Investigating Violation Behavior at Intersections using Intelligent Transportation Systems: A Feasibility Analysis on Vehicle/Bicycle-to-Infrastructure Communications as a Potential Countermeasure

By

Arash Jahangiri

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In

Civil Engineering

Thomas Dingus

Hesham Rakha

Zachary Doerzaph

Ihab El-Shawarby

July 21, 2015, Blacksburg, VA

Keywords: *Intersection safety, driver/cyclist violation prediction, transportation mode recognition, machine learning*

Investigating Violation Behavior at Intersections using Intelligent Transportation Systems: A Feasibility Analysis on Vehicle/Bicycle-to-Infrastructure Communications as a Potential Countermeasure

Arash Jahangiri

Abstract

The focus of this dissertation is on safety improvement at intersections and presenting how Vehicle/Bicycle-to-Infrastructure Communications can be a potential countermeasure for crashes resulting from drivers' and cyclists' violations at intersections. The characteristics (e.g., acceleration capabilities, etc.) of transportation modes affect the violation behavior. Therefore, the first building block is to identify the users' transportation mode. Consequently, having the mode information, the second building block is to predict whether or not the user is going to violate. This step focuses on two different modes (i.e., driver violation prediction and cyclist violation prediction). Warnings can then be issued for users in potential danger to react or for the infrastructure and vehicles so they can take appropriate actions to avoid or mitigate crashes.

A smartphone application was developed to collect sensor data used to conduct the transportation mode recognition task. Driver violation prediction task at signalized intersections was conducted using observational and simulator data. Also, a naturalistic cycling experiment was designed for cyclist violation prediction task. Subsequently, cyclist violation behavior was investigated at both signalized and stop-controlled intersections. To build the prediction models in all the aforementioned tasks, various Artificial Intelligence techniques were adopted. K-fold Cross-Validation as well as Out-of-Bag error was used for model selection and validation.

Transportation mode recognition models contributed to high classification accuracies (e.g., up to 98%). Thus, data obtained from the smartphone sensors were found to provide important information to distinguish between transportation modes. Driver violation (i.e., red light running) prediction models were resulted in high accuracies (i.e., up to 99.9%). Time to intersection (TTI), distance to intersection (DTI), the required deceleration parameter (RDP), and velocity at the onset of a yellow light were among the most important factors in violation prediction. Based on logistic regression analysis, movement type and presence of other users were found as significant factors affecting the probability of red light violations by cyclists at signalized intersections. Also, presence of other road users and age were the significant factors affecting violations at stop-controlled intersections. In case of stop-controlled intersections, violation prediction models resulted in error rates of 0 to 10 percent depending on how far from the intersection the prediction task is conducted.

Dedication

This work is dedicated to my wife for her constant love and support and to my parents for their endearment and encouragement throughout this long process.

Acknowledgment

I would like to express my sincere gratitude to my advisors, Dr. Hesham Rakha and Dr. Thomas Dingus, for their continuous encouragement, support, and understanding throughout the entire course of this dissertation. Their guidance helped me in all the time of research and writing of this dissertation. This dissertation would have not been possible without their constant care.

Besides my advisors, I would like to thank the rest of my PhD committee: Dr. Ihab El-Shawarby and Dr. Zachary Doerzaph for their insightful comments and encouragement.

Last but not the least, I would like to thank all who have helped me all the way through the past five years: my wife, my parents, my colleagues, and my friends.

Attribution

Chapter 3, 4, 5, and 6:

Hesham Rakha, PHD, Department of Civil and Environmental Engineering at Virginia Tech, is currently a professor at Virginia Tech. Dr. Rakha was a co-author on the papers in these chapters and helped through the entire process.

Chapter 4, 5, and 6:

Thomas Dingus, PHD, Department of Civil and Environmental Engineering at Virginia Tech, is currently a professor at Virginia Tech. Dr. Dingus was a co-author on the papers in these chapters and helped through the entire process.

Table of Contents

Abstract ii
Dedication iii
Acknowledgment iv
Attribution v
Table of Contents vi
List of Figures xii
List of Tables xiv
Chapter 1: Introduction 1
Problem Statement 4
Research plans 5
Dissertation layout 5
References 6
Chapter 2: Model Development 7
Identifying three main modules 8
Chapter 3: Transportation Mode Recognition 12
Developing a Support Vector Machine (SVM) Classifier for Transportation Mode Identification using Mobile Phone Sensor Data 13
Abstract 14
Introduction 15
Relevant work 15
Data collection and preprocessing 18
Model development 19
Attribute Selection 20
Results 21
Conclusions 25
Acknowledgements 25
References 25
Machine Learning Transportation Mode Recognition using Mobile Phone Sensor Data 27
Abstract 27
Introduction 27
Relevant work 28

Data collection, preprocessing and feature extraction 29
Model development 30
K-Nearest Neighbor (KNN) 31
Support Vector Machines (SVMs) 31
Tree based models 32
Feature Selection 32
K-fold Cross-Validation 33
Results 33
KNN Model 33
SVM Model 34
Tree-based models 34
Feature Importance 36
Model Comparison 36
Feature Combination 36
Conclusions 37
Acknowledgments 38
References 38
Transportation Mode Recognition using a Distributed Learning Approach 41
Abstract 42
Introduction 43
Transportation mode recognition 43
Distributed learning in Transportation 45
Data Collection 46
Model development 46
Methods 46
Feature selection 48
Results 49
Measures of comparison 49
Comparison 49
Conclusion 51
Acknowledgements 51
References 52
Chapter 4: Driver Violation Prediction 54

Predicting Red-light Running Violations at Signalized Intersections using Machine Learning Techniques 55
Abstract 56
Introduction 57
Data Collection 59
Model development 59
Methods 59
Time window corresponding to the yellow onset 60
Factor selection 62
Results 63
Conclusion 64
Acknowledgements 65
References 65
Adopting Machine Learning Methods to Predict Red-light Running Violations 68
Abstract 68
Introduction 68
Background 69
Data Description 69
Model Development 69
Methods 69
Monitoring period corresponding to the yellow onset 70
Feature selection 71
Results 71
Conclusion 72
Acknowledgment 73
References 73
Red-light Running Violation Prediction using Observational and Simulator Data 76
Abstract 77
Introduction 78
Dilemma Zone and Influential Factors 78
Data Collection Methods 80
Study Focus and Objectives 81
Relevant Work 83

Data Description 85
Observational Data 85
Driving Simulator Data 86
Model Development 87
RF Method 87
Monitoring Period 88
Factor Creation 91
Factor Selection 92
Results 93
Observational Data Results 93
Simulator Data Results 98
Model comparison: observational data vs. simulator data 101
Implementation Considerations 102
Conclusions 102
Acknowledgements 104
References 104
Chapter 5: Bicycle Naturalistic Data Collection 108
Developing a System Architecture for Cyclist Violation Prediction Models Incorporating Naturalistic Cycling Data 109
Abstract 109
Introduction 110
Background 110
Examining Countermeasures 110
Investigating contributing factors 110
Important Factors 113
Data Collection and Analysis 113
Cycling Naturalistic Data Collection System 113
Data Visualization Tool 113
Violation Prediction Models 114
Intersection bicycle-car crash prediction system 114
Conclusions and Future Work 115
Acknowledgements 115
References 116

Chapter 6: Bicycle Violation Prediction 118
Investigating Cyclist Violations at Intersections using Naturalistic Cycling data 119
Abstract 120
Introduction 121
Literature review 121
Examining countermeasures 121
Investigating contributing factors 122
Naturalistic cycling experiment 124
Pre-screening 124
Data collection 124
Data reduction 125
Model development 125
Multivariate logistic regression (MLR) 125
Random forest (RF) 126
Monitoring period 126
Variable selection 127
Intersection bicycle-car crash prediction system 127
Results 128
Signal-controlled intersections 128
Stop-controlled intersections 130
Conclusion 132
Acknowledgements 133
References 133
Chapter 7: Conclusions and Future Recommendations 135
Transportation Mode Recognition 136
Conclusions 137
Future recommendations 137
Driver Violation Prediction 138
Conclusions 138
Future recommendations 139
Cyclist Violation Prediction 140
Conclusions 140
Future recommendations 141

Appendix A: Transportation mode recognition - Error analysis 142
Appendix B: Transportation mode recognition - Analysis extension 150
Appendix C: Driver violation prediction - Sensitivity analysis 152
Appendix D: Model performance: Violation prediction 154

List of Figures

Figure 1 Relationship between different parts: building blocks, variables, and data sources 4
Figure 2 Road users approaching a sign-controlled or signalized intersection 8
Figure 3 Transportation mode recognition starts at a specific point 9
Figure 4 Violation prediction starts after transportation modes are detected 9
Figure 5 Safety messages can be sent to users and or infrastructure after violations are predicted 10
Figure 6: Impacts of the Regularization and the Gaussian Parameters on Model Accuracy for Scenarios 5 and Scenario 6 23
Figure 7 Impact of number of neighbors on model misclassification error 34
Figure 8 Impacts of regularization and Gaussian parameters on model Accuracy 34
Figure 9 Illustration of a single Decision Tree 35
Figure 10 Impact of number of trees on the misclassification error 35
Figure 11 Impact of number of features on the misclassification error 35
Figure 12 Feature importance for two different measures 36
Figure 13 Model comparison results 37
Figure 14: Accuracy comparison - SVM models 50
Figure 15: Accuracy Comparison - RF models 50
Figure 16: determination of the monitoring time window; vehicle v as the RLR violator and vehicle e as the endangered vehicle 61
Figure 17: Time to Intersection at the Onset of Yellow Light 62
Figure 18: Selecting the number of trees - RF model 63
Figure 19: Selecting the number of factors for each tree - RF model 64
Figure 20: Conducting the model selection - SVM model 64
Figure 21 determination of the monitoring period; vehicle v as the RLR violator and vehicle e as the endangered vehicle 71
Figure 22 Distance to Intersection at the Onset of Yellow Light 71
Figure 23 Selecting the number of trees - RF model 72
Figure 24 Selecting the number of features for each tree - RF model 72
Figure 25 Conducting the model selection - SVM model 72
Figure 26: Determination of the monitoring period; vehicle *v* as the RLR violator and vehicle *e* as the endangered vehicle 90
Figure 27: Observational Data - Selecting the required number of trees 94
Figure 28: Observational Data - Selecting the number of factors for each tree 94
Figure 29 Parameters to select monitoring periods 95

Figure 30 Observational Data - OOB error rate in percentage using all 17 observational data factors 96
Figure 31 Observational Data - Factor importance change - Acceleration at Onset of Yellow 96
Figure 32 OOB error for different monitoring periods and number of top factors 98
Figure 33 Simulator Data - Selecting the required number of trees 99
Figure 34 Simulator Data - Selecting the number of factors for each tree 99
Figure 35 Simulator Data - Factor Importance 100
Figure 36 OOB error for different number of top factors used 101
Figure 37 Naturalistic cycling data collection system 113
Figure 38 Hawkeye Software as Data Visualization tool for Data Reduction 114
Figure 39 Intersection bicycle-car crash prediction: System Architecture 115
Figure 40 Naturalistic cycling data collection system 125
Figure 41 Factors to define a monitoring period 127
Figure 42 Intersection bicycle-car crash prediction: System Architecture 128
Figure 43 RF models – overall prediction accuracies 132
Figure A.1 Car misclassified as Bus (type 1) 143
Figure A.2 Bus misclassified as Car (type 2) 144
Figure A.3 Bike misclassified as Walk (type 3) 145
Figure A.4 Walk misclassified as Bike (type 4) 146
Figure A.5 Bus misclassified as Bike (type 5) 147
Figure A.6 Run misclassified as Walk (type 6) 148
Figure A.7 Car misclassified as Bike (type 7) 149
Figure C.1 Sensitivity Analysis on the monitoring period; vehicle ***v*** as the RLR violator and vehicle ***e*** as the endangered vehicle 152
Figure C.2 Sensitivity Analysis on different monitoring periods 153

List of Tables

Table 1 Bicycle Crash Types at intersections 2
Table 2 Summary of Past Efforts 17
Table 3 Key Features of Studies using a Time Window Less than a Minute 18
Table 4 sets of attributes from different sensors 21
Table 5 overall accuracy and key points of different scenarios, constant values of regularization and Gaussian parameters considered 22
Table 6 Confusion matrices in percentage for scenarios 5 and 6 23
Table 7 Robustness of the Developed Model 24
Table 8 Comparison with the Most Relevant Study 25
Table 9 MEASURES USED TO CREATE FEATURES 30
Table 10 CONFUSION MATRIX - KNN MODEL 34
Table 11 CONFUSION MATRIX - SVM MODEL 34
Table 12 CONFUSION MATRIX – ENSEMBLE OF SVM MODELS 34
Table 13 CONFUSION MATRIX - DECISION TREE MODEL 35
Table 14 confusion matrix - Pruned Decision Tree model 35
Table 15 CONFUSION MATRIX - RANDOM FOREST 35
Table 16 CONFUSION MATRIX - BAGGING 36
Table 17 IMPORTANT FEATURES 36
Table 18: Measures used to create features 49
Table 19 List of the examined factors 62
Table 20 List of the examined features 71
Table 21 Influential factors affecting driver behaviour when approaching signalized intersections 79
Table 22 Factors used from the observational data (CICAS-V) 86
Table 23 Factors used from the simulator data 86
Table 24 Observational data - list of the examined factors 91
Table 25 simulator data - list of the examined factors 92
Table 26 Factor Importance Change 97
Table 27 Simulator Data - Factor Ranking 99
Table 28 Bicycle Crash Types at intersections 110
Table 29 Violation rates in different countries 121
Table 30 Data collection through video cameras at intersections - summary of past studies 122
Table 31 Naturalistic cycling data collection - summary of past studies 123

Table 32 variables obtained from data reduction 125
Table 33 Cyclist violation types part 1 129
Table 34 Cyclist violation types part 2 130
Table 35 Logistic regression model - signalized intersections 130
Table 36 Logistic regression model - sign-controlled intersections 131
Table 37 list of factors to develop the RF model 131
Table A.1 Confusion matrix – Random forest model 142
Table A.2 Misclassification errors 142
Table B.1 without using GPS - 95.1% overall accuracy 150
Table B.2 with using GPS - 96.3% overall accuracy 150
Table B.3 Bus and Car combined - without GPS - 97.02% overall accuracy 151
Table B.4 Bus and Car combined - with GPS - 98.5% overall accuracy 151
Table D.1 Confusion matrix 154
Table D.2 Confusion matrix - Driver violation prediction using observational data 154
Table D.3 Confusion matrix - Driver violation prediction using simulator data 154
Table D.4 Confusion matrix - Cyclist violation prediction at 4-way stop signs 155
Table D.5 Confusion matrix - Cyclist violation prediction at 2-way stop signs 155

Chapter 1: Introduction

Introduction

According to National Highway Traffic Safety Administration (NHTSA) report, during 2012, more than 2.5 million intersection-related crashes occurred in the United States, of which 2,850 were fatal crashes and 680,000 were injurious crashes [1]. Specifically, statistics demonstrate that a large number of crashes occur at signalized intersections due to traffic violations, of which running red lights has been reported to be a serious issue. According to the Insurance Institute for Highway Safety (IIHS), 683 people were killed and an estimated 133,000 were injured in crashes in the United States during 2012 due to running red lights [2]. The AAA Foundation for Traffic Safety surveyed 2,000 United States residents aged 16 and older. The survey showed that approximately 93% of drivers believe that running through a red light is unacceptable if it is possible to stop safely. However, one-third mentioned they ran through a red light during the past 30 days. This shows that, although drivers are generally aware of the dangers of this type of violation, they are likely to occasionally run a red light [3].

According to the FARS[1] database, an average of more than 30% of cyclists' fatalities has occurred at intersections during the past 5 years (2008-20012). Failure to obey traffic signs, signals, or officer was reported as the forth common factor (10.6%) leading to fatalities. However, no more details were provided in FARS regarding vehicle-bicycle crash types. The following two studies provided more details on bicycle-vehicle crash types: Crash data from 2005 to 2009 in North Carolina showed that 43.5 percent of the crashes that involved bicyclists occurred at intersections [4]. Similarly, from an older (early 1990's) but more comprehensive (Data from six US states) study, almost half of the bicycle-motor vehicle crashes took place at intersections [5]. This research was a Federal Highway Administration (FHWA) research study that was conducted by the University of North Carolina Highway Safety Research Center. The data set used in this study was a sample of crash data obtained from six US states. More specifically, the following crash types were recognized for the bicycle related crashes that occurred at intersections as shown in Table 1 [4, 6]. As demonstrated by statistics, bicycle safety at intersections has been a serious issue. Further, the growing number of bicycle commuters makes the problem even more important; from 2000 to 2011, bicycle commuting rates in the US increased: by 80 percent in large Bicycle Friendly Cities (BFCs), by 32 percent in non-BFCs, and by the national average of 47 percent [7].

Table 1 Bicycle Crash Types at intersections

Crash Type	NC state (2005-2009)	Six US states (early 1990's)
	Percent of all vehicle-bicycle crashes	
Motorist drive out : Sign-Controlled Intersection	9.7%	9.3%
Bicyclist ride out : Sign-Controlled Intersection	7.9%	9.7%
Bicyclist ride out : Signalized Intersection	4.7%	7.1%
Motorist drive out: Signalized Intersection	2.6%	2%

The focus of this dissertation is on safety improvement at intersections and presenting how Vehicle/Bicycle-to-Infrastructure Communications can be a potential countermeasure for crashes resulting from drivers' and cyclists' violations at intersections. The transportation mode

[1] FATALITY ANALYSIS REPORTING SYSTEM (FARS) ENCYCLOPEDIA

characteristics such as acceleration/deceleration capabilities, physical shape, etc. affect the violation behavior. Therefore, the first building block is to identify the users' transportation mode. In case an individual is using an instrumented mode (e.g., a vehicle equipped with devices capable of sending transportation mode information), the mode information can be easily obtained. However, it will take years that all vehicles will be instrumented with such equipment. Also, for some transportation modes such as bicycles and pedestrians it may not be possible to instrument all bikes or any pedestrians. Most individuals, however, have their cell phone with them all the time. Another advantage of smartphones over on-board units is that when using on-board units, the GPS requires a warm-up time that leads to not having valid GPS data for the start of the trips (usually more than 5-10 minutes), but smartphones do not require that (i.e., warm-up time is usually less than one minute). Thus, transportation mode recognition task using cell phones is considered as an important task. Consequently, having the mode information, the second building block is to predict whether or not the user is going to violate. In other words, violation prediction models need to be developed for each transportation mode. This step in the dissertation focuses on two different modes (i.e., driver violation prediction and cyclist violation prediction). Warnings can then be issued for users in potential danger to react or for the infrastructure and vehicles so they can take appropriate actions to avoid or mitigate crashes.

Figure 1 presents a flowchart that shows the two building blocks, the required variables, data sources, and how they are connected. The first row presents the two building blocks, namely transportation mode recognition and violation prediction. The violation prediction in this dissertation only focuses on two modes as indicated in red color (i.e. passenger cars & bicycles). The required variables as shown in the second row of this figure, how they are created, and how they are selected are discussed for each task in the corresponding chapter. The third row in this figure presents different data sources that can be adopted to obtain the required variables. The data sources that are written in red represent the sources used in this dissertation. For implementation testing in real world conditions, only the smartphones and onboard equipment are desirable because in case a potential crash is predicted, warnings can only be sent to the smartphones and on board equipment (i.e., warnings cannot be sent when data are collected through video cameras. Also, simulators are not applicable). Moreover, simulator data may not reflect the natural user behavior. However, simulators are needed for testing certain scenarios in which users might be in dangerous situations or when examining factors such as age, gender, using cell phones.

Smartphones, nowadays, are equipped with powerful sensors such as GPS, accelerometer, gyroscope, light sensors, temperature sensors, etc. Having such powerful sensors all embedded in a small device carried in everyday life activities has enabled researchers to investigate new research areas. The advantages of these smart devices include ubiquity, ability to send and receive data through various ways (e.g. Wi-Fi/cellular network/Bluetooth), providing alerts, and storing/processing data. Furthermore, smartphones will soon be equipped such that they will be capable of sending/receiving DSRC[2]. Therefore, to appreciate the value of smartphones, for the mode recognition task, data were obtained from smartphones. The detailed explanation of the data, how different factors were created and selected are presented in the corresponding chapter. Further, for the driver violation prediction task, observational data (i.e. through video cameras) and simulator data were adopted. For the cyclist violation behavior, a naturalistic cycling data collection method (i.e., through on board equipment) was used. Detailed explanations regarding data

[2] Dedicated Short Range Communications

collection, what factors were included, how different factors were created and selected are discussed in the corresponding chapters.

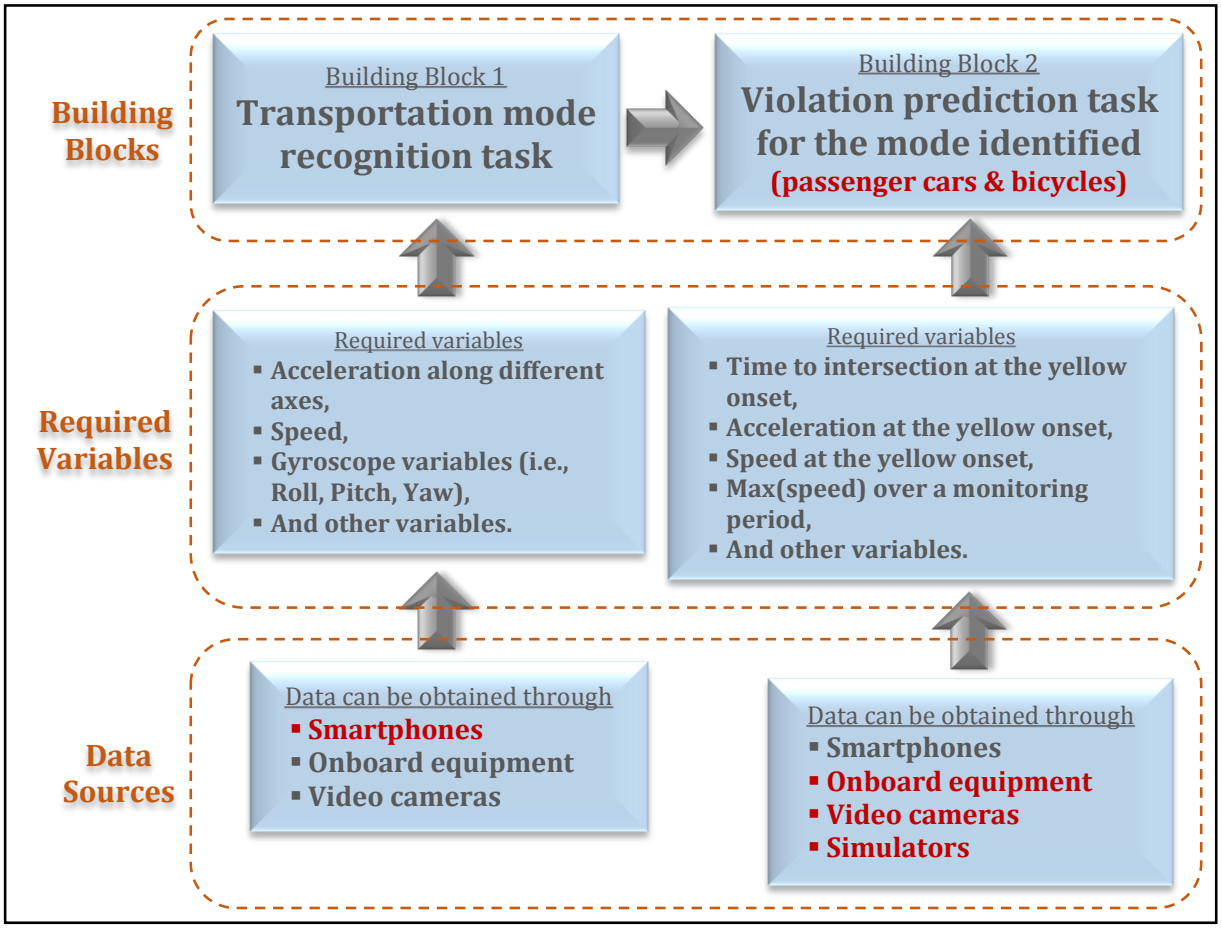


Figure 1 Relationship between different parts: building blocks, variables, and data sources

Problem Statement

For different reasons (e.g. distraction, judgment, etc.), drivers and cyclists clearly fail to obey traffic rules at both signalized and sign-controlled intersections. Hence the problem is how to prevent/mitigate these intersection-related crashes. The failure to comply need to be identified before they occur so actions can be taken to alleviate the consequences. To better understand the problem, it can be divided into two sub-problems. Thus, the following research questions are expected to be answered throughout the dissertation.

1. What is the transportation mode of the road user?
2. When approaching an intersection, how can we predict whether the driver/cyclist is going to violate the red light or stop sign?

It should be noted that in this dissertation, the driver violation prediction was conducted at signalized intersections and the cyclist violation prediction was conducted at sign-controlled intersections. Similar procedures can be followed to conduct driver violation prediction at stop signs and cyclist violation prediction task at signalized intersections.

Research plans

In order to answer the research questions, the research plans in this dissertation include the following:

1. Collect transportation mode data from different users when using various transportation modes through a smartphone application.
2. Develop a model to identify the transportation mode using smart phone data.
3. Analyze pre-collected observational passenger car data of different drivers to assess their violation behavior when approaching a signalized intersection.
4. Develop a model to predict if a driver is going to violate a red light using observational and data simulator data.
5. Design a naturalistic cycling experiment and collect bicycle data for different riders to assess their behavior when approaching intersections.
6. Analyze naturalistic cycling data to assess cyclist violation behavior when approaching signalized intersections and sign-controlled intersections.
7. Asses the applicability of the collected bicycle naturalistic data to develop cyclist violation prediction models at sign-controlled intersections.
8. Identify significant factors to predict violations at intersections.

Dissertation layout

The manuscript format was used for this dissertation for which a brief description of each chapter is presented below.

Chapter 1 - Introduction: this chapter gives an introduction, states the problem, and summarizes the research objectives. It also provides the proposal layout.

Chapter 2 - Model Development: this chapter shows how the problem is divided into three main tasks (i.e., transportation mode recognition, driver violation prediction, and cyclist violation prediction) and visually presents and discusses each part for which models were developed. These tasks are all sub-sections of the model development chapter. However, instead of having sub-chapters for these tasks and making a long single chapter for model development, four standalone chapters are provided.

Chapter 3 – Transportation Mode Recognition: this chapter includes three papers, co-authored by Dr. Hesham Rakha, aiming at developing models to recognize the mode of transportation using data from smartphone sensors.

Chapter 4 - Driver Violation Prediction: this chapter presents three papers, co-authored by Dr. Hesham Rakha and Dr. Thomas Dingus, which use observational data from a signalized intersection as well as simulator data to develop models for predicting Red Light Running (RLR) violations.

Chapter 5 - Bicycle Naturalistic Data Collection: this chapter contains a paper, co-authored by Dr. Hesham Rakha and Dr. Thomas Dingus, which explains the naturalistic data collection procedure for bicycles.

Chapter 6 - Cyclist Violation Prediction: this chapter includes a paper, co-authored by Dr. Hesham Rakha and Dr. Thomas Dingus, which concentrates on analyzing the bicycle naturalistic data to assess cyclist violation behavior and to evaluate the capability of developing violation prediction models for cyclists.

Chapter 7 - Conclusions and Future recommendations: this chapter presents the conclusions and future recommendations.

References

[1] NHTSA, *Traffic safety facts 2012*. 2014, National Center for Statistics and Analysis, US Department of Transportation, Washington, DC.

[2] IIHS. *Red light running*. 2012; Available from: http://www.iihs.org/iihs/topics/t/red-light-running/topicoverview.

[3] Insurance Institute for Highway Safety (IIHS), *Status Report: Public seeks safer roads but still takes risks*. 2010.

[4] The-University-of-North-Carolina-Highway-Safety-Research-Center. *North Carolina Bicycle Crash Types 2005 - 2009*. 2011; Available from: http://www.ncdot.gov/bikeped/download/summary_bike_types05-09.pdf.

[5] Hunter, W.W., et al., *Pedestrian and bicycle crash types of the early 1990's*. 1996.

[6] Tan, C., *Crash-type manual for bicyclists.* Publication No. FHWA-RD-96-104, 1996.

[7] McLeod, K., D. Flusche, and A. Clarke, *Where We Ride: Analysis of Bicycling in American Cities.* 2013.

Chapter 2: Model Development

Model Development

In order to prevent/mitigate intersection-related crashes that involve bicycles, violations (by both driver and rider) at intersections need to be identified before they occur, so appropriate warnings can be issued to the users in potential danger or to the infrastructure and consequently appropriate actions can be taken. Several factors influence the drivers'/riders' behavior when approaching intersections. These include the vehicle/bicycle speed [1], Time to Intersection (TTI) [1], Distance to Intersection (DTI) [1], age [2, 3], gender [3, 4], direction of travel [3, 5], presence of other road users [2, 3, 5], helmet use [6], and etc. The driver-related factors (e.g. age, gender) are more difficult to obtain in practice. On the other hand, kinetic factors (e.g. speed, acceleration) can be obtained by monitoring the movement of vehicles through video cameras installed on the infrastructure or through on-board devices installed on the vehicles. Hence, the problem of interest is to develop models to predict violations using kinetic information of individual bicycles/vehicles.

Identifying three main modules

To construct the models, the problem was divided into three main parts: (1) Transportation Mode Recognition (2) Driver Violation prediction (3) Cyclist Violation Prediction. The goal is to first identify the mode of transportation. Subsequently, violation prediction is conducted for the drivers and the cyclists. Figures 1 through 4 visually presents how these models perform. Figure 2 illustrates a situation in which three road users (shown as green, blue, and yellow arrows) are approaching a sign-controlled or signalized intersection. At this point, the transportation modes of the users are unknown.

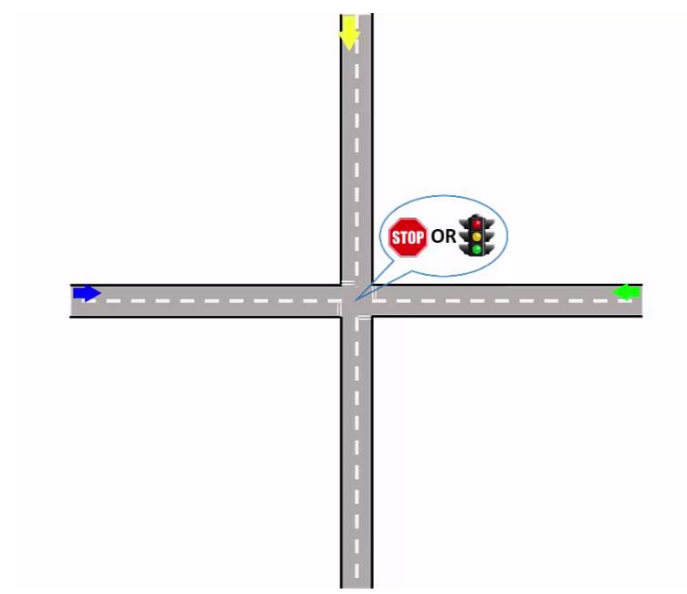


Figure 2 Road users approaching a sign-controlled or signalized intersection

As these users approach the intersection, at a desired Time To Intersection (TTI) or Distance To Intersection (DTI), transportation mode recognition starts as shown in Figure 3, which obtains sensor information such as accelerometer and gyroscope from the user for a short period of time.

Consequently, the users' modes of transport are identified as shown in Figure 4 and subsequently, violation prediction starts.

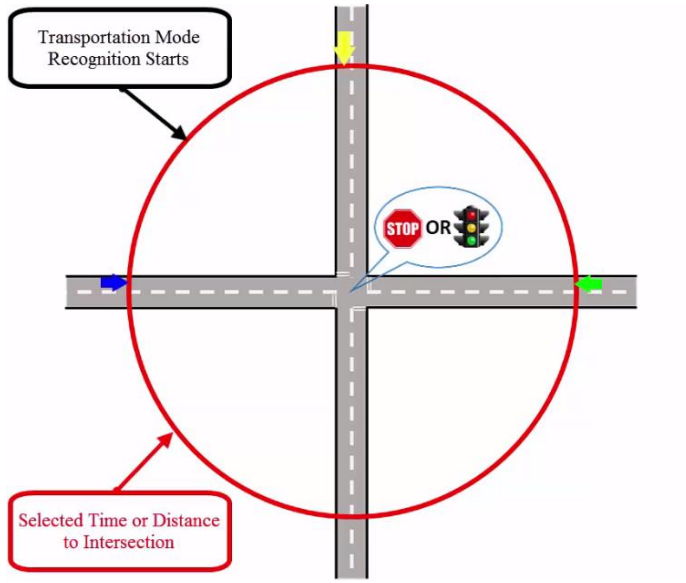


Figure 3 Transportation mode recognition starts at a specific point

The time period required to recognize the transportation mode is shown in Figure 4, which is a short timestamp. However, this task can be undertaken further away from the intersection and as the user becomes closer to the intersection, the model would check a number of times to make sure the user has not changed his/her mode and to increase the reliability of the mode recognition. Changing modes may occur specially when there is a bus station or a bike share station near the intersection.

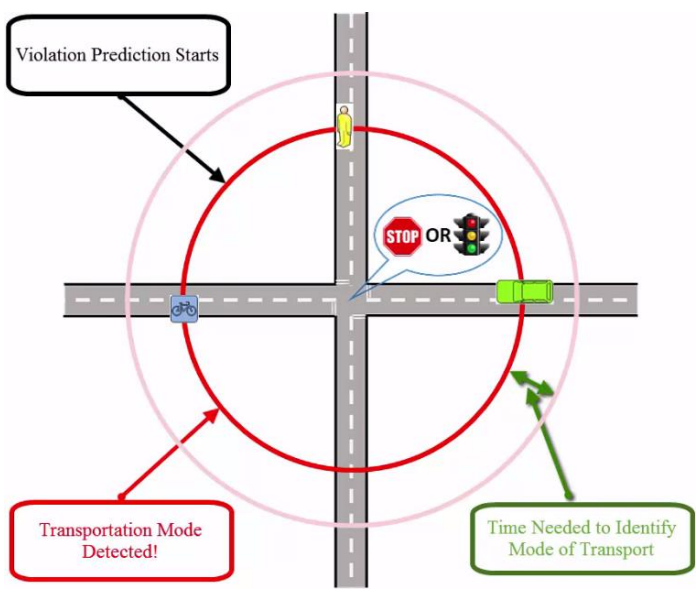


Figure 4 Violation prediction starts after transportation modes are detected

After the transportation mode is identified, the violation prediction is conducted for the mode identified. The focus of this dissertation is on the bicycle and the car modes. Hence, the capability of developing violation prediction models for only these two modes will be assessed. in order to carry out the prediction, a time window as shown in Figure 5 is selected, which gathers information such as speed, accelerometer, TTI at onset of yellow, and etc. from which the violation prediction models are developed.

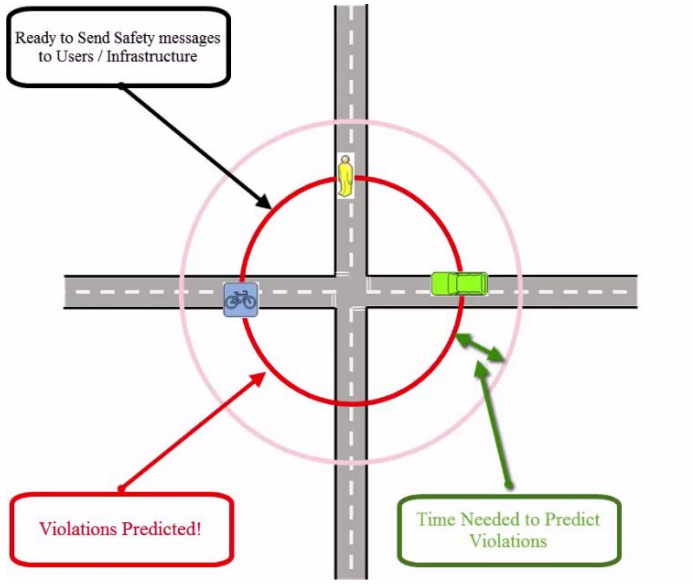


Figure 5 Safety messages can be sent to users and or infrastructure after violations are predicted

Once the prediction task is made, users in potential danger as well as the infrastructure can be notified to take appropriate actions with the aim of reducing/mitigating crashes.

In the next four chapters, the three main tasks (i.e., transportation mode recognition, driver violation prediction, cyclist violation prediction) that were visually shown in this chapter are presented.

References

[1] Gates, T.J., et al., *Analysis of driver behavior in dilemma zones at signalized intersections.* Transportation Research Record: Journal of the Transportation Research Board, 2007. **2030**(1): p. 29-39.

[2] Wu, C., L. Yao, and K. Zhang, *The red-light running behavior of electric bike riders and cyclists at urban intersections in China: an observational study.* Accident Analysis & Prevention, 2012. **49**: p. 186-192.

[3] Johnson, M., et al., *Why do cyclists infringe at red lights? An investigation of Australian cyclists' reasons for red light infringement.* Accident Analysis & Prevention, 2013. **50**: p. 840-847.

[4] Johnson, M., J. Charlton, and J. Oxley. *Cyclists and red lights—a study of the behaviour of commuter cyclist in Melbourne*. in *Australasian Road Safety Research, Policing and Education Conference, Adelaide*. 2008.

[5] Johnson, M., et al., *Riding through red lights: The rate, characteristics and risk factors of non-compliant urban commuter cyclists.* Accident Analysis & Prevention, 2011. **43**(1): p. 323-328.

[6] Pai, C.-W. and R.-C. Jou, *Cyclists' red-light running behaviours: An examination of risk-taking, opportunistic, and law-obeying behaviours.* Accident Analysis & Prevention, 2014. **62**: p. 191-198.

Chapter 3: Transportation Mode Recognition

(Paper 1 accepted to: Transportation Research Board 93rd Annual Meeting, 2014)

(Paper 2 published in: IEEE Transactions on Intelligent Transportation Systems)

(Paper3 accepted to: ITS World Congress, 2015)

Developing a Support Vector Machine (SVM) Classifier for Transportation Mode Identification using Mobile Phone Sensor Data

Arash Jahangiri
Center for Sustainable Mobility, Virginia Tech Transportation Institute
3500 Transportation Research Plaza, Blacksburg, VA 24061
E-mail: arashj@vt.edu
Phone: (540) 200-7561

Hesham Rakha (corresponding author)
Center for Sustainable Mobility, Virginia Tech Transportation Institute
3500 Transportation Research Plaza, Blacksburg, VA 24061
E-mail: HRakha@vtti.vt.edu
Phone: (540) 231-1505

Word count: 5,002 + 2,000 (1 Figures + 7 Tables) = 7,002

Submitted for presentation at the 93rd Annual Meeting of the Transportation Research Board and publication in the *Transportation Research Record*

Abstract

Identifying the transportation mode can offer several advantages in different fields of transportation engineering such as transportation planning and intelligent transportation systems which lead to a broad range of environmental and safety applications. Support vector machine, as a supervised learning method, is adopted in this paper to develop a multi-class classifier to distinguish between different transportation modes including driving a car, riding a bicycle, taking a bus, walking, and running. Data from different mobile phone sensors were trained and tested to evaluate the model. Sensors from which the data were obtained include accelerometer, gyroscope, rotation vector, and Global Positioning System (GPS). A Gaussian kernel was applied as part of the classifier and unlike some ambiguity seen in the literature, a complete model selection is conducted. A small window size of one second was considered, so the model can be useful in a broader range of applications. For the first time, the data from gyroscope and rotation vector sensors were used in experiments based on individual sensor data. The study showed that such data can contribute to high classification rates. It was found that including attributes that have similar behavior among different modes can negatively impacts the classification rates. When using multiple sensors, high average overall accuracies of 98.86% and 97.89% were achieved with and without using the GPS data, respectively. These results offer improvements compared to what is reported in the literature. The bus mode was the most difficult mode to differentiate due to some similarities to the car and the bike mode data.

Keywords: transportation mode; support vector machine; mobile phone sensor data; machine learning

Ch. 3 - Transportation Mode Recognition

Introduction

Recognizing different types of physical activities using sensor data has been a recent research topic that has received considerable attention [1, 2]. Transportation mode classification can be considered as an activity recognition task in which data from smartphone sensors carried by users are utilized to infer what transportation mode the individuals are using. Micro-electromechanical systems (MEMS), such as accelerometers and gyroscopes are embedded in most smartphone devices [3] from which the data can be obtained at high frequencies. Smartphones, nowadays, are equipped with powerful sensors such as GPS, accelerometer, gyroscope, light sensors, temperature sensors, etc. Having such powerful sensors all embedded in a small device carried in everyday life activities has enabled researchers to investigate new research areas. Other advantages of these smart devices are their ubiquity, their ability to send and receive data through Wi-Fi/cellular network/Bluetooth, to provide alerts, and store data as well as to process the data [4].

The knowledge of individuals' mode of transport can facilitate some tasks and also can be adopted in several applications. Knowing the mode of transportation is an essential part of urban transportation planning, which is usually investigated through questionnaires/travel diaries/telephone interviews [4, 5] .This traditional way of surveying is usually expensive, erroneous, limited to a specific area, and does not incorporate the latest information [6]. As an environmental application, the carbon footprint as well as the amount of calories burnt of individuals can be determined by obtaining the mode of transport. Other applications include providing users with real-time information using the knowledge of speed and transport mode from the users as probes [4, 7], Providing individuals with customized advertisements and messages based on the transportation mode they are using [4], physical activity and health monitoring, tracking the hazard exposure and assessing the environmental impact of one's activities, and profile based recruitment for distributed data gathering [8].

Many studies have used GPS for classification purposes. However, several limitations are associated with the use of GPS sensors. These limitations include: GPS information is not available in shielded areas (e.g. tunnels) and the GPS signals may be lost especially at high dense locations which results in inaccurate position information. Moreover, the GPS sensor consumes significant power that sometimes users turn it off to save the battery [6, 7]. This paper focuses on developing a classifier using the support vector machine method and data obtained from smartphone sensors including accelerometer, gyroscope, rotation vector, and GPS data. Consideration of multiple sensors is beneficial in that even without using GPS the transportation modes can be identified. The unique contributions of this research effort are:

1. Exploiting data from sensors other than those used in the literature including gyroscope and rotation vector data,
2. Increasing the prediction accuracies with almost real-time prediction (time window of one second), and
3. Developing a complete model selection procedure of support vector machine using Gaussian kernels.

The remainder of the paper is organized in the following five sections. Relevant literature is reviewed in the next section followed by the data collection section. Subsequently, the development of the proposed model is discussed using support vector machine techniques. Subsequently, the results of the study are presented and finally the conclusions of the study are presented.

Relevant work

Table 2 presents a summary of past studies focusing on identifying transportation modes. Almost all studies used data from GPS sensors that have the aforementioned drawbacks. Also, all studies took advantage of Artificial Intelligence (AI) tools such as Fuzzy Expert Systems as in [9] Decision Trees as in

[4-8, 10], Bayesian Networks as in [4, 10], Random Forests as in [4], Naïve Bayesian techniques as in [4, 8], Neural Networks as in [11, 12], and Support Vector Machine (SVM) techniques as in [8, 10, 13-16], of which the Decision Tree and SVM methods were used the most. To improve the model performance some other techniques were also combined with machine learning methods such as Discrete Hidden Markov Models as in [8] and Bootstrap aggregating as in [17]. Other than AI tools, statistical methods were also applied such as the Random Subspace Method by [18]. Some studies have used additional information from GIS maps as in [4, 9, 19, 20]. However, GIS data is not always available, and also this approach may not be suitable for real-time applications because it mostly relies on the knowledge of the entire trip with respect to the GIS features such as bus stops, subway entrances, and rail lines.

The Decision Tree method was identified as the best method by [8, 10] compared to some other methods including SVM. However, when applying SVM, several factors can greatly influence the model performance, which have not been considered in previous work. For example, a linear kernel was used in [8, 10] as part of the method, but generally for a certain type of problems and depending on the size of the available data and features, SVM can produce better results with more advanced kernels such as Gaussian kernel. Also, when applying Gaussian kernel, it was shown that if complete model selection is conducted with Gaussian there is no need to consider the linear kernel [21]. It is also unclear whether feature scaling and regularization were adopted in the most studies using SVM. Feature scaling is used to normalize the range of different features (or attributes), which leads to higher model performance and training speed and the regularization is incorporated into the model to deal with the issue of over-fitting (high variance). Gaussian kernel was only used in three studies; however, [14] did not conduct the complete model selection. In other words, constant values for the regularization parameter and the Gaussian parameter were used. It appears that [15] did not consider regularization parameter, and also they mentioned that Gaussian parameter should be optimized, but the optimized value was not reported. [16] reported the best regularization parameter (or cost parameter) to be 3, but the method by which they obtained this value is unclear. In addition, the value of Gaussian parameter they applied is not stated.

Depending on the application of interest, different window sizes have been used for predicting the mode of transport. For example, [12] found that longer monitoring durations lead to higher accuracy. Intuitively, the bigger the window size the easier the prediction becomes since with bigger window sizes more information is available. If the application is only a survey for demand analysis the window size can be as large as trip duration, whereas if the application provides real-time information for environmental or some transit applications, then smaller window sizes are more desirable. The size should be as small as possible for some safety applications (e.g. crash prevention/mitigation). A study [13] used 200-meter and 150-second segments in their experiment. Whereas another study [6] used 10-second time windows to separate walking from non-walking segments and then applied a maximum size of 2 minutes. Other than the window size, the overlaps of two consecutive windows have also been considered. Reference [7] obtained the best window size and overlap to be 10.24 seconds and 50%, respectively. The entire trip duration appears to be considered in [5, 9, 11, 16, 20].

Table 2 presents different classes, the data, and the overall accuracy of the prediction models for different studies; however, the overall accuracy was not reported in some of them for which the averages of the reported values are considered in here. Also, it should be noted that, high classification rates were achieved for some of the classes (not all), as such, accuracy of 98% and 92% were obtained by [18] for bicycle and walk classes, respectively. Also, the reported values by [12] are for a 10-minute window size and one ping every 2-minutes. The studies showed higher accuracies were achieved by increasing these two parameters.

Higher accuracies are achieved by increasing the window size as shown in [12]. Since the focus of the present study is on small window sizes, in order to ensure a fair comparison of the various studies only those with window sizes less than a minute are considered, as summarized in Table 3. Thus, the application would include a broader range of applications such as environmental and safety applications.

Table 2 Summary of Past Efforts

Studies	Classes	Data	Window Size	Accuracy	Studies	Classes	Data source	Window Size	Accuracy
[13]	1- Car 2- Walk 3- Bus 4- Bike	1- GPS	200-meter and 150-second segments	83.6[a]	[7]	1- Bus 2- Metro 3- Walk 4- Bicycle 5- Train 6- Car 7- Still 8- Motorcycle	1-Accelerometer rate of 25 Hz	10.24 seconds , 50% overlap	82.14
[6]	1- Walk 2- Bike 3- Motorcycle 4- Car 5- Bus 6- Tram 7- above train 8- subway	1- GPS 3- Accelerometer data	10-second / maximum of 2 minutes	75.8[a]	[11]	1- car 2- bus 3- walk	1- GPS data	Entire trip	91.23
[19]	1- walk 2- car 3- bus 4- subway 5- commuter rail	1- GPS-based travel survey 2- GIS data from local agencies	Developed rules to identify trip segments	82.6	[12]	1- Car highway 2- Car arterial 3- Bus arterial 4- streetcar 5- walk	1- GPS logger	1/5/10/15/20 minutes	82.2[b]
[4]	1- car 2- bus 3- aboveground train 4- walking 5- bike 6- stationary	1- GPS 2- GIS	30 seconds	93.5	[14]	1- walk 2- bike 3- run 4- car 5- train 6- bus	1- Accelerometer	5 seconds , 50% overlap	93.88[c]
[8]	1- stationary 2- walk 3- run 4- bike 5- motorized transport	1- GPS 2- Accelerometer	1 second	93.6	[22]	1- Walk 2- jog/run 3- bike 4- inline skating 5- car	1-GPS 2-Accelerometer	Entire trip	97.7
[9]	1- Stationary 2- Walk 3- Car 4- Train 5- Tram 6- Underground 7- Bicycle 8- Bus 9- Ferry 10- Sail boat 11 - Aircraft	1- GPS 2- GIS	Entire trip	91.6	[16]	1- walk 2- Car 3- Train 4- Bicycle 5- Bus 6- Tube	1- GPS	Entire trip	88
[18]	1- Walk 2- Car 3- Train 4- Tram 5- Metro 6- Bicycle 7- Bus 8- Motorcycle	1- GPS 2- Accelerometer	>20 seconds	61.75 /78.8[d]	[15]	1- Car 2- Train 3- Pedestrian	1-Accelerometer	4 seconds – 50% overlap	96.9 / 97.3[e]

[a] the overall accuracy not reported. Here, the average of the reported recall values are used

[b] the overall accuracy not reported. Here, the average of the reported precision values are used

[c] it appears that the reported accuracy is for the first four classes

[d] the overall accuracy not reported. Here, the average of the reported recall values are used. Also first value is for when 8 classes are considered and the second value is for when 6 classes are considered, meaning that classes 3, 4, and 5 are combined as a single class

[e] 96.9 obtained with the time window of 4 seconds / 97.3 obtained considering ten consecutive windows that leads to window size of 40 seconds

Other than the window size, several factors are shown in Table 3 that also influence the model performance as follows:

(1) Number of classes: as the number of classes increases, class differentiation becomes more difficult.
(2) Use of accelerometer/GPS/GIS data: the level of model dependency on different sources of data is considered as an important factor. Less dependent models are more desirable as they can be applicable even with limited sources of data. In this case, sensors such as accelerometers and gyroscopes are more reliable since their data are available most of the time.
(3) Ability to distinguish between motorized classes: as different motorized classes have similar characteristics such as speed and acceleration, a model capable of differentiating between these modes is of great value. For example, distinguishing the bus mode from the car mode is significantly more difficult than discriminating walking from driving.
(4) Sensor positioning: it shows how realistic the experiments are conducted. Positioning the devices at certain locations increases the prediction accuracy because the movements can be monitored in more detail, but may not reflect realistic behavior. Some of the studies required that the participants attach sensors/smartphones to different parts of their body.

The highest reported accuracy of 96.9% is achieved by [15] with a window size of 4 seconds. In this approach only accelerometer data were used and they did not rely on GPS and GIS data. Their method is capable of differentiating between motorized modes (car and train) and no specific sensor positioning was applied. Nevertheless, they only considered three classes. The second best accuracy is obtained by [14]. They also used accelerometer data without relying on GPS/GIS data. However, although different motorized modes were mentioned in the paper, it seems that the reported accuracies show only one motorized mode. Also, subjects in their study were asked to keep their device in their pocket of the non-dominant hip while collecting data which is more realistic compared to attaching sensors to the body, but still does not reflect a complete realistic behavior. [8] reported the accuracy of 93.6 which is ranked third in the table. They applied the lowest window size throughout the literature. However, their approach was dependent on data from GPS sensors. Moreover, different motorized classes were not considered.

Table 3 Key Features of Studies using a Time Window Less than a Minute

Study	Number of classes	accelerometer	GPS	GIS	Different motorized	positioning	Window size (seconds)	Overall Accuracy (%)
[4]	6	no	yes	yes	yes	Not specific requirements	30	93.5
[8]	4	yes	yes	no	no	Not specific requirements	1	93.6
[7]	8	yes	no	no	yes	Not specific requirements	10.24	82.14
[14]	4	yes	no	no	no	In pocket of non-dominant hip	5	93.88
[15]	3	yes	no	no	yes	Not specific requirements	4	96.9
[18]	8/6	yes	yes	no	yes	Not specific requirements	>20	61.75/78.8

Data collection and preprocessing

A smartphone application was developed for the purpose of data collection. To collect the data, the transportation mode should be selected before starting the logging process, and then the application stores the data coming from smartphone's sensors including GPS, Accelerometer, Gyroscope, and Rotation Vector at the highest possible frequency. In order to ensure that the data are gathered at identical sampling rates linear interpolation was applied to the data similar to [7] to produce continuous data sets and finally the data were re-sampled at the desired rate (rate of 100 Hz was applied). Data collection was carried out

by three individuals using two different android phones (i.e. Galaxy Nexus and Nexus 4). A total of 7 hours of data were stored and used for training and testing purposes. The data in minutes were comprised of about 50, 20, 270, 15, and 70 for Car, Bus, Bike, Run, and Walk modes, respectively.

Model development

SVM is known as a large margin classifier, which means when classifying data, it determines the best possible decision boundary that provides the largest possible gap between classes. This characteristic contributes to a higher confidence in solving classification problems. To implement SVM, the LibSVM library of SVMs was applied. For multiclass classification, considering h classes, LibSVM applies one-against-one method in which $h(h-1)/2$ binary models are built. Among these, LibSVM chooses the parameters that achieve the highest overall performance. Another well-known method is called one-against-all which is more intuitive and has similar performance. However, LibSVM takes advantage of one-against-one because of its shorter training time. Using the LibSVM package, a data set can be trained to build a prediction model for classification, and then evaluate the model by testing it on another data set [23].

To construct the model, the following factors are taken into account: using a Gaussian kernel with complete model selection, which entails consideration of the regularization parameter and the Gaussian parameter, applying feature scaling, and examining several features. The accuracy is obtained using three metrics, namely: overall accuracy, precision and recall. These three metrics are used for model evaluation. The entire data set is divided into two groups; one for training and the other for testing or evaluating how well the model is performing. The overall accuracy is calculated by dividing the total number of correct predictions by the total number of test data. The recall is calculated by dividing the total number of true positives by the total number of actual positives. The precision is computed by dividing the total number of true positives by the total number of predicted positives.

Equation 1 presents the SVM formulation to solve the classification problem and the associated constraints are shown in Equation 2 and Equation 3 [24]. The objective function is comprised of two terms: minimizing the first term is basically equivalent to maximizing the margin between classes, and the second term consists of an error term multiplied by the regularization (penalty) parameter denoted by C. The C parameter should be determined to provide the relative importance between the two terms. Equation 2 ensures that margin of at least 1 exist with consideration of some violations. The value of 1 was resulted from normalizing w. Equation 3 restricts the data points to the points that have positive errors.

$$\min_{w,b,\xi}\left(\frac{1}{2}w^T w + C\sum_{n=1}^{N}\xi_n\right) \qquad \text{Equation 1}$$

Subject to:

$$y_n(w^T\phi(x_n)+b) \geq 1-\xi_n\,, n=1,\dots,N \qquad \text{Equation 2}$$

$$\xi_n \geq 0\,, n=1,\dots,N \qquad \text{Equation 3}$$

Where,

w	Parameters to define decision boundary between classes
C	Regularization (or penalty) parameter
ξ_n	Error parameter to denote margin violation
b	Intercept associated with decision boundaries
$\phi(x_n)$	Function to transform data from X space into some Z space

Kernels are functions that are adopted to create the features based on the provided attributes in a higher maybe infinite dimensional Z space. So, basically, for a function $\phi(x_n)$ that transfers data from X space into the higher dimensional Z space, the kernel corresponds to the vector inner products in the Z space. Different types of kernels exist such as linear kernel, polynomial kernels, and Gaussian kernel. Linear kernel, as applied in [8, 10], is the basic mode which means no kernels are actually taken into account. In other words, vector inner product as appears in the dual formulation of the problem are considered without transforming data into another space. According to our data size and attribute size, Gaussian kernel was believed to be the most appropriate kernel [25], and as noted earlier, if a complete model selection is carried out, there is no need to test the linear kernel because the results obtained from the Gaussina kernel include the results obtained from the linear kernel. In fact, when using Gaussian kernel, If $\sigma^2 \to \infty$ and $C = C^L\sigma^2$ where C^L is fixed then the SVM classifier behaves like an SVM classifier with a linear kernel with regularization parameter C^L [21]. In this paper, the $\phi(x_n)$ function which corresponds to the Gaussian kernel has an infinite dimensional space. The formulation of the Gaussian kernel is shown in Equation 4.

$$K(x, x') = exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) \qquad \text{Equation 4}$$

Where,

x, x'	n-dimensional vectors
$\|x - x'\|$	Euclidean distance between vectors x, x'
σ	Gaussian parameter

n-dimensional vectors are basically vectors of attributes. In other words, each vector is an instance of the available data consisting of different attributes. For example, an instance of the training dataset with only time and acceleration data is a 4-dimentional vector as shown in Equation 5 below.

$$x^{train} = (x_1^{train}, x_2^{train}, x_3^{train}, x_4^{train}) = (t, a_x, a_y, a_z) \qquad \text{Equation 5}$$

Where,

t	The timestamp at which the data are stored
a_x, a_y, a_z	Accelerations along the x, y, z axes

Attribute Selection

At first glance, the velocity seems to be a feature by which transportation modes can be easily identified. However, traffic conditions and weather conditions can greatly influence the speed in a way that similar speed values are observed from different modes. Also, driving on local roads and riding on bicycle on the same routes may have similar velocities [17].

