even with this improvement the model produced a high false positive rate, particularly during driving events where very little driver input was required. Despite the fact that the results contradict the primary hypothesis and unimpressive performance of the model, this study provides several contributions to the drowsiness detection, drowsiness mitigation, drowsy driving, and predictive modeling literatures.

The primary result of this study relative to the drowsiness detection literature is the importance of including driving maneuver-level context relative to the broader road context on performance. In some sense adding broad road context directly in to the model seems to bias the model toward false positives on

segments of road where a drowsiness incident is likely to occur. Although maneuver-level context still retains this bias, the results here suggest it has a smaller effect. The lack of substantial difference between the structures explored here suggests a need for repetition of this analysis with a larger data set including a less structured series of drives. In a more general sense the performance of these context-based models relative to their non-contextual counterparts does provide rather strong evidence to support the importance of at least some level of driving context in all detection models. Further explorations should consider broader approaches to including this context, perhaps even training multiple models for each type of road context. In the context of drowsiness mitigation systems, these results inspire some level of pessimism. The algorithm designed in this study partitions situations characterized by driver steering and pedal behavior during particular contexts. The lack of separation between drowsy and awake drivers in these instances suggests that even state-of-the-art drowsiness detection algorithms will have non-zero false positive rates. Although this could be due to the strict definition of ground truth drowsiness as a drowsy-related lane departure and suggests a need for more thorough definitions of drowsiness, it is still important to consider the implications of imperfect algorithms. The implication of this for mitigation system designs suggests the need for tiered mitigation systems with an escalation of warnings, or those that increase driver monitoring with an initial false positive classification.

The results from this study can be extended beyond drowsiness detection and mitigation to a more general understanding of drowsy driving behavior. The increase in drowsiness detection performance derived from the inclusion of driving context into the model emphasizes the role of driving context in drowsy driving crashes, and provides support for theories that consider the role of context in driver behavior. The perplexing stability of false positives in the final algorithm might be interpreted as an indication of strong similarities in driving behavior between drowsy and alert drivers on certain roads. This similarity might in turn be used to explain the lack of consistent findings relating driving behavior to drowsiness (MacLean et al., 2003; Williamson et al., 2011). Furthermore the similarity suggests that some measure of road context should be included as a covariate in subsequent analyses of drowsy driving

behavior and provides impetus for further analyses of the link between subjective drowsiness and performance decrements.

From a general predictive modeling standpoint, it is interesting that the structure that combined random forests from each type of measure generally outperformed the model that compiled all of the features into a single random forest model. This could simply be an artifact of an ill fit between the inductive bias of the random forest and the partitioning of the training data, but it might also provide further evidence for the use of ensembles of models, particularly those that involve different types of measures. This idea warrants further exploration with a larger data set. The results here also provide further support for SAX as a feature generation method.

It is challenging to provide a general conclusion for this study. Although the results show an additional method to improve drowsiness detection algorithms, there is also evidence that highlights the limitations of this approach. The results of this study are encouraging in that they prompt further exploration into the integration of context and drowsiness detection models, but they do not highlight a clear path toward improvement. Despite this the study does provide several contributions across several literatures and perhaps a guideline and benchmark for further analyses.

Chapter 8 Additional analyses and explorations

Throughout this dissertation several choices were made that limited the level of exploration in the previous three chapters. The goal of this chapter is to address some of these limitations through extended explorations and provide directions for further model improvement or extended support of the claims made in previous chapters. This chapter focuses on four areas: analyzing overlapping windows for capturing drowsiness behavior and improving model timeliness, developing models to predict Retrospective Sleepiness Scores, analyzing the effects of loosening the constraints on the zero-one loss function, and model interpretation.

8.1 Window overlap analysis

The 60 s non-overlapping window approach used in the previous analyses is limited by the timeliness of predictions. In real time the models could make one prediction each minute, meaning that in the worst case an alert could be generated 59 s after a lane departure. One way to improve this timeliness is to employ overlapping windows, where each successive window contains some percentage of the same data as the previous window. The problem with sliding windows is their use violates the assumption of HMMs that all observations are conditionally independent. However there is some evidence to suggest that minor violations of this assumption may not affect detection performance (Vail, Veloso, & Lafferty, 2007). The evidence and limited timeliness of algorithm prediction motivates a revisit to the previous studies with sliding windows.

The sliding window analysis explored here was conducted with 50 % overlapping windows, meaning that half of the data in each prediction were repeated for the successive prediction and that predictions could be made every 30 s rather than once per minute. This sliding window analysis was applied to all three studies discussed in the previous chapters. Beyond the change in training data the analysis process was consistent. Figure 42, Figure 43 and, Figure 44 show the ROC curve results of these analyses. Each figure represents the results from a single study and the studies are presented in order. Generally these results are quite similar to those from the original studies. The first and third studies show

the exact same ordering of the models. The second study results suggest that the distributional features slightly outperform the SAX features, which contradicts the results from the original analyses, however the difference is small (delta AUC = 0.01). The SAX model also produces fewer false positives than the distributional parameters at the threshold that detects all drowsy-related lane departures. Interestingly this performance does not extend to the third study where the SAX model actually produces lower false positives than the maneuver-level context model at the threshold that detects all drowsy-related lane departures.

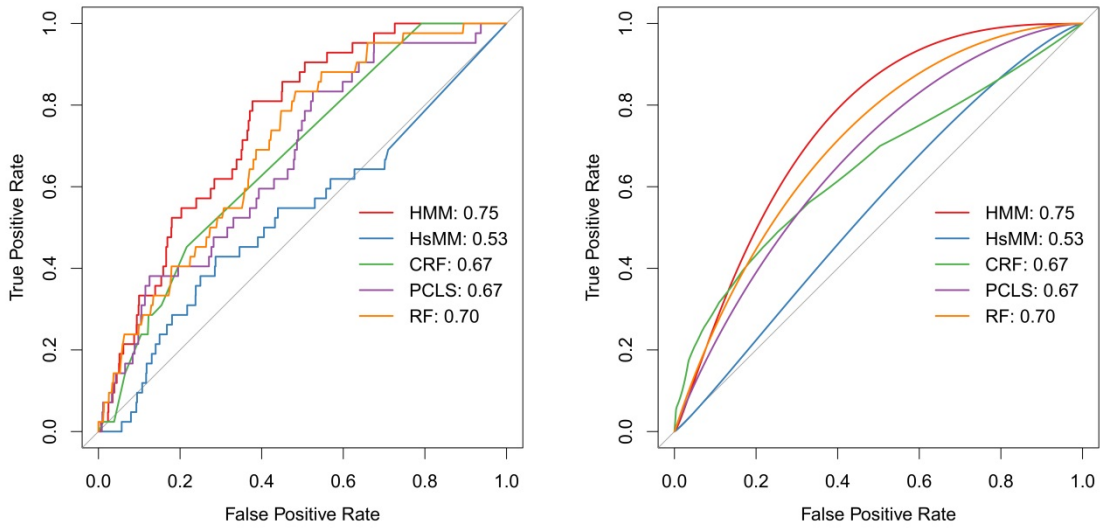

Figure 42 ROC and smooth ROC curves for the study 1 models trained on sliding window data.

Bootstrapped significance tests suggest that in the first study the only significant difference in AUC is the difference between the HMM and the HsMM models ($D(2000) = 3.34$, $p < 0.001$). This result is consistent with the non-overlapping windows. In contrast, the significance in the latter two studies changes between the non-overlapping and overlapping analysis. In the second study the difference between the SAX model and the PERCLOS model is significant at the p = 0.1 level, $D(2000) = 1.91$, p = 0.06. In the third study, the difference between the MLC HMM model and both PERCLOS and the static steering model was significant, $D(2000) = 2.48$, $p = 0.01$ and $D(2000) = 2.08$, $p = 0.04$ respectively.

