APPROVAL SHEET

Title of Dissertation: **ANOMALY DETECTION ACROSS MULTI SCALE TEMPORAL DATA STREAMS FOR HUMAN BEHAVIOR MODELING**

Name of Candidate: Faisal Quader
Doctor of Philosophy, 2020

Dissertation and Abstract Approved: _____

Dr. Vandana Janeja
Professor and Chair
Department of Information Systems

Date Approved: _____

ABSTRACT

Title of Document: **ANOMALY DETECTION ACROSS MULTI SCALE TEMPORAL DATA STREAMS FOR HUMAN BEHAVIOR MODELING**

Faisal Quader, Doctor of Philosophy, 2020

Directed By: Professor and Chair, Dr. Vandana Janeja
Department of Information Systems

This research focuses on developing computational models for human behavior from multi scale temporal data to detect anomalous behavior, evaluating car driving behavior as a case study. Human behavioral patterns capture frequent or repeated behaviors of users in the data. Here the usage data can be multi-scale from devices, computer networks or even vehicle driving data. We define novel computational models that represent these patterns in data in order to associate them with the human behaviors for the detection of the anomalous situations. Such behavioral patterns can be associated with identifying anomalies which we believe are precursors or even indicators of impending or ongoing unexpected behavior. Current state of the art in driving behavior has not focused on assessing behavioral models from multiple streams of temporal data which might be complex. This is an important problem in the domain of driving pattern detection at large because human behavior is impacted by the environment in and around the car. Thus, it is important to study any type of this usage data in combination across multiple data streams to understand a human behavioral perspective.

Through our research we aim to address the discovery of anomalous human behavioral patterns in the driving domain. We present time series based anomaly detection utilizing car

telematics data, eye gaze distraction data and health vital statistics data to provide a comprehensive view of the driver behavioral patterns. We analyze different scales and resolutions of time from seconds to minutes and the anomalous variations and their intensities in the data streams also impacted accordingly from minor fluctuations to major spikes.

We identify anomalies, which might be precursors or even indicators of impending or ongoing unexpected behavior and detect anomalous activities in different settings to understand behaviors in reactive environments such as automobiles. We attribute the anomalous behavior to distraction, driver health, vehicular state or other external factors. Our results indicate that each of the heterogeneous temporal data streams of Telematics data, Eye tracking data, Driver vital health data individually detect anomalies in driving states. However, gaze data is more representative of the anomalies than Health and Telematics data in individual streams. In general, we also found that all data stream combinations are useful, however, presence of anomalies in eye gaze data is more indicative of anomalous behavior. We are also able to supplement the anomalous state information through the combination and overlap of anomalies in the three data streams. We compare results from our methodology with traditional data mining methods and found that some of these overlapping anomalies across the three data streams and unique anomalies in gaze data are missed.

Auto industry and the auto insurance companies gather and analyze mostly Telematics data to gage driver's driving behavior and categorize the safety of the driver according to their driving speed, abrupt acceleration, abrupt deceleration, sharp turns, etc. Our research helps provide a more comprehensive view of the driving behavior by incorporating multiple heterogeneous data streams.

ANOMALY DETECTION ACROSS MULTI SCALE TEMPORAL DATA STREAMS FOR HUMAN BEHAVIOR MODELING

By

Faisal Quader

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, Baltimore County, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2020

This dissertation is dedicated to my wife, my son, my daughter, my brothers & sisters and above all my departed parents for their endless love, inspiration, motivation, and patience. Without them, I would not have been able to complete this process.

Acknowledgements

I am eternally grateful to my supervisor Professor Dr. Vandana Janeja for her guidance and endless patience in bearing with me and my research work. I have learned so much from my professor Dr. Janeja from the very beginning of my PhD course work and other publications. She has guided me every step of the way in my research work in spite of her extremely busy schedule.

I am also thankful to my committee members Professors Dr. Aryya Gangopadhyay, Dr. Zhiyuan Chen, Dr. Wayne Lutters and Dr. Andrea Kleinsmith. All my committee professors have provided guidance and support with my research and dissertation work. They have given me their valuable and unparalleled guidance to make my research work a great success.

Lastly, I am grateful to my university – UMBC for giving me the opportunity to pursue my PhD. Dr. Janeja has been instrumental with her extraordinary teaching and guidance with my dissertation and without whom I would not have been able to come this far in my work toward my Doctor of Philosophy degree. I am equally grateful to my fellow student colleague, a graduate student at UMBC, Mr. Adam Podskalny who assisted me with data collection and experiment as well as working with me on data preprocessing. I am also grateful to my family for their sacrifice and for giving me time and inspiration to work on my PhD.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION

In this research we propose novel methods for anomaly detection across multi-scale temporal data streams. We utilize the automobile-based data streams as a primary domain to establish this research and its major contributions.

The safety of a vehicle depends on the mechanical state of the vehicle as well as the operator's driving pattern for a given time period. Driver behavior can make the driving conditions anomalous by speeding, accelerating, irregular breaking, distracting, poor physical health, etc. along with mechanical failures. Therefore, it is critical to study human behavioral aspects to detect anomalous activities in different settings to understand anomalies in more reactive environments such as automobiles. Our research problem is to identify the anomalous driving behavior and the driving states which are indicators of a variety of unsafe driving conditions. Consequently, the question to this research problem is how we measure the anomalies in driving behavior. The answer to this research question is the measures and analysis of driver behavior data from multiple data streams are predictors of unsafe driving. Moreover, the scale of anomaly from certain domains may change the driving state significantly from safe to unsafe.

This research aims to address the discovery of computational models for human behavioral patterns in automobile driver data. We capture the following three different data streams for the overall driver behavior analysis:

(1) Telematics data with driving patterns

(2) Driver distraction data

(3) Driver vital health data while driving

Three data streams as time series are collected. Anomalies are identified on each data stream. These heterogeneous but related data streams are analyzed to present a holistic approach to detect anomalies in the driver behavior. We study anomaly detection in smart cars to provide driver behavioral patterns. Specifically, we deal with different scales and resolutions of time where computational models for human behavior appear in a heterogeneous manner. These computational models for human behavior would help us identify driver performance.

Data in such system represents humans and their behaviors. We discover anomalies by analyzing the data generated by users in different types of environments to study these behaviors. Thus, the goal of our research is anomaly detection in varying time scales with key focus on human behavior. This human behavior could be "slow online behavior" or "real time behavior" or a hybrid scenario with "slow online and real time behavior". The scope of this research is focused on the 3rd category, which is the driver behavior on a slow online and real time behavior. The hybrid between the slow online and real time behavior is a driving pattern that includes driver behavior on distraction, vital health and automobile's mechanical state. In order to capture such behaviors, we analyze time-based anomaly detection of a smart car and detect any anomalous state of the car that includes driver distraction, driver's anomalous health condition and/or an unusual mechanical state of the car. We explore an example of this time-based human behavior next in the Motivation section.

1.1 Motivation

Data is people, meaning human behavior is portrayed in the data collected in systems in use by humans. We are always alarmed with road safety while we are driving. Driver behavior would be anomalous if there is any kind of cyber threat, mechanical failure, driver's reckless driving,

driver's distraction or driver's degraded health condition. Therefore, monitoring driver behavior in real time and raising alerts would alleviate unsafe environments in automobiles and on the road as well. A majority of road accidents happen in the U.S. because of speeding and other driver behavior related activities [10]. Consequently, it is critical to monitor driver behavior on the road at all times.

This research aims to help organizations take a proactive approach to study driver behavior in near real time. Vehicles (car, bus, fleet) are an absolute necessity in our lives to go to places and/or move goods from one place to the other. We are also making our vehicles smart, fully connected with the Internet to view real time traffic and weather, talk on the phone with Bluetooth, listen to the radio, watch video as well as getting real time status of an automobile's mechanical functions. However, this smart interface comes with a price, which are the vulnerability to threats, distractions, driver's health degradation as well as malfunctions to mechanical parts of the vehicle. We like to exploit the power of the connectivity within the vehicle with machine learning to analyze, alert and secure the vehicle from all sorts of accidents and dangerous situations.

So far, we see the focus on the driving behavior is mainly based on the driver's speed, sharp turns, abrupt acceleration, abrupt deceleration as well as the mechanical state of the automobile through the Telematics data stream. However, telematics by itself does not provide a comprehensive view of driver behavior and the driving state unless we analyze other key data streams like distraction data and driver's vital health data while driving. Consequently our research includes all three data streams of Telematics, Distraction data and vital health data. Each sub stream individually may not give us a comprehensive view of the overall driving behavior and driving state. Anomaly detection on an individual stream will only provide with one aspect of the anomalous driving. In order to have a clear picture of the driver behavior and to safeguard the

driver from getting into an accident, we evaluate not only the full set of key attributes from each domain, but also study multi-domain heterogeneous data streams and identify the relationships among each other.

1.2 Contribution of Thesis

We study multi-scale temporal data in this research to detect both individual and overlapping anomalies across these data streams. We focus on driver behavior along with telematics data, Eye Gaze distraction data, health vital data and loosely the environmental data (weather conditions) to detect anomalous behavior. We also flag the anomalous behavior against baseline data generated from multiple data captures from the driver. Our methodology focuses on discovering driving patterns and finding anomalies from telematics, gaze and health data to determine the safety of the driving state.

The research problems we are solving here are to understand the human behavior and their motive that may cause anomalous behavior, discover computational models for human behavior from usage data to detect anomalous behavior, identify and predict anomalies which might be precursors or even indicators of impending or ongoing unexpected behavior, detect anomalous activities in different settings to understand anomalies in more reactive environments such as automobiles and analyze driver behavior to predict and safeguard from getting into accidents.

Therefore, the objectives of this thesis are to develop a methodology to discover driving patterns using vehicular data, detect anomalous driving states, utilize telematics data, Eye Gaze distraction data, and health vital data to detect anomalies, discover deviations from baseline driver behavior to detect anomalies, and attribute the anomalous behavior to distraction, driver health, vehicular state.

In order to achieve the above objectives, we evaluate Telematics data stream (T), Eye tracking data stream (D), and Driver vital health (H) individually to help detect anomalies in driving states. We then combine Telematics data stream (T), Eye tracking data stream (D), Driver vital health (H) to help detect anomalies and find the relationships among different data streams. As anomalies are indicative of unsafe driving conditions, we find that Distraction (D) and Telematics data (T) are related to each other in detecting anomalies, and Distraction (D) and Health data (H) are also related to each other. We also find that Distraction (D) data is the most prominent indicator of anomalous driving state. Therefore, the measures of driver behavior from multiple data streams are predictors of unsafe driving. The scale of anomaly is important here and even a single domain can influence the safety of a driving condition.

CHAPTER 2

RELATED WORK

Extensive research has been done on APT, Social Engineering as well as Smart Car safety in general, however, we lack the security aspect as well as the real time anomaly detection on the human interaction as time progresses for the above areas. Driver behavior and safety depend on the driver's attention to driving, his/her health as well as the automobile's mechanical state. Vehicle's mechanical state may change significantly from safe state to unsafe anomalous state if there is mechanical failure or driver's high speed driving or cyber-attack on the automobile. What happens if I lose control of the car while driving because of the engine being taken over by a cyber intruder? Therefore, along with our research on driver behavior, it is also important to study cyber threats to the smart cars. Our research provides a good foundation on cyber threats to automobiles for further research. Following is the synopsis of related work in these areas.

2.1 Persistent Threat Patterns

Slow online behavior could be persistent and very slow online activities. For instance, Advanced Persistent Threat (APT) on a smart car is a slow online persistent threat to penetrate into the network. A group of cyber hackers were able to get into Tesla's ECU (Electronic Control Unit) of an S model after trying to hack into the system for a while and literally hijacked the brake and steering wheel [33].

Detecting persistent threats and finding the related patterns are complex. Mandiant [75] has investigated cyber-attacks since 2004 and finally came to the conclusion that a large organization from China including the Chinese government was involved in massive advanced persistent threats on hundreds of organizations including the US military. Mandiant uncovered that

97% of 1905 APT incidents were coming from Shanghai IP's. The report however does not show specific patterns by which we can predict and identify possible persistent threats.

Binde et al. [12] have provided various APT detection strategies including Open Source Tools (such as Snort, Scapy, Splunk, Sguil and Squert), Rule Sets, Statistical and Correlation Methods. However, these approaches do not identify consistently repetitive unusual threats.

Finally, Namayanja and Janeja [84] study the discovery of persistent threat structure through temporal and geo-spatial characterization by monitoring central nodes to determine consistency and inconsistency in their availability across time periods. However, this approach does not address repeated threat patterns over time but primarily focuses on shifts in network communication over time.

2.2 Social Engineering

The real time behavior could be social engineering like attack that is an act of psychological persuasion of gaining access to their system by getting their trust. This can very well happen to an intelligent connected automobile and its RSU (Road Side Unit) within its vehicular network by malicious nodes that would impersonate RSUs in an attempt to trick users into divulging their authentication details [91, 92, 14].

Social engineering is defined as a mechanism of getting people to comply in order to gain access with computer systems and the information that resides there in an unauthorized manner. The goal of social engineering is to obtain information that will allow the hacker to gain unauthorized access to a system in order to commit fraud, identity theft, disrupt or compromise a network, or to commit industrial espionage [91]. Types of social engineering are categorized as Human based and Computer Mediated methods [91, 92]. Human Based methods include

impersonation, the overly helpful helpdesk, third-party authorization, tech support, roaming the halls, repairman, trusted authority figure, snail mail, etc.

As mobile, cloud computing and social media technologies as well as smart cars continue to grow, the security concerns associated with them also grow significantly due to lot of users not complying with the security policies that makes it easy for hackers to target the key stakeholders [102]. Bloomberg et al. [14] highlighted the fact that social engineering can turn into a cyber-war where it involves multiple countries. The attacks on Google and other companies in China in 2009 were initiated through phishing – the underlying technical exploit is often trivial but social engineering is always the entry strategy [59]. Information security is more than an engineering challenge as people are essential part of the critical infrastructure. Therefore, understanding and addressing human behavior is essential to building a genuine security culture. Bad actors purposely build trust to enable their actions [128]. The insider threat like espionage and data leakage involving computer networks is among the most pressing cybersecurity challenges within government and private sectors. Greitzer, et al. has implemented a psychological reasoning for human behavior in cyber-attacks and categorized the attackers with disgruntlement, difficulty accepting feedback, anger management issues, disengagement, and disregard for authority [47, 48]. Greitzer, et al. also revealed that current or former employees and contractors are the second greatest cybersecurity threat, exceeded only by hackers, and that the number of security incidents is continuing to increase [48]. Yang et al. showed models to detect possible human cyber threats by virtue of building behavioral profiles. Any anomaly deviates from the baseline would raise the security risk [129]. Finally, Anderson et al. emphasized the fact that the user base for home computers is exponentially increasing. As a result, the system is becoming vulnerable and susceptible to all kinds of cyber-attacks. They hypothesized that psychological ownership of one's

computer is positively related to behavioral intentions to protect one's own computer [9]. Likewise, the technological advancements of the automobile industry and connectivity through the Internet inside the car can make the automobile susceptible to social engineering attacks where intruders can hack into the known vehicles.

However, we believe that all these studies have neglected one crucial element: how do we detect one of the most common social engineering attacks in the network using a computational model? Our study here would build the foundation to help detect the anomaly in driving condition which could be the effect of the social engineering attack and/or any other cybersecurity threat.

2.3 Anomaly Detection in Smart Cars

Anomaly detection of smart cars is a common phenomenon these days and quite a bit of research has been done to alert drivers for mechanical malfunctions of the engine. For instance, Foster et al. [40] talks about collecting telematics data from OBD2 (On-board Diagnostic) device that reads mechanical condition of the car and alerts telematic failure as it happens. Dr. Miller and Chris Valasek showed how to simulate a cyber threat to the automobile remotely and take over the control of mechanical engine with break and acceleration [81]. Koscher et al. [68] demonstrate that an attacker who is able to infiltrate virtually any Electronic Control Unit (ECU) can leverage the ability to completely circumvent a broad array of safety-critical systems. This helps the car manufacturer taking care of those vulnerabilities for future models. Research shows that an automobile using a GPS based system can be vulnerable to different kinds of attack, including blocking, jamming, spoofing, and physical attacks [57]. New car manufacturers are taking these findings and making sure the new automobiles are less susceptible to all these cyber threats. In spite of these precautions and mitigations to vulnerabilities, the hackers are smart enough to infiltrate to the system and cause damages.

Telematics is just one of the many aspects to derive the overall anomalies in driving state of a driver. For instance, distraction is another major impact that cause unsafe driving state at any point of time. According to a WHO (World Health Organization) report, more than a million deaths are caused by the road accidents in the world because of some sort of distraction [24]. Among these casualties, majority of them are distraction related [73]. Moreover, the health condition of the driver may impact the driver behavior [98, 49]. For example, if the blood pressure or blood sugar goes outside of normal range while driving, that may not be a safe environment for the driver as well as the surrounding traffic. Our research analyzes the temporal driver behavior that includes both distraction and vital health and detects anomalous state of the driving state as well as the anomaly from automobile's basic mechanical functions. We alert the driver in near real time, so the driver can take immediate action to avoid adverse effects of accidents. We bring in human behavioral aspects of the driver and predict the anomalous behavior looking at the pattern on a time series.

2.4 Survey of Cyber Attacks and Leading Factors

We performed a comprehensive survey of cyber threats that includes cyber-attacks on automobiles. Our background research focuses on understanding the characteristics and causes of multiple types of cyber-attacks through a comprehensive evaluation of case studies of real-world cyber-attacks. For each type of attack, we studied attack type, their characteristics and disclose the causes of that attack. Key characteristics included in our study are: type of industry, financial intensity of the attack, non-financial intensity impacts, volume of impacted customers, users' trust & loyalty impacted. In addition, key causes included are: human behavioral aspects leading to attacks, cultural factors at play, security policies adapted, technology adoption and investment by

the business, training & awareness of all stakeholders including users, customers, employees and investments in cybersecurity.

This study can help take a proactive approach to analyze relevant cyber threats and educate the organizations to become more knowledgeable and sophisticated on how to minimize damage caused by such relevant cyber-attacks. Our findings indicate that human behavioral aspects leading to attacks are the weakest link to successful prevention of cyber threats. We focus on human behavior that is responsible for major cyber-attacks and concentrate on the mitigation strategy. So far, we see discoveries on specific cyber threats and its mitigation plan. However, we perform a comprehensive study of most of the cyber threats in the industry and study them with human behavior that relates to the persistent threats, social engineering and predictive driver behavior. Extensive research has been conducted to identify cyber threats to the network and the computer systems of an organization. We also see extensive study pertaining to the security of self-driven car. However minimal research has been done to mitigate cyber threats targeted toward a smart automobile that are driver operated [72]. Consequently, we provide a foundation of studying driver behavior that would assist doing further research on the cyber threat related anomalous driver behavior.

2.5 Temporal Anomaly Detection

Our research focuses on anomaly detection across multi scale temporal data streams for human behavior as it relates to the driving behavior. Temporal anomaly detection is a classical way to detect anomalies in a time series. We analyze signal data from multi-data streams on a time scale and compare the anomalies from different data streams. In order to validate the anomalies, we apply matrix profile (MP) to visualize the distance (variance) on the subsequent data attributes that shows the significant changes in the data stream and hence shows the anomalies.

If we experience anomalies in automobile data, study shows various ways to detect unusual behavior on the automotive network via temporal anomaly detection. When we experience statistically significant deviation from the training set for CAN (Controller Area Network) bus attributes, we detect anomalous state in the automotive network [106]. We can also reduce high dimensionality on a time series by Piecewise Aggregation Approximation (PAA). The PAA applies to the distribution of the points of each equal sized segment on a time series to provide focused data attributes [50]. We see a similar Multi-Domain Anomalous Temporal Association (Multi-DATA) to our proposed anomaly detection strategy where Shukla et. al compare the anomalies in a single domain with other domains on a time scale [104]. If the anomalies do not match on the exact time series, this paper considers the anomaly on the closed proximity of the time windows. In our case, data coming from three different domains are binned on a specific time scale to compare the similarities and correlations on anomalies from all three domains. Usually anomalies on one data stream correspond to the other data streams to some extent. We apply different data analytics mechanisms to validate the results for high levels of accuracy.