Features are basically generated by the kernel function using the training data set. In other words, every single data point is used by the kernel function to create a new feature. Different data attributes (also called features/indicators) such as speed, acceleration, etc. are introduced to the model for feature creation. Attributes are basically used to differentiate between transportation modes.

Some attributes are considered to be basic/traditional attributes (e.g. mean speed), which are more intuitive to be influential and are widely used in the literature and some are considered to be more advanced attributes (e.g. heading change rate) as presented by [13]. Some methods have been applied to select the most relevant attributes to use such as ANOVA tests used in [16], correlation based feature selection (CFS) used in [8], and Chi Squared and Information gain methods applied in [4]. A similar approach to what [15] applied was used in our study. While preprocessing the data [15], for each time

window, computed the standard deviation, the maximum value, the norm, and the number of sign changes of the cumulative acceleration values $(a_x + a_y + a_z - g)$, where a_x, a_y, a_z are accelerations along the x, y, z axes and g is the gravitational acceleration. The total acceleration values of each time window $(\sqrt{{a_x}^2 + {a_y}^2 + {a_z}^2} - g)$ were also used to create sets of attributes, and finally the combination of all sets of attributes was also examined. In this paper, however, instead of adding acceleration values, individual values were considered to account for the individual effects. The total acceleration was also included without gravity acceleration because the linear acceleration sensor was used from which the gravity force is already excluded. A similar procedure was applied to data obtained from the gyroscope, rotation vector, and GPS sensors. To see the effects of individual sensors, the set of attributes computed for each sensor was examined by itself and finally the entire sets of attributes were examined. Table 4 presents the sets of attributes.

Table 4 sets of attributes from different sensors

set 1 - Accelerometer	set 2 - Gyroscope	set 3 - Rotation Vector	set 4 - GPS
$\overline{a_x}$	$\overline{g_x}$	$\overline{rv_x}$	$\bar{v}$
$\overline{a_y}$	$\overline{g_y}$	$\overline{rv_y}$	$range(v)$
$\overline{a_z}$	$\overline{g_z}$	$\overline{rv_z}$	$stdv(v)$
$\sqrt{\overline{a_x}^2 + \overline{a_y}^2 + \overline{a_z}^2}$	$\sqrt{\overline{g_x}^2 + \overline{g_y}^2 + \overline{g_z}^2}$	$\sqrt{\overline{rv_x}^2 + \overline{rv_y}^2 + \overline{rv_z}^2}$	$iqr(v)$
$range(a_x)$	$range(g_x)$	$range(rv_x)$	
$range(a_y)$	$range(g_y)$	$range(rv_y)$	
$range(a_z)$	$range(g_z)$	$range(rv_z)$	
$stdv(a_x)$	$stdv(g_x)$	$stdv(rv_x)$	
$stdv(a_y)$	$stdv(g_y)$	$stdv(rv_y)$	
$stdv(a_z)$	$stdv(g_z)$	$stdv(rv_z)$	
$iqr(a_x)$	$iqr(g_x)$	$iqr(rv_x)$	
$iqr(a_y)$	$iqr(g_y)$	$iqr(rv_y)$	
$iqr(a_z)$	$iqr(g_z)$	$iqr(rv_z)$	
$number\ of\ sign\ changes(a_x)$	$number\ of\ sign\ changes(g_x)$	$number\ of\ sign\ changes(rv_x)$	
$number\ of\ sign\ changes(a_y)$	$number\ of\ sign\ changes(g_y)$	$number\ of\ sign\ changes(rv_y)$	
$number\ of\ sign\ changes(a_z)$	$number\ of\ sign\ changes(g_z)$	$number\ of\ sign\ changes(rv_z)$	

Results

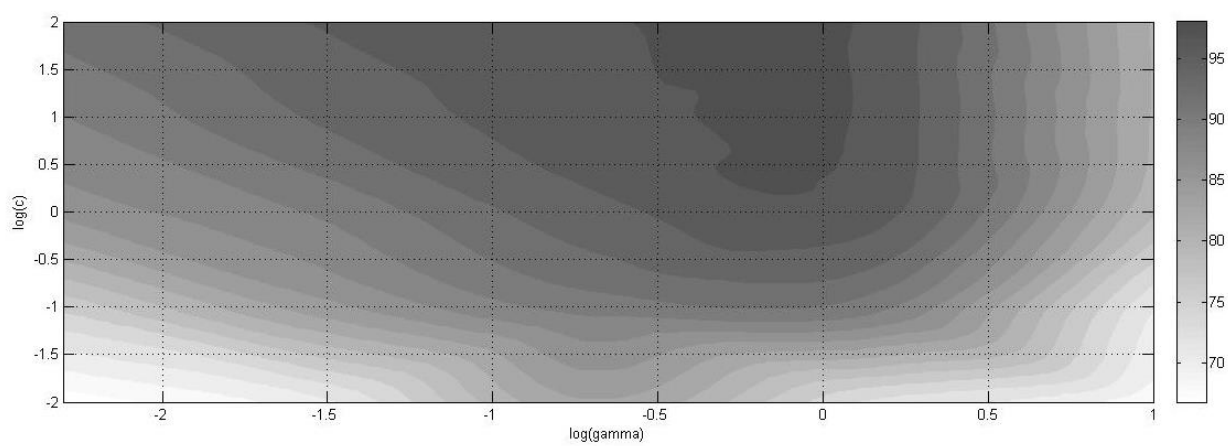
The data gathered were divided into a training set (70 percent of the data) and a testing set (30 percent of the data). The distinction between the training and testing set was conducted randomly across all five modes of travel. Six scenarios were assessed based on the set of attributes used. Table 5 presents the overall accuracy as well as other key factors associated with each scenario. These results are obtained for the testing set. Scenarios 1 through 4 accounts for attributes obtained from the accelerometer, gyroscope, rotation vector, and GPS sensors, respectively, and evaluate the individual sensor effects. Scenario 5 and 6 reflect combined effects of using multiple sensors. Scenario 6 uses data from all sensors, while scenario 5 uses accelerometer, gyroscope, and rotation vector sensors excluding the data from the GPS sensor.

Scenario 6 clearly achieved the best accuracy while scenario 5 also reached accuracies close to scenario 6. The advantage of scenario 5 is that it does not rely on data from the GPS sensor and thus requires less power since considerable battery usage is associated with GPS sensors. For these preliminary results, constant values of 1 and 0.01 are considered for the regularization and the Gaussian parameter, respectively.

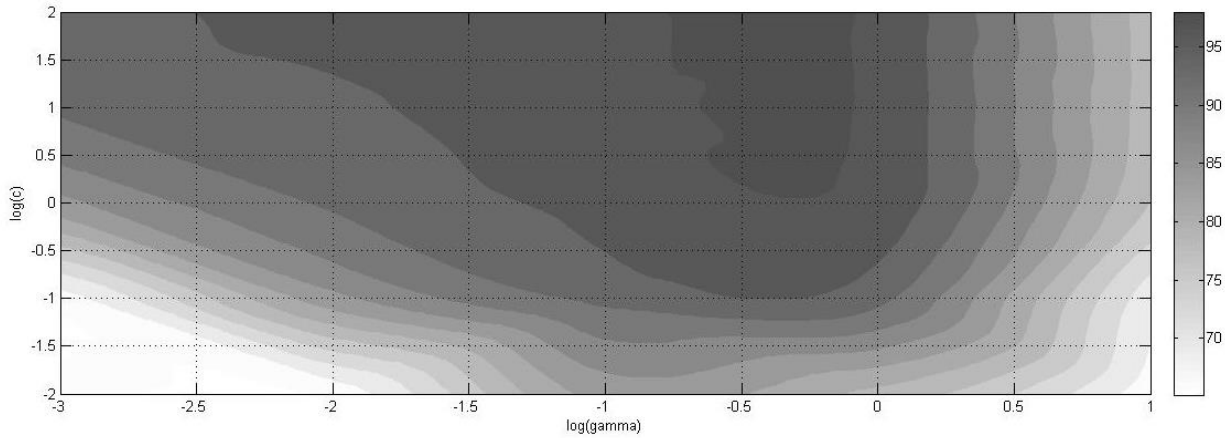
Table 5 overall accuracy and key points of different scenarios, constant values of regularization and Gaussian parameters considered

Scenarios	attributes	Number of classes	Accelerometer/ gyroscope / rotation vector	GPS	GIS	Different motorized	Positioning	Window size (seconds)	Overall Accuracy
1	set 1	5	yes	no	no	yes	No specific requirements	1	83.46
2	set 2	5	yes	no	no	yes	No specific requirements	1	80.45
3	set 3	5	yes	no	no	yes	No specific requirements	1	75.02
4	set 4	5	yes	yes	no	yes	No specific requirements	1	83.40
5	set 1,2,3	5	yes	no	no	yes	No specific requirements	1	88.66
6	set 1,2,3,4	5	yes	yes	no	yes	No specific requirements	1	93.92

Even higher accuracies were achieved by conducting the complete model selection. In order to complete the model selection, the regularization parameter (parameter c) as well as the Gaussian parameter should be optimized. The Gaussian kernel formulation used in libSVM is slightly different from Equation 4; in their formulation, the parameter $gamma$ was used instead of $\frac{1}{2\sigma^2}$. Figure 6 presents contour plots that illustrate how different values of the regularization (c) and the Gaussian ($gamma$) parameters impact the performance of the models used in scenario 5 and 6, respectively. The optimal values for ($gamma, c$) were found to be (0.63, 63.1) and (0.4, 63.1) for scenarios 5 and 6 that led to the high overall accuracies of 98.23% and 98.78%, respectively. Parameter c deals with the issues of over fitting and under fitting. Choosing a too large/too small value for the regularization parameter results in under fitting/over fitting. In other words, the model will suffer from high bias if too small values of the regularization parameter are applied, and on the other hand, if too large values of the regularization parameter are used the model will suffer from high variances. The Gaussian parameter also impacts bias and variance seen in the model. With small values of $gamma$ (or large values of σ), features can vary more smoothly leading to higher bias and lower variance. Also, when using large values of $gamma$, features can vary less smoothly which results in lower bias and higher variance.



(a) Scenario 5



(b) Scenario 6

Figure 6: Impacts of the Regularization and the Gaussian Parameters on Model Accuracy for Scenarios 5 and Scenario 6

Table 6 presents confusion matrices for scenarios 5 and 6, which shows the classification rates (i.e. true positives and true negatives in percentage based on actual values) for each mode as well as the misclassification rates (i.e. false positives and false negatives in percentage based on actual values). Since true positives are reported in percentages based on actual values, they are essentially the recall values. The highest recall of more than 99% was obtained for the bike mode in both scenarios. Moreover, the model predicts the other modes with high recalls. However, the lowest accuracy, as expected, was for the bus mode. In scenario 5, more than 7% of the time the bus mode were misclassified as bike and car modes. In scenario 6, more than 7% of the time the bus mode was misclassified as the car mode, which was the highest misclassification rate. Similarly, high precision accuracies of different modes show that the models performed very well.

Table 6 Confusion matrices in percentage for scenarios 5 and 6

Scenario 5		Actual						Scenario 6		Actual					
		Bike	**Car**	**Walk**	**Run**	**Bus**	**Precision**			**Bike**	**Car**	**Walk**	**Run**	**Bus**	**Precision**
Predicted	**Bike**	99.30	0.50	4.47	0.52	3.56	98.63	Predicted	**Bike**	99.68	0.37	1.35	1.04	1.62	99.48
	Car	0.06	98.38	0.00	0.00	3.88	98.13		**Car**	0.00	97.63	0.00	0.00	7.77	97.02
	Walk	0.60	0.25	95.45	1.04	0.97	96.84		**Walk**	0.28	0.00	98.40	1.56	0.65	98.40
	Run	0.00	0.00	0.00	98.44	0.32	99.47		**Run**	0.00	0.00	0.00	97.40	0.32	99.47
	Bus	0.04	0.87	0.08	0.00	91.26	96.58		**Bus**	0.04	2.00	0.25	0.00	89.64	92.95
	Recall	99.30	98.38	95.45	98.44	91.26			**Recall**	99.68	97.63	98.40	97.40	89.64	

The features used in scenario 5 were a subset of features used in scenario 6. When using additional information obtained from the GPS sensor (as in scenario 6) the recall values of the walk and run modes increased by more than 3%. In addition, the recall value of the bus mode improved by slightly more than 1% and there were minor improvements in the recall values of the bike and the car modes (less than 1%). These changes make sense since the additional features, which are all speed related variables, are better indicators for distinguishing between the walk and run mode from the other modes due to the obvious speed differences, but they may not be good indicators to distinguish between the bus and car modes due to their similar speeds. It should be noted that in general, the improvement obtained by including the GPS data was not significant (1.78% change in average recall value and almost no change in average precision value).

The entire training and testing procedures were conducted ten more times using the optimal values obtained from the model selection task for scenarios 5 and 6 to show the robustness of the model. In this case 70 percent of the data were used for training and the remaining 30 percent were used for testing procedures. These 70 and 30 percent were randomly selected for each of the repetitions. Standard deviation or the recall values is applied as an indicator to show how the accuracies vary in different runs as shown in Table 7. Small values of standard deviation show that the models in both scenarios are extremely robust.

Table 7 Robustness of the Developed Model

	Scenario 5						Scenario 6					
	Bike	**Car**	**Walk**	**Run**	**Bus**	**Average**	**Bike**	**Car**	**Walk**	**Run**	**Bus**	**Average**
base	99.30	98.38	95.45	98.44	91.26	96.57	99.68	97.63	98.40	97.40	89.64	96.55
1	99.22	96.88	95.03	96.88	89.64	95.53	99.56	96.00	98.65	97.92	91.26	96.68
2	99.18	97.75	94.95	92.71	89.97	94.91	99.74	96.88	98.90	99.48	92.23	97.45
3	99.18	97.63	96.29	94.79	89.00	95.38	99.64	98.50	98.90	97.92	91.26	97.24
4	98.98	97.00	95.20	93.75	90.29	95.04	99.60	97.63	98.99	98.44	91.91	97.31
5	98.74	96.75	95.96	96.88	86.73	95.01	99.72	99.72	99.16	98.44	90.29	97.46
6	99.12	98.38	96.29	95.31	89.32	95.68	99.62	98.13	98.48	97.92	87.06	96.24
7	99.00	96.75	95.70	92.71	89.64	94.76	99.56	97.38	98.23	97.92	90.94	96.80
8	98.98	97.00	95.11	92.19	87.06	94.07	99.50	98.25	99.33	97.92	90.94	97.19
9	99.30	98.38	95.45	98.44	91.26	96.57	99.58	97.63	98.90	98.44	89.64	96.84
10	99.20	96.63	96.97	93.75	89.32	95.17	99.64	97.75	98.65	97.40	89.97	96.68
Average	99.11	97.41	95.67	95.08	89.41	95.34	99.62	97.77	98.78	98.11	90.47	96.99
Standard Deviation	0.17	0.71	0.64	2.29	1.45	0.65	0.07	0.94	0.33	0.58	1.42	0.40

Higher accuracies were obtained when comparing the present study with similar studies as listed in Table 3. The study carried out by [8] was considered to be the most similar research effort for the sake of a fair comparison since their study was the only one that chose a one-second time window as done in our study. This comparison is shown in Table 8. Also, only scenario 5 is presented in this table to show that even without using data from the GPS sensor, a higher accuracy was achieved. Furthermore, as mentioned earlier, [8] did not consider differentiating between motorized modes and their method also relied on GPS data. Other than the accelerometer, the present study took advantage of data from the gyroscope and rotation vector sensors. It should be noted that they used a larger dataset collected from 16 users. A larger dataset probably include more variability and thus more difficult to distinguish between modes. The higher accuracies obtained in this paper might be due to having less data from only three users. However, it might be due to conducting a complete model selection or examining a large number of features in this paper or both.

Table 8 Comparison with the Most Relevant Study

studies	Number of classes	Accelerometer/ gyroscope / rotation vector	GPS	GIS	Different motorized	positioning	Window size (seconds)	Overall Accuracy
[8]	4	yes	yes	no	no	Not specific requirements	1	93.60
Present Study	5	yes	no	no	yes	Not specific requirements	1	95.34

Conclusions

A classifier was developed using the support vector machine learning technique to identify different transportation modes including bike, car, walk, run, and bus. To train and test the classifier, data were obtained from smartphone sensors such as accelerometer, gyroscope, rotation vector, and GPS sensors. This effort is the first application to use gyroscope and the rotation vector sensors for the purpose of transportation mode classification. Individual experiments showed that both of them are significant indicators for distinguishing different modes. A Gaussian kernel was applied to create features from different sets of attributes coming from different sensors. When using multiple sensors simultaneously, a complete model selection was conducted to obtain the optimal regularization parameter and the optimal Gaussian parameter resulting in very accurate and extremely robust models. A time window of one second was chosen, so the model can fit in a broader range of applications. Comparing to the only study in which a time window of one second was used, higher accuracies were achieved. The focus of the future work will be on error analysis to identify any patterns that lead to misclassifications, and then to incorporate that knowledge into the prediction model for obtaining even higher accuracies.

Acknowledgements

This research effort was funded by the Mid-Atlantic University Transportation Center (MAUTC) and the Connected Vehicle Initiative UTC (CVI-UTC).

References

[1] Bao, L. and S.S. Intille, *Activity recognition from user-annotated acceleration data*, in *Pervasive Computing, Proceedings*, A. Ferscha and F. Mattern, Editors. 2004. p. 1-17.

[2] Kwapisz, J.R., G.M. Weiss, and S.A. Moore, *Activity recognition using cell phone accelerometers.* SIGKDD Explor. Newsl., 2011. **12**(2): p. 74-82.

[3] Susi, M., V. Renaudin, and G. Lachapelle, *Motion Mode Recognition and Step Detection Algorithms for Mobile Phone Users.* Sensors, 2013. **13**(2): p. 1539-62.

[4] Stenneth, L., et al. *Transportation mode detection using mobile phones and GIS information*. in *19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM SIGSPATIAL GIS 2011, November 1, 2011 - November 4, 2011*. 2011. Chicago, IL, United states: Association for Computing Machinery.

[5] Yu, X., et al. *Transportation activity analysis using smartphones*. in *Consumer Communications and Networking Conference (CCNC), 2012 IEEE*. 2012.

[6] Widhalm, P., P. Nitsche, and N. Brandie. *Transport mode detection with realistic Smartphone sensor data*. in *2012 21st International Conference on Pattern Recognition (ICPR 2012), 11-15 Nov. 2012*. 2012. Piscataway, NJ, USA: IEEE.

[7] Manzoni, V., et al., *Transportation mode identification and real-time CO2 emission estimation using smartphones*. 2010, Technical report, Massachusetts Institute of Technology, Cambridge.

[8] Reddy, S., et al., *Using Mobile Phones to Determine Transportation Modes.* Acm Transactions on Sensor Networks, 2010. **6**(2).

[9] Biljecki, F., H. Ledoux, and P. van Oosterom, *Transportation mode-based segmentation and classification of movement trajectories.* International Journal of Geographical Information Science, 2013. **27**(2): p. 385-407.

[10] Zheng, Y., et al., *Learning transportation mode from raw gps data for geographic applications on the web*, in *Proceedings of the 17th international conference on World Wide Web*. 2008, ACM: Beijing, China. p. 247-256.

[11] Gonzalez, P.A., et al., *Automating mode detection for travel behaviour analysis by using global positioning systems-enabled mobile phones and neural networks.* Iet Intelligent Transport Systems, 2010. **4**(1): p. 37-49.

[12] Byon, Y.J., B. Abdulhai, and A. Shalaby, *Real-Time Transportation Mode Detection via Tracking Global Positioning System Mobile Devices.* Journal of Intelligent Transportation Systems, 2009. **13**(4): p. 161-170.

[13] Zhang, L., M. Qiang, and G. Yang, *Mobility transportation mode detection based on trajectory segment.* Journal of Computational Information Systems, 2013. **9**(8): p. 3279-3286.

[14] Nham, B., K. Siangliulue, and S. Yeung, *Predicting mode of transport from iphone accelerometer data*. 2008, Tech. report, Stanford Univ.

[15] Nick, T., et al. *Classifying means of transportation using mobile sensor data*. in *Neural Networks (IJCNN), The 2010 International Joint Conference on*. 2010. IEEE.

[16] Bolbol, A., et al., *Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification.* Computers, Environment and Urban Systems, 2012.

[17] Zheng, Y., et al., *Understanding transportation modes based on GPS data for web applications.* ACM Transactions on the Web (TWEB), 2010. **4**(1): p. 1.

[18] Nitsche, P., et al., *A strategy on how to utilize smartphones for automatically reconstructing trips in travel surveys.* Procedia-Social and Behavioral Sciences, 2012. **48**: p. 1033-1046.

[19] Gong, H., et al., *A GPS/GIS method for travel mode detection in New York City.* Computers, Environment and Urban Systems, 2012. **36**(2): p. 131-139.

[20] Lester, J., et al., *MobileSense-Sensing modes of transportation in studies of the built environment.* UrbanSense08, 2008: p. 46-50.

[21] Keerthi, S.S. and C.-J. Lin, *Asymptotic behaviors of support vector machines with Gaussian kernel.* Neural computation, 2003. **15**(7): p. 1667-1689.

[22] Troped, P.J., et al., *Prediction of activity mode with global positioning system and accelerometer data.* Medicine and Science in Sports and Exercise, 2008. **40**(5): p. 972-978.

[23] Chang, C.-C. and C.-J. Lin, *LIBSVM: a library for support vector machines.* ACM Transactions on Intelligent Systems and Technology (TIST), 2011. **2**(3): p. 27.

[24] Hsu, C.-W. and C.-J. Lin, *A comparison of methods for multiclass support vector machines.* Neural Networks, IEEE Transactions on, 2002. **13**(2): p. 415-425.

[25] Hsu, C.-W., C.-C. Chang, and C.-J. Lin, *A practical guide to support vector classification*. 2003.

Machine Learning Transportation Mode Recognition using Mobile Phone Sensor Data

Arash Jahangiri and Hesham A. Rakha, *Member, IEEE*

Abstract

The paper adopts different supervised learning methods from the field of machine learning to develop multi-class classifiers to distinguish between different transportation modes including driving a car, riding a bicycle, taking a bus, walking, and running. Methods that were used include K-Nearest Neighbor (KNN), Support Vector Machines (SVMs), and tree-based models that comprise a single Decision Tree (DT), Bagging (Bag), and Random Forest (RF) methods. For training and validating purposes, data were obtained from smartphone sensors including the accelerometer, gyroscope, and rotation vector sensors. K-fold Cross-Validation as well as Out-of-Bag error was used for model selection and validation. Several features were created from which a subset was identified through minimum Redundancy Maximum Relevance (mRMR) method as the most representative features. Data obtained from the smartphone sensors were found to have important information to distinguish between transportation modes. The performance of different methods were evaluated and compared to each other. The Random Forest (RF) and Support Vector Machine (SVM) methods were found to perform the best. Feature importance of different features was determined for the Random Forest model.

***Index Terms*—Cellular phone sensor data, machine learning algorithms, transportation mode recognition.**

Introduction

DISTINGUISHING between different types of physical activities using sensor data has been a recent research topic that has received considerable attention [1, 2]. Transportation mode detection can be considered as an activity recognition task in which data from smartphone sensors carried by users are utilized to infer what transportation mode the individuals have used. Micro-electromechanical systems (MEMS), such as accelerometers and gyroscopes are embedded in most smartphone devices [3] from which the data can be obtained at high frequencies. Smartphones, nowadays, are equipped with powerful sensors such as GPS, accelerometer, gyroscope, light sensors, etc. Having such powerful sensors all embedded in a small device carried in everyday life activities has enabled researchers to investigate new research areas. The advantages of these smart devices include ubiquity, ability to send and receive data through various ways (e.g. Wi-Fi/cellular network/Bluetooth), and storing/processing data [4].

The knowledge of individuals' mode of transport can facilitate some tasks and also can be adopted in several applications as follows:

1) Knowing the mode of transportation is an essential part of urban transportation planning, which is usually investigated through questionnaires/travel diaries/ telephone interviews [4, 5]. This traditional way of surveying is usually expensive, erroneous, limited to a specific area, and not so up-to-date [6].
2) As environmental applications, the carbon footprint as well as the amount of calories burnt of individuals can be determined by obtaining the mode of transport. Also, physical activities and health can be monitored, the hazard exposure can be tracked, and the environmental impacts of one's activities can be assessed [7].
3) Other applications include providing users with real-time information using the knowledge of speed and transport mode from the users as probes [4, 8], providing individuals with customized advertisements and messages based on the transportation mode they are using [4].

Many studies have used Global Positioning System (GPS) data for classification purposes. However, several limitations are associated with the use of GPS sensors. These limitations include: GPS information is not available in shielded areas (e.g. tunnels) and the GPS signals may be lost especially in high dense locations, which results in inaccurate position information. Moreover, the GPS sensor consumes significant power that sometimes users turn it off to save the battery [6, 8]. This paper focuses on developing detection models using machine learning techniques and data obtained from smartphone sensors including accelerometer, gyroscope, and rotation vector, without GPS data. Consideration of multiple

Manuscript submitted May 11, 2014. This work was supported in part by the Connected Vehicle Initiative University Transportation Center (CVI-UTC), the Mid-Atlantic University Transportation Center (MAUTC), and the TranLIVE University Transportation Center.

A. Jahangiri and H. A. Rakha are with the Center for Sustainable Mobility, Virginia Tech Transportation Institute, 3500 Transportation Research Plaza, Blacksburg, VA 24061 USA (e-mail: arashj@vt.edu; hrakha@vtti.vt.edu).

sensors is beneficial in that even without using GPS, the transportation modes can be identified.

The present research demonstrates how to apply several machine learning techniques, including: K-Nearest Neighbor (KNN), Support Vector Machines (SVMs), and tree-based models that comprise a single Decision Tree (DT), Bagging (Bag), and Random Forest (RF) methods to identify transportation modes using data obtained from smartphone sensors. The data include acceleration extracted from the accelerometer sensor, rate of device rotation extracted from the gyroscope sensor, and device's orientation extracted from the rotation vector sensor; all these were extracted around different coordinate axes; more details on feature selection are presented in sections "data collection, preprocessing, and feature extraction" on page 4 and "feature selection" on page 6. Previous studies lacked adequate simultaneous consideration of several factors, whereas this study is unique in that it comprehensively and simultaneously considers all these factors to obtain a naturalistic data which better reflects real world situations. To the best of our knowledge, items 4, 5, 6, and 8 (listed below) have not been considered in the past literature. To summarize, these factors include:

1) The research considered both motorized (car and bus) and non-motorized modes of travel (bike, walk and run).
2) The research did not require that the travelers maintain a fixed location for their phone as was done in other studies (e.g. phone must be in the traveler's pocket).
3) The research did not use data from GPS sensors because GPS sensors can deplete the phone battery and the signal may be lost in urban areas.
4) The research made use of data from gyroscope and rotation vector sensors, which was never used in previous transportation mode detection studies. Some features were created based on data from these two sensors.
5) The research considered motorized travel (car and bus) on different road types with different speed limits (e.g. 15, 25, 35, 45, and 65 mph speed limits). This wide range of speeds was selected to ensure that the algorithms developed would be robust to different travel conditions.
6) The data collection required travelers to collect bus, car, and bike data along routes where they had to stop at different intersections and thus data included data in traffic jam conditions.
7) The research considered all common machine learning procedures in the development of the models, namely; complete model selection, regularization (applied when using SVM), feature selection, and feature scaling (applied when using SVM and KNN).
8) The research created a large number of features from which the most representative features were selected for model development. More details about the examined features are presented in sections "data collection, preprocessing, and feature extraction" on page 4 and "feature selection" on page 6.
9) The research identified the features based on statistical measures of dispersion as well as derivatives to obtain variations over the time window of interest and consequently incorporated this knowledge (i.e. feature time dependency) into the models.

Some of the study challenges included: (1) Data synchronization; since it was not possible to store the data from different sensors at specific times and thus data were resampled at a desired frequency, (2) High frequency of data; due to having data at very high frequencies. Many data points were available even in small time window sizes. Consequently, using statistical measures of dispersion, raw data values were replaced with those measures corresponding to the time window of interest, (3) High computations; optimizing the parameters of different models in the model selection task, required high computations and thus to alleviate this problem, statistical measures of dispersion were replaced with all the data points within the time window of interest to decrease the number of data points. In addition, the mRMR feature selection method was employed to select the most representative features before developing different models. A characteristic of this method was that it was independent of the models and it was not required to carry out the feature selection task for each model; hence it was conducted only once, and (4) Data noise; a preprocessing task was conducted for noise reduction.

The remainder of the paper is organized in the following sections: Relevant literature is reviewed in the next section followed by the data collection section. Subsequently, it is shown how the detection models were developed. Subsequently, the results of the study are presented and finally the conclusions of the study are presented.

Relevant work

Almost all studies used data from GPS sensors that have the aforementioned drawbacks. Also, they took advantage of Artificial Intelligence (AI) tools such as Fuzzy Expert Systems as in [9], Decision Trees as in [4-8, 10], Bayesian Networks as in [4, 10], Random Forests as in [4], Naïve Bayesian techniques as in [4, 7], Neural Networks as in [11, 12], and Support Vector Machine (SVM) techniques as in [7, 10, 13-16], of which the Decision Tree and SVM methods were used the most. To improve the model performance, other techniques were also combined with machine learning methods such as Discrete Hidden Markov Models as in [7] and Bootstrap aggregating as in [17]. Other than AI tools, statistical methods were also applied such as the Random Subspace Method in [18]. Some studies have used additional information from Geographic Information System (GIS) maps as in [4, 9, 19, 20]. However, GIS data is not always available, and also this approach may not be suitable for real-time applications because it mostly relies on the knowledge of the entire trip with respect to the GIS features such as bus stops, subway entrances, and rail lines.

The Decision Tree method was identified as the best method in [7, 10] compared to other methods including SVM. However, in developing the models, several factors need to be

considered to obtain the best possible model performance. It appears that similar studies lack at least one of the following:

- Conducting a complete model selection
- Considering regularization
- Using feature selection methods
- Considering feature scaling

A complete model selection is equivalent to incorporating all the tuning parameters in order to obtain the best detection accuracy. Regularization is included in the model to deal with the issue of over-fitting (high variance). Feature selection methods are adopted to use the most representative features. Feature scaling is applied to normalize the range of different features (or attributes), which leads to higher model performance and training speed. However, it should be noted that in general not all these factors are always required. For example, selecting features based on intuition or expert knowledge may lead to as good results as using a feature selection method. It is also possible that these factors specially feature scaling (since it is a simple procedure) were part of the software package that was used in their work, but the authors were not clear whether the factors were applied or they just did not emphasize or focus on the importance of these factors. Nevertheless, these are important factors to be considered when solving machine learning problems.

Depending on the application of interest, different time window sizes have been used for detecting the mode of travel. For example, [12] found that longer monitoring durations lead to higher accuracy. Intuitively, the bigger the time window size the easier the detection becomes since with bigger window sizes more information is available. If the application is a survey for demand analysis the time window size can be as large as trip duration, whereas if the application provides real-time information for environmental or some transit applications, then smaller time window sizes are more desirable. The size should be as small as possible for some safety applications (e.g. crash prevention). An earlier study [13] used 200-meter and 150-second segments in their experiment. Whereas another study [6] used 10-second time windows to separate walking from non-walking segments and then applied a maximum size of 2 minutes. Other than the time window size, the overlaps of two consecutive windows have also been considered. Reference [8] obtained the best time window size and overlap to be 10.24 seconds and 50%, respectively. The entire trip duration appears to be considered in [5, 9, 11, 16, 20]. Higher accuracies are achieved by increasing the time window size as shown in [12]. However, the focus of this study is on small time window sizes, so the developed models have the potential to be used in a broader range of applications such as environmental and safety applications.

Other than the time window size, several factors that also influence the model performance are as follows:

1) *Number of classes:* as the number of classes increases, class differentiation becomes more difficult.
2) *Model dependency on data sources:* Less dependent models are more desirable as they can be applicable even with limited sources of data. In this case, sensors such as accelerometers and gyroscopes are more reliable since their data are always available. Whereas, GPS, as mentioned earlier, has its own drawbacks.
3) *Ability to distinguish between motorized classes:* as different motorized classes have similar characteristics such as speed and acceleration, a model capable of differentiating between these modes is of great value. For example, distinguishing the bus mode from the car mode is significantly more difficult than discriminating walking from driving.
4) *Sensor positioning:* it shows how realistic the experiments are conducted. Positioning the devices in certain locations increases the detection accuracy because the movements monitored by the sensors show the movements of the transportation mode (or the person) they are attached to. However, it may not reflect realistic behavior of the travelers. Some of the studies required that the participants attach sensors/smartphones to different parts of their body.

Different detection accuracies have been reported by different studies. Although in almost all of them including our previous work, comparisons were drawn between the accuracies obtained from their approaches with those of others, such comparisons were excluded in the present study because of two reasons. First, in different studies, models were developed on different data sets. Second, several factors can affect model performance (e.g. time window size, number of classes, etc.). Here are some examples: Excluding those studies that assumed the time window size to be the entire trip, the highest reported accuracy of 96.9% was achieved by [15] with a time window size of 4 seconds. In their approach, they only used accelerometer data and did not rely on GPS and GIS data. Their method was capable of differentiating between motorized modes (car and train) and no specific sensor positioning was applied. Nevertheless, they only considered three classes. The second best accuracy was obtained by [14]. They also used accelerometer data without relying on GPS/GIS data. However, although different motorized modes were mentioned in the paper, it seems that the reported accuracies show only one motorized mode. Also, subjects in their study were asked to keep their device in their pocket of the non-dominant hip while collecting data which is more realistic compared to attaching sensors to the body, but still does not reflect a complete realistic behavior. An accuracy of 93.6% was reported by [7]. They applied the lowest time window size throughout the literature which is one second. However, their approach was dependent on data from GPS sensors. Further, different motorized classes were not considered.

Data collection, preprocessing and feature extraction

A smartphone application was developed for the purpose of data collection. The application stores the data coming from smartphone sensors including GPS, Accelerometer, Gyroscope, and Rotation Vector at the highest possible

frequency. To collect the data, ten employees at Virginia Tech Transportation Institute (VTTI) were asked to carry a smartphone (two devices were used: a Galaxy Nexus and a Nexus 4) with the application installed on it on multiple trips. They were asked to select the travel mode they intend to use before starting the logging process, and then using the application buttons they were able to start and stop data logging. Although smartphones can be carried in other places, to make sure the data collection is less dependent on the sensor positioning, the travelers were asked to carry the smartphone in different positions that they normally do such as in pocket, in palm, in backpack, and different places inside car (e.g. on front right seat, coffee holder alongside of the driver) as they reported after the data collection. However, the amount of time that was spent for different positions were unknown since the participants were not asked to collect data in a particular position for a certain amount of time and the reason was to make the data collection as natural as possible. Data collection was conducted on different workdays (Mon through Fri) during working hours (8 AM to 6 PM) on different road types with different speed limits (i.e. car mode on roads with 15, 25, 35, 45, and 65 mph; bus mode on roads with 15, 25, 35, and 45 mph; bike mode on roads with 15, 25, and 35 mph) in Blacksburg, Virginia. Thirty minutes worth of data for each mode per person were collected. The original data frequency was about 25 Hz (for accelerometer, gyroscope, and rotation vector sensors), but the data from different sensors were not synchronized. Thus, in order to ensure that the data were gathered at identical sampling rates, linear interpolation was first applied to the data similar to [8] to produce continuous data sets and then the data were re-sampled at the desired rate (rate of 100 Hz was applied). Since the original frequency of 25 Hz was not a constant rate (i.e. a constant frequency was not possible to set for collecting data), the choice of 100 was made to make sure no information is lost. Furthermore, a low pass filter was used for noise reduction. In total, 25 hours of data (30 minutes per mode per person) were stored and used for training and testing purposes. In other words, total of ten travelers collected 30 minutes of data for each mode that equals (30x10x5)/60 = 25 hours.

Some features are considered to be basic/traditional features (e.g. mean speed, mean acceleration), which are more intuitive to be influential and were widely used in the literature and some are considered to be more advanced features (e.g. heading change rate) as presented by [13]. In our previous work [21], we used approximately 60 features that we created from the sensor data, mostly based on some statistical measures of dispersion. In the present study, we created a set of features that include those 60 features with some modifications. First, a feature should have a meaningful relationship to the transportation modes. Therefore, absolute values of the rotation vector sensor are excluded from the feature set because the absolute values correspond to the device's orientation and are unrelated to the transportation modes. Second, since a time window is being monitored, other features that can describe variations in time were created to incorporate the features' time dependency (e.g. based on derivatives). Also, spectral entropy was added, which can be used as a measure to show the peaky spots of a distribution [22]. Peaky spots are important since this measure can be different for different transportation modes. Intuitively, an abrupt braking (which in reflected in the accelerometer data) in a car mode is peakier than in the bike mode. A High value of spectral entropy for a distribution shows that the distribution is somewhat flat. Conversely, the spectral entropy decreases when the distribution becomes less flat [23]. In addition, the data from the sensors were treated as signals, consequently, the energy of the signal within the time window of interest was added to the feature set [24]. Also, the data from the GPS were excluded to only focus on the scenario where no GPS data are available. To summarize, using the data from different sensors and for each time window, Table 9 shows the measures that were used to create the feature set. Other than the "spectral entropy" and "energy" that was mentioned above, other measure include: mean (or average), max (maximum), min (minimum), var (variance), std (standard deviation), range, iqr (interquartile range), and signChange (number of times the sign of a feature changes over the time window). Also in this table, x_i^t represents the data array for the ith feature (e.g. acceleration) from the time window t. Also, $\dot{x}_i^t$ represent the derivative of x_i^t. A total of 165 features were created: out of the 18 measures presented in Table 9, all the 18 measures were applied to each of the sensor values; 7 measures were applied to rotation vector sensor values; 16 measures were applied to the summation values from accelerometer and gyroscope sensors (e.g. $m_{acceleration} = \sqrt{a_x{}^2 + a_y{}^2 + a_z{}^2}$); 4 measures were applied to the summation values from rotation vector sensor. As a result, the total number of features reached 18*6+7*3+16*2+4*1 = 165 features.

Table 9 MEASURES USED TO CREATE FEATURES

No.	Measure	No.	Measure
1	$mean(x_i^t)$	10	$spectralEntropy(x_i^t)$
2	$max(x_i^t)$	11	$mean(\dot{x}_i^t)$
3	$min(x_i^t)$	12	$max(\dot{x}_i^t)$
4	$var(x_i^t)$	13	$min(\dot{x}_i^t)$
5	$std\ (x_i^t)$	14	$var(\dot{x}_i^t)$
6	$range(x_i^t)$	15	$std\ (\dot{x}_i^t)$
7	$iqr(x_i^t)$	16	$range(\dot{x}_i^t)$
8	$signChange(x_i^t)$	17	$iqr(\dot{x}_i^t)$
9	$energy(x_i^t)$	18	$signChange(\dot{x}_i^t)$

Model development

Three methods were considered to construct the detection models. Support Vector Machine (SVM) and Decision Tree have been used in most of the literature to classify transportation modes, and some papers found the Decision Tree to be the best method. Consequently, tree-based models

(single Decision Tree, Bagging, and Random Forest) and SVM were selected for model construction. In addition, the K-Nearest Neighbor (KNN) method was considered for the purpose of comparison given that it is a simple technique. Several methods were adopted in the model development process; maximum dependency minimum redundancy (mRMR) for feature selection, K-fold Cross- Validation for model selection, and Scaling for normalization. To conduct feature scaling, the feature values were normalized to be within the range of [-1, 1] (Scaling was conducted only when applying SVM and KNN).

K-Nearest Neighbor (KNN)

A simple yet effective method, namely the K-Nearest Neighbor (KNN), which has been applied to numerous classification and regression problems in different fields, was adopted to identify transportation modes. For each test observation (X_j^{test}) that includes different features (such as x_i^t), this method first identifies the K nearest train observations (X_j^{train}) in the training data set to the test observation and stores them in the N_K set. Taking the majority vote of the classes for the K nearest points identifies the class of the test observation. Calculating the average in the Equation 6 is equivalent to taking the majority vote in the case of classification (versus regression). y_j^{train} and y_j^{test} are the response (or target) values corresponding to the observations X_j^{train} and X_j^{test}, respectively. K is a tuning parameter that needs to be determined [25].

$$y_j^{test} = \frac{1}{K} \sum_{X_j^{train} \in N_K} y_j^{train} \qquad \mathbf{6}$$

Support Vector Machines (SVMs)

SVM is known as a large margin classifier, which means when classifying data, it determines the best possible decision boundary that provides the largest possible gap between classes. This characteristic contributes to a higher confidence in solving classification problems. To construct the SVM model, the following factors are taken into account: using a Gaussian kernel with complete model selection (which entails consideration of the regularization parameter and the Gaussian parameter), and applying feature scaling.

Equation 7 presents the SVM formulation to solve the classification problem and the associated constraints are shown in Equations 8 and 9 [26]. The objective function is composed of two terms: minimizing the first term is basically equivalent to maximizing the margin between classes, and the second term consists of an error term multiplied by the regularization (penalty) parameter denoted by C. The C parameter should be determined to provide the relative importance between the two terms. Equation 8 ensures that margin of at least 1 exists with consideration of some violations. The value of 1 was resulted from normalizing w. Equation 9 restricts the data points to the points that have positive errors.

$$\min_{w,b,\xi} \left(\frac{1}{2} w^T w + C \sum_{n=1}^{N} \xi_n \right) \qquad \mathbf{7}$$

Subject to:

$$y_n (w^T \phi(x_n) + b) \geq 1 - \xi_n \,, n = 1, \dots, N \qquad 8$$

$$\xi_n \geq 0 \,, n = 1, \dots, N \qquad 9$$

Where,

w	Parameters to define decision boundary between classes
C	Regularization (or penalty) parameter
ξ_n	Error parameter to denote margin violation
b	Intercept associated with decision boundaries
$\phi(x_n)$	Function to transform data from X space into some Z space
y_n	Target value for the nth observation

SVM applies the function $\phi(.)$ to transform data from the current n-dimensional X space into a higher dimensional Z space in which the decision boundaries between classes are easier to identify. This transformation could be computationally very expensive; consequently, to solve the problem, the SVM only needs to obtain vector inner products in the space of interest. Hence, SVM takes advantage of some functions known as Kernels that return the vector inner product in the desired Z space. Different types of kernels exist such as linear kernel, polynomial kernels, and Gaussian kernel. Linear kernel, as applied in [7, 10], is the basic mode which means no kernels are actually taken into account. In other words, vector inner product as appears in the dual formulation of the problem is considered without transforming data into another space. For a certain type of problems, SVM can produce better results with more advanced kernels such as Gaussian kernel. According to our data size and number of features, Gaussian kernel was believed to be the most appropriate kernel [27]. Also, if a complete model selection is carried out, there is no need to test the linear kernel because the results obtained from the Gaussian kernel include the results obtained from the linear kernel [28]. In this paper, the $\phi(x_n)$ function corresponds to the Gaussian kernel. The formulation of the Gaussian kernel is shown in Equation 10. When using this type of kernel, the tuning parameters are the Gaussian parameter (σ) and the regularization parameter (C) that should be determined to obtain the best possible detection performance.

$$K(x, x') = exp\left(-\frac{\|x - x'\|^2}{2\sigma^2} \right) \qquad 10$$

Where,

x, x'	n-dimensional vectors
$\|x - x'\|$	Euclidean distance between vectors x, x'
σ	Gaussian parameter

Two approaches were examined: (1) Developing a single

SVM model using the entire dataset, (2) Developing an ensemble of SVM models using a smaller data set for each model. Similar to the idea behind the Bag approach, instead of developing a single SVM model, a series of SVM models was developed and the final result was determined based on the majority vote obtained from the SVM models. A number of studies have considered taking advantage of an ensemble of SVMs [29, 30]. As the number of observations in the training dataset (n) increases the training time increases with the power of two (n^2) [31]. Thus, if the data set is sufficiently large developing a number of SVM models using a subset of data can be faster than developing a single SVM model using the entire data.

Tree based models

1) Decision Tree

Decision Trees were introduced for classification and regression problems in the mid-80s [32]. These approaches have several advantages; among all, they are easy to explain and interpret, they reflect the human decision making process, they can be graphically displayed, and there is no need to create dummy variables for qualitative predictors. However, as the tree becomes larger, it may over-fit the data and show poor performance on the test data set. Consequently, some strategies are used in the R and CART software to construct a large tree using recursive binary splitting and then pruning back to obtain a good sub-tree. This approach is known as Cost Complexity Pruning or Weakest Link Pruning. In the Recursive binary splitting method, a root node is the starting point where a predictor (feature) needs to be selected with a cut point to split the data into two parts or nodes. This procedure of selecting a feature and splitting is carried out successively to grow the tree. Different criteria can be used to choose the best split at each node, including: classification error rate, Gini index, and Cross-Entropy. In practice the two latter methods result in better performance. Consequently, Cross-Entropy was used in this study. Having K classes, at each node m which receives N^m observations (x_i^m, y_i^m) from its parent node, Cross-Entropy can be obtained through Equation 11 [33].

$$\sum_{k=1}^{K} P_k^m \log P_k^m \qquad 11$$

where,

P_k^m	Proportion of class k observations in node m
$P_k^m =$	$\frac{1}{N^m}\sum_{x_i^m} I(y_i^m = k)$
y_i^m	Target value of nth observation in node m
$I(y_i^m = k)$	1 if $y_i^m = k$, and 0 otherwise

2) Bagging

Bagging or Bootstrap aggregating method, introduced in 1996 [34], takes advantage of aggregating results from different models to reduce the variance. The detection/prediction results of different models constructed on different training sets can be averaged. However, in practice, we usually have only one training set. Instead, bootstrapped training data (Pseudo training sets) can be obtained by taking repeated samples from a single training set [35] and a tree model can be constructed for each. Afterwards, the average performance of all models represents the overall performance, which is called Bagging or Bootstrap aggregation. There is no need for pruning of trees as the variance is reduced by averaging. Averaging is equivalent to taking a majority vote for classification problems, which is the case in the present study. The detection/prediction for a single data point x is obtained by averaging (taking a majority vote) the detections resulted from all bootstrapped samples as shown in Equation 12. The trees can be as large as possible, thus the only parameter to be determined is the number of trees.

$$\hat{\boldsymbol{y}}_{bag}(\boldsymbol{x}) = \frac{\mathbf{1}}{\boldsymbol{B}}\sum_{b=1}^{B} \hat{\boldsymbol{y}}_b(\boldsymbol{x}) \qquad 12$$

where,

$\hat{\boldsymbol{y}}_{bag}(\boldsymbol{x})$	Target value resulted from averaging
$\hat{\boldsymbol{y}}_b(\boldsymbol{x})$	Detected target value for observation $\boldsymbol{x}$ in bootstrap sample $\boldsymbol{b}$
$\boldsymbol{B}$	Total bootstrap samples

3) Random Forest

Similar to the Bagging method, the random forest method, as proposed in 2001 [36], creates an ensemble of trees and the result is obtained based on the majority votes. One issue relating to the Bagging is that the trees can be very similar since all the features are used to construct each tree; consequently, the trees can be highly correlated. To tackle this problem random forest restricts the number of features by randomly selecting a subset of features to grow each tree. The parameters to be determined are the number of features to use and the number of trees. Interestingly, in Random Forest and also Bagging approaches, adding more trees does not lead to over-fitting, but at some point not much benefit is gained by including more trees [33].

Feature Selection

Feature selection is considered to be a critical task as it can reduce the dimensionality of the problem, reduce the noise, identify more important predictors, and lead to more interpretable features [37].

Some methods have been applied to select the most relevant features to use such as ANOVA tests used in [16], correlation based feature selection (CFS) used in [7], and Chi Squared and Information gain methods applied in [4]. Using mutual information or some statistical tests to select the top-ranked features may not be sufficient as the selected features could be highly correlated among themselves [37, 38]. In other words, not much benefit is gained by combining highly correlated features. In the present study, the selection of the most representative features entailed using the minimum redundancy maximum relevance (mRMR) approach. This approach was used to deal with this issue; when selecting the

x_i feature from the feature set M, assuming set F has the already selected features, the goal is to simultaneously maximize the relevance between the feature and the target class (i.e. x_i, c) as shown in Equation 13 and to minimize the redundancy between that feature and the already selected features (i.e. x_i, x_j) as shown in Equation 14 [37]. Hence, using mRMR all the features that were created were ranked to choose the most useful ones; the top 80 features were selected out of the entire 165 features. The number 80 was chosen by experimenting different values. In other words, it was desired to achieve a good level of detection accuracy and at the same time to exclude less useful features.

$$\max_{x_i \in (M-F)} MI(x_i, c) \quad 13$$

$$\min_{x_i \in (M-F)} \frac{1}{|F|} \sum_{x_j \in F} MI(x_i, x_j) \quad 14$$

where,

$MI(x, y)$	Mutual Information of x and y
x_i	The feature to be examined
x_j	A previously selected feature
M	Set of all features
F	Set of the selected features
c	Target class

K-fold Cross-Validation

The K-fold cross-validation is a powerful technique to estimate the detection/prediction error. Consequently, it is used to select the best model and to determine the model parameters. The idea is to randomly divide the data of n observations into K approximately equal parts or folds ($F_1, F_2, \dots, F_K$) with n_k observations in each fold. Subsequently, the data of the first fold are set aside as the validation data set, and the data of the remaining $K-1$ folds are used as the training data set to construct a model. The same procedure is conducted K times, each time a different validation data set is chosen. The performance of each model is evaluated on the corresponding validation set and the average detection error is obtained over the K models as shown in Equation 15. The special case is when the number of folds is exactly the same as the number of observations; this is called Leave-One-Out Cross-Validation (LOOCV). LOOCV requires high computations as it needs to construct the model n times which in this case is the number of observations. Also, since only one observation is left out at each k stage, the training sets are almost the same for each model and thus the estimates are highly correlated, consequently, the average over K folds can have high variance. In practice, the best choice for the number of folds is 5 or 10 [39, 40].

$$CV_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} Err_k \quad 15$$

Where,

$$Err_k = \sum_{i \in F_k} I(y_i \neq \hat{y}_i)/n_k$$

n	Number of observations
n_k	Number of observations in k^{th} fold
F_k	Set of observations in k^{th} fold
y_i	Actual target value
$\hat{y}_i$	Detected target value
$I(y_i \neq \hat{y}_i)$	1 if $y_i \neq \hat{y}_i$, and 0 otherwise

Results

Using mRMR, 80 features were selected as the most relevant features, which were used to construct the models. The performance of each model was quantified using different metrics depending on the model, namely: misclassification error obtained from 5-fold Cross-Validation and Out-Of-Bag error. Cross-Validation, as mentioned earlier, is a good technique to estimate the detection/prediction errors. Out-Of-Bag error is an accurate estimate of the errors suitable for Bagging and Random Forest that is almost identical to the Cross-Validation accuracy [33]. Moreover, in developing different models, 30% of the data were set aside to obtain a test error, and the remaining 70% were used as the training set for model development. Subsequently, Confusion matrices were obtained for each model, which shows the classification rates, misclassification rates, recall, and precision values. The recall measure is calculated by dividing the total number of true positives by the total number of actual positives. The Precision measure is computed by dividing the total number of true positives by the total number of predicted positives. Finally the models were compared to each other using different performance measures such as F-Score, Youden's index, and the discriminant power that will be presented in the "model comparison" section.

KNN Model

The KNN model using 5-fold Cross-Validation was implemented using the R software [41] and the class package [42]. The only tuning parameter in the KNN method is the number of K neighbors. Figure 7 shows the misclassification error obtained from a 5-fold Cross-Validation (applied in the training set) of 25 runs for different numbers of neighbors. The highest accuracy was achieved when K was 7 resulting in an error rate of 8.8% (accuracy of 91.2%). The confusion matrix of the KNN model including the Recall and Precision values is shown in Table 10.

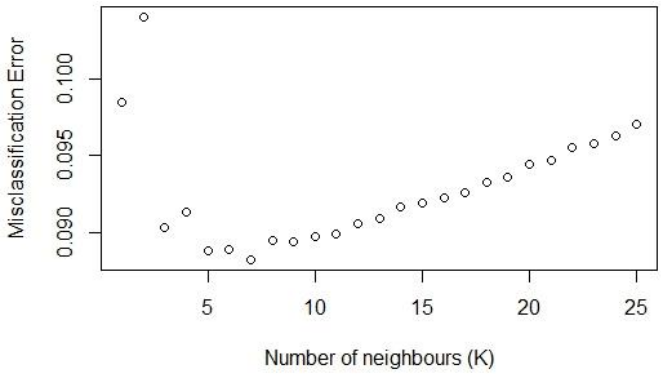


Figure 7 Impact of number of neighbors on model misclassification error

Table 10 CONFUSION MATRIX - KNN MODEL

KNN		Actual					
		Bike	Car	Walk	Run	Bus	**Precision**
Predicted	Bike	93.31	2.33	2.30	0.99	2.96	91.72
	Car	2.73	84.07	0.77	0.38	11.74	84.32
	Walk	2.37	0.20	96.51	1.50	0.25	95.72
	Run	0.07	0.00	0.15	97.00	0.05	99.71
	Bus	1.51	13.40	0.27	0.13	85.01	84.69
	Recall	93.31	84.07	96.51	97.00	85.01	

SVM Model

In implementing the SVM, the LibSVM library of SVMs was used. For multiclass classification, considering h classes, LibSVM applies one-against-one method in which $h(h-1)/2$ binary models are built. Among these, LibSVM chooses the parameters that achieve the highest overall performance. Another well-known method is called one-against-all which is more intuitive and has similar performance. However, LibSVM takes advantage of one-against-one because of its shorter training time. [43]. Furthermore, 5-fold Cross-Validation was used for model development and assessment.

Single SVM

The 5-fold Cross-Validation was applied on the training set to develop a single SVM model. In order to conduct a complete model selection, the regularization parameter (c) as well as the Gaussian parameter (σ) should be optimized. The Gaussian kernel formulation used in libSVM [43] is slightly different from Equation 10; in their formulation, the parameter gamma was used instead of $\frac{1}{2\sigma^2}$. Figure 8 presents a contour plot that illustrates how different values of the regularization (c) and the Gaussian (gamma) parameters impact the performance of the SVM model. The optimal values for (gamma, c) were found to be (2.828, 512) that led to the overall accuracy of 94.62%. Parameters c and σ (or gamma) deals with the issues of over-fitting and under-fitting which is a bias-variance tradeoff. Detailed information regarding the bias-variance tradeoff can be found in [29]. Table 11 presents the confusion matrix for the SVM model.

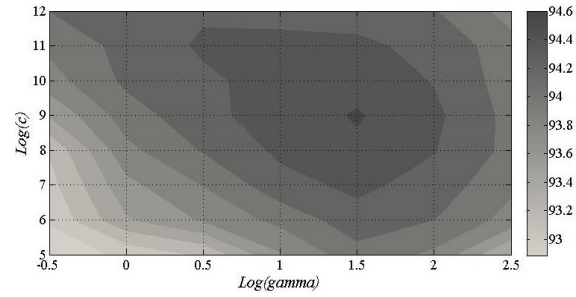


Figure 8 Impacts of regularization and Gaussian parameters on model Accuracy

Table 11 CONFUSION MATRIX - SVM MODEL

SVM		Actual					
		Bike	Car	Walk	Run	Bus	**Precision**
Predicted	Bike	95.11	0.89	1.69	0.45	0.94	96.28
	Car	0.96	93.58	0.18	0.15	6.90	91.74
	Walk	1.89	0.28	97.11	1.34	0.55	95.79
	Run	0.37	0.15	0.77	97.55	0.17	98.50
	Bus	1.66	5.10	0.26	0.50	91.44	92.42
	Recall	95.11	93.58	97.11	97.55	91.44	

Ensemble of SVMs (E.SVM)

For the ensemble of SVM models, the 5-fold cross validation was adopted in a slightly different fashion: the training set (70% part) was divided into 5 folds; one fold was set aside as the validation set, and about 25% of the remaining 4 folds were used to train the first SVM model. Similarly, 25% of the 4 folds was sampled (bootstrap sample; sampling with replacement) to train the second SVM model. The procedure continues until 200 models were constructed. In addition, trial and error was used to pick model parameters (gamma, c) for each SVM model. The average of these 200 models was validated with the data fold that was set aside. All these steps were carried out 5 times, each time with a different data fold as the validation set. Averaging the results of the five folds represented the cross validation results of the ensemble of SVM models. This method led to an overall accuracy of 94.41%. The confusion matrix corresponding for this approach is shown in Table 12.