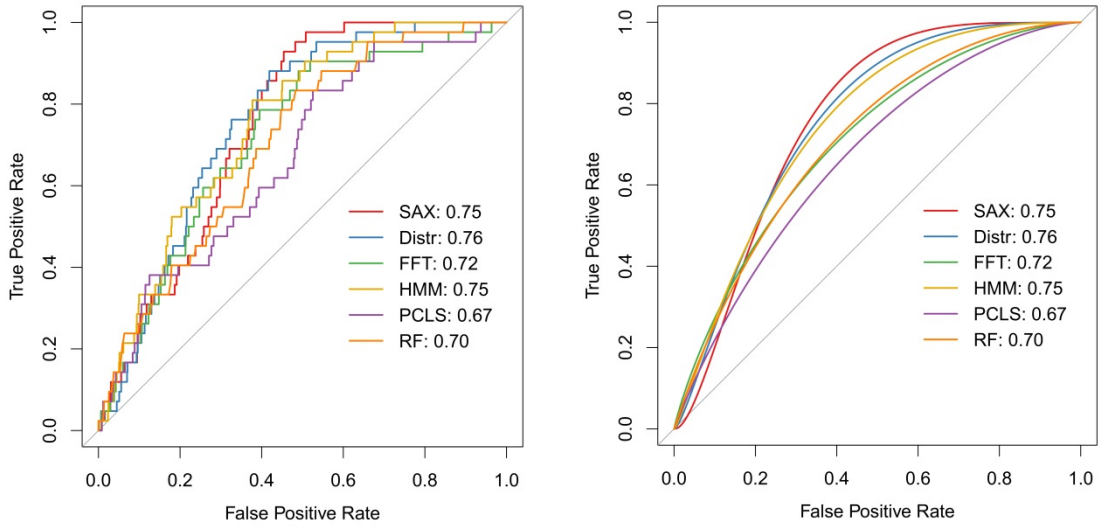

Figure 43 ROC and smooth ROC curves for the study 2 models analyzed with sliding windows

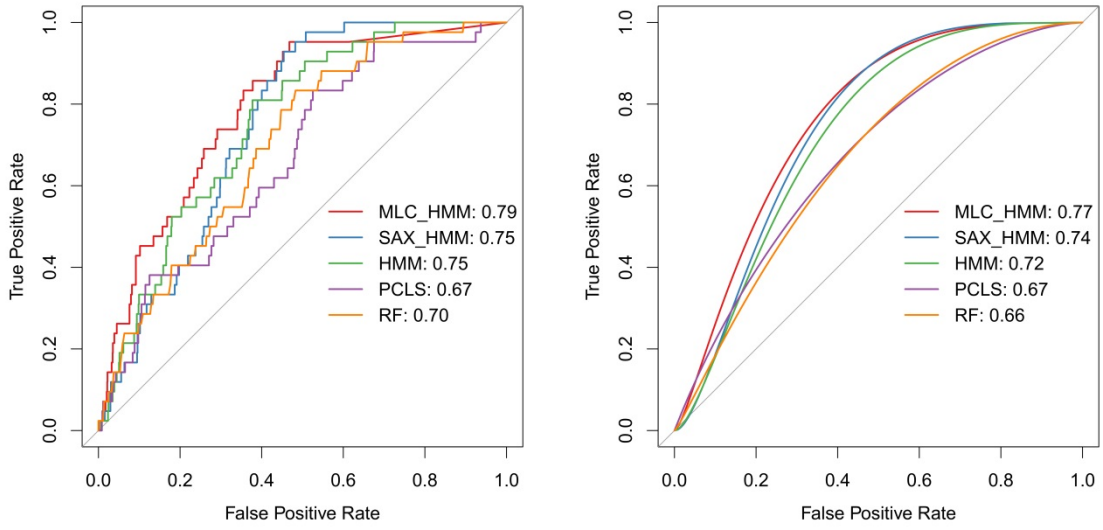

Figure 44 ROC and smooth ROC curves for the study 3 models analyzed with sliding windows.

It is interesting that compared to the non-overlapping window results, the window overlap actually increases performance. This increase is best reflected in the AUC of the HMM which improved from 0.72 to 0.75 between the non-overlapping and overlapping window analyses. The SAX and the Distributional parameter models from the second study also improved although the difference was not substantial (0.74 and 0.73 respectively to 0.75). The maneuver-level contextual model in the third study

also improved from 0.77 to 0.79. Although none of these differences were significant based on a bootstrapped significance test, the results show a promising addition to the modeling architecture.

These results are encouraging in that they represent a substantial improvement in the timeliness of the algorithm prediction without cost to the algorithm's prediction accuracy. Although it is unclear exactly how this benefit is achieved, the result may be an artifact of the use of random forest meta-features. These features originally served as a noise reduction tool, but they may also work to effectively reduce the effects of dependence of overlapping windows. This suggestion warrants further investigation with a larger and more robust data set.

8.2 Predicting Retrospective Sleepiness Scores with HsMM

The poor performance of the Hidden semi-Markov models in the analysis from the first study and the sliding window analysis is surprising and one of the more theoretically contradictory results found in the three studies. One possible explanation for these results is that the binary and strict definition of ground truth as drowsy-related lane departures inhibits the benefits of HsMM. Furthermore this definition is a simplification of the transition from alert to drowsy driving, which in actuality is a gradual descent through several states (Dinges, 1995). Expanding the hidden state space and definition of ground truth to accommodate these transitions might result in improved HsMM performance and further insight into driver drowsiness.

To evaluate the performance of HsMM across a more gradually transitioning ground truth measure, we performed an analysis using Retrospective Sleepiness Scores (RSS) as ground truth drowsiness. RSS is a subjective measure of drowsiness conducted immediately following each drive. The scale is measured from 1 to 9 and participants provided a score for each simulator event in the drive. For this analysis, RSS was slightly modified from this original measurement so that it was both consistent with the windowed data approach and simplified from a classification perspective. Consistency between event-based RSS was achieved by simply merging the event and window datasets so each window contained a single event and a single RSS score. The simplification of RSS consisted of reducing the

score to a one to three scale. This reduction was performed to balance the frequency of each score and provide some level of normalization across individual drivers. Beyond these conversions, the method of training and testing algorithms was consistent with the previous studies.

Two models were fit for this analysis: a HMM and an HsMM. The models used the same multivariate continuous density observations, which consisted of random forest meta-features developed from steering and pedal features. State durations for the HsMM model were modeled with Gamma distributions. These state durations and the associated transition probabilities represented the only differences between the two models. Following specification both models were trained using expectation maximization. The trained models were evaluated with the same held aside test set as previous drives. The primary evaluation criterion for the models was accuracy of predictions. This change is due to the fact that ROC curves cannot be generated for a detection problem with more than two hidden states. The accuracy of both the HMM and HsMM models were 0.52, which is not significantly better than random according to a Fisher's Exact Test ($p = 0.44$ for the HsMM, $p = 0.38$ for the HMM).

The limited predictive power of these models makes it difficult to generalize the results and connect them to theory however comparing the performance of the models in this analysis to those in the previous does suggest that RSS are not a valid measure of drowsiness. Given the similarity in study design, one would expect the steering and pedal features to provide similar predictive power and therefore decreases in the predictive power as a whole are likely due to inconsistencies in the measurement of drowsiness. Even with these constraints on measurement it is worthwhile to note that this study provides some further evidence that there is no benefit of including state duration in drowsiness detection algorithms. Continuing the same logic from the first study this may reflect regularity in state transitions into and out of drowsiness however the duration of these states is highly irregular and random.

8.3 Analyzing the ground truth and false positives

The zero-one loss function employed in the previous analyses of this dissertation strictly defines drowsiness as a window of data containing an uncorrected drowsy-related lane departure. This definition

is somewhat limited in that drowsy-related lane departures may occur after an extended period of extreme drowsiness, meaning that in some instances a driver may exhibit drowsiness but be classified as awake by the ground truth. Although there are several reasons for maintaining this strict definition it warrants further analysis, particularly with a focus on false positive classifications. The strict definition of ground truth effectively creates two types of false positives: cases where the driver is clearly awake and the algorithm predicts that the driver is drowsy and cases where the driver may be drowsy but did not have a lane departure and the algorithm predicts drowsy. Figure 45 shows an example of these two types of false positives and the somewhat ambiguous line between them. In each plot the figure shows the ground truth drowsiness with a dashed line and the predictions of the maneuver-level context HMM from the previous chapter at the threshold that correctly identifies all drowsiness instances. Note that in nearly all of the ground truth drowsiness instances, the algorithm predicts a lane departure prior to the actual occurrence. This highlights a potential weakness of the ground truth and in one sense indicates the need for more advanced definitions of drowsiness that are beyond the scope of this dissertation. Alternatively these ambiguous false positives motivate an exploration of a broader definition of drowsiness. The goal of this section is to perform such an exploration.

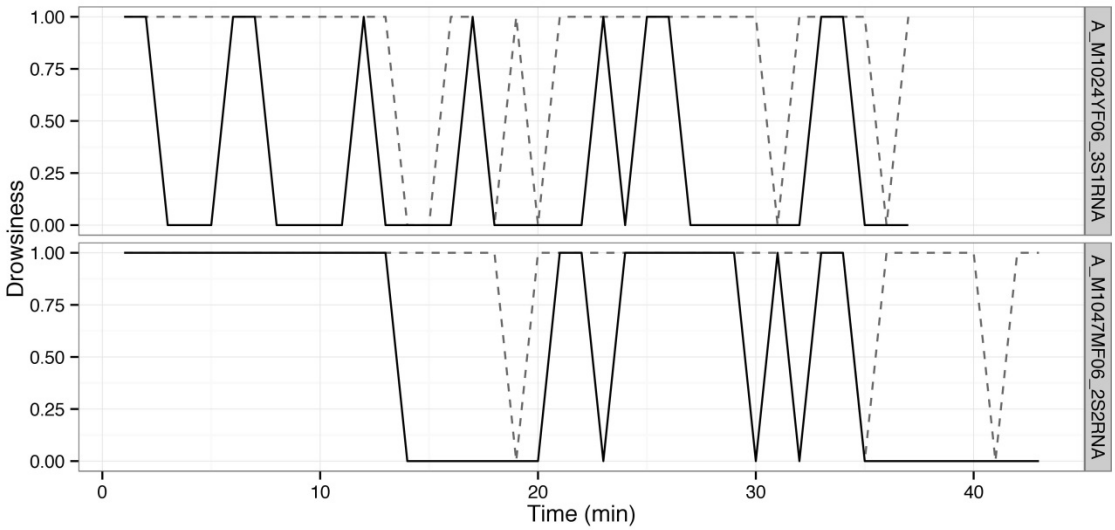

Figure 45 Temporal patterns of predictions for two drives with the MLC HMM predictions as a black line and the ground truth as a dotted line.