2.6 Overall Discussion on Related Work

We have discussed five different aspects of anomaly detection for smart cars that are related to the analysis of driving behavior. Cybersecurity and cyber threats may not be directly related to driver behavior; however, the driving state of a driver, his/her driving behavior as well as the safety of the driver are impacted by cyber-attacks. We talk about Advanced Persistent Threats (APT) hacking into the systems on an automobile and taking over control by an outsider while driving. Likewise, Social Engineering is another form of cyber intrusion in an automobile and having access to the Road Side Unit (RSU) of a smart car and changes the safety features of the car to make the driving state very unsafe. We see extensive research on cyber threats and its survey for

autonomous cars, however, we do not see much research on the security of a smart car that is not self-driven. Consequently, our research will help identify the state of the driver and its safety by looking at the telematics data along with driver behavior to accurately find the anomalies of a driver. Cybersecurity attack mostly will show up through mechanical engine issues as captured through telematics anomalies. For instance, if an intruder attacks on the brake system of a vehicle and takes over the brake system and speeds up the vehicle abnormally, then our model will detect the anomaly the same way as if the driver speeds up the vehicle intentionally. Hence, it would be difficult to pinpoint the reason for telematics anomaly whether it happens because of the engine malfunction by the driver or by the engine wear and tear or by the cyber attacker. We can potentially indicate cyber threats in an engine if the telematics reading is extremely anomalous in a very short period of time. However, that has to be further investigated as future research. Therefore, our research provides a good foundation for cyber threats related anomalies for future work. Telematics by itself would not give us the full picture of the driver behavior, rather we have added two more key attributes – distraction and the vital health of the driver to identify the anomalies for a driver behavior. We study multi-dimensional data attributes related to the driver behavior to present an accurate measure of anomalous driving behavior. Anomaly detection on drivers was done on specific single threaded attribute like telematics only. However, we do not see a comprehensive study on multi-threaded attributes like gaze data and driver's vital health data along with telematics to present an accurate driving state and its dependencies. Finding the relationships between different driving data attributes on a time scale to detect the anomalous behavior of a driver is the main objective of our study.

CHAPTER 3

METHODOLOGY

In our study, we have identified a thematic focus corresponding to time variations on human behavior for anomaly detection. We study driver behavior on a defined time series and predict the driver's driving behavior to ensure driver's safety. Consequently, time and human behavior are the common themes for anomaly detection in this research.

Studying driver behavior for potential cyber threats and malfunction on the automobile engine and detecting driver's anomalous driving behavior because of this along with distraction and health issues are very important aspects for driver's safety. We baseline safe driving patterns and any deviation from that would be considered anomaly.

3.1 Overall Research & Methodology Approach

In our research, we have a common theme of detecting human behavior and anomalous behavior on multiple time scales. In order to accomplish that, we gather real time driver data from three main data streams: Telematics data from OBD2 device, eye gaze data (as a measure of distraction) from Tobii Eye Glass 2 and vital health data from E4 Empatica. (1) The driver wears the Tobii Eye glass that is connected to the data collection laptop as shown in Figure 1. (2) The driver also wears a wrist band E4 watch-like device to capture health related data stream. (3) The automobile is also connected to the OBD2 device to read telematics data in real time.

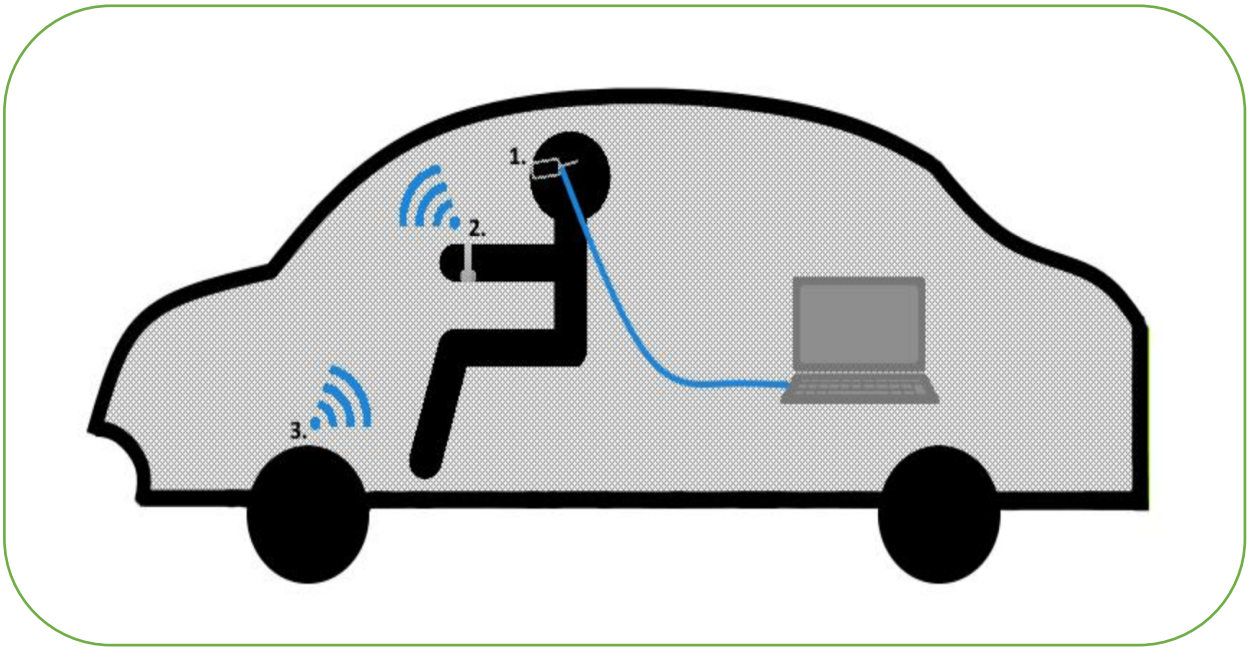


Figure 1: Driver & Automobile Setup for Data Collection

We categorize the driving behavior into the following four driving states:

- **Safest**: Most safe condition
    - All Telematics data attributes are within the normal range, and
    - Eye gaze is exclusively to the target (on the road), and
    - Health data is within the normal range (standard)

Normal range is determined by the attributes close to the mean. Any value outside of 1 level of standard deviation is considered out of normal range and anomalous. Since the scale of anomaly is very important here especially for gaze data, a small deviation from the mean can be unsafe. Therefore, one standard deviation is the benchmark for detecting an anomaly. For the Safest driving condition, all Telematics, Eye Gaze and Health data attributes are close to the normal range (mean).

- **Safer**: Relatively safer condition
    - Almost all the Telematics data attributes are within the normal range, and

- Eye Gaze is mostly to the target (on the road), and
- Health data is almost within the normal range

Safer condition signifies almost all the data attributes from all three data streams to be close to the mean with very few exceptions. For instance, the driver is very attentive and focused on driving with very slight distraction, the health of the driver is within the normal range with minor deviation on the vitals like heart rate being little higher than normal and above all the speed, acceleration, auto engine measures are mostly in the normal range.

- **Safe**: Within the safe condition
  - Most of the Telematics data attributes are within the normal range, and
  - Eye Gaze is to the target (on the road) except occasional distractions, and
  - Health data is mostly within the normal range

Safe condition signifies most of the data attributes from all three data streams to be close to the mean with very few exceptions. For instance, the driver is attentive and focused on driving with a few occasional distractions, the health of the driver is within the normal range with minor deviation on the vitals like heart rate being higher than normal at times and above all, the speed, acceleration, auto engine measures are mostly in the normal range.

- **Unsafe**: Not in a safe condition
  - Several Telematics data attributes are not within the normal range and/or
  - Eye Gaze is not to the target (on the road) several times and/or
  - Health data is not within the normal range

Unsafe driving condition occurs when several data attributes from all three data streams are outside the mean that cause the driver to be in a very insecure condition. For instance, the driver is speeding

and doing several abrupt accelerations, abrupt decelerations, distracted with phone, texting, adjusting the navigation and radio, health vitals may be outside the range of a normal driver. Therefore, the baseline of a safe driving state as shown in table 1 should be as follows:

Table 1: Safe Driving Condition

Safe Driving Condition
Telematics data attributes are within the normal range (close to the Mean value) and
Eye gaze is to the target (on the road) and
Health data is within the normal range (standard) and
After normalizing the data attributes, the data points are close to the mean value

Unsafe: Not in a safe driving state away from the baseline. We have a well-defined set of the ranges for normal Telematics and vital health data. As part of the normalization strategy for the data attributes, we normalize the attributes between 0 to 1 where the mean is the baseline for normal (safe) data attributes. Table 2 depicts the unsafe driving condition.

Table 2: Unsafe Driving Condition

Unsafe Driving Condition
Several Telematics data attributes are not within the normal range and/or
Eye Gaze is not to the target (on the road) several times and/or *(For Eye Gaze data, a small variation from the mean for a very short period of time may change the driving state from safe to unsafe. As a rule of thumb, data attributes outside of 1 standard deviation are considered anomalous)*
Health data is not within the normal range and
After normalizing the data attributes, the data points are not close to the mean value

3.2 Data Analytics Rules & Collective Anomaly Assessment

Based on our driving conditions, we have come up with specific rules that establish the validity of our theory. We are focusing on Collective Anomaly as we are analyzing multi-scale anomalies among Telematics, Health and Distraction data streams that can be interrelated. Following is the set of definitions and the respective rules for our driving conditions:

(a) <u>Data Stream:</u>

Data stream is the set of data attributes coming from different heterogeneous multi-scale domains.

Definition:

Data Stream D, such that each data stream i in D is d^i where i = {1,2,3,…,n} data streams such that |D| = n, i.e. number of data streams is n.

Example:

D = {d^M, d^G, d^H}, n = 3, d^M is Telematics, d^G is Gaze data stream (Distraction) and d^H is Health data.

(b) <u>Anomaly:</u>

Anomaly is the data attributes from the data stream that are deviated from the norm.

Definition:

Anomaly A, such that for each data stream d^i, anomalies generated are represented as a^i where i = {1,2,3,…,n} data streams such that |A| = n, i.e. number of anomaly sets is n corresponding to each data stream

Example:

A= {a^M, a^D, a^H}, n = 3

where for any data stream i, $a^i = \{d^i_j \mid d^i_j$ *occurs at time* $t_j \;\&\&\; d^i_j > +\sigma \;||\; d^i_j < -\sigma\}$ *where* σ *(standard deviation) varies from 1 to* 3σ *and* a_t^{di} *represents anomaly for the corresponding time period t in the data stream i*

The data attributes - Telematics (M), Distractions (D) and Health (H) are anomalous if they are between 1 and 3 standard deviation from the mean.

(c) <u>Collective Anomaly:</u>

Collective anomaly is the anomalous data attributes across data streams that are deviated from the norm comparing to the other data streams.

Definition:

Collective Anomaly C_A *spans across data streams such that* $C_A = \{a_t^{d1} \cap\ a_t^{d2} \cap a_t^{d3} \cap \ldots . a_t^{dn}\}$

such that $a_t^d \subset a^{d1}$ *and* $a_t^{d2} \subset a^{d2} \ldots\ldots\ a_t^{di} \subset a^{di}$

Example:

$C_A = \{a_t^M \cap a_t^G \cap a_t^H\}$ *for Telematics, Gaze and Health data streams*

such that $a_t^M \subset a^M$, $a_t^G \subset a^G$ *and* $a_t^H \subset a^H$

(d) <u>Measuring the Anomaly Spread:</u>

We measure the anomaly spread across multi domains for overlapping anomalies.

Definition:

$$DA_\Omega = \frac{Count^d(|\ a_t^{d1} \neq \emptyset\ |\cap|\ a_t^{d2} \neq \emptyset\ |\cap \ldots\ldots |\ a_t^{di} \neq \emptyset\ |)}{n}$$

d is the number of overlapping data streams

$$if\ DA_\Omega == 1, then\ Collective\ Spread\ Impact$$

$$if\ DA_\Omega < 1\ \&\&\ > 0, then\ Limited\ Collective\ Spread\ Impact$$

$$if\ DA_\Omega == 0, then\ Non-overlapping\ Spread\ Impact$$

$$if\ DA_\Omega == 0, a_t^{di} == 0, then\ Normal\ State$$

Collective Anomaly Spread Impact usually makes the driving state to be unsafe.

Limited Collective Spread Impact usually makes the driving state to be somewhat safe.

Non-overlapping Stream Impact usually makes the driving state safer.

Normal State with no overlap and no anomaly makes the driving state the safest.

Example:

$a_t^M = 15$, $a_t^G = 12$, $a_t^H = 10$

All three data streams have at least one overlapping anomaly, then $DA_\Omega = 3/3 = 1$ => this is a collective spread impact (Unsafe).

The more comprehensive data points with real world examples will be furnished in Chapter 4 - Experimental Results section.

(e) <u>Measuring the Temporal Overlaps:</u>

We measure the anomaly overlaps on a time series for multi domain data streams.

Definition:

$$OA_\Omega = \frac{Count^a|(a_t^{d1} \neq \emptyset) \cap (a_t^{d2} \neq \emptyset) \cap \ldots\ldots (a_t^{di} \neq \emptyset)|}{n}$$

a is the anomaly on the overlapping data streams

$$if\ OA_\Omega \geq 1, OA_\Omega\ in\ \left(1, \frac{|t|}{n}\right) then\ Collective\ Temporal\ Overlaps$$

$$if\ OA_\Omega < 1\ \&\&\ > 0, then\ Limited\ Collective\ Temporal\ Overlaps$$

$$if\ OA_\Omega == 0, then\ Non - overlapping\ Anomaly$$

$$if\ All\ |a_t^{di}| == 0, then\ Normal\ State$$

Collective Temporal Overlaps usually make the driving state to be unsafe.

Limited Collective Temporal Overlaps usually make the driving state to be somewhat safe.

Non-overlapping Anomaly usually makes the driving state safer.

Normal State with no overlap and no anomaly makes the driving state the safest.

Example:

$a_t^M = 15$, $a_t^G = 12$, $a_t^H = 10$

Intersect with 10 anomalies, then $OA_\Omega = 10/3 = 3.3$ => Collective Temporal overlaps (Unsafe)

(f) <u>Measuring the Intensity of Anomalies</u>

Intensity of anomaly depends on either high # of anomalies in one domain or high overlaps across multiple domains. Higher the number, the more intense the single stream impact is.

Definition:

$$IA_\Omega = \frac{max(|\, a_t^{d1} \neq \emptyset \,|,\ |\, a_t^{d2} \neq \emptyset \,|,\ldots\ldots|\, a_t^{di} \neq \emptyset \,|)}{|t|}$$

Max is the anomaly in every t instance

$if\ IA_\Omega \geq 0.05, then\ Collective\ Intensity, IA_\Omega\ in(0.05,1)$

$if\ IA_\Omega < 0.05\ \&\&\ > 0, then\ Limited\ Collective\ Intensity$

$if\ All\ |a_t^{di}| == 0, then\ normal\ state$

Collective Intensity usually makes the driving state to be unsafe.

Limited Collective Intensity usually make the driving state to be somewhat safe.

Normal State with no overlap and no anomaly makes the driving state the safest.

Example:

$a_t^M = 200$, $a_t^G = 150$, $a_t^H = 100$, $t = 600$ seconds

Then $IA_\Omega = 200/600 = 0.33$ => Collective Intensity (Unsafe)

(g) <u>Measuring Weighted Anomaly Spread</u>

This rule measures the weighted anomaly spread that drives the safety of the driving behavior. We combine the anomaly spread with the intensity of anomalies by multiplying them to find the weighted anomaly spread as shown in figure 2.

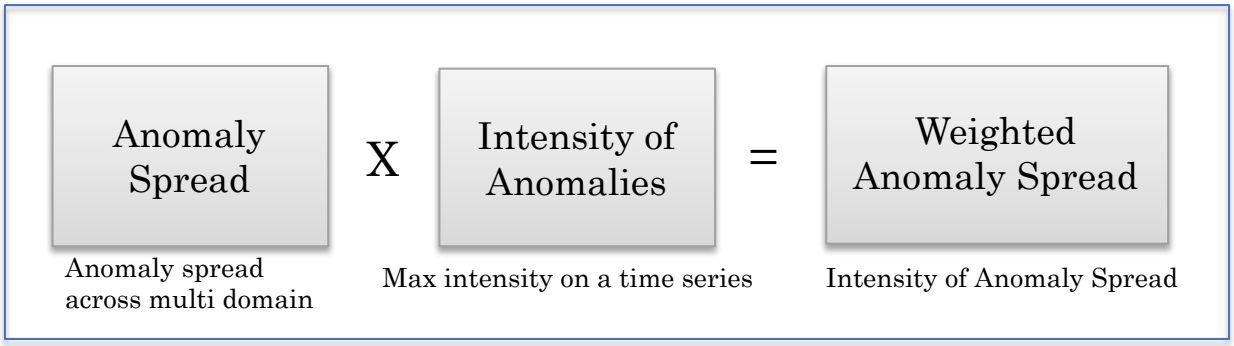


Figure 2: Collective Anomaly Assessment (Weighted Anomaly Spread)

Definition:

$WDA_{\Omega} = DA_{\Omega} \times IA_{\Omega}$

$if\ WDA_{\Omega} \geq 0.05, then\ Collective\ Weighted\ Spread, WDA_{\Omega}\ in(0.05,1)$

$if\ WDA_{\Omega} < 0.05\ \&\&\ > 0, then\ Limited\ Collective\ Weighted\ Spread$

$if\ \ WDA_{\Omega} == 0, then\ non - overlapping\ weighted\ stream$

$when\ DA_{\Omega}\ has\ non - overlapping\ anomaly$

$if\ \ WDA_{\Omega} == 0, then\ normal\ state\ when\ both\ DA_{\Omega}\ \&\ IA_{\Omega}\ are\ 0$

Collective Weighted Spread usually makes the driving state to be unsafe.

Limited Collective Weighted Spread usually makes the driving state to be somewhat safe.

Non-overlapping Anomaly usually makes the driving state safer.

Normal State with no overlap and no anomaly makes the driving state the safest.

Example:

$a_t^M = 15,\ a_t^G = 12,\ a_t^H = 10,\ t = 600\ seconds$

All three data streams have at least one overlapping anomaly

Then $DA_{\Omega} = 3/3 = 1$ => this is a collective spread impact

$IA_{\Omega} = 15/600 = 0.025$

WDA_{Ω} = *1 x .025 = .025* => Limited Collective Weighted spread (somewhat Safe)

(h) <u>Measuring Weighted Temporal Overlaps</u>

This rule measures the relative overlaps for anomalous states across different domains. We combine the anomaly overlaps on a time series with the intensity of anomalies by multiplying them to find the weighted temporal overlaps as shown in figure 3.

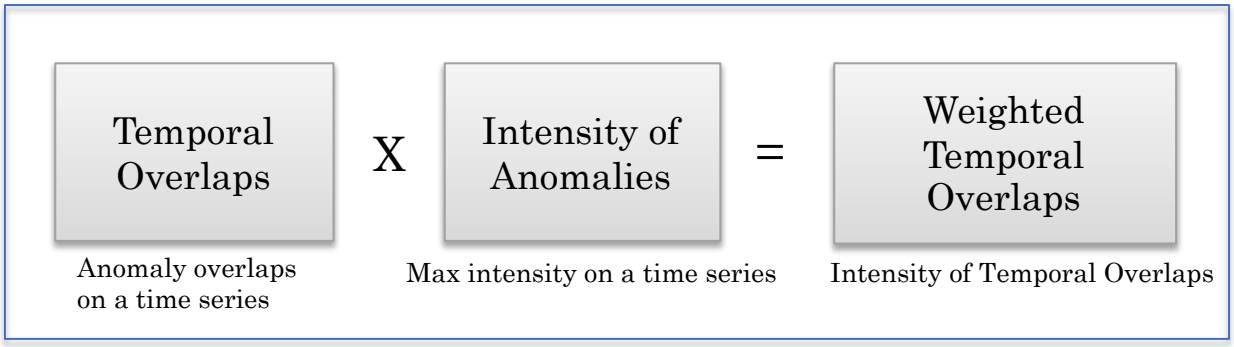


Figure 3: Collective Anomaly Assessment (Weighted Temporal Overlaps)

Definition:

$WOA_{\Omega} = OA_{\Omega} \times IA_{\Omega}$

$if\ WOA_{\Omega} \geq 0.05, then\ Collective\ Weighted\ Temporal\ Overlaps$

$if\ WOA_{\Omega} < 0.05\ \&\&\ > 0, then\ Limited\ Collective\ Weighted\ Temporal\ Overlaps$

$if\ \ WOA_{\Omega} == 0, then\ non - overlapping\ weighted\ Temporal\ Anomaly$

$when\ OA_{\Omega}\ has\ non - overlapping\ Weighted\ Temporal\ Anomaly$

$if\ \ WOA_{\Omega} == 0, then\ normal\ state\ when\ both\ OA_{\Omega}\ \&\ IA_{\Omega}\ are\ 0$

$if\ All\ |a_t^{di}| == 0, then\ normal\ state$

Collective Weighted Temporal Overlaps make the driving state to be unsafe.