Table 12 CONFUSION MATRIX – ENSEMBLE OF SVM MODELS

Ensemble of SVMs		Actual					
		Bike	Car	Walk	Run	Bus	**Precision**
Predicted	Bike	95.63	0.68	1.71	0.56	1.08	96.07
	Car	0.83	91.72	0.13	0.15	7.75	91.04
	Walk	1.60	0.48	97.16	1.27	0.74	95.91
	Run	0.46	0.30	0.75	97.82	0.32	98.12
	Bus	1.48	6.83	0.25	0.20	90.11	91.25
	Recall	95.63	91.72	97.16	97.82	90.11	

Tree-based models

Decision Tree (DT)

The decision tree method was implemented in the R software along with two packages ("tree" and "maptree" packages) for tree analysis [41, 44, 45]. The resultant single

tree was a very large tree with 48 terminal nodes with an overall accuracy of 87.27%. Table 13 shows the confusion matrix of the decision tree model.

Table 13 CONFUSION MATRIX - DECISION TREE MODEL

Decision Tree		Actual					
		Bike	Car	Walk	Run	Bus	**Precision**
Predicted	Bike	85.32	1.78	5.21	0.85	3.07	88.96
	Car	1.39	79.30	0.26	0.14	12.14	85.03
	Walk	8.65	0.10	91.99	2.87	0.13	88.54
	Run	0.40	0.00	1.17	95.30	0.00	98.32
	Bus	4.24	18.83	1.36	0.85	84.65	76.92
	Recall	85.32	79.30	91.99	95.30	84.65	

Since the tree is very large, Cost Complexity Pruning was applied to prune the tree from 48 terminal nodes to 24 without much loss in performance. Table 14 shows the confusion matrix of the pruned tree resulted in 86.3% overall accuracy. This model was called DT.P to abbreviate the model title of the pruned decision tree.

Table 14 confusion matrix - Pruned Decision Tree model

Decision Tree - Pruned		Actual					
		Bike	Car	Walk	Run	Bus	**Precision**
Predicted	Bike	84.37	3.12	5.87	1.38	4.02	85.79
	Car	0.61	78.50	0.33	0.05	12.77	85.07
	Walk	9.80	0.07	90.44	2.85	0.10	87.42
	Run	0.53	0.00	2.30	95.12	0.00	97.04
	Bus	4.69	18.31	1.06	0.59	83.11	77.02
	Recall	84.37	78.50	90.44	95.12	83.11	

Because the tree was big to illustrate, the tree was pruned again to reduce the number of terminal nodes to 9 just for illustration purposes, as illustrated in Figure 9. In this case the accuracy of the model is 82.1%.

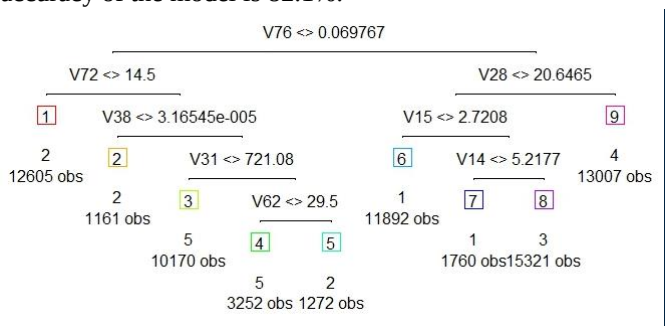


Figure 9 Illustration of a single Decision Tree

Bagging (Bag) and Random Forest (RF) models

In implementing the Bagging and Random Forest methods, the R software and the package "RandomForest" were used, respectively [41, 46]. These two methods were examined together since the Bagging is in fact a special case of a random forest when the number of selected features equals the total number of features. As mentioned before, adding more trees will not cause over-fitting. However, a sufficient number of trees are needed. Figure 10 shows a series of Random Forest models with different number of trees, from 1 to 500, using 5 features for each tree. After approximately 200 trees, no benefit is gained by including more trees. Thus, to apply these approaches, 400 trees were used, which is a sufficiently large number. On the far left of the diagram, when the number of trees is 1, it is equivalent to having a single decision tree.

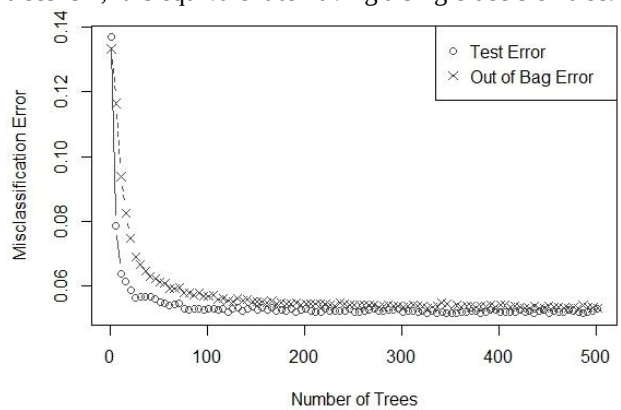


Figure 10 Impact of number of trees on the misclassification error

For the random forest method, other than the number of trees, the number of features needs to be determined as well. Figure 11 shows a series of random forest models with different number of features for each tree. Since a total of 80 features were used (as identified by mRMR), a total of 80 random forest models were constructed to find the best number of features to use. The far right of the figure shows the results of the Bagging approach, where all the 80 features were used. The minimum error rate was obtained with 12-25 features in use. The model with 12 features was selected as the best model since a less complex model with less features is always more disirable. The Confusion matrix for the bagging and the best random forest models are shown in Table 15 and Table 16. The overal accuracy of the best random forest model and the bagging model, obtained from 5-fold Cross-Validation, were 95.1% and 94.4% respectively.

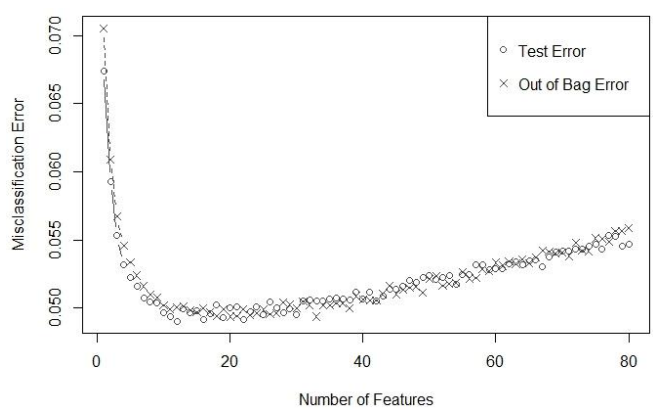


Figure 11 Impact of number of features on the misclassification error

Table 15 CONFUSION MATRIX - RANDOM FOREST

Random Forest	Actual					
	Bike	Car	Walk	Run	Bus	**Precision**

Predicted	Bike	95.47	1.46	2.63	0.97	2.29	93.06
	Car	0.37	93.84	0.12	0.05	4.47	94.93
	Walk	2.93	0.13	96.23	1.59	0.12	95.24
	Run	0.03	0.00	0.40	96.81	0.00	99.55
	Bus	1.19	4.57	0.63	0.58	93.12	93.02
	Recall	95.47	93.84	96.23	96.81	93.12	

Table 16 CONFUSION MATRIX - BAGGING

Bagging		Actual					
		Bike	Car	Walk	Run	Bus	**Precision**
Predicted	Bike	94.63	1.48	2.81	1.00	2.34	92.75
	Car	0.48	92.64	0.12	0.03	5.08	94.18
	Walk	3.43	0.13	95.95	1.63	0.22	94.62
	Run	0.03	0.00	0.58	96.79	0.02	99.34
	Bus	1.42	5.74	0.55	0.54	92.34	91.76
	Recall	94.63	92.64	95.95	96.79	92.34	

Feature Importance

As was mentioned earlier, a total of 80 features were identified as the most relevant features by mRMR method. Figure 12 shows the actual importance of the best 20 features associated to the best random forest model. The importance of the features were assessed based on two measures: (1) Mean Decrease Accuracy that shows how the detection accuracy is decreased if a feature was excluded, averaged over all trees, and normalized by the standard deviation of the differences in accuracy and (2) Mean Decrease Gini that shows how a single feature contributed to decrease the Gini index over all the trees. Table 17 shows the feature names in the order of importance. Since the two measures determine the feature importance in different ways the identified features by the two measures are different. While both measures have been used in the literature, there have been arguments concerning the preference for one measure over another. It is recommended that the first method (i.e. Mean Decrease Accuracy) is more suitable for causal interpretations. More details about the arguments and some contradictions regarding these measures can be found in [47].

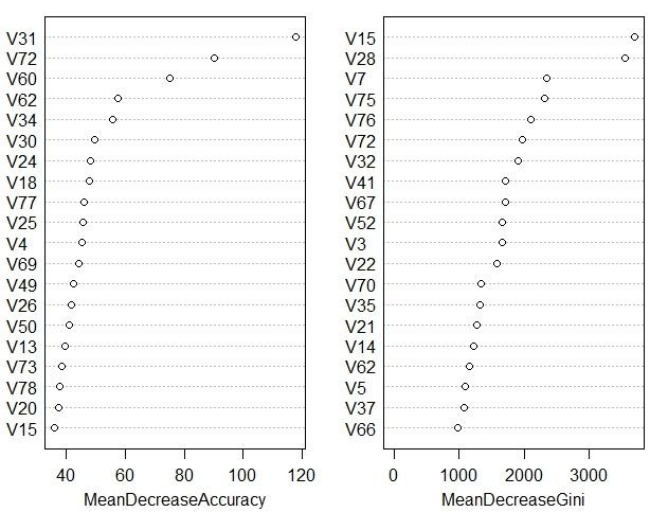


Figure 12 Feature importance for two different measures

Table 17 IMPORTANT FEATURES

No.	Feature Name	No.	Feature Name
1	$spectralEntropy(a_x)$	11	$mean(a_z)$
2	$range(a_y)$	12	$iqr(\dot{a_x})$
3	$max(a_y)$	13	$var(\dot{g_x})$
4	$max(g_y)$	14	$min(\dot{a_y})$
5	$min(g_y)$	15	$range(\dot{a_x})$
6	$range(\dot{g_x})$	16	$energy(a_x)$
7	$spectralEntropy(a_y)$	17	$range(g_x)$
8	$max(a_z)$	18	$mean(g_z)$
9	$mean(\dot{g_x})$	19	$std(\dot{a_y})$
10	$min(\dot{a_z})$	20	$spectralEntropy(g_x)$

Model Comparison

The performance of the models was evaluated using four metrics, namely: the overall accuracy, the F-Score, Youden's index, and the Discriminant Power (DP). The overall accuracy is calculated by dividing the total number of correct detections by the total number of test data. The F-Score is a combined measure of the Recall and the Precision. The Youden's index is a measure to assess the ability of a model to avoid failure. The discriminant power shows how well a model discriminates between different classes by summarizing sensitivity and specificity of the model; the model is a poor discriminant if DP <1, limited if DP <2, fair if DP <3, good – in other cases. The sensitivity and specificity assess model performance on a single class, and are equivalent to the recall. By definition, assuming two classes (positive and negative) sensitivity is exactly the same as the Recall measure. Specificity is also the same metric but for the negative class. Figure 13 illustrates a visual comparison between the models using different performance measures.

Feature Combination

An effort was made to develop a new additional feature which is a combination of other features. This was carried out by combining two approaches; a Meta heuristic approach called Simulated Annealing (SA) [48] and the Random Forest techniques. The new feature was created by multiplying other features; SA was adopted to select the best features to combine. The steps of this approach are as follows:

1. Define an initial solution: two random features were selected and placed in a set called *CF*. These features were combined (by multiplying by each other) to create a new feature. Subsequently, a RF model was developed using the previously used features (i.e. 80 features) and the newly defined feature to obtain the error rate for this initial solution.

2. Choose the algorithm's settings: trial and error was carried out to determine these algorithm parameters.

- Initial temperature ($t_{initial}$); (The term "temperature" is basically a control parameter which affects the probability of accepting or rejecting new solutions.)
- Final temperature (t_{final}) and stopping criteria
- Number of iterations at each temperature (M_k)

- Cooling schedule

After several trials, 5, 0.1, 15, and 0.8 were chosen as the $t_{initial}$, t_{final}, M_k, and the temperature reduction multiplier, respectively.

3. Repeat until stopping criteria are met
   - Repeat until n=M_k, (n is a counter, starting from 0)
     - Generate a new solution: the new solution is generated by either randomly removing an already selected feature in the CF set or randomly adding a new feature to the CF set. Thereafter, the new feature is updated by multiplying all the features in the CF set.
     - Calculate Δ, the relative difference between the new and current error rates
     - If Δ≤0, the new solution is accepted, otherwise, the new solution can still be accepted with the probability of $e^{-(\Delta/\text{temperature})}$
     - *n = n+1*
   - Decrease the temperature according to the cooling schedule: the temperature was decreased by multiplying the temperature value to 0.8 at each stage. Each stage corresponds to M_k iterations.

The error rate obtained by this approach was 4.7% which shows a very small improvement comparing to the results that was previously obtained by the RF model (i.e. 4.9%). The results showed that combining different features did not enhance the RF model significantly. The error could be attributed to having very similar data for different modes; cars, buses, and bicycles waiting at a traffic light; a traveler collecting the run mode may have stopped just a bit to catch their breath or stopped at a traffic light or a stop sign, which would be similar to the walk mode; a bus and a car travelling on the same road with very similar kinetic variables such as speed and acceleration.

Conclusions

Different classifiers were developed using machine learning techniques to identify different transportation modes including bike, car, walk, run, and bus. In training and testing the classifier, data were obtained from smartphone sensors such as accelerometer, gyroscope, and rotation vector which were found to have important information for the purpose of mode recognition. A time window of one second was chosen, so the model can fit in a broader range of applications. For each method, parameters that needed to be optimized were examined to conduct a complete model selection. K-fold Cross-Validation and Out-Of-Bag error were used for model evaluations. Also, some performance measures such as the F-Score, Youden's index, and Discriminant Power were applied to assess model performances on the individual modes. Considering misclassification rates, the car and bus modes were the most difficult ones to distinguish, as would be expected. Even using more complex models such as SVM and RF, the car mode was misclassified as the bus mode in about 4-6% of the time. The Random Forest method was found to produce the best overall performance. However, for specific modes (i.e. walk and Run), the SVM outperformed the RF method. Several features were created and examined; among which 80 features were identified using the mRMR method as the most relevant feature. Other than some statistical measures of dispersion (e.g. range, max, var etc.), spectral entropy and energy were among the most important features. The focus of the future work will be on error analysis to identify any patterns that lead to misclassifications, and then to incorporate that knowledge into detection models for obtaining even higher detection accuracies.

Some recommendations for future directions that applies to this and similar research problems include: adding more data, applying approaches to examine the data as a sequence, considering more transportation modes (e.g. metro), and conducting error analysis[3] to gain some insights about where different models fail to correctly classify the data and consequently incorporate that knowledge into the models to enhance the detection performance.

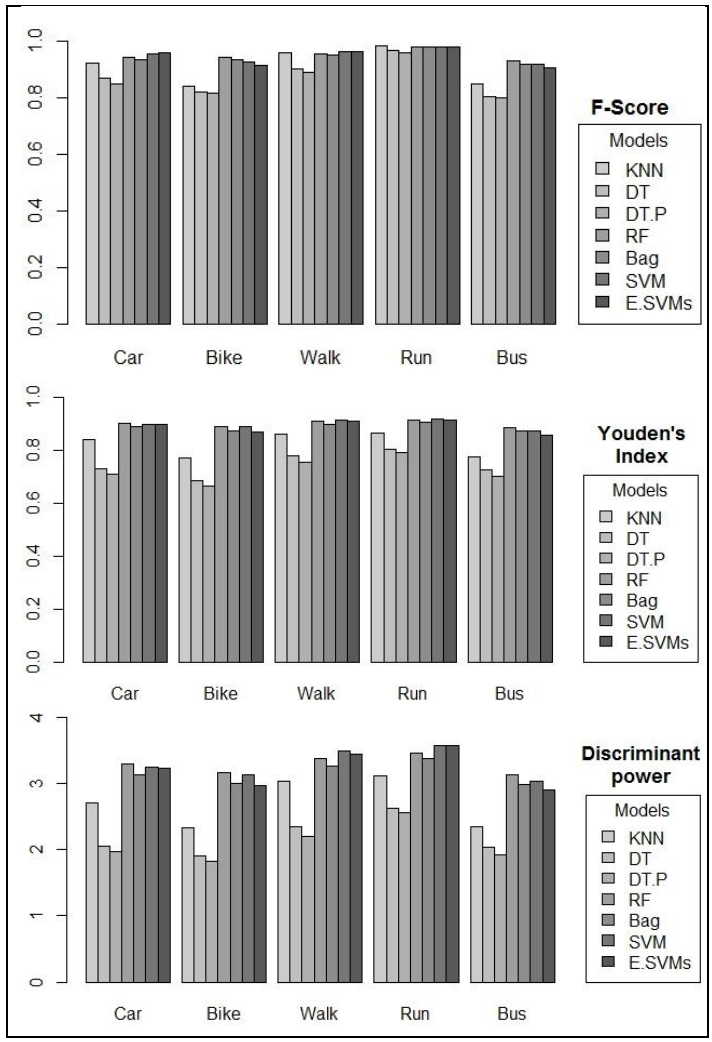


Figure 13 Model comparison results

[3] See appendix A and B for error analysis

Acknowledgments

This research was co-funded by the Mid-Atlantic University Transportation Center (MAUTC) and the Connected Vehicle Initiative UTC (CVI-UTC).

References

[1] Bao, L. and S.S. Intille, *Activity recognition from user-annotated acceleration data*, in *Pervasive Computing, Proceedings*, A. Ferscha and F. Mattern, Editors. 2004. p. 1-17.

[2] Kwapisz, J.R., G.M. Weiss, and S.A. Moore, *Activity recognition using cell phone accelerometers.* SIGKDD Explor. Newsl., 2011. **12**(2): p. 74-82.

[3] Susi, M., V. Renaudin, and G. Lachapelle, *Motion Mode Recognition and Step Detection Algorithms for Mobile Phone Users.* Sensors, 2013. **13**(2): p. 1539-62.

[4] Stenneth, L., et al. *Transportation mode detection using mobile phones and GIS information*. in *19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM SIGSPATIAL GIS 2011, November 1, 2011 - November 4, 2011*. 2011. Chicago, IL, United states: Association for Computing Machinery.

[5] Yu, X., et al. *Transportation activity analysis using smartphones*. in *Consumer Communications and Networking Conference (CCNC), 2012 IEEE*. 2012.

[6] Widhalm, P., P. Nitsche, and N. Brandie. *Transport mode detection with realistic Smartphone sensor data*. in *2012 21st International Conference on Pattern Recognition (ICPR 2012), 11-15 Nov. 2012*. 2012. Piscataway, NJ, USA: IEEE.

[7] Reddy, S., et al., *Using Mobile Phones to Determine Transportation Modes.* Acm Transactions on Sensor Networks, 2010. **6**(2).

[8] Manzoni, V., et al., *Transportation mode identification and real-time CO2 emission estimation using smartphones*. 2010, Technical report, Massachusetts Institute of Technology, Cambridge.

[9] Biljecki, F., H. Ledoux, and P. van Oosterom, *Transportation mode-based segmentation and classification of movement trajectories.* International Journal of Geographical Information Science, 2013. **27**(2): p. 385-407.

[10] Zheng, Y., et al., *Learning transportation mode from raw gps data for geographic applications on the web*, in *Proceedings of the 17th international conference on World Wide Web*. 2008, ACM: Beijing, China. p. 247-256.

[11] Gonzalez, P.A., et al., *Automating mode detection for travel behaviour analysis by using global positioning systems-enabled mobile phones and neural networks.* Iet Intelligent Transport Systems, 2010. **4**(1): p. 37-49.

[12] Byon, Y.J., B. Abdulhai, and A. Shalaby, *Real-Time Transportation Mode Detection via Tracking Global Positioning System Mobile Devices.* Journal of Intelligent Transportation Systems, 2009. **13**(4): p. 161-170.

[13] Zhang, L., M. Qiang, and G. Yang, *Mobility transportation mode detection based on trajectory segment.* Journal of Computational Information Systems, 2013. **9**(8): p. 3279-3286.

[14] Nham, B., K. Siangliulue, and S. Yeung, *Predicting mode of transport from iphone accelerometer data*. 2008, Tech. report, Stanford Univ.

[15] Nick, T., et al. *Classifying means of transportation using mobile sensor data*. in *Neural Networks (IJCNN), The 2010 International Joint Conference on*. 2010. IEEE.

[16] Bolbol, A., et al., *Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification.* Computers, Environment and Urban Systems, 2012.

[17] Zheng, Y., et al., *Understanding transportation modes based on GPS data for web applications.* ACM Transactions on the Web (TWEB), 2010. **4**(1): p. 1.

[18] Nitsche, P., et al., *A strategy on how to utilize smartphones for automatically reconstructing trips in travel surveys.* Procedia-Social and Behavioral Sciences, 2012. **48**: p. 1033-1046.

[19] Gong, H., et al., *A GPS/GIS method for travel mode detection in New York City.* Computers, Environment and Urban Systems, 2012. **36**(2): p. 131-139.

[20] Lester, J., et al., *MobileSense-Sensing modes of transportation in studies of the built environment.* UrbanSense08, 2008: p. 46-50.

[21] Jahangiri, A. and H. Rakha. *Developing a Support Vector Machine (SVM) Classifier for Transportation Mode Identification by Using Mobile Phone Sensor Data*. in *Transportation Research Board 93rd Annual Meeting*. 2014.

[22] Misra, H., et al. *Spectral entropy based feature for robust ASR*. in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*. 2004. IEEE.

[23] Lu, H., et al. *The Jigsaw continuous sensing engine for mobile phone applications*. in *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*. 2010. ACM.

[24] Shmaliy, Y., *Continuous-time signals*. 2006: Springer.

[25] Friedman, J.H., F. Baskett, and L.J. Shustek, *An algorithm for finding nearest neighbors.* IEEE Transactions on computers, 1975. **24**(10): p. 1000-1006.

[26] Hsu, C.-W. and C.-J. Lin, *A comparison of methods for multiclass support vector machines.* Neural Networks, IEEE Transactions on, 2002. **13**(2): p. 415-425.

[27] Hsu, C.-W., C.-C. Chang, and C.-J. Lin, *A practical guide to support vector classification*. 2003.

[28] Keerthi, S.S. and C.-J. Lin, *Asymptotic behaviors of support vector machines with Gaussian kernel.* Neural computation, 2003. **15**(7): p. 1667-1689.

[29] Valentini, G. and T.G. Dietterich, *Bias-variance analysis of support vector machines for the development of SVM-based ensemble methods.* The Journal of Machine Learning Research, 2004. **5**: p. 725-775.

[30] Kim, H.-C., et al., *Constructing support vector machine ensemble.* Pattern recognition, 2003. **36**(12): p. 2757-2767.

[31] Claesen, M., et al., *EnsembleSVM: A Library for Ensemble Learning Using Support Vector Machines.* Journal of Machine Learning Research, 2013(accepted).

[32] Breiman, L., et al., *Classification and regression trees*. 1984: CRC press.

[33] Hastie, T., et al., *The elements of statistical learning*. Vol. 2. 2009: Springer.

[34] Breiman, L., *Bagging predictors.* Machine learning, 1996. **24**(2): p. 123-140.

[35] Efron, B. and R.J. Tibshirani, *An introduction to the bootstrap*. Vol. 57. 1994: CRC press.

[36] Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.

[37] Ding, C. and H. Peng, *Minimum redundancy feature selection from microarray gene expression data.* Journal of bioinformatics and computational biology, 2005. **3**(02): p. 185-205.

[38] Peng, H., F. Long, and C. Ding, *Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2005. **27**(8): p. 1226-1238.

[39] Kohavi, R. *A study of cross-validation and bootstrap for accuracy estimation and model selection*. in *IJCAI*. 1995.

[40] James, G., et al., *An introduction to statistical learning*. 2013: Springer.

[41] R Core Team, *R: A Language and Environment for Statistical Computing*. 2014, R Foundation for Statistical Computing.

[42] Venables, W.N. and B.D. Ripley, *Modern applied statistics with S*. 2002: Springer.

[43] Chang, C.-C. and C.-J. Lin, *LIBSVM: a library for support vector machines.* ACM Transactions on Intelligent Systems and Technology (TIST), 2011. **2**(3): p. 27.

[44] Ripley, B. *tree: Classification and regression trees*. 2014.

[45] Gramacy, D.W.a.R.B., *maptree: Mapping, pruning, and graphing tree models.* 2012.

[46] Liaw, A. and M. Wiener, *Classification and Regression by randomForest.* R news, 2002. **2**(3): p. 18-22.

[47] Neville, P.G., *Controversy of Variable Importance in Random Forests.* Journal of Unified Statistical Techniques, 2013. **1**(1).

[48] Glover, F. and G.A. Kochenberger, *Handbook of metaheuristics*. 2003: Springer.

Arash Jahangiri received the M.Sc. in civil and environmental engineering from Virginia Tech, Blacksburg, in 2012.
He is currently a PhD candidate with the Charles E. Via, Jr. Department of Civil and Environmental Engineering at Virginia Tech. As a Graduate Research Assistant, he is currently with the Center for Sustainable Mobility (CSM) at Virginia Tech Transportation Institute (VTTI). His research interest comprises Intelligent Transportation Systems, Traffic Safety, Environmental impacts of Transportation, Artificial Intelligence, and Traffic Flow Theory.

Hesham A. Rakha (M'04) received the B.Sc. degree (with honors) in civil engineering from Cairo University, Cairo, Egypt, in 1987 and the M.Sc. and Ph.D.

degrees in civil and environmental engineering from Queen's University, Kingston, ON, Canada, in 1990 and 1993, respectively. He is currently the Samuel Reynolds Pritchard Professor of Engineering with the Charles E. Via, Jr. Department of Civil and Environmental Engineering, a Courtesy Professor with the Bradley Department of Electrical and Computer Engineering at Virginia Tech, Blacksburg, and the Director of the Center for Sustainable Mobility at the Virginia Tech Transportation Institute. He has authored/coauthored 4 books, more than 300 refereed publications in the areas of traffic flow theory, traffic modeling and simulation, traveler and driver behavior modeling, artificial intelligence, dynamic traffic assignment, traffic control, energy and environmental modeling, and safety modeling. Dr. Rakha in addition to being a member of IEEE is a member of the ITE, the ASCE, and the Transportation Research Board (TRB). He is a Professional Engineer in Ontario, Canada.

Transportation Mode Recognition using a Distributed Learning Approach

Arash Jahangiri
Center for Sustainable Mobility, Virginia Tech Transportation Institute
3500 Transportation Research Plaza, Blacksburg, VA 24061
E-mail: arashj@vt.edu
Phone: (540) 200-7561

Hesham Rakha (corresponding author)
Center for Sustainable Mobility, Virginia Tech Transportation Institute
3500 Transportation Research Plaza, Blacksburg, VA 24061
E-mail: HRakha@vtti.vt.edu
Phone: (540) 231-1505

Abstract

This study focuses on adopting machine learning techniques in a distributed learning approach to develop transportation mode detection models. When applying machine learning methods, the goal is to build models developed based on some data that includes a number of observations. Each observation contains a response (s) which is the dependent variable or the target value and an instance which contains some predictors or independent variables. In the case of transportation mode detection problem, the response is categorical (i.e. Car, Bus, Walk, etc.) and therefore it is considered as a classification problem. The goal is to classify transportation modes in a distributed approach (local level) and compare it to the detection models developed in a centralized manner (global level). In most transportation related problems, the data come from human subjects, which makes the prediction (or detection) more difficult due to the disparities between humans' behavior. Therefore, this paper employs a distributed learning approach in which detection models are developed for each individual instead of developing a single model for the entire data as is conducted in centralized systems. The number of models in the distributed approach is equal to the number of human subjects in the study who collected the data. As machine learning methods, support vector machine (SVM) and random forest (RF) were employed. Moreover, cross validation and out of bag error were applied to measure the performance of the models. Based on the data used in this study, it was found the distributed learning approach contributes to more accurate models compared to the centralized approach.

Keywords: distributed learning; support vector machine; random forest; machine learning; transportation mode recognition

Introduction

When dealing with Machine Learning problems in a centralized approach, the traditional way is to collect some relevant data and develop a single model based on the entire data. However, the data can be divided and different models can be developed on different parts of the data. Having a distributed system rather than a centralized system has been an active research topic in the computer science field, but other fields also have shown interest in adopting distributed approaches. There are several reasons why a distributed approach can be beneficial to the centralized approach as follows [1-4].

1. Dealing with a large amount of data in a centralized system is too difficult to be handled.
2. Data sources may be physically from different locations, and therefore too expensive to be directed to a centralized system
3. Sometimes the data from various sources cannot be shared due to privacy, security, and data ownership issues
4. Sometimes it is more efficient to have learning activities in parallel

In such conditions, a desired approach is to design a knowledge acquisition system that can analyze parts of the data wherever available and then the analysis results can be transmitted if needed. In other words, the knowledge can be acquired from different data parts and if required the results can be aggregated [3, 4]. Furthermore, specifically in the transportation domain, recent methods of data collection such as probe vehicles have been proposed to collect traffic data as a cost-effective alternative way compared to the more traditional ways such as loop detectors and video cameras [5]. In fact the combination of the traditional data collection methods (on-road sensors) with these new methods (on-board sensors) provide high quality data sets to be utilized [6]. In the new methods, instead of the infrastructure (e.g. loop detector, video cameras, etc.), individuals (e.g. probe vehicles, smartphones, etc.) collect the data. As a result, using these methods, there is an opportunity to analyze the data for each individual in a distributed manner and if required the analysis results can be aggregated.

Distributed data collection has been applied in several studies. In the Mobile century project [5], the feasibility of a traffic monitoring system for freeways by adopting GPS-enabled mobile phones was evaluated. In the Mobile Millennium study [7], arterial traffic conditions were estimated using GPS-enabled devices by applying statistical models. While these and other similar studies have been using a distributed data collection method, the data analysis has been conducted in a centralized system. In other words, each vehicle is equipped with mobile phones or GPS throughout the road network and transmits data (e.g. speed, location, etc.) to a central location where the data are processed and different inferences can be drawn such as travel time prediction, alternative routes, and etc. [6]. In the present study, other than collecting data in a distributed manner, a distributed data analysis or learning is adopted.

The paper is organized as follows. The first section presents an overview of the past studies on transportation mode detection as the example problem used to investigate the distributed learning approach. The second section formulates the distributed learning problem. The data collection process is presented in the third section. In the fourth section, the model development is explained. Finally, the results that draw a comparison between the distributed and centralized approach, are presented in the fifth section followed by the sixth section in which the conclusion is given.

Transportation mode recognition

Recognizing different types of physical activities using sensor data has been a recent research topic that has received considerable attention [8, 9]. Transportation mode detection can be considered as an activity recognition task in which data from smartphone sensors carried by users are utilized to infer what transportation mode the individuals have used. Micro-electromechanical systems (MEMS), such as accelerometers and gyroscopes are embedded in most smartphone devices [10] from which the data can be obtained at high frequencies. Smartphones, nowadays, are equipped with powerful sensors such as GPS, accelerometer, gyroscope, light sensors, temperature sensors, etc. Having such powerful sensors all

embedded in a small device carried in everyday life activities has enabled researchers to investigate new research areas. Other advantages of these smart devices are their ubiquity, their ability to send and receive data through Wi-Fi/cellular network/Bluetooth, and store data as well as to process the data [11].

The knowledge of individuals' mode of transport can facilitate some tasks and also can be adopted in several applications. Knowing the mode of transportation is an essential part of urban transportation planning, which is usually investigated through questionnaires/travel diaries/telephone interviews [11]. This traditional way of surveying is usually expensive, erroneous, limited to a specific area, and not so up-to-date [12]. As an environmental application, the carbon footprint as well as the amount of calories burnt of individuals can be determined by obtaining the mode of transport. Other applications include providing users with real-time information using the knowledge of speed and transport mode from the users as probes [11, 13], Providing individuals with customized advertisements and messages based on the transportation mode they are using [11], physical activity and health monitoring, tracking the hazard exposure and assessing the environmental impact of one's activities, and profile based recruitment for distributed data gathering [14].

Many studies have used GPS for classification purposes. However, several limitations are associated with the use of GPS sensors. These limitations include: GPS information is not available in shielded areas (e.g. tunnels) and the GPS signals may be lost especially at high dense locations which results in inaccurate position information. Moreover, the GPS sensor consumes significant power that sometimes users turn it off to save the battery [13]. Almost all studies used data from GPS sensors that have the aforementioned drawbacks. Also, all studies took advantage of Artificial Intelligence (AI) tools such as K-Nearest Neighbor as in [15], Decision Trees as in [11, 13-15], Bayesian Networks as in [11, 16], Random Forests as in [11, 15], Naïve Bayesian techniques as in [11, 14], Neural Networks as in [17], and Support Vector Machine (SVM) techniques as in [14-16, 18], of which the Decision Tree and SVM methods were used the most. Some studies have used additional information from GIS maps as in [11, 19].However, GIS data is not always available, and also this approach may not be suitable for real-time applications because it mostly relies on the knowledge of the entire trip with respect to the GIS features such as bus stops, subway entrances, and rail lines.

The Decision Tree method was identified as the best method by [14, 16] compared to some other methods including SVM. However, when applying SVM, several factors can greatly influence the model performance, which have not been considered in their studies. For example, a linear kernel was used in [14, 16] as part of the method, but generally for a certain type of problems and depending on the size of the available data and features, SVM can produce better results with more advanced kernels such as Gaussian kernel. Also, when applying Gaussian kernel, it was shown that if complete model selection is conducted with Gaussian there is no need to consider the linear kernel [20]. It is also unclear whether feature scaling and regularization were adopted in the most studies using SVM. Feature scaling is used to normalize the range of different features (or attributes), which leads to higher model performance and training speed and the regularization is incorporated into the model to deal with the issue of over-fitting (high variance).

Depending on the application of interest, different window sizes have been used for predicting the mode of transport. For example, [21] found that longer monitoring durations lead to higher accuracy. Intuitively, the bigger the window size the easier the prediction becomes since with bigger window sizes more information is available. If the application is only a survey for demand analysis the window size can be as large as trip duration, whereas if the application provides real-time information for environmental or some transit applications, then smaller window sizes are more desirable. The size should be as small as possible for some safety applications (e.g. crash prevention/mitigation). The time window in our previous works [15, 18] as well as the present study was assumed to be one second so that the potential application would include a broader range of applications such as environmental and safety applications. Other than the window size, other factors also influence the model performance as follows.

(5) Number of classes: as the number of classes increases, class differentiation becomes more difficult.
(6) Use of accelerometer/GPS/GIS data: the level of model dependency on different sources of data is considered as an important factor. Less dependent models are more desirable as they can be applicable even with limited sources of data. In this case, sensors such as accelerometers and gyroscopes are more reliable since their data are always available.
(7) Ability to distinguish between motorized classes: as different motorized classes have similar characteristics such as speed and acceleration, a model capable of differentiating between these modes is of great value. For example, distinguishing the bus mode from the car mode is significantly more difficult than discriminating walking from driving.
(8) Sensor positioning: it shows how realistic the experiments are conducted. Positioning the devices at certain locations increases the prediction accuracy because the movements can be monitored in more detail, but may not reflect realistic behavior. Some of the studies required that the participants attach sensors/smartphones to different parts of their body.

Distributed learning in Transportation

In predicting or estimating different measures, attributes, and/or behavior in the transportation science, in order to apply machine learning techniques, prediction/detection models are developed based on some data that include a number of observations. Each observation contains a response (s) (i.e. the different measures such as travel time, transportation mode, crash probability) which is the dependent variable or the target value and an instance which contains some predictors or independent variables. The predictors are specific to each problem but examples are vehicle speed, vehicle type, vehicle acceleration, signal setting, and etc. In terms of the general concept, a part of the data is used for training and a part is set aside for testing and validating (well-known techniques can be applied such as Cross Validation). The goal in the centralized approach is to develop a single model to predict or estimate the responses based on the data instances. An example from our previous works [15, 18] is used to show how a centralized approach is applied compared to the distributed approach. The example is recognizing transportation modes based on the data from smartphone sensors as explained above. In the transportation mode detection problem, the response or the dependent variable is the transportation mode (e.g. Bike, Bus, Car, and etc.) and the data instance includes the predictors (also called attributes, features, and independent variables) such as acceleration that were obtained from the smartphone sensors (e.g. accelerometer, gyroscope, etc.).

Although the distributed data collection has been adopted in some studies, not much research has been conducted on distributed learning or analysis in the transportation domain. In this approach, different models are developed for different individuals. In other words, instead of using the entire data to develop a single prediction model, a model is developed for each individual based on the data collected from that specific individual. The resultant prediction models are expected to be more accurate. Three important factors have motivated us to apply the distributed learning approach. First, as mentioned earlier, sometimes the available data set is very big and sometime it is physically difficult or impossible to handle them in a centralized system. Second, new methods of collecting data (e.g. via smartphones), as explained earlier, have enabled researchers to collect the data in a distributed manner. In addition, thanks to the technology development, recent handheld devices have the capability of analyzing the data as well as collecting them. Third, in most transportation related problems, the data come from human subjects, which makes the prediction more difficult due to the disparities between humans' behavior. Some studies try to account for more predictors to describe the differences between humans (e.g. age, sex), but it is more difficult in practice to obtain such data. Also, even two similar human subjects may behave much differently (e.g. aggressive vs. conservative driving). In the distributed learning approach, the developed

models for individuals reflect these differences (i.e. differences in age, sex, aggressiveness, and etc.) without having to include additional variables in the model, which sometimes is difficult or impossible to accurately measure (e.g. aggressiveness).

To better understand the distributed learning approach compared to the centralized learning approach, the formulations of these two approaches are discussed as follows. As mentioned earlier, the data contains a response (s) which is the dependent variable or the target value and an instance which contains some predictors or independent variables. In the case of transportation mode detection problem, the response is categorical (i.e. Car, Bus, Walk, etc.) and therefore it is considered as a classification problem. In the centralized learning, given a dataset D, the learning algorithm determines a hypothesis h from a set of hypothesis H (i.e. $h \subset H$) that optimizes a selected performance measure P. The dataset D contains N observations and each observation includes F features (or attributes/variables) and a response or target class denoted by C which can be any value from the set of all classes. In the case of transportation mode recognition, the classes are Car, Bike, Bus, Walk, and Run. In classification problems, a hypothesis h is a decision boundary that classifies different classes. There are many decision boundaries to classify (i.e. set of H); Support Vector Machine (SVM) and Random Forest(RF) in this study were applied to find the best decision boundary (i.e. hypothesis h) that optimizes a performance measure P. As for the performance measure, cross-validation error rate and out-of-bag error were used for SVM and RF techniques, respectively.

Data Collection

The data that was used in this study were from our previous works in which a smartphone application was developed to collect the required data from smartphones [15, 18]. To collect the data, the transportation mode (Car, Bike, Bus, Walk, and Run) should be selected before starting the logging process, and then the application stores the data coming from smartphone's sensors including GPS, Accelerometer, Gyroscope, and Rotation Vector at the highest possible frequency. Data collection was carried out by five individuals. While the app collected the data from the aforementioned sensors, the data from the GPS sensor were not used in this study. About 150 minutes of data were gathered by each person while using different transportation modes. The data were equally gathered amongst transportation modes (i.e 30 minutes for each mode). A total of 750 minutes of data were stored and used for model development. There was no restriction on placing the smartphone (i.e. attaching the smartphone to part of the body) while collecting the data so the approach becomes independent of the device placement.

Model development

Methods

To compare the proposed approach against the traditional approach, machine learning methods namely Random Forest (RF) and Support Vector Machine (SVM) were applied. Random Forest method, as proposed in 2001 [22], creates an ensemble of decision trees from which a majority vote make the predictions. There are two parameters to be determined in order to apply this method, namely the number of decision trees and the number of variables (features) to use in each tree. Recursive binary splitting is the method used for growing trees. Using this method, a predictor (or feature) is selected to divide the data into two parts at each step of growing the tree; the decision on data separation can be based on different criteria such as Gini index, Cross-Entropy, or classification Error rate. Gini index and Cross-Entropy have been recommended to have a better performance. Cross-Entropy that was used in this study can be obtained through **Equation 16** [23].

$$\boldsymbol{G = \sum_{k=1}^{K} P_k^m \log P_k^m} \qquad \textbf{Equation 16}$$

Where,

$$\boldsymbol{P_k^m = \frac{1}{N^m} \sum_{x_i^m} I(y_i^m = k)}$$

N^m	Number of observations received at node m
y_i^m	The response value corresponding to the observation i at node m
x_i^m	The feature vector corresponding to the observation i at node m
k	class

A single Random Forest model was developed using the entire data set to represent the traditional approach. Also, to assess the proposed approach, five Random Forest models were developed, each of which was based on the data from the corresponding data obtained from each individual. To implement RF method, the R software and RandomForest packages were used [24, 25].

SVM is known as a large margin classifier, which means when classifying data, it determines the best possible decision boundary that provides the largest possible gap between classes. This characteristic contributes to a higher confidence in solving classification problems. To implement SVM, the LibSVM library of SVMs was applied. For multiclass classification, considering K classes, LibSVM applies one-against-one method in which $K(K-1)/2$ binary models are built. Among these, LibSVM chooses the parameters that achieve the highest overall performance. Another well-known method is called one-against-all which is more intuitive and has similar performance. However, LibSVM takes advantage of one-against-one because of its shorter training time. [26]. To construct the model, the following factors are taken into account: using a Gaussian kernel with complete model selection, which entails consideration of the regularization parameter and the Gaussian parameter, applying feature scaling, and examining several features.

Equation 21 presents the SVM formulation to solve the classification problem and the associated constraints are shown in Equation 22 and Equation 23 [27]. The objective function is comprised of two terms: minimizing the first term is basically equivalent to maximizing the margin between classes, and the second term consists of an error term multiplied by the regularization (penalty) parameter denoted by C. The C parameter should be determined to provide the relative importance between the two terms. Equation 22 ensures that margin of at least 1 exist with consideration of some violations. The value of 1 was resulted from normalizing w. Equation 23 restricts the data points to the points that have positive errors.

$$\min_{w,b,\xi} \left(\frac{1}{2} w^T w + C \sum_{n=1}^{N} \xi_n \right) \qquad \textbf{Equation 17}$$

Subject to:

$$y_n(w^T \phi(x_n) + b) \geq 1 - \xi_n \,, n = 1, \dots, N \qquad \textbf{Equation 18}$$

$$\xi_n \geq 0 \,, n = 1, \dots, N \qquad \textbf{Equation 19}$$

Where,

w	Parameters to define decision boundary between classes
C	Regularization (or penalty) parameter
ξ_n	Error parameter to denote margin violation
b	Intercept associated with decision boundaries
$\phi(x_n)$	Function to transform data from X space into some Z space

Kernels are functions that are adopted to solve the SVM problem. When solving the problem, a function $\phi(x_n)$ is applied to transfer the data from the current X space into a higher maybe infinite dimensional Z space. So, basically, for the function $\phi(x_n)$, the kernel corresponds to the vector inner products in the Z space. Different types of kernels exist such as linear kernel, polynomial kernels, and Gaussian kernel. Linear kernel, as applied in [14, 28], is the basic mode which means no kernels are actually taken into account. In other words, vector inner product as appears in the dual formulation of the problem are considered without transforming data into another space. According to our data size and feature size, Gaussian kernel was believed to be the most appropriate kernel [29], and if a complete model selection is carried out, there is no need to test the linear kernel because the results obtained from the Gaussian kernel include the results obtained from the linear kernel. In fact, when using Gaussian kernel, If $\sigma^2 \to \infty$ and $C = C^L\sigma^2$ where C^L is fixed then the SVM classifier behaves like an SVM classifier with a linear kernel with regularization parameter C^L [20]. In this paper, the $\phi(x_n)$ function which corresponds to the Gaussian kernel has an infinite dimensional space. The formulation of the Gaussian kernel is shown in **Equation 20**.

$$K(x, x') = exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) \qquad \textbf{Equation 20}$$

Where,

x, x'	n-dimensional vectors
$\|x - x'\|$	Euclidean distance between vectors x, x'
σ	Gaussian parameter

Feature selection

Feature selection is considered as a critical task as it can reduce the dimensionality of the problem, reduce the noise, identify more important predictors, and lead to more interpretable features [30]. In our previous work [18], we used approximately 60 features that we created from the sensor data, mostly based on some statistical measures of dispersion. In the present study and similar to our most recent work [15], we created a set of features that include those 60 features with some modifications. First, a feature should have a meaningful relationship to the transportation modes. Therefore, absolute values of the rotation vector sensor are excluded from the feature set. Second, since a time window is being monitored, other features that can describe variations in time were added to the feature set. Also, spectral entropy was added, which can be used as a measure to show the peaky spots of a distribution [31]. A High value of Spectral Entropy for a distribution shows that the distribution is somewhat flat. Conversely, the Spectral Entropy decreases when the distribution becomes less flat [32]. In addition, the data from the sensors were treated as signals, consequently, the energy of the signal within the time window of interest was added to the feature set [33]. Also, as mentioned earlier, the data from the GPS were excluded to only focus on the scenario where no GPS data are available. To summarize, using the data from different sensors and for each time window, Table 18 shows the measures that were used to create the feature set. A total of 167 features were created. Then, to make a selection of the most representative features the minimum redundancy maximum relevance (mRMR) approach was applied. As can be inferred from its title, this method tries to find the most relevant features to the target class and simultaneously tries to avoid redundant features [30].

Table 18: Measures used to create features

No.	Measure	No.	Measure	No.	Measure	No.	Measure
1	$mean(x_i^t)$	6	$range(x_i^t)$	11	$max\left(\dot{x_i^t}\right)$	16	$iqr\left(\dot{x_i^t}\right)$
2	$max(x_i^t)$	7	$iqr(x_i^t)$	12	$min\left(\dot{x_i^t}\right)$	17	$signChange\left(\dot{x_i^t}\right)$
3	$min(x_i^t)$	8	$signChange(x_i^t)$	13	$var\left(\dot{x_i^t}\right)$	18	$spectralEntropy(x_i^t)$
4	$var(x_i^t)$	9	$energy(x_i^t)$	14	$std\left(\dot{x_i^t}\right)$		
5	$std\ (x_i^t)$	10	$mean\left(\dot{x_i^t}\right)$	15	$range\left(\dot{x_i^t}\right)$		

Results

Measures of comparison

To be able to evaluate how the proposed approach performs compared to the traditional approach k-fold cross validation error rate and out of bag error were used when assessing the SVM and the RF models, respectively. K-fold cross validation is a great technique to evaluate how well a model is performing. The classic way is to divide data into two data sets, develop the model based on one set, namely the training set, and evaluate the model using the other set called the test set. Using k-fold cross validation, the data is divided into k parts. The model is developed based on k-1 part and evaluated based on the remaining untouched part. The procedure is repeated k times, each time with a different part as the test set and the remaining parts as the training set. Subsequently, the final result is averaged over the results obtained from the k models. In practice, the best choice for the number of folds is 5 or 10 [34, 35]. Out-Of-Bag error is an accurate estimate of the prediction error suitable for tree based models such as Bagging and Random Forest and is almost identical to the Cross-Validation accuracy [23].

Comparison

Using SVM or RF methods, the proposed approach resulted in 5 different models as this approach develops different models for different individuals. Since in the present study data obtained from total of 5 people were used, total of 5 different models were constructed using each method (i.e. SVM and RF). On the other hand, the traditional approach led to a single model. The cross validation error rates of the SVM models are shown in Figure 14. Similarly, the out of bag error rates of the RF models are shown in Figure 15. For the sake of comparison, the average of the 5 models associated with different individuals was also obtained that can be seen in the figures. The horizontal axis in Figure 14 and Figure 15 represent the number of features (variables) that were used to develop the models. Feature selection method, namely mRMR, that was described earlier was applied to choose the most representative features.

Looking at the error rates of all models as shown in Figure 14 and Figure 15, all individual models as well as the average model resulted in a lower error rate compared to the model obtained from the traditional approach except in only one case when using SVM and the number of variables were 9. In this special case, it seems that the number of variables is not sufficient to develop a good model as high error rates were resulted. Therefore, this case can be excluded simply because not enough information is provided to the models. It can be seen that when using 36 variables or above, SVM models achieved low error rates which shows these models have good performances. Thus, at these points (above 36 variables) comparison can be drawn. Similarly, RF models resulted in more accurate models when the proposed approach was adopted as opposed to the traditional approach. As shown in Figure 15, the proposed approach always led to a lower error rate. Using both methods (i.e. SVM, RF), the more variables (features) were included the more accurate models were achieved. Furthermore, when having more variables, the performance of the traditional approach becomes closer to that of the proposed approach.

However, the traditional approach was always fallen behind even when using 100 variables. In fact, considering all models, not much benefit is gained by including more than 49 variables.

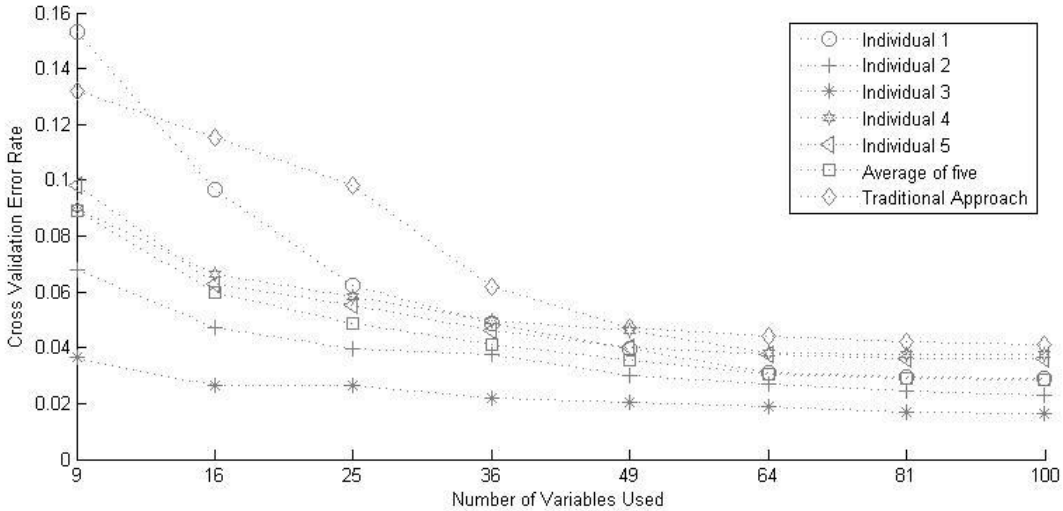


Figure 14: Accuracy comparison - SVM models

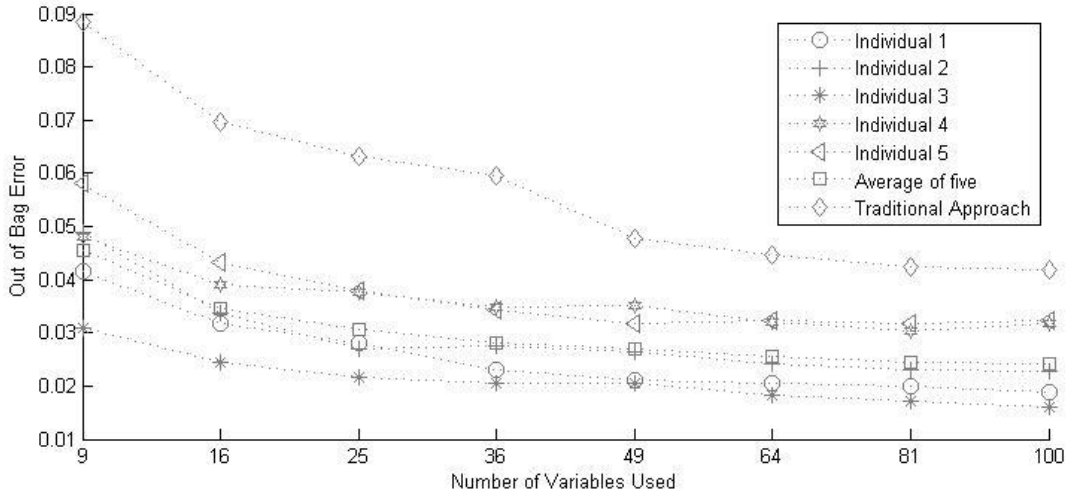


Figure 15: Accuracy Comparison - RF models

When comparing the two approaches as presented in Figure 14 and Figure 15, three points can be made. First, the data used in the traditional approach is much bigger (five times in this case) than the data used in the proposed approach because in the traditional approach the data from all individuals (five in this case) were used. This is not always true but the bigger data favors the traditional approach because in machine leaning problems more data may produce better results. Even though, the proposed approach led to more accurate models. Also, in reality data from many individuals need to be dealt with. As a result, when using the traditional approach, a much bigger data set needs to be handled. At some point, it becomes too difficult or even impossible to develop models based on a huge data set due to the memory

and other computational limitations. On the other hand, the proposed approach does not suffer from these limitations since models are developed based on data from each individual. Thus, if the data from 1000 individuals were available 1/1000 of the entire data set would be used to develop models in the proposed approach that is a much smaller data set. Moreover, depending on the problem of interest, it may be possible to implement models in individuals' handheld devices and the results from the devices can be transmitted if needed. Second, it may seem that the accuracies from the proposed approach do not differ significantly from those of the traditional approach. However, even small improvements can be vital in certain applications. For example, users of a safety application that notifies them from a potential hazard may lose their trust even when they experience small percentages of false alarm rates. Third, the five individuals who collected the data for this study had the same sex and age and therefore, their behavior is somewhat similar to each other. It is expected that if different behavior caused by being in other age groups, different sex, and/or other reasons such as individuals' aggressiveness the results would reveal the benefits of the proposed approach even more.

Conclusion

The distributed data collection methods in transportation (e.g. using probe vehicles, smartphones, etc.) have become important methods of collecting high quality data. Furthermore, handheld devices have becoming more capable in terms of the memory and processing the data. An approach, namely the distributed learning, was proposed to solve machine learning problems in the transportation domain. In general, a distributed system is advantageous to a centralized system because of some limitations associated with centralized systems; difficulty in dealing with large data sets, issues with sharing data due to privacy, security, and data ownership, etc. Instead of the traditional way (i.e. using a centralized system) in which a single model is developed based on the entire data set, in the proposed approach, a model is developed for each individual based on the data corresponding to that specific individual. Therefore, the number of developed models in the proposed approach is equivalent to the number of individuals who collected the data. Since in many transportation problems the data come from human subjects, much variation exists in the collected data because of the differences in humans' behavior. In the traditional approach, these variations need to be captured by including additional features (variables) such as age and sex. Including these variables, however, is costly (i.e. need to collect extra features) or sometimes impossible to capture (e.g. variables like aggressive driving behavior) in practice. On the other hand, the proposed approach can explain any differences between human subjects without the need to have those additional variables (e.g. age, sex, etc.) because models developed in the proposed approach are based on data from only one individual. As a result, any difference in the subjects' behavior is an inherent part of the models. Transportation mode detection problem as an example was used in order to show how the proposed approach performs compared to the traditional approach. Also, Support vector Machine (SVM) and Random Forest (RF) methods, as two well-known machine learning techniques, were adopted to develop models using each approach. Models developed using both SVM and RF methods in the proposed approach contributed to more accurate models compared to the models in the traditional approach. The accuracies obtained using the proposed approach were not significantly higher than those of the traditional approach, but even small improvements are considered vital in certain applications (e.g. safety). More accurate models were obtained using the proposed approach even though the following situations were in favor of the traditional approach; a bigger data set were used in the traditional approach, and the five individuals, who collected the data, had somewhat similar characteristics as all had the same age and sex.

Acknowledgements

This research effort was funded by the Mid-Atlantic University Transportation Center (MAUTC) and the Connected Vehicle Initiative UTC (CVI-UTC).

References

[1] Weiß, G. *A multiagent perspective of parallel and distributed machine learning*. in *International Conference on Autonomous Agents: Proceedings of the second international conference on Autonomous agents*. 1998.

[2] Hall, L.O., et al., *Learning rules from distributed data*, in *Large-Scale Parallel Data Mining*. 2000, Springer. p. 211-220.