This analysis focused on a simple expansion of the ground truth, which includes both windows containing an uncorrected drowsy-related lane departure and the previous window. The analysis involved a single model, the maneuver-level context (MLC) HMM, which included steering and pedal behavior along with maneuver-level context defined by SAX as input. Each of these input variables was integrated into the model with an independent random forest meta-feature and the random forest votes were combined with a multivariate Normal observation density function. The remainder of the modeling and analysis process followed the previous chapters. The model generated by this process was compared to the MLC HMM model from the previous study using the original ground truth definition. Figure 46 shows the ROC curves for both models. The figure illustrates that despite the broadening of the ground truth definition, the new model actually performs worse than with the original definition. A bootstrapped significance test suggests that this difference in AUC is significant at the $p = 0.1$ level ($D(2000) = 1.65$, $p = 0.1$). Despite the lack of significance, the figure clearly shows that the model predicting the previous ground truth data outperforms the model with the new ground truth definition at nearly every threshold.

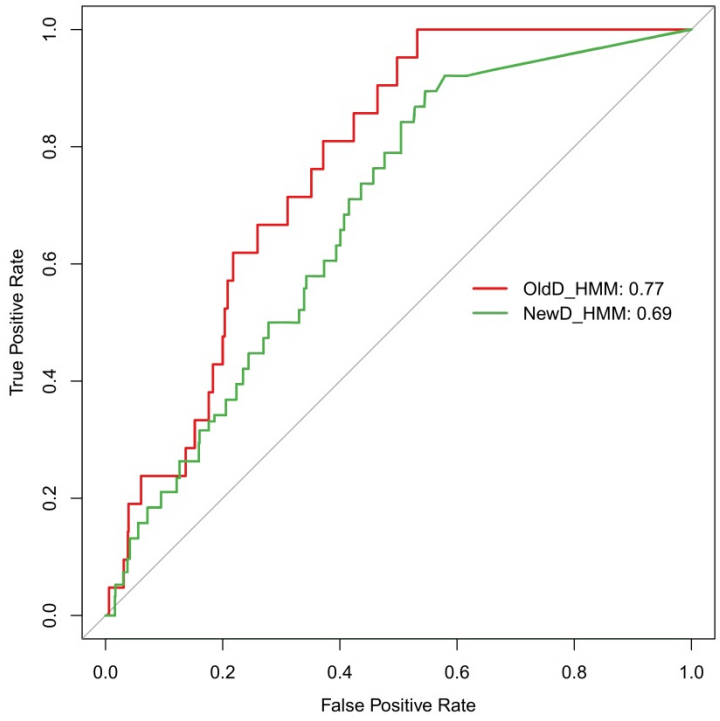

Figure 46 ROC curves for the original ground truth definition of drowsiness (OldD_HMM) and the new definition of drowsiness (NewD_HMM), which includes both windows containing an uncorrected lane departure and the preceding window.

These results suggest that the simple expansion of ground truth drowsiness is not a beneficial extension. They also highlight that the ratio of false positives from the windows preceding a drowsy-related lane departure to true positives activating at exactly the window of drowsy-related lane departures is approximately equivalent or at least not as extreme as expected. At a deeper level this change highlights the need for new and more robust definitions of ground truth drowsiness. Although the definition used in this work highlights the link between drowsiness and its consequences, it does not sufficiently address the process leading to drowsiness-related consequences. Future work should include this area of the definition, with a focus on tiered but discrete states.

8.4 Model interpretation

The layered combination of the random forest meta-features, contextual variables, and the HMM dynamics in the MLC HMM model provides little insight into the association between driving behavior and the model predictions. Such an association is critical to model interpretation because it provides a

measure of model validity. The goal of this analysis is to establish this association through plotting windows of unfiltered measures by model predictions for the MLC HMM model. Although this approach is somewhat limited by the fact that it does not illustrate the full process of the model, it does give some indication of the model's understanding of the raw data.

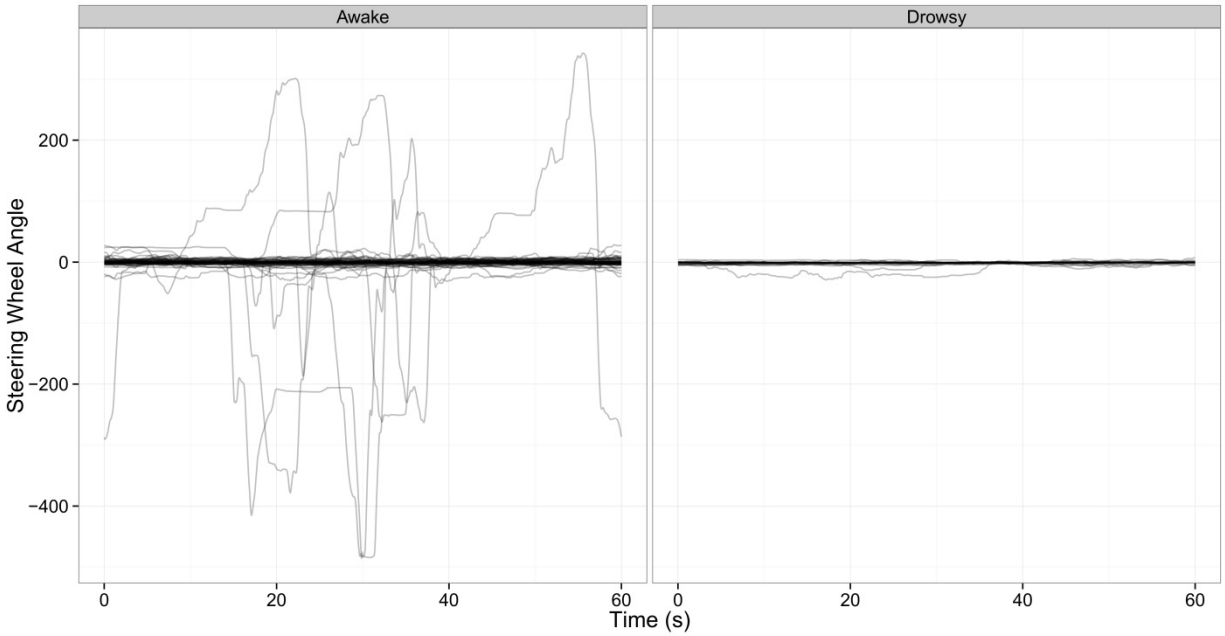

Figure 47 Patterns of steering angle data confidently identified by the MLC HMM model as Awake and Drowsy.

Figure 47, Figure 48, and Figure 49 show 60 s windows of data separated by model predictions for steering angle, accelerator pedal position, and brake pedal position respectively. In isolation each of these figures presents a portion of the model's interpretation of awake versus drowsy behavior. Figure 47 strongly suggests that the model interprets windows of large magnitude steering input as awake. Additionally there is some association between a lack of input and drowsy predictions. It is important to note that this lack of input does not preclude an awake prediction, as there are also several patterns of awake prediction data with very little input. In contrast to Figure 47, Figure 48 is much more difficult to decipher. It is interesting to note that all drowsiness instances have at least one non-zero accelerator input. This pattern seems to be consistent with the prevalence of drowsy-related lane departures in the rural straight and highway driving conditions. Figure 49 is also somewhat difficult to interpret yet like the steering angle data there is some correlation between large input values and alertness. Interestingly there

seems to be substantially more variance in input in the alert case. This might suggest that the model is partitioning the classes based on a reduction in reaction time effects.