Limited Collective Weighted Temporal Overlaps usually make the driving state to be somewhat safe.

Non-overlapping Weighted Temporal Overlaps usually make the driving state safer.

Normal State with no overlap and no anomaly makes the driving state the safest.

Example:

$a_t^M = 15, a_t^G = 12, a_t^H = 10, t = 600$ *seconds*

Intersect with 10 anomalies

Then $OA_\Omega = 10/3 = 3.3$

$IA_\Omega = 15/600 = 0.025$

$WOA_\Omega = 3.3 \times 0.025 = .0825$ => Collective Weighted Temporal Overlaps (Unsafe)

Intensity and the impact of distraction like behavior are very high with low scale of anomaly. Therefore, the threshold for both the Collective Intensity of Anomaly, Collective Weighted Anomaly Spread and Collective Weighted Temporal Overlaps are set to only 5% to capture most of the high impact anomalous driving behavior. We will misclassify the collective anomaly if we make the threshold higher and mistakenly categorize the high impact anomalous behavior to mid to low impact anomalies.

(i) <u>Measuring the relative presence of certain domain anomaly across data streams</u>

This rule provides the relative (%) presence of anomaly for each domain across different data streams shown in figure 4. The higher the percentage the more dominating presence the domain has. The relative presence of a domain anomaly is an independent percentage for a given time series that may have a ceiling values of 1 {bounded by 1 => 600 anomalies (a) in 600 seconds (t)}

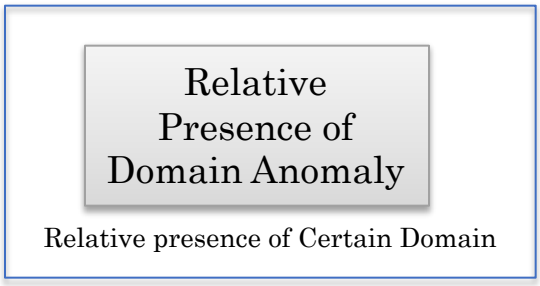


Figure 4: Collective Anomaly Assessment (Relative Presence of Certain Domain)

Definition:

$\delta_i = \frac{a_t^{di}}{|t|}$, δ_i *is the relative anomaly for a data stream i*

Example: $\delta_M = \frac{a_t^M}{|t|}$, *where M=Telematics(400)* , $\delta_H = \frac{a_t^H}{|t|}$, *where H=Health(200)* , $\delta_G = \frac{a_t^G}{|t|}$,

where G=Gaze (Distraction) (300), t = 600 seconds

δ_M*=400/600 = 0.67 => 67%*

δ_H*=200/600 = 0.33 => 33%*

δ_G*=300/600 = 0.50 => 50%*

Most prevalent (67%) anomaly is coming from Telematics data stream

(j) ***Measuring the weighted intensity of Anomalies across data streams***

Here we measure the intensity of a certain anomaly as in figure 5. This measure provides the influence of one anomalous data attribute over others in collective data streams. This rule depicts which data stream has the most prominent indicator of anomalous driving state.

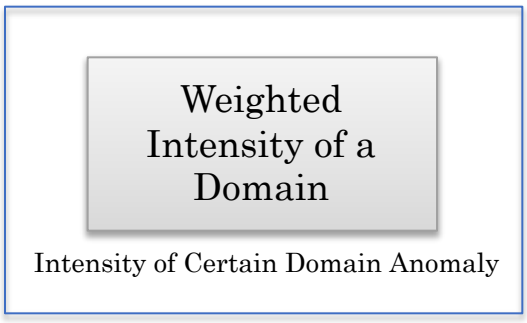


Figure 5: Collective Anomaly Assessment (Weighted Intensity of a Domain)

Definition:

$WAI_{\Omega}^{di} = \delta_i * \partial_i$

δ_i is the relative presence of the domain *i* among different domains

∂_i *i*s relative weight of anomalies for domain *i* among different domains

Relative weight is determined by the landscape of different domains that can influence anomalous driving behavior based on the literature and domain expertise.

Example:

Based on the domain expertise and extensive literature, $\partial_G > \partial_H > \partial_M$, meaning the influence of Gaze anomalies is higher than the influence of Health anomalies that has higher influence than Telematics anomalies. From our extensive research and domain knowledge, we see almost 50% of the accidents happen because of some sort of distraction, about 30% of the accidents happen for driver's health related issues, and about 20% of the accidents happen for Telematics related anomalies.

Therefore, we set $\partial_G = 0.5,\ \partial_H = 0.3,\ \partial_M = 0.2$

$a_t^M = 400,\ a_t^G = 200,\ a_t^H = 300,\ t = 600$ *seconds*

$\delta_M = 400/600 = 0.67 => 67\%$

$\delta_H = 200/600 = 0.33 => 33\%$

$\delta_G = 300/600 = 0.50 => 50\%$

$WIA_{\Omega}^{di} = WIA_{\Omega}^{M} = 0.67 \times 0.20 = 0.13$

$WIA_{\Omega}^{di} = WIA_{\Omega}^{H} = 0.33 \times 0.30 = 0.10$

$WIA_{\Omega}^{di} = WIA_{\Omega}^{G} = 0.50 \times 0.50 = 0.25$

Weighted Intensity for Gaze is 25%, Weighted Intensity for Health is 10%, and Weighted Intensity for Telematics is 13% on a time series t (600). Even though Gaze has a smaller number of anomalies than that of Telematics, the intensity of Gaze is higher (25%) than the intensity of Telematics (13%). Therefore, Gaze anomaly (25%) creates more unsafe driving state than Health (10%) anomaly and Telematics (13%) anomaly.

3.3 Data Capture and Analysis

We have divided our data capture and analysis into two parts. First, we captured only the OBD2 data and analyzed the data streams individually. Afterwards, we added Tobii eye glass and the E4 wristband to include both distraction data and vital health data to have a comprehensive driver behavior analysis.

We have utilized the following process as shown on the figure 6 when we perform a standalone telematics data analysis: First we collect data, then preprocess the data. After that, we apply feature selection to extract and focus on the key attributes. Next, we perform temporal binning to segment the datasets by time series and we apply clusters and clustering & classification based association rule mining to get to the final result of anomaly detection.

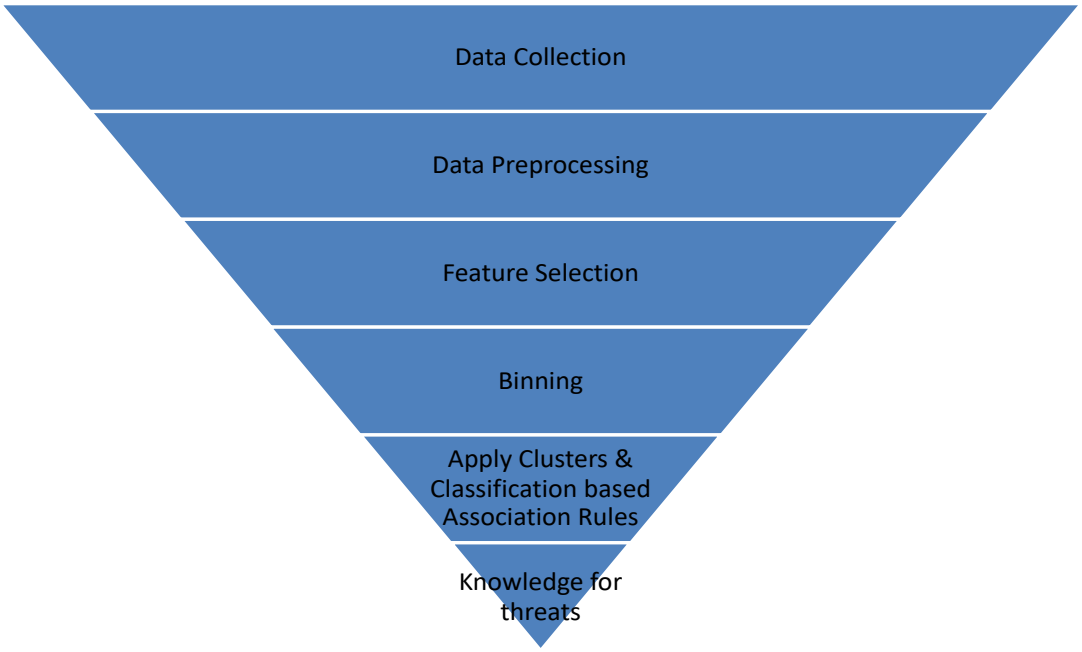


Figure 6: Overall Methodology Approach

Following is the overall methodology for driver behavior as shown in Figure 7, where we capture and preprocess data from four different data streams from four subjects (drivers) and identify the key attributes using Feature Selection methodology. We baseline the driver behavior via clustering and compare the driver attributes with the baseline to identify anomalous behavior. We then validate the result by applying the Association Rule Mining (ARM) with Classification based Association.

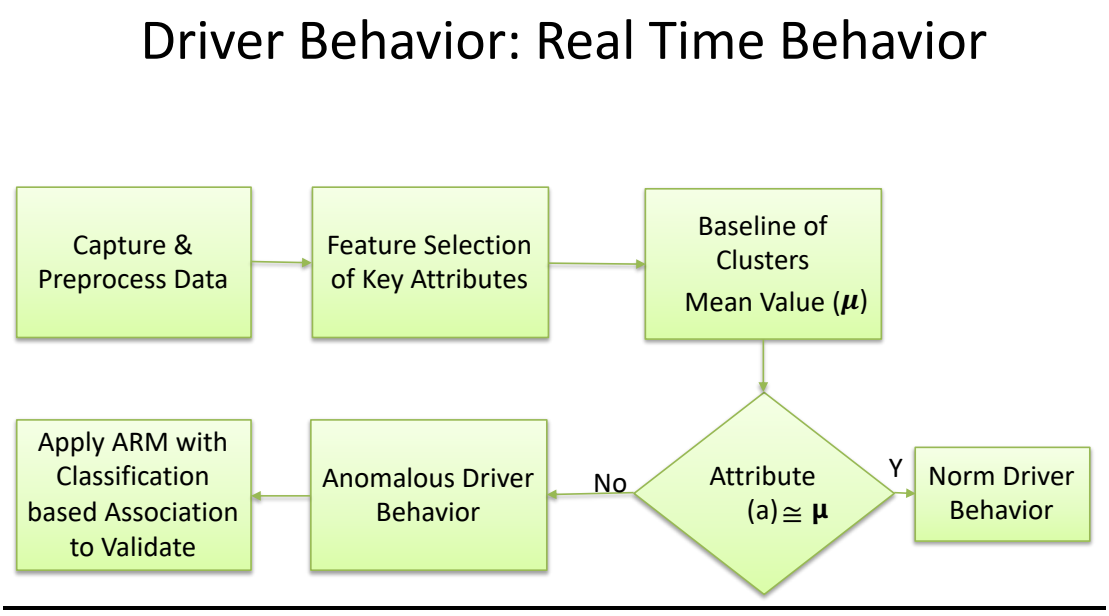


Figure 7: Driver Behavior Anomaly Detection: Real Time Behavior

In figure 8 below, we are showing 4 drivers from whom we collect the driving data. For every driver, we create the baseline cluster for each data stream. For anomaly detection on just telematics data stream, baseline cluster refers to the K-means clustering based on well-defined industry wide telematics attributes that are within the normal range. We also create a baseline for the combined data streams as the second part of the experiment with multi-scale data. The baseline in this part refers to the centroid (data attributes close to the mean 0.5) of the normalized data streams between 0 and 1 for all three data streams. Any deviation from the baseline is categorized as unsafe driving.

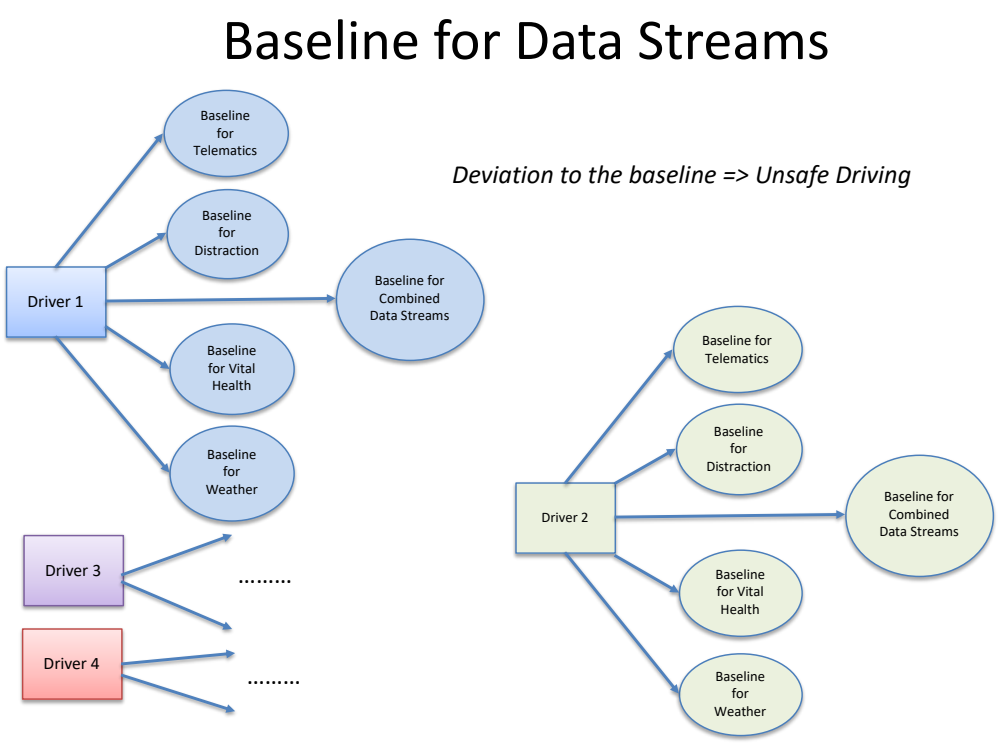


Figure 8: Baseline for Data Streams

In addition to the anomalies from the baseline, there can be direct relationship/correlation between the attributes of one data stream to the attributes of other data streams as shown in Figure 9. For instance, both visual and cognitive distractions (attributes of distraction data stream D_1, D_2, etc.) can cause sharp turns and/or abrupt acceleration/deceleration (Telematics data attributes T_1, T_2, etc.). Likewise, fast heartbeat (V_4) and distraction (D_4) can be correlated. Moreover, the adverse weather condition (W_2) can cause tension (V_2) and distraction (D_2) to the drivers. We have not studied weather impact closely in this research and leave this attribute for future work.

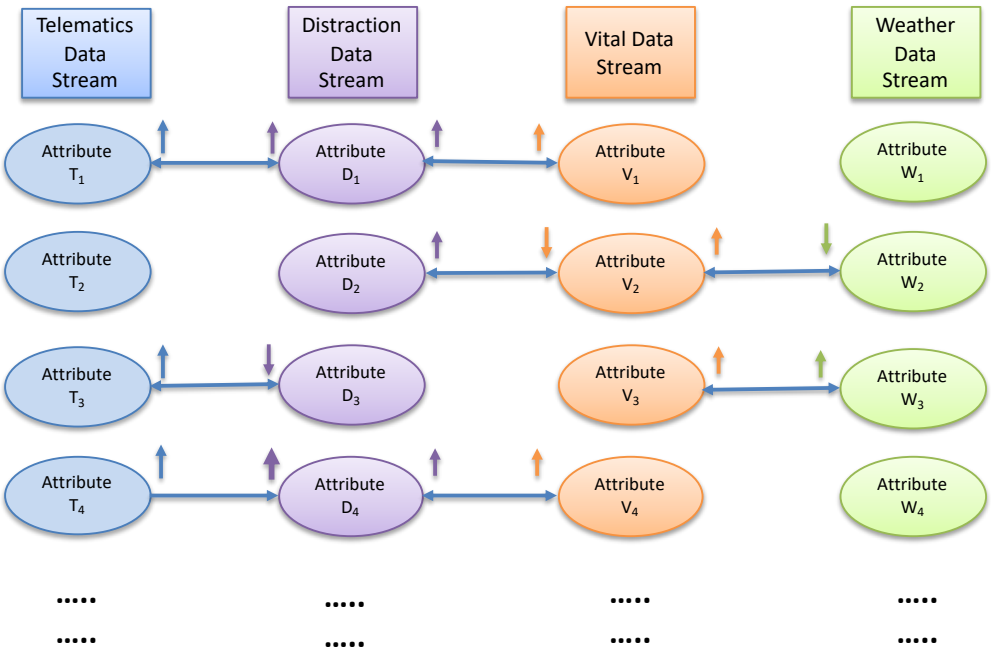


Figure 9: Direct Correlation & Overlaps

Our research would assist the drivers to see their driving patterns and take appropriate actions to become safe drivers. For instance, our children's safety depends on the school bus drivers' safe driving practice. The school transportation can utilize our dissertation to distinguish the unsafe drivers from the safe drivers and initiate mandatory safety training for their unsafe driving behavior.

3.4 Study Design:

Following are the overall design of the study:

As described above, we initially conduct anomaly detection on Telematics data only and then we combine all three data streams as the second part of the experimentation.

(1) Part One: Analyze driver behavior based on Telematics Data only

Data Analytics on Telematics (T) only

For Anomaly Detection on just Telematics data, (1) we first collect telematics data using the OBD2 device connected to the automobile. (2) We then preprocess the raw data by cleaning up the dataset for Feature Extraction. (3) Apply Feature Selection methodology to extract extraneous data attributes and focus on the key attributes. (4) We perform Temporal binning to segment the datasets by time series. (5) We then apply clusters and K-mean clustering. (6) Finally, we apply Classification based Association Rule Mining to derive the final result of anomaly detection.

(2) Part Two: Combined Distraction data and Vital Health data with Telematics data to detect anomalies in driver behavior

Data Analytics on Eye Tracking (D), Health (H) & Telematics (T) data combined

For anomaly detection on a combined data stream, (1) we first collect Eye Tracking data from Tobii eye gaze device, collect health data with E4 Empatica wrist band and collect Telematics data OBD2 device. (2) We preprocess data by cleaning the noise data and taking care of the missing data. (3) We then apply feature selection to extract and focus on the key attributes. (4) We normalize all three data streams. (5) We perform matrix profile (MP) on the combined data sets to identify the drastic changes in the data attributes. (6) We then compare the signal data among different datasets on a given time scale (Bin). (7) We compare the MP data among different datasets on a given time scale (Bin). (8) We then find the correlation/similarities from the signal data and the MP data. (9) We also compare the multi-scale data on the same time span to detect the relationships. (10) We also validate the relationships using ARM and identify the key rules for among different data streams (11) Finally we identify anomalous driving states. Different domains

have different ranges and scales, and hence we normalize all the heterogenous data attributes under the same range so we can compare and run data analytics models effectively.

3.5 Driver Behavior based on Telematics Data only

The automobile industry has come a long way to introduce new technology along with self-driven capabilities, Forward-Collision Warning (FCW), Automatic Emergency Braking (AEB), A backup camera, Blind-spot monitoring, Bluetooth connectivity, voice controls, to name a few [83]. All these sophistications come with vulnerabilities and we need to ensure that we overcome these vulnerabilities and we can secure our vehicles. The moment we are connected to the Internet in our vehicles, we are susceptible to intrusions and threats from outside. Therefore, we need to make sure we have proper anomaly detection mechanism in place in real time to take immediate action to avoid any accidents. Driver behavioral patterns over a long period of time create a training set and any deviation to the norm would impose a threat to the safety of the driver.

Therefore, we are focusing on anomaly detection of a smart car data and detect any anomalous state of the car to alert and protect the drivers and the riders in near real time. We will study several critical scenarios of risks, perform behavioral analytics on a time series, and derive health check of the vehicle in real time to ensure the safety of the drivers and the riders. Driver Behavior analysis is driven by the following key questions:

- Can we identify the features that capture driver's behavior?
- Can we differentiate driver behavior vs. intrusion?
- Can we extract driver behavior?
- Can we detect anomalies in driver behavior?
- Is it natural driver behavior or is it something outside of driver behavior?

Following is the set of steps taken to derive answers to the above questions as shown in figure 10.