[3] Caragea, D., A. Silvescu, and V. Honavar, *Decision tree induction from distributed heterogeneous autonomous data sources*, in *Intelligent Systems Design and Applications*. 2003, Springer. p. 341-350.

[4] Khedr, A.M., *Learning k-nearest neighbors classifier from distributed data.* Computing and Informatics, 2012. **27**(3): p. 355-376.

[5] Herrera, J.C., et al., *Evaluation of traffic data obtained via GPS-enabled mobile phones: The< i> Mobile Century</i> field experiment.* Transportation Research Part C: Emerging Technologies, 2010. **18**(4): p. 568-583.

[6] Leduc, G., *Road traffic data: Collection methods and applications.* Working Papers on Energy, Transport and Climate Change, 2008. **1**: p. 55.

[7] Herring, R., et al. *Using mobile phones to forecast arterial traffic through statistical learning*. in *89th Transportation Research Board Annual Meeting, Washington DC*. 2010.

[8] Bao, L. and S.S. Intille, *Activity recognition from user-annotated acceleration data*, in *Pervasive Computing, Proceedings*, A. Ferscha and F. Mattern, Editors. 2004. p. 1-17.

[9] Kwapisz, J.R., G.M. Weiss, and S.A. Moore, *Activity recognition using cell phone accelerometers.* SIGKDD Explor. Newsl., 2011. **12**(2): p. 74-82.

[10] Susi, M., V. Renaudin, and G. Lachapelle, *Motion Mode Recognition and Step Detection Algorithms for Mobile Phone Users.* Sensors, 2013. **13**(2): p. 1539-62.

[11] Stenneth, L., et al. *Transportation mode detection using mobile phones and GIS information*. in *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 2011. ACM.

[12] Widhalm, P., P. Nitsche, and N. Brandie. *Transport mode detection with realistic Smartphone sensor data*. in *2012 21st International Conference on Pattern Recognition (ICPR 2012), 11-15 Nov. 2012*. 2012. Piscataway, NJ, USA: IEEE.

[13] Manzoni, V., et al., *Transportation mode identification and real-time CO2 emission estimation using smartphones.* SENSEable City Lab, Massachusetts Institute of Technology, nd, 2010.

[14] Reddy, S., et al., *Using Mobile Phones to Determine Transportation Modes.* Acm Transactions on Sensor Networks, 2010. **6**(2).

[15] Jahangiri, A. and H. Rakha, *Applying Machine Learning Techniques to Transportation Mode Recognition using Mobile Phone Sensor Data.* ieee transactions on intelligent transportation systems, 2014.

[16] Zheng, Y., et al. *Learning transportation mode from raw gps data for geographic applications on the web*. in *Proceedings of the 17th international conference on World Wide Web*. 2008. ACM.

[17] Gonzalez, P.A., et al., *Automating mode detection for travel behaviour analysis by using global positioning systemsenabled mobile phones and neural networks.* Intelligent Transport Systems, IET, 2010. **4**(1): p. 37-49.

[18] Jahangiri, A. and H. Rakha. *Developing a Support Vector Machine (SVM) Classifier for Transportation Mode Identification by Using Mobile Phone Sensor Data*. in *Transportation Research Board 93rd Annual Meeting*. 2014.

[19] Gong, H., et al., *A GPS/GIS method for travel mode detection in New York City.* Computers, Environment and Urban Systems, 2012. **36**(2): p. 131-139.
[20] Keerthi, S.S. and C.-J. Lin, *Asymptotic behaviors of support vector machines with Gaussian kernel.* Neural computation, 2003. **15**(7): p. 1667-1689.
[21] Byon, Y.-J., B. Abdulhai, and A. Shalaby, *Real-time transportation mode detection via tracking global positioning system mobile devices.* Journal of Intelligent Transportation Systems, 2009. **13**(4): p. 161-170.
[22] Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.
[23] Hastie, T., et al., *The elements of statistical learning*. Vol. 2. 2009: Springer.
[24] R Core Team, *R: A Language and Environment for Statistical Computing*. 2014, R Foundation for Statistical Computing.
[25] Liaw, A. and M. Wiener, *Classification and Regression by randomForest.* R news, 2002. **2**(3): p. 18-22.
[26] Chang, C.-C. and C.-J. Lin, *LIBSVM: a library for support vector machines.* ACM Transactions on Intelligent Systems and Technology (TIST), 2011. **2**(3): p. 27.
[27] Hsu, C.-W. and C.-J. Lin, *A comparison of methods for multiclass support vector machines.* Neural Networks, IEEE Transactions on, 2002. **13**(2): p. 415-425.
[28] Zheng, Y., et al., *Learning transportation mode from raw gps data for geographic applications on the web*, in *Proceedings of the 17th international conference on World Wide Web*. 2008, ACM: Beijing, China. p. 247-256.
[29] Hsu, C.-W., C.-C. Chang, and C.-J. Lin, *A practical guide to support vector classification*. 2003.
[30] Ding, C. and H. Peng, *Minimum redundancy feature selection from microarray gene expression data.* Journal of bioinformatics and computational biology, 2005. **3**(02): p. 185-205.
[31] Misra, H., et al. *Spectral entropy based feature for robust ASR*. in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*. 2004. IEEE.
[32] Lu, H., et al. *The Jigsaw continuous sensing engine for mobile phone applications*. in *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*. 2010. ACM.
[33] Shmaliy, Y., *Continuous-time signals*. 2006: Springer.
[34] Kohavi, R. *A study of cross-validation and bootstrap for accuracy estimation and model selection*. in *IJCAI*. 1995.
[35] James, G., et al., *An introduction to statistical learning*. 2013: Springer.

Chapter 4: Driver Violation Prediction

(Paper accepted to: Transportation Research Board 94th Annual Meeting, 2015)

(Paper accepted to: 2015 IEEE 18th International Conference on Intelligent Transportation Systems)

(Paper submitted to: Accident Analysis & Prevention journal)

Predicting Red-light Running Violations at Signalized Intersections using Machine Learning Techniques

Arash Jahangiri
Center for Sustainable Mobility, Virginia Tech Transportation Institute
3500 Transportation Research Plaza, Blacksburg, VA 24061
E-mail: ArashJ@vt.edu
Phone: (540) 200-7561

Hesham Rakha (corresponding author)
Center for Sustainable Mobility, Virginia Tech Transportation Institute
3500 Transportation Research Plaza, Blacksburg, VA 24061
E-mail: HRakha@vtti.vt.edu
Phone: (540) 231-1505

Thomas A. Dingus
Virginia Tech Transportation Institute
3500 Transportation Research Plaza, Blacksburg, VA 24061
E-mail: TDingus@vtti.vt.edu
Phone: (540) 231-1501

Word count: 5,603 + 1500 (5 Figures + 1 Table) = 7,113

Submitted for presentation at the 94th Annual Meeting of the Transportation Research Board and publication in the *Transportation Research Record*

Abstract

Statistics demonstrate that a large number of crashes occur at signalized intersections due to traffic violations, specifically red light running (RLR). In order to prevent/mitigate intersection-related crashes, these violations need to be identified before they occur, so appropriate warnings can be issued. Several factors influence the drivers' behavior when approaching intersections. These include the vehicle speed, Time to Intersection (TTI), Distance to Intersection (DTI), age, gender, etc. However, the driver-related factors (i.e. age, gender) are more difficult to obtain in practice. On the other hand, kinetic factors (e.g. speed, acceleration) can be obtained by monitoring the movement of vehicles through video cameras installed on the infrastructure or through on-board devices installed on the vehicles. Hence, the problem of interest is to develop models to predict red light running (RLR) violations using kinetic information of individual drivers/vehicles. Machine learning techniques, namely Support Vector Machine (SVM) and Random Forest (RF), were adopted to develop prediction models. The minimum Redundancy Maximum Relevance (mRMR) factor selection technique was used to identify the most important factors for model development. To evaluate the performance of the models the K-fold cross-validation and out-of-bag (OOB) errors were used for the SVM and RF models, which contributed to high prediction accuracies of 96.7 and 94.2 percent, respectively. It was shown that other than the critical instant at which the traffic signal changes to yellow, an appropriate time window with respect to the yellow onset can provide additional useful information ensuring that the driver decision occurs during that time window.

Keywords: driver violation; red light running; signalized intersection; violation prediction; support vector machine; random forest; machine learning

Introduction

In 2012, more than 2.5 million intersection related crashes occurred, of which 2,850 were fatal crashes and 680,000 were injury crashes [1]. More specifically, statistics demonstrate that a large number of crashes occur at signalized intersections due to traffic violations, of which running red lights has been reported as a serious issue. According to the Insurance Institute for Highway Safety, in 2012, 683 people were killed and an estimated 133,000 were injured in crashes due to running red lights [2]. A survey of 2,000 U.S. residents age 16 and older that was conducted by the AAA Foundation for Traffic Safety showed that about 93% of drivers believe that running through a red light is unacceptable if it is possible to stop safely. However, one-third mentioned they ran through a red light in the past 30 days. This shows that although the drivers are generally aware of the dangers of this type of violation, they are likely to run a red light on occasion [3].

Drivers approaching signalized intersections need to make a decision whether to stop or go when the traffic signal changes from green to yellow. In case they decide to go, if the signal turns to red before they pass the stop bar they are considered as red light violators. In another scenario, in case they abruptly stop, if a following driver have made a conflicting decision (i.e. decision to go) rear-end crashes may occur. The area in which the drivers should decide what to do is known as the dilemma zone which can be defined in space or time [4]. Dilemma zone problem is a classic problem first introduced in [5] followed by many other studies throughout the literature such as [4, 6-14].

A number of studies have attempted to investigate factors that affect the driver behavior when approaching the signalized intersections [4, 8, 12-17]. These factors influence the drivers' decision whether to stop or go when facing a yellow light and consequently have an impact on the crash risk for rear-end as well as right-angle crashes. Factors that have been studied throughout the literature include but not limited to: the driver perception-reaction time; the driver's acceptable deceleration rate; the driver's age; the driver's gender; the time-to-intersection (TTI) at the onset of yellow; the distance-to-intersection (DTI) at the onset of yellow; approach speed; vehicle type; presence of side-street vehicles, pedestrians, bicycles, or opposing vehicles waiting to turn left; flow rate; length of yellow interval; cycle length; presence of police; and pavement condition.

Data collection methods restrict the factors that can be considered: using driving simulator as applied in [14, 15, 18] provides the capability of examining many factors. However, the behavior of the drivers may not reflect their natural behavior when driving in real world conditions. On the other hand, using other methods such as video cameras as adopted in [8, 12, 13], the drivers' natural behavior can be captured as the drivers are not aware that their data are being collected. Nevertheless, some factors such as age and gender cannot be captured. Also, some other factors such as the presence of bicycles or police may not occur at all while recording the data. Somewhere between the naturalistic data collection and the driving simulator data collection lays the data collection using an experimental test track that is conducted in a controlled environment. The behavior of the drivers is more natural than the driving simulator but their behavior may be affected because after all they know they are in an experiment. The examples in which the data were collected in a controlled field environment include [4, 16, 17, 19].

Some studies concentrate on investigating the characteristics of red light violators as well as conditions in which the drivers are more or less likely to violate [20-25]. For example drivers who do not use safety belt and non-Caucasian drivers were more likely to violate the red light. Moreover, larger intersections and higher traffic volumes are associated with higher RLR violation rates [20]. Some of these studies developed models to estimate the frequency of red light violators. For instance, a regression model was developed in [23] using several factors as model variables such as flow rate, cycle length, yellow duration, and etc.

Artificial Intelligent (AI) techniques have been adopted to solve many problems in the transportation domain: Real-time detection of driver cognitive distraction [26], lane detection and tracking [27] transportation mode recognition [28], traffic sign detection [29], and Incident detection [30] are some examples in which applying AI methods have shown promising results. AI methods have also been used to predict RLR violations at signalized intersections [31, 32]. However, more research is

needed to enhance current and traditional prediction models. When using AI methods, different terms may be used to refer to the factors that are employed to build models; these terms that all have the same meaning include the words “factors”, “features”, “variables”, and “predictors”. In this paper, the term factor is used for consistency. Other than AI tools, statistical and probabilistic approaches have also been applied [6, 33]. For example, a probabilistic model was developed in [33] to predict RLR violations, taking into consideration minimizing both false alarm rate and missing error.

To predict RLR violations a classification tree model was applied in [32]and the vehicle’s distance at the onset of yellow, operating speed at the onset of yellow, and position in traffic flow were found to be the most important factors. They examined a limited number of factors which were obtained from the onset of yellow instant. In addition to the factors included in their study, there are several factors seem to be important in developing RLR prediction models. A systematic method is required to include potential factors and determine the importance of them in improving the performance of the model. In this study, a proper factor selection method is used to identify the most useful factors. Another point to be mentioned is related to the time instant or period from which factors are extracted. Although the yellow onset is an important instant (i.e. because it is the moment the drivers encounter the yellow and consequently they need to make a decision to stop or go) the drivers’ decision are not made at that instant but during a short time period. Therefore, in addition to the factors from the yellow onset, factors that describe the drivers’ behavior in a time period immediately after the yellow onset should also be examined. This has been taken into account in the present study.

In another study [31] aiming at predicting RLR violations, Support Vector Machine (SVM) and Hidden Markov Model (HMM) were applied. They showed how their models outperformed some traditional methods of prediction. They did not use any factor selection method; Using SVM, They found by experimenting different combinations that three factors, namely DTI, speed, and acceleration, led to the best result. Similarly, when applying HMM method they tested different combinations of factors to find the best ones to use; DTI; speed; acceleration; TTI; and required deceleration parameter (RDP). In addition, they considered a point in time after which the prediction becomes useless as not enough time is available for a driver in a potential collision to react. They logic is that if a safety system employs any prediction model, the prediction task needs to be carried out before a certain time to provide sufficient time for endangered drivers to respond. Furthermore, they considered a time window from which the required factors were extracted for model development. This time window was chosen right before the aforementioned critical point in time. They determined this critical point based on two criteria (whichever happened first): (1) minimum time threshold: three different values were selected based on a human response time distribution as discussed in [34]; 1, 1.6, and 2 seconds sufficient for 45%, 80%, and 90% of the population to respond, respectively and (2) minimum distance threshold: in case vehicles’ approaching speed was very low, vehicles get too close to the intersection and thus a minimum DTI was assumed.

However, the minimum time threshold is not enough to avoid a possible collision since it only corresponds to the driver response time without considering the vehicle response time. In other words to avoid a potential collision two time periods should be available; the driver response time and the vehicle response time. Moreover, the factors that they selected to use did not reflect the interaction between the drivers and the signal setting. For example, it seems that the factors such as DTI and speed in their SVM model were extracted from a time window that they defined without knowing when the yellow light starts (i.e. the yellow light may start before or after their defined time window). As mentioned earlier, both the yellow onset and the time window right after that are very important as they contain the information reflecting the drivers’ decision.

A closely related topic to the present study is predicting driver decision (i.e. stop or proceed) when the traffic signal turns to yellow from green as studied in [35, 36]. However, not all decisions to proceed lead to red light violations (i.e. passing the intersection during yellow time). The present study focuses on developing prediction models aiming at identifying red light running (RLR) violations.

Therefore, to summarize, the objectives of the present study is threefold: (1) Creating several factors and using a factor selection method, namely mRMR, to select the most useful factors. (2)

Determining an appropriate time window corresponding to the onset of yellow light to capture the information that reflects drivers' decision. (3) Investigating how SVM and RF methods can predict RLR violations before they occur proving enough time for endangered drivers to respond.

The remainder of the paper is organized as follows. The first section describes the data collection process. The second section explains the model development that includes the machine learning methods that were employed, the selected time window, as well as the adopted factor selection method. The results are presented in the third section. Finally, the conclusion is given in the fourth section.

Data Collection

The naturalistic data used in this research came from the Cooperative Intersection Collision Avoidance Systems for Violations (CICAS-V) project. As part of the CICAS-V project, different equipment such as radars, video cameras, Signal phase sniffer, and etc. were included in the data acquisition systems (DASs) that was designed and developed by the Center for Technology Development (CTD) at the Virginia Tech Transportation Institute (VTTI). The DAS equipment were installed at six stop-controlled intersections and three signalized intersections in the New River Valley area of southwest Virginia, of which a data sample from one of the signalized intersections (the intersection of Franklin Street and Depot Street) was used in the present study. The details of data collection are presented in [37]. The sample used in this study includes about 500 observations for the violation behavior and 500 observations reflecting the compliant behavior. For each individual vehicle approaching an intersection data such as speed, acceleration, distance to intersection (DTI), signal setting information, and etc. were collected at high resolution.

Model development

Methods

Machine learning methods namely Random Forest (RF) and Support Vector Machine (SVM) were applied to predict RLR violations. Random Forest method, as proposed in 2001 [38], is similar to the decision tree method; instead of one single tree, RF uses an ensemble of decision trees from which a majority vote makes the predictions. Two model parameters, namely the number of decision trees and the number of variables (or factors) to use in each tree, should be determined in order to apply RF method. To construct each tree, Recursive Binary Splitting method in which factors are selected to divide the data into different parts, was adopted. Different criteria may be used to determine how to separate the data. The recommended criteria are the Gini index and Cross-Entropy. The Cross-Entropy criteria was used in this study as presented in Equation 21 [39].

$$G = \sum_{k=1}^{K} P_k^m \log P_k^m \qquad \text{Equation 21}$$

Where,

$$P_k^m = \frac{1}{N^m} \sum_{x_i^m} I(y_i^m = k)$$

N^m	Number of observations received at node m
y_i^m	The response value corresponding to the observation i at node m
x_i^m	The factor vector corresponding to the observation i at node m
k	class

Support Vector Machine (SVM) is a relatively complex method that is used for solving both regression and classification problems. In a classification problem, which is the case in the present study, this method tries to find the best possible decision boundary between different classes. This corresponds to determining the boundary that largest possible gap between different classes is obtained and thus a good generalization result is achieved. The formulation of the SVM method is shown in Equation 22 which presents minimization of two parts; the first part corresponds to maximizing the gap between classes and the second part represents some error terms that are controlled by a regularization parameter denoted by C. The regularization parameter (sometimes referred to as the penalty) deals with the issue of over-fitting and determines the relative importance between the two terms in the objective function. The constraints of the objective function are presented in Equation 23 and Equation 24 [40].

$$\min_{w,b,\xi}\left(\frac{1}{2}w^T w + C\sum_{n=1}^{N}\xi_n\right) \qquad \text{Equation 22}$$

Subject to:

$$y_n(w^T\phi(x_n) + b) \geq 1 - \xi_n\,, n = 1, \dots, N \qquad \text{Equation 23}$$

$$\xi_n \geq 0\,, n = 1, \dots, N \qquad \text{Equation 24}$$

Where,

w	Parameters to define decision boundary between classes
C	Regularization (or penalty) parameter
ξ_n	Error parameter to denote margin violation
b	Intercept associated with decision boundaries
$\phi(x_n)$	Function to transform data from X space into some Z space

When solving an SVM problem, the function $\phi(x_n)$ as shown in Equation 23, transfers data from the current X space into a higher dimensional Z space where data separation becomes an easier task. The vector inner product in the Z space, known as kernel, is an important term which appears in the dual formulation of the problem. Different types of kernels exist such as linear kernel, polynomial kernels, and Gaussian kernel. According to our data size and factor size, Gaussian kernel was believed to be the most appropriate kernel [41]. In this paper, the $\phi(x_n)$ function which corresponds to the Gaussian kernel has an infinite dimensional space. The formulation of the Gaussian kernel is shown in Equation 25.

$$K(x, x') = exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) \qquad \text{Equation 25}$$

Where,

x, x'	n-dimensional vectors
$\|x - x'\|$	Euclidean distance between vectors x, x'
σ	Gaussian parameter

Time window corresponding to the yellow onset

As discussed earlier, a time window needs to be examined to extract the information reflecting driver behavior when approaching a signalized intersection. Figure 16 illustrates a signalized intersection that shows two vehicles: (1) violator vehicle: that presumably is going to violate the red light (2) endangered vehicle: that is going to be at risk of a right-angle crash due to the violator vehicle's RLR violation. The time window (t^v_{mon}) that needs be monitored to predict RLR violation of the violator vehicle should include the time at which the yellow light starts (yellow may start anytime within $t^v_{yellow\ onset}$). However,

some considerations needed to be accounted. Thus, this time window (i.e. t^v_{mon}) was determined by appropriately choosing its starting and ending points; the starting point of the time window should not be selected too early in order to exclude unnecessary information (i.e. when the drivers are very far from the intersection, their behavior may not reflect their decision on violating the red light). In fact, the behavior of the drivers before the yellow onset may not be related to their decision whether to stop or go. Figure 17 shows different approaching vehicles with the corresponding time to intersections at yellow onset. As the data suggests only few vehicles had the TTI of greater than 6 seconds at the onset of yellow light. Therefore, the TTI of 6 seconds was chosen as the starting point of t^v_{mon}. Although it is desirable to choose the ending point so that TTI of all vehicles at the yellow onset would be greater than the ending point, the ending point is restricted by t^v_{min} which is the minimum time required for the endangered vehicle to respond if a possible collision is predicted. In fact, t^v_{min} is equivalent to the sum of two terms: the time required for the driver to respond (t^e_{driver}) and the time required for the vehicle to stop ($t^e_{vehicle}$). The goal is to provide enough time for the endangered vehicle to avoid the collision. As mentioned earlier, using a distribution of the human response time presented in [34], 1.6 seconds was chosen as the t^e_{driver}, which is sufficient for 80% of the population to react. The $t^e_{vehicle}$ was obtained based on the equations of motion and the assumptions made in calculating the stopping sight distance [42]. The value of 1.9 seconds was obtained for $t^e_{vehicle}$ with regard to the design speed of 35 mph. As a result, sum of the two terms (t^e_{driver} and $t^e_{vehicle}$) was found to be 3.5 seconds that was essentially dictated the ending point of the t^v_{mon} to be the TTI of 3.5 seconds leading to the monitoring time window (t^v_{mon}) between the TTI of 3.5 and 6 seconds[4]. Accordingly, the prediction needs to be carried out within this time window.

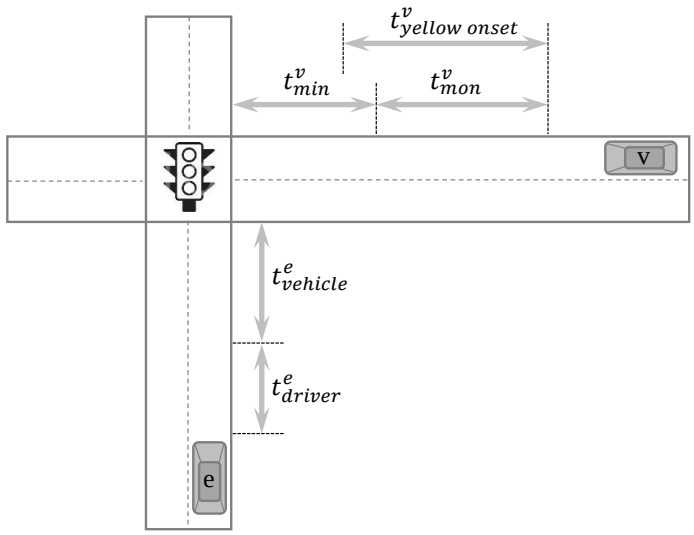


Figure 16: determination of the monitoring time window; vehicle v as the RLR violator and vehicle e as the endangered vehicle

[4] See Appendix C for sensitivity analysis result when changing the monitoring period

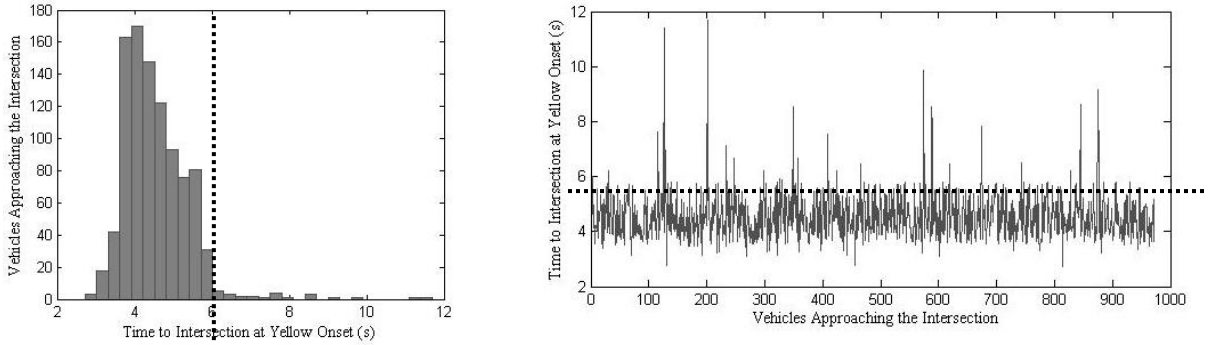


Figure 17: Time to Intersection at the Onset of Yellow Light

Factor selection

In model development process, those factors that can provide useful information for prediction need to be identified. Advantages of an appropriate factor selection include reducing the dimensionality of the problem, reducing the noise, identifying more important factors, and obtaining more interpretable factors. In this paper the minimum redundancy maximum relevance (mRMR) approach was adopted to choose the most representative factors. This approach attempts to find those factors that have the highest level of relevance and at the same time selects those factors that minimizes the redundancies between them [43].

The vehicles' kinetic information obtained from the data collection process includes velocity, acceleration, time to intersection (TTI), and distance to intersection (DTI). As discussed in the previous section, the time window to be examined for each vehicle contains the information between TTI of 3.5 and 6 seconds. Other than the information at the yellow onset, some statistical measures of dispersion were applied to describe changes in the defined time window to capture the driver behavior since it is expected that any change in driver behavior (due to the drivers' decisions to stop/go) can directly affect the kinetic information within the defined time window. Table 19 presents the list of factors obtained or created based on the acquired data. It should be pointed out that the Required Deceleration Parameter (RDP) is the deceleration value required for a vehicle to be able to stop at the stop bar and was obtained through Equation 26 as follows [44].

$$RDP = \frac{V^2}{2.DTI.g} \qquad \text{Equation 26}$$

Where,

V	Vehicle's instantaneous velocity
DTI	Distance to Intersection
g	Gravitational constant

Table 19 List of the examined factors

No.	Factor	No.	Factor
1	Distance to Intersection (DTI) at Onset of Yellow	10	$std(Velocity)$ over the t_{mon}^{v}
2	Velocity at Onset of Yellow	11	$mean(Acceleraiton)$ over the t_{mon}^{v}
3	Acceleration at Onset of Yellow	12	$range(Acceleraiton)$ over the t_{mon}^{v}
4	Time to Intersection (TTI) at Onset of Yellow	13	$max(Acceleraiton)$ over the t_{mon}^{v}
5	Required Deceleration Parameter (RDP) at Onset of Yellow	14	$min(Acceleraiton)$ over the t_{mon}^{v}

6	$mean(Velocity)$ over the t_{mon}^{v}	15	$std(Acceleraiton)$ over the t_{mon}^{v}
7	$range(Velocity)$ over the t_{mon}^{v}	16	$mean(DTI)$
8	$max(Velocity)$ over the t_{mon}^{v}	17	$mean(TTI)$
9	$min(Velocity)$ over the t_{mon}^{v}		

Results

To assess the performance of the SVM and the RF models 5-fold cross-validation accuracy and the out-of-bag (OOB) error were used, respectively. In order to apply K-fold cross validation, the data is divided into K parts. The model is developed based on K-1 part and evaluated based on the remaining untouched part. The procedure is repeated K times, each time with a different part as the test set and the remaining parts as the training set. Subsequently, the final result is averaged over the results obtained from the K models [45, 46]. When evaluating tree based models such as Bagging and Random Forest, there is no need to use K-fold cross-validation as the unbiased estimation of the error, namely the Out-Of-Bag (OOB) error, is obtained internally and is almost identical to the cross-validation accuracy [39].

As stated earlier, mRMR method was applied to use the most useful factors. The number of used factors was chosen to be 5 by experimenting different numbers. It was tried to choose a number that resulted in good performance. However, it was avoided to use many factors to focus on the most important ones. These 5 factors were found to be the Distance to Intersection (DTI) at the onset of yellow, mean(Velocity) over the t_{mon}^{v}, Required Deceleration Parameter (RDP) at the onset of yellow, max(Acceleraiton) over the t_{mon}^{v}, and min(Velocity) over the t_{mon}^{v}. Therefore, using these 5 factors the SVM and the RF models were developed.

To implement RF method, the R software and RandomForest packages were used [47, 48]. When developing the RF model, different number of trees was examined as shown in Figure 18. Increasing the number of trees beyond 100 did not seem to be beneficial. However, since increasing number of trees does not create problems such as over-fitting the value of 500 was selected to use to make sure sufficient number of trees was applied. Another parameter that needed to be determined was the number of features that each tree requires to grow. As Figure 19 illustrates, different number of factors resulted in different error rates and thus the value of 3 was chosen for this parameter as it led to the lowest error rate. To implement SVM algorithm, the LibSVM library of SVMs was applied [49]. To construct the SVM model, a Gaussian kernel was used with complete model selection that entails consideration of the regularization parameter and the Gaussian parameter to obtain the best possible performance. The complete model selection of the SVM model is demonstrated in Figure 20. The parameter $Gamma$ shown in this figure is equivalent to $\frac{1}{2\sigma^2}$ that is a part of the Gaussian kernel formulation in Equation 25. Furthermore, factor scaling was carried out in both the RF and the SVM models to normalize the factors.

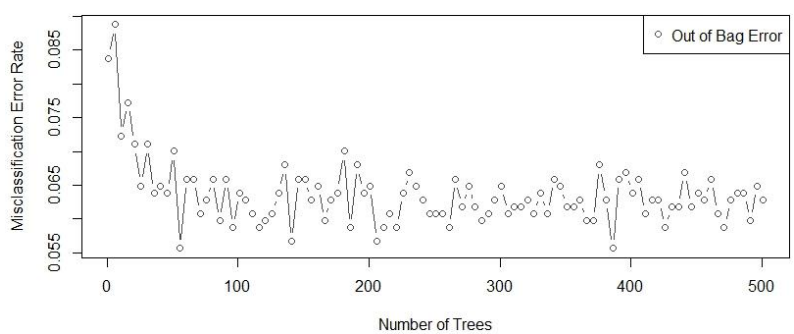


Figure 18: Selecting the number of trees - RF model

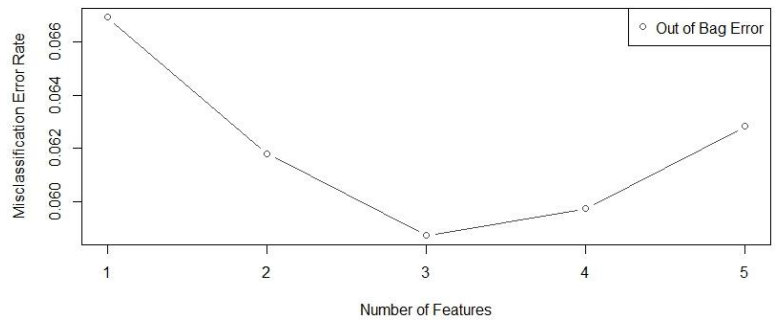


Figure 19: Selecting the number of factors for each tree - RF model

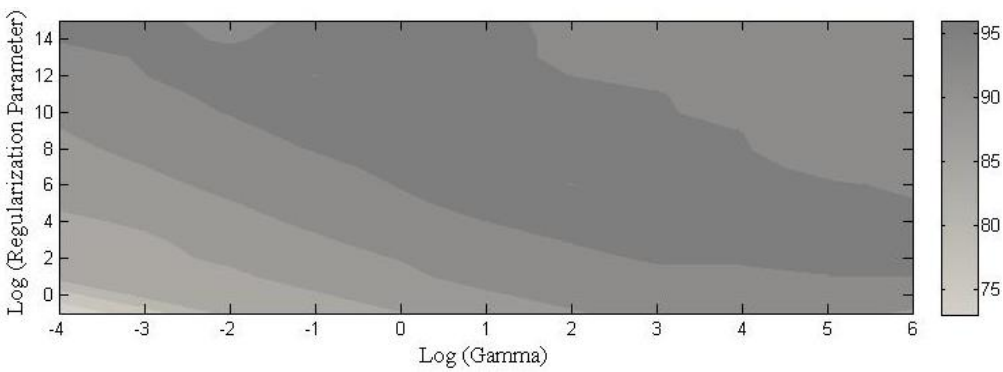


Figure 20: Conducting the model selection - SVM model

Both the SVM model and the RF model resulted in high accuracies of about 96.7% and 94.2%, respectively. As discussed throughout the paper, several considerations (i.e. determining an appropriate time window) as well as some tuning parameters needed to be determined which all contributed to the high prediction accuracies. The mRMR method identified 5 factors as the most useful ones; two of them (i.e. the Distance to Intersection (DTI) at Onset of Yellow and the Required Deceleration Parameter (RDP) at Onset of Yellow) were related to a point in time which is the instant at which the traffic light changes from green to yellow. Hence, this point in time plays a critical role in predicting the RLR violations. Moreover, the other three factors (i.e. $mean(Velocity)$ over the t_{mon}^{v}, $max(Acceleraiton)$ over the t_{mon}^{v}, and $min(Velocity)$ over the t_{mon}^{v}) described some statistical quantities over the time window that was determined earlier. This shows that other than the yellow onset, an appropriate time window can provide useful information reflecting the driver behavior when approaching a signalized intersection.

Conclusion

Prediction models were developed to identify RLR violations before they occur to provide sufficient time for an endangered driver to respond. This was determined based on three points: (1) the time required for the endangered driver to react, which is equivalent to the perception reaction time (2) the time required for the endangered driver to stop at the stop bar, which was obtained with respect to the stopping sight distance and equations of motion and (3) the information before the yellow onset was excluded since it was expected that the driver decision is only affected after the traffic light changes to yellow. A total of 17 factors were examined, of which 5 were identified using the mRMR method as the most representative factors. These factors included the Distance to Intersection (DTI) at the onset of yellow, $mean(Velocity)$ over the $\mathrm{t}_{\mathrm{mon}}^{\mathrm{v}}$, the Required Deceleration Parameter (RDP) at the onset of yellow, $max(Acceleraiton)$ over the t_{mon}^{v}, and the $min(Velocity)$ over the t_{mon}^{v} in the order of importance. This showed that both the critical time instant (e.g. the yellow onset) and the monitoring time window provide essential

information for predicting RLR violations. Machine learning methods, namely Support Vector Machine (SVM) and Random Forest (RF), were employed to construct the prediction models. Using Out-of-Bag (OOB) error, the RF model produced a classification accuracy of 94.2 percent. The tuning parameters for the RF model (e.g. the number of trees and the number of factors used in each tree) were also determined to obtain the best possible OOB error. In case of the SVM model, 5-fold cross validation was adopted to conduct a complete model selection (i.e. accounting for both the regularization and the Gaussian parameter) which resulted in a classification accuracy of 96.7 percent.

Acknowledgements

This research effort was funded by the Connected Vehicle Initiative UTC (CVI-UTC).

References

[1] National Highway Traffic Safety Administration, *Traffic safety facts 2012*. 2014, National Center for Statistics and Analysis, US Department of Transportation, Washington, DC.

[2] Insurance Institute for Highway Safety (IIHS). *Red light running*. Available from: http://www.iihs.org/iihs/topics/t/red-light-running/topicoverview.

[3] Insurance Institute for Highway Safety (IIHS), *Status Report: Public seeks safer roads but still takes risks*. 2010.

[4] Rakha, H., I. El-Shawarby, and J.R. Setti, *Characterizing driver behavior on signalized intersection approaches at the onset of a yellow-phase trigger.* Intelligent Transportation Systems, IEEE Transactions on, 2007. **8**(4): p. 630-640.

[5] Gazis, D., R. Herman, and A. Maradudin, *The problem of the amber signal light in traffic flow.* Operations Research, 1960. **8**(1): p. 112-132.

[6] Sheffi, Y. and H. Mahmassani, *A model of driver behavior at high speed signalized intersections.* Transportation Science, 1981. **15**(1): p. 50-61.

[7] Bonneson, J.A., et al., *Intelligent detection-control system for rural signalized intersections*. 2002, Texas Transportation Institute, Texas A&M University System.

[8] Gates, T.J., et al., *Analysis of driver behavior in dilemma zones at signalized intersections.* Transportation Research Record: Journal of the Transportation Research Board, 2007. **2030**(1): p. 29-39.

[9] Pant, P.D., et al., *Field testing and implementation of dilemma zone protection and signal coordination at closely-spaced high-speed intersections*. 2005, University of Cincinnati.

[10] Chang, M.-S., C.J. Messer, and A.J. Santiago, *Timing traffic signal change intervals based on driver behavior*. 1985.

[11] Zegeer, C.V., *GREEN-EXTENSION SYSTEMS AT EHGI-I-SPEED INTERSECTIONS.* 1978.

[12] Liu, Y., et al., *Empirical observations of dynamic dilemma zones at signalized intersections.* Transportation Research Record: Journal of the Transportation Research Board, 2007. **2035**(1): p. 122-133.

[13] Wei, H., et al., *Quantifying Dynamic Factors Contributing to Dilemma Zone at High-Speed Signalized Intersections.* Transportation Research Record: Journal of the Transportation Research Board, 2011. **2259**(1): p. 202-212.

[14] Ghanipoor Machiani, S. and M. Abbas. *Dynamic Driver's Perception of Dilemma Zone: Experimental Design and Analysis of Driver's Learning in a Simulator Study*. in *The 93nd Annual Meeting of the Transportation Research Board*. 2014. Washington, DC.

[15] Caird, J.K., et al., *The effect of yellow light onset time on older and younger drivers' perception response time (PRT) and intersection behavior.* Transportation research part F: traffic psychology and behaviour, 2007. **10**(5): p. 383-396.

[16] El-Shawarby, I., et al. *Age and gender impact on driver behavior at the onset of a yellow phase on high-speed signalized intersection approaches*. in *Transportation Research Board 86th Annual Meeting*. 2007.

[17] Li, H., H. Rakha, and I. El-Shawarby, *Designing Yellow Intervals for Rainy and Wet Roadway Conditions.* International Journal of Transportation Science and Technology, 2012. **1**(2): p. 171-190.

[18] Mussa, R.N., et al., *Simulator evaluation of green and flashing amber signal phasing.* Transportation Research Record: Journal of the Transportation Research Board, 1996. **1550**(1): p. 23-29.

[19] Amer, A., H. Rakha, and I. El-Shawarby, *Agent-based stochastic modeling of driver decision at onset of yellow light at signalized intersections.* Transportation Research Record: Journal of the Transportation Research Board, 2011. **2241**(1): p. 68-77.

[20] Porter, B.E. and K.J. England, *Predicting red-light running behavior: a traffic safety study in three urban settings.* Journal of Safety Research, 2000. **31**(1): p. 1-8.

[21] Retting, R.A. and A.F. Williams, *Characteristics of red light violators: results of a field investigation.* Journal of Safety Research, 1996. **27**(1): p. 9-15.

[22] Porter, B.E. and T.D. Berry, *A nationwide survey of self-reported red light running: measuring prevalence, predictors, and perceived consequences.* Accident Analysis & Prevention, 2001. **33**(6): p. 735-741.

[23] Bonneson, J.A. and H.J. Son, *Prediction of expected red-light-running frequency at urban intersections.* Transportation Research Record: Journal of the Transportation Research Board, 2003. **1830**(1): p. 38-47.

[24] Dingus, T.A., et al., *The 100-car naturalistic driving study, Phase II-results of the 100-car field experiment*. 2006.

[25] Retting, R.A., S.A. Ferguson, and C.M. Farmer, *Reducing red light running through longer yellow signal timing and red light camera enforcement: results of a field investigation.* Accident Analysis & Prevention, 2008. **40**(1): p. 327-333.

[26] Liang, Y., M.L. Reyes, and J.D. Lee, *Real-time detection of driver cognitive distraction using support vector machines.* Intelligent Transportation Systems, IEEE Transactions on, 2007. **8**(2): p. 340-350.

[27] Kim, Z., *Robust lane detection and tracking in challenging scenarios.* Intelligent Transportation Systems, IEEE Transactions on, 2008. **9**(1): p. 16-26.

[28] Jahangiri, A. and H. Rakha. *Developing a Support Vector Machine (SVM) Classifier for Transportation Mode Identification by Using Mobile Phone Sensor Data*. in *Transportation Research Board 93rd Annual Meeting*. 2014.

[29] Balali, V. and M. Golparvar-Fard. *Video-Based Detection and Classification of US Traffic Signs and Mile Markers using Color Candidate Extraction and Feature-Based Recognition*. in *Computing in Civil and Building Engineering (2014)*. ASCE.

[30] Yuan, F. and R.L. Cheu, *Incident detection using support vector machines.* Transportation Research Part C: Emerging Technologies, 2003. **11**(3): p. 309-328.

[31] Aoude, G.S., et al., *Driver behavior classification at intersections and validation on large naturalistic data set.* Intelligent Transportation Systems, IEEE Transactions on, 2012. **13**(2): p. 724-736.

[32] Elmitiny, N., et al., *Classification analysis of driver's stop/go decision and red-light running violation.* Accident Analysis & Prevention, 2010. **42**(1): p. 101-111.

[33] Zhang, L., et al., *Prediction of red light running based on statistics of discrete point sensors.* Transportation Research Record: Journal of the Transportation Research Board, 2009. **2128**(1): p. 132-142.

[34] McLaughlin, S.B., J.M. Hankey, and T.A. Dingus, *A method for evaluating collision avoidance systems using naturalistic driving data.* Accident Analysis & Prevention, 2008. **40**(1): p. 8-16.

[35] Elhenawy, M., H. Rakha, and I. El-Shawarby. *Enhancing Driver Stop/Run Modeling at the Onset of a Yellow Indication using Historical Behavior and Machine Learning Techniques*. in *Transportation Research Board 93rd Annual Meeting*. 2014.

[36] Ghanipoor Machiani, S. and M. Abbas. *Predicting Drivers Decision in Dilemma Zone in a Driving Simulator Environment using Canonical Discriminant Analysis*. in *The 93nd Annual Meeting of the Transportation Research Board*. 2014. Washington, DC.

[37] Doerzaph, Z.R. and V. Neale. *Data acquisition method for developing crash avoidance algorithms through innovative roadside data collection*. in *Transportation Research Board 89th Annual Meeting*. 2010.

[38] Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.

[39] Hastie, T., et al., *The elements of statistical learning*. Vol. 2. 2009: Springer.

[40] Hsu, C.-W. and C.-J. Lin, *A comparison of methods for multiclass support vector machines.* Neural Networks, IEEE Transactions on, 2002. **13**(2): p. 415-425.

[41] Hsu, C.-W., C.-C. Chang, and C.-J. Lin, *A practical guide to support vector classification*. 2003.

[42] Layton, R. and K. Dixon, *Stopping Sight Distance.* Kiewit Center for Infrastructure and Transportation, Oregon Department of Transportation, 2012.

[43] Ding, C. and H. Peng, *Minimum redundancy feature selection from microarray gene expression data.* Journal of bioinformatics and computational biology, 2005. **3**(02): p. 185-205.

[44] Doerzaph, Z.R., V. Neale, and R. Kiefer. *Cooperative intersection collision avoidance for violations: threat assessment algorithm development and evaluation method*. in *Transportation Research Board 89th Annual Meeting*. 2010.

[45] Kohavi, R. *A study of cross-validation and bootstrap for accuracy estimation and model selection*. in *IJCAI*. 1995.

[46] James, G., et al., *An introduction to statistical learning*. 2013: Springer.

[47] R Core Team, *R: A Language and Environment for Statistical Computing*. 2014, R Foundation for Statistical Computing.

[48] Liaw, A. and M. Wiener, *Classification and Regression by randomForest.* R news, 2002. **2**(3): p. 18-22.

[49] Chang, C.-C. and C.-J. Lin, *LIBSVM: a library for support vector machines.* ACM Transactions on Intelligent Systems and Technology (TIST), 2011. **2**(3): p. 27.

Adopting Machine Learning Methods to Predict Red-light Running Violations

Arash Jahangiri, Hesham A. Rakha, and Thomas A. Dingus

Abstract

Statistics demonstrate that a large number of crashes occur at signalized intersections due to traffic violations, specifically red light running (RLR). In order to prevent/mitigate intersection-related crashes, these violations need to be identified before they occur, so appropriate actions can be taken. Several factors such as vehicle speed, Time to Intersection (TTI), Distance to Intersection (DTI), age, gender, etc. influence the drivers behavior when approaching intersections. However, the driver-related factors (i.e. age, gender) are more difficult to obtain in practice. On the other hand, kinetic factors (e.g. speed, acceleration) can be obtained by monitoring the movement of vehicles through video cameras installed on the infrastructure or through on-board devices installed on the vehicles. Hence, the problem of interest is to develop models to predict RLR violations using kinetic information of vehicles. A monitoring period was defined to extract data from each vehicle before reaching the intersection. Machine learning techniques, namely Support Vector Machine (SVM) and Random Forest (RF), were adopted to develop prediction models. The minimum Redundancy Maximum Relevance (mRMR) feature selection method was used to identify the most important factors for model development. To evaluate the performance of the models the *K*-fold cross-validation and out-of-bag (OOB) errors were used for the SVM and RF models, which contributed to high prediction accuracies of 97.9 and 93.6 percent, respectively. It was shown that other than the critical instant at which the traffic signal changes to yellow, an appropriate monitoring period with respect to the yellow onset can provide additional useful information ensuring that the driver decision occurs during that period.

Keywords— driver violation; red light running; signalized intersection; violation prediction; support vector machine; random forest; machine learning

Introduction

In 2012, more than 2.5 million intersection related crashes occurred resulted in 2,850 fatal crashes and 680,000 injury crashes [1]. Also, RLR violation has been reported as a serious issue causing intersection related crashes; the Insurance Institute for Highway Safety reported that in the US in 2012, running red lights led to 683 fatalities and an estimated 133,000 injured [2].

A driver is considered a RLR violator when he passes the stop bar at a signalized intersection while the traffic light is red. The area in which the driver needs to make a decision as to stopping or proceeding is known as the dilemma zone which is a classic problem, first introduced in [3] and have been studied by many researchers [4-13]. Several factors that influence the driver behavior when approaching a signalized intersection have been studied in the literature [6, 7, 11-16]. These factors include: the driver perception-reaction time; the driver's acceptable deceleration rate; the driver's age; the driver's gender; TTI at the onset of yellow; DTI at the onset of yellow; approach speed; vehicle type; presence of side-street vehicles, pedestrians, bicycles, or opposing vehicles waiting to turn left; flow rate; length of yellow interval; cycle length; presence of police; and pavement condition.

The factors that can be examined are limited depending on the data collection method; some studies such as [13, 14, 17] applied driving simulators which are suitable to construct hypothetical scenarios in which many factors can be examined. However, driver behavior may not be as natural compared to the real world situations. When using video cameras to collect data as adopted in [6, 11, 12], the driver behavior is completely natural as the drivers do not realize that they are in an experiment. Nevertheless, it may not be feasible to collect some driver related factors (e.g. age, gender) through video cameras. Moreover, it may not be possible to study some other factors such as the presence of bicycles or police since these situations may not occur at all during the data collection. Some studies collected the data using test tracks [7, 15, 16, 18]. It appears that when using test tracks, the driver behavior is more natural compared to using driving simulators, but still since the drivers know that they are in an experiment, they may adjust their behavior.

The present study focuses on developing RLR violation prediction models using a naturalistic data set (i.e. data collected through video cameras). A monitoring period was defined in space (this period was defined in time in our previous work [19]) that corresponds to the location where the driver makes the decision to stop or go. The particular RLR violation scenario is when no vehicle is waiting at the intersection, the traffic light changes from green to yellow for a vehicle approaching the intersection. The vehicle would fail to cross the stop bar before the red light starts, and thus RLR violation would occur. To summarize, the objectives of the

*Research funded by the Connected Vehicle Initiative UTC (CVI-UTC). This work was carried out by Virginia Tech Transportation Institute, Virginia, USA.

Arash Jahangiri is a PhD candidate with the Civil and Environmental Engineering Department at Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA (e-mail:Arashj@vt.edu).

Hesham A. Rakha (corresponding author) is a Professor with the Civil and Environmental Engineering Department at Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA (Ph: 540-231-1505, email: hrakha@vt.edu).

Thomas A. Dingus is a Professor with the Civil and Environmental Engineering Department at Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA (Ph: 540-231-1501, email: tdingus@vt.edu).

present study is threefold: (1) Using a feature selection method, namely mRMR, to select the most useful features. (2) Determining an appropriate monitoring period corresponding to the onset of yellow light to capture the information that reflects drivers' decision. (3) Investigating how SVM and RF methods can predict RLR violations before they occur proving enough time for endangered drivers to respond.

The remainder of the paper is organized as follows. Section II reviews relevant studies in the literature. Section III presents the data collection process. Section IV explains the model development that includes the machine learning methods that were employed, the selected monitoring period, as well as the adopted feature selection method. The results are presented in section V. Finally, the conclusion is given in section VI.

Background

Some studies investigated some factors that affects the probability of running red lights [20-25]. For example, drivers who do not use safety belts, non-Caucasian drivers, larger intersections, and higher traffic volumes increases the probability of RLR violations [20]. Moreover, estimation models were developed in some studies; using factors such as flow rate, cycle length, yellow duration, a regression model was developed in [23] to estimate the frequency of RLR violations. Some other studies as in [26, 27] concentrated on modeling stop/run behavior at the yellow onset, which is basically predicting the driver decision to stop or go. However, not all decisions to go results in RLR violations (i.e. passing during yellow time).

There is a growing trend in applying Artificial Intelligent (AI) methods that are capable of developing prediction models with high performance; some examples are: Real-time detection of driver cognitive distraction [28], lane detection and tracking [29] transportation mode recognition [30, 31], traffic sign detection [32], and Incident detection [33]. AI methods have been used to predict RLR violations at signalized intersections [34, 35]. However, more research is needed to enhance current and traditional models. Other than AI tools, statistical and probabilistic approaches have also been applied [4, 36]. For example, a probabilistic model was developed in [36] to predict RLR violations, taking into consideration minimizing both false alarm rate and missing error.

Limited studies focused on RLR violation predictions using AI techniques: To predict RLR violations a classification tree model was applied in [35] and the vehicle's distance at onset of yellow, operating speed at the onset of yellow, and position in traffic flow were found to be the most important factors (or features). No feature selection method was applied to select the most useful features. However, in order to examine several features (factors), the ability to select the most relevant features can improve the performance of the model. Furthermore, they did not use a monitoring period. It appears that the drivers decision to stop or go are made within a short monitoring period instead of in an instant, and thus features obtained from this time period should also be considered. In another study [34], SVM and Hidden Markov Model (HMM) were applied to predict RLR violations. Features including DTI, speed, acceleration, TTI, and required deceleration parameter (RDP) were found to be the most useful features. However, these features were identified by experimenting on different combinations of features without using any feature selection methods. They also applied a monitoring period to extract some features for developing prediction models. This period was defined with respect to a minimum time threshold based on human perception reaction time. However, to avoid a potential collision a vehicle response time should also be considered. In addition, it appears that the yellow onset was not considered as a factor. Nevertheless, onset of yellow is a critical point in time that affects the driver decision to stop or proceed and needs to be included in model development.

Data Description

The data used in this paper is from the Cooperative Intersection Collision Avoidance Systems for Violations (CICAS-V) project. In the CICAS-V project, data acquisition systems (DASs) ,developed by the Center for Technology Development (CTD) at the Virginia Tech Transportation Institute (VTTI), along with other equipment such as radars, video cameras, Signal phase sniffer, etc. were installed at different intersections in the New River Valley area of southwest Virginia that resulted in a large data set of vehicle trajectories approaching the intersections. In the present study, a data sample from one of the signalized intersections (the intersection of Franklin Street and Depot Street) was used that includes about 500 observations for the violation behavior and 500 observations for the compliant behavior. Different features such as speed, acceleration, distance to intersection (DTI), signal setting information, and etc. were extracted for each individual vehicle in the data sample. A more detailed explanation of the data collection process is presented in [37].

Model Development

Methods

Machine learning methods namely Random Forest (RF) and Support Vector Machine (SVM) were employed to model RLR violations. RF method, as proposed in 2001 [38], is a learning algorithm that takes advantage of averaging multiple learners. In fact, RF obtains the results from an ensemble of decision trees and subsequently, a majority vote would contribute to the final result. To apply RF method, the number of decision trees and the number of variables (or factors) to use in each tree are the two parameters that need to be determined. Each decision tree is built by splitting the data on several stages using the Recursive Binary Splitting method. Data splitting is carried out based on different criteria, among which the Gini index and Cross-Entropy are the recommended criteria. The Gini index was applied in the present study as shown in (27) [39].

$$G = \sum_{k=1}^{K} P_k^m (1 - P_k^m) \tag{27}$$

Where,

$P_k^m = \frac{1}{N^m} \sum_{x_i^m} I(y_i^m = k)$

N^m Number of observations received at node m

y_i^m	The response value corresponding to the observation i at node m
x_i^m	The feature vector corresponding to the observation i at node m
k	class

SVM is a relatively complex method that is used for solving both regression and classification problems. In a classification problem, which is the case in the present study, this method tries to find the best decision boundary between different classes. This corresponds to determining the boundary that largest possible gap between different classes is obtained and thus a good generalization result is achieved. The formulation of the SVM method is shown in (28) which presents minimization of two parts; the first part corresponds to maximizing the gap between classes and the second part represents some error terms that are controlled by a regularization parameter denoted by C. The regularization parameter (sometimes referred to as the penalty) deals with the issue of over-fitting and determines the relative importance between the two parts in the objective function. The constraints of the objective function are presented in (29) and (30) [40].

$$\min_{w,b,\xi}\left(\frac{1}{2}w^T w + C\sum_{n=1}^{N}\xi_n\right) \quad (28)$$

Subject to:

$$y_n(w^T\phi(x_n) + b) \geq 1 - \xi_n\,, n = 1, \dots, N \quad (29)$$

$$\xi_n \geq 0\,, n = 1, \dots, N \quad (30)$$

Where,

w	Parameters to define decision boundary between classes
C	Regularization (or penalty) parameter
ξ_n	Error parameter to denote margin violation
b	Intercept associated with decision boundaries
$\phi(x_n)$	Function to transform data from X space into some Z space

When solving an SVM problem, the function $\phi(x_n)$ as shown in (29), transfers data from the current X space into a higher dimensional Z space where data separation becomes an easier task. The vector inner product in the Z space, known as kernel, is an important term which appears in the dual formulation of the problem. Different types of kernels exist such as linear kernel, polynomial kernels, and Gaussian kernel. Based on our data size and feature size, Gaussian kernel was identified as the most suitable kernel [41]. The formulation of the Gaussian kernel is shown in (31).

$$K(x, x') = exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) \quad (31)$$

Where,

x, x'	n-dimensional vectors
$\|x - x'\|$	Euclidean distance between vectors x, x'
σ	Gaussian parameter

Monitoring period corresponding to the yellow onset

As discussed earlier, a monitoring period needs to be examined to extract the information reflecting driver behavior when approaching a signalized intersection. Figure 21 illustrates a situation in which two vehicles are approaching a signalized intersection: (1) a violator vehicle that is going to violate the red light (2) an endangered vehicle that is going to be in a possible right-angle crash due to the RLR violation. The monitoring period (p^v_{mon}) should include the time at which the yellow light starts (yellow may start anytime within $t^v_{yellow\ onset}$). This period (i.e. p^v_{mon}) was determined by appropriately choosing its starting and ending points; the starting point should not be selected too early to exclude unnecessary information. In other words, before the yellow onset (i.e. When traffic light is green) the drivers do not consider whether to stop and thus their possible RLR violation behavior would occur after the yellow onset. Figure 22 shows different approaching vehicles with the corresponding DTI at yellow onset. As the data suggests, only few vehicles had the DTI of greater than 100 meters at the onset of yellow light. Therefore, the DTI of 100 meters was chosen as the starting point of p^v_{mon}. Although it is desirable to choose the ending point so that DTI of all vehicles at the yellow onset would be greater than the ending point, the ending point is restricted by t^v_{min} which is the minimum time required for the endangered vehicle to respond if a possible collision is predicted. In fact, t^v_{min} consists of two parts: the time required for the driver to respond (t^e_{driver}) and the time required for the vehicle to stop ($t^e_{vehicle}$). The goal is to provide enough time for the endangered vehicle to avoid the collision. Using a distribution of the human response time as presented in [42], 1.6 seconds was chosen as the t^e_{driver}, which is sufficient for 80% of the population to react. The $t^e_{vehicle}$ was obtained based on the equations of motion and the assumptions made in calculating the stopping sight distance [43]. The value of 1.9 seconds was obtained for $t^e_{vehicle}$ with regard to the design speed of 35 mph. Sum of the two terms (t^e_{driver} and $t^e_{vehicle}$) was found to be 3.5 seconds that was essentially the same as t^v_{min}. Since the p^v_{mon} was defined in space rather than time, the DTI for the corresponding t^v_{min}, denoted by d^v_{min}, was calculated assuming that the violator vehicle speed is 35 mph (i.e. the speed limit). Hence, d^v_{min} was found to be about 55 meters ($d^v_{min} = 35 * t^v_{min}$) which marked the end point of the monitoring period. This resulted in the monitoring period (p^v_{mon}) between the DTI of 55 and 100 meters. Accordingly, the prediction needs to be carried out within this monitoring period.