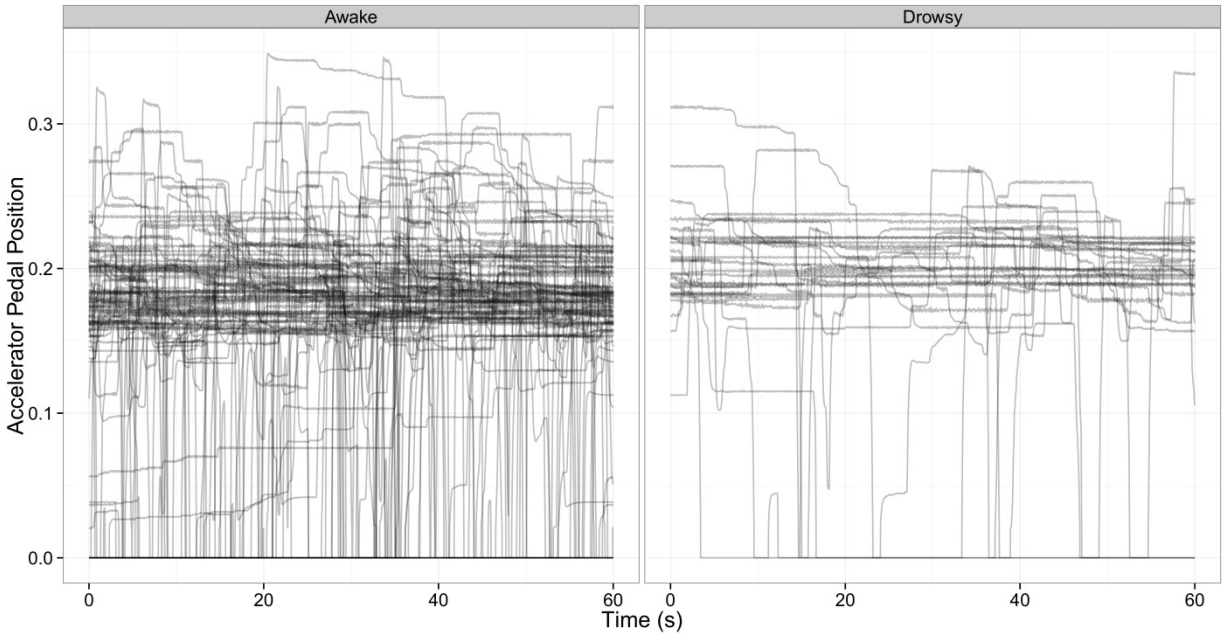

Figure 48 Patterns of accelerator pedal data confidently identified by the MLC HMM model as Awake and Drowsy.

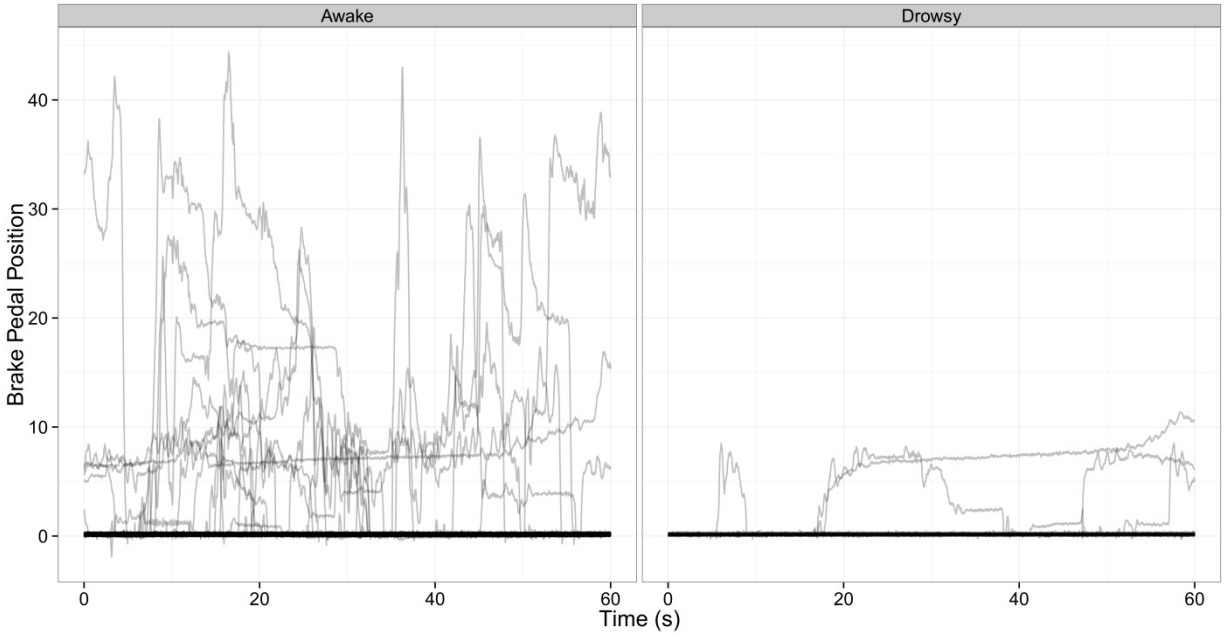

Figure 49 Patterns of brake pedal data confidently identified by the MLC HMM model as Awake and Drowsy.

Chapter 9 General conclusions, limitations, and future work

Drowsy driving causes a significant number of crashes, injuries, and fatalities. A promising direction for mitigating these crashes introduces technology into vehicles that detects and provides alerts to drivers prior to crashing. The detection algorithm is a critical part of this system because its performance influences the design space of the warning system and ultimately drives drivers' trust in the system. Despite a breadth of research on drowsiness detection algorithms, no clearly superior algorithm has emerged. Analyses of the drowsiness detection literature and theories of driver behavior suggest three gaps in current detection algorithms: 1) Most detection algorithms do not consider temporal dependencies, 2) No algorithm considers the effects of on-road context, and 3) No algorithm considers hierarchical context and the interplay between on-road context and maneuver-level context. This dissertation iteratively addresses these three gaps through a standardized process repeated for three studies and a series of supplemental analyses.

The standardized process used for algorithm development and evaluation was optimized for immediate implementation and effective driver feedback. This process included focusing on driving simulator data, consistent partitioning of training and testing data via a clustering algorithm, defining ground truth drowsiness as uncorrected drowsy-related lane departures, focusing on steering and pedal input measures, and a standard evaluation process focused on AUC, as well as point value false positive rates. At each level of this process the selected methods produced tradeoffs, and it is important to understand the impact of these tradeoffs and their effects on future work. The choice of simulator data in this case is beneficial due to its lack of noise but it is limited in the lack of variability in the drives as well as the sample size. This sample size is limiting because it does not ensure stability within the model estimates. The lack of stability negatively biases the estimation of AUC, thus it is reasonable to expect an increase in detection performance with a larger dataset. The definition of ground truth as uncorrected drowsy-related lane departures is beneficial because the measures are repeatable and simply communicated to drivers however the decision inhibits the algorithm's ability to capture the full spectrum

of drowsiness. The use of steering and pedal input facilitate immediate application of the algorithms developed here but also suggest portions of these algorithms may become obsolete with automated vehicles. Despite these limitations, the standardization of the algorithm evaluation process and decisions made are beneficial because they provide a guide for future work and allow the three studies to be consistently comparable.

The first of the three studies explored the effects of three temporal modeling structures: Hidden Markov Models, Hidden semi-Markov Models, and Conditional Random Fields on drowsiness detection performance, and compared these three structures to PERCLOS, an established benchmark, and a random forest steering algorithm that did not consider temporal dependencies. The study found that Hidden Markov Models outperformed all others on detecting drowsiness over a wide range of thresholds. Although the differences in AUC were not significant, an analysis of false positive rates demonstrated that the HMM could predict the same number of true positives with significantly fewer false positives than PERCLOS and the random forest algorithm. Furthermore the HsMM model performed the worst of all of the models and the CRF models did not provide a substantial benefit despite using many more parameters. These differences strongly suggest that temporal dependencies should be considered for future algorithm development, specifically in an HMM framework. In a more general sense these results suggest that although the transitions between levels of drowsiness is relatively regular, the duration of drowsiness states is highly irregular and random. Beyond these contributions to drowsiness detection and the understanding of the process of drowsiness, this study contributes to the predictive modeling literature in that it demonstrates the benefits of HMMs versus sliding window methods, and provides a benchmark for HMM effectiveness in time-series anomaly detection. Future work in this area should explore the relationships between window size and HsMM performance.

The second study explored the effects of time-series feature generation methods on-road context generation and model performance. Three feature generation methods: distributional parameters, Fourier transforms, and SAX time-series analysis. Each method was used to generate features describing on-road context generated from speed and acceleration data. These features were then integrated into an HMM

model. All of the feature generation methods improved drowsiness detection relative to the context-free HMM from the previous study however the difference in AUC was not significant. Extended analysis of the false positive rate from the SAX HMM model showed that the model significantly reduced false positives relative to a context-free HMM at the maximum performance threshold. Despite this increase in performance, the model did not significantly reduce false positives in road contexts that required very little driver input. These results provide support for the value of including road-context into models, but also illustrate its limitations. The model's limited benefit in certain contexts suggests a need for more advanced and thorough measures of driving context. This need extends to analyses of drowsy driving crash data that often focus on broad definitions of road context rather than more micro-level analyses of driving context such as maneuver frequency. Beyond these contributions the study also contributes to the predictive modeling literature by providing further support for the value of SAX for time-series feature generation. Future work in this area might explore Wavelet transforms as an alternative to Fourier transforms as the Wavelet transforms do not assume stationarity of the signal and may capture maneuver-level context in more detail.