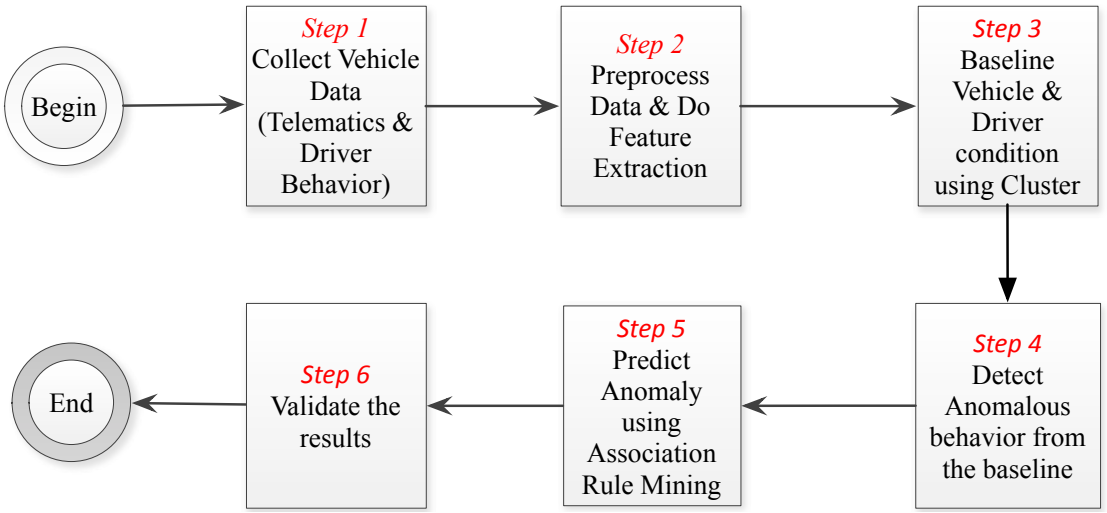


Figure 10:Block diagram for Anomaly Detection of Vehicle and Driver Behavior

3.5.1 Data Collection

Driver behavior analysis along with capturing the mechanical state of the engine is done through a data collection mechanism from our automobiles as shown in figure 11 below. We collect data using an OBD2 based mobile application called Torque that collects telematics data in real time. OBD2 device is connected to the automobile and the telematics data is collected from the device to the Android based mobile app via Bluetooth connection.

Figure 11: Telematics data Collection Mechanism

Through an android based custom application developed at Technuf LLC we capture raw OBD data attributes from the vehicle in real time and computes “derived” data like Sharp Turns, Sudden Acceleration, etc. as indicators of key driver behavior as shown in figure 12.

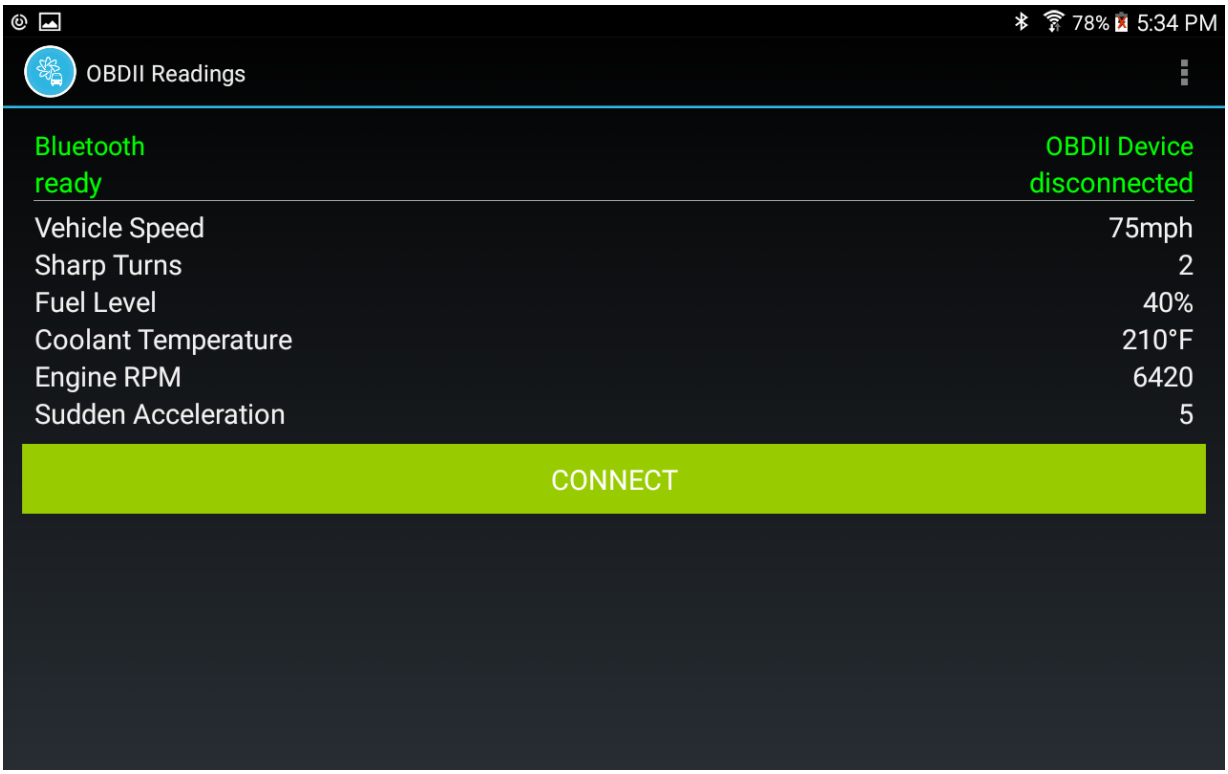


Figure 12: Custom Mobile App to Capture "derived" Driver data

3.5.2 Feature Selection

We captured numerous attributes from the vehicle to study driver behavior on a time series. The sources are Telematics data through OBD2 device and derived data like sharp turns, abrupt acceleration, abrupt deceleration, Engine RPM, etc. We applied feature extraction to shortlist the key number of attributes. The objective of variable selection is three-fold:

- Improving the prediction performance of the predictors
- Providing faster and more cost-effective predictors
- Providing a better understanding of the underlying process that generated the data. [51]

Therefore, in our research on driver behavior, feature extraction helps us narrowing down the key attributes among several data points captured from the vehicle and helps us identify the anomalies in an efficient way.

Automatic feature selection methods can be used to build many models with different subsets of a dataset and identify those attributes that are and are not required to build an accurate model. A popular automatic method for feature selection provided by the caret R package is called Recursive Feature Elimination or RFE. We use RFE method with Random Forest algorithm on each iteration to evaluate our model. The algorithm is configured to explore all possible subsets of the attributes. The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) is a frequently used method that measures the differences between values (sample and population values) predicted by a model or an estimator and the values actually observed. We apply RMSE (Root Mean Square Error) as the Y-axis and the Variables in the x-axis. All 49 telematics attributes are selected here, although as an example, the plot in figure 13 shows the accuracy of the different attribute subset, and we can see that 9 attributes like 2, 11, 19, 24, 29, 31, 32, 38, 40 give almost comparable results for a subset. Therefore, we can pick either one of these 9 attributes and eliminate the rest.

[2] "Longitude"

[11] "G.calibrated."

[19] "Distance.to.empty..Estimated..miles."

[24] "Fuel.flow.rate.minute.gal.min."

[29] "GPS.Latitude..."

[31] "GPS.Satellites"

[32] "Horsepower..At.the.wheels..hp."

[38] "Percentage.of.Highway.driving..."

[40] "Speed..GPS..mph."

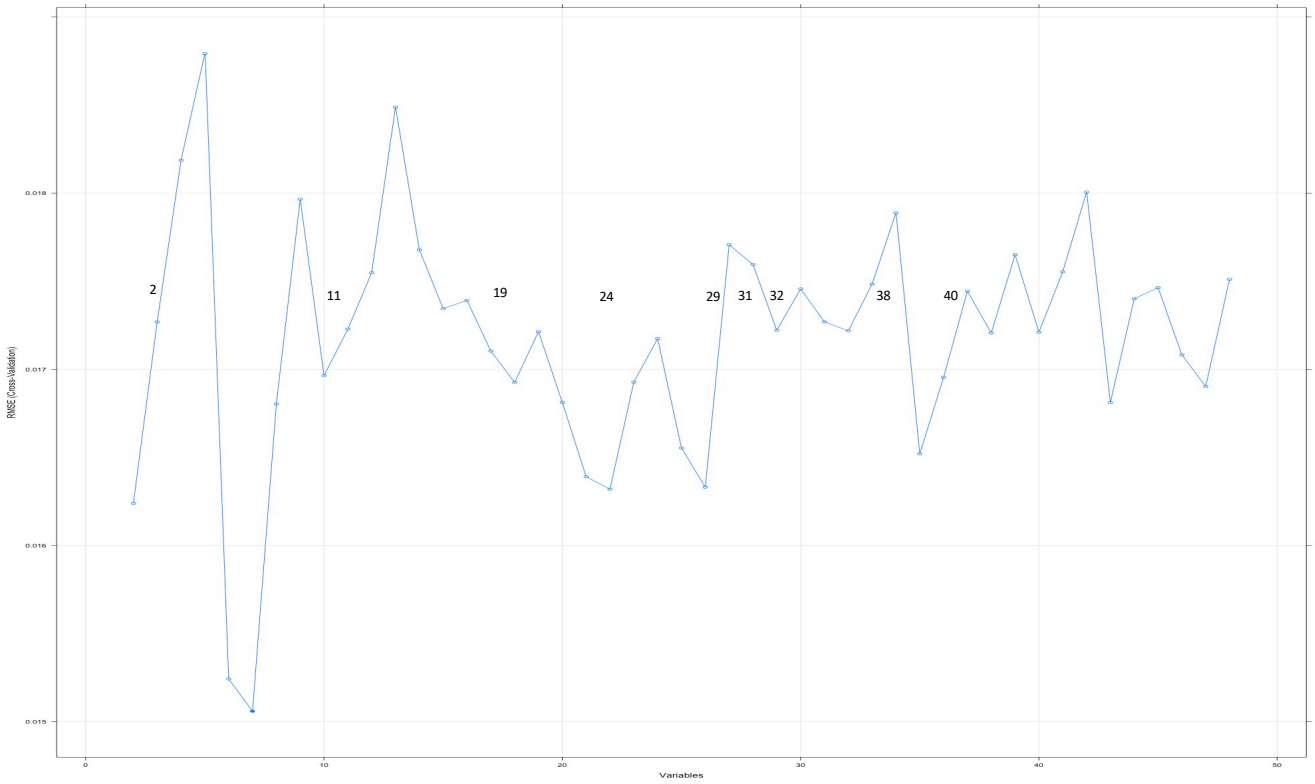


Figure 13: Plot for Estimating of Variable Importance

3.5.2.1 Correlation Matrix:

Data may contain attributes that are highly correlated with each other. Many methods perform better if highly correlated attributes are removed. Generally, we want to remove attributes with an absolute correlation of 0.75 or higher. We conduct correlation matrix to perform feature selection and reduce the number of attributes. We do not remove any key attribute that is a deterministic data attribute for anomaly detection. For instance, the variables according to the column number which are highly correlated are:

[17] “Average.trip.speed.whilst.moving.only..mph.”

[38] “Percentage.of.Highway.driving... “

They have a correlation of 0.98349208. So one of the variables can be removed.

Figure 14 shows a matrix depicting the correlations between all pairs of attributes.

Interpretation of the Diagram and Ellipses are as follows:

The narrow Ellipses have high correlation.

The positive slope and blue are for positive correlations.

The bar in the right side represent the correlation values according to the color.

The other plot in figure 15 represents the correlation with numeric values for the attributes.

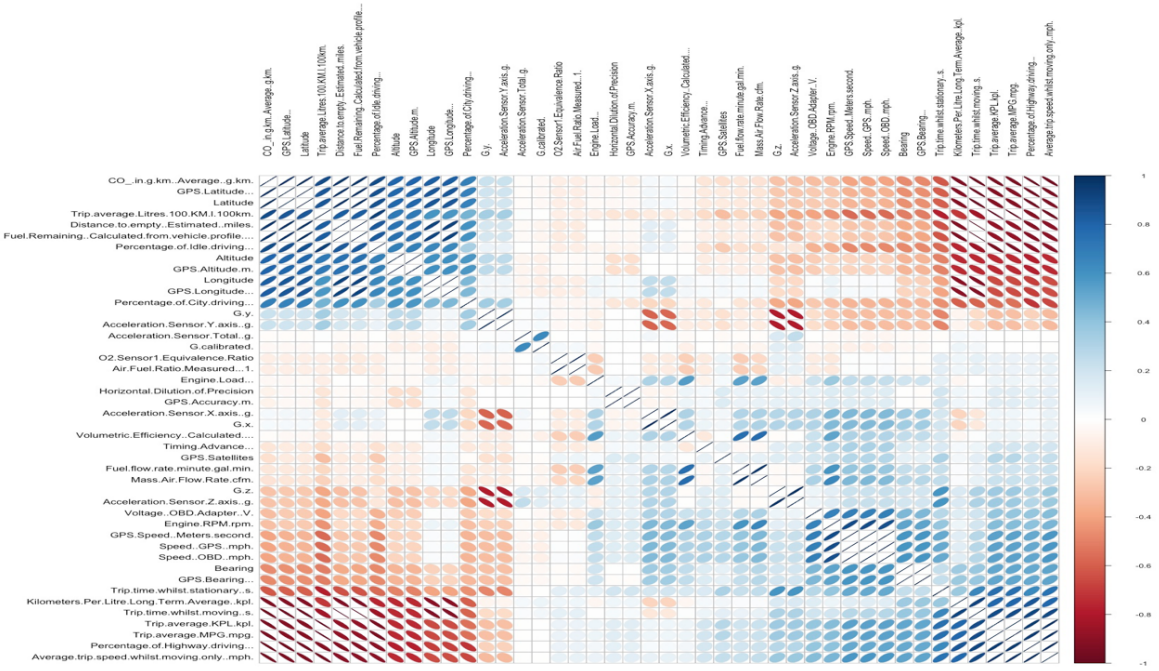


Figure 14:Correlation Matrix for All 49 Attributes

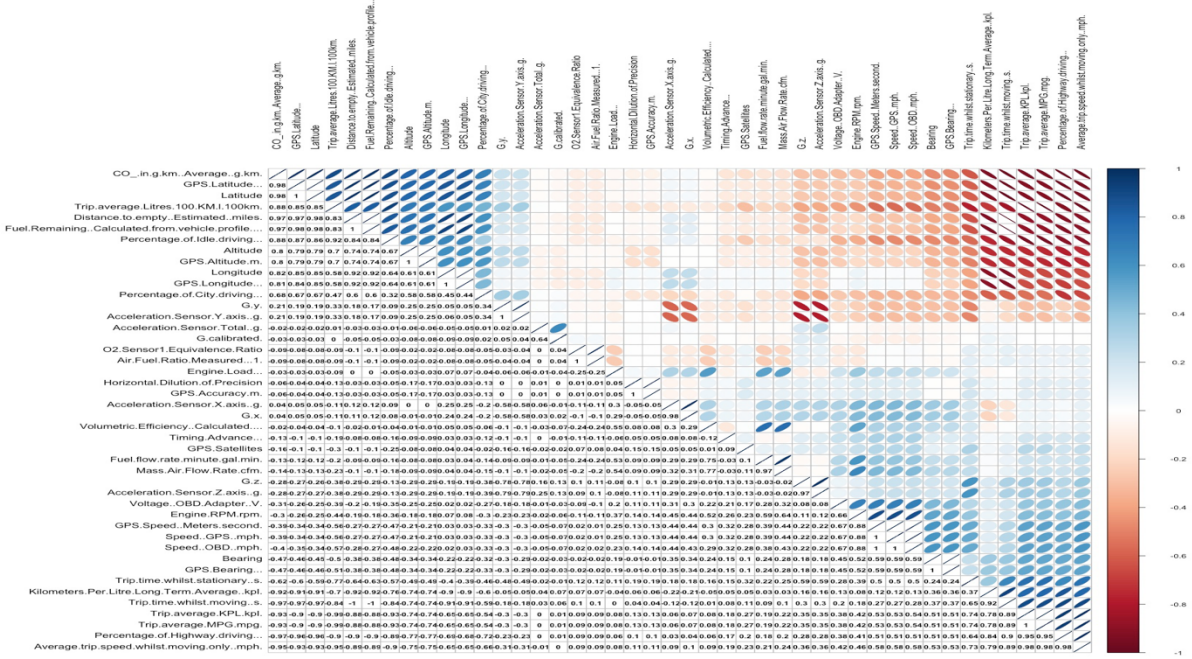


Figure 15: Correlation Matrix with Numeric Values

Likewise, we can eliminate similar attributes as an attribute reduction strategy. For instance, we analyze the attributes from the OBD2 data (file: 'trackLog-2017-Jun-11_16-57-43.csv') and find the [1]acceleration related attributes to be [12] "Acceleration.Sensor.Total..g.", [13] "Acceleration.Sensor.X.axis..g.", [14]"Acceleration.Sensor.Y.axis..g.", [15]"Acceleration.Sensor.Z.axis..g."

Similarly, [4]"GPS.Speed..Meters.second." and [40]"Speed..GPS..mph." are similar to [16]Speed (OBD).

Lastly, [33]Trip speed is related to [43]"Trip.average.KPL.kpl.", [44]"Trip.average.Litres.100.KM.l.100km.", [45]"Trip.average.MPG.mpg.", [46] "Trip.time.whilst.moving..s." and [47]"Trip.time.whilst.stationary..s."

3.5.2.2 Telematics and derived data attributes analysis by cars:

We compared telematics and derived data attributes from 4 (Mercedes, BMW, Toyota and Honda) cars and highlighted the data attributes with common denominator. Based on our analysis and the data coverage from all 4 cars, we are focusing on the following key attributes in Table 3:

Table 3: Focused Key Attributes

Acceleration Sensor
Actual Engine % Torque – Engine RPM
Air Fuel Ratio
Avg Trip Speed
Engine Coolant Temperature
Engine RPM
Fuel Flow Rate/min
Fuel Remaining
Fuel Used
Speed (OBD)
Driver Behavior (Sharp Turn/Acceleration/Brake)
Engine speed / threshold violation – Engine Coolant Temperature
Low Battery – Voltage Adapter
Live weather and traffic updates
Real time Heart Rate Monitor - *Future*

3.5.3 Baseline of Driver Data

We need to first baseline the driver behavior and the vehicle's normal condition. We are doing clustering to get a good baseline. We use the elbow method to find the number of clusters. After identifying the K value, we perform K means clustering on the data and present a good set of baseline data for each subject. We use K-means clustering because it is relatively simple to implement and works well for large data sets like a large number of telematics data records. Moreover, we do not have any hierarchical nature of the data. K-means clustering is also fast and efficient in terms of computational cost.

We categorize the baseline clusters into the following 4 different types of safety condition, such as:

- **Safest:** *Most safe condition*
- **Safer:** *Relatively safer condition*
- **Safe:** *Within the safe condition*
- **Unsafe:** *Not in a safe condition*

We categorize the baseline clusters into the following 4 different types of safety conditions, such as:

- **Safest:** $\mu_{CS} - \mu_B \cong 0$
- **Safer:** $\mu_{CS} - \mu_B = \delta_1$ (deviation from mean)
- **Safe:** $\mu_{CS} - \mu_B = \delta_2$ (more deviation from mean)
- **Unsafe:** $\mu_{CS} - \mu_B = \delta_3$ (Significant deviation from mean)

3.5.4 Anomaly Detection of the Driver Behavior

We compare new driver data with their baseline clusters and categorize to the appropriate safe condition. If we find clusters outside of these conditions, then we trigger them as anomalous situations. As we perform single stream anomaly detection on each subject, we find the baseline for that subject. Then we also compare the baselines from each to see the similarities and differences. We can also combine all the subjects and perform anomaly detection on the combined data streams. Finally, we applied ARM (Association Rule Mining) capability on our vehicle data. We first converted the numeric data to categorical data. We introduced different ranges of data applicable to the numeric dataset. These ARM rules along with the Classification Based Association (CBA) rules imply specific driving behavior based on certain parameters on the vehicle.

3.6 Driver Behavior based on Combined Data Streams - Distraction, Health & Telematics

To address the emerging phenomenon of autonomous vehicles that perform the task of driving at a level comparable to the most capable driver, the quantifiable and repeatable process of driving must be closely examined. Achieving proficiency as a human driver requires intuition built by the awareness of every sensory component used in the process. The model that replaces this intuition must be built to consider as many minute changes in both the human driver and mechanical vehicle as can be observed. This research study employs technology to observe and quantify the behavior of the test subject and vehicle, and with conduct time-series analysis using Matrix Profile subsequence detection to identify how to a standard of safety for autonomous vehicles can be defined.