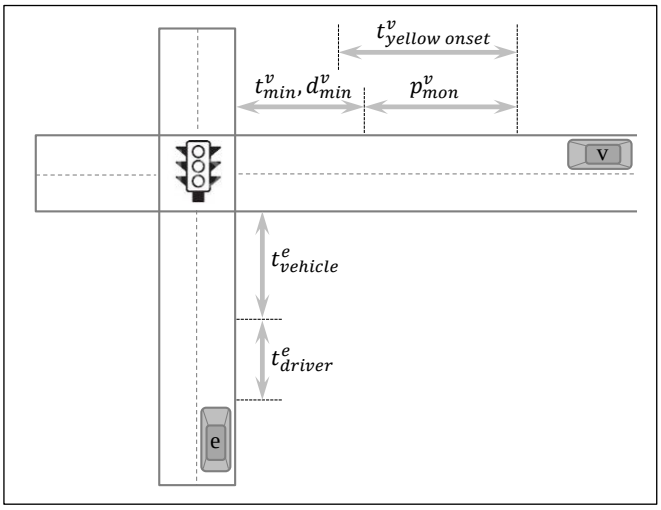


Figure 21 determination of the monitoring period; vehicle v as the RLR violator and vehicle e as the endangered vehicle

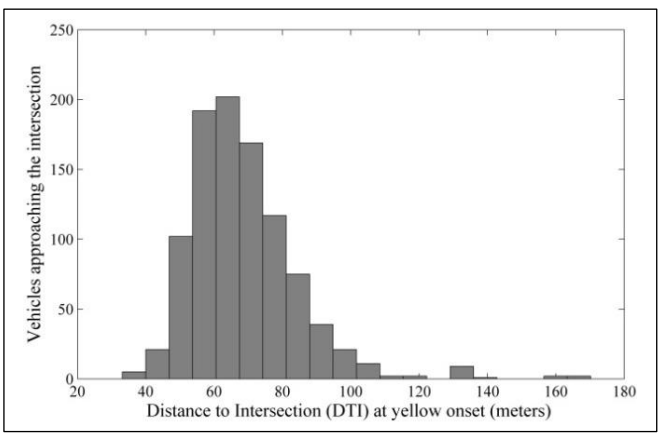


Figure 22 Distance to Intersection at the Onset of Yellow Light

Feature selection

Selecting the most useful features is considered as an important task in the model development process. Feature selection has several advantages: reducing the dimensionality of the problem; reducing the noise; identifying more important predictors (features); and obtaining more interpretable features. In this paper the minimum redundancy maximum relevance (mRMR) approach was adopted to choose the most representative features. This approach attempts to find features that have the highest level of relevance to the target value (i.e. the value that needs to be predicted; RLR violations in this case) and selects features that minimizes the redundancies between them [44].

The vehicles information obtained from the data collection process includes velocity, acceleration, time to intersection (TTI), and distance to intersection (DTI). As discussed in the previous section, we determined a monitoring period (i.e. between DTI of 55 and 100 meters at yellow onset) from which vehicle information was extracted. Other than the information at the yellow onset, some statistical measures of dispersion were applied to describe changes in the defined monitoring period to capture the driver behavior since it is expected that any changes in the driver behavior can affect the variables (e.g. speed, acceleration) within the defined monitoring period. Table 20 presents the list of features obtained or created based on the acquired data. The Required Deceleration Parameter in this table (RDP) represents the deceleration value required for a vehicle to be able to stop at the stop bar and was obtained through (32) as follows [45].

$$RDP = \frac{V^2}{2.DTI.g} \tag{32}$$

Where,

V	Vehicle's instantaneous velocity
DTI	Distance to Intersection
g	Gravitational constant

Table 20 List of the examined features

No.	Feature	No.	Feature
1	Distance to Intersection (DTI) at Onset of Yellow	10	$min(Velocity)$ over the t^v_{mon}
2	Velocity at Onset of Yellow	11	$mean(Acceleraiton)$ over the t^v_{mon}
3	Acceleration at Onset of Yellow	12	$range(Acceleraiton)$ over the t^v_{mon}
4	Time to Intersection (TTI) at Onset of Yellow	13	$std(Acceleraiton)$ over the t^v_{mon}
5	Required Deceleration Parameter (RDP) at Onset of Yellow	14	$max(Acceleraiton)$ over the t^v_{mon}
6	$mean(Velocity)$ over the t^v_{mon}	15	$min(Acceleraiton)$ over the t^v_{mon}
7	$range(Velocity)$ over the t^v_{mon}	16	$mean(DTI)$
8	$std(Velocity)$ over the t^v_{mon}	17	$mean(TTI)$
9	$max(Velocity)$ over the t^v_{mon}		

Results

To assess the performance of the SVM and the RF models 5-fold cross-validation accuracy and the OOB error were used, respectively. In order to apply K-fold cross validation, the data is divided into K parts. A model is developed based on K-1 part and evaluated based on the remaining untouched part. The procedure is repeated K times, each time with a different part as the test set and the remaining parts as the training set. Subsequently, the final result is averaged over the results obtained from the K models [46, 47]. When evaluating tree based models such as Bagging and RF, there is no need to use K-fold cross-validation as the unbiased estimation of the error, namely the OOB error, is obtained internally and is almost identical to the cross-validation accuracy [39].

As stated earlier, mRMR method was applied to use the most useful features. The number of features to use was chosen to be 5 by experimenting different numbers: It was tried to choose a number that resulted in good performance. However, it was avoided to use many features to focus on the most important ones. These 5 features were found to be the Distance to Intersection (DTI) at the onset of yellow, mean(Velocity) over the t^{v}_{mon}, max(Velocity) over the t^{v}_{mon}, std(Acceleration) over the t^{v}_{mon}, and Required Deceleration Parameter (RDP) at the onset of yellow. Therefore, using these 5 features the SVM and the RF models were developed.

The R software and the "RandomForest" package were used to implement the RF method [48, 49]. In model development process, total number of trees as well as number of features in each tree were varied to achieve the best possible accuracy as shown in Figure 23 and Figure 24; going beyond 100 trees did not decrease the error rate. However, given that adding more trees does not lead to over-fitting, 500 was selected as the number of trees to build the model. Also, the number of features used in each tree was selected to be 2 as this resulted in lowest error rate.

The LibSVM library of SVMs was adopted to implement SVM model [50]. To conduct a complete model selection, the regularization parameter and the Gaussian parameter were optimized as shown in Figure 25. The parameter Gamma in the figure is the same as the term $\frac{1}{2\sigma^2}$ presented in (31) which is a part of the Gaussian kernel formulation.

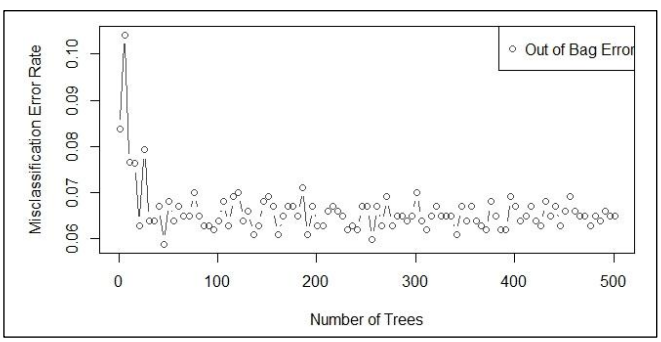


Figure 23 Selecting the number of trees - RF model

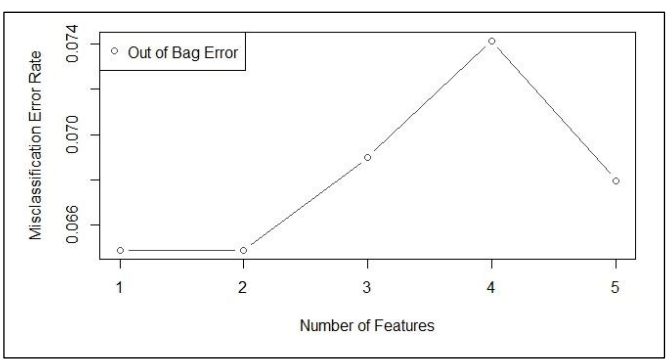


Figure 24 Selecting the number of features for each tree - RF model

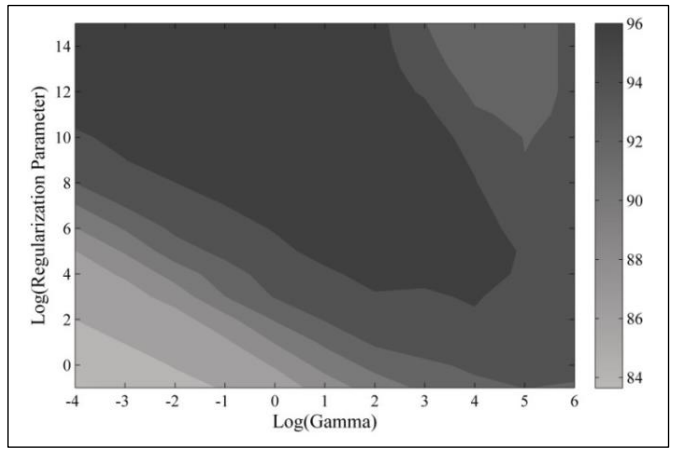


Figure 25 Conducting the model selection - SVM model

Both the SVM and the RF models resulted in high accuracies of about 97.9% and 93.6%, respectively. As discussed throughout the paper, several considerations (i.e. determining an appropriate monitoring period) as well as some tuning parameters needed to be determined which all contributed to the high prediction accuracies. The mRMR method identified 5 features as the most useful ones; two of them (i.e. the Distance to Intersection (DTI) at Onset of Yellow and the Required Deceleration Parameter (RDP) at Onset of Yellow) were variables at a point in time (i.e. onset of the yellow indication). Hence, this instant plays a critical role in predicting the RLR violations. Moreover, the other three important features (i.e. $mean(Velocity)$ over the t^{v}_{mon}, $max(Velocity)$ over the t^{v}_{mon}, and std(Acceleration) over the t^{v}_{mon}) described some statistical measures over the monitoring period. This shows that in addition to the yellow onset, an appropriate monitoring period can provide useful information to predict RLR violations.

Conclusion

RLR violation prediction models were developed in order to notify the endangered drivers of potential crashes, providing enough time for them to react. The models were developed based on three points: (1) the time required for the endangered driver to react, which is equivalent to the perception reaction time (2) the time required for the endangered driver to stop at the stop bar, which was obtained with respect to the stopping sight distance and equations of motion and (3) the monitoring period corresponding to the yellow onset. The information before the yellow onset was excluded since it was expected that the driver decision is only affected after the traffic light changes to yellow. A total of 17 features were examined, of which 5 were identified using the mRMR method as the most important features. These features included the Distance to Intersection (DTI) at the onset of yellow, mean(Velocity) over the t^{v}_{mon}, max(Velocity) over the t^{v}_{mon}, std(Acceleration) over the t^{v}_{mon}, and Required Deceleration Parameter (RDP) at

the onset of yellow in the order of importance. This showed that both the critical time instant (e.g. the yellow onset) and the monitoring period provide essential information for predicting RLR violations. To construct the prediction models, SVM and RF methods were employed. Using the OOB error, the RF model produced a classification accuracy of 93.6 percent. The tuning parameters for the RF model (e.g. the number of trees and the number of features used in each tree) were also determined to obtain the best possible OOB error. In case of the SVM model, 5-fold cross validation was adopted to conduct a complete model selection (i.e. accounting for both the regularization and the Gaussian parameter) which resulted in a classification accuracy of 97.9 percent.

Acknowledgment

This research effort was funded by the Connected Vehicle Initiative UTC (CVI-UTC).

References

[1] National Highway Traffic Safety Administration, *Traffic safety facts 2012*. 2014, National Center for Statistics and Analysis, US Department of Transportation, Washington, DC.

[2] Insurance Institute for Highway Safety (IIHS). *Red light running*. Available from: http://www.iihs.org/iihs/topics/t/red-light-running/topicoverview.

[3] Gazis, D., R. Herman, and A. Maradudin, *The problem of the amber signal light in traffic flow.* Operations Research, 1960. **8**(1): p. 112-132.

[4] Sheffi, Y. and H. Mahmassani, *A model of driver behavior at high speed signalized intersections.* Transportation Science, 1981. **15**(1): p. 50-61.

[5] Bonneson, J.A., et al., *Intelligent detection-control system for rural signalized intersections*. 2002, Texas Transportation Institute, Texas A&M University System.

[6] Gates, T.J., et al., *Analysis of driver behavior in dilemma zones at signalized intersections.* Transportation Research Record: Journal of the Transportation Research Board, 2007. **2030**(1): p. 29-39.

[7] Rakha, H., I. El-Shawarby, and J.R. Setti, *Characterizing driver behavior on signalized intersection approaches at the onset of a yellow-phase trigger.* Intelligent Transportation Systems, IEEE Transactions on, 2007. **8**(4): p. 630-640.

[8] Pant, P.D., et al., *Field testing and implementation of dilemma zone protection and signal coordination at closely-spaced high-speed intersections*. 2005, University of Cincinnati.

[9] Chang, M.-S., C.J. Messer, and A.J. Santiago, *Timing traffic signal change intervals based on driver behavior*. 1985.

[10] Zegeer, C.V., *GREEN-EXTENSION SYSTEMS AT EHGI-I-SPEED INTERSECTIONS.* 1978.

[11] Liu, Y., et al., *Empirical observations of dynamic dilemma zones at signalized intersections.* Transportation Research Record: Journal of the Transportation Research Board, 2007. **2035**(1): p. 122-133.

[12] Wei, H., et al., *Quantifying Dynamic Factors Contributing to Dilemma Zone at High-Speed Signalized Intersections.* Transportation Research Record: Journal of the Transportation Research Board, 2011. **2259**(1): p. 202-212.

[13] Ghanipoor Machiani, S. and M. Abbas. *Dynamic Driver's Perception of Dilemma Zone: Experimental Design and Analysis of Driver's Learning in a Simulator Study*. in *The 93nd Annual Meeting of the Transportation Research Board*. 2014. Washington, DC.

[14] Caird, J.K., et al., *The effect of yellow light onset time on older and younger drivers' perception response time (PRT) and intersection behavior.* Transportation research part F: traffic psychology and behaviour, 2007. **10**(5): p. 383-396.

[15] El-Shawarby, I., et al. *Age and gender impact on driver behavior at the onset of a yellow phase on high-speed signalized intersection approaches*. in *Transportation Research Board 86th Annual Meeting*. 2007.

[16] Li, H., H. Rakha, and I. El-Shawarby, *Designing Yellow Intervals for Rainy and Wet Roadway Conditions.* International Journal of Transportation Science and Technology, 2012. **1**(2): p. 171-190.

[17] Mussa, R.N., et al., *Simulator evaluation of green and flashing amber signal phasing.* Transportation Research Record: Journal of the Transportation Research Board, 1996. **1550**(1): p. 23-29.

[18] Amer, A., H. Rakha, and I. El-Shawarby, *Agent-based stochastic modeling of driver decision at onset of yellow light at signalized intersections.* Transportation Research Record: Journal of the

Transportation Research Board, 2011. **2241**(1): p. 68-77.

[19] Jahangiri, A., H. Rakha, and T.A. Dingus, *Predicting Red-light Running Violations at Signalized Intersections Using Machine Learning Techniques*. 2015.

[20] Porter, B.E. and K.J. England, *Predicting red-light running behavior: a traffic safety study in three urban settings.* Journal of Safety Research, 2000. **31**(1): p. 1-8.

[21] Retting, R.A. and A.F. Williams, *Characteristics of red light violators: results of a field investigation.* Journal of Safety Research, 1996. **27**(1): p. 9-15.

[22] Porter, B.E. and T.D. Berry, *A nationwide survey of self-reported red light running: measuring prevalence, predictors, and perceived consequences.* Accident Analysis & Prevention, 2001. **33**(6): p. 735-741.

[23] Bonneson, J.A. and H.J. Son, *Prediction of expected red-light-running frequency at urban intersections.* Transportation Research Record: Journal of the Transportation Research Board, 2003. **1830**(1): p. 38-47.

[24] Dingus, T.A., et al., *The 100-car naturalistic driving study, Phase II-results of the 100-car field experiment*. 2006.

[25] Retting, R.A., S.A. Ferguson, and C.M. Farmer, *Reducing red light running through longer yellow signal timing and red light camera enforcement: results of a field investigation.* Accident Analysis & Prevention, 2008. **40**(1): p. 327-333.

[26] Elhenawy, M., H. Rakha, and I. El-Shawarby. *Enhancing Driver Stop/Run Modeling at the Onset of a Yellow Indication using Historical Behavior and Machine Learning Techniques*. in *Transportation Research Board 93rd Annual Meeting*. 2014.

[27] Ghanipoor Machiani, S. and M. Abbas. *Predicting Drivers Decision in Dilemma Zone in a Driving Simulator Environment using Canonical Discriminant Analysis*. in *The 93nd Annual Meeting of the Transportation Research Board*. 2014. Washington, DC.

[28] Liang, Y., M.L. Reyes, and J.D. Lee, *Real-time detection of driver cognitive distraction using support vector machines.* Intelligent Transportation Systems, IEEE Transactions on, 2007. **8**(2): p. 340-350.

[29] Kim, Z., *Robust lane detection and tracking in challenging scenarios.* Intelligent Transportation Systems, IEEE Transactions on, 2008. **9**(1): p. 16-26.

[30] Jahangiri, A. and H. Rakha. *Developing a Support Vector Machine (SVM) Classifier for Transportation Mode Identification by Using Mobile Phone Sensor Data*. in *Transportation Research Board 93rd Annual Meeting*. 2014.

[31] Jahangiri, A. and H.A. Rakha, *Applying Machine Learning Techniques to Transportation Mode Recognition Using Mobile Phone Sensor Data.* Intelligent Transportation Systems, IEEE Transactions on, 2015. **PP**(99): p. 1-12.

[32] Balali, V. and M. Golparvar-Fard. *Video-Based Detection and Classification of US Traffic Signs and Mile Markers using Color Candidate Extraction and Feature-Based Recognition*. in *Computing in Civil and Building Engineering (2014)*. ASCE.

[33] Yuan, F. and R.L. Cheu, *Incident detection using support vector machines.* Transportation Research Part C: Emerging Technologies, 2003. **11**(3): p. 309-328.

[34] Aoude, G.S., et al., *Driver behavior classification at intersections and validation on large naturalistic data set.* Intelligent Transportation Systems, IEEE Transactions on, 2012. **13**(2): p. 724-736.

[35] Elmitiny, N., et al., *Classification analysis of driver's stop/go decision and red-light running violation.* Accident Analysis & Prevention, 2010. **42**(1): p. 101-111.

[36] Zhang, L., et al., *Prediction of red light running based on statistics of discrete point sensors.* Transportation Research Record: Journal of the Transportation Research Board, 2009. **2128**(1): p. 132-142.

[37] Doerzaph, Z.R. and V. Neale. *Data acquisition method for developing crash avoidance algorithms through innovative roadside data collection*. in *Transportation Research Board 89th Annual Meeting*. 2010.

[38] Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.

[39] Hastie, T., et al., *The elements of statistical learning*. Vol. 2. 2009: Springer.

[40] Hsu, C.-W. and C.-J. Lin, *A comparison of methods for multiclass support vector machines.* Neural Networks, IEEE Transactions on, 2002. **13**(2): p. 415-425.

[41] Hsu, C.-W., C.-C. Chang, and C.-J. Lin, *A practical guide to support vector classification*. 2003.

[42] McLaughlin, S.B., J.M. Hankey, and T.A. Dingus, *A method for evaluating collision avoidance systems using naturalistic driving data.* Accident Analysis & Prevention, 2008. **40**(1): p. 8-16.

[43] Layton, R. and K. Dixon, *Stopping Sight Distance.* Kiewit Center for Infrastructure and Transportation, Oregon Department of Transportation, 2012.

[44] Ding, C. and H. Peng, *Minimum redundancy feature selection from microarray gene expression data.* Journal of bioinformatics and computational biology, 2005. **3**(02): p. 185-205.

[45] Doerzaph, Z.R., V. Neale, and R. Kiefer. *Cooperative intersection collision avoidance for violations: threat assessment algorithm development and evaluation method*. in *Transportation Research Board 89th Annual Meeting*. 2010.

[46] Kohavi, R. *A study of cross-validation and bootstrap for accuracy estimation and model selection*. in *IJCAI*. 1995.

[47] James, G., et al., *An introduction to statistical learning*. 2013: Springer.

[48] R Core Team, *R: A Language and Environment for Statistical Computing*. 2014, R Foundation for Statistical Computing.

[49] Liaw, A. and M. Wiener, *Classification and Regression by randomForest.* R news, 2002. **2**(3): p. 18-22.

[50] Chang, C.-C. and C.-J. Lin, *LIBSVM: a library for support vector machines.* ACM Transactions on Intelligent Systems and Technology (TIST), 2011. **2**(3): p. 27.

Red-light Running Violation Prediction using Observational and Simulator Data

Arash Jahangiri

Center for Sustainable Mobility, Virginia Tech Transportation Institute

3500 Transportation Research Plaza, Blacksburg, VA 24061

E-mail: ArashJ@vt.edu

Phone: (540) 200-7561

Hesham Rakha (corresponding author)

Center for Sustainable Mobility, Virginia Tech Transportation Institute

3500 Transportation Research Plaza, Blacksburg, VA 24061

E-mail: HRakha@vtti.vt.edu

Phone: (540) 231-1505

Thomas A. Dingus

Virginia Tech Transportation Institute

3500 Transportation Research Plaza, Blacksburg, VA 24061

E-mail: TDingus@vtti.vt.edu

Phone: (540) 231-1501

Abstract

In the United States, 683 people were killed and an estimated 133,000 were injured in crashes due to running red lights during 2012. To help prevent/mitigate crashes caused by running red lights, these violations need to be identified before they occur, so both the road users (i.e., drivers, pedestrians, etc.) in potential danger and the infrastructure can be notified and actions can be taken accordingly. Two different data sets were used to assess the feasibility of developing red-light running (RLR) violation prediction models: (1) Observational data and (2) Simulator data. Both data sets included common factors, such as time to intersection (TTI), distance to intersection (DTI), and velocity at the onset of a yellow light. However, the observational data set provided additional factors that the simulator data set did not, and vice versa. The observational data included vehicle information (e.g., speed, acceleration, etc.) for several different time frames. For each vehicle approaching an intersection in the observational data set, required data were extracted from several time frames as the vehicle drew closer to the intersection. However, since the observational data were inherently anonymous, driver factors such as age and gender were unavailable in the observational data set. Conversely, the simulator data set contained age and gender. In addition, the simulator data included a secondary (non-driving) task factor and a treatment factor (i.e., incoming/outgoing calls while driving). The simulator data only included vehicle information for certain time frames (e.g., yellow onset); the data did not provide vehicle information for several different time frames as vehicles were approaching an intersection. In this study, random forest (RF) as a machine-learning technique was adopted to develop RLR violation prediction models. Factor importance was obtained for different models and different data sets to show how differently the factors influence the performance of each model. A sensitivity analysis showed that the factor importance to identify RLR violations changed when data from different time frames were used to develop the prediction models. TTI, DTI, the required deceleration parameter (RDP), and velocity at the onset of a yellow light were among the most important factors identified by both models constructed using observational data and simulator data. Furthermore, in addition to the factors obtained from a point in time (i.e., yellow onset),

valuable information suitable for RLR violation prediction was obtained from defined monitoring periods. It was found that period lengths of 2 to 6 meters contributed to the best model performance.

Keywords: driver violation; red-light running; signalized intersection; violation prediction; observational data; simulator data; random forest; machine learning

Introduction

According to National Highway Traffic Safety Administration (NHTSA) report, during 2012, more than 2.5 million intersection-related crashes occurred in the United States, of which 2,850 were fatal crashes and 680,000 were injurious crashes [1]. Specifically, statistics demonstrate that a large number of crashes occur at signalized intersections due to traffic violations, of which running red lights has been reported to be a serious issue. According to the Insurance Institute for Highway Safety (IIHS), 683 people were killed and an estimated 133,000 were injured in crashes in the United States during 2012 due to running red lights [2]. The AAA Foundation for Traffic Safety surveyed 2,000 United States residents aged 16 and older. The survey showed that approximately 93% of drivers believe that running through a red light is unacceptable if it is possible to stop safely. However, one-third mentioned they ran through a red light during the past 30 days. This shows that, although drivers are generally aware of the dangers of this type of violation, they are likely to occasionally run a red light [3].

Dilemma Zone and Influential Factors

Drivers approaching signalized intersections need to make a decision whether to stop or proceed when the traffic signal changes from green to yellow. If they decide to proceed and the signal turns red before the driver passes the stop bar (or before clearing the intersection), the driver is considered a red-light violator. Violating a red light may lead to crashes with the side-street traffic. In another scenario, if the driver abruptly stops at the onset of a yellow light while a following vehicle makes a conflicting decision (i.e., decision to proceed), rear-end crashes may occur. The area in which drivers decide what action to take when the traffic signal changes from green to yellow is known as the dilemma zone, which can be defined in space or time [4]. Specifically, Dilemma zone is an area ahead of the signalized

intersection stop line that exists when the minimum stopping distance is larger than maximum clearing distance. In this case, if drivers distance to the stop bar falls between the minimum stopping distance and the maximum clearing distance while encountering the yellow signal, they are neither able to stop nor to proceed through the intersection safely. Alternatively, option zone is formed when the minimum stopping distance is less than maximum clearing distance. In this case, if the approaching vehicles' distance to the stop bar at the onset of yellow is between the minimum stopping distance and the maximum clearing distance, drivers have two options to safely stop at the stop bar or proceed through the intersection[5]. The dilemma zone is a classic problem first introduced by Gazis, Herman [6], followed by numerous studies [4, 7-16].

When predicting red-light running (RLR) violations, factors that influence driver behavior in the dilemma zone need to be taken into account. A number of studies have attempted to investigate factors that affect driver behavior when approaching signalized intersections. These factors influence the driver's decision to stop or proceed when facing a yellow light and, consequently, have an impact on the probability of rear-end and right-angle crashes. A summary of these factors are listed in Table 21. A comprehensive list of influential factors that have been investigated throughout relevant literature can be found in a study conducted by Abbas, Machiani [17].

Table 21 Influential factors affecting driver behaviour when approaching signalized intersections

Factor name	**Studies**
Perception-reaction time	[4, 9, 13, 14, 18-20]
Acceleration/deceleration rate	[9, 13, 14, 18, 19, 21, 22]
Age	[4, 18-20, 23, 24]
Gender	[4, 18-20, 23, 24]
Time to intersection (TTI) at the onset of yellow	[4, 18, 19, 21, 22, 25]
Distance to intersection (DTI) at the onset of yellow	[9, 19, 21, 22, 26]
Approach speed	[9, 13, 14, 20-22, 24, 26]
Vehicle type	[9, 24, 27]

Presence of side-street vehicles, pedestrians, bicycles, or opposing vehicles waiting to turn left	[9, 15]
Flow rate	[9, 24]
Length of yellow interval	[9, 19]
Cycle length	[9, 24]
Presence of police	[15, 25]
Pavement and weather conditions (wet, rainy)	[19, 27]
cell phone use	[24, 26]
Speed limit	[9, 20]
Roadway grade	[19, 20]
Driver aggressiveness	[27]
Driver's learning through different signal settings	[15]
Required deceleration parameter (RDP) at onset of yellow	[21, 22, 28]
Time lost/gained	[25]

Data Collection Methods

Data collection methods restrict factors that can be considered during analyses. For example, using driving simulators as applied in many studies [15, 18, 25, 26, 29-31] facilitates the examination of many factors (e.g. age, gender, presence of police, work zone, distraction, cell phone use, etc.). However, the behavior of the drivers in a simulator may not reflect their natural behaviors when driving in real-world conditions. Using observational data collection methods (i.e., through video cameras), such as [9, 13, 14], the drivers' natural behaviors are captured as the drivers are not aware that their data are being collected. Also, when using a naturalistic data collection method (i.e. having participants drive/ride instrumented vehicles/bicycles) [32], the drivers' natural behaviors can be captured as the drivers become quickly accustomed to the instrumentation and drive as they normally would. Nevertheless, factors such as age and gender cannot be captured in observational (i.e., video recording at infrastructure) studies due

to the inherent anonymity of observational data collection methods. Also, other specific environmental factors, such as the presence of bicycles or police, may not be available in naturalistic or observational data, whereas these scenarios can be easily developed using driving simulators. Another data collection method involves using an experimental test track and running scenarios in a controlled environment. This data collection method essentially combines the capabilities of observational (or naturalistic) and simulator data collection methods. That is, the behavior of the drivers in a test-track environment is more natural than their behavior in a driving simulator. However, their behavior may be affected because the participants know they are in an experiment and may adjust their behavior. A test-track environment also allows specific scenarios to be run, whereas observational or naturalistic data are limited to what scenarios occur in real-world driving. Studies in which data were collected in a controlled field environment include [4, 19, 23, 33].

Ideally, naturalistic or observational data should be used to test and validate an RLR violation prediction model because natural driver behavior occurring in the real world can only be observed in naturalistic or observational data. Nevertheless, several factors such as distraction and cell phone use that affect driver behavior at the onset of a yellow light can only be investigated through driving simulators because using other data collection methods (i.e. observational, naturalistic, and test track experiments) to examine such factors, would place human subjects in danger. Also, there are several factors, such as age, gender, presence of vulnerable road users at intersections, etc., that may not be available in observational or naturalistic data. These factors should also be examined to determine how they influence driver behavior at the onset of a yellow light. Thus, simulators or test tracks are more suitable to provide such data. Furthermore, driving simulators are considered vital for assessing factors related to new vehicle technologies [30].

Study Focus and Objectives

Some studies concentrate on investigating the characteristics of red-light violators and conditions in which drivers are more or less likely to violate [34-39]. For example, drivers who do not use safety belts and non-Caucasian drivers were more likely to violate the red light. Moreover, larger intersections

and higher traffic volumes were associated with higher RLR violation rates [34]. Several studies developed models to estimate the frequency of red-light violators. For instance, Bonneson and Son [37] developed a regression model using several factors, such as flow rate, cycle length, and yellow duration. Also, a topic closely related to the present study is predicting driver decision (i.e., stop or proceed) when the traffic signal turns from green to yellow; this topic was the focus of such studies as [40, 41]. However, not all driver decisions to proceed lead to red-light violations (i.e., passing the intersection during yellow time). Statistical and probabilistic approaches have also been applied [7, 42]. For example, Zhang, Zhou [42] developed a probabilistic model to predict RLR violations, taking into consideration minimizing both false alarm rates and missing errors.

The present study focuses on developing prediction models aiming at identifying RLR violations using two different data sets (i.e., observational data set and driving simulator data set). Random forest (RF) as an artificial intelligent (AI) technique was employed to construct RLR violation prediction models. AI techniques have been adopted to solve many problems in the transportation domain, such as real-time detection of driver cognitive distraction [43], lane detection and tracking [44], transportation mode recognition [45], traffic sign detection [46], and incident detection [47]. These studies illustrate that applying AI methods can lead to promising results. However, studies that apply AI techniques to develop RLR violation prediction models are limited, thus more research is needed. When using AI methods, different terms may be used to refer to the factors that are employed to build models, such as “factors,” “features,” “variables,” “attributes,” and “predictors.” These terms may be used interchangeably. However, the word “factor” is used throughout this paper for consistency and clarity.

The objectives of the present study are as follows: (1) Create several factors in model development and use a selection method to determine the most useful factors. (2) Conduct a sensitivity analysis to determine an appropriate monitoring period corresponding to the onset of a yellow light to capture the information that reflects drivers’ decisions. (3) Investigate how the RF method can predict RLR violations using different monitoring periods while providing enough time for endangered drivers

and/or the infrastructure to respond. (4) Use observational data and driving simulator data to develop prediction models. (5) Identify important factors in predicting RLR violations.

Relevant work is reviewed in the second section of this paper, while the third section describes the observational data and the simulator data used. The fourth section explains the model development, which includes the RF method that was employed, the selected monitoring period, and the adopted factor selection method. The results are presented in the fifth section. Finally, the conclusion is provided in the sixth section.

Relevant Work

Elmitiny, Yan [48] developed a classification tree model to predict RLR violations and found the most important factors to be the vehicle distance at the onset of a yellow light, the operating speed at the onset of a yellow light, and position in traffic flow. However, only a limited number of factors were examined, thus Elmitiny, Yan [48] did not use any factor selection method to determine the most useful factor in predicting RLR violations. To examine several factors, a proper selection becomes critical towards using the most relevant factors. Moreover, the factors examined and found to be important by Elmitiny, Yan [48] were obtained from an instant in time (i.e., the onset of a yellow light). Although the yellow onset is an important instant (i.e., because it is the moment the drivers encounter the yellow and, consequently, must decide whether to stop or proceed), the drivers' decisions are not made at that instant but, rather, are made during a short time period. Therefore, other factors that can explain drivers' behaviors in a time period immediately following the yellow onset should also be examined when predicting RLR violations. The current study makes an effort to first construct such factors and then use a selection method to identify the most useful factors.

Aoude, Desaraju [22] applied the support vector machine (SVM) and hidden Markov model (HMM) to build RLR violation prediction models. They showed how their models outperform traditional methods of prediction. Similarly to Elmitiny, Yan [48], they did not use any factor selection method. Using SVM, Aoude, Desaraju [22] found by experimenting on different combinations that three factors,

namely DTI, speed, and acceleration, led to the best result (i.e., lowest error). Similarly, when applying the HMM method, they tested different combinations of factors to find the ones that led to the lowest error, namely DTI, speed, acceleration, TTI, and the RDP.

Aoude, Desaraju [22] considered a point in time after which the prediction becomes invalid because not enough time is available for a driver in a potential collision to react. Their logic was that, if a safety system employs any prediction model, the prediction task needs to be conducted before a certain period to provide sufficient time for endangered drivers to respond. Furthermore, they considered a monitoring period during which the required factors were extracted for model development. This period was chosen right before the aforementioned critical point in time. They determined this critical point based on two criteria, whichever was the first to occur: (1) Minimum time threshold. Three different values were selected based on a human response time distribution as discussed by McLaughlin, Hankey [49]: 1, 1.6, and 2 seconds, which are sufficient for 45%, 80%, and 90% of the population to respond, respectively; and (2) Minimum distance threshold. If the vehicles' approaching speeds were very low and vehicles drew too close to the intersection, it is unlikely that the minimum time threshold criterion meets, thus a minimum DTI was assumed. However, the minimum time threshold is not enough to avoid a possible collision since it only corresponds to the driver response time without considering the vehicle response time. Consequently, to avoid a potential collision, two time periods were taken into account in the present research: (1) The driver response time and (2) The vehicle response time.

Aoude, Desaraju [22] did not use any factors that reflect the interaction between the drivers and the signal setting. For example, it appears that factors incorporated into the Aoude, Desaraju [22] SVM model, such as DTI and speed, were extracted from the monitoring period that they defined without knowing the yellow light onset (i.e., the yellow light may start before or after their defined monitoring period). In the present paper, the yellow onset and the monitoring period immediately after onset were both accounted for as such time periods contain the information reflecting the drivers' decisions when encountering a yellow light.

Zhang, Zhou [42] and Zhang, Wang [50] proposed a probabilistic framework to identify RLR violations, which was adopted in dynamic all-red extension (DARE) as an intersection collision avoidance method. Their prediction model was based on vehicle speed measurements from a minimum set of two point sensors and their corresponding event time stamps. They applied Neyman–Pearson (NP) criterion to maximize the probability of prediction while keeping the false-alarm rate equal to or lower than a given level.

Data Description

Observational Data

The observational data used in this research were derived from the Cooperative Intersection Collision Avoidance Systems for Violations (CICAS-V) project. As part of the CICAS-V project, different equipment, such as radars, video cameras, and signal phase sniffers, were included in the data acquisition systems (DASs) designed and developed by the Center for Technology Development (CTD) at the Virginia Tech Transportation Institute (VTTI). The DAS units were installed at six stop-controlled intersections and three signalized intersections in the New River Valley area of Southwest Virginia. A data sample from one of the signalized intersections (i.e., the intersection of Franklin Street and Depot Street) was used in the present study. Doerzaph and Neale [51] provide additional details about the data collection of CICAS-V. The sample used in this study includes approximately 500 observations for the RLR violation behavior and 500 observations reflecting compliant behavior. RLR violations are defined based on the location of the vehicles when the traffic light changes to red. According to National Cooperative Highway Research Program (NCHRP) report, there are two definitions for RLR violations: *"under "permissive" yellow law, drivers may enter the intersection during the entire duration of the yellow change interval and legally be in the intersection while the red signal indication is displayed, so long as entrance occurred before or during the yellow signal indication. Under the "restrictive" yellow law, (1) drivers may not enter the intersection during the yellow signal indication unless it can be entirely cleared prior to the onset of the red signal indication, or (2) drivers may not enter the intersection unless it is impossible or unsafe to stop."* [52]. In the present research, it was assumed that the "permissive"

yellow law is followed. That is, drivers were identified as RLR violators if the light was red the moment they crossed the stop line. For each individual vehicle approaching an intersection, data such as speed, acceleration, DTI, and signal setting information were collected at high resolution. Table 22 presents the factors used in this study.

Table 22 Factors used from the observational data (CICAS-V)

No.	Factor	No.	Factor
1	DTI	6	Velocity at Onset of Yellow
2	TTI	7	Acceleration at Onset of Yellow
3	DTI at Onset of Yellow	8	Vehicle speed
4	TTI at Onset of Yellow	9	Vehicle Acceleration
5	RDP at Onset of Yellow		

Driving Simulator Data

The data set from a driving simulator study provided for a data contest at the 93rd Annual Meeting of the Transportation Research Board was obtained through the *Journal of Accident Analysis & Prevention* website. The data contain several factors, of which a subset was used in this study, as shown in Table 23.

Table 23 Factors used from the simulator data

No.	Factor
1	Age group: young (18-25), middle-aged (30-45), and older (50-60)
2	Gender
3	The secondary (or non-driving) task condition: • Using handheld wireless for dialing and conversing. • Using hands-free wireless for voice dialing using digits, and hands-free using external speaker kit for conversing. • Using the phone for voice dialing using digits, and the headset for hands-free conversing.
4	Treatment: Baseline (no call), Outgoing call, and Incoming call

5	Frame number at which the traffic light changes from Yellow to Red
6	Distance from stop line when the vehicle comes to stop. Notes: Drivers who stopped beyond the stop bar as well as the drivers who did not stop were also coded.
11	The frame number for when the participant had an accelerator pedal change of greater than 10% percent.
12	Acceleration Pedal Change Direction: -1=released, 1=depressed
7	Max deceleration between the 10% increase in Acceleration Pedal and when driver goes past intersection.
8	Max acceleration between the 10% increase in Acceleration Pedal and when driver goes past intersection.
9	Velocity at onset of yellow.
10	Distance from stop line at onset of yellow.
11	Frame number when participant reached stop line.

According to the provided data description, the practice runs denoted as "FAMILIAR" were omitted from the data set. The RLR violations were identified as follows: Looking at the "Distance from stop line when the vehicle comes to stop" (factor 6), drivers who did not stop and those who stopped beyond the stop line were extracted first. Subsequently, considering only this subset of observations, those with a frame number at the stop line (factor 11) greater than the frame number at yellow to red (factor 5) represent violators. To summarize, violators in the simulator data are those who passed the stop line when the light was red, no matter if they passed the intersection completely or stopped just beyond the stop line.

Model Development

RF Method

A machine-learning method, namely RF, was employed in the present research to predict RLR violations. RF was used as it offers several advantages [53, 54]. Namely, the model performance is as good as (and sometimes better than) other powerful methods, such as Adaboost, discriminant analysis, SVMs, and neural networks. RF is robust against over-fitting and is relatively robust to outliers and noise. It is faster than bagging or boosting, it provides useful internal error estimates known as out-of-bag (OOB) errors, and it provides factor importance. In addition, RF is easy to tune and requires only two

parameters, thus it can be simply optimized. The RF method, as proposed by Breiman [54], is an ensemble learning approach based on predictions of a number of decision trees. Instead of a single decision tree, RF uses a group of decision trees from which a majority vote makes the predictions. Two model parameters, namely the number of decision trees and the number of factors to use in each tree, should be determined to apply the RF method. To construct each tree, the recursive binary splitting method was adopted in which factors are selected to divide the data into different parts. Different criteria may be used to determine how to separate the data. The Gini index criterion was used in this study, as presented in Equation 33. It should be noted that the Gini index and cross-entropy criteria act similarly and are both recommended approaches [55].

$$G = \sum_{k=1}^{K} P_k^m (1 - P_k^m) \qquad \text{Equation 33}$$

Where,

$$P_k^m = \frac{1}{N^m} \sum_{x_i^m} I(y_i^m = k)$$

P_k^m	Proportion of class k observations in node m
N^m	Number of observations received at node m
y_i^m	The response value corresponding to the observation i at node m
x_i^m	The feature vector corresponding to the observation i at node m
k	class

Monitoring Period

Figure 26 illustrates two vehicles approaching a signalized intersection: (1) A violator vehicle denoted by v, which is presumably going to violate the red light; and (2) An endangered vehicle denoted by e, which is going to be at risk of a right-angle crash due to the RLR violation of vehicle v. In previous work [21, 56], fixed monitoring periods were determined to extract the information reflecting driver behavior when approaching a signalized intersection. This fixed period was defined in both time (i.e.,

based on TTI; [21])and in space (i.e., based on DTI; [56]). Consequently, prediction models were developed based on the information obtained from those periods. In the present study, the monitoring period was defined in space similar to [56], but different period lengths were examined instead of a fixed period. DTI was used here instead of TTI due to convenience. That is, identifying the monitoring period based only on the TTI becomes difficult, especially when speeds are very low that lead to very high TTI values. Moreover, using DTI makes it easier to conduct a sensitivity analysis with different monitoring periods. The monitoring period for each observation was defined depending on the DTI of each driver at the yellow onset instead of defining a fixed period for all drivers, as was performed in a previous study [56]. It should be noted that the monitoring period was only applied when using the observational data. Although simulators in general can record data at 60 Hz, the particular simulator data used in this study did not include factors (e.g., speed, acceleration) for all data frames. Hence, it was not possible to define a monitoring period in the simulator data.

The monitoring period (t_{mon}^{v}) was defined by its start and end points based on DTI values, as illustrated in Figure 26. For example, selecting DTI for start and end points as 40 and 25 meters, respectively, lead to a monitoring period of 15 meters. The start point of the monitoring period should not be selected too early to exclude unnecessary information (i.e., when the drivers are very far from the intersection, their behavior may not reflect their decision to violate the red light). The behavior of the drivers before the yellow onset may not be related to their decision to stop or proceed. Therefore, DTI at the yellow onset was selected as the start point for each observation. The end point was restricted in previous work to a point (i.e., t_{min}^{v} as shown in Figure 26) that provides sufficient time for the endangered vehicle to respond if a possible collision is predicted; t_{min}^{v} is equivalent to the sum of two terms, namely the time required for the endangered driver to respond (t_{driver}^{e}) and the time required for the endangered vehicle to stop ($t_{vehicle}^{e}$). As a result, the sum of the two terms (t_{driver}^{e} and $t_{vehicle}^{e}$) dictated the end point of the t_{mon}^{v} [21, 56]. In the present study, however, a sensitivity analysis was conducted to investigate the

effects of different monitoring periods on predicting RLR violations. The sensitivity analysis was performed to show how accurately the prediction models can perform with different monitoring periods.

Crashes resulting from an RLR violation can occur in two scenarios: (1) When no vehicle is waiting at the intersection, the endangered vehicle is approaching the intersection, and the traffic light turns green before the endangered vehicle needs to stop. This scenario is illustrated in Figure 26 and was the focus of previous work with a fixed monitoring period [21]; and (2) When the endangered vehicle is already at the intersection waiting for the light to turn green. This vehicle is in front of a possible waiting queue and is the first vehicle to enter the intersection as soon as the light turns green. The present study focuses on both scenarios with consideration of different monitoring periods. For the first scenario, if enough time is available, the endangered driver can be notified so he or she can avoid a possible collision. However, if there is not enough time for the driver to react, or if the second scenario occurs, the infrastructure can be notified so appropriate action can be taken (e.g., providing a red clearance interval such as what Zhang, Wang [50] proposed).

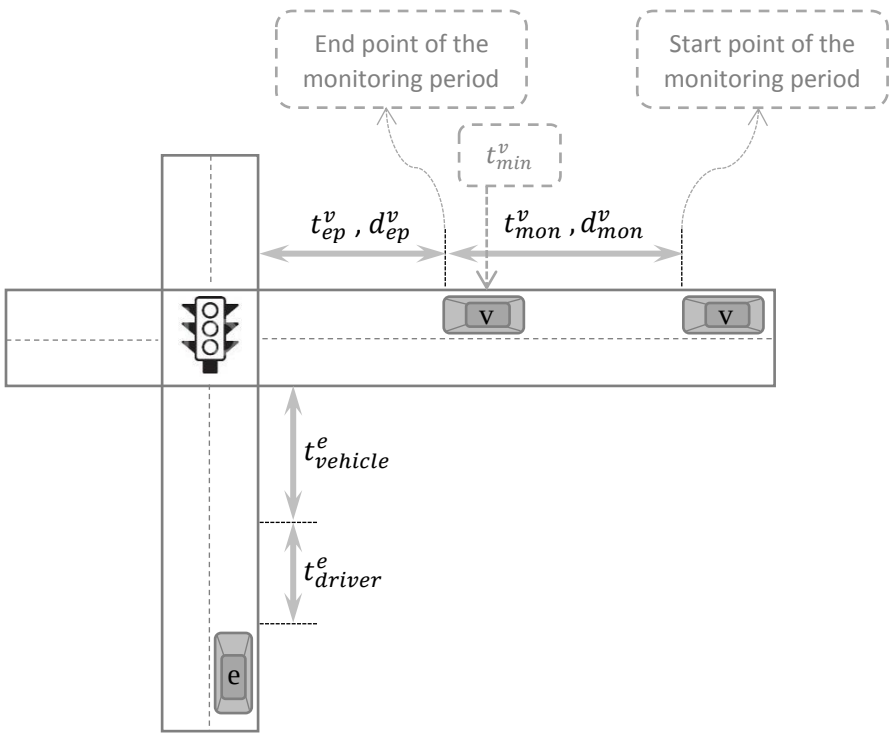


Figure 26: Determination of the monitoring period; vehicle *v* as the RLR violator and vehicle *e* as the endangered vehicle

Factor Creation

In addition to the available factors in both observational and simulator data sets, others were created as follows:

Using the observational data

Using the factors from observational data, as presented in Table 22, additional factors were created for examination. Table 24 lists all of the factors that were used for model development. The vehicle information for different time frames included velocity, acceleration, TTI, and DTI. As discussed in the previous section, a monitoring period to be examined for each vehicle contains the information between the start and end points of the period. To describe changes in the defined monitoring period, the additional factors were created using mostly statistical measures of dispersion (e.g., maximum, minimum, and standard deviation). The idea was that any change in driver behavior due to the drivers' decisions to stop or proceed can directly affect the kinetic information within the monitoring period. It should be noted that the RDP is the deceleration value required for a vehicle to be able to stop at the stop bar. The RDP was obtained through Equation 34 as follows [28].

$$RDP = \frac{V^2}{2.DTI.g} \qquad \text{Equation 34}$$

Where,

V	Vehicle's instantaneous velocity
DTI	Distance to intersection
g	Gravitational constant

Table 24 Observational data - list of the examined factors

No.	Factor	No.	Factor
1	DTI at Onset of Yellow	10	$std(Velocity)$ over the t_{mon}^{v}
2	Velocity at Onset of Yellow	11	$mean(Acceleraiton)$ over the t_{mon}^{v}
3	Acceleration at Onset of Yellow	12	$range(Acceleraiton)$ over the t_{mon}^{v}
4	TTI at Onset of Yellow	13	$max(Acceleraiton)$ over the t_{mon}^{v}

5	RDP at Onset of Yellow	14	$min(Acceleraiton)$ over the t_{mon}^{v}
6	$mean(Velocity)$ over the t_{mon}^{v}	15	$std(Acceleraiton)$ over the t_{mon}^{v}
7	$range(Velocity)$ over the t_{mon}^{v}	16	$mean(DTI)$ over the t_{mon}^{v}
8	$max(Velocity)$ over the t_{mon}^{v}	17	$mean(TTI)$ over the t_{mon}^{v}
9	$min(Velocity)$ over the t_{mon}^{v}		

Using the simulator data

The additional factors created in the simulator data included TTI at the onset of a yellow light and the RDP at the onset of a yellow light. Total factors examined from this data set are shown in Table 25. As mentioned earlier, the data points within a defined monitoring period were not available for this particular data set. Therefore, unlike the observational data, no factors were created based on the monitoring period.

Table 25 simulator data - list of the examined factors

No.	Factor	No.	Factor
1	Gender	7	Velocity at Onset of Yellow
2	Age	8	DTI at Onset of Yellow
3	Acceleration Pedal Changed 10%	9	The secondary task condition
4	Acceleration Pedal Change Direction	10	Treatment
5	Max deceleration	11	TTI at Onset of Yellow
6	Max Acceleration	12	RDP at Onset of Yellow

Factor Selection

In the model development process, factors that can provide useful information for prediction need to be identified. Advantages of an appropriate factor selection method include reducing the dimensionality of the problem, reducing the noise, and identifying more important and more interpretable factors. In a previous study [21], the minimum redundancy maximum relevance (mRMR) approach was adopted to select the five most representative factors in predicting RLR violations. This approach

attempts to find factors that have both the highest level of relevance and the lowest level of redundancies between factors [57]. In the present study, a different approach was used to identify the most useful factors. While using the RF method for developing prediction models, the individual contribution of each factor, called the factor importance, was obtained. Thus, the factors were ranked based on that measure. In other words, all of the factors were taken into account to develop prediction models; as a result, the importance of individual factors was obtained. Subsequently, a desirable subset of only important factors (e.g., top 5, or top 10) can be identified based on the factor importance measure. The factor selection based on the factor importance metric leads to selecting the most representative factors in a more accurate way when compared to other techniques, such as mRMR. This is because the actual contribution of each factor is assessed while developing the prediction models. However, the factor selection using factor importance is most valuable when using the RF method as it internally calculates the factor importance.

Results

For each data set, prediction models were developed, and the individual factor importance was obtained. To assess the model performances, the OOB error was used. When evaluating tree-based models such as RF, the unbiased estimation of the error, namely the OOB error, is obtained internally and is nearly identical to the cross-validation accuracy [55].

Observational Data Results

To implement the RF method, the R software and RandomForest package were used [53, 58]. First, all of the factors that were obtained or created as listed in Table 24 were taken into account to develop a prediction model. When developing the RF model, different numbers of trees was examined, as shown in Figure 18. The error rate became stable by increasing the number of trees beyond 400. However, since increasing the number of trees does not lead to over-fitting, the value of 800 was selected to ensure that a sufficient number of trees was applied. Another parameter that needed to be determined was the number of factors that each tree requires to grow. As Figure 19 illustrates, the effect of the

number of factors was negligible, but the lowest error rate was achieved when six factors were used to grow each tree.

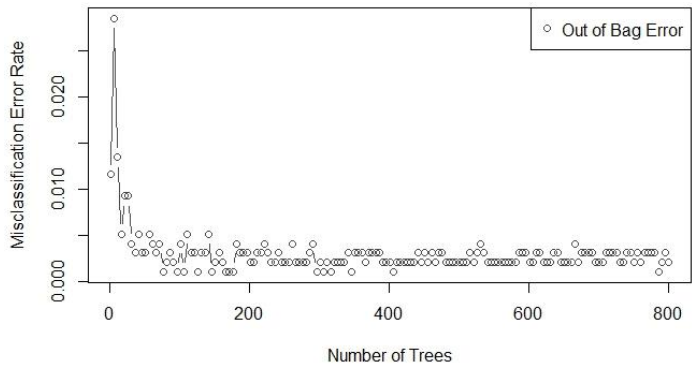


Figure 27: Observational Data - Selecting the required number of trees

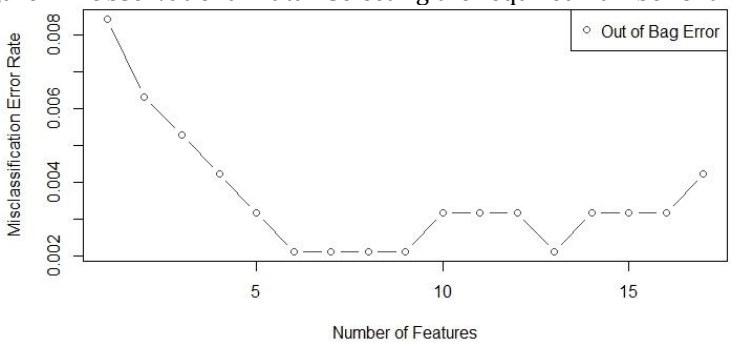


Figure 28: Observational Data - Selecting the number of factors for each tree

Furthermore, to conduct the sensitivity analysis, different monitoring periods were evaluated. As mentioned earlier, a monitoring period is defined by two parameters: the start and end points. The start point is always the DTI at the yellow onset. Therefore, different monitoring periods were obtained by changing the end point, as shown in Figure 29. As a result, monitoring periods with different lengths (i.e., from 2 to 30 meters) were assessed.

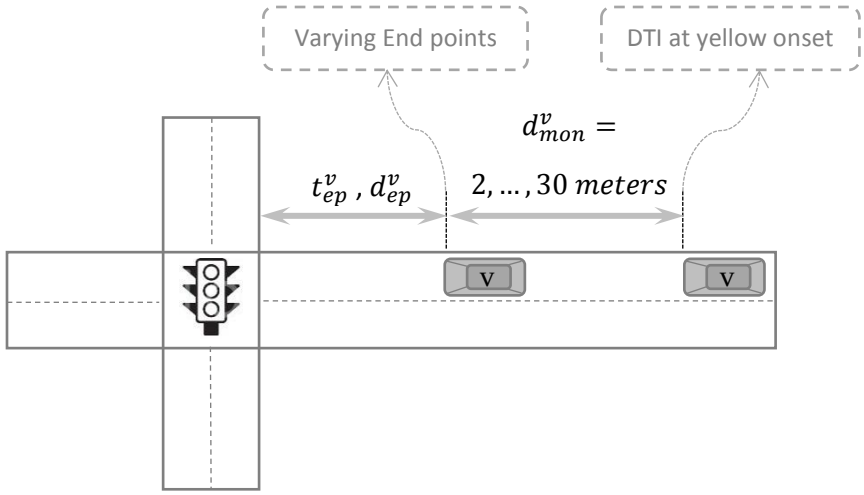


Figure 29 Parameters to select monitoring periods

Using all 17 factors in the observational data, the models contributed to low error rates[5] (0.11-0.53%) for all monitoring periods, as shown in Figure 30. The factor importance was obtained to rank all factors and identify the most useful ones. Factor importance can be assessed based on two measures: (1) Mean decrease accuracy, which shows how the detection accuracy is decreased if a factor is excluded, averaged across all trees, and normalized by the standard deviation of the differences in accuracy; and (2) Mean decrease Gini, which shows how a single factor contributes to decrease the Gini index across all of the trees. The factors identified by these two measures were found to be the same. Thus, only the mean decrease Gini was used for evaluations. The order of importance for the factors was found to be different when different monitoring periods were assessed. For example, Figure 31 illustrates how the importance of the third factor identified in Table 24 (i.e., acceleration at the onset of a yellow light) changes for different monitoring periods. The figure shows that this factor was recognized as more important when the monitoring period of 4 to 7 meters was used compared to when longer monitoring periods were applied (e.g., > 10 meters). Therefore, the factor importance change was taken into account when a different number of factors was applied, as shown in Table 26. For example, when using the top four factors, factor numbers 4, 17, and 1 were used for all monitoring periods, factor number 16 was used for monitoring periods of 2 to 29 meters, and factor number 15 was used for the monitoring period of 30

[5] See appendix D for a discussion on false positives and false negative rates

meters. As a result, RF models were developed for different combinations of monitoring periods and the number of factors used, as illustrated in Figure 32.