The third study expanded the notion of driving context to include both on-road context, but also maneuver-level context. Both contexts were defined with SAX time-series features derived from speed and acceleration measures. The on-road context followed from the previous study and used the same parameter settings. The maneuver-level context was generated by locally normalizing speed and acceleration data over windows of 5 s, 10 s, or 30 s. The study explored a wide range of SAX input values and four model structures composed of a factorial combination of two factors. The first factor, hierarchy, reflected the inclusion or exclusion of on-road context. The second factor, context integration, explored the inclusion of maneuver-level context in either an independent random forest or a combined random forest with steering and pedal features. The modeling results surprisingly showed that the independent random forest model including only maneuver-level context performed the best, although the difference in AUC was not in general significant. However the model did predict all of the drowsiness instances correctly with a significantly lower false positive rate. These results further substantiate the importance of

driving context in drowsiness detection development, and illustrate a further need for more refined definitions of the hierarchy of driving context. They further suggest that while road-contexts, such as highway driving, are considered a significant factor in drowsiness induction, it might be more appropriate to define the relationship at the maneuver level, i.e. highway driving without lane changes. Beyond these contributions this study provides a contribution to the predictive modeling literature in the comparisons between the ensemble of independent random forests and the combined random forest. The results here suggest that the inductive bias of the random forest may not be conducive to partition indicators from several different measurement sources.

Following the planned analyses, three additional analyses were conducted to address model limitations and more firmly establish conclusions. These analyses consisted of an analysis of overlapping windows, a comparison between HMM and HsMM models for detecting Retrospective Sleepiness Scores, and an analysis of a broader definition of ground truth drowsiness. The overlapping window analysis compared the original modeling results to results generated by allowing successive classifications to be based on 30 s of overlapping observation data. This change violates the assumption of HMM models that requires all observations to be conditionally independent given the hidden states. Despite this violation, the overlapping windows did not have a significant effect on the model performance. This result is important from an algorithm development perspective because it improves the timeliness of the algorithms without cost to prediction performance. The goal of the RSS modeling analysis was to verify the observation from the first study that HsMM models do not provide a benefit for drowsiness predictions because of the irregularity of the duration of drowsiness states. The analysis followed the same process as the first study yet substituted drivers' subjective ratings of sleepiness for drowsy-related lane departures as ground truth. The analysis did not show a benefit of either model and effectively demonstrated that RSS scores are a poor measure of ground truth, given the variability in their meaning across drivers. The final analysis focused on expanding the ground truth definition of drowsiness in to include both windows containing uncorrected drowsy-related lane departures and the preceding windows. This analysis showed that this change actually reduced the effectiveness of the model predictions. This

change highlights the need for further exploration into ground truth measures, with a particular focus on capturing the gradual transitions between alertness and drowsiness.

Considered together this body of results clarifies the effects of including time, context, and contextual hierarchies on drowsiness-detection algorithms. They demonstrate that context and temporal dependencies can be used to significantly improve drowsiness detection, even when the definition of drowsiness is highly specific. In addition the analysis process here represents a standardization that could be used in future work to facilitate comparisons between algorithms. Beyond these improvements the performance of the models here facilitates the generation of more advanced theory of drowsiness detection and provides benchmarks for further predictive modeling analyses with similar data.

References

Agrawal, R., Faloutsos, C., & Swami, A. (1993). Efficient similarity search in sequence databases. In D. B. Lomet (Ed.), *4th International Conference, FODO '93 Chicago, Illinois, USA, October 13–15, 1993 Proceedings* (Vol. 8958546, pp. 69–84). Chicago, IL: Springer Berlin Heidelberg. doi:10.1007/3-540-57301-1_5

Anderson, J., & Lebiere, C. (1998). *The atomic components of thought*. New York, NY: Psychology Press.

Armstrong, K., Obst, P., Banks, T., & Smith, S. (2010). Managing driver fatigue: education or motivation? *Road & Transportation Research: A Journal of Australian and New Zealand Research and Practice*, *19*(3), 14–20.

Balkin, T. J., Horrey, W. J., Graeber, R. C., Czeisler, C. A., & Dinges, D. F. (2011). The challenges and opportunities of technological approaches to fatigue management. *Accident Analysis & Prevention*, *43*(2), 565–72. doi:10.1016/j.aap.2009.12.006

Barbu, A. (2009). Training an active random field for real-time image denoising. *IEEE Transactions on Image Processing : A Publication of the IEEE Signal Processing Society*, *18*(11), 2451–62. doi:10.1109/TIP.2009.2028254

Boyle, L. N., Tippin, J., Paul, A., & Rizzo, M. (2008). Driver Performance in the Moments Surrounding a Microsleep. *Transportation Research. Part F, Traffic Psychology and Behaviour*, *11*(2), 126–136. doi:10.1016/j.trf.2007.08.001

Brand, M., Oliver, N., & Pentland, A. (1997). Coupled hidden Markov models for complex action recognition. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 994–999). IEEE Comput. Soc. doi:10.1109/CVPR.1997.609450

Brown, I. D. (1994). Driver Fatigue. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *36*(2), 298–314. doi:10.1177/001872089403600210

Brown, I. D. (1997). Prospects for technological countermeasures against driver fatigue. *Accident Analysis & Prevention*, *29*(4), 525–531. doi:10.1016/S0001-4575(97)00032-8

Brown, T., Lee, J., Schwarz, C., Fiorentino, D., & McDonald, A. D. (2011). *Final Report: Advanced Countermeasures for Multiple Impairments*. Washington D.C.

Chan, K., & Fu, A. W. (1999). Efficient time series matching by wavelets. *Proceedings of the 15th International Conference on Data Engineering*, 126–133. doi:10.1109/ICDE.1999.754915

Dagli, I., Brost, M., & Breuel, G. (2003). Action recognition and prediction for driver assistance systems using dynamic belief networks. In J. G. Carbonell, J. Siekmann, R. Kowalczyk, J. P. Muller, T. Huaglory, & R. Unland (Eds.), *Agent Technologies, Infrastructures,*

Tools, and Applications for E-Services (pp. 179–194). Berlin: Springer Berlin Heidelberg. doi:10.1007/3-540-36559-1_15

Das, D., Zhou, S., & Lee, J. D. (2012). Differentiating Alcohol-Induced Driving Behavior Using Steering Wheel Signals. *IEEE Transactions on Intelligent Transportation Systems*, *13*(3), 1355–1368. doi:10.1109/TITS.2012.2188891

Dawson, D., Noy, Y. I., Härmä, M., Akerstedt, T., & Belenky, G. (2011). Modelling fatigue and the use of fatigue models in work settings. *Accident Analysis & Prevention*, *43*(2), 549–64. doi:10.1016/j.aap.2009.12.030

Dietterich, T. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, *10*(7), 1895–1923. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9744903

Dietterich, T. (2002). Machine learning for sequential data: A review. *Structural, Syntactic, and Statistical Pattern Recognition*, *2396*, 15–30. doi:10.1007/3-540-70659-3_2

Dinges, D. F. (1995). An overview of sleepiness and accidents. *Journal of Sleep Research*, *4*(S2), 4–14. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17017552

Dinges, D. F., & Grace, R. (1998). *PERCLOS: A valid psychophysiological measure of alertness as assessed by psychomotor vigilance*. *Federal Highway Administration*. Washington D.C. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:PERCLOS+:+A+Valid+Psychophysiological+Measure+of+Alertness+As+Assessed+by+Psychomotor+Vigilance#0

Dinges, D. F., Maislin, G., Brewster, R., Krueger, G., & Carroll, R. (2005). Pilot Test of Fatigue Management Technologies. *Transportation Research Record*, *1922*(1), 175–182. doi:10.3141/1922-22

Dinges, D. F., & Mallis, M. M. (1998). Managing fatigue by drowsiness detection: can technological promises be realized? In L. R. Hartley (Ed.), *Managing Fatigue in Transportation: Proceedings of the Third International Conference on Fatigue and Transportation*. Oxford: Elsevier.

Dinges, D. F., Mallis, M. M., Maislin, G., & Powell IV, J. W. (1998). *Evaluation of techniques for ocular measurement as an index of fatigue and the basis for alertness management*. Washington DC: NHTSA.