Along with Telematics data, we need to add 2 or 3 more key factors that can influence driver behavior: they are Distraction data stream, vital Health data stream and weather data. We use Tobii Eye Glass to measure distraction of the driver. The attributes selected in the Tobii Glasses 2 data stream represent the change in distance from fixation on a given point in three-dimensional space (Figure 16). In addition, the amount of time spent with fixation on this point and the measured size of the subject's pupil represent the level of attention focused on this point.

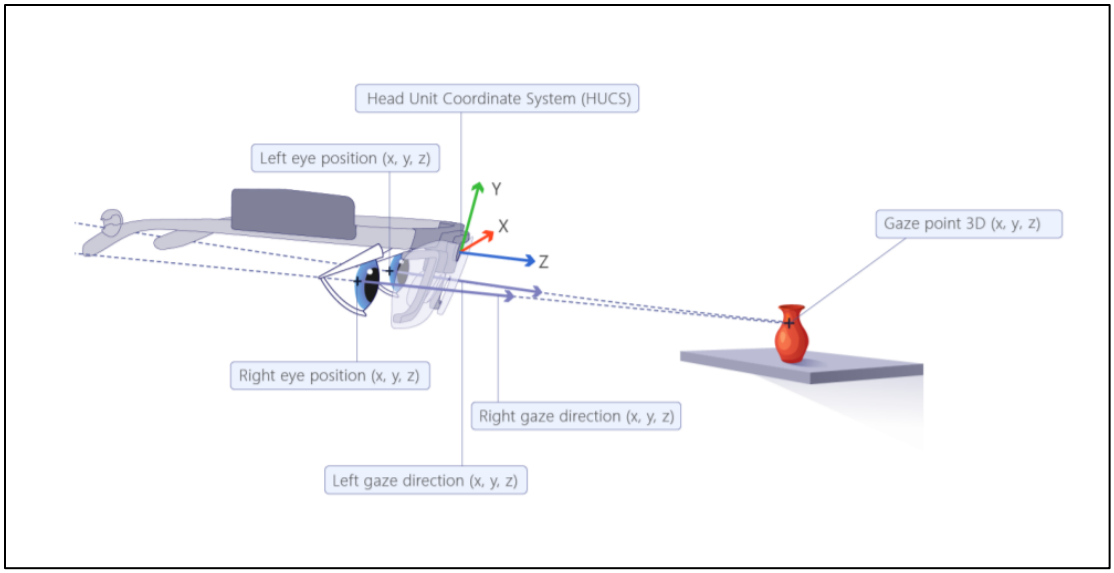


Figure 16: Anomaly Detection for Combined Data Stream

We use the E4 wristband to measure vital health attributes of a driver while he/she is driving. The health data attributes would determine whether the driver is fit to drive with no issue. On the contrary, the driver behavior could be influenced by the state of health of the driver. For instance, if a driver is having a fever or tensed with high heart palpitation or high/low blood pressure than normal, that may make the driver drive in an anomalous way. Lastly, the weather conditions like rain, snow, cloudy, dark, sunny situation may influence the driver to some extent although we are leaving the weather condition as an attribute to study as a future work. We perform the following steps as laid out in figure 17 to capture different data streams from different devices, extract the key attributes, preprocess the baseline data and detect anomalies.

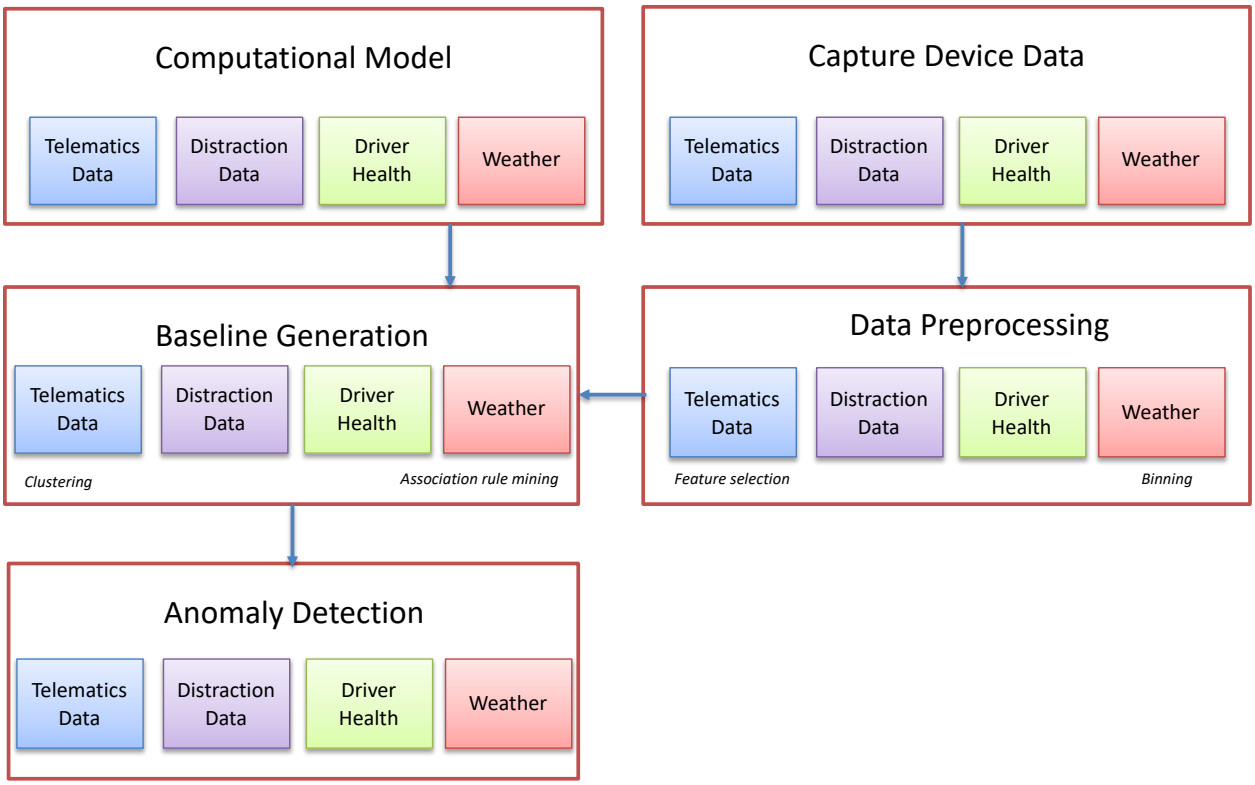


Figure 17: Anomaly Detection for Combined Data Stream

3.6.1 Data Collection

Following 4 data streams shown in table 4 are being captured simultaneously while driving from 2 vehicles of luxury and non-luxury class: BMW and Toyota Corolla

Table 4:Four Key Data Streams

Data Stream	**Data Attributes**	**Measurable & Accessible**
Data Stream 1: State of the Vehicle & Driving Pattern while driving	Telematics Data (Air Fuel, CO, Engine Temp, RPM, etc.) Sudden Acceleration Sharp Turns	Telematics + Custom App
Data Stream 2: Distraction	Eye Gaze (*Looking away from the road – How many times on a specific time span*) Lane deviation (*How many times on a specific time span*) Steering smoothness (*How many times on a specific time span*) Texting with driving (*How many times on a specific time span*)	Tobii Pro Glasses 2

Data Stream 3: Driver Health	Heartbeat Blood Pressure Pulse Fast Breathing	E4 Wristband from Empatica
Data Stream 4: Weather	Sunny (*Map to 1*) Cloudy (*Map to 2*) Rain (*Map to 3*) Snow (*Map to 4*)	Smart Phones

Figure 18: Driver wearing Tobii Eye Glass 2 for Distraction data attributes

The subject in figure 18 is wearing the Eye Glass that is connected to the laptop inside the car to capture the gaze data of the driver. At the same time, the driver is wearing the E4 Empatica Wristband to retrieve health condition related data from the driver while driving as shown in figure 19. At the same token, as described above, we would still have the Telematics OBD2 device plugged in to the automobile to capture automobile data on our android mobile app – Torque and our custom mobile app. We capture the data streams at the same time of the day with similar weather so we can defer the weather related influence for future research.

Figure 19: Driver wearing E4 Wristband for vital health data attributes

3.6.2 Feature Selection

From the raw data coming out of all three data streams, we discarded the noise data, took care of the missing data, and feature extracted the key attributes. We show the data reduction strategy in figure 20. We want to focus on the high impact health data attributes related to driver behavior. Hence, we feature selected HR, EDA and BVP from our health data stream. We receive some noise data in our reading which are completely out of the norm. We record every reading and replay the recordings to find the mis-match with the noise data and discard them. Noise data occurs because of the wristband not making proper contact with the skin while turning the steering wheel repeatedly. We have had to calibrate and adjusted a few times while doing trial driving. Even though the EDA reading had delays by couple seconds, it does not impact the overall rules and the anomaly detection results and the driving behavior on a large time series. We also clean and replace the missing data with the average value of the data stream. Consequently, as a smoothing strategy, we use similar to low pass filtering which is replacing the

missing records with average value. Also we discard the irregular, out of range data outside of the possible range using a python script. Moreover, depending on the sun light and/or darkness on the road, the pupil diameter reacts differently. Therefore, in order to keep the data reading consistent, we choose to do the driving with all the subjects at the same time of the day during the same season.

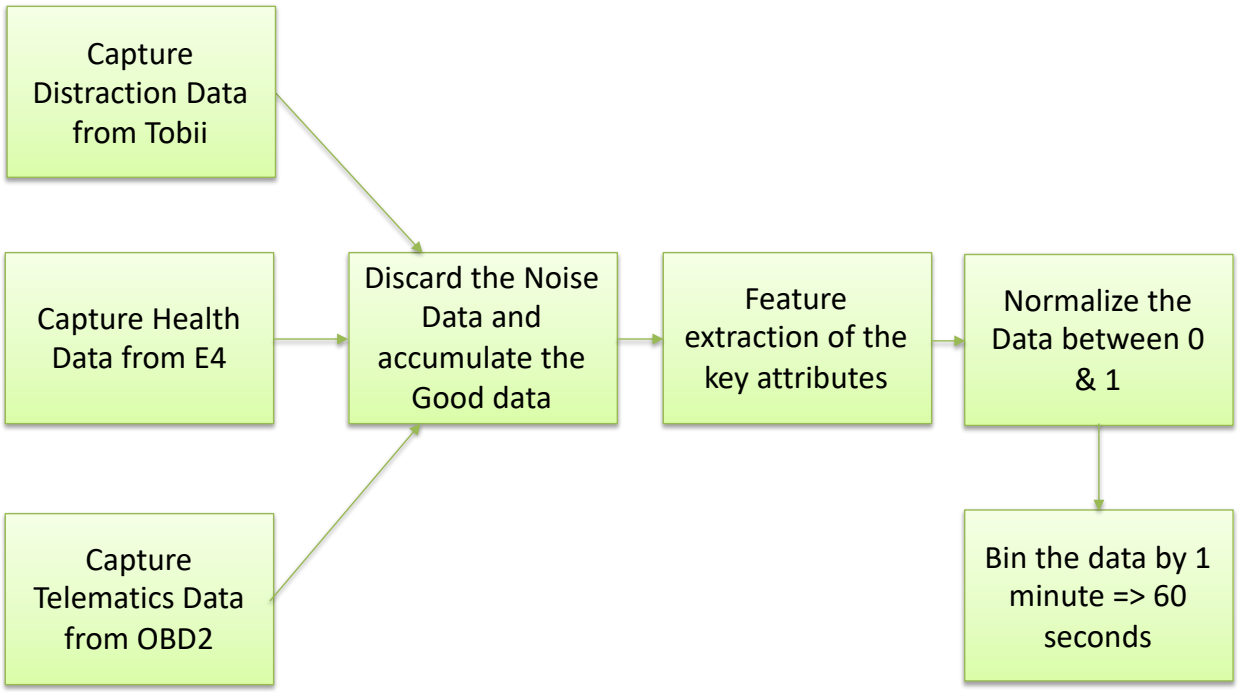


Figure 20: Data Reduction Strategy

Following is the list of features extracted key attributes from all three data streams:

- Distraction data source – Tobii Gaze Data
    - Accelerometer – Acceleration measure
    - Gyroscope – Measure angular velocity
    - Gaze Direction - Eye Tracker (Attention)
    - Pupil Position - Eye Tracker (Fixation)
    - Pupil Diameter - Eye Tracker (Fixation)

- Health Vital data source – E4 Data
    - HR - Heart Rate
    - EDA - Electrodermal activity sensor expressed as microsiemens (μS)
    - BVP - Data from photoplethysmography (Blood Volume Pulse)
- Telematics Data – OBD2 Data
    - Acceleration – Automobile acceleration
    - AirFuelRatio – Measure the oxygen content in the exhaust
    - Engine RPM – Engine Revolutions per Minute
    - Speed – Speed of a car

3.6.3 Baseline of Driving Data

We normalize the data attributes using Pandas and Scikit-Learn machine learning library and set the values from 0.0000 to 1.0000. Values close to mean (**μ**) are the centroid and they are considered normal. The values away from the center (**μ**) are considered anomalous as shown in figure 21.

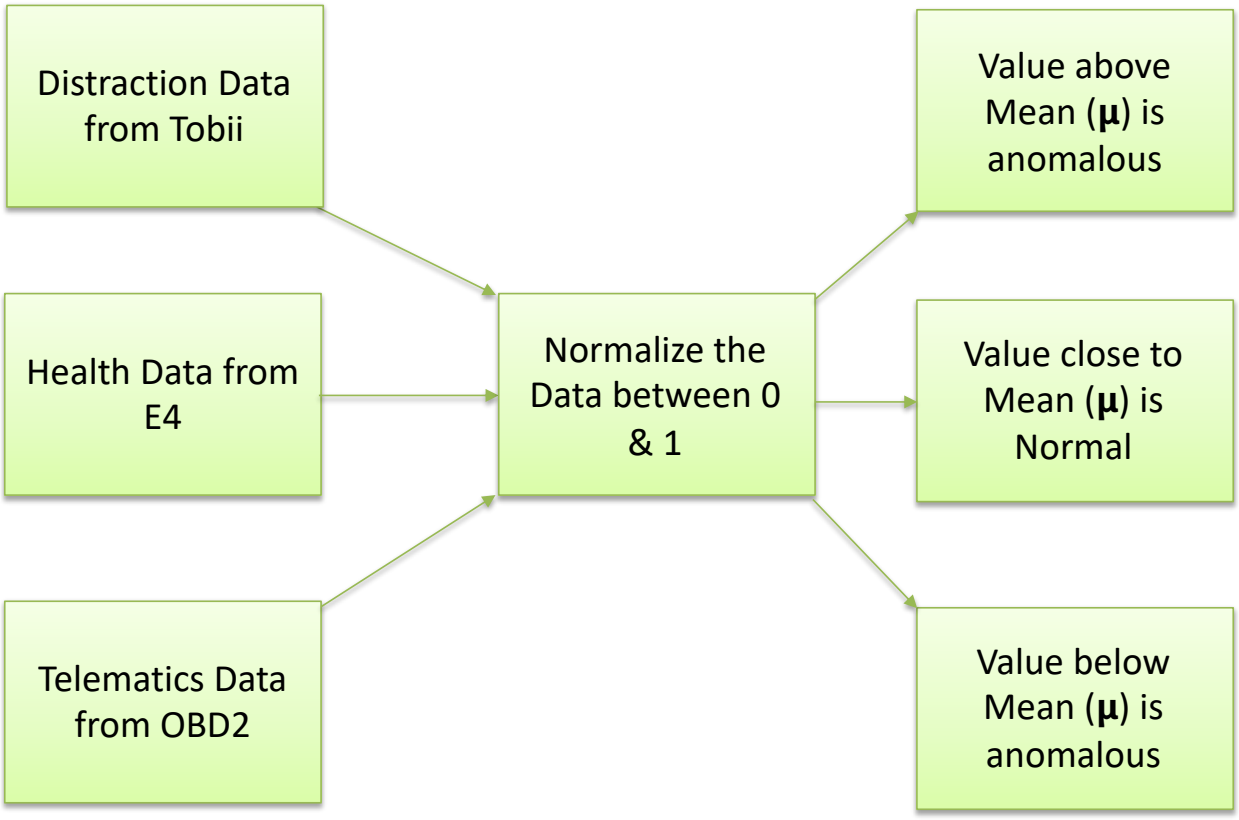


Figure 21: Normalization of Key Data Attributes

3.6.4 Anomaly Detection of the Driver Behavior

When we combine anomalies coming from different data streams, we can validate driver behavior more accurately. The anomaly can happen for various reasons. Following is a list of few common reasons for driver behavioral anomalies. High engine temperature, brake malfunction, reckless driving like abrupt acceleration, abrupt deceleration, making sharp turns and any other engine malfunction for cyber-attacks may adversely impact the safety for the driver. Likewise, running high or low blood pressure, having major hypertension while driving may also impact the driving behavior. Lastly, a distracted driver who is talking or texting on the phone while driving could be very damaging to the safety of the driver and the neighboring cars on the road.

- Telematics
    - Engine malfunction
    - Cyber attack
    - Reckless Driving
- Vital Health
    - Health problem
    - Tensed Mind
- Eye Gaze
    - Distraction with texting
    - Talking on the phone
    - Adjusting the music
    - Setting up the navigation
    - Eating while driving

Among these three data streams, the scale of anomaly is not the same for all. For example, a very small deviation from focusing on the road could be devastating if not paying attention on the road. On the contrary, driver could be more tolerant to moderate deviation from the normal pulse rate, heartbeat and/or blood pressure. Moreover, high speed may or may not be as bad if the surrounding traffic is very less. However, if the brake malfunctions in the rainy and icy road, that could be catastrophic. Following list depicts the scale of anomaly for different data streams.

- Distraction data source – Tobii Gaze Data
    - Scale of Anomaly is very small & impactful
    - Little deviation from the norm could be devastated
    - Very short period of anomaly is also devastating

- Health Vital data source – E4 Data
    - Scale of Anomaly is moderate
    - Little deviation is not catastrophic
- Telematics Data - OBD2 Data
    - Scale of Anomaly is moderate to small
    - Little deviation may not be catastrophic, but could be risky

Based on our observation and research, tolerance level (scale) of anomaly on Telematics is much higher than Health anomaly which is much higher than distraction anomaly. In other words, very minimal scale of anomaly on Gaze data could be dangerous than both health and telematics data as depicted in the following equations. Therefore, the intensity of distraction is much higher than that of Health anomaly and Telematics anomaly.

- *Tolerance of (Anomalous Telematics > Anomalous Health > Anomalous Gaze)*
- *Tolerance of* $(A_T > A_H > A_G)$
- *Impact on (Change of Anomaly for Gaze > Change of Anomaly for Health > Change of Anomaly for Telematics)*
- *Impact on* $(\partial_G > \partial_H > \partial_T)$

We first perform anomaly detection on just Telematics data stream by itself and identify the driver behavior as safe vs. unsafe using clustering and association rule mining. In order to bring in a more holistic approach for studying driver behavior with other contributing factors, we add two more key data attributes – Distraction data stream and Vital Health data stream. We study all three heterogeneous data streams and find the relationships between the three data attributes and detect anomalies for driver behavior. We study the driver behavior with OBD2 connected to the automobile, Tobii eye glass and E4 vital health monitoring watch connected to the driver's eye and

wrist respectively. We record the driving data and identify the anomalies on a video. We apply the real numbers on our derived formulas to identify the anomalies, the weighted intensity. We then identify the anomalous time span and study them closely by analyzing the signal (raw) data and compare the deviations on the Matrix Profile data streams and match the subsequent distance on a time series. We also use K-means clustering on the data streams to identify safe vs. unsafe driving clusters. We then analyze and validate our rules with machine learning techniques on heterogenous data streams using Association Rule Mining with Classification based Association. We calculate the mean as the norm and label anomalous data that are outside of +-1 standard deviation. The scale of anomaly is so low and impactful that even 1 degree of standard deviation gives good result of anomaly even though it is rather simplistic. We do not have to set two or three standard deviation here as the scale of anomaly for distraction data is low and therefore outside of only 1 standard deviation would give us anomalous driver behavior that includes both two or three standard deviation. We then label data based on the anomalies for each domain as shown in figure 22. We use J48 pruned tree as well as the decision tree to identify the path to anomalous driving states. We run and compare the Random Forrest and ROC curve to validate the precision of our results. Finally, we apply our Segmented Overlap and Collective Anomaly models on multi-stream heterogenous data and validate our results.