According to Table 26, factor numbers 4 (TTI at the onset of a yellow light), 17 ($mean(TTI)$ over t_{mon}^{v}), 1 (DTI at the onset of a yellow light), and 16 ($mean(DTI)$ over t_{mon}^{v}) were found to be the four most important factors for all monitoring periods except one case; there was one situation during which factor 15 was identified as the fourth important factor when using a monitoring period of 30 meters. Factor numbers 5, 7, 15, and 16 were identified as the fifth most important factor, depending on which monitoring period was used.

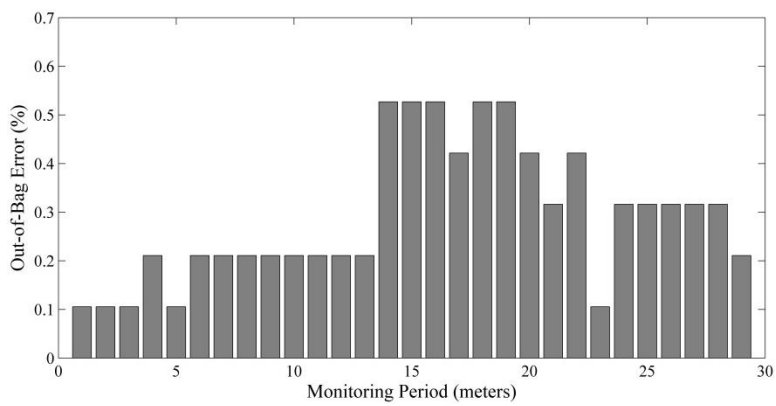


Figure 30 Observational Data - OOB error rate in percentage using all 17 observational data factors

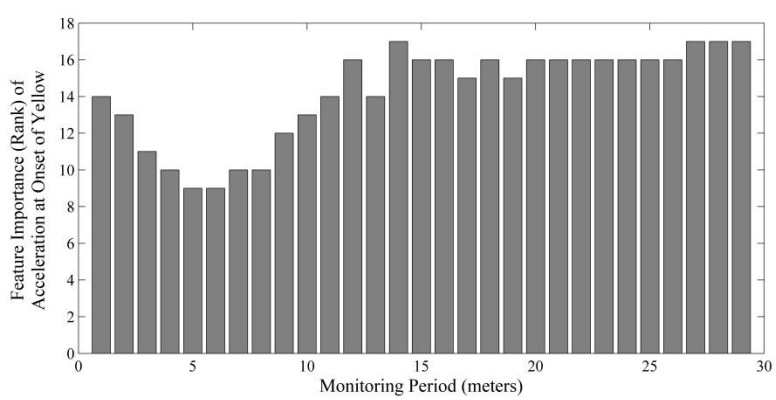


Figure 31 Observational Data - Factor importance change - Acceleration at Onset of Yellow

Table 26 Factor Importance Change

Monitoring Period (meters)	Factor Rank																
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***	***14***	***15***	***16***	***17***
2	4	17	1	16	5	2	7	10	6	9	8	13	12	3	15	11	14
3	4	17	1	16	5	7	2	9	10	6	8	13	3	15	14	11	12
4	4	17	1	16	5	2	9	6	10	7	3	13	14	15	8	12	11
5	4	17	1	16	5	2	9	6	10	3	7	15	8	13	11	14	12
6	4	17	1	16	7	8	5	2	3	6	15	9	10	11	13	12	14
7	4	17	1	16	7	8	5	2	3	6	9	15	10	11	13	12	14
8	4	17	1	16	7	8	5	2	9	3	6	15	10	11	13	12	14
9	4	17	1	16	7	8	5	2	11	3	15	6	9	10	13	12	14
10	4	17	1	16	7	8	5	2	15	13	9	3	6	11	10	12	14
11	4	17	1	16	7	5	8	2	13	11	9	6	3	10	15	12	14
12	4	17	1	16	5	7	2	13	8	9	6	15	11	3	10	12	14
13	4	17	1	16	5	7	2	13	15	9	8	6	10	11	12	3	14
14	4	17	1	16	7	5	2	15	8	13	9	6	11	3	12	10	14
15	4	17	1	16	5	7	2	15	8	9	6	13	10	11	12	14	3
16	4	17	1	16	7	5	15	2	12	9	8	11	6	10	13	3	14
17	4	17	1	16	5	7	15	2	12	9	6	11	13	10	8	3	14
18	4	17	1	16	5	7	15	2	12	9	6	10	13	11	3	14	8
19	4	17	1	16	5	15	7	2	12	9	6	11	13	10	8	3	14
20	4	17	1	16	5	15	7	2	12	9	6	11	10	13	3	8	14
21	4	17	1	16	15	5	2	12	13	7	9	6	11	10	8	3	14
22	4	17	1	16	15	5	12	13	2	9	11	6	7	8	10	3	14
23	4	17	1	16	15	5	12	2	11	13	9	8	6	7	10	3	14
24	4	17	1	16	5	15	13	12	2	11	6	9	8	10	7	3	14
25	4	17	1	16	15	5	11	12	2	9	8	13	6	10	7	3	14
26	4	17	1	16	15	5	12	11	13	2	9	6	8	10	7	3	14
27	4	17	1	16	15	5	11	13	12	2	9	6	10	7	8	3	14
28	4	17	1	16	15	5	11	13	12	2	9	6	8	10	7	14	3

	29	4	17	1	16	15	5	11	13	2	9	8	6	7	10	12	14	3
	30	4	17	1	15	16	11	5	13	2	9	8	6	10	7	12	14	3

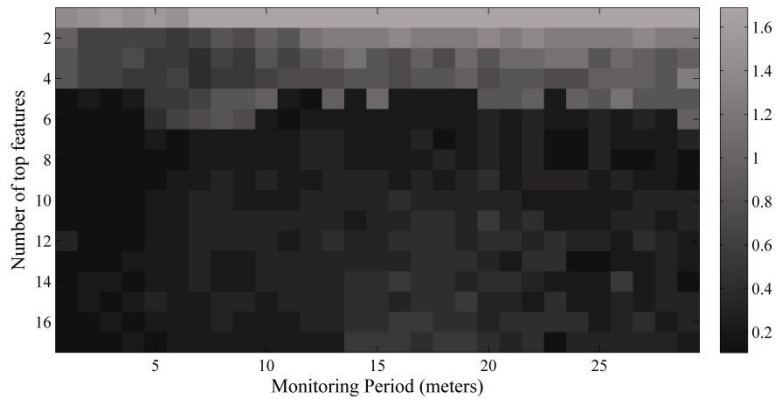


Figure 32 OOB error for different monitoring periods and number of top factors

Based on Figure 32, low error rates were obtained using different monitoring periods and number of top factors (i.e., from 0.1~1.6%), with darker regions representing lower error rates. Therefore, the prediction models achieved the lowest error rates when using more than five factors and the monitoring period of 2 to 6 meters. As the monitoring period increases with a specific number of factors (e.g., top 10), the error rate generally increases. As mentioned earlier, the start point of the monitoring period is the DTI at the yellow onset. Hence, it can be inferred that a short monitoring period (i.e., 2 to 6 meters) immediately following the yellow onset is the most appropriate period that should be monitored to predict RLR violations. Moreover, when using more factors with a particular monitoring period (e.g., 5 meters), the error rate decreases. However, minimal benefit was gained when employing more than six factors.

Simulator Data Results

As mentioned earlier, factors listed in Table 25 were obtained from the simulator data. To develop RF models, a procedure was used similar to that of the model development using observational data. The number of trees and number of factors used for each tree in the simulator data were determined to be 700 and 3, respectively, as shown in Figure 33 and Figure 34. Since it was not possible to define a

monitoring period for the simulator data set, no sensitivity analysis was conducted. The importance of different factors was obtained and is illustrated in Figure 35; the associated factor ranking is presented in Table 27. TTI at the onset of a yellow light was found to be the most important factor, followed by the RDP at the onset of a yellow light and the DTI at the onset of a yellow light. Driver factors (i.e., age and gender), the treatment factor, and the secondary task factor were among the least important factors.

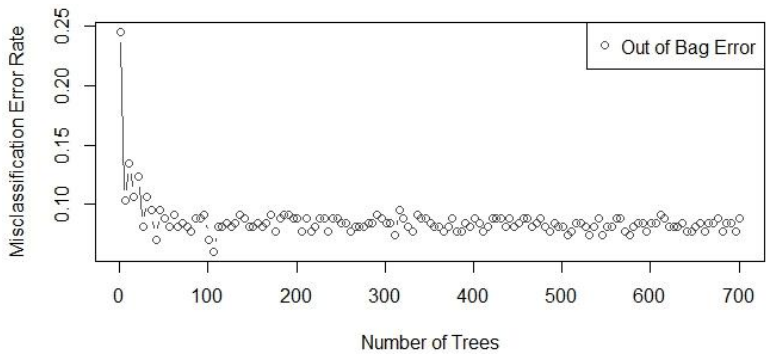


Figure 33 Simulator Data - Selecting the required number of trees

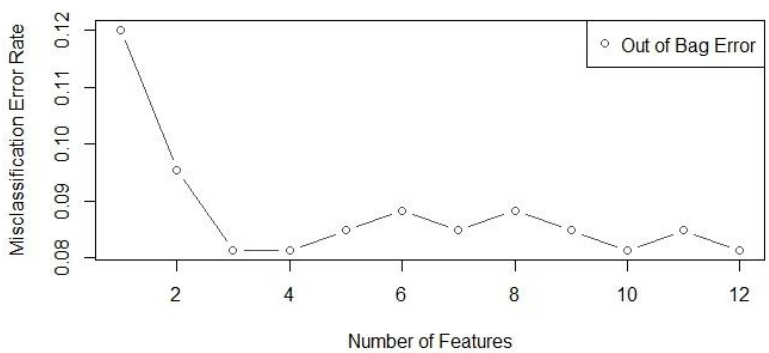


Figure 34 Simulator Data - Selecting the number of factors for each tree

Table 27 Simulator Data - Factor Ranking

No.	Factor	Rank	No.	Factor	Rank
1	Gender	***10, 11***	7	Velocity at Onset of Yellow	***4, 5***
2	Age	***7***	8	DTI at Onset of Yellow	***3***
3	Acceleration Pedal Changed 10%	***8, 11***	9	The secondary task condition	***8, 9***
4	Acceleration Pedal Change Direction	***10, 12***	10	Treatment	***9, 12***
5	Max deceleration	***4, 6***	11	TTI at Onset of Yellow	***1***
6	Max Acceleration	***5, 6***	12	RDP at Onset of Yellow	***2***

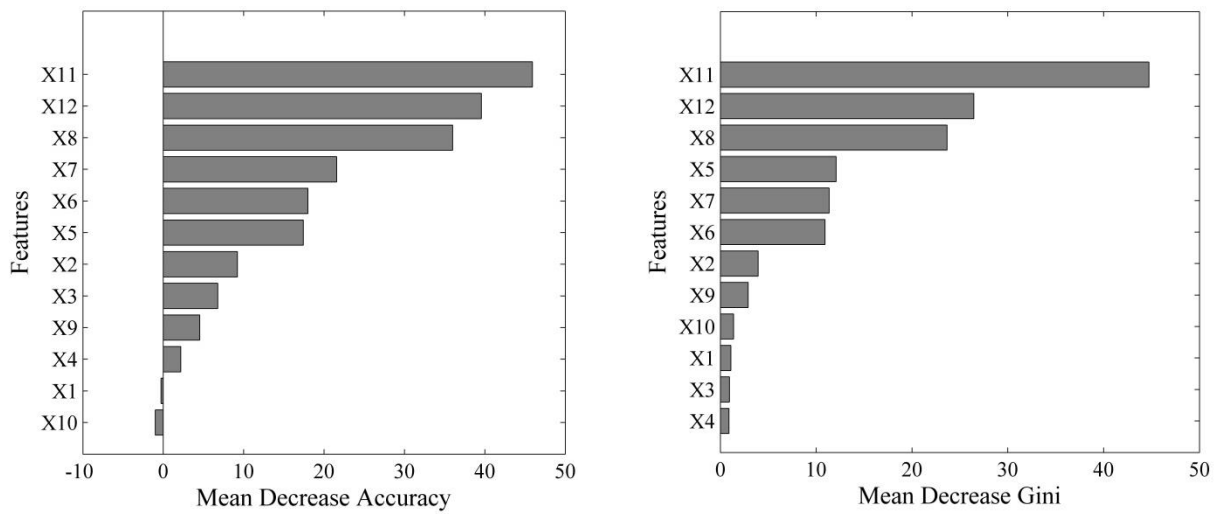


Figure 35 Simulator Data - Factor Importance

Similarly to the observational data model, factor rankings of the simulator data were obtained using both mean decrease Gini and mean decrease accuracy criteria, as shown in Figure 35. According to Figure 35 and Table 27, factor numbers 11 (TTI at the onset of a yellow light), 12 (RDP at the onset of a yellow light), and 8 (DTI at the onset of a yellow light) were found by both criteria to be the three most important factors. Different models were developed using a different number of factors with respect to importance. For example, a model was developed using only the top five factors, and so forth. Error rates of 5.9% to 17.9% were obtained depending on the number of top factors used in model development, which shows a relatively poor performance[6]. After using the top four factors, minimal benefit was gained, as illustrated in Figure 36. There is even a slight increase in the error rate after using more than four factors, suggesting that adding more factors does not necessarily result in lower error rates. This shows the significance of factor selection, for which obtaining factor importance was shown to be a useful method when developing RF models.

[6] See appendix D for a discussion on false positives and false negative rates

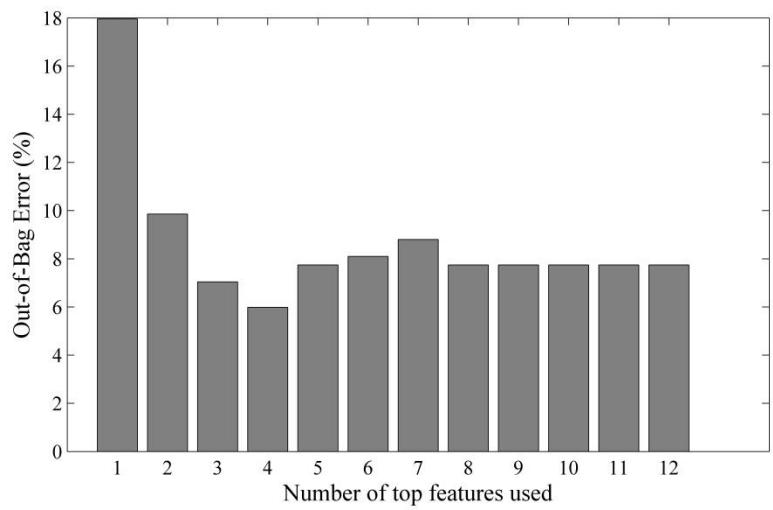


Figure 36 OOB error for different number of top factors used

Model comparison: observational data vs. simulator data

A direct comparison between the models developed using the observational data set and the simulator data set is not meaningful. This is because the models were constructed using different data sets, and some factors were available in the observational data set that were unavailable in the simulator data set, and vice versa. However, two points can be made: (1) Important factors identified by models that were developed using both data sets were similar. TTI at the onset of a yellow light, DTI at the onset of a yellow light, RDP at the onset of a yellow light, and velocity at the onset of a yellow light were among the most important factors identified by models constructed using both data sets; and (2) Models developed using the observational data achieved lower error rates compared to those constructed by the simulator data. It initially appears that the observational data models are more accurate (i.e., lower error rates) because the observational data provided information for several frame numbers, thus enabling the use of a monitoring period. However, even when using one factor (i.e., TTI at yellow onset) that was not based on the monitoring period in the observational data, the error rate was significantly lower than that of the simulator data models (i.e., 1.6%). This suggests that the monitoring period may not be the reason the observational data models show higher performance rates. In other words, using TTI at yellow onset, RLR violations were identified with a high accuracy (i.e., 1.6% error rate) with the observational data, while using the same factor in the simulator data model led to a poor performance (i.e., error rate of

17.95%). Thus, it seems that the driver behavior in an observational situation may be significantly different than driver behavior in a simulator condition. This suggests that a direct comparison between simulator data and observational data is necessary in future research endeavors. For this to be a meaningful comparison, the same participants would be required to collect both observational and simulator data.

Implementation Considerations

Vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) technology can be considered as a potential countermeasure to crashes resulting from RLR violations. In situations where the endangered driver has sufficient time, a warning can be issued to the driver to respond. In cases where insufficient time is available, the infrastructure can take appropriate actions (e.g., providing a red clearance interval). Connected-vehicle technology uses dedicated short-range communications (DSRC) that provide a reliable and fast communication (latency of less than 100 ms) and a range of less than 1,000 meters. Once the prediction model was developed, the prediction time for a single observation was extremely fast (between 1 to 10 ms on an i5-3230M CPU at 2.6 Hz and RAM of 8 GB). It may take a significant amount of time to develop a model depending on data size and frequency. However, the model is constructed using pre-collected (i.e., historical) data, thus the development time is excluded in real-time applications. Therefore, it appears that the future intersection safety systems will be capable of performing violation predictions and notifying endangered drivers and/or the infrastructure in a fraction of a second.

Conclusions

Prediction models were developed to identify RLR violations before they occur so the endangered driver and/or the infrastructure can be notified. The prediction models were developed using two data sets: an observational data set and a simulator data set. A machine-learning technique, namely the RF method, was adopted to develop prediction models that resulted in very high prediction accuracies in the case of the observational data set (i.e., error rates of 0.1% to 1.6%) and poor accuracies in the case of the simulator data set (i.e., error rates of 5.9% to 17.9%). When using the observational data, statistical

measures of dispersion and central tendency were applied to create factors reflecting driver behavior (i.e., decision to stop or proceed at the onset of a yellow light) across a monitoring period from which the data were extracted. Two types of factors were evaluated: (1) Factors that describe quantities at a point in time (e.g., TTI at yellow onset) and (2) Factors that describe quantities across a time period that occurs during the defined monitoring period (e.g., $max(Velocity)$ over t_{mon}^{v}).

Results from the observational data showed that both types of factors were significant in predicting RLR violations. Individual factor importance was obtained, and it was shown that the factor importance may change depending on the monitoring period being assessed. The tuning parameters for the RF model (e.g., the number of trees and the number of factors used in each tree) were also determined to obtain the best possible model performance. TTI at the onset of a yellow light, the $\mathrm{mean}(TTI)$ over $\mathrm{t}_{\mathrm{mon}}^{\mathrm{v}}$, and DTI at the onset of a yellow light were identified as the most important factors (i.e. top three) of the observational data. Additional important factors in the observational data that described statistical quantities across the monitoring period were $mean(TTI)$ over t_{mon}^{v}, $mean(DTI)$ over t_{mon}^{v}, $range(Velocity)$ over t_{mon}^{v}, and $std(Acceleraiton)$ over t_{mon}^{v}. This shows that, in addition to the yellow onset, an appropriate monitoring period can provide useful information reflecting driver behavior (i.e., decision to stop or proceed) when approaching a signalized intersection.

Simulator data had the advantage of accounting for driver factors (i.e., age and gender) and specific hypothetical scenarios, as indicated by the treatment and the secondary task condition factors. However, these factors were found to be among the least important in predicting RLR violations. In fact, models developed without these factors had similar performances compared to the models that used these factors. TTI at the onset of a yellow light, RDP at the onset of a yellow light, and DTI at the onset of a yellow light were found to be the most important factors from the simulator data set. These factors were all related to a point in time, which is the instant at which the traffic light changes from green to yellow. Hence, this point in time plays a critical role in predicting RLR violations, which is consistent with existing literature.

Comparing the models developed using observational and simulator data, TTI at the onset of a yellow light, DTI at the onset of a yellow light, RDP at the onset of a yellow light, and velocity at the onset of a yellow light were among the most important factors identified by models constructed using both data sets. Moreover, models developed using observational data contributed to higher prediction accuracies. Using only one common factor (i.e. TTI at yellow onset) in model development, the observational data model resulted in %1.6 error, whereas the simulator data model led to a high error rate of %17.95. However, a direct comparison between models developed using two different data sets may not be legitimate. To have a meaningful comparison, same human subjects would be required to collect both observational and simulator data.

Acknowledgements

This research effort was funded by the Tier 1 U.S. Department of Transportation Connected Vehicle/Infrastructure University Transportation Center (CVI-UTC).

References

[1] NHTSA, *Traffic safety facts 2012*. 2014, National Center for Statistics and Analysis, US Department of Transportation, Washington, DC.

[2] IIHS. *Red light running*. 2012; Available from: http://www.iihs.org/iihs/topics/t/red-light-running/topicoverview.

[3] Insurance Institute for Highway Safety (IIHS), *Status Report: Public seeks safer roads but still takes risks*. 2010.

[4] Rakha, H., I. El-Shawarby, and J.R. Setti, *Characterizing driver behavior on signalized intersection approaches at the onset of a yellow-phase trigger.* Intelligent Transportation Systems, IEEE Transactions on, 2007. **8**(4): p. 630-640.

[5] Elhenawy, M., et al., *Modeling driver stop/run behavior at the onset of a yellow indication considering driver run tendency and roadway surface conditions.* Accident Analysis & Prevention, 2015. **83**: p. 90-100.

[6] Gazis, D., R. Herman, and A. Maradudin, *The problem of the amber signal light in traffic flow.* Operations Research, 1960. **8**(1): p. 112-132.

[7] Sheffi, Y. and H. Mahmassani, *A model of driver behavior at high speed signalized intersections.* Transportation Science, 1981. **15**(1): p. 50-61.

[8] Bonneson, J.A., et al., *Intelligent detection-control system for rural signalized intersections*. 2002, Texas Transportation Institute, Texas A&M University System.

[9] Gates, T.J., et al., *Analysis of driver behavior in dilemma zones at signalized intersections.* Transportation Research Record: Journal of the Transportation Research Board, 2007. **2030**(1): p. 29-39.

[10] Pant, P.D., et al., *Field testing and implementation of dilemma zone protection and signal coordination at closely-spaced high-speed intersections*. 2005, University of Cincinnati.

[11] Chang, M.-S., C.J. Messer, and A.J. Santiago, *Timing traffic signal change intervals based on driver behavior*. 1985.

[12] Zegeer, C.V., *GREEN-EXTENSION SYSTEMS AT EHGI-I-SPEED INTERSECTIONS.* 1978.

[13] Liu, Y., et al., *Empirical observations of dynamic dilemma zones at signalized intersections.* Transportation Research Record: Journal of the Transportation Research Board, 2007. **2035**(1): p. 122-133.

[14] Wei, H., et al., *Quantifying Dynamic Factors Contributing to Dilemma Zone at High-Speed Signalized Intersections.* Transportation Research Record: Journal of the Transportation Research Board, 2011. **2259**(1): p. 202-212.

[15] Ghanipoor Machiani, S. and M. Abbas. *Dynamic Driver's Perception of Dilemma Zone: Experimental Design and Analysis of Driver's Learning in a Simulator Study*. in *The 93nd Annual Meeting of the Transportation Research Board*. 2014. Washington, DC.

[16] Ghanipoor Machiani, S. and M. Abbas, *Safety surrogate histograms (SSH): A novel real-time safety assessment of dilemma zone related conflicts at signalized intersections.* Accident Analysis & Prevention, 2015(0).

[17] Abbas, M., et al., *Modeling the Dynamics of Driver's Dilemma Zone Perception Using Machine Learning Methods for Safer Intersection Control*. 2014.

[18] Caird, J.K., et al., *The effect of yellow light onset time on older and younger drivers' perception response time (PRT) and intersection behavior.* Transportation research part F: traffic psychology and behaviour, 2007. **10**(5): p. 383-396.

[19] Li, H., H. Rakha, and I. El-Shawarby, *Designing Yellow Intervals for Rainy and Wet Roadway Conditions.* International Journal of Transportation Science and Technology, 2012. **1**(2): p. 171-190.

[20] Amer, A., H. Rakha, and I. El-Shawarby, *Novel stochastic procedure for designing yellow intervals at signalized intersections.* Journal of transportation engineering, 2011. **138**(6): p. 751-759.

[21] Jahangiri, A., H. Rakha, and T.A. Dingus. *Predicting Red-light Running Violations at Signalized Intersections using Machine Learning Techniques*. in *Transportation Research Board 93rd Annual Meeting*. 2015.

[22] Aoude, G.S., et al., *Driver behavior classification at intersections and validation on large naturalistic data set.* Intelligent Transportation Systems, IEEE Transactions on, 2012. **13**(2): p. 724-736.

[23] El-Shawarby, I., et al. *Age and gender impact on driver behavior at the onset of a yellow phase on high-speed signalized intersection approaches*. in *Transportation Research Board 86th Annual Meeting*. 2007.

[24] Liu, Y., G.-L. Chang, and J. Yu, *Empirical study of driver responses during the yellow signal phase at six maryland intersections.* Journal of transportation engineering, 2011. **138**(1): p. 31-42.

[25] Ghanipoor Machiani, S. and M. Abbas, *Assessment of Driver Stopping Prediction Models Before and After the Onset of Yellow Using Two Driving Simulator Datasets.* Accident Analysis & Prevention, 2015.

[26] Haque, M.M., et al., *Decisions and actions of distracted drivers at the onset of yellow lights.* Accident Analysis & Prevention, 2015.

[27] Elhenawy, M., et al., *Classification of driver stop/run behavior at the onset of a yellow indication for different vehicles and roadway surface conditions using historical behavior*, in *6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015*. 2015: Las Vegas, Nevada, USA.

[28] Doerzaph, Z.R., V. Neale, and R. Kiefer. *Cooperative intersection collision avoidance for violations: threat assessment algorithm development and evaluation method*. in *Transportation Research Board 89th Annual Meeting*. 2010.

[29] Mussa, R.N., et al., *Simulator evaluation of green and flashing amber signal phasing.* Transportation Research Record: Journal of the Transportation Research Board, 1996. **1550**(1): p. 23-29.

[30] Boyle, L.N. and J.D. Lee, *Using driving simulators to assess driving safety.* Accident Analysis & Prevention, 2010. **42**(3): p. 785-787.

[31] Peng, Y., et al., *Factors affecting glance behavior when interacting with in-vehicle devices: Implications from a simulator study.* 7th International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, Bolton Landing, NY, 2013.

[32] Dingus, T.A., et al., *The 100-car naturalistic driving study, Phase II-results of the 100-car field experiment*. 2006.

[33] Amer, A., H. Rakha, and I. El-Shawarby, *Agent-based stochastic modeling of driver decision at onset of yellow light at signalized intersections.* Transportation Research Record: Journal of the Transportation Research Board, 2011. **2241**(1): p. 68-77.

[34] Porter, B.E. and K.J. England, *Predicting red-light running behavior: a traffic safety study in three urban settings.* Journal of Safety Research, 2000. **31**(1): p. 1-8.

[35] Retting, R.A. and A.F. Williams, *Characteristics of red light violators: results of a field investigation.* Journal of Safety Research, 1996. **27**(1): p. 9-15.

[36] Porter, B.E. and T.D. Berry, *A nationwide survey of self-reported red light running: measuring prevalence, predictors, and perceived consequences.* Accident Analysis & Prevention, 2001. **33**(6): p. 735-741.

[37] Bonneson, J.A. and H.J. Son, *Prediction of expected red-light-running frequency at urban intersections.* Transportation Research Record: Journal of the Transportation Research Board, 2003. **1830**(1): p. 38-47.

[38] Neale, V.L. and C.C. McGhee, *Intersection decision support: evaluation of a violation Warning system to mitigate straight crossing path collisions*. 2006, Virginia Transportation Research Council.

[39] Retting, R.A., S.A. Ferguson, and C.M. Farmer, *Reducing red light running through longer yellow signal timing and red light camera enforcement: results of a field investigation.* Accident Analysis & Prevention, 2008. **40**(1): p. 327-333.

[40] Elhenawy, M., H. Rakha, and I. El-Shawarby. *Enhancing Driver Stop/Run Modeling at the Onset of a Yellow Indication using Historical Behavior and Machine Learning Techniques*. in *Transportation Research Board 93rd Annual Meeting*. 2014.

[41] Ghanipoor Machiani, S. and M. Abbas. *Predicting Drivers Decision in Dilemma Zone in a Driving Simulator Environment using Canonical Discriminant Analysis*. in *The 93nd Annual Meeting of the Transportation Research Board*. 2014. Washington, DC.

[42] Zhang, L., et al., *Prediction of red light running based on statistics of discrete point sensors.* Transportation Research Record: Journal of the Transportation Research Board, 2009. **2128**(1): p. 132-142.

[43] Liang, Y., M.L. Reyes, and J.D. Lee, *Real-time detection of driver cognitive distraction using support vector machines.* Intelligent Transportation Systems, IEEE Transactions on, 2007. **8**(2): p. 340-350.

[44] Kim, Z., *Robust lane detection and tracking in challenging scenarios.* Intelligent Transportation Systems, IEEE Transactions on, 2008. **9**(1): p. 16-26.

[45] Jahangiri, A. and H. Rakha, *Applying Machine Learning Techniques to Transportation Mode Recognition Using Mobile Phone Sensor Data.*

[46] Balali, V. and M. Golparvar-Fard. *Video-Based Detection and Classification of US Traffic Signs and Mile Markers using Color Candidate Extraction and Feature-Based Recognition*. in *Computing in Civil and Building Engineering (2014)*. ASCE.

[47] Yuan, F. and R.L. Cheu, *Incident detection using support vector machines.* Transportation Research Part C: Emerging Technologies, 2003. **11**(3): p. 309-328.

[48] Elmitiny, N., et al., *Classification analysis of driver's stop/go decision and red-light running violation.* Accident Analysis & Prevention, 2010. **42**(1): p. 101-111.

[49] McLaughlin, S.B., J.M. Hankey, and T.A. Dingus, *A method for evaluating collision avoidance systems using naturalistic driving data.* Accident Analysis & Prevention, 2008. **40**(1): p. 8-16.

[50] Zhang, L., et al., *Dynamic all-red extension at a signalized intersection: a framework of probabilistic modeling and performance evaluation.* Intelligent Transportation Systems, IEEE Transactions on, 2012. **13**(1): p. 166-179.

[51] Doerzaph, Z.R. and V. Neale. *Data acquisition method for developing crash avoidance algorithms through innovative roadside data collection*. in *Transportation Research Board 89th Annual Meeting*. 2010.

[52] NCHRP, *Guidelines for Timing Yellow and All-Red Intervals at Signalized Intersections*. 2012.

[53] Liaw, A. and M. Wiener, *Classification and Regression by randomForest.* R news, 2002. **2**(3): p. 18-22.

[54] Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.

[55] Hastie, T., et al., *The elements of statistical learning*. Vol. 2. 2009: Springer.

[56] Jahangiri, A., H.A. Rakha, and T.A. Dingus, *Adopting Machine Learning Methods to Predict Red-light Running Violations*, in *18th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. 2015.

[57] Ding, C. and H. Peng, *Minimum redundancy feature selection from microarray gene expression data.* Journal of bioinformatics and computational biology, 2005. **3**(02): p. 185-205.

[58] R Core Team, *R: A Language and Environment for Statistical Computing*. 2014, R Foundation for Statistical Computing.

Chapter 5: Bicycle Naturalistic Data Collection

Available online at www.sciencedirect.com

ScienceDirect

Procedia Manufacturing 00 (2015) 000–000

Procedia MANUFACTURING

www.elsevier.com/locate/procedia

6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015

Developing a System Architecture for Cyclist Violation Prediction Models Incorporating Naturalistic Cycling Data

Arash Jahangiri[a], Hesham A. Rakha[a]7, Thomas A. Dingus[a]

[a]*Virginia Tech Transportation Institute, 3500 Transportation Research Plaza, Blacksburg, VA 24061, USA*

Abstract

More than 40% of crashes that involve bicycles have occurred at intersections. According to the FARS database, an average of more than 30% of cyclist fatalities occurred at intersections from 2008 to 2012. Furthermore, up to 16% of bicycle-related crashes resulted from cyclist violations (i.e. bicyclists ride out at signalized or sign-controlled intersections). Not only has bicycle safety at intersections been a serious issue, but also the growing number of bicycle commuters makes the problem even more important. For example from 2000 to 2011, bicycle commuting rates in the US increased by 80% in large Bicycle Friendly Cities (BFCs), by 32% in non-BFCs, and overall by 47%. Moreover, improving bicycle safety should be considered more seriously to promote sustainable and eco-friendly modes of transport. For several different reasons (e.g. inattention, distraction, etc.), cyclists fail to obey traffic rules at both signalized and sign-controlled intersections. Hence the problem is how to prevent/mitigate these intersection-related crashes that involve bicycles. Failures to comply need to be identified before they occur so actions can be taken to alleviate the consequences. The focus of this paper is to present a system architecture that incorporates naturalistic cycling data to develop cyclist violation prediction models at intersections.

Peer-review under responsibility of AHFE Conference.

Keywords: Type your keywords here, separated by semicolons ;

* Corresponding author. Tel.: (540) 231-1505; fax: (540) 231-1555.
E-mail address: HRakha@vt.edu

Ch. 5 - Naturalistic Cycling Data Collection

Introduction

Crash data from 2005 to 2009 in North Carolina showed that 43.5 percent of the crashes that involved bicyclists occurred at intersections [1]. Similarly, from an older (early 1990's) but more comprehensive (Data from six US states) study, almost half of the bicycle-motor vehicle crashes took place at intersections [2]. This research was a Federal Highway Administration (FHWA) research study that was conducted by the University of North Carolina Highway Safety Research Center. The data set used in this study was a sample of crash data obtained from six US states. Moreover, according to the FARS database, average of more than 30% of cyclists' fatalities have occurred at intersections during the past 5 years (2008-20012). More specifically, the following crash types were recognized for the bicycle related crashes that occurred at intersections as shown in Table 28 [1, 3].

Table 28 Bicycle Crash Types at intersections

Crash Type	NC state (2005-2009)	Six US states (early 1990's)
Motorist drive out : Sign-Controlled Intersection	10.4%	9.3%
Bicyclist ride out : Sign-Controlled Intersection	6.6%	9.7%
Bicyclist ride out : Signalized Intersection	3.9%	7.1%
Motorist drive out: Signalized Intersection	2.8%	2%

As demonstrated by statistics, bicycle safety at intersections has been a serious issue. Further, the growing number of bicycle commuters makes the problem even more important; from 2000 to 2011, bicycle commuting rates in the US increased: by 80 percent in large Bicycle Friendly Cities (BFCs), by 32 percent in non-BFCs, and by the national average of 47 percent [4]. In addition, more attention should be given to enhancing bicycle safety to promote sustainable and eco-friendly modes of travel. For several different reasons (e.g. inattention, distraction, etc.), drivers and cyclists clearly fail to obey traffic rules at both signalized and sign-controlled intersections. Hence the problem is how to prevent/mitigate these intersection-related crashes that involves bicycles. The failure to comply need to be identified before they occur so actions can be taken to alleviate the consequences.

The remainder of the paper is organized as follows: first, past efforts on bicycle safety at intersections are reviewed. Then, the naturalistic cycling data collection method is presented along with the data analysis approach. Subsequently, the system architecture for developing cyclist violation prediction models is presented, and finally, conclusion and future work are presented.

Background

Only a few number of studies focused on the bicycle safety at intersections in the past; they can be categorized into two main parts: (1) examining countermeasures (2) investigating contributing factors.

Examining Countermeasures

In limited number of studies, a strategy was examined in order to reduce/mitigate crashes involving bicycles at intersections. Phillips, Bjørnskau [5] took 57 hours of video data at a Norwegian road–cycle path intersection to examine the effects of a cycle path. Yielding and conflict events were assessed 2 months, 4 years and 10 years following the introduction of the path, which resulted in a significant decrease in overall conflict levels after 4 years and further decrease after 10 years. Zhang and Wu [6] evaluated the impact of having a sunshield for cyclists at the intersection was evaluated at two sites across the city of Hangzhou, China. This was an observational study in which two video cameras were used to examine the crossing behavior of cyclists; 2477 riders were captured from the video recordings. Logistic regression and analysis of variance were applied to understand how the sunshield as a factor influenced red light running behavior. It was found that the red light infringement was reduced when having the sunshield both on sunny and cloudy days with the positive effect larger on sunny weather compared to the cloudy weather. As another example, the impact of a regulation change on bicycle safety was evaluated using video recording data. The videos were obtained before and after the regulation was imposed [7].

Investigating contributing factors

Contributing factors have been evaluated through analysis of three different kinds of data as presented below.

Naturalistic Data

In naturalistic data collections, video cameras are used unobtrusively to capture users' behavior, there is no experimenter, and no special instructions are given to the participants (if study needs participants). As a result, realistic behaviors of users are obtained [8]. Limited research has been conducted using naturalistic cycling data in the literature. Two naturalistic data collection methods have been used to study bicycle safety: (1) Collecting data through unobtrusive video cameras installed at infrastructure, and (2) Collecting data through instrumented bicycles. However, the results from the first approach cannot be generalized as the data collected in this approach are from certain locations and thus any conclusion might be location specific. In the studies in which the second approach was adopted, the data collection methods did not particularly target the cyclist violation behavior at intersections. Consequently, it appears that no sufficient data were collected to investigate these violations. In the present research, the second approach (i.e. data collection using instrumented bicycles) was adopted to ensure enough data can be collected; first, the potential participants were pre-screened to understand their weekly trip patterns by bicycles, and second, only those who ran through many intersections were recruited to run the experiment.

Collecting data through unobtrusive video cameras installed at infrastructure:

Johnson, Newstead [9] used video cameras at 10 locations throughout metropolitan Melbourne, Australia, for about 7 months to capture cyclists' behavior at intersections. They captured the crossing behavior of 4225 cyclists who faced the red light, of which 6.9% violated the red light. A single binary logistic regression analysis model was used for data analysis; the main predictive factor was found to be the direction of travel, turning left (in Australia, traffic travels on the left-side). Moreover, it was more likely that a cyclist violates when no other road user was present.

Wu, Yao [10] used video cameras at three signalized intersections in China to study the red light running behavior of bicycle riders; a total of 541 observations were captured of which 222 were e-bike riders and 229 were cyclists. Crossing behavior of the cyclists was classified into three distinct groups: law-obeying (49%) cyclists who would stop by obeying the red light, risk-taking (28%) cyclists who would ignore the red light and travel through the intersection without stopping (but may slow down), and opportunistic (23%) cyclists who would first wait at red lights but would not be patient enough to wait until the green light and subsequently cross the intersection as they would find gaps between crossing traffic. By applying logistic regression method, they found that age was a significant factor; the young and middle-aged riders were more likely to run against the red light than the old. Moreover, the following conditions increased the violation probability: when the rider was alone, when there were fewer riders waiting, and when there were riders already violating the red light [10].

Johnson, Charlton [11] captured 5,420 cyclists through video cameras at two intersections, of which the morning and afternoon red light violation rate were 3% and 11%, respectively. Similar to [10], three behavior types were recognized: (1) the "racers" who encountered an amber light, accelerated, but failed to pass before the light turns to red (25%), (2) the "impatients" who initially stopped, but then could not wait until the end of red phase (33%), and (3) the "runners" who rode through the red phase without stopping (42%). It was found that males are more likely than females to run against the red light and most of these male violators fell into the "runners" category.

Pai and Jou [12] installed video cameras at selected intersections in Taiwan to observe cyclists' behavior. using a mixed logit model, the following factors were found to increase the crash probability: intersections with short red-light duration, T/Y intersections, when riders were pupils in uniform, when riders were riding electric bicycles, and when riders did not use helmet. They adopted the cyclist behavior classifications as defined in [10], Out of 11,410 regular riders, 4.7% had risk-taking, 9.5% had opportunistic, and the rest (85.8%) had law-obeying behavior.

Collecting data through instrumented bicycles:

Inspired by the 100 car study [13], Integrated Vehicle Based Safety System (IVBSS) [14], and euroFOT [15], Dozza, Werneke [16], [17] conducted a naturalistic bicycle study. The instrumented bicycles used in the study were equipped with several sensors including two cameras, a GPS, a 3-axis accelerometer, a 3-axis gyroscope, and a 3-axis magnetometer, two pressure brake sensors, and a speed sensor. The criteria to participate were: age between 25 and 70, ride more than 40 minutes per day on weekdays, bicycle is the transportation mode they used for commuting, and the participant were asked not to carry children on the bicycle during the experiment. They designed and installed a human machine interface on the handlebar so the cyclist could record the time of critical

events (e.g. near-crashes and crashes) through a push button. It was shown how naturalistic data can be used to understand cycling behavior and bicycle dynamics. However, using a push button by the participant may negatively impact the naturalistic way of collecting data.

Johnson, Charlton [18] conducted a naturalistic bicycle study in Melbourne, Australia using techniques from the 100 car study [13]. The criteria for recruiting participants were: age over 18, regularly commuted (by cycling) to and from work, rode the majority (70%) of trips on the paved roads during commutes, could collect 12 hours of data over 4 weeks. It was found that: the most frequent event type was sideswipe (40.7%), the most events took place at an intersection/intersection-related location (70.3%), and the driver was the violator in the majority of events (87.0%) [18]. The data used in this study was limited to the video recordings and other sensors such as or Global Positioning System (GPS) and accelerometer were not adopted and thus analysis of kinematic information was not feasible.

Gustafsson and Archer [19] recruited 16 commuter cyclists to ride on 17 different major cycle routes for a bicycle study. Participants were required to ride the major part of their trips during morning (07:00 - 09:00) and afternoon (16:00 - 18:00) peak hours. Bicycles were equipped with GPS and cameras (installed on the handlebar) from which date, timestamp, GPS coordinates, speed, and video recordings were obtained. Hard-braking, swerving or acceleration to avoid a collision were considered to recognize safety problems for which important factors were identified such as: involvement of other road users, the event location, the responsible part, as well as the frequency of the events. Total of 220 safety problems, that included conflicting interactions such as Cycle-Car or Cycle-Bus interactions and other problems such as construction work or design of facility, were identified. However, no violations at intersections were mentioned in their paper. Their results showed that the most unsafe cycle-car conflicts took place at intersections and also lack of consideration from the drivers to the cyclists. Since each trip included several stops at traffic lights and other locations, participants were given strict instructions to obey traffic rules so the delays at intersections were counted. Nevertheless, this has a negative effect on the participants' natural crossing behavior at intersections where they occasionally may violate the red light or stop signs.

Police Reported Data

Schepers, Kroeze [20] classified Bicycle-Motor Vehicle (BMV) crashes of the police reported data into two categories based on who had priority: (1) type I crashes in which the cyclist had the priority, and (2) type II crashes in which the motorist had the priority. The focus of their study was on investigating the relationship between crashes and road factors. Results from negative binomial regression models showed that more crashes of type I was seen at intersections with two-way bicycle tracks, well-marked, and reddish colored bicycle crossings. Also, presence of raised bicycle crossings (e.g. speed hump) and other speed reducing measures were associated with less crashes of type I. Further, intersections with cycle track approaches deflected between 2 and 5 meters away from the main road corresponded to less crashes of type I. However, there were no road factors significantly affecting cashes of type II. Martínez-Ruiz, Lardelli-Claret [21] analyzed 19,007 collisions between a bicycle and another vehicle using police reported crash data in Spain. In these collisions, only one of the parties (the driver or the cyclist) violated the traffic law. Results from logistic regression and multinomial regression analyses showed that age from 10 to 19 years, male gender, alcohol or drug consumption, and non-helmet use increased the probability of crashes.

Surveys and Interviews

Lacherez, Wood [22] carried out a survey study in which 184 cyclists from Australia who had been involved in motor vehicle crashes were asked about visibility factors affecting bicyclist-motor-vehicle crashes. While the main focus of their paper was on the perceived cause of the collision, ambient weather and general visibility, as well as the clothing and bicycle lights used by the bicyclist, the most common sites in which the crashes took place were identified; sign-controlled intersections and signalized intersections were found to be the third (~17%) and fourth (~9%) common crash sites. Although the crash location information can also be obtained through other methods of data collection (e.g. police reported data), survey studies are beneficial as the public opinion and risk perceptions can reveal important information. For instance, it was found that the cyclist underrated visibility aids as a mean of improving traffic safety. In another study, Johnson, Charlton [23] surveyed 2061 Australian cyclists regarding behavioral, attitudinal and traffic factors contributing to red light infringement. A total of 37.3% reported they had violated a red light. Results from a multinomial logistic regression model showed that males are more likely to run through a red light and the old are less likely to violate the light. The following reasons were obtained from participants for riding against a red light: turning left (32%), which is consistent with the results from their other

work [9]; inductive loop detector failed to detect their bike (24.2%); absence of other road users (16.6%); at a pedestrian crossing (10.7%); and "Other" (16.5%) [23].

Important Factors

As found by different studies, red light infringement rates can be very different in different locations (7-9% in Melbourne [9], 56% in China [23], 21% in Taiwan [12], and self-reported rate of 38.4% in Brazil [9]). Even in the same country, riders may be more prone to run against a red light at an intersection compared to another intersection (e.g. large cities vs. small towns). Other important factors affecting cyclists' crossing behavior at intersections are as follows: Age [10, 12, 21, 23]; Gender [10, 11, 21, 23]; Direction of travel [9, 23]; Presence of other road users [9, 10, 23]; Signal timing [12]; Intersection type [12]; Helmet use [12, 21]; Detector failure [23]; Design characteristics [20]; and Consumption alcohol or drug [21].

Data Collection and Analysis

Cycling Naturalistic Data Collection System

Virginia Tech Transportation Institute (VTTI) is recognized as a pioneer in adopting naturalistic data collection by conducting the "100 car study" performed by Dingus, Klauer [13]. Typically, in a naturalistic data collection, passenger cars (or other modes of transport) are instrumented with a data acquisition system (DAS) and no special instruction is given to the drivers. Focusing on bicycles, VTTI has developed a smaller DAS (compared to the ones used in the "100 car study") as shown in Figure 37, known as Mini-DAS that has capabilities similar to the DAS used for passenger cars. The mini-DAS includes two cameras; one for capturing the forward roadway scene and the other captures the rider face and partial body. In addition, the mini-DAS contains sensors such as accelerometer, gyroscope, and GPS. To provide power, a removable battery was also designed that looks like a water bottle and needs to be charged occasionally.

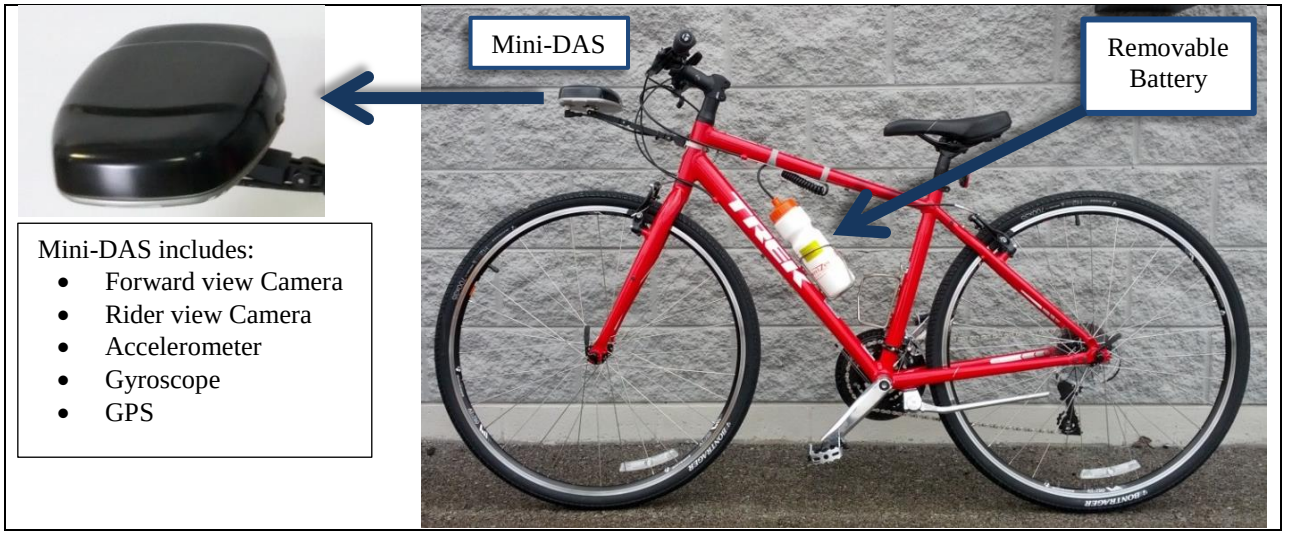


Figure 37 Naturalistic cycling data collection system

Data Visualization Tool

Hawkeye software[8] was used as a data visualization tool that has the capability of integrating and presenting different data (i.e. video data from both angles as well as sensor data) simultaneously as shown in Figure 38. On the left hand side, different variables are shown, and then video data from both angles are presented with their associated time stamps. To the right side of the videos selected variables are shown in diagrams (i.e. values against time stamps). The four diagrams in this figure illustrate acceleration along x axis, bicycle speed, latitude, and longitude. Finally, on the far right of the figure, different trips are listed. The software environment enables us to

[8] Hawkeye is a specialized software program developed by Virginia Tech Transportation Institute for data reduction

reduce data conveniently and extract additional variables that are useful for predicting violations (e.g. Time to Intersection (TTI)).

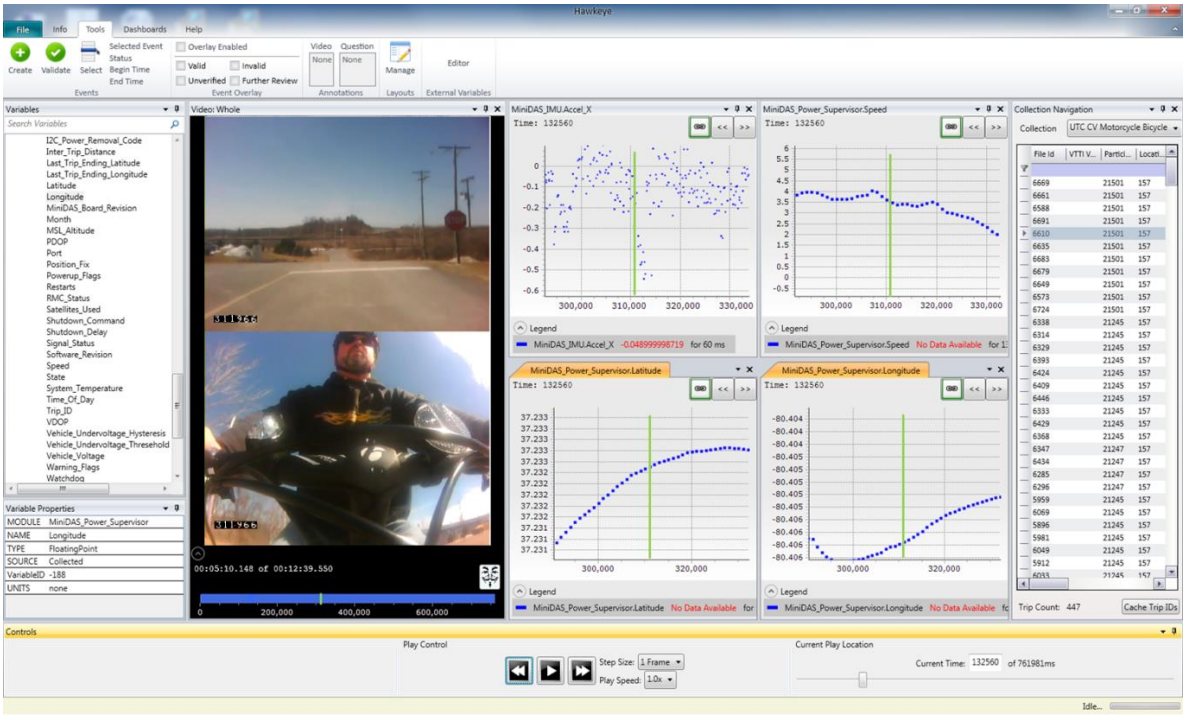


Figure 38 Hawkeye Software as Data Visualization tool for Data Reduction

Violation Prediction Models

Initial collected variables such as bicycle speed and acceleration as well as newly extracted variables such as TTI can be used to develop violation prediction models. In order to prevent/mitigate intersection-related crashes, these violations need to be identified before they occur, so appropriate actions can be taken. Machine learning techniques such as Support Vector Machine (SVM) and Random Forest (RF) can be applied to develop such models. In our previous work [24], we developed red light running (RLR) violation models for passenger cars. Violation prediction models for bicycles can be constructed in a similar fashion and is an ongoing task.

Intersection bicycle-car crash prediction system

This paper focuses on the system architecture for developing cyclist violation prediction models using naturalistic data and discusses different system components in a connected environment as shown in Figure 39. Vehicle to Vehicle (V2V) and Vehicle to Infrastructure (V2I) technology has been a highly active area of research. However, the focus has been more on passenger cars and thus bicycle as an important transportation mode has been given less attention. Hence, in this paper, bicycles are incorporated into the connected environment. For bicycles, V2X is used which is an acronym referring to “vehicle to other” (i.e. pedestrian, bicycle, etc.). The data collection system that was discussed in previous sections is used to collect and extract required variables to develop violation prediction models. However, in practice the prediction model is already developed; in other words the model development is conducted using historical data and therefore the mini-DAS or OBE (on-board equipment) does not need to collect video data. In fact, video data are only required for model development and after the prediction model is constructed, bicycles and other transportation modes only need to send their sensor data (e.g. speed, acceleration, location) to the infrastructure as depicted in Figure 39. The missing data such as TTI at yellow onset that was extracted using video data for model development is still needed as an input to the developed prediction model. Therefore, signal phase and timing, known as SPaT should be provided to obtain such missing data as shown in the figure.

Violation prediction models for different modes of transport should be constantly monitoring individuals approaching the intersection. When a potential threat is predicted, different actions can be taken depending on the situation; in situations where the endangered driver (or rider) has sufficient time, a warning can be issued and sent from roadside equipment (RSE) to the driver (or rider) to respond. In cases where not enough time is available, the infrastructure can take appropriate actions by changing the signal control through the traffic light (e.g. providing all red clearance) as shown in Figure 39.

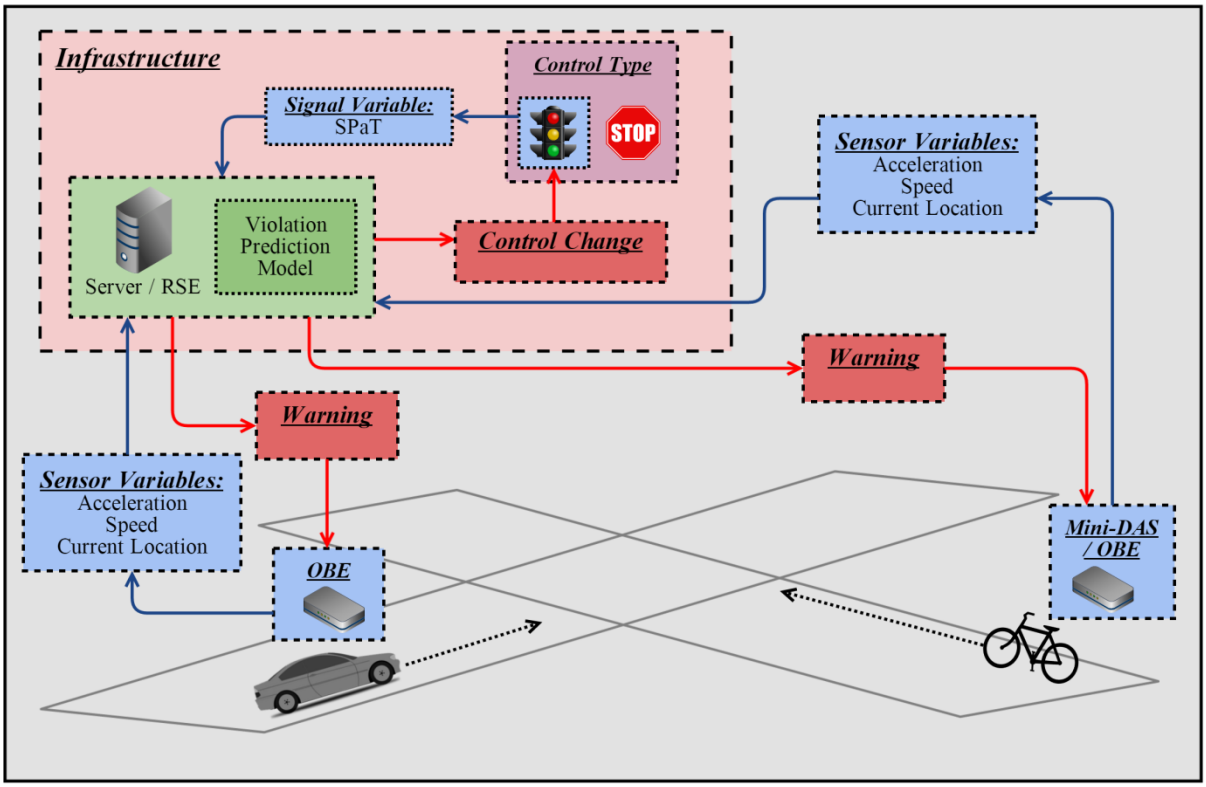


Figure 39 Intersection bicycle-car crash prediction: System Architecture

Conclusions and Future Work

This paper presented the naturalistic cycling data collection system that can be used to develop bicycle violation prediction models. For model development, required data were extracted using naturalistic data collection. Further, Hawkeye software as a data visualization tool was employed for data reduction. Subsequently, the system architecture that embodied such violation models was demonstrated. It was shown how connected vehicle technology can be adopted for different parts to communicate amongst themselves. Communication between different system entities was shown to have different purposes: (1) sending required variables for violation prediction models such as bicycle speed, acceleration, current location, and SPaT (2) sending warning to the drivers/riders in potential danger (3) sending a “control change” order to change the signal setting. Future work will focus on developing violation prediction models for cyclists. As mentioned earlier, we already developed violation prediction models for passenger cars. Therefore, to complement our previous work [24], we are applying machine learning techniques to develop cyclist violation prediction models. These models are currently under development and will be completed by the end of this year.