Doshi, A., Morris, B. T., & Trivedi, M. M. (2011). On-road prediction of driver's intent with multimodal sensory cues. *IEEE Pervasive Computing*, *10*(3), 22–34. doi:10.1109/MPRV.2011.38

Eskandarian, A., & Mortazavi, A. (2007). Evaluation of a Smart Algorithm for Commercial Vehicle Driver Drowsiness Detection. In *2007 IEEE Intelligent Vehicles Symposium* (pp. 553–559). Istanbul, Turkey: IEEE. doi:10.1109/IVS.2007.4290173

Fawcett. (2004). ROC Graphs: Notes and Practical Consideration for Researchers. *ReCALL*, *31*(HPL-2003-4), 1–38.

Fletcher, a, McCulloch, K., Baulk, S. D., & Dawson, D. (2005). Countermeasures to driver fatigue: a review of public awareness campaigns and legal approaches. *Australian and New Zealand Journal of Public Health*, *29*(5), 471–6. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16255451

Forsman, P. M., Vila, B. J., Short, R. A., Mott, C. G., & Van Dongen, H. P. A. (2013). Efficient driver drowsiness detection at moderate levels of drowsiness. *Accident Analysis & Prevention*, *50*(null), 341–50. doi:10.1016/j.aap.2012.05.005

Fuller, R. (2005). Towards a general theory of driver behaviour. *Accident Analysis & Prevention*, *37*(3), 461–72. doi:10.1016/j.aap.2004.11.003

Furman, G. D., Baharav, A., Cahan, C., & Akselrod, S. (2008). Early detection of falling asleep at the wheel: A Heart Rate Variability approach. *Computers in Cardiology*. Bologna, Italy.

Gander, P., Hartley, L., Powell, D., Cabon, P., Hitchcock, E., Mills, A., & Popkin, S. (2011). Fatigue risk management: Organizational factors at the regulatory and industry/company level. *Accident Analysis & Prevention*, *43*(2), 573–90. doi:10.1016/j.aap.2009.11.007

Geist, G. F., Merkt, R., & Altamuro, S. N.J.S.2C:11-5 (2002). State of New Jersey.

Ghazizadeh, M., Peng, Y., Lee, J. D., & Boyle, L. N. (2012). Augmenting the Technology Acceptance Model with Trust: Commercial Drivers' Attitudes towards Monitoring and Feedback. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *56*(1), 2286–2290. doi:10.1177/1071181312561481

Gibson, J. J. (1966). *The senses considered as perceptual systems*. Boston: Houghton Mifflin.

Grace, R., Byrne, V. E., Bierman, D. M., Legrand, J.-M., Gricourt, D., Davis, B. K., … Carnahan, B. (1996). A drowsy driver detection system for heavy vehicles. In *17th DASC. AIAA/IEEE/SAE. Digital Avionics Systems Conference. Proceedings (Cat. No.98CH36267)* (Vol. 2, pp. I36/1–I36/8). IEEE. doi:10.1109/DASC.1998.739878

Grace, R., & Stewart, S. (2001). The Co-Pilot: A Low-Cost Drowsy Driver and Driver Inattention Monitor. In *The International Driving Symposium on Human Factors in Driver Assessment, Training And Vehicle Design*. Aspen, CO.

Grassberger, P., & Procaccia, I. (1983). Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena*, *9*(1-2), 189–208. doi:10.1016/0167-2789(83)90298-1

Halpert, J. (2012). Drowsiness detection systems. *Jean Knows Cars*. Retrieved from http://www.jeanknowscars.com/cool-tech/car-tech-news/drowsiness-detection-systems/

Hanowski, R. J., Bowman, D., Alden, A., Wierwille, W. W., & Carroll, R. (2008). PERCLOS+: Development of a Robust Field Measure of Driver Drowsiness. *15th World Congress on Intelligent Transport Systems and ITS America's 2008 Annual Meeting*. New York.

Hartley, L., Horberry, T., & Mabbott, N. (2000). *Review of Fatigue Detection and Prediction Technologies*. Melbourne.

Hoddes, E., Zarcone, V., Smythe, H., Phillips, R., & Dement, W. C. (1973). Quantification of Sleepiness: A New Approach. *Psychophysiology*, *10*(4), 431–436. doi:10.1111/j.1469-8986.1973.tb00801.x

Hu, S., & Zheng, G. (2009). Driver drowsiness detection with eyelid related parameters by Support Vector Machine. *Expert Systems with Applications*, *36*(4), 7651–7658. doi:10.1016/j.eswa.2008.09.030

Jabon, M. E., Bailenson, J. N., Pontikakis, E., Takayama, L., & Nass, C. (2011). Facial expression analysis for predicting unsafe driving behavior. *IEEE Pervasive Computing*, *10*(4), 84–95. doi:10.1109/MPRV.2010.46

Jap, B. T., Lal, S., Fischer, P., & Bekiaris, E. (2009). Using EEG spectral components to assess algorithms for detecting fatigue. *Expert Systems with Applications*, *36*(2), 2352–2359. doi:10.1016/j.eswa.2007.12.043

Ji, Q., Lan, P., & Looney, C. (2006). A probabilistic framework for modeling and real-time monitoring human fatigue. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *36*(5), 862–875. doi:10.1109/TSMCA.2005.855922

Ji, Q., Zhu, Z., & Lan, P. (2004). Real-Time Nonintrusive Monitoring and Prediction of Driver Fatigue. *IEEE Transactions on Vehicular Technology*, *53*(4), 1052–1068. doi:10.1109/TVT.2004.830974

Kaufman, L., & Rousseeuw, P. (1990). Clustering Large Applications (Program CLARA). In *Finding Groups in Data: An Introduction to Cluster Analysis* (pp. 126–163). doi:10.1002/9780470316801.ch3

Keogh, E., Chakrabarti, K., Pazzani, M., & Mehrotra, S. (2001). Locally adaptive dimensionality reduction for indexing large time series databases. *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data - SIGMOD '01*, 151–162. doi:10.1145/375663.375680

Keogh, E., & Kasetty, S. (2003). On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Mining and Knowledge Discovery*, *7*, 349–371. Retrieved from http://link.springer.com/article/10.1023/A:1024988512476

Khushaba, R. N., Kodagoda, S., Lal, S., & Dissanayake, G. (2011). Driver drowsiness classification using fuzzy wavelet-packet-based feature-extraction algorithm. *IEEE Transactions on Bio-Medical Engineering*, *58*(1), 121–31. doi:10.1109/TBME.2010.2077291

Klauer, S., Dingus, T., & Neale, V. (2006). *The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data*. Washington D.C. Retrieved from http://trid.trb.org/view.aspx?id=786825

Knipling, R. R., & Wang, J. (1994). *Crashes and Fatalities Related to Driver Drowsiness/Fatigue*. Washington D.C.

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2007). Supervised Machine Learning : A Review of Classification Techniques. *Informatica*, *31*(3), 501–520.

Krajewski, J., Golz, M., & Sommer, D. (2009). Detecting sleepy drivers by pattern recognition based analysis of steering wheel behaviour. *Der Mensch Im Mittelpunkt Technischer Systeme*, 285–288. Retrieved from https://www.tu-berlin.de/fileadmin/f25/dokumente/8BWMMS/15.5-Krajewski.pdf

Krajewski, J., & Sommer, D. (2009). Steering wheel behavior based estimation of fatigue. In *Proceedings of the 5th International Driving Symposium on Human Factors in Driver Assessment and Design* (pp. 118–124). Retrieved from http://www.researchgate.net/publication/228688504_Steering_Wheel_Behavior_Based_Estimation_of_Fatigue/file/79e4150a91cc037eae.pdf

Kuge, N., Yamamura, T., Shimoyama, O., & Liu, A. (2000). A driver behavior recognition method based on a driver model framework. *SAE Transactions*, *109*(6), 469–476. Retrieved from http://dandelion-patch.mit.edu/people/amliu/Papers/SAE2000_Kuge.pdf

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York, New York, USA: Springer.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., & Engelhardt, A. (2011). caret: Classification and Regression Training. Retrieved from http://cran.r-project.org/package=caret

Kutila, M. H., Jokela, M., Mäkinen, T., Viitanen, J., Markkula, G., & Victor, T. W. (2007). Driver cognitive distraction detection: Feature estimation and implementation. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, *221*(9), 1027–1040. doi:10.1243/09544070JAUTO332

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields : probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML-2001)* (pp. 282–289). Morgan Kaufmann. Retrieved from http://ci.nii.ac.jp/naid/10014960504/en/

Lal, S. K. L., & Craig, A. (2001). A critical review of the psychophysiology of driver fatigue. *Biological Psychology*, *55*(3), 173–94. doi:10.1016/S0301-0511(00)00085-5

Lal, S. K. L., Craig, A., Boord, P., Kirkup, L., & Nguyen, H. (2003). Development of an algorithm for an EEG-based driver fatigue countermeasure. *Journal of Safety Research*, *34*(3), 321–328. doi:10.1016/S0022-4375(03)00027-6

Lee, J. D., Fiorentino, D., Reyes, M. L., Brown, T., Ahmad, O., Fell, J., … Dufour, R. (2010). *Assessing the Feasibility of Vehicle-Based Sensors to Detect Alcohol Impairment*. Washington DC: National Highway Traffic Safety Administration.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80.