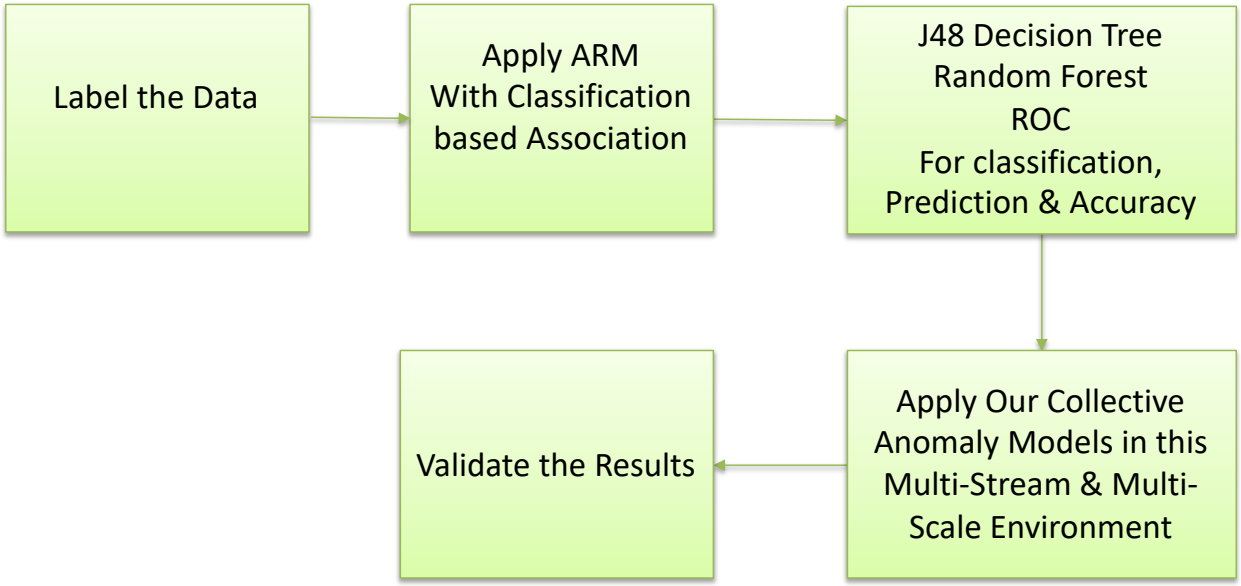


Figure 22: Anomaly Detection on Multi Stream Data Sets

CHAPTER 4

EXPERIMENTAL RESULTS

We will discuss the nature of the experiments, the datasets, the results and the Analysis & Observations in this section. Like the methodology section, we are showing the telematics experimental results by itself first and then we will show the combined experiments and the results from both Gaze data, Health Data and Telematics data. In our experimental study with different subjects (drivers) and their vehicles, we utilize the vehicles that the driver is used to and is their own, not with an unknown and unfamiliar automobile. Otherwise, we would introduce new variables on the drivers' driving patterns with the vehicles that are new to them. Consequently, the mode of Human Machine Teaming (HMT) between the driver and the vehicle is important here.

4.1 Driver Behavior from Telematics Data Stream

We have collected data on 3 different vehicles with 3 different drivers. Initially we collect only Telematics data from OBD2 device and from our custom application for derived data attributes like Sharp Turn, Abrupt Acceleration, Abrupt Deceleration, etc. We show data analytics and anomaly detection on the first subject based on more datasets and more realistic scenario like on the main road along with couple of highway rides as oppose to the other subjects being on the local roads and parking lots for safety.

In general, the data collection process is little challenging as the subjects are not usually comfortable being recorded, wearing different data collection devices and above all driving on the road while we sit next to the subject monitoring his/her driving behavior. Couple subjects deny

continuing with the runs as they start to feel uncomfortable. We somehow manage to collect good amount of data from one subject that we are showing here in Table 5.

Table 5: Data Collection Details for Telematics

Subject	Age	Gender	Vehicle	# of Excursions	Time	# of Readings	Data Stream
Driver 1	42	Female	Toyota Corolla	4	4 PM – 6 PM	5,000 (9 Runs: 7 to 9 min each run)	Telematics & Custom App
Driver 2	22	Male	BMW 528i	2	4 PM – 6 PM	4,000 (7 Runs: 5 to 8 min each run)	Telematics & Custom App
Driver 3	56	Male	Honda Odyssey	2	4 PM – 6 PM	4,500 (9 Runs: 8 to 9 min each run)	Telematics & Custom App

We highlight on 1 vehicle and its baseline here as follows:

Vehicle: Toyota Corolla

Subject: 42 year old female

We collect data for a few days on similar time (between 4 PM & 6 PM) and append data in one dataset. Then we preprocess the dataset using Feature Selection methods. After that, we cluster the data set on Telematics data and driver behavior data. We picked 4 clusters as the most efficient number from the following elbow outcome as shown in figure 23. We then cluster the dataset with

4 clusters and create a baseline. We categorize the baseline according to the safety measures of the telematics attributes and driver behavior attributes. We finally compare the subject with this baseline for further comparison and analysis. We perform clustering on individual drivers' data related to his/her own driver behavior. Therefore, the clusters on one driver may vary from the clusters on others.

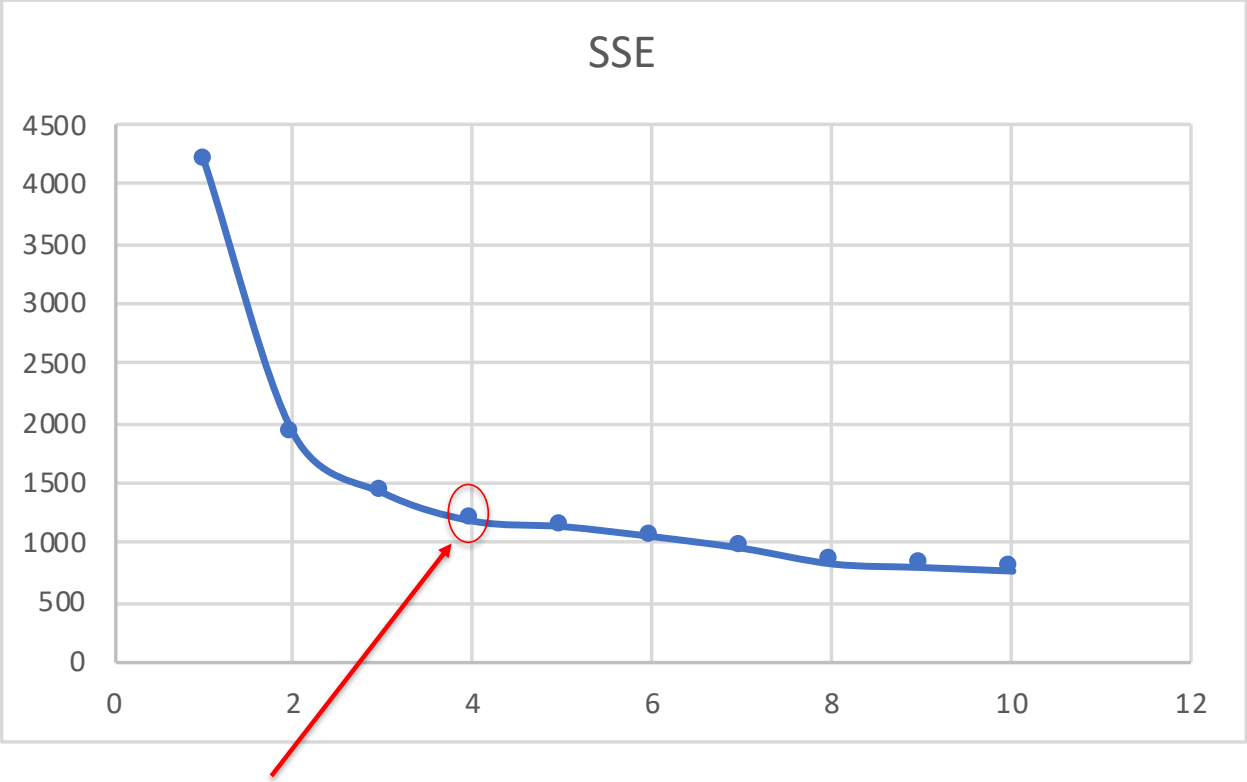


Figure 23: Elbow method to determine the optimal number of clusters(4) for k-means clustering

Following are the 4 clusters from the vehicle data:

We take the feature extracted Telematics data and run K means clustering with 4 clusters based on the above Elbow method. We analyze each cluster and locate the unsafe situations in each cluster. For instance, in the first cluster in Table 6, we see high emission of carbon monoxide (CO), high intake air temperature, low volumetric efficiency, high frequency of sharp turns and sudden acceleration. Therefore, this cluster (cluster 0) is labeled as UNSAFE cluster. On the contrary, the

second cluster (cluster 1) has no anomalous situation and hence this cluster is the SAFEST cluster. Third cluster (cluster 3) has high percentage of engine load as well as high speed. Lastly, the 4th cluster (cluster 3) has little high average trip speed. Therefore, cluster 2 & 3 have been labeled as SAFER and SAFE clusters respectively. Consequently, we have labeled the clusters according to their relative anomalous states.

Table 6: Clusters from Vehicle Data (Time of the day: between 4 PM & 6 PM)

Attributes	Full Data (3217)	Cluster 0 (716)	Cluster 1 (536)	Cluster 2 (994)	Cluster 3 (971)	Comment
Safety Categorization	*Overall*	*UNSAFE*	*SAFEST*	*SAFER*	*SAFE*	
Acceleration Sensor(Total)(g)	-0.0014	0	-0.0027	-0.0038	0.0006	
Air Fuel Ratio(Measured)(:1)	14.9527	14.7175	15.0472	14.8252	15.2044	
CO_ in g/km (Average)(g/km)	262.1087	319.812	294.1554	245.0225	219.3601	High emmision in cluster 0
Average trip speed(whilst moving only)(mph)	37.7857	13.194	31.7344	46.804	50.0275	Trip speed is high in cluser 3
Distance to empty (Estimated)(miles)	98.3274	110.6275	105.3646	96.6266	87.1139	
Engine Coolant Temperature(F)	193.5647	195.0961	192.4284	193.674	192.9508	
Engine Load(%)	34.0744	27.1301	38.7013	41.3587	29.184	High engine load in cluster 2
Engine RPM(rpm)	1840.6632	1109.2685	2146.6852	2315.1491	1725.3303	
Fuel flow rate/minute(gal/min)	0.0202	0.0115	0.0231	0.0288	0.0161	
Fuel Remaining (Calculated from vehicle profile)(%)	21.7179	24.5172	23.3882	21.3146	19.1447	
Fuel used (trip)(gal)	0.4984	0.0504	0.2312	0.5629	0.9103	
Intake Air Temperature(F)	101.0379	118.2668	96.7664	95.4111	96.4515	High temperature in cluster 0
Kilometers Per Litre(Long Term Average)(kpl)	14.4527	14.3978	14.3766	14.4662	14.5214	
Percentage of City driving(%)	43.0016	53.2712	51.0463	36.6352	37.5053	
Percentage of Highway driving(%)	32.9031	0	17.9611	46.6256	51.3661	
Speed (OBD)(mph)	43.6867	11.6151	57.4801	59.889	43.1357	High speed in cluster 2
Trip average MPG(mpg)	31.3372	25.7464	29.3618	33.4055	34.4329	
Voltage (OBD Adapter)(V)	12.6947	12.5162	12.7886	12.7728	12.6943	
Volumetric Efficiency (Calculated)(%)	43.0454	38.0922	44.8078	49.5563	39.0597	Low efficiency in cluster 0
Sharp Turns	0.0103	0.0377	0.0056	0	0.0031	High frequency in cluster 0
Sudden Acceleration	0.0096	0.0321	0	0.003	0.0051	High frequency in cluster 0

We also applied Association Rules on this dataset to derive the correlation between the attributes and identify the key rules to predict driver activities.

4.1.1 Results based on solely Telematics data stream

This section depicts the final outcome of our research for Driver Behavior based on Telematics data only.

We baseline Vehicle data and categorize the clusters by Safe, Safer, Safest and Unsafe types.

Cluster 0: This cluster is categorized as "UNSAFE" because of high CO emission, high air temperature, low volumetric efficiency, high frequency of sharp turns and high frequency of

sudden accelerations. All these parameters are way outside of the normal threshold and hence this cluster in UNSAFE.

Cluster 1: This cluster is categorized as "SAFEST" because of high efficiency driving patterns and very low to no incident of adverse driver behavior as well as the engine condition.

Cluster 2: This cluster is categorized as "SAFER" because of the majority of the attributes being in the safe range except occasional speeding and high engine load because of the speed.

Cluster 3: This cluster is categorized as "SAFE" because of most of the attributes being in the efficient range with the exception of occasional speeding as well as sharp turns and abrupt acceleration.

We compare this baseline of SAFE vs. UNSAFE clusters of data attributes with our subject (driver and the similar car – Toyota) for another instance of driving behavior and we notice the following anomalous behavior in table 7:

Table 7: New Instance of Driver & Engine Behavior

Attributes	Full Data (1938)	Cluster 0 (1088)	Cluster 1 (147)	Cluster 2 (414)	Cluster 3 (289)	Comment
Safety Categorization	***Overall***	***UNSAFE***	***UNSAFE***	***UNSAFE***	***UNSAFE***	***ANOMALY***
CO_ in g/km (Average)(g/km)	243.5672	260.3344	70.0431	253.3363	254.7126	High emmision in cluster 0
Engine Coolant Temperature(F)	211.3373	213.5893	182.8844	210.2362	218.9094	Very high temperature in every cluster
Sharp Turns	0.2389	0.2243	0.3946	0.2802	0.1557	High frequency in every cluster
Sudden Acceleration	0.1615	0	0.1633	0	1	High frequency in cluster 3

Regarding this subject (driver and the automobile), we notice unusually high CO emission, unusually high Engine coolant temperature, several sharp turns as well as many sudden accelerations. This is very anomalous to the same driver's baseline clusters. Therefore, there could be a potential cyber threat taken place on the automobile along with the driver's unusual driving patters. We also run association rule mining that discovers the following important rules shown in Table 9:

Sample converted dataset:

We take the raw data from vehicles and classify them with three different buckets with appropriate ranges. For classification rules, we need to convert the numeric data to categorical data so we can properly categorize the driving state in the right bucket. We classify the data attribute range in three equal segments to classify them by Low, Medium and High. For instance, Engine Coolant Temperature fluctuates between 185 to 200 degrees. Therefore, we divide the data attributes in 3 equal buckets as shown in table 8:

Table 8:Numeric to Categorical Data Mapping

Engine Coolant Temperature(F) (Numeric)	Engine Coolant Temperature(F) (Categorical)
185	ECT 185 to 189.9
190.4	ECT 190 to 194.9
195.8	ECT 195 to 200

We have classified the remaining data attributes the same way. We sort each data attribute from low to high and put them on Histogram to identify the proper range for categorical data.

4.1.2 Analysis & Validation

We run the categorical data on Apriori based association rule mining using Weka to identify the best suited rules with its associated attributes. For instance, following is a list of few key rules. These rules assist us identifying anomalies in driving behavior based on the Telematics data stream. This helps us understand the human behavior and their motive that may cause anomalous behavior. This cluster based single stream anomaly detection discovers computational models for human behavior from usage data to detect anomalous behavior. It identifies and predicts anomalies which might be precursors or even indicators of impending or ongoing unexpected behavior.

Table 9:Few Key Rules as Sample

Rules	*Confidence*	*Result*
Acceleration Sensor(Total)(g)=AS -0.07 to 0.2 ==> Sharp Turns=NO Sudden Acceleration =NO	0.97	**Low** Acceleration Sensor implies **no** sharp turn and **no** sudden acceleration
Average trip speed(mph)=ATS 34 to 51 Percentage of City driving(%)=PCD 33 to 65.9 ==> Sudden Acceleration =NO 1940	1.00	**Low** average trip speed with a **low** city driving implies **no** abrupt acceleration
Air Fuel Ratio=AFR 16.1 to 19.0 Average trip speed (mph)=ATS 34 to 51 Engine Coolant Temperature (F)=ECT 195 to 200.0 Percentage of Highway driving(%)=PHD 0 to 32.9 Speed (OBD)(mph)=Speed 50 to 75 ==> Sudden Acceleration =YES	0.93	**High** Air Fuel with relatively **high** speed with **high** Engine Coolant temperature would definitely imply Sudden Acceleration (**Yes**)
Average trip speed(mph)=ATS 34 to 51 Engine Coolant Temperature(F)=ECT 190 to 194.9 Percentage of Highway driving(%)=PHD 0 to 32.9 203 ==> Sudden Acceleration =YES	0.93	**High** trip speed with moderately **high** Engine temperature would imply sudden acceleration (**Yes**)

Air Fuel Ratio=AFR 16.1 to 19.0 **CO_ in g/km (Average)(g/km)=CO 245 to 285.9** **Percentage of Highway driving(%)=PHD 0 to 32.9 ==>** **Sudden Acceleration =YES**	0.92	**High** Air Fuel with a very **high** CO emission imply abrupt acceleration (**Yes**)
Air Fuel Ratio=AFR 16.1 to 19.0 **CO_ in g/km (Average)(g/km)=CO 245 to 285.9** **Percentage of Highway driving(%)=PHD 0 to 32.9 ==>** **Sharp Turns =YES**	0.92	**High** Air Fuel with a very **high** CO emission imply Sharp Turns (**Yes**)

The above rules that are anomalous (high measures) from normal driver behavior fall under the "Unsafe" cluster as discovered from the baseline cluster in Table 6. Likewise, the safe driving rules show low number of intensity which rightfully fall under the "Safe" cluster of the baseline.

4.2 Driver Behavior from Multi Source Data Stream

We have collected data on 2 different vehicles with 2 different drivers. We connect OBD2 device for Telematics data, custom mobile app on Android connected to the OBD2 by Bluetooth for derived Telematics data, Tobii for Eye Gaze Distraction data and E4 for vital Health data with the subject drivers. It was little intimidating to the subject drivers to be driving with all 4 data retrieval gadgets around them. Since both subjects' driving patterns are similar and the roads used for data collection are identical, we compile both data streams into one and apply anomaly detection on them to present results for general population. Following table 10 shows the two subjects we used to collects real world driving data.

Table 10: Data Collection Details for Telematics, Distraction and Health

Subject	Age	Gender	Vehicle	# of Excursions	Time	# of Readings	Data Stream
Driver 1	47	Male	BMW 528i	3	6 PM – 8 PM	5,000 (9 Runs: 7 to 9 min each run)	Telematics & Custom App, Distraction Data and Health Data
Driver 2	38	Male	Toyota Corolla	3	6 PM – 8 PM	4,000 (7 Runs: 5 to 7 min each run)	Telematics & Custom App, Distraction Data and Health Data

We combined the data streams from both drivers and the baselines are here as follows:

Vehicle: BMW 528i

Subject: 45 year old male

Vehicle: Toyota Corolla

Subject: 38 year old male

We collect data for a few days on similar time (between 6 PM & 8 PM) and append data in one dataset. We normalize the combined data stream and we preprocess the dataset using Feature Selection methods. After that, we look at the dataset and the video recording to see which pockets of time series the driver is distracted at. From the video, it is pretty apparent that the driver is

looking up and down, left and right and is pre-occupied with other things during the following time frames. From the observation in the data for Health and Telematics, we notice the following variations for the given distraction as shown in Table 11.

Table 11: Anomaly Comparison

Time	Eye Gaze	Health	Telematics
2:04 – 2:30	Looking down at stereo, adjusting the navigation	Medium-High BVP, High EDA	Low Speed, medium Acceleration
4:38 – 5:00	Looking down at the phone, texting, looking at the text	BVP – Mid-high EDA - High HR – Mid-High	High Speed, mid acceleration
6:17 – 6:30	Looking around, not focused on the wheel and the road	BVP – Mid-high EDA - High HR – High	High Speed, mid acceleration
9:05 – 9:30	Highly distracted, talking, looking to the passenger seat	BVP – High EDA - High HR – Mid	Mid-high Speed, Mid acceleration

We graph the signal data from the three data streams and perform matrix profile on the data stream to visualize the subsequent changes in the value as a result of the distraction.

4.2.1 Results based on multi-scale datastream

Following is a comprehensive result set for multi-scale data streams. We compare the signal/sensor data streams from all three data sources with the Matrix Profile (MP) data streams for specific time series and compare the anomalous driver behavior. We also compare the attributes of one data stream to the attributes of another data stream to locate the similarities and the dependencies for a specific state of the driver.