Acknowledgements

This research effort was funded by the Connected Vehicle Initiative UTC (CVI-UTC).

References

[1] The-University-of-North-Carolina-Highway-Safety-Research-Center. *North Carolina Bicycle Crash Types 2005 - 2009*. 2011; Available from: http://www.ncdot.gov/bikeped/download/summary_bike_types05-09.pdf.

[2] Hunter, W.W., et al., *Pedestrian and bicycle crash types of the early 1990's*. 1996.

[3] Tan, C., *Crash-type manual for bicyclists.* Publication No. FHWA-RD-96-104, 1996.

[4] McLeod, K., D. Flusche, and A. Clarke, *Where We Ride: Analysis of Bicycling in American Cities.* 2013.

[5] Phillips, R.O., et al., *Reduction in car–bicycle conflict at a road–cycle path intersection: Evidence of road user adaptation?* Transportation research part F: traffic psychology and behaviour, 2011. **14**(2): p. 87-95.

[6] Zhang, Y. and C. Wu, *The effects of sunshields on red light running behavior of cyclists and electric bike riders.* Accident Analysis & Prevention, 2013. **52**: p. 210-218.

[7] Räsänen, M., I. Koivisto, and H. Summala, *Car driver and bicyclist behavior at bicycle crossings under different priority regulations.* Journal of Safety Research, 1999. **30**(1): p. 67-77.

[8] Neale, V.L., et al., *An overview of the 100-car naturalistic study and findings.* National Highway Traffic Safety Administration, Paper, 2005(05-0400).

[9] Johnson, M., et al., *Riding through red lights: The rate, characteristics and risk factors of non-compliant urban commuter cyclists.* Accident Analysis & Prevention, 2011. **43**(1): p. 323-328.

[10] Wu, C., L. Yao, and K. Zhang, *The red-light running behavior of electric bike riders and cyclists at urban intersections in China: an observational study.* Accident Analysis & Prevention, 2012. **49**: p. 186-192.

[11] Johnson, M., J. Charlton, and J. Oxley. *Cyclists and red lights—a study of the behaviour of commuter cyclist in Melbourne*. in *Australasian Road Safety Research, Policing and Education Conference, Adelaide*. 2008.

[12] Pai, C.-W. and R.-C. Jou, *Cyclists' red-light running behaviours: An examination of risk-taking, opportunistic, and law-obeying behaviours.* Accident Analysis & Prevention, 2014. **62**: p. 191-198.

[13] Dingus, T.A., et al., *The 100-car naturalistic driving study, Phase II-results of the 100-car field experiment*. 2006.

[14] Sayer, J., et al., *Integrated vehicle-based safety systems field operational test final program report.* 2011.

[15] Benmimoun, M., et al. *Safety Analysis Method for Assessing the Impacts of Advanced Driver Assistance Systems within the European Large Scale Field Test euroFOT*. in *8th ITS European Congress, Lyon, France*. 2011.

[16] Dozza, M., J. Werneke, and A. Fernandez. *Piloting the naturalistic methodology on bicycles*. in *Proceeding ot the st International Cycling Safety Conference, Helmond NL, Nov 7-8 2012.* 2012.

[17] Dozza, M. and A. Fernandez, *Understanding Bicycle Dynamics and Cyclist Behavior From Naturalistic Field Data (November 2012).* 2014.

[18] Johnson, M., et al. *Naturalistic cycling study: identifying risk factors for on-road commuter cyclists*. in *Annals of Advances in Automotive Medicine/Annual Scientific Conference*. 2010. Association for the Advancement of Automotive Medicine.

[19] Gustafsson, L. and J. Archer, *A naturalistic study of commuter cyclists in the greater Stockholm area.* Accident Analysis & Prevention, 2013. **58**: p. 286-298.

[20] Schepers, J., et al., *Road factors and bicycle–motor vehicle crashes at unsignalized priority intersections.* Accident Analysis & Prevention, 2011. **43**(3): p. 853-861.

[21] Martínez-Ruiz, V., et al., *Risk factors for causing road crashes involving cyclists: An application of a quasi-induced exposure method.* Accident Analysis & Prevention, 2013. **51**: p. 228-237.

[22] Lacherez, P., et al., *Visibility-related characteristics of crashes involving bicyclists and motor vehicles–Responses from an online questionnaire study.* Transportation research part F: traffic psychology and behaviour, 2013. **20**: p. 52-58.

[23] Johnson, M., et al., *Why do cyclists infringe at red lights? An investigation of Australian cyclists' reasons for red light infringement.* Accident Analysis & Prevention, 2013. **50**: p. 840-847.

[24] Jahangiri, A., H. Rakha, and T.A. Dingus, *Predicting Red-light Running Violations at Signalized Intersections Using Machine Learning Techniques*. 2015.

Chapter 6: Bicycle Violation Prediction

Investigating Cyclist Violations at Intersections using Naturalistic Cycling data

Arash Jahangiri
Center for Sustainable Mobility, Virginia Tech Transportation Institute
3500 Transportation Research Plaza, Blacksburg, VA 24061
E-mail: ArashJ@vt.edu
Phone: (540) 200-7561

Hesham Rakha (corresponding author)
Center for Sustainable Mobility, Virginia Tech Transportation Institute
3500 Transportation Research Plaza, Blacksburg, VA 24061
E-mail: HRakha@vtti.vt.edu
Phone: (540) 231-1505

Thomas A. Dingus
Virginia Tech Transportation Institute
3500 Transportation Research Plaza, Blacksburg, VA 24061
E-mail: TDingus@vtti.vt.edu
Phone: (540) 231-1501

Ch. 6 - Cyclist Violation Prediction

Abstract

Improving bicycle safety is considered as a growing concern for two reasons. First, in the United States in in recent years, about 700 cyclists were killed and about 48000 were injured in bicycle motor vehicle crashes each year, which in total resulted in over four billion dollars per year. Regarding crash location, from 2008 to 2012 in the United States, more than 30% of cyclist fatalities occurred at intersections. Furthermore, up to 16% of bicycle-related crashes were due to cyclist violations at intersections. Second, from 2000 to 2011, bicycle commuting rates in the United States has increased: by 80% in large Bicycle Friendly Cities (BFCs), by 32% in non-BFCs, and overall by 47%. Also, cycling as one of the sustainable and eco-friendly modes of transport is receiving more attention than before. Thus, more research is needed to improve bicycle safety. The focus of this paper is on investigating factors affecting cyclist behavior and predicting cyclist violations at intersections. For several different reasons (e.g. inattention, distraction, etc.), cyclists fail to obey traffic rules at both signalized and stop-controlled intersections. Hence the problem is how to prevent/mitigate these intersection-related crashes that involve bicycles. Failures to comply need to be identified before they occur so actions can be taken to alleviate the consequences. Naturalistic cycling data were used to assess the feasibility of developing cyclist violation prediction models. Based on logistic regression analysis, movement type and presence of other users were found as significant factors affecting the probability of red light violations by cyclists. Also, presence of other road users and Age were the significant factors affecting violations at stop-controlled intersections. In case of stop-controlled intersections, violation prediction models were developed based on kinetic information of cyclists approaching the intersection. Prediction error rates of 0 to 10 percent were obtained depending on how far from the intersection the prediction task is conducted. The error rate of 6 percent was obtained when the violating cyclist is at time to intersection of about 2 seconds which is sufficient for most endangered riders/drivers to respond.

Keywords: cyclist violation; intersections; violation prediction; logistic regression; random forest; machine learning

Ch. 6 - Cyclist Violation Prediction

Introduction

According to National Highway Traffic Safety Administration (NHTSA) report, in the United States in 2013, 742 cyclists were killed and about 48000 were injured in bicycle motor vehicle crashes, which in total resulted in over four billion dollars per year [1]. Although bicycle fatalities represent less than two percent of total traffic fatalities, trips made by bicycles constitute only one percent of all trips. There is also considerable uncertainty as to how cycling compares to other modes of transport in terms of safety. The uncertainty is due to unreliable source of exposure data because there is no data regarding bicycle miles traveled and the time it takes to travel those miles. Regarding crash locations, from 2008 to 2012 in the U.S., average of more than 30% of cyclists' fatalities have occurred at intersections [2]. Particularly, cyclist violations at intersections resulted in up to %16 of bicycle-related crashes [3, 4]. Therefore, more attention should be given to enhancing bicycle safety not only to reduce bicycle related crashes, but also to promote this sustainable and eco-friendly mode of travel. Cycling is already a growing activity and is going to attract more people in the future; from 2000 to 2011, bicycle commuting rates in the US increased: by 80 percent in large Bicycle Friendly Cities (BFCs), by 32 percent in non-BFCs, and by the national average of 47 percent [5].

Many different factors influence cyclist violation behavior at intersections such as Age [6-9]; Gender [6, 7, 9, 10]; Direction of travel [7, 11]; Presence of other road users [6, 7, 11]; Signal timing [8]; Intersection type [8]; Helmet use [8, 9]; Detector failure [7]; Design characteristics [12]; and alcohol or drug Consumption [9]. Location is another factor that shows violation rates could be significantly different for different locations as presented in Table 29.

Table 29 Violation rates in different countries

No	Violation rate	Country	Reference
1	7-9%	Australia	[11]
2	56%	China	[7]
3	21%	Taiwan	[8]
4	38.4%	Brazil (self-reported)	[11]

For several different reasons (e.g. inattention, distraction, etc.), drivers and cyclists clearly fail to obey traffic rules at both signalized and stop-controlled intersections. Hence the problem is how to prevent/mitigate these intersection-related crashes that involves bicycles. The failure to comply need to be identified before they occur so actions can be taken to alleviate the consequences. In this research, naturalistic cycling data were adopted to assess the feasibility of developing cyclist violation prediction models at intersections. The present work is the continuation of a previous work in which a system architecture was presented that incorporates naturalistic cycling data for developing cyclist violation prediction models [13].

The remainder of the paper is organized as follows: first, relevant work on bicycle safety at intersections is reviewed. Then, the naturalistic cycling data collection method is presented followed by the model development section. Subsequently, the system architecture for developing cyclist violation prediction models is reviewed. Finally, results and conclusions are presented.

Literature review

The focus of the present study is on feasibility analysis of developing cyclist violation prediction models at intersections and evaluation on different factor impacts. Relevant studies are divided into two parts and are briefly presented in this section.

Examining countermeasures

A limited number of studies focused on introducing countermeasures to reduce bicycle related crashes at intersections. Phillips, Bjørnskau [14] examined the impacts of having a cycle path at a cycle-road intersection based on the change in the number of yielding and conflict events after introducing the path.

Zhang and Wu [15] adopted logistic regression and analysis of variance to assess the effects of having sunshields for cyclists at two intersections. Räsänen, Koivisto [16] conducted a before-after study to evaluate the impact of a regulation change on bicycle related crashes.

Investigating contributing factors

Many studies focused on assessing factors and conditions that influence crossing behavior of cyclists at intersections. These studies are categorized based on the type of data they employed as follows.

Naturalistic data through unobtrusive video cameras installed at infrastructure

In this approach, video cameras are installed at intersections to capture how cyclists approach intersections. Since individuals are unaware of the data collection, cyclist behavior is realistic. However, the results may not be generalized if the data collected are for limited locations. In the following studies as summarized in Table 30, video cameras were used to capture crossing behavior of cyclists. Using a single binary logistic regression analysis, Johnson, Newstead [11] found that the most important factor to predict red light runners is the direction of travel, turning left (in Australia, traffic travels on the left-side) which is equivalent to turning right in the United States. By applying logistic regression, Wu, Yao [6]found the following factors to increase the probability of running red lights: younger riders, when the rider was alone, when there were fewer riders waiting, and when there were riders already violating the red light. Johnson, Charlton [10] found gender to be an important factor with males more likely to violate. Pai and Jou [8] used a mixed logit model to examine different factors. Factors to increase the crash probability were found to be intersections with short red-light duration, T/Y intersections, when riders were pupils in uniform, when riders were riding electric bicycles, and when riders did not use helmet.

Table 30 Data collection through video cameras at intersections - summary of past studies

No	Observations	Behavior classification of riders	Reference
1	4225	▪ 6.9% violation rate - no further classification	Johnson, Newstead [11]
2	229	▪ 56% violation rate ▪ Risk-taking[a] (28%); law-obeying[b] (49%); opportunistic[c] (23%)	Wu, Yao [6]
3	5,420	▪ 3% morning 11% afternoon violation rates ▪ Racers[d] (25%); runners[e] (42%); impatients[f] (33%)	Johnson, Charlton [10]
4	11,410	▪ Risk-taking (4.7%); law-obeying (85.8%); opportunistic (9.5%)	Pai and Jou [8]

[b] **law-obeying:** cyclists who would stop by obeying the red light
[a] **risk-taking:** cyclists who would ignore the red light and travel through the intersection without stopping (but may slow down)
[c] **opportunistic:** cyclists who would first wait at red lights but would not be patient enough to wait until the green light and subsequently cross the intersection as they would find gaps between crossing traffic.
[d] **racers:** who encountered an amber light, accelerated, but failed to pass before the light turns to red
[e] **impatients:** who initially stopped, but then could not wait until the end of red phase
[f] **runners:** who rode through the red phase without stopping

Naturalistic Data through instrumented bicycles

Bicycles are instrumented in this approach so their data can be collected for cyclists' entire trips. No special instructions as to how to ride, when to ride, and where to ride are given to the participants. Thus, the data collected in this approach not only reflects realistic cyclist behavior, it also leads to better generalization as the data include many different locations. There have been a few studies that adopt naturalistic cycling data collection techniques similar to the 100 car study [17], Integrated Vehicle Based Safety System (IVBSS) [18], and euroFOT [19]. Table 31 presents a summary regarding these studies.

Table 31 Naturalistic cycling data collection - summary of past studies

No.	Equipment	Participant criteria	Reference
1	two cameras, a GPS, a 3-axis accelerometer, a 3-axis gyroscope, a 3-axis magnetometer, two pressure brake sensors, and a speed sensor	Age between 25 and 70, ride more than 40 minutes per day on weekdays, bicycle is the transportation mode they used for commuting, and the participant were asked not to carry children on the bicycle during the experiment.	Dozza, Werneke [20], [21]
2	Video cameras	Age over 18, regularly commuted by cycling to and from work, rode the majority (70%) of trips on the paved roads during commutes, could collect 12 hours of data over 4 weeks.	Johnson, Charlton [22]
3	GPS and cameras	Commuter cyclists to ride on 17 different major cycle routes. Participants were required to ride the major part of their trips during morning (07:00 - 09:00) and afternoon (16:00 - 18:00) peak hours.	Gustafsson and Archer [23]

To identify critical events such as crashes and near crashed, Dozza, Werneke [20], [21] designed and installed a human machine interface on the handlebar so the cyclist could record the time of such events using a push button. However, giving participants extra tasks than the riding could impact the realistic riding behavior. Johnson, Charlton [22] identified sideswipe as the most frequent event type (40.7%) and regarding the event location, the majority of events occurred in intersection/intersection-related locations. Also, drivers were recognized as violators in most of the cases. However, Johnson, Charlton [22] did not uses sensors in their data collection system to obtain and analyze kinetic data. Gustafsson and Archer [23] defined safety problems as Hard-braking, swerving or acceleration to avoid a collision. For these safety problems, factors such as involvement of other road users, the event location, the responsible part, as well as the frequency of the events were identified. Although it was shown that the most unsafe cycle-car conflicts took place at intersections no violations at intersections was reported. Also, the participants were given strict instructions to obey traffic rules at intersections. Nevertheless, this negatively impacts participants' realistic riding behavior.

Police Reported Data

Schepers, Kroeze [12] used police reported data to study two bicycle motorist crash types at stop-controlled intersections. Crashes in which the cyclist had the priority were classifies as type I and the crashes in which the motorist had the priority were classifies as type II. They employed negative binomial regression models to study factors such as intersections with two-way bicycle tracks, well-marked, reddish colored bicycle crossings, presence of raised bicycle crossings (e.g. speed hump), and other speed reducing measures. Martínez-Ruiz, Lardelli-Claret [9] applied logistic regression and multinomial regression analyses to study 19,007 collisions between a bicycle and another vehicle using police reported crash data. age younger cyclists (i.e., 10-19), male gender, alcohol or drug consumption, and non-helmet use were identified as factors to increase the risk of crashes.

Surveys and Interviews

Comparing to other data collection methods, surveys and interviews are important as public opinion and risk perceptions can be obtained. A survey study was carried out by Lacherez, Wood [24] to investigate visibility factors that impact the crashes [24]. 184 cyclists who had been involved in motor vehicle crashes were surveyed in this study. Moreover, regarding crash locations, stop-controlled intersections and signalized intersections were found to be the third (~17%) and fourth (~9%) common crash sites. Red light infringement was examined in another survey study [7]. Out of 2061 cyclist in the study, 37.3% reported that they ran against a red light. Participants answered the following reasons for their violation behavior: turning left (32%), inductive loop detector failed to detect their bike (24.2%); absence of other

road users (16.6%); at a pedestrian crossing (10.7%); and "Other" (16.5%). Also, males and younger participants were associated with higher violation probability based on a multinomial logistic regression analysis.

Naturalistic cycling experiment

The data used in the present research is the naturalistic data collected through instrumented bicycles. It appears that the past naturalistic cycling studies did not particularly target the cyclist violation behavior at intersections. similar to naturalistic driving studies (e.g. "100 car study" performed by Dingus, Klauer [17]), bicycles (instead of passenger cars) were instrumented and given to the participants to ride. The participants were asked to ride the bicycles as they normally would without providing any special instructions. The naturalistic cycling experiment was conducted in three steps as follows.

Pre-screening

Because the objective of the present study was to investigate cyclist violation behavior at intersections, the potential participants were pre-screened to understand their weekly trip patterns by bicycles. Subsequently, only those who ran through many intersections were recruited to run the experiment. To identify these cyclists, a series of questions were asked over the phone. After reviewing the answers, those who encountered maximum number of intersections were selected. Other factors to identify eligible participants included:

- Participants must make most of their trips on paved roads rather than on sidewalks and bicycle trails.
- Participants must commute/travel by bicycle at least 3 times per week on average in the Blacksburg, Christiansburg, and/or Radford areas
- Participants were not allowed to transport children by the instrumented bicycle
- Participants must be 18-30 or 45-65 years of age from both gender

Data collection

Virginia tech Transportation Institute (VTTI) developed a data acquisition system (DAS), called min-DAS, to instrument bicycles for data collection as shown in Figure 40. The original mini-DAS included two cameras (e.g. one for forward roadway view from bike and the other for rider view) and sensors such as accelerometer, gyroscope, and GPS. A removable battery placed in the water bottle cage was included, which needed to be charged by the participants using a battery charger. The bicycles were hybrid models (Trek 7.2 FX) available in three sizes: small (15"), medium (17.5"), and large (20"). All the participants needed to do was to make sure the battery was charged and to make sure they had the mini-DAS turned on while riding. Analysis on sample data showed that the acceleration data had too much noise (i.e. due to possible DAS vibration when riding) and difficult to work with. Therefore, a speed sensor was added as shown in Figure 40 that basically measures the distance as the bicycle wheels roll. Consequently, the speed and acceleration data were derived from the distance data.

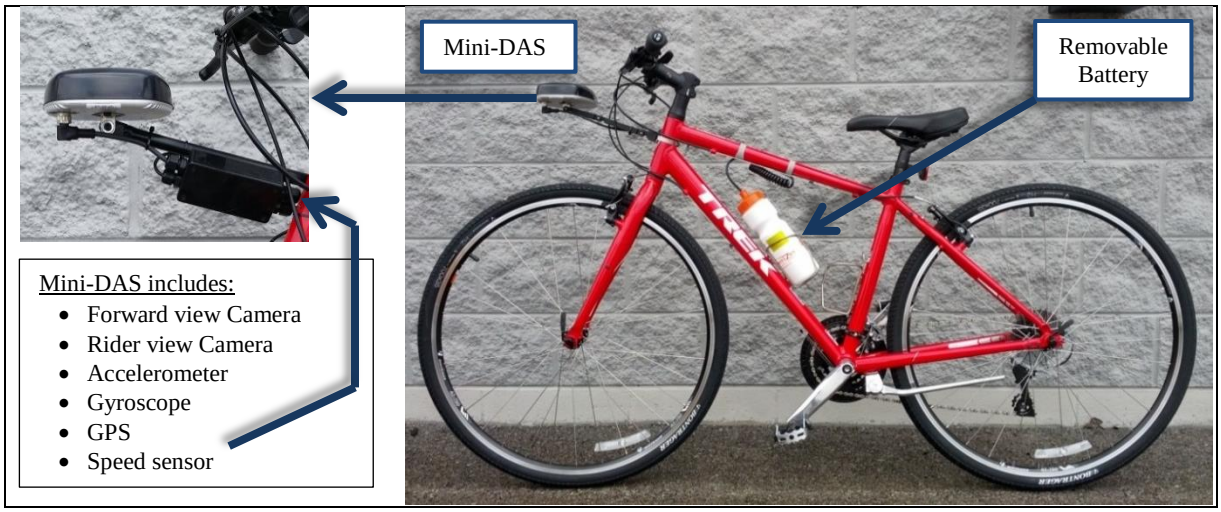


Figure 40 Naturalistic cycling data collection system

Data reduction

Hawkeye software, a data visualization tool, was used for data reduction. An "event" was defined as crossing an intersection and for each event, several variables were extracted. Table 32 lists the variables obtained through data reduction for all events. As the data collection is still an ongoing task, the data used in the present research include data reduced for 7 participants (i.e., 3 participants with 4 weeks and 4 participants with 2 weeks of data).

Table 32 variables obtained from data reduction

No	Variable	No	Variable
1	Time 1 (morning/noon/evening)	6	Weather 1(warm/cool)
2	Time 2 (weekend/weekday)	7	Weather 2(cloudy/rainy/clear)
3	Road slope (uphill/downhill/flat)	8	yellow onset
4	Movement type (right/through/left)	9	red onset
5	presence of other road users(side/opposing/front/adjacent)		

Model development

Two methods, namely multivariate logistic regression (MLR) and random forest (RF), were applied for model development. Relationship models were developed using the MLR approach to investigate impacts of different factors. Violation prediction models were developed using the RF technique which is a supervised learning approach. The prediction problem is a behavioral classification with binary responses (i.e. 1 as violation 0 as compliance). Brief overviews of these approaches are provided as follows.

Multivariate logistic regression (MLR)

MLR is applied for predicting a binary response using multiple variables. The logistic regression model produces the probability that an observation belongs to a particular response as presented in Equation 35. In the case of the present research, an observation is explained with multiple variables extracted in the data reduction. Also, the binary response in this case is defined as cyclist violation (i.e., denoted by 1) versus cyclist compliance (i.e., denoted by 0) at intersections. The goal is to develop a relationship model to investigate variable impacts on violation behavior of cyclists at intersections.

$$p(Y=1|X) = \frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n}}{1+e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n}} \qquad \text{Equation 35}$$

Where,

X	$(X_1, \dots, X_n)$
$X_1, \dots, X_n$	n variables
$\beta_0, \dots, \beta_n$	Model coefficients
Y	Binary response
$p(Y = 1\|X)$	Probability that a response is 1 given X

Random forest (RF)

RF as a supervised machine learning technique was applied to predict cyclists' violations. Random forest was introduced by Breiman [25], and is considered as an ensemble learning approach based on decision tree method; RF creates a number of decision trees and the outcome is obtained from averaging the results from each tree in the case of regression problems or taking the majority votes in the case of classification problems. To grow each tree, data are divided into two parts in several steps until a desirable separation between classes (i.e., the two classes in the case of present research are violation and compliance) is achieved. Data separation in each step is carried out by Recursive Binary Splitting method in which different criteria can be used to split the data. Gini index as one of the recommended approaches was employed in this study as presented in Equation 36 [26].

$$G = \sum_{k=1}^{K} P_k^m (1 - P_k^m) \qquad \text{Equation 36}$$

Where,

$$P_k^m = \frac{1}{N^m} \sum_{x_i^m} I(y_i^m = k)$$

N^m	Number of observations received at node m
y_i^m	The response value corresponding to the observation i at node m
x_i^m	The feature vector corresponding to the observation i at node m
k	class

Monitoring period

Similar to driver violation prediction models as developed by Jahangiri, Rakha [27], monitoring periods were defined and data such as speed and acceleration were obtained to develop the prediction models. Figure 41 illustrates different variables for defining a monitoring period for modeling violation prediction models. A cyclist (i.e., violating bicycle) is going to violate at an intersection on one approach and as a result, a vehicle or cyclist (i.e., endangered vehicle/bicycle) on a conflicting approach is going to be at risk of a right-angle crash. The monitoring period (i.e., t_{mon}^{v}, from point *a* to point *b*) is defined by its start (i.e., point *a*) and end points (i.e., point *b*). The start point should not be selected too early to exclude unnecessary information. The end point is restricted by t_{min}^{v} (i.e., from point *b* to point *c*), which is the minimum time required for the endangered vehicle/bicycle to avoid the possible crash. Therefore, t_{min}^{v} is equivalent to the time required for the endangered driver/rider to react (i.e., $t_{driver/rider}^{e}$, from point *x* to point *y*) and the time required for the endangered vehicle/bicycle to come to a complete stop (i.e., $t_{vehicle/bicycle}^{e}$, from point *y* to point *z*). According to a distribution of the human response time as presented by McLaughlin, Hankey [28], values from 0.5 to 2.5 seconds were chosen for the driver/rider response time (i.e., $t_{driver/rider}^{e}$) that corresponds to about 5 to 95 percent of population to respond. It should be noted that the distribution of the human response time is for

the drivers. However, response times for the cyclists were shown to be similar to the drivers as discussed by Landis, Petritsch [29]. Also, values from 1.9 to 3.4 seconds (in case of passenger cars) and 1.6 to 2.8 seconds (in case of bicycles) can be considered for the vehicle/bicycle response time (i.e., $t^{e}_{vehicle/bicycle}$) that corresponds to vehicles approaching at velocities from 25 to 45 mph and bicycles approaching at velocities from 20 mph to 35mph, respectively[30][9]. However, in case of 4-way stop signs, the endangered vehicle/bicycle is going to stop regardless of what the violating bicycle indents to do (i.e., normal crossing at stop signs). Consequently, there is no need to account for the vehicle/bicycle response time and $t^{e}_{vehicle/bicycle}$ would be zero. In the 2-way stop signs, however, the $t^{e}_{vehicle/bicycle}$ should also be added.

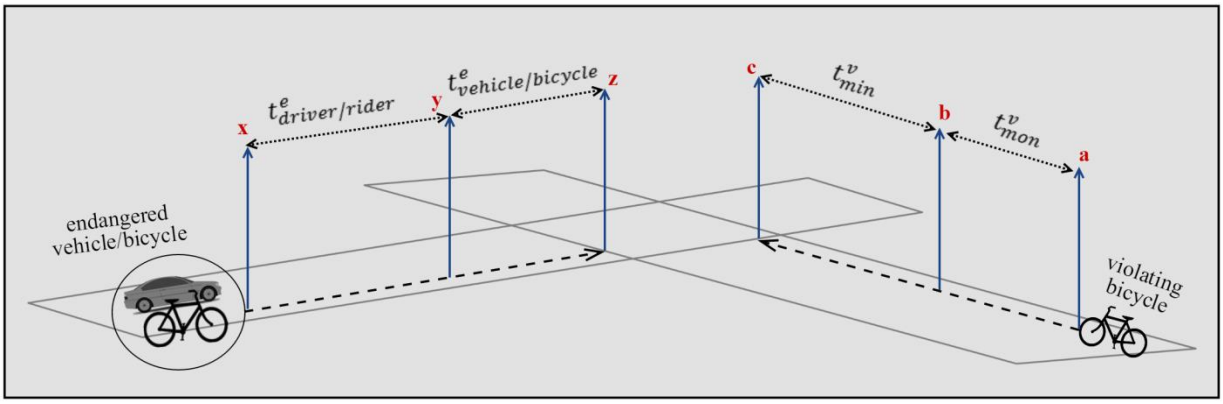


Figure 41 Factors to define a monitoring period

Variable selection

In addition to the variables extracted from the data reduction step, speed and acceleration data within the monitoring period were obtained from the mini-DAS. Subsequently, statistical measures (e.g., mean, range, max, etc.) were applied within the monitoring period to create kinetic related variables by which cyclist behavior (i.e., stopping vs proceeding) could be explained.

Intersection bicycle-car crash prediction system

Figure 42 presents a system architecture that incorporates violation prediction models developed using naturalistic cycling data at intersections. Bicycle/vehicle-to-infrastructure communication is required to send variables such as speed and acceleration to the violation prediction models implemented on the infrastructure side. When a potential violation is predicted, warnings can be issued for the users in potential danger or the intersection control settings can be changed in the case of signalized intersections.

[9] for example, a bicycle with approaching speed of 35 mph or 15.64 m/s^2 and maximum deceleration rate of 5.5 m/s^2, requires 2.8 seconds to stop: $t^{e}_{vehicle/bicycle} = velocity/maximum\ acceleration = 15.64/5.5 = 2.8$ s

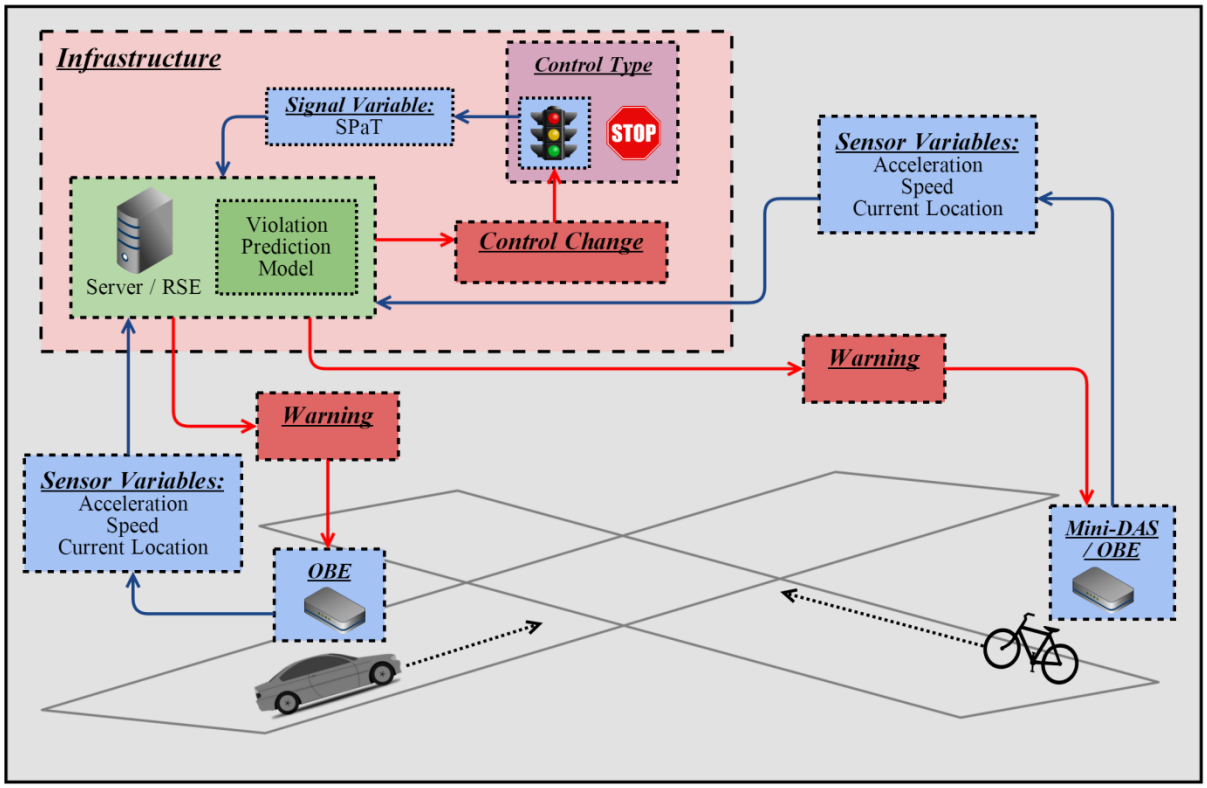


Figure 42 Intersection bicycle-car crash prediction: System Architecture

Results

After conducting data reduction for 7 participants as explained earlier, 718 crossings at signal-controlled intersections and 875 crossings at stop-controlled intersections were obtained. Hence, this section is divided into two parts: (1) signal-controlled intersections and (2) stop-controlled intersections.

Signal-controlled intersections

Out of 718 crossings at signalized intersections, 300 encountered red lights, of which 80 violated the red lights, resulted in 26.67% violation rate. Total of seven different violation types were identified as shown in Table 33 and Table 34. Based on the logistic regression model, movement type (i.e., right turns), presence of other users (i.e., side and front) were found as significant factors as presented in Table 35. This shows that it is more likely to violate a red light when making right turns. Also, the probability of red light violation decreases when there is side traffic or traffic in front.

Table 33 Cyclist violation types part 1

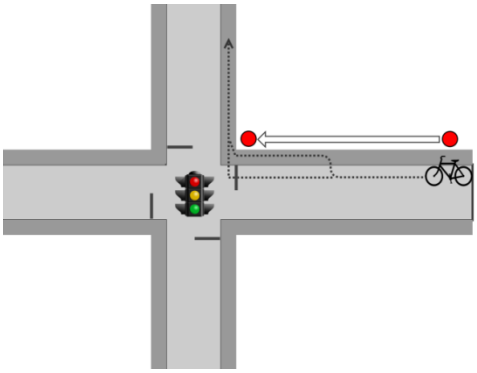
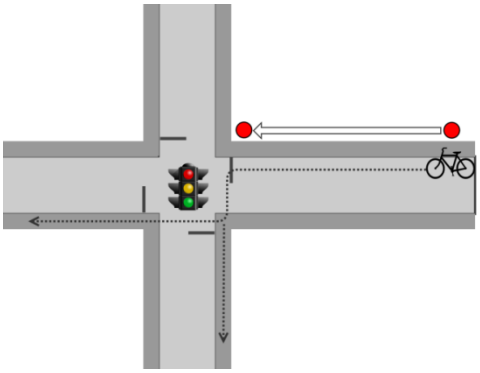
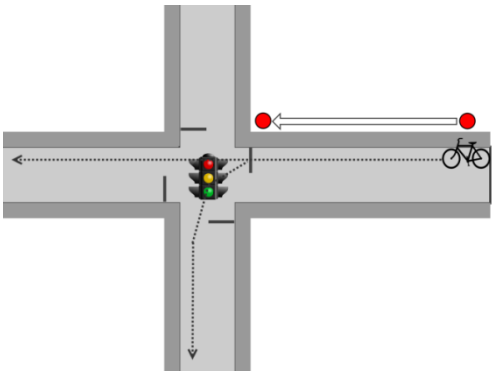
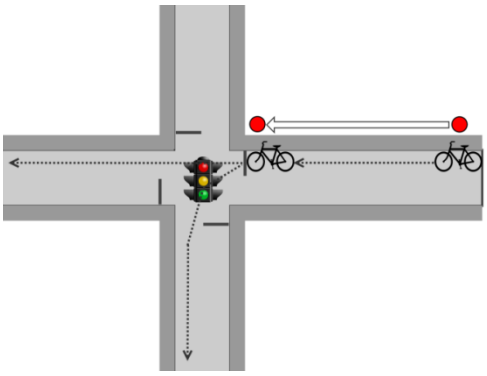
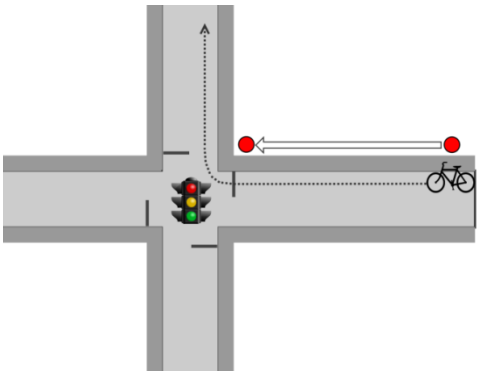
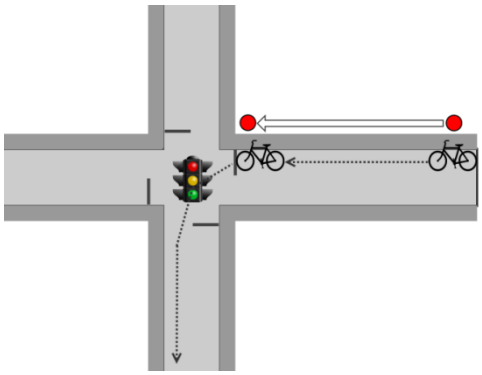
In the following violation types, it appears that the traffic light is already red (i.e., the cyclist does not see the yellow and red onsets or the cyclist sees the red onset when he/she is too far from the intersection) when the cyclist reaches the intersection.	
1) The cyclist goes on the sidewalk to the right side of the street and makes a right turn on the sidewalk to avoid waiting at the traffic light (13.75%). 	2) The cyclist goes to the pedestrian crossing and makes a left turn to the sidewalk. Subsequently, continues on the sidewalk or makes a right turn on the pedestrian crossing (5%). 
3) The cyclist goes through the intersection without stopping when no other road users with a potential conflict are present at the intersection (17.5%). 	4) The cyclist stops first, but after a while, he/she proceeds while the light is still red (6.25%). 
5) The cyclist makes a right turn without stopping (40%). 	6) The cyclist intends to turn left, stops, and waits for the light to change to green. However, the cyclist is not detected by the camera at the intersection and thus the light stays red. Consequently, the cyclist violates the red light as he/she finds a gap (7.5%). 

Table 34 Cyclist violation types part 2

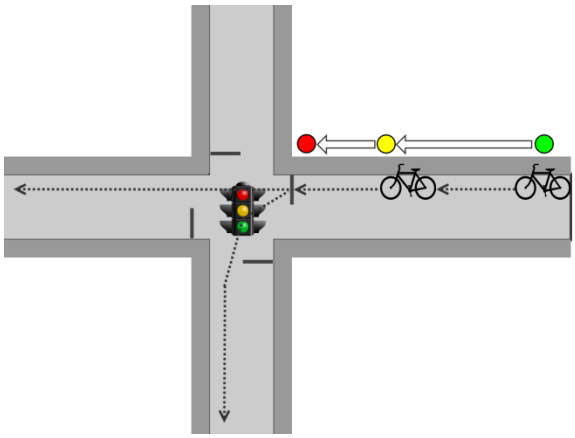
The remaining violation type occurred as a result of being in the dilemma zone as follows.
7) The cyclist sees the yellow onset and decides to proceed. However, the light turns red before reaching the stop bar (10%). 

Table 35 Logistic regression model - signalized intersections

Variables	coefficients	p-value
Morning trip	-0.03435	0.95777
Noon trip	0.61314	0.38447
Evening trip	0.62602	0.33467
Trip on weekend	0.44978	0.47341
Uphill	-0.19133	0.64751
Downhill	0.03481	0.95352
Right turn	3.56206	3.09e-13 ***
Left turn	0.55233	0.23097
Presence of side traffic	-1.70405	6.82e-05 ***
Presence of opposing traffic	-0.49125	0.19340
Presence of front traffic	-1.46949	0.00348 **
Presence of adjacent traffic	-0.53757	0.15174
Weather-warm	-0.20388	0.63081
Weather-cloudy	0.47513	0.26014
Weather-rainy	1.29592	0.05783
Age-younger	0.82284	0.20092
Gender-male	1.12173	0.10146

Stop-controlled intersections

Out of 875 crossings at stop-controlled intersections, almost all cyclists violated the stop sign. However, the number of violations depends on how the violation is defined at the intersection. The traffic rule for cyclists is similar to the motorized vehicles, which is to come to a complete stop before the stop bar[10]. However, only few observations had speed at stop bar of zero. To have more complying observations, the threshold to determine violations was increased from zero to 1.2 meters per second which is similar to the walking speed. As a result, 834 observations were coded as violating behavior and 41 observations as complying behavior. Based on a logistic regression developed, the presence of other road users (i.e., Presence of side traffic) and Age (i.e., younger) were found to be statistically significant

[10] In most US states including Virginia, bicycles have to come to a complete stop at stop signs just like other motorized vehicles

as presented in Table 36. It is less likely to violate a stop sign when there is side traffic or when the cyclist is older.

Table 36 Logistic regression model - sign-controlled intersections

Variables	coefficients	p-value
Morning trip	-0.86396	0.3039
Noon trip	-0.16461	0.7341
Evening trip	0.18996	0.6453
Trip on weekend	-0.28707	0.5417
Uphill	0.76272	0.1002
Downhill	0.37034	0.4589
Right turn	0.67876	0.1592
Left turn	-0.05242	0.9020
Presence of side traffic	-2.12399	4.61e-07 ***
Presence of opposing traffic	0.03304	0.9491
Presence of front traffic	-0.65949	0.1394
Presence of adjacent traffic	-0.91196	0.0913 .
Weather-warm	0.47931	0.2009
Weather-cloudy	0.32370	0.4269
Weather-rainy	0.92193	0.3922
Age-younger	2.28660	0.0328 *
Gender-male	1.89259	0.0861 .

RF method was applied to develop violation prediction models at stop-controlled intersections based on the kinetic information of the cyclists. Two model parameters, namely the number of decision trees and the number of variables (or factors) to use in each tree, needed for model development were determined to be 500 and 5, respectively. Statistical measures were applied to the kinetic data (e.g. speed and acceleration) to create factors for model development. In addition to speed and acceleration variables, time-to-intersection (TTI), distance-to-intersection (DTI), and required deceleration parameter (RDP) were used to create more factors. Table 37 presents the list of all examined factors.

Table 37 list of factors to develop the RF model

No.	Factor	No.	Factor
1	$mean(TTI)$ over the t^{v}_{mon}	14	$max(acceleraiton)$ over the t^{v}_{mon}
2	$range(TTI)$ over the t^{v}_{mon}	15	$min(acceleraiton)$ over the t^{v}_{mon}
3	$std(TTI)$ over the t^{v}_{mon}	16	$mean(DTI)$ over the t^{v}_{mon}
4	$max(TTI)$ over the t^{v}_{mon}	17	$range(DTI)$ over the t^{v}_{mon}
5	$min(TTI)$ over the t^{v}_{mon}	18	$std(DTI)$ over the t^{v}_{mon}
6	$mean(speed)$ over the t^{v}_{mon}	19	$max(DTI)$ over the t^{v}_{mon}
7	$range(speed)$ over the t^{v}_{mon}	20	$min(DTI)$ over the t^{v}_{mon}
8	$std(speed)$ over the t^{v}_{mon}	21	$mean(RDP)$ over the t^{v}_{mon}
9	$max(speed)$ over the t^{v}_{mon}	22	$range(RDP)$ over the t^{v}_{mon}
10	$min(speed)$ over the t^{v}_{mon}	23	$std(RDP)$ over the t^{v}_{mon}
11	$mean(acceleraiton)$ over the t^{v}_{mon}	24	$max(RDP)$ over the t^{v}_{mon}
12	$range(acceleraiton)$ over the t^{v}_{mon}	25	$min(RDP)$ over the t^{v}_{mon}
13	$std(acceleraiton)$ over the t^{v}_{mon}		

As mentioned in the monitoring period section, there is no need to account for the driver/rider response time in the case of stop-controlled intersections. Therefore, only the vehicle/bicycle response (i.e., $t^{e}_{driver/rider}$) with a range of 0.5 to 2.5 seconds was a factor to determine the monitoring period, which basically dictated the end point of the monitoring period. A sensitivity analysis was conducted to understand how the model performance changes with different monitoring periods. For the sake of

comparison the end point of the monitoring period was extended from 2.5 to 4 seconds and for each end point, 4 different start points were considered to examine four different monitoring period lengths of 1, 2, 3, and 4 seconds. Furthermore, since the there was only 41 complying behavior, the data set was synthesized to have a balanced data set using SMOTE[11] oversampling technique [31]. Figure 43 illustrates error rates (i.e., overall error rates on the left and error rates for violations only on the right) of RF models with different monitoring periods[12]. As the end point of the monitoring period is selected closer to the intersection the classification error decreases, which shows cyclists behavior can be predicted with higher accuracies as they become closer to the intersection. However, the prediction accuracies of almost 100% (e.g., when the end point is 0.25 seconds) may not be practical as the time is not sufficient for the endangered rider/driver to respond. In the case of 4-way stop signs where the $t^{e}_{vehicle/bicycle}$ was excluded, the error rates of about 6 percent were obtained when the end point was selected 2 to 2.5 seconds that is sufficient for most individuals to react. For these periods (i.e., with end point of 2-2.5), the most important factors in predicting violations were found to be $std(speed)$ over the t^{v}_{mon}, $range(speed)$ over the t^{v}_{mon}, $min(acceleraiton)$ over the t^{v}_{mon}, $mean(acceleraiton)$ over the t^{v}_{mon}, $std(TTI)$ over the t^{v}_{mon}, $min(TTI)$ over the t^{v}_{mon}, and $range(RDP)$ over the t^{v}_{mon}. However, when different monitoring periods are examined different factors were found as important. In the case of 2-way stop signs, the prediction needs to be conducted sooner because $t^{e}_{driver/rider}$ should be added to $t^{e}_{vehicle/bicycle}$. For example, if the endangered vehicle has an approaching speed of 35 mph, about 2.6 seconds is required for the vehicle to stop. Adding 2 seconds for the $t^{e}_{vehicle/bicycle}$ would result in 4.6 seconds away from the intersection. Thus, as shown in Figure 43, the overall accuracy is about 8 percent.

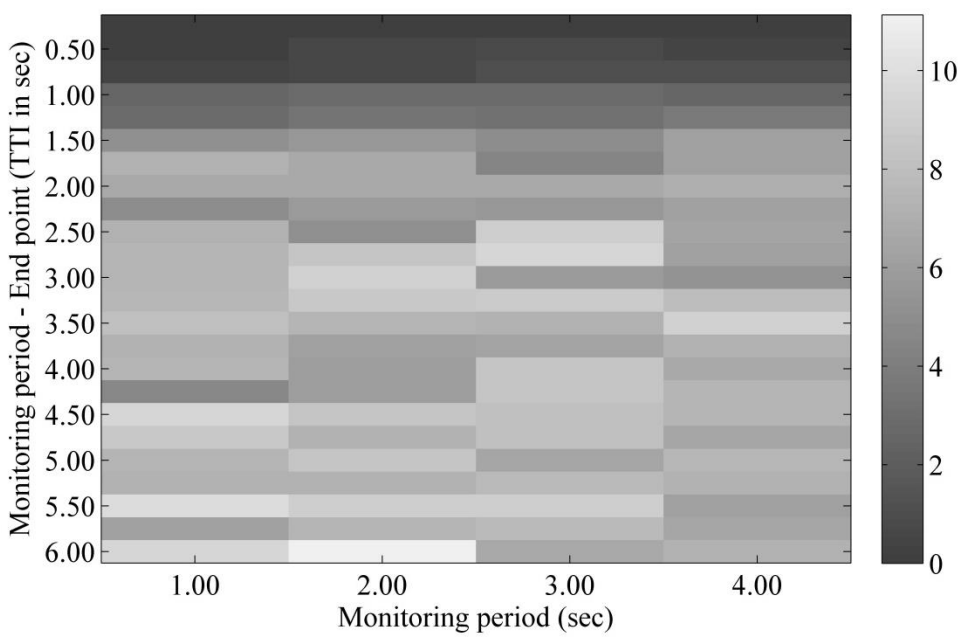


Figure 43 RF models – overall prediction accuracies

Conclusion

Investigating cyclist violations at intersections using naturalistic cycling data was the focus of this paper. Different factors that affect cyclist violations at both signalized and stop-controlled intersections were examined. Logistic regression was carried out to identify statistically significant factors; it was found that it is more likely that a cyclist violates a red light when making right turns at signalized

[11] Synthetic Minority Over-sampling Technique

[12] See appendix D for a discussion on false positives and false negative rates

intersections. Also, the probability of red light violation decreases when there is side traffic at the intersection or when there is traffic in front of the cyclist. In case of stop-controlled intersections, the likelihood of violating a stop sign (i.e., crossing the stop bar at over 1.2 meters per second) increases when there is no side traffic or when the cyclist is younger. Violation prediction models were developed for stop-controlled intersections using RF method and based on kinetic information. The kinetic data such as speed and acceleration were obtained through instrumented bicycle as part of the naturalistic cycling experiment. Different monitoring periods to extract kinetic data were considered. Different period length (i.e., from 1 to 4 seconds) were tested. However, it appeared that the monitoring period length did not change model performance. Another factor to define a monitoring period was the end point of the period which is basically the time at which the prediction task is completed. The closer the end point was to the intersection the higher the prediction accuracy was achieved. However, the trade-off was that higher accuracies are associated with less time for endangered users to react. The error rates of about 6 percent were obtained when the end point was selected 2 to 2.5 seconds that is sufficient for most individuals to react.

Acknowledgements

This research effort was funded by the Connected Vehicle Initiative UTC (CVI-UTC).

References

[1] NHTSA, *'Bicyclists and Other Cyclists' Traffic Safety Facts*. 2015, National Center for Statistics and Analysis, US Department of Transportation, Washington, DC.

[2] FARS. [cited 2015; Available from: http://www-fars.nhtsa.dot.gov/People/PeoplePedalcyclists.aspx.

[3] The-University-of-North-Carolina-Highway-Safety-Research-Center. *North Carolina Bicycle Crash Types 2005 - 2009*. 2011; Available from: http://www.ncdot.gov/bikeped/download/summary_bike_types05-09.pdf.

[4] Tan, C., *Crash-type manual for bicyclists.* Publication No. FHWA-RD-96-104, 1996.

[5] McLeod, K., D. Flusche, and A. Clarke, *Where We Ride: Analysis of Bicycling in American Cities.* 2013.

[6] Wu, C., L. Yao, and K. Zhang, *The red-light running behavior of electric bike riders and cyclists at urban intersections in China: an observational study.* Accident Analysis & Prevention, 2012. **49**: p. 186-192.

[7] Johnson, M., et al., *Why do cyclists infringe at red lights? An investigation of Australian cyclists' reasons for red light infringement.* Accident Analysis & Prevention, 2013. **50**: p. 840-847.

[8] Pai, C.-W. and R.-C. Jou, *Cyclists' red-light running behaviours: An examination of risk-taking, opportunistic, and law-obeying behaviours.* Accident Analysis & Prevention, 2014. **62**: p. 191-198.

[9] Martínez-Ruiz, V., et al., *Risk factors for causing road crashes involving cyclists: An application of a quasi-induced exposure method.* Accident Analysis & Prevention, 2013. **51**: p. 228-237.

[10] Johnson, M., J. Charlton, and J. Oxley. *Cyclists and red lights—a study of the behaviour of commuter cyclist in Melbourne*. in *Australasian Road Safety Research, Policing and Education Conference, Adelaide*. 2008.

[11] Johnson, M., et al., *Riding through red lights: The rate, characteristics and risk factors of non-compliant urban commuter cyclists.* Accident Analysis & Prevention, 2011. **43**(1): p. 323-328.

[12] Schepers, J., et al., *Road factors and bicycle–motor vehicle crashes at unsignalized priority intersections.* Accident Analysis & Prevention, 2011. **43**(3): p. 853-861.

[13] Arash, J., R. Hesham, and D. Thomas, *Developing a System Architecture for Cyclist Violation Prediction Models Incorporating Naturalistic Cycling Data*, in *6th International Conference on*

Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE. 2015: Las Vagas.

[14] Phillips, R.O., et al., *Reduction in car–bicycle conflict at a road–cycle path intersection: Evidence of road user adaptation?* Transportation research part F: traffic psychology and behaviour, 2011. **14**(2): p. 87-95.

[15] Zhang, Y. and C. Wu, *The effects of sunshields on red light running behavior of cyclists and electric bike riders.* Accident Analysis & Prevention, 2013. **52**: p. 210-218.

[16] Räsänen, M., I. Koivisto, and H. Summala, *Car driver and bicyclist behavior at bicycle crossings under different priority regulations.* Journal of Safety Research, 1999. **30**(1): p. 67-77.

[17] Dingus, T.A., et al., *The 100-car naturalistic driving study, Phase II-results of the 100-car field experiment*. 2006.

[18] Sayer, J., et al., *Integrated vehicle-based safety systems field operational test final program report.* 2011.

[19] Benmimoun, M., et al. *Safety Analysis Method for Assessing the Impacts of Advanced Driver Assistance Systems within the European Large Scale Field Test euroFOT*. in *8th ITS European Congress, Lyon, France*. 2011.

[20] Dozza, M., J. Werneke, and A. Fernandez. *Piloting the naturalistic methodology on bicycles*. in *Proceeding ot the st International Cycling Safety Conference, Helmond NL, Nov 7-8 2012.* 2012.

[21] Dozza, M. and A. Fernandez, *Understanding Bicycle Dynamics and Cyclist Behavior From Naturalistic Field Data (November 2012).* 2014.

[22] Johnson, M., et al. *Naturalistic cycling study: identifying risk factors for on-road commuter cyclists*. in *Annals of Advances in Automotive Medicine/Annual Scientific Conference*. 2010. Association for the Advancement of Automotive Medicine.

[23] Gustafsson, L. and J. Archer, *A naturalistic study of commuter cyclists in the greater Stockholm area.* Accident Analysis & Prevention, 2013. **58**: p. 286-298.

[24] Lacherez, P., et al., *Visibility-related characteristics of crashes involving bicyclists and motor vehicles–Responses from an online questionnaire study.* Transportation research part F: traffic psychology and behaviour, 2013. **20**: p. 52-58.

[25] Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.

[26] Hastie, T., et al., *The elements of statistical learning*. Vol. 2. 2009: Springer.

[27] Jahangiri, A., H. Rakha, and T.A. Dingus, *Predicting Red-light Running Violations at Signalized Intersections Using Machine Learning Techniques*. 2015.

[28] McLaughlin, S.B., J.M. Hankey, and T.A. Dingus, *A method for evaluating collision avoidance systems using naturalistic driving data.* Accident Analysis & Prevention, 2008. **40**(1): p. 8-16.

[29] Landis, B., et al., *Characteristics of emerging road and trail users and their safety.* Transportation Research Record: Journal of the Transportation Research Board, 2004(1878): p. 131-139.

[30] Wilson, D.G. and J. Papadopoulos, *Bicycling science*. 2004: Mit Press.

[31] Chawla, N.V., et al., *SMOTE: synthetic minority over-sampling technique.* Journal of artificial intelligence research, 2002: p. 321-357.