Leslie, C., Eskin, E., & Noble, W. S. (2002). The spectrum kernel: A string kernel for SVM protein classification. In *Pacific Symposium on Biocomputing* (Vol. 7, pp. 566–575). Lihue, Hawaii.

Liang, Y., Reyes, M. L., & Lee, J. D. (2007). Real-Time Detection of Driver Cognitive Distraction Using Support Vector Machines. *IEEE Transactions on Intelligent Transportation Systems*, *8*(2), 340–350. doi:10.1109/TITS.2007.895298

Lin, C., Wu, R., Liang, S., Chao, W., Chen, Y., & Jung, T. (2005). EEG-based drowsiness estimation for safety driving using independent component analysis. *IEEE Transactions on Circuits and Systems I: Regular Papers*, *52*(12), 2726–2738. doi:10.1109/TCSI.2005.857555

Lin, C.-T., Chen, Y.-C., Huang, T.-Y., Chiu, T.-T., Ko, L.-W., Liang, S.-F., … Duann, J.-R. (2008). Development of wireless brain computer interface with embedded multitask scheduling and its application on real-time driver's drowsiness detection and warning. *IEEE Transactions on Bio-Medical Engineering*, *55*(5), 1582–91. doi:10.1109/TBME.2008.918566

Lin, J., Keogh, E., Wei, L., & Lonardi, S. (2007). Experiencing SAX: a novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, *15*(2), 107–144. doi:10.1007/s10618-007-0064-z

MacLean, A. W., Davies, D. R. ., & Thiele, K. (2003). The hazards and prevention of driving while sleepy. *Sleep Medicine Reviews*, *7*(6), 507–521. doi:10.1016/S1087-0792(03)90004-9

Maycock. (1997). Sleepiness and driving: The experience of U.K. car drivers. *Accident Analysis & Prevention*, *29*(4), 453–462.

McCall, J. C., & Trivedi, M. M. (2007). Driver Behavior and Situation Aware Brake Assistance for Intelligent Vehicles. *Proceedings of the IEEE*, *95*(2), 374–387. doi:10.1109/JPROC.2006.888388

McCartt, A. T., Rohrbaugh, J. W., Hammer, M. C., & Fuller, S. Z. (2000). Factors associated with falling asleep at the wheel among long-distance truck drivers. *Accident Analysis & Prevention*, *32*(4), 493–504. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10868752

McDonald, A. D., Lee, J. D., Aksan, N. S., Dawson, J. D., Tippin, J., & Rizzo, M. (2013a). Highway Healthcare: How Naturalistic Driving Data Index Adherence to CPAP Therapy

in Obstructive Sleep Apnea. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*(1), 1859–1863. doi:10.1177/1541931213571415

McDonald, A. D., Lee, J. D., Aksan, N. S., Dawson, J. D., Tippin, J., & Rizzo, M. (2013b). The Language of Driving. *Transportation Research Record: Journal of the Transportation Research Board*, *2392*(-1), 22–30. doi:10.3141/2392-03

McDonald, A. D., Lee, J. D., Schwarz, C., & Brown, T. L. (2013). Steering in a Random Forest: Ensemble Learning for Detecting Drowsiness-Related Lane Departures. *Human Factors: The Journal of the Human Factors and Ergonomics Society*. doi:10.1177/0018720813515272

McLaurin, E., McDonald, A. D., Lee, J. D., Aksan, N. S., Dawson, J. D., Tippin, J., & Rizzo, M. (2014). Variations on a theme : Topic modeling of naturalistic driving data. In *Proceedings of the 2014 International Annual Meeting of the Human Factors and Ergonomics Society*. Chicago, IL.

McRuer, D., Allen, R., Weir, D., & Klein, R. (1977). New results in driver steering control models. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *19*(4), 381–397. doi:10.1177/001872087701900406

McRuer, D., & Weir, D. H. (1969). Theory of manual vehicular control. *Ergonomics*, *12*(4), 599–633. doi:10.1080/00140136908931082

Michon, J. A. (1986). A critical view of driver behavior models: what do we know, what should we do? In *Human Behavior and Traffic Safety* (pp. 485–524). Springer US. doi:10.1007/978-1-4613-2173-6_19

Michon, J. A. (1989). Explanatory pitfalls and rule-based driver models. *Accident Analysis & Prevention*, *21*(4), 341–53. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18248389

Mitchell, T. M. (1997). *Machine Learning* (1st ed.). McGraw-Hill Science/Engineering/Math.

Mitler, M. M., Miller, J. C., Lipsitz, J. J., Walsh, J. K., & Wylie, C. D. (1997). The sleep of long-haul truck drivers. *The New England Journal of Medicine*, *337*(11), 755–61. doi:10.1056/NEJM199709113371106

Murphy, K. P. (2002). Dynamic bayesian networks. In M. Jordan (Ed.), *Probabilistic Graphical Models*. Retrieved from http://webdocs.cs.ualberta.ca/~greiner/C-366/366-SLIDES/dbn-murphy.pdf

Murphy, K. P. (2006). General Conditional Random Field (CRF) Toolbox for Matlab. Retrieved from http://www.cs.ubc.ca/~murphyk/Software/CRF/crfGeneralOld.html

NADS. (2010). National Advanced Driving Simulator. Retrieved from http://www.nads-sc.uiowa.edu/sim_nads1.php

National Sleep Foundation. (2009). 2009 Sleep In AmericaTM Poll: Summary Of Findings. (N. S. Foundataion, Ed.). Washington DC.

Neale, V. L., Dingus, T. A., Klauer, S. G., Sudweeks, J., & Goodman, M. (2005). *An Overview of the 100-Car Naturalistic Study and Findings*. *19th International Technical Conference on the Enhanced Safety of Vehicles*. Washington, D.C.

Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems 14* (2nd ed.). Cambridge, MA: The MIT Press. Retrieved from http://ai.stanford.edu/~ang/papers/nips01-discriminativegenerative.pdf

NHTSA. (2011). *Traffic Safety Facts: Drowsy Driving*. Washington D.C. Retrieved from http://www-nrd.nhtsa.dot.gov/pubs/811449.pdf

NHTSA. (2012). Fatality Analysis Reporting System (FARS) Database. Retrieved from ftp://ftp.nhtsa.dot.gov/fars/

Noy, Y. I., Horrey, W. J., Popkin, S. M., Folkard, S., Howarth, H. D., & Courtney, T. K. (2011). Future directions in fatigue and safety research. *Accident Analysis & Prevention*, *43*(2), 495–7. doi:10.1016/j.aap.2009.12.017

O'Connell, J., & Højsgaard, S. (2011). Hidden semi Markov models for multiple observation sequences: the mhsmm package for R. *Journal of Statistical Software*, *39*(104). Retrieved from http://core.kmi.open.ac.uk/download/pdf/5652927.pdf

Oliver, N., & Pentland, A. (2000). Graphical models for driver behavior recognition in a SmartCar. In *Proceedings of the IEEE Intelligent Vehicles Symposium 2000 (Cat. No.00TH8511)* (pp. 7–12). IEEE. doi:10.1109/IVS.2000.898310

Pack, A. I., Pack, A. M., Rodgman, E., Cucchiara, A., Dinges, D. F., & Schwab, C. W. (1995). Characteristics of crashes attributed to the driver having fallen asleep. *Accident Analysis & Prevention*, *27*(6), 769–775. doi:10.1016/0001-4575(95)00034-8

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics. Part A, Systems and Humans : A Publication of the IEEE Systems, Man, and Cybernetics Society*, *30*(3), 286–97. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11760769

Patel, M., Lal, S. K. L., Kavanagh, D., & Rossiter, P. (2011). Applying neural network analysis on heart rate variability data to assess driver fatigue. *Expert Systems with Applications*, *38*(6), 7235–7242. doi:10.1016/j.eswa.2010.12.028

Pérez-Chada, D., Videla, A. J., O'Flaherty, M. E., Palermo, P., Meoni, J., Sarchi, M. I., … Durán-Cantolla, J. (2005). Sleep habits and accident risk among truck drivers: a cross-

sectional study in Argentina. *Sleep*, *28*(9), 1103–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16268379

Philip, P. (2005). Sleepiness of occupational drivers. *Industrial Health*, *43*(1), 30–3. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15732301

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286. doi:10.1109/5.18626

Rabiner, L. R., & Juang, B. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine*, *3*(1), 4–16. doi:10.1002/0471250953.bia03as18

Raksincharoensak, P., Khaisongkram, W., Nagai, M., Shimosaka, M., Mori, T., & Sato, T. (2010). Integrated driver modelling considering state transition feature for individual adaptation of driver assistance systems. *Vehicle System Dynamics*, *48*(sup1), 55–71. doi:10.1080/00423111003668229

Reason, J. (1990). *Human Error*. Cambridge, UK: Cambridge University Press.