We use matrix profile to find the gradual change of the attributes over time. We first look at the driving record (signal data) and identify the timespan when we notice the anomalies. We run

matrix profile on the same dataset and identify the change of the driving behavior over time. Matrix profile helps us identifying the gradual shift of the driving pattern for gaze data, health data as well as the telematics data stream. We locate the subsequent changes over time and find the respective relationships between distraction, vital health and telematics data.

4.2.1.1 Anomaly Comparison with Gaze Data

We first graph the Eye Gaze (ACCZ) signal data and then we draw the graph for Matrix Profile data for the same time series. Then we identify the distracted timespan on the signal data to validate the distraction. Likewise, we notice the drastic changes in the subsequent dataset on matrix profile (MP) data for the same timespan that validate the distraction anomalies. For instance, the subject was looking down at the stereo and adjusting the navigation between 2 minute 4 second and 2 minute 30 second. On the signal graph, we notice high and low values of gaze data and a major change in the MP value on that duration. Similar anomalies are observed from 4 min 38 sec to 5 min, from 6 min 17 sec to 6 min 30 sec and 9 min5 sec to 9 min 30 sec as shown in figure 24 below.

Time	Eye Gaze
2:04 – 2:30	Looking down at stereo, adjusting the navigation
4:38 – 5:00	Looking down at the phone, texting, looking at the text
6:17 – 6:30	Looking around, not focused on the wheel and the road
9:05 – 9:30	Highly distracted, talking, looking to the passenger seat

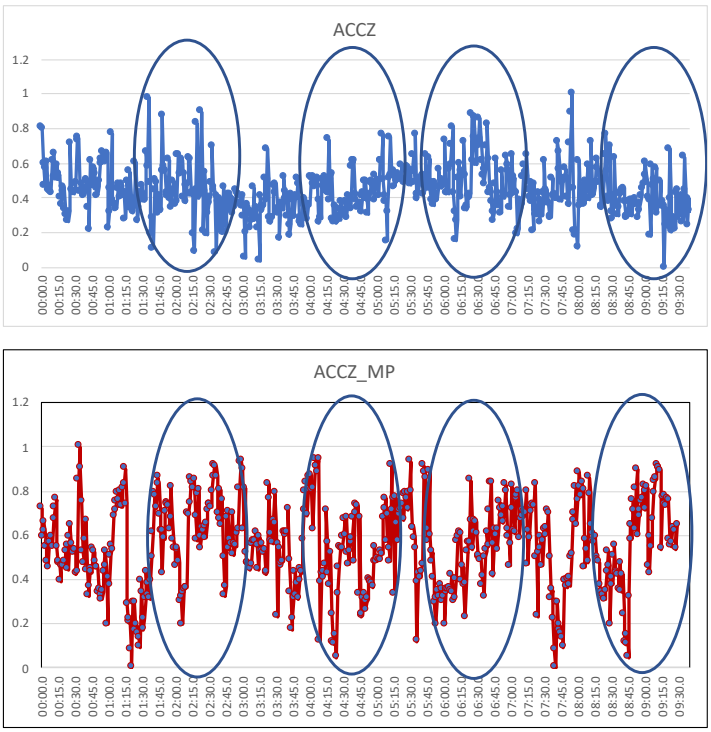


Figure 24: Anomaly Comparison for Gaze Data

Visually, signal data gives us some indication of anomalies, however, it is rather difficult and complex to derive specific anomalous behavior with so much data. Matrix Profile (MP) helps us comparing the subsequent distance for different attributes on a time series. We compute matrix profile on a slide window of a time series to locate the discords (long distance) which are anomalies. As we see the similarities on anomalous time span between the signal and MP data sets from the observation above, we like to validate the anomalies statistically by calculating the mean to identify the norm and the standard deviation to identify the anomalies on the MP data. We perform (+/-)1 level standard deviation based outlier detection on the matrix profile (MP) data and identify the anomalous timeseries that are less than (mean - 1 Std Dev) and greater than (mean + 1 Std Dev). We map that back to the actual distracted time frames and validate. These pockets of anomalies on the MP data for Eye Gaze shows in figure 25 below. For example, we see a drastic

change in distraction (ACCZ) value from 9 min 8 sec to 9 min 21 sec because of the driver's high distraction, which is also reflected on the graph 24.

Similar result is prevalent for other gaze data attributes like PupilDiameter.

Time	Eye Gaze
2:04 – 2:30	Looking down at stereo, adjusting the navigation
4:38 – 5:00	Looking down at the phone, texting, looking at the text
6:17 – 6:30	Looking around, not focused on the wheel and the road
9:05 – 9:30	Highly distracted, talking, looking to the passenger seat

Time	ACCZ	ACCZ - 1 SD
02:04.0	0.3033	Anomaly-L
02:05.0	0.3185	Anomaly-L
02:06.0	0.1945	Anomaly-L
02:07.0	0.343	Anomaly-L
02:08.0	0.3502	Anomaly-L
02:09.0	0.3616	Normal
02:10.0	0.3628	Normal
02:11.0	0.6972	Normal
02:12.0	0.6934	Normal
02:13.0	0.7458	Normal
02:14.0	0.8592	Anomaly-H
02:15.0	0.8352	Anomaly-H
02:16.0	0.812	Anomaly-H
02:17.0	0.8503	Anomaly-H
02:18.0	0.6846	Normal
02:19.0	0.5771	Normal
02:20.0	0.7829	Anomaly-H
02:21.0	0.8062	Anomaly-H

Mean	0.55248737
Std Dev	0.1973742
1-Lower	0.35511317
1-Upper	0.74986157

Time	ACCZ	ACCZ - 1 SD
09:05.0	0.5903	Normal
09:06.0	0.5472	Normal
09:07.0	0.6706	Normal
09:08.0	0.7917	Anomaly-H
09:09.0	0.8446	Anomaly-H
09:10.0	0.8563	Anomaly-H
09:11.0	0.9177	Anomaly-H
09:12.0	0.9023	Anomaly-H
09:13.0	0.9043	Anomaly-H
09:14.0	0.8917	Anomaly-H
09:15.0	0.5372	Normal
09:16.0	0.7689	Anomaly-H
09:17.0	0.737	Normal
09:18.0	0.7794	Anomaly-H
09:19.0	0.7326	Normal
09:20.0	0.7621	Anomaly-H
09:21.0	0.7562	Anomaly-H

Figure 25: Anomaly outside of 1 STD DEV

4.2.1.2 Anomaly Comparison on Health Data

Next, we graph the Health (BVD) signal data and then we draw the graph for Matrix Profile data for the same time series. Then we identify the anomalous BVD data timespan on the signal data to validate the irregular state of health. Likewise, we notice the drastic changes in the subsequent dataset on matrix profile (MP) data for the same timespan that validate the health anomalies. For instance, the subject's blood pressure was mid to low between 9 minute 5 second and 9 minute 30 second. On the signal graph, we notice mid and low values of BVD (Blood pressure) data and a major change in the MP value on that duration. Similar anomalies are observed from 2 min 4 sec

to 2 min 30 sec, from 4 min 38 sec to 5 min and from 6 min 17 sec to 6 min 30 sec as shown in figure 26.

Time	Health
2:04 – 2:30	Medium-High BVP, High EDA
4:38 – 5:00	BVP – Mid-high EDA - High HR – Mid-High
6:17 – 6:30	BVP – Mid-high EDA - High HR – High
9:05 – 9:30	BVP – Mid-Low EDA - High HR – Mid

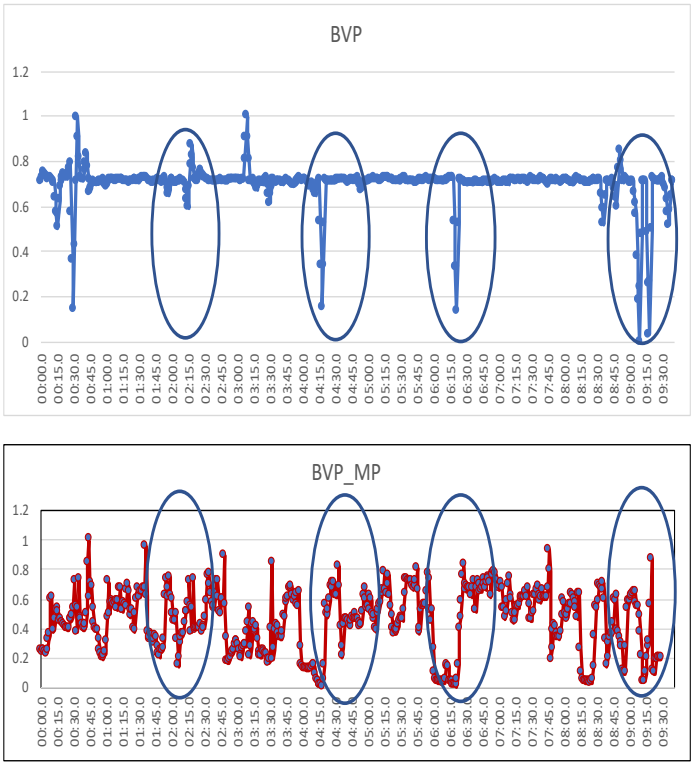


Figure 26: Anomaly Comparison for Gaze Data

Now similarly we perform (+/-)1 standard deviation based anomaly detection on the matrix profile (MP) data and identify the anomalous timeseries that are less than (mean - 1 Std Dev) and greater than (mean + 1 Std Dev). We map that back to the actual health (BVD) related anomalous time frames and validate. These pockets of anomalies on the MP data on BVD shows in figure 27. For instance, between 9 min 11 sec and 9 min 29 sec, we see a drastic change in the BVD measure as the blood pressure went down significantly that is related to the anomalous distraction.

Similar result is prevalent for other Health data attributes like EDA and HR.

Time	Health
2:04 – 2:30	Medium-High BVP, High EDA
4:38 – 5:00	BVP – Mid-high EDA - High HR – Mid-High
6:17 – 6:30	BVP – Mid-high EDA - High HR – High
9:05 – 9:30	BVP – Mid-Low EDA - High HR – Mid

Time	BVP	BVP - 1 SD
06:17.0	0.0339	Anomaly-L
06:18.0	0.0116	Anomaly-L
06:19.0	0.0107	Anomaly-L
06:20.0	0.011	Anomaly-L
06:21.0	0.0064	Anomaly-L
06:22.0	0.0499	Anomaly-L
06:23.0	0.1522	Anomaly-L
06:24.0	0.3963	Normal
06:25.0	0.4659	Normal
06:26.0	0.5869	Normal
06:27.0	0.7538	Anomaly-H
06:28.0	0.8345	Anomaly-H
06:29.0	0.6895	Anomaly-H
06:30.0	0.6548	Normal
06:31.0	0.6783	Anomaly-H
06:32.0	0.6676	Anomaly-H
06:33.0	0.6713	Anomaly-H
06:34.0	0.6709	Anomaly-H
06:35.0	0.6184	Normal
06:36.0	0.6689	Anomaly-H
06:37.0	0.7135	Anomaly-H

Time	BVP	BVP - 1 SD
09:11.0	0.2089	Anomaly-L
09:12.0	0.0397	Anomaly-L
09:13.0	0.0358	Anomaly-L
09:14.0	0.0372	Anomaly-L
09:15.0	0.0969	Anomaly-L
09:16.0	0.1976	Anomaly-L
09:17.0	0.2744	Normal
09:18.0	0.3038	Normal
09:19.0	0.5545	Normal
09:20.0	0.8708	Anomaly-H
09:21.0	0.1092	Anomaly-L
09:22.0	0.0957	Anomaly-L
09:23.0	0.0935	Anomaly-L
09:24.0	0.187	Anomaly-L
09:25.0	0.1893	Anomaly-L
09:26.0	0.1932	Anomaly-L
09:27.0	0.1933	Anomaly-L
09:28.0	0.195	Anomaly-L
09:29.0	0.2009	Anomaly-L

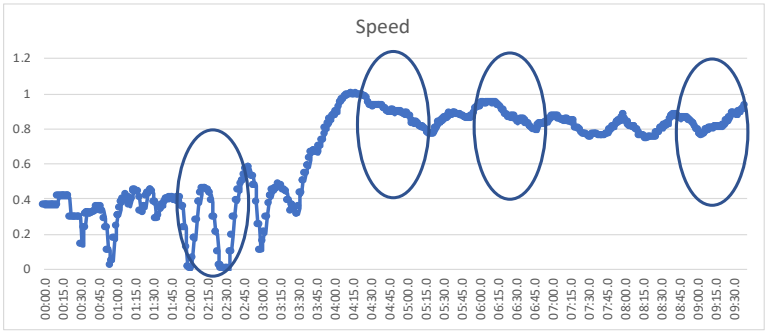
Mean	0.44758807
Std Dev	0.2099233
1-Lower	0.23766477
1-Upper	0.65751137

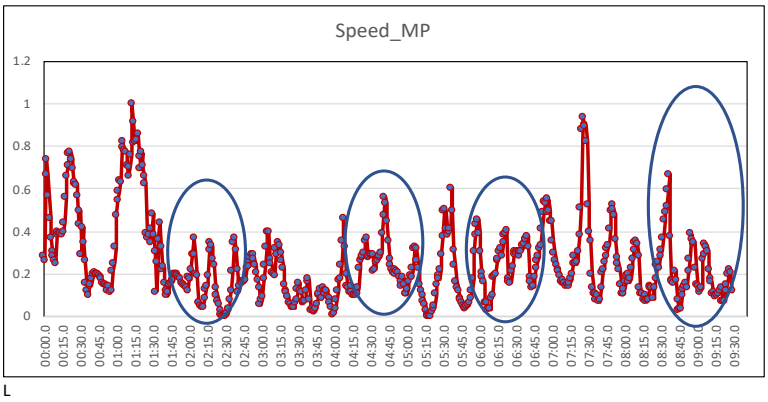
Figure 27: Anomaly outside of 1 STD DEV

4.2.1.3 Anomaly Comparison on Telematics Data

Next we analyze the Telematics (Speed) signal data by plotting them on a two dimensional graph and then draw the same for Matrix Profile data for the same time series as shown in figure 28. Then we identify the high-speed timespan on the signal data to validate the telematics anomalies. Likewise, we notice the drastic changes in the subsequent dataset on matrix profile (MP) data for the same timespan that validate the speed anomalies. For instance, the subject was driving with a high speed between 4 minute 38 second and 5 minutes. On the signal graph, we notice high values of Telematics (Speed) data and a major change in the MP value on that duration. Similar anomalies are observed from 2 min 4 sec to 2 min 30 sec, from 6 min 17 sec to 6 min 30 sec and 9 min 5 sec and 9 min 30 sec.

Time	Telematics
2:04 – 2:30	Low Speed, medium Acceleration
4:38 – 5:00	High Speed, mid acceleration
6:17 – 6:30	High Speed, mid acceleration
9:05 – 9:30	Mid-high Speed, Mid acceleration





L

Figure 28: Anomaly Comparison for Gaze Data

Next we perform (+/-)1 standard deviation based anomaly detection on the matrix profile (MP) data and identify the anomalous timeseries that are less than (mean - 1 Std Dev) and greater than (mean + 1 Std Dev). We map that back to the actual Telematics (Speed) related anomalous time frames and validate. These pockets of anomalies on the MP data on Speed shows in figure 29 below. For instance, between 2 min 8 sec and 2 min 12 sec, we see a drastic change in the Speed as that is related to the anomalous Telematics state which could be related to the high level of distraction.

Similar result is prevalent for other Telematics data attributes like Acceleration and Air Fuel Ratio.

Time	Telematics
2:04 – 2:30	Low Speed, medium Acceleration
4:38 – 5:00	High Speed, mid acceleration
6:17 – 6:30	High Speed, mid acceleration
9:05 – 9:30	Mid-high Speed, Mid acceleration

Time	Speed	Speed - 1 SD
02:05.0	0.2929	Normal
02:06.0	0.1982	Normal
02:07.0	0.1904	Normal
02:08.0	0.0655	Anomaly-L
02:09.0	0.065	Anomaly-L
02:10.0	0.0528	Anomaly-L
02:11.0	0.0459	Anomaly-L
02:12.0	0.0454	Anomaly-L
02:13.0	0.0878	Normal
02:14.0	0.1297	Normal
02:15.0	0.1425	Normal
02:16.0	0.1901	Normal
02:17.0	0.3108	Normal
02:18.0	0.3464	Normal
02:19.0	0.3326	Normal
02:20.0	0.2728	Normal

Mean	0.2546
Std Dev	0.18759
1-Lower	0.06701
1-Upper	0.44219

Time	Speed	Speed - 1 SD
04:37.0	0.2757	Normal
04:38.0	0.2814	Normal
04:39.0	0.301	Normal
04:40.0	0.388	Normal
04:41.0	0.4779	Anomaly-H
04:42.0	0.563	Anomaly-H
04:43.0	0.5302	Anomaly-H
04:44.0	0.4436	Anomaly-H
04:45.0	0.3537	Normal
04:46.0	0.2673	Normal
04:47.0	0.2382	Normal
04:48.0	0.2203	Normal
04:49.0	0.2031	Normal
04:50.0	0.1986	Normal
04:51.0	0.2173	Normal

Figure 29: Anomaly outside of 1 STD DEV

4.2.1.4 Anomaly Comparison on Three Diferent Data Sreams

We compare the matrix profile data streams from Gaze data, Health data and Telematics data on a specific time scale to see the similar changes on the distance and hence the relationships among the three multi scale data streams. For instance, on a time scale between 9:05 and 9:30 in table 12, we notice a highly distracted state of a driver who has a high blood flow and relatively high speed. Similarly, we compare the subsequent changes on the matrix profile data sets for gaze data (Figure 30), Health data (Figure 31) and Telematics data (Figure 32) between the same time scale of 9:05 to 9:30 and we notice very similar trends among the three as derived above. We see high change of distraction on ACCZ MP data during that time along with high change of BVD MP data and mid to high change of speed. High change means the data fluctuates close to the lowest limit to the highest limit. Mid to high change in speed means the speed fluctuates from mid-range to

highest data point. Looking at the recorded video starting the 9th minute through 9 minute 30 second, we see the driver talking on the phone and looking at the passenger seat instead of looking on the road. The data reflects that same behavior as we see a significant subsequent distance change on ACCX distraction data during that time in figure 30. We see a very similar overlap on health data in figure 31 where the subsequent distance changes on health data. Lastly, the 3rd data stream from Telematics data shows minimal Acceleration change in figure 32 as the distraction and impactful health did not have a big impact on the telematics data.

Table 12: Anomaly Comparison among three data streams

Time	Eye Gaze	Health	Telematics
9:05 – 9:30	Highly distracted, talking, looking to the passenger seat	BVP – High EDA - High HR – Mid	Mid-high Speed, Mid Acceleration

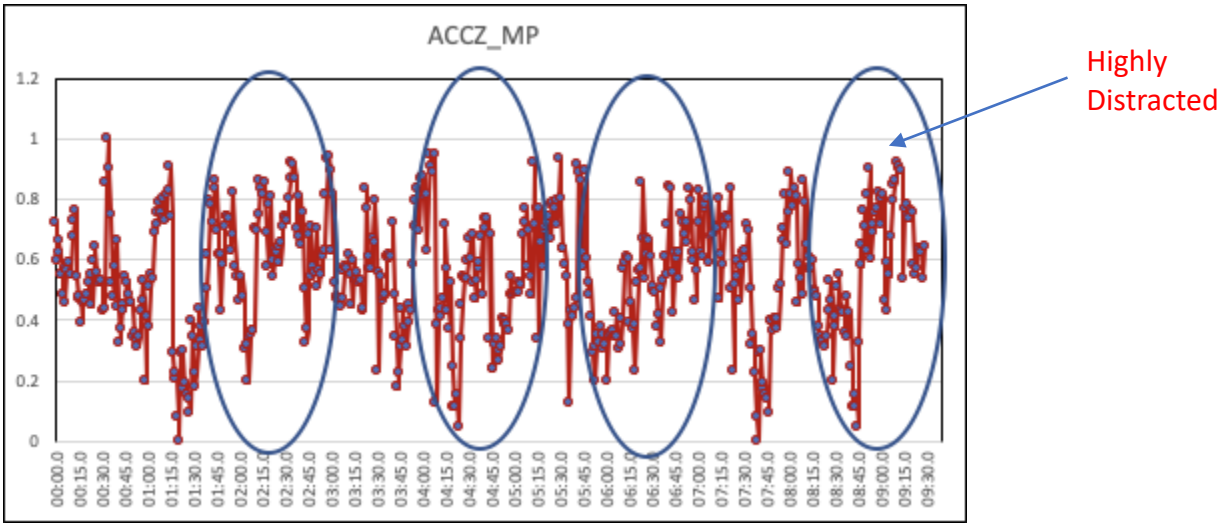


Figure 30: Anomaly for Gaze Data

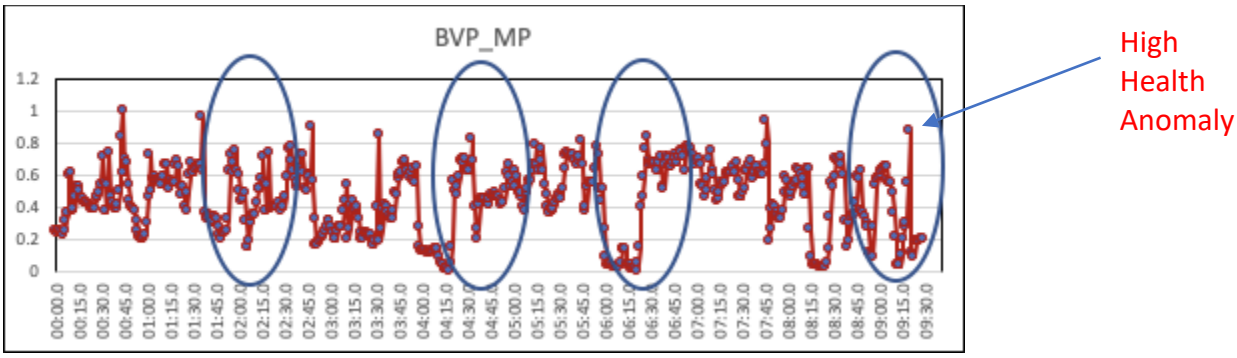


Figure 31:Anomaly for Health Data

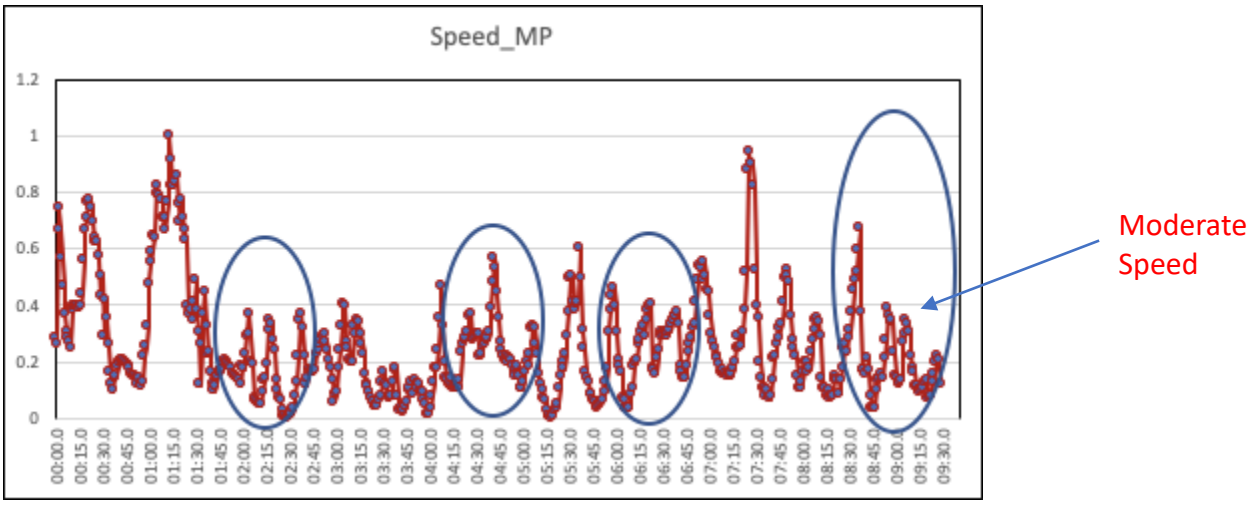


Figure 32: Anomaly for Telematics Data

4.2.1.5 Clusters on Hetergenious Data Stream

We perform K-means clustering on three combined data streams to identify the "Unsafe" driving state from the "Safe" driving state. Unfortunately, the result shows very similar clusters and hence it is rather difficult to distinguish Safe driver cluster to Unsafe driver cluster unlike the clusters identified in just Telematics data shown earlier.

For instance, as shown in Table 13 & 14, the safe driving conditions usually center around the mid points (mean 0.5) and the unsafe driving conditions move away from the center. Among the 4

clusters, the number of anomalous attributes are very similar. Therefore, clustering on heterogenous data streams is not a good identifier for safe vs. unsafe driving condition. Relatively speaking, the clusters 1 & 3 are slightly safer than clusters 2 & 4.

Table 13: Clusters from Distraction, E4 & Telematics - Signal Data

Attributes	Full Data (581)	Cluster 1 (338)	Cluster 2 (85)	Cluster 3 (79)	Cluster 4 (79)2
ACCX	0.569511	0.547082	0.556056	0.565019	0.582967
ACCY	0.364915	0.379246	0.469763	0.384775	0
ACCZ	0.389225	0.431127	0.426679	0.25612	0.038952
GYRX	0.376564	0.376564	0.316837	0.337884	0.365188
GYRY	0.567408	0	0.175523	0.248322	0.405647
GYRZ	0.592333	0.083277	0	0.304727	0.415905
GazeDirectionLeftX	0.404098	0.404098	0	0.130994	0.059294
GazeDirectionLeftY	0.455748	0.455748	0.092083	0.129543	0.263466
GazeDirectionLeftZ	0.984243	0.987197	0.930345	0.947766	0.938367
GazeDirectionRightX	0.499914	0.499914	0.124601	0.316964	0.28132
GazeDirectionRightY	0.563498	0.563498	0.371474	0.014592	0
GazeDirectionRightZ	0.99574	0.975688	0.903038	0.916402	0.916402
GazePoint3DX	0.757604	0.757604	0.740785	0.720581	0.640343
GazePoint3DY	0.043078	0.043078	0.044435	0.033601	0.032421
GazePoint3DZ	0.006397	0.006539	0.005777	0.005269	0.005137
GazePointX	0.537313	0.537313	0.399057	0.507463	0.4674
GazePointY	0.311159	0.461373	0.47103	0.287554	0.374464
PupilPositionLeftX	0.505495	0.452119	0.430141	0.516484	0.365777
PupilPositionLeftY	0.55106	0.473988	0.472062	0.44894	0.539499
PupilPositionLeftZ	0.220026	0.220026	0.205534	0.221344	0.216074
PupilPositionRightX	0.532995	0.51269	0.475127	0.527919	0.44264
PupilPositionRightY	0.410372	0.420519	0.413754	0.435175	0.401353
PupilPositionRightZ	0.616346	0.598077	0.629808	0.592308	0.624038
PupilDiameterLeft	0.525424	0.502825	0.536723	0.525424	0.553672
PupilDiameterRight	0.475728	0.402913	0.451456	0.34466	0.470874
X	0.439655	0.439655	0.517241	0.439655	0.474138
Y	0.752809	0.730337	0.752809	0.719101	0.775281
Z	0.796748	0.747967	0.796748	0.796748	0.764228
BVP	0.710757	0.710757	0.713907	0.706348	0.713862
EDA	0.848739	0.848739	0.890758	0.823529	0.840338
HR	0.275893	0.275893	0.415816	0.326913	0.240051
AccelerationSensor	0.47619	0.47619	0.444444	0.531746	0.539683
AirFuelRatio	1	1	1	1	1
EngineRPM	0.241031	0.136287	0.241031	0.413436	0.01517
Speed	0.767926	0.794654	0.750108	0.866073	0.821526

Observations:

1. Rather difficult to identify safe driving cluster vs. unsafe driving cluster

2. Mid point close to 0.5 is the safe driving condition

3. Attributes away from 0.5 is Anomalous

- Cluster 1:
  - Anomalous attributes = 13
- Cluster 2:
  - Anomalous attributes = 15
- Cluster 3:
  - Anomalous attributes = 13
- Cluster 4:
  - Anomalous attributes = 15

- Safer clusters are 1 & 3

Table 14: Clusters from Distraction, E4 & Telematics – Signal Nominal Data

Attributes	Full Data (581)	Cluster 0 (192)	Cluster 1 (123)	Cluster 2 (155)	Cluster 3 (111)
ACCX	Mid	Mid (1)	Mid (1)	Mid (1)	Mid (1)
ACCY	Mid-Low	Mid-Low (1)	Mid-Low (1)	Mid (2)	Mid-Low (1)
ACCZ	Mid	Mid (2)	Mid (2)	Mid-Low (1)	Mid (2)
GYRX	Mid-Low	Mid-Low (2)	Mid-Low (2)	Mid (3)	Mid-Low (2)
GYRY	Mid	Mid (3)	Mid (3)	Mid (4)	Mid (3)
GYRZ	Mid	Mid-High (1)	Mid (4)	Mid (5)	Mid (4)
GazeDirectionLeftX	Mid	Mid (4)	Mid-High (1)	Mid (6)	Mid-Low (3)
GazeDirectionLeftY	Mid	Mid (5)	Mid (5)	Mid (7)	Mid-High (1)
GazeDirectionLeftZ	High	High (1)	High (1)	High (1)	High (1)
GazeDirectionRightX	Mid	Mid (6)	Mid-High (2)	Mid-High (1)	Mid (5)
GazeDirectionRightY	Mid-High	Mid-High (2)	Mid-High (3)	Mid-High (2)	Mid-High (2)
GazeDirectionRightZ	High	High (2)	High (2)	High (2)	High (2)
GazePoint3DX	Mid-High	Mid-High (3)	Mid-High (4)	Mid-High (3)	Mid-High (3)
GazePoint3DY	Low	Low (1)	Low (1)	Low (1)	Low (1)
GazePoint3DZ	Low	Low (2)	Low (2)	Low (2)	Low (2)
GazePointX	Mid	Mid (7)	Mid-Low (3)	Mid (8)	Mid-High (4)
GazePointY	Mid	Mid (8)	Mid (5)	Mid (9)	Mid-Low (4)
PupilPositionLeftX	Mid	Mid (9)	Mid (7)	Mid (10)	Mid-Low (5)
PupilPositionLeftY	Mid	Mid (10)	Mid (8)	Mid (11)	Mid (6)
PupilPositionLeftZ	Mid-Low	Mid-Low (3)	Mid-Low (4)	Mid-Low (2)	Mid-Low (6)
PupilPositionRightX	Mid	Mid (11)	Mid (9)	Mid (12)	Mid (7)
PupilPositionRightY	Mid	Mid (12)	Mid (10)	Mid (13)	Mid-Low (7)
PupilPositionRightZ	Mid-High	Mid-High (4)	Mid-High (5)	Mid-High (4)	Mid-High (5)
PupilDiameterLeft	Mid	Mid (13)	Mid (11)	Mid (14)	Mid (8)
PupilDiameterRight	Mid	Mid (14)	Mid (12)	Mid (15)	Mid (9)
X	Mid	Mid (15)	Mid (13)	Mid (16)	Mid (10)
Y	Mid-High	Mid-High (5)	Mid-High (6)	Mid-High (5)	Mid-High (6)
Z	Mid-High	Mid-High (6)	Mid-High (7)	Mid-High (6)	Mid-High (7)
BVP	Mid-High	Mid-High (7)	Mid-High (8)	Mid-High (7)	Mid-High (8)
EDA	High	High (3)	High (3)	High (3)	High (3)
HR	Mid-Low	Mid-Low (4)	Mid (14)	Mid-Low (3)	Mid-Low (8)
AccelerationSensor	Mid	Mid (16)	Mid (15)	Mid (17)	Mid (11)
AirFuelRatio	Low	Low (3)	Low (3)	Low (3)	Low (3)
EngineRPM	Mid-Low	Mid-Low (5)	Mid (16)	Mid-Low (4)	Mid-Low (9)
Speed	High	High (4)	High (4)	High (4)	High (4)

Observations:

1. Rather difficult to identify safe driving cluster vs. unsafe driving cluster

2. Mid point close to 0.5 is the safe driving condition

3. Attributes away from 0.5 is Anomalous

- Cluster 1:
  - Mid = 16, High = 4, Low = 3
- Cluster 2:
  - Mid = 16, High = 4, Low = 3
- Cluster 3:
  - Mid = 17, High = 4, Low = 3
- Cluster 4:
  - Mid = 11, High = 4, Low = 3

- Safer clusters are 1, 2 & 3

Contrary to the single stream clusters, the multidomain clusters do not work very well with multi-scale heterogenous data sets as the data attributes are so different from each domain.

4.2.2 Analysis & Validation

We validate our hypothesis and results by applying specific machine learning methodologies on our heterogeneous data streams as follows:

- Association Rule Mining (ARM) with Classification based Association
- ARM with Normalized Signal Data
- ARM on Signal Data with (+-)1 Standard Deviation as Anomaly
- ARM on Normalized Matrix Profile Data
- ARM on Matrix Profile Data with (+-)1 Standard Deviation as Anomaly

We analyze and validate the results from ARM with classification based analytics as follows:

- J48 to generate decision tree to depict key rules from heterogeneous data streams
- Decision Tree to validate the key rules
- Random Forrest for classification and prediction

We evaluate the classification accuracy with ROC Curve.

4.2.2.1 ARM on Normalized Signal Data

We run Association Rule Mining (ARM) with Classification based Association on the heterogeneous signal data stream to validate the relationships between the data attributes from three different data streams. For instance, as depicted in Table 15 and Figure 33, we validate the following rules:

- Distraction may imply Tension (Anomalous Blood Pressure) [Distraction to Health]
- Distraction may or may not have an impact on Acceleration [Distraction to Telematics]
- Distraction may impact Speed [Distraction to Telematics]

Table 15: ARM on Normalized Signal Data

Index	*Rules*	*Confidence*	*Result*
1	GazeDirectionLeftZ=High GazeDirectionRightZ=High GazePoint3DY=Low ==> BVP=Mid-High	0.92	High Distraction implies mid to high blood pressure
2	GazeDirectionLeftZ=High GazeDirectionRightZ=High GazePoint3DX=Mid-High GazePoint3DZ=Low ==> BVP=Mid-High	0.92	High Distraction implies mid to high blood pressure
3	GYRX=Mid-Low GazeDirectionLeftZ=High GazePointX=Mid Z=Mid-High ==> AccelerationSensor=Mid	0.90	Mid to high Distraction implies normal acceleration
4	GazeDirectionLeftZ=High GazeDirectionRightZ=High GazePoint3DX=Mid-High	0.98	High Distraction and high Heart Beat implies high Speed

	GazePoint3DY=Low X=Mid Y=Mid-High HR=High ==> Speed=High		
5	GazeDirectionLeftZ=High GazeDirectionRightZ=High GazePoint3DY=Low X=Mid Y=Mid-High HR=High ==> Speed=High	0.98	High Distraction and high Heart Beat implies high Speed

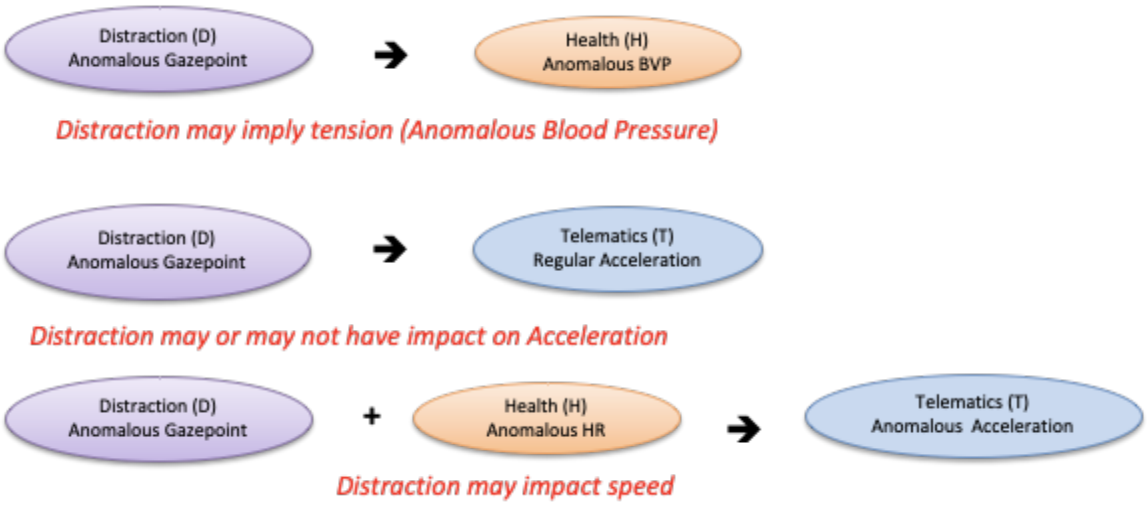


Figure 33: ARM on Normalized Signal Data

4.2.2.2 ARM on Normalized Signal Data with (+-)1 Standard Deviation as Anomaly

We run Association Rule Mining (ARM) with Classification based Association on the heterogeneous signal data stream with labeled anomalies that are one standard deviation away and validate the relationships between the data attributes from three different data streams. For instance, as depicted in Table 16 and Figure 34, we validate the following rules:

- Distraction may imply Tension (High Heartbeat) [Distraction to Health]
- Distraction may imply Tension (Abnormal Blood Pressure) [Distraction to Health]

Table 16: ARM on Normalized Signal Data with (+-)1 Standard Deviation

Index	*Rules*	*Confidence*	*Result*
1	PupilPositionRightZ-SD=Normal PupilDiameterLeft-SD=Anomaly-H HR-SD=Normal ==> Speed-SD=Anomaly-L	0.91	Distraction may imply Tension with impact on Telematics (High Heartbeat) [Distraction to Health]
2	ACCZ-SD=Anomaly-H EDA-SD=Normal EngineRPM-SD=Normal Speed-SD=Normal ==> HR-SD=Anomaly-H	0.91	Distraction may imply Tension (High Heartbeat) [Distraction to Health]
3	ACCZ-SD=Anomaly-H BVP-SD=Normal EDA-SD=Normal Speed-SD=Normal ==> HR-SD=Anomaly-H	0.91	Distraction may imply Tension (High Heartbeat) [Distraction to Health]
4	PupilDiameterLeft-SD=Anomaly-H ==> BVP-SD=Anomaly-L	0.91	Distraction may imply Tension (Abnormal Blood Pressure) [Distraction to Health]

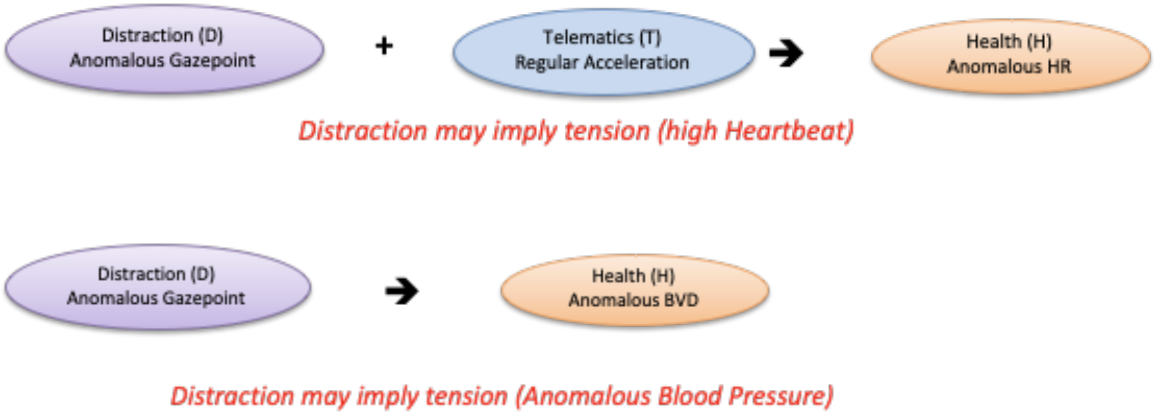


Figure 34: ARM on Signal Data with +- STD DEV

4.2.2.3 ARM on Normalized Matrix Profile (MP) Data

We run Association Rule Mining (ARM) with Classification based Association on the heterogeneous Matrix Profile (MP) data stream to validate the relationships between the data attributes from three different data streams. For instance, as depicted in Table 17 and Figure 35, we validate the following rules:

- Distraction may imply Tension (High Blood Pressure) [Distraction to Health]
- Distraction may imply Emotion (Anomalous EDA) [Distraction to Health]

Table 17: ARM on Normalized Matrix Profile (MP) Data

Index	*Rules*	*Confidence*	*Result*
1	GazePoint3DY=Low Z=Mid-High Speed=Mid ==> HR=Mid-Low	0.90	Distraction may imply Tension (High Blood Pressure) [Distraction to Health]
2	GazePoint3DZ=Low Z=