Chapter 7: Conclusions and Future Recommendations

Ch. 7 – Conclusions and Future recommendations

This chapter summarizes the dissertation, highlights the conclusions of different chapters, and provides future recommendations. The main goal of this dissertation was to improve safety at intersections by predicting individuals' violations. When users (e.g., drivers, cyclists, pedestrians, etc.) are approaching an intersection, the first question is: what is their transportation mode? This makes the first building block in the dissertation. The reason for this to be the first block is that the violations by different transportation modes are different due to disparities between modes in terms of acceleration/deceleration capabilities and perceived level of safety while using certain transportation modes. For this part, five different transportation modes were considered (i.e., passenger car, bus, bicycle, walking, running). Subsequently, the second question is: given the transportation mode, how to predict whether the user is going to violate traffic rules at the intersection? To help prevent/mitigate crashes caused by violations, violations need to be identified before they occur, so both the road users (i.e., drivers, pedestrians, etc.) in potential danger and the infrastructure can be notified and actions can be taken accordingly. For this part, two different transportation modes (e.g. passenger cars and bicycles) were studied. Hence, the model development was divided into three main tasks: (1) Transportation mode recognition, (2) Driver violation prediction, and (3) Cyclist violation prediction.

Different data collection methods can be adopted to undertake each task. For transportation mode recognition task, smartphones, on-board equipment, and video cameras can be used to obtain required information for model development. For violation prediction task, smartphones, on-board equipment, video cameras, and simulator can be used. In any case, user information such as speed and acceleration can be obtained and used to develop recognition and prediction models. Smartphones are now capable of sending and receiving data through various ways (e.g. Wi-Fi/cellular network/Bluetooth), providing alerts, and storing/processing data. Furthermore, DSRC-enabled smartphones will soon be in the market. Thus, to recognize the value of smartphones, they were used to collect data for the transportation mode recognition task. For the driver violation prediction task, observational data (e.g. through video cameras) and simulator data were adopted. Observational data has the advantage of capturing natural driver behavior as the drivers are not aware that their data are being collected. Simulator data may not reflect natural driver behavior, but it enabled us to examine additional factors such as age and gender that were unavailable through the observational data. In the cyclist violation prediction task, a naturalistic cycling experiment was designed through instrumenting bicycles by on-board equipment (i.e., mini-DAS) to capture cyclist behavior in real world condition as they cross intersections. It should be noted that although each data collection method has its own advantages (e.g., examining risky scenarios by simulators, examining locations with high violation rate through video cameras), smartphones and on-board equipment are desirable for implementation testing because in these ways warnings can be sent to the users in potential danger. Smartphones have an additional advantages over on-board equipment: most individual carry one in their pocket no matter what transportation mode they are using. This is especially important for bicycle and walk modes for which instrumentation might be difficult or not applicable. Also, when using on-board units, the GPS requires a warm-up time that leads to not having valid GPS data for the start of the trips.

Transportation Mode Recognition

An android application was developed to collect sensor data from smartphones. The sensors include GPS, accelerometer, gyroscope, and rotation vector sensors. Using the acquired data, detection models were developed to recognize the mode of transportation. The transportation

mode recognition task conducted in this dissertation is different from previous work in that it comprehensively accounts for the following factors. This task: concentrated on both motorized (car, bike, and bus) and non-motorized modes (walk and run); did not require the travelers to maintain the phone's position at a specific location (such as in pocket or attached to their body) during data collection; made use of the information from gyroscope and rotation vector sensors, which has not been used in previous studies for the purpose of transportation mode detection; required the travelers to collect the car and bus data on different road types with different speed limits (e.g. 15, 25, 35, 45, and 65 mph); required the travelers to collect the bus, car, and bike data on some path where they had to stop at different intersections and thus data similar to traffic jam condition were included as well; applied all common machine learning procedures, namely, complete model selection, regularization, feature selection, and feature scaling; created a large number of features of which the most representative ones were selected for model development; created the features based on statistical measures of dispersion as well as derivatives to obtain variations over the time window of interest and consequently incorporated this knowledge (i.e. features' time dependency) into the models. Here the conclusions and future recommendations are summarized:

Conclusions

- Data obtained from smartphone sensors were found to have invaluable information for the purpose of transportation mode recognition.
- Considering misclassification rates, the car and bus modes (i.e., motorized modes) were the most difficult ones to distinguish. Even using more complex models such as SVM and RF, the car mode was misclassified as the bus mode in about 4-6% of the time.
- The Random Forest method was found to produce the best overall performance. However, for specific modes (i.e. walk and Run), the SVM outperformed the RF method.
- A large number of features were examined, of which top twenty were identified as the most useful features for the transportation mode recognition model.
- Error analysis was conducted to gain possible insights about where different models fail to correctly classify the data. The only discovered pattern found was that including speed data into the model improved the overall classification accuracy from 95.1% to 96.3%.

Future recommendations

- Including more data in model development would result in making better generalizations.
- Applying approaches to examine the data as a sequence could be helpful in correcting misclassifications.
- Including other transportation modes (e.g. Motorcycles) would make the problem more difficult but would lead to a more comprehensive understanding of different transportation modes.

- Since speed was found to be a factor to improve detection model performance, estimating speed data from sensors such as accelerometer and including that information in the model would improve the classification accuracy. Therefore, even without having the GPS data, the models can be improved.
- Implementation testing: To better understand the model effectiveness, the recognition models should be tested to see their performance and other impacts such as battery usage.

Driver Violation Prediction

The driver violation prediction task focuses on developing prediction models aiming at identifying RLR violations using two different data sets (i.e., observational data set and driving simulator data set). The observational data came from the Cooperative Intersection Collision Avoidance Systems for Violations (CICAS-V) project. For this project, data acquisition systems (DASs) at different intersections, provided video, radar, and traffic signal data. The driving simulator data came from a study conducted by University of Iowa.

Some previous studies concentrated on investigating the characteristics of red-light violators and conditions in which drivers are more or less likely to violate. Several studies developed models to estimate the frequency of red-light violators. Statistical and probabilistic approaches have also been applied. However, studies that apply AI techniques to develop RLR violation prediction models are limited. Hence, the driver violation prediction task in this dissertation: created several factors in model development and used a selection method to determine the most useful factors; conducted a sensitivity analysis to determine an appropriate monitoring period corresponding to the onset of a yellow light to capture the information that reflects drivers' decisions; investigated how the RF method can predict RLR violations using different monitoring periods while providing enough time for endangered drivers and/or the infrastructure to respond; used observational data and driving simulator data to develop prediction models; identified important factors in predicting RLR violations. Here the conclusions and future recommendations are summarized:

Conclusions

- Prediction models resulted in very high prediction accuracies in the case of the observational data set (i.e., error rates of 0.1% to 1.6%) and poor accuracies in the case of the simulator data set (i.e., error rates of 5.9% to 17.9%).
- The prediction models from the observational data achieved the lowest error rates when using more than five factors and the monitoring period of 2 to 6 meters immediately after encountering the yellow light.
- Both types of factors were found to be important in predicting RLR violations: (1) Factors that describe quantities at a point in time (e.g., TTI at yellow onset) and (2) Factors that describe quantities across a time period that occurs during the defined monitoring period (e.g., $max(Velocity)$ over t_{mon}^{v}).
- Individual factor importance was obtained, and it was shown that the factor importance may change depending on the monitoring period being assessed.

- TTI at the onset of a yellow light, the mean(TTI) over t^{v}_{mon}, and DTI at the onset of a yellow light were identified as the most important factors (i.e. top three) of the observational data. Additional important factors in the observational data that described statistical quantities across the monitoring period were $mean(TTI)$ over t^{v}_{mon}, $mean(DTI)$ over t^{v}_{mon}, $range(Velocity)$ over t^{v}_{mon}, and $std(Acceleraiton)$ over t^{v}_{mon}. This shows that, in addition to the yellow onset, an appropriate monitoring period can provide useful information reflecting driver behavior (i.e., decision to stop or proceed) when approaching a signalized intersection.

- Simulator data had the advantage of accounting for driver factors (i.e., age and gender) and specific hypothetical scenarios, as indicated by the treatment and the secondary task condition factors. However, these factors were found to be among the least important in predicting RLR violations.

- TTI, RDP, and DTI at the onset of a yellow light were found to be the most important factors from the simulator data set.

- Comparing the models developed using observational and simulator data, TTI, DTI, RDP, and velocity at the onset of a yellow light were among the most important factors identified by models constructed using both data sets.

- Connected-vehicle technology uses dedicated short-range communications (DSRC) that provide a reliable and fast communication (latency of less than 100 ms) and a range of less than 1,000 meters. Once the prediction model was developed, the prediction time for a single observation was extremely fast (between 1 to 10 ms on an i5-3230M CPU at 2.6 Hz and RAM of 8 GB). Hence, the processing time and communication latency would account for a very short period of time. Consequently, in situations where the endangered driver has sufficient time, a warning can be issued to the driver to respond. In cases where insufficient time is available, the infrastructure can take appropriate actions (e.g., providing a longer red clearance interval).

Future recommendations

- Developing violation prediction models for different locations for making better generalizations.

- Developing violation prediction models for other transportation modes such as motorcycles and trucks should be considered to have prediction models for all modes of transport.

- Testing and implementing a crash violation prediction system that incorporated driver violation prediction models.

- Developing and testing warning systems including the human machine interface in real world.

Cyclist Violation Prediction

A naturalistic cycling experiment was designed to investigate violation behavior of cyclists at intersections. For several different reasons (e.g. bad judgment, distraction, etc.), cyclists also fail to obey traffic rules at intersections. Hence the problem is how to prevent/mitigate these intersection-related crashes that involves bicycles. The failure to comply need to be identified before they occur so actions can be taken to alleviate the consequences. A limited number of studies focused on introducing countermeasures to reduce bicycle related crashes at intersections. Many studies focused on assessing factors and conditions that influence crossing behavior of cyclists at intersections. However, it appears that the past naturalistic cycling studies did not particularly target the cyclist violation behavior at intersections. The cyclist violation prediction task in this dissertation: conducted naturalistic cycling data collection; identified eligible participants (i.e. those who ran through many intersections) through pre-screening ; applied Hawkeye software for data reduction to extract required variables for each crossing behavior; adopted multivariate logistic regression to identify significant factors that influence cyclist behavior when crossing intersections; employed random forest method to predict whether or not a cyclist is going to violate a stop sign; and identified important factors in predicting the violations.

In the naturalistic cycling experiment, participants were provided with instrumented bicycles capable of collecting video data (using two cameras: a forward view camera and a rider view camera) as well as sensor data (e.g. GPS, acceleration, gyroscope). Participants were asked to ride these bicycles whenever, wherever, and however they normally would. Data collection is still running at the time of writing this document. Therefore, data from 8 participants (i.e., 4 participants with 4 weeks and 4 participants with 2 weeks of data) were included in analyses. The entire data from these participants included 718 and 875 observations at signalized and stop-controlled intersections, respectively. Here the conclusions and future recommendations are summarized:

Conclusions

- A system architecture that incorporates naturalistic cycling data to develop cyclist violation prediction models at intersections was presented and it was shown how connected vehicle technology can be adopted for different system entities to communicate amongst themselves.
- Logistic regression analysis showed that it is more likely that a cyclist violates a red light when making right turns at signalized intersections. Also, the probability of red light violation decreases when there is side traffic at the intersection or when there is traffic in front of the cyclist.
- The likelihood of violating a stop sign (i.e., crossing the stop bar at over 1.2 meters per second) increases when there is no side traffic or when the cyclist is younger.
- Violation prediction models were developed for stop-controlled intersections using RF method and based on kinetic information. The prediction models contributed to error rates of 0 to 10 percent depending on how far from the intersection the prediction task was conducted.
- Different monitoring periods were examined as the cyclists approach the intersections. The monitoring period length was found to be insignificant. Also, the

closer the end point of monitoring period was to the intersection the higher the prediction accuracy was achieved. However, the trade-off was that higher accuracies are associated with less time for endangered users to react. The error rates of about 6 percent were obtained when the end point was selected 2 to 2.5 seconds that is sufficient for most individuals to react.

- For monitoring periods of about 2-2.5 seconds, the most important factors in predicting violations were found to be $std(speed)$ over the t_{mon}^{v}, $range(speed)$ over the t_{mon}^{v}, $min(acceleraiton)$ over the t_{mon}^{v}, $mean(acceleraiton)$ over the t_{mon}^{v}, $std(TTI)$ over the t_{mon}^{v}, $min(TTI)$ over the t_{mon}^{v}, and $range(RDP)$ over the t_{mon}^{v}. However, when different monitoring periods are examined different factors were found as important.

Future recommendations

- Including more data from additional human subjects and from different locations (i.e., including congested areas seen in bigger cities than in Blacksburg) for making better generalizations. Also, identifying intersections with high violation rates and including human subjects who cross these intersections as they commute.

- Speed sensors that were added to the instrumented bicycles significantly improved data quality comparing to the GPS and accelerometer sensors. Therefore, it is recommended to use speed sensors for similar data collection studies. The speed sensor used was basically based on distance data as the bicycle wheels roll. speed and acceleration data can be derived based on the distance data.

- Improving the performance of the prediction models by reducing false negatives (i.e., actual violations predicted as compliances) while maintaining false positive (i.e., actual compliances predicted as violations) rate below 5% as recommended by the auto industries.

- Violation prediction models for different modes of transport should be constantly monitoring individuals approaching intersections. When a potential threat is predicted, different actions can be taken depending on the situation; when the endangered rider has sufficient time, a warning can be issued and sent from roadside equipment (RSE) to the rider to respond. In cases where not enough time is available, the infrastructure can take appropriate actions by changing the signal control through the traffic light (e.g. providing longer all red clearance). More research is required to develop warning systems and implement such violation prediction models in real world.

Appendix A: Transportation mode recognition - Error analysis

Random Forest model that led to the best overall performance was selected to conduct the error analysis. The confusion matrix resulted from the RF model is shown in Table A.1. As presented in this matrix, different types of errors were identified. Table A.2 lists all types of errors in the order of importance (i.e., highest errors on top of the list).

Table A.1 Confusion matrix – Random forest model

Random Forest		Actual					
		Bike	Car	Walk	Run	Bus	**Precision**
Predicted	Bike	95.47	1.46	2.63	0.97	2.29	93.06
	Car	0.37	93.84	0.12	0.05	4.47	94.93
	Walk	2.93	0.13	96.23	1.59	0.12	95.24
	Run	0.03	0.00	0.40	96.81	0.00	99.55
	Bus	1.19	4.57	0.63	0.58	93.12	93.02
	Recall	95.47	93.84	96.23	96.81	93.12	

Table A.2 Misclassification errors

No.	Description	Number of Errors	Recall (%)
1	Car misclassified as Bus	281	4.57
2	Bus misclassified as Car	264	4.47
3	Bike misclassified as Walk	183	2.93
4	Walk misclassified as Bike	164	2.63
5	Bus misclassified as Bike	133	2.29
6	Run misclassified as Walk	94	1.59
7	Car misclassified as Bike	88	1.46
8	Bike misclassified as Bus	74	1.19
9	Run misclassified as Bike	56	0.97
10	Walk misclassified as Bus	39	0.63
11	Run misclassified as Bus	34	0.58
12	Walk misclassified as Run	27	0.40
13	Bike misclassified as Car	23	0.37
14	Car misclassified as Walk	7	0.13
15	Walk misclassified as Car	6	0.12
16	Bus misclassified as Walk	6	0.12
17	Run misclassified as Car	3	0.05
18	Bike misclassified as Run	2	0.03
19	Car misclassified as Run	0	0
20	Bus misclassified as Run	0	0

Since a single observation includes many variables, error analysis becomes a difficult task by looking at individual variables. Here the top seven error types as highlighted in Table A.2 were reviewed; Focusing on different variables, no obvious pattern in general can be distinguished as shown in the following figures. However, for particular variables, it appears that the errors occurred with the value of that variable close to zero as shown in a few cases such as figures A.6(c), A.4(d), A.4(i), and A.5(d). Although being close to zero is a pattern, including this information cannot be helpful in model improvement. There are many observations with similar values (i.e. close to zero) and the model was able to correctly classify them as shown in the same figures.

Considering the mean of speed variable, it appears that the errors could have been avoided to some degree if this variable had been used. For example, there are many bike observations with mean speed of over 2 m/s that were misclassified as the walk mode as shown in figure A.3(j). Therefore, knowing that an observation with a mean speed of over 2 m/s cannot belong to a walk mode can improve the classification algorithm. Similar examples include figures A.1(j), A.5(j), A.6(j), and A.7(j).

Appendix A

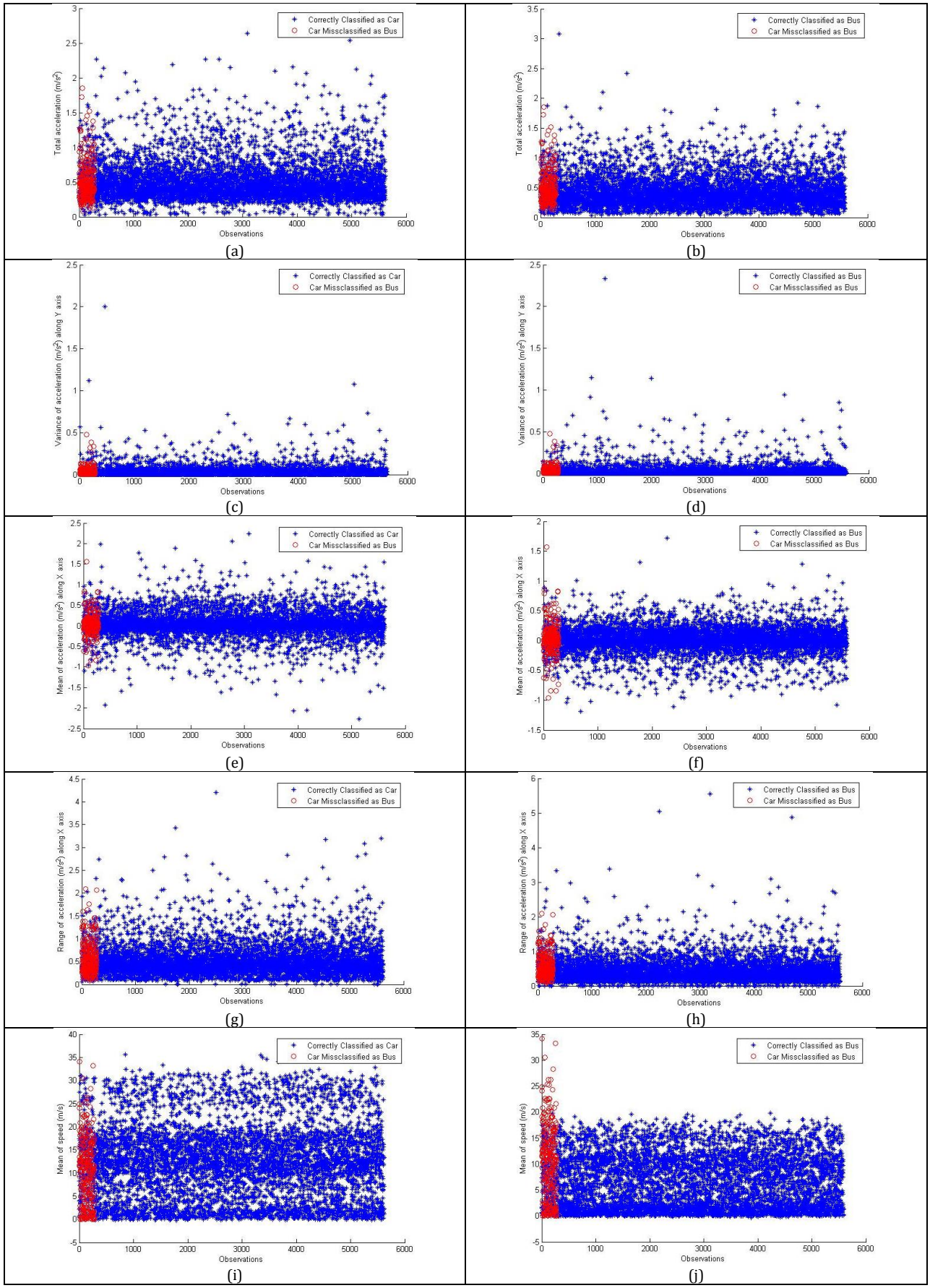


Figure A.1 Car misclassified as Bus (type 1)

Appendix A

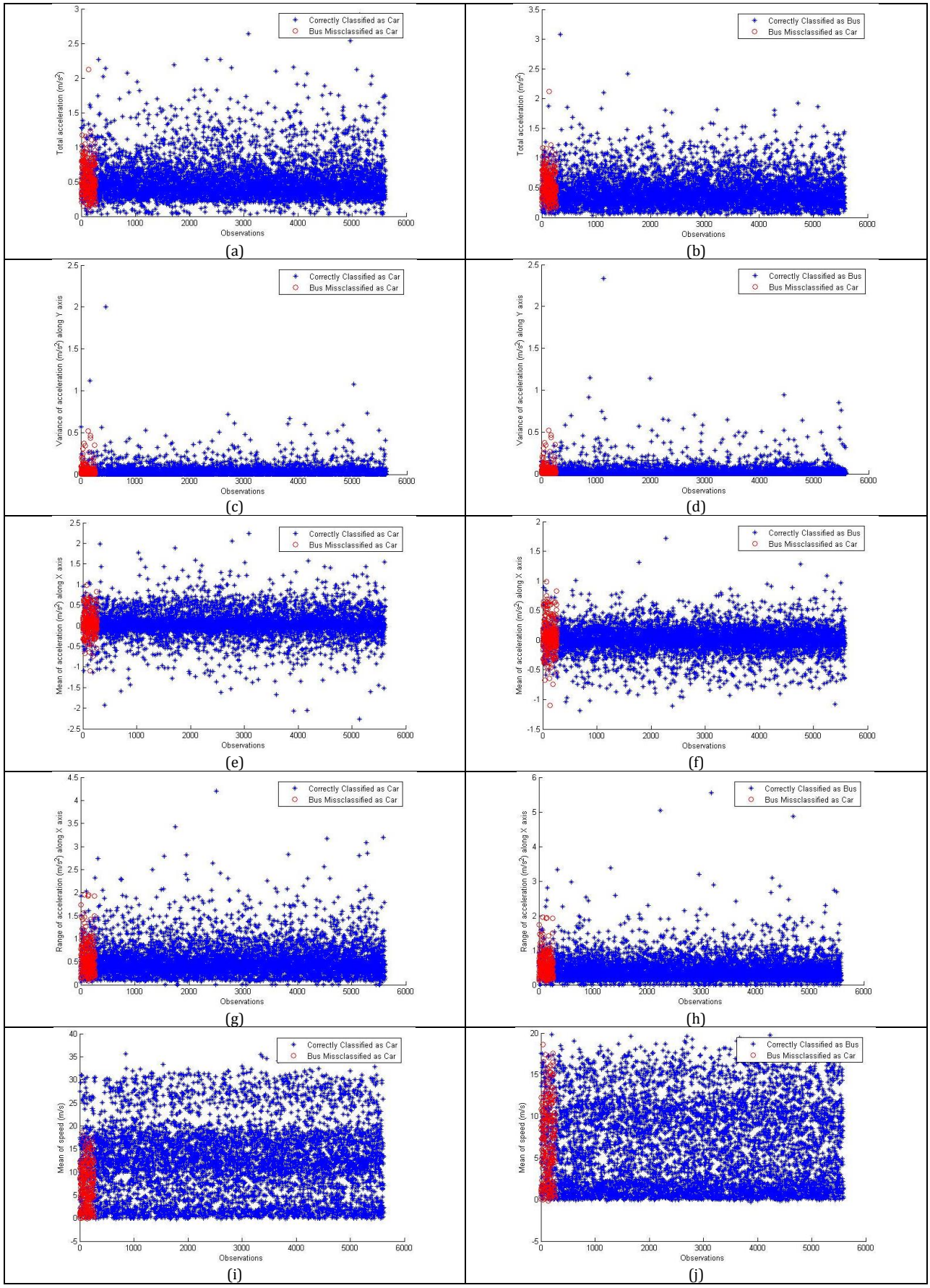


Figure A.2 Bus misclassified as Car (type 2)

Appendix A

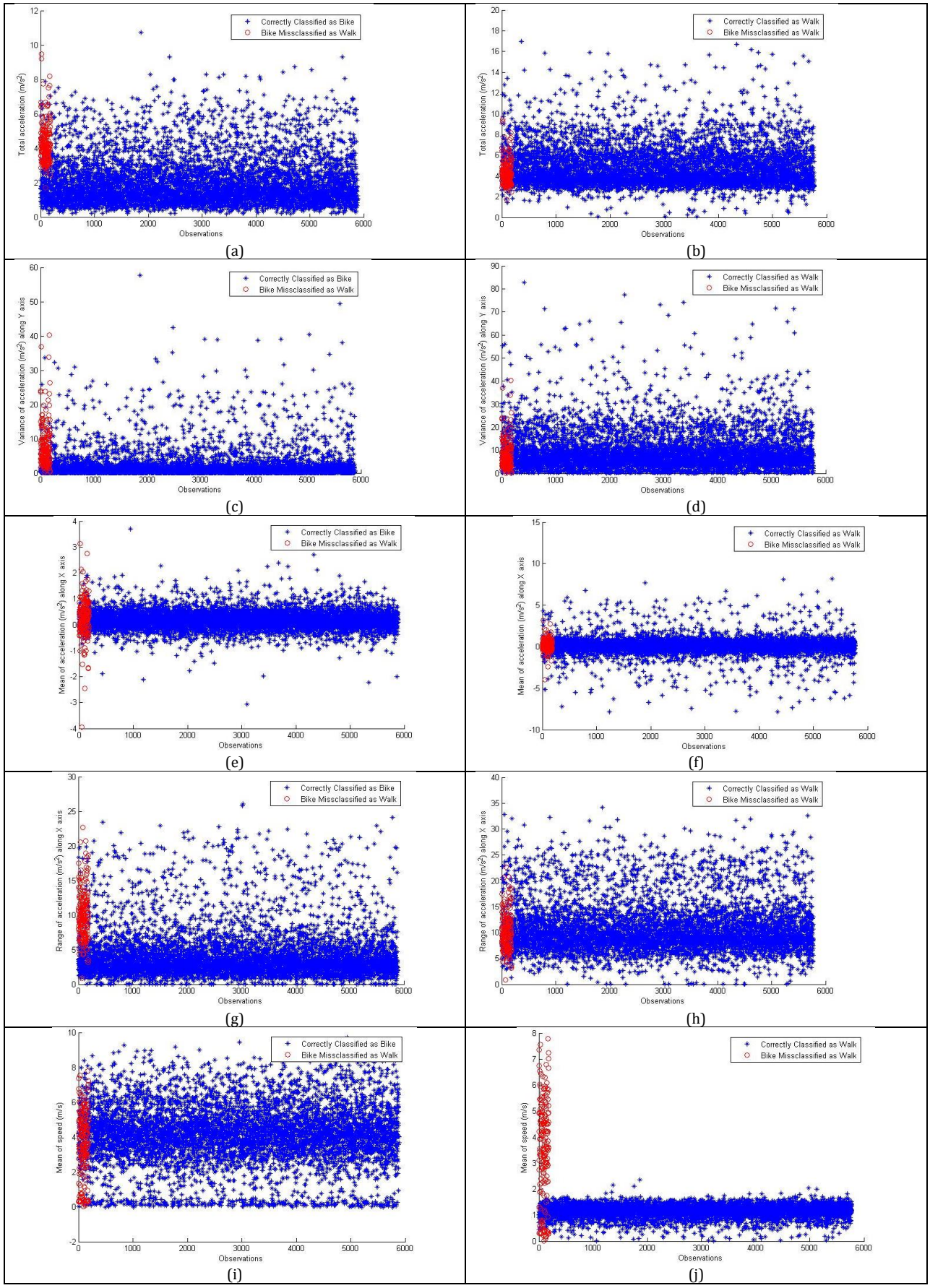


Figure A.3 Bike misclassified as Walk (type 3)

Appendix A

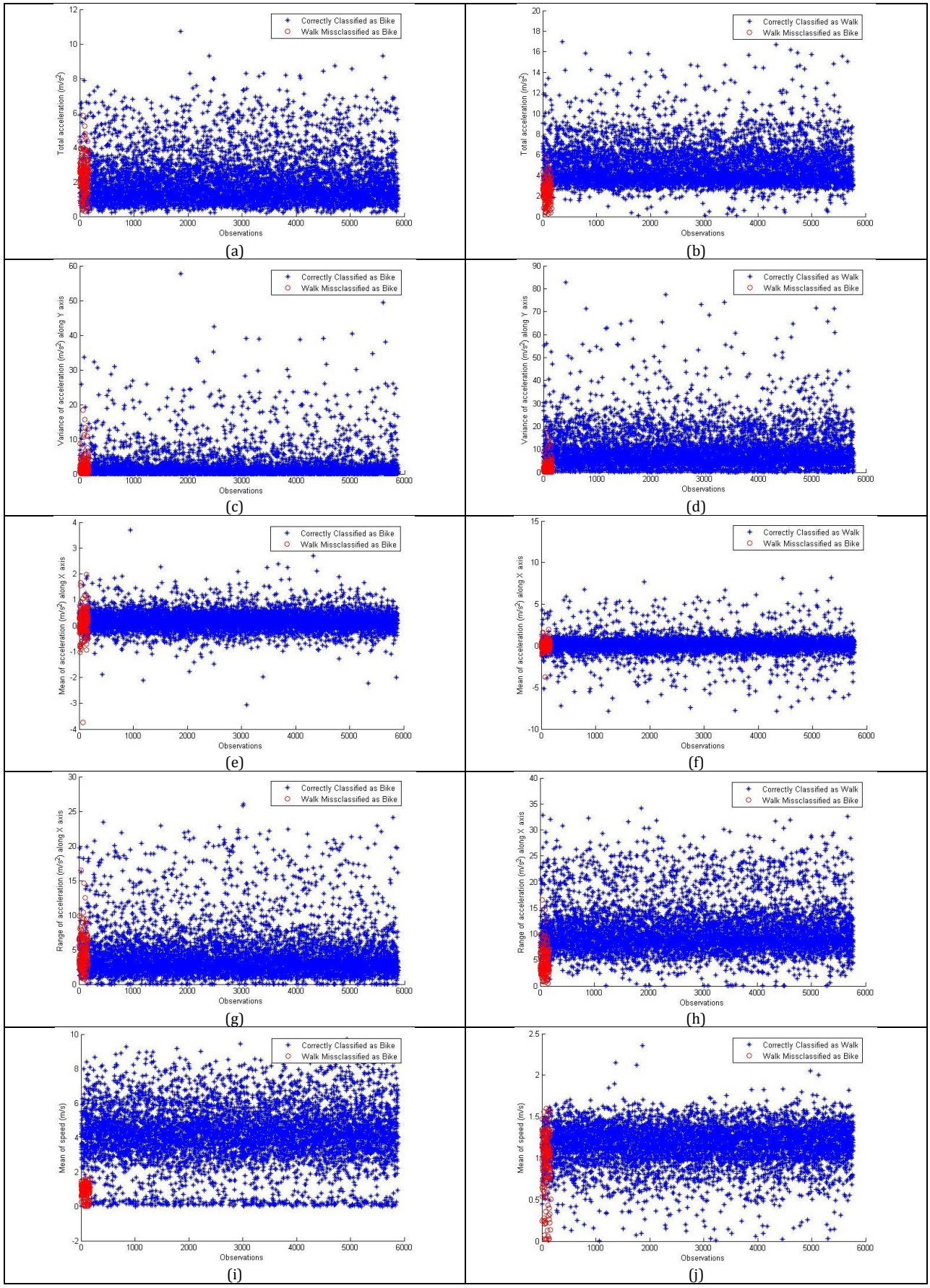


Figure A.4 Walk misclassified as Bike (type 4)

Appendix A

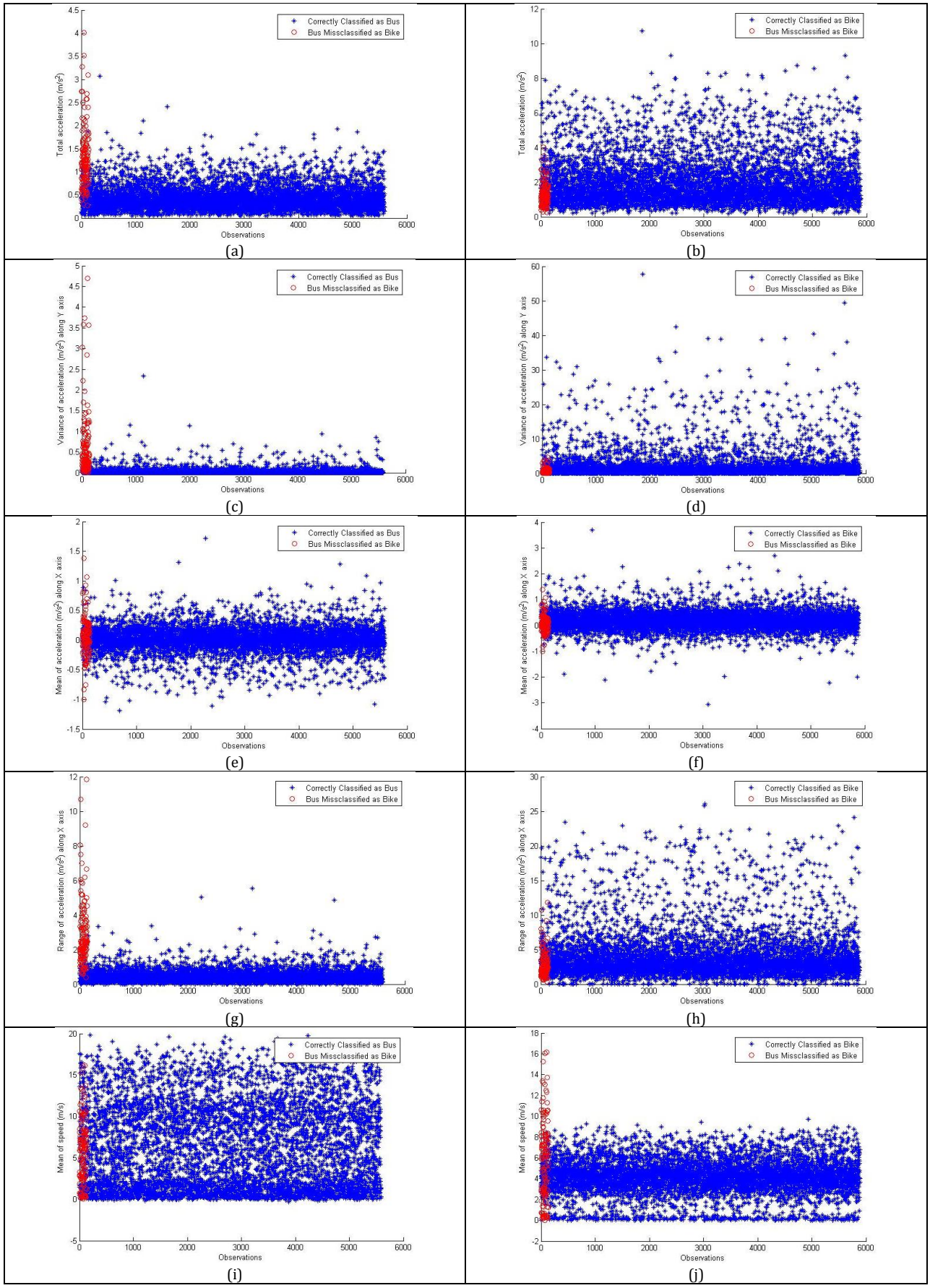


Figure A.5 Bus misclassified as Bike (type 5)

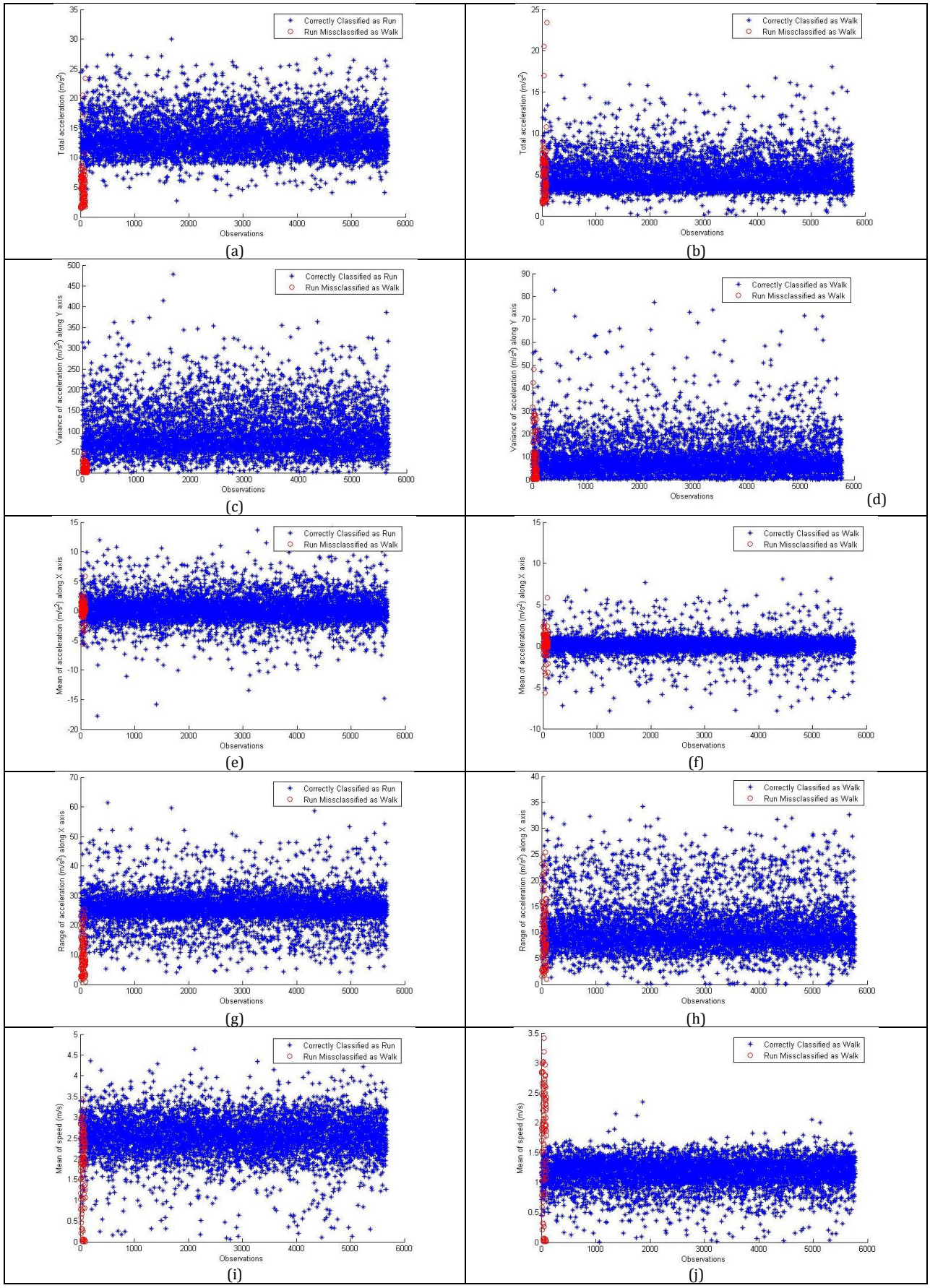


Figure A.6 Run misclassified as Walk (type 6)

Appendix A

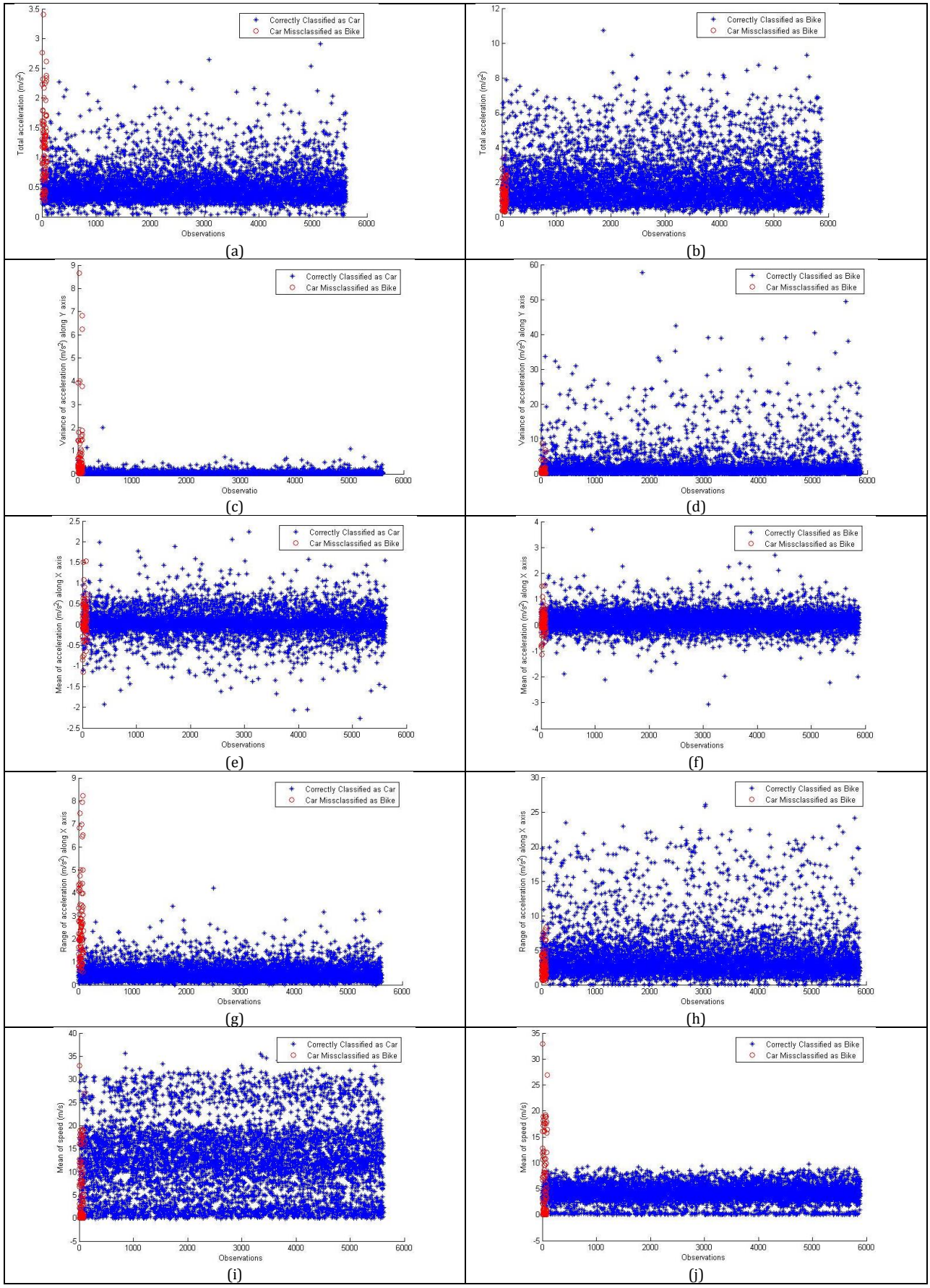


Figure A.7 Car misclassified as Bike (type 7)

Appendix B: Transportation mode recognition - Analysis extension

This section presents further analysis on the transportation mode recognition task. The second paper of Chapter 4 (i.e., Ieee journal paper) showed recognition models developed to distinguish five transportation modes (i.e., Car, Bus, Bicycle, Walk, Run) using smartphone sensor data without using GPS. Two additional cases were considered and assessed as follows. In both cases, a mode detection model was developed using random forest method that was found in the paper as the best model based on overall accuracy.

Case 1) *When GPS information is available in addition to other sensor data*

Although the goal was to exclude the GPS data in model development due to the disadvantages of GPS sensor, it may occur in many situations in which the GPS data are already available (e.g., when navigating) and thus it can lead to improving the detection models. Therefore, confusion matrices were obtained with and without using the GPS to understand the impacts of including the GPS information as presented in Table B.1 and Table B.2. The overall model accuracy improved from 95.1% to 96.3%. However, including the GPS data did not necessarily improved the detection accuracy for all modes. For example, Bus was misclassified as car in 4.47% of the time without using GPS. However, including GPS data led to almost the misclassification rate of 4.61%. Even the misclassification of car as bus increased from 4.57% to 5.87%. The reason for this is the fact that the speed data for car and bus mode can be very similar as these modes are the motorized modes and thus produce similar speed data. Consequently, including speed information may even confuse the detection models.

Table B.1 without using GPS - 95.1% overall accuracy

Random Forest		Actual					
		Bike	Car	Walk	Run	Bus	**Precision**
Predicted	Bike	95.47	1.46	2.63	0.97	2.29	93.06
	Car	0.37	93.84	0.12	0.05	4.47	94.93
	Walk	2.93	0.13	96.23	1.59	0.12	95.24
	Run	0.03	0.00	0.40	96.81	0.00	99.55
	Bus	1.19	4.57	0.63	0.58	93.12	93.02
	Recall	95.47	93.84	96.23	96.81	93.12	

Table B.2 with using GPS - 96.3% overall accuracy

Random Forest		Actual					
		Bike	Car	Walk	Run	Bus	**Precision**
Predicted	Bike	98.46	0.80	1.00	1.10	1.49	95.86
	Car	0.10	93.37	0.05	0.00	4.61	95.10
	Walk	0.70	0.05	98.20	0.60	0.03	98.58
	Run	0.05	0.03	0.37	97.96	0.05	99.48
	Bus	0.70	5.87	0.37	0.34	93.82	92.88
	Recall	98.46	93.37	98.20	97.96	93.82	

Looking at table 40, the high accuracy of 96.3% was obtained. However, it is not possible to see how the misclassifications affect the violation prediction. For example, the bike mode is misclassified as the bus mode in 0.7% of the time. For this 0.7%, if the user violates at the intersection, the violation prediction task may lead to true positives (i.e., correct classifications) or false negatives. Similarly,

for this 0.7%, if the user complies at the intersection, the violation prediction task may lead to true negatives (i.e., correct classifications) or false positives. Therefore, it is recommended to test both models (i.e., transportation mode recognition model and violation prediction model) on the same dataset to see the impacts. In this dissertation, however, the available data sources were not the same and thus the misclassification impacts on the performance of violation prediction models are unknown.

Case 2) *when the car and bus modes are combined as a single mode*

Depending on the problem of interest, it may not be necessary to differentiate the car mode from the bus mode. In these cases, car and bus modes can be combined an assessed as a single mode (i.e., motorized mode) because of their similarities. Therefore, this section presents the results of mode recognition models when having car and bus modes as a single class. Similar to case 1, two conditions, with and without the GPS data, were evaluated as presented in Table B.3 and Table B.4. When the GPS data are not available, the overall accuracy improved from 95.1% (i.e., from Table B.1) to 97.02% was obtained. The motorized mode (i.e., car/bus) was correctly classified in 98.71% of the time that showed about 5% improvement compared to when having car and bus modes individually (i.e., 93.84% correctly classified as car, 93.12% correctly classified as bus). When using GPS data, the overall accuracy improved even more (i.e., 98.5%). Also, it contributed to a high accuracy of 99.01% for the motorized mode.

Table B.3 Bus and Car combined - without GPS - 97.02% overall accuracy

Random Forest		Actual				
		Bike	Car/Bus	Walk	Run	**Precision**
Predicted	Bike	94.47	1.15	2.43	0.86	94.61
	Car/Bus	2.25	98.71	1.05	0.58	98.08
	Walk	3.24	0.14	96.22	1.58	94.82
	Run	0.03	0.00	0.29	96.98	99.66
	Recall	94.47	98.71	96.22	96.98	

Table B.4 Bus and Car combined - with GPS - 98.5% overall accuracy

Random Forest		Actual				
		Bike	Car/Bus	Walk	Run	**Precision**
Predicted	Bike	98.33	0.91	0.97	1.08	96.34
	Car/Bus	1.04	99.01	0.49	0.34	99.07
	Walk	0.58	0.04	98.11	0.55	98.75
	Run	0.05	0.04	0.43	98.02	99.43
	Recall	98.33	99.01	98.11	98.02	

Appendix C: Driver violation prediction - Sensitivity analysis

This section presents a sensitivity analysis conducted to show how different monitoring periods (i.e., as defined in Chapter 5, in the paper entitled "Predicting Red-light Running Violations at Signalized Intersections using Machine Learning Techniques") affect violation prediction models. As stated earlier in this paper, a monitoring period needs to be examined to extract the information reflecting driver behavior when approaching a signalized intersection. This period (i.e. t^{v}_{mon} as shown in Figure C.1) was determined by choosing its starting and ending points; the starting point of the time window should not be selected too early in order to exclude unnecessary information (i.e. when the drivers are very far from the intersection, their behavior may not reflect their decision on violating the red light). In fact, the behavior of the drivers before the yellow onset may not be related to their decision whether to stop or go. Therefore, the TTI of 6 seconds was chosen as the starting point of t^{v}_{mon} so almost all observations include the yellow onset. Although it is desirable to choose the ending point so that TTI of all vehicles at the yellow onset would be greater than the ending point, the ending point is restricted by t^{v}_{min} which is the minimum time required for the endangered vehicle to respond if a possible collision is predicted. In fact, t^{v}_{min} is equivalent to the sum of two terms: the time required for the driver to respond (t^{e}_{driver}) and the time required for the vehicle to stop ($t^{e}_{vehicle}$). In the aforementioned paper, fixed values for these two terms were assumed (i.e., 1.6 seconds for t^{e}_{driver} and 1.9 seconds for $t^{e}_{vehicle}$). According to a distribution of the human response time as presented by [1], values from 0.5 to 2.5 seconds were chosen for the driver response time (i.e., t^{e}_{driver}) that corresponds to about 5 to 95 percent of population to respond. Also, values from 1.9 to 3.4 seconds were chosen for the vehicle response time (i.e., $t^{e}_{vehicle}$) that corresponds to vehicles approaching at velocities of from 25 to 45 mph. Since the driver response time and vehicle response time need to be added (i.e., $t^{v}_{min} = t^{e}_{driver} + t^{e}_{vehicle}$) to find the end point of the monitoring period, t^{v}_{min} was found to vary from 2.4 to 5.9 seconds. Moreover, the starting point was increased from 6 to 8 seconds because the maximum ending point was found to be 5.9 seconds and the monitoring period of 5.9 to 6 would have been too short. Consequently, the monitoring periods of 2.5 to 5.5 seconds were examined as shown in Figure C.2.

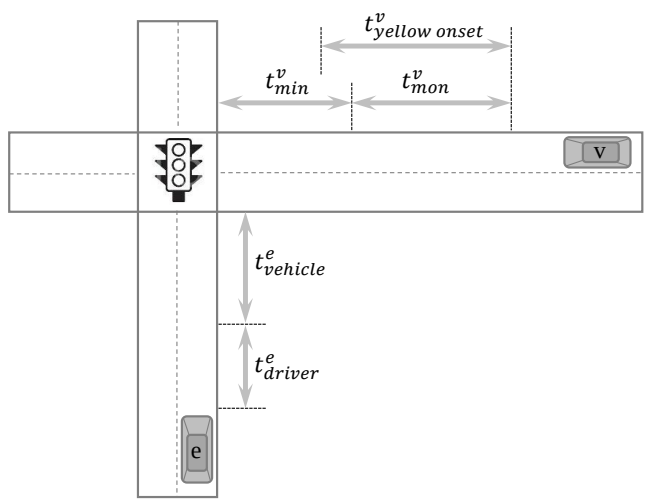


Figure C.1 Sensitivity Analysis on the monitoring period; vehicle ***v*** as the RLR violator and vehicle ***e*** as the endangered vehicle

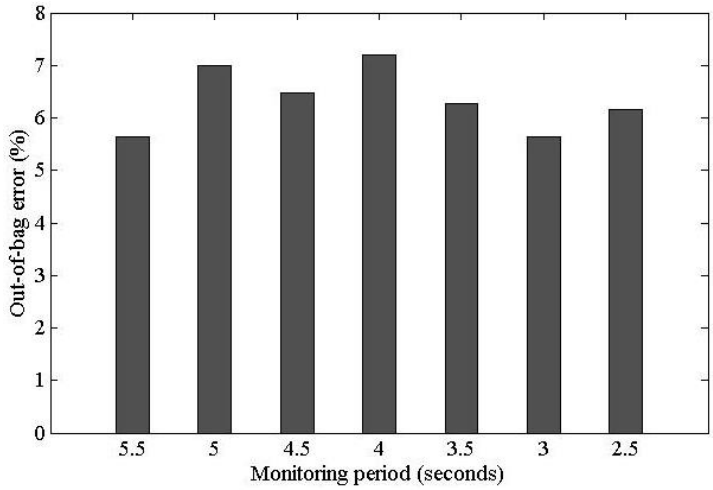


Figure C.2 Sensitivity Analysis on different monitoring periods

Longer monitoring periods are expected to produce lower error rates because more information is available for longer periods. However, such a pattern was not observed in the results as shown in Figure C.2. Two explanations can be thought of as follows: (1) including more information as a result of longer monitoring periods is not sufficient to reduce errors. In other words, although we are adding more information we are still unable to reduce the error. (2) Since the starting points of all monitoring periods were the same, it might have led to including unnecessary information (i.e. the information before encountering the yellow). Unnecessary information might have negatively affected the model performance as this information does not reflect the diver decision as to stop or proceed. To examine if this is the case, sensitivity analysis was conducted in a different way. The starting points were not selected at the same pints and instead they were defined for each individual. In other words, the starting point was different for each driver and it was defined as the point where the driver encounters the yellow light. The results are presented in the third paper (i.e., the AAP journal paper) of Chapter 5 in the dissertation.

References

[1] McLaughlin, S.B., J.M. Hankey, and T.A. Dingus, *A method for evaluating collision avoidance systems using naturalistic driving data.* Accident Analysis & Prevention, 2008. **40**(1): p. 8-16.

Appendix D: Model performance: Violation prediction

After developing the violation prediction models, overall accuracy is one measure to see the model performance, but confusion matrices can also present false positives and false negatives as shown in Table D.1. In this table, a and d represent correct classifications and b and c represent misclassifications (i.e., a: true positives; b: false positives; c: false negatives; d: true negatives). Overall accuracy can be calculated as follows:

$$Overall\ accuracy = (a+b)/(b+c)$$

It should be pointed out the difference between false positives and false negatives. It is desirable to reduce both, but false negatives (i.e., when actual violations are predicted as compliance) are more important in terms of safety and false positives (i.e., actual compliances are predicted as violations) are more important in terms of users relying on the system without getting annoyed. Based on automotive industry recommendations, false positive rates of 5% and below is acceptable to most users [1].

Table D.1 Confusion matrix

		Actual	
		Violation	Compliance
Predicted	Violation	a	b
	Compliance	c	d

As an example, one of the best driver violation prediction models using observational data was when using a monitoring period of four meters and eight factors resulted in overall accuracy of 0.1%. Table D.2 presents the confusion matrix that shows 0% for the false positive rate and 0.2% for the false negative rate. When using the simulator data, the best overall accuracy achieved was about six percent when using four factors. Table D.3 presents the corresponding confusion matrix that shows 5.3% for the false positive rate and 7.1% for the false negative rate. The false positive rate in this case of simulator data is a bit over 5% that is recommended by the auto industry.

Table D.2 Confusion matrix - Driver violation prediction using observational data

		Actual	
		Violation	Compliance
Predicted	Violation	99.8%	0%
	Compliance	0.2%	100%

Table D.3 Confusion matrix - Driver violation prediction using simulator data

		Actual	
		Violation	Compliance
Predicted	Violation	92.9%	5.3%
	Compliance	7.1%	94.7%

Regarding cyclist violation prediction at 4-way stop signs, assuming the end point of the monitoring period is 2, the overall accuracy is about 6%. The corresponding confusion matrix shows the false positive rate and false negative rate to be 6.5% and 7% respectively as shown in Table D.4. The false negative and false positive rates in this case are not as low as those of the driver violation prediction. Also, the false positive rate is 1.5% above the 5% threshold as explained earlier.

Table D.4 Confusion matrix - Cyclist violation prediction at 4-way stop signs

		Actual	
		Violation	Compliance
Predicted	Violation	93%	6.5%
	Compliance	7%	93.5%

For 2-way stop signs, assuming the end point of the monitoring period is 4.6 as discussed in the corresponding paper (i.e., chapter 6), the overall accuracy is about 8%. The corresponding confusion matrix shows the false positive rate and false negative rate to be 6.5% and 12.8%, respectively as shown in Table D.5. The false positive rate did not change compared to the 4-way stop sign models, but false negative rate increased as expected because in the case of 2-way stop signs the prediction task is conducted farther from the intersection compared to the case of 4-way stop signs.

Table D.5 Confusion matrix - Cyclist violation prediction at 2-way stop signs

		Actual	
		Violation	Compliance
Predicted	Violation	87.2%	6.5%
	Compliance	12.8%	93.5%

References

[1] Aoude, G.S., et al., *Driver behavior classification at intersections and validation on large naturalistic data set.* Intelligent Transportation Systems, IEEE Transactions on, 2012. **13**(2): p. 724-736.