Regan, M. A., Lee, J. D., & Young, K. L. (2009). Driver distraction injury prevention countermeasures-Part 2: Education and training. In *Driver Distraction: Theory Effects and Mitigation* (pp. 559–578). Boca Raton, FL: CRC Press.

Richman, J. S., & Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology. Heart and Circulatory Physiology*, *278*(6), H2039–49. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10843903

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(Swets 1973), 77. doi:10.1186/1471-2105-12-77

Sagberg, F. (1999). Road accidents caused by drivers falling asleep. *Accident Analysis & Prevention*, *31*(6), 639–49. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10487339

Sahayadhas, A., Sundaraj, K., & Murugappan, M. (2012). Detecting driver drowsiness based on sensors: a review. *Sensors (Basel, Switzerland)*, *12*(12), 16937–53. doi:10.3390/s121216937

Salamin, H., & Vinciarelli, A. (2011). Introduction to Sequence Analysis for Human Behavior Understanding. In A. A. Salah & T. Gevers (Eds.), *Computer Analysis of Human Behavior* (pp. 21–40). London: Springer London. doi:10.1007/978-0-85729-994-9

Salvucci, D. (2006). Modeling Driver Behavior in a Cognitive Architecture. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *48*(2), 362–380. doi:10.1518/001872006777724417

Salvucci, D., Boer, E., & Liu, A. (2001). Toward an Integrated Model of Driver Behavior in Cognitive Architecture. *Transportation Research Record*, *1779*(1), 9–16. doi:10.3141/1779-02

Sandberg, D., Akerstedt, T., Anund, A., Kecklund, G., & Wahde, M. (2011). Detecting Driver Sleepiness Using Optimized Nonlinear Combinations of Sleepiness Indicators. *IEEE Transactions on Intelligent Transportation Systems*, *12*(1), 97–108. doi:10.1109/TITS.2010.2077281

Sathyanarayana, A., Boyraz, P., & Hansen, J. H. L. (2008). Driver behavior analysis and route recognition by Hidden Markov Models. In *2008 IEEE International Conference on Vehicular Electronics and Safety* (pp. 276–281). IEEE. doi:10.1109/ICVES.2008.4640874

Sayed, R., & Eskandarian, A. (2001). Unobtrusive drowsiness detection by neural network learning of driver steering. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, *215*(9), 969–975. doi:10.1243/0954407011528536

Summala, H. (1988). Risk control is not risk adjustment: the zero-risk theory of driver behaviour and its implications. *Ergonomics*, *31*(4), 491–506. doi:10.1080/00140138808966694

Sutton, C., & McCallum, A. (2006). An introduction to conditional random fields for relational learning. In L. Getoor & B. Taskar (Eds.), *Introduction to statistical relational learning*. Cambridge, MA: MIT Press.

Tijerina, L., Gleckler, M., & Stoltzfus, D. (1999). *A preliminary assessment of algorithms for drowsy and inattentive driver detection on the road* (Vol. 808). Washington D.C. Retrieved from http://trid.trb.org/view.aspx?id=502406

Torkkola, K., Venkatesan, S., & Liu, H. (2005). Sensor Sequence Modeling for Driving. In *FLAIRS Conference* (pp. 721–727). Clearwater Beach, FL. Retrieved from http://www.aaai.org/Papers/FLAIRS/2005/Flairs05-118.pdf

Turaga, P., Chellappa, R., Subrahmanian, V. S., & Udrea, O. (2008). Machine Recognition of Human Activities: A Survey. *IEEE Transactions on Circuits and Systems for Video Technology*, *18*(11), 1473–1488. doi:10.1109/TCSVT.2008.2005594

Vail, D. L., Veloso, M. M., & Lafferty, J. D. (2007). Conditional random fields for activity recognition. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems - AAMAS '07* (p. 1). New York, New York, USA: ACM Press. doi:10.1145/1329125.1329409

Van Kasteren, T., Englebienne, G., & Krose, B. (2010). Activity recognition using semi-markov models on real world smart home datasets. *Journal of Ambient Intelligence and Smart Environments*, *2*(3), 311–325. doi:10.3233/AIS-2010-0070

Vuckovic, A., Radivojevic, V., Chen, A. C. N., & Popovic, D. (2002). Automatic recognition of alertness and drowsiness from EEG by an artificial neural network. *Medical Engineering & Physics*, *24*(5), 349–360. doi:10.1016/S1350-4533(02)00030-9

Wang, Q., Yang, J., Ren, M., & Zheng, Y. (2006). Driver Fatigue Detection: A Survey. In *2006 6th World Congress on Intelligent Control and Automation* (pp. 8587–8591). Dalian, China: IEEE. doi:10.1109/WCICA.2006.1713656

Weir, D., & McRuer, D. (1970). Dynamics of driver vehicle steering control. *Automatica*, *6*(1), 87–98. Retrieved from http://www.sciencedirect.com/science/article/pii/0005109870900774

Wierwille, W. W., Wreggit, S. S., Kirn, C. L., Ellsworth, L. A., & Fairbanks, R. J. (1994). *Research on vehicle-based driver status/performance monitoring: Development, validation, and refinement of algorithms for detecting driver drowsiness*. Washington D.C. Retrieved from http://trid.trb.org/view.aspx?id=448128

Wilde, G. J. S. (1982). The Theory of Risk Homeostasis: Implications for Safety and Health. *Risk Analysis*, *2*(4), 209–225. doi:10.1111/j.1539-6924.1982.tb01384.x

Wilde, G. J. S. (1988). Risk homeostasis theory and traffic accidents: propositions, deductions and discussion of dissension in recent reactions. *Ergonomics*, *31*(4), 441–468. doi:10.1080/00140138808966691

Wilde, G. J. S. (1994). *Target risk*. PDE Publications. Retrieved from http://www.engr.mun.ca/~cdaley/6003/Target Risk_Wilde.pdf

Wilkinson, R. T., & Houghton, D. (1982). Field Test of Arousal: A Portable Reaction Timer with Data Storage. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *24*(4), 487–493. doi:10.1177/001872088202400409

Williamson, A., Lombardi, D. a, Folkard, S., Stutts, J., Courtney, T. K., & Connor, J. L. (2011). The link between fatigue and safety. *Accident Analysis & Prevention*, *43*(2), 498–515. doi:10.1016/j.aap.2009.11.011

Wilson, B. J., & Bracewell, T. D. (2000). Alertness monitor using neural networks for EEG analysis. In *Neural Networks for Signal Processing X. Proceedings of the 2000 IEEE Signal Processing Society Workshop (Cat. No.00TH8501)* (Vol. 2, pp. 814–820). Sydney, NSW: IEEE. doi:10.1109/NNSP.2000.890161

Wolf, A., Swift, J. B., Swinney, H. L., & Vastano, J. A. (1985). Determining Lyapunov exponents from a time series. *Physica D: Nonlinear Phenomena*, *16*(3), 285–317. doi:10.1016/0167-2789(85)90011-9

Yang, G., Lin, Y., & Bhattacharya, P. (2010). A driver fatigue recognition model based on information fusion and dynamic Bayesian network. *Information Sciences*, *180*(10), 1942–1954.

Yang, J. H., Tijerina, L., Pilutti, T., Coughlin, J. F., & Feron, E. (2009). Detection of Driver Fatigue Caused by Sleep Deprivation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *39*(4), 694–705. doi:10.1109/TSMCA.2009.2018634

Yeo, M. V. M., Li, X., Shen, K., & Wilder-Smith, E. P. V. (2009). Can SVM be used for automatic EEG detection of drowsiness during car driving? *Safety Science*, *47*(1), 115–124. doi:10.1016/j.ssci.2008.01.007

Zhao, C., Zhao, M., Liu, J., & Zheng, C. (2012). Electroencephalogram and electrocardiograph assessment of mental fatigue in a driving simulator. *Accident Analysis & Prevention*, *45*, 83–90. doi:10.1016/j.aap.2011.11.019

Zhao, C., Zheng, C., Zhao, M., Tu, Y., & Liu, J. (2011). Multivariate autoregressive models and kernel learning algorithms for classifying driving mental fatigue based on electroencephalographic. *Expert Systems with Applications*, *38*(3), 1859–1865. doi:10.1016/j.eswa.2010.07.115

Zilberg, E., Xu, Z. M., Burton, D., Karrar, M., & Lal, S. K. L. (2007). Methodology and initial analysis results for development of non-invasive and hybrid driver drowsiness detection systems. In *The 2nd International Conference on Wireless Broadband and Ultra Wideband Communications (AusWireless 2007)* (pp. 16–16). Sydney, NSW: IEEE. doi:10.1109/AUSWIRELESS.2007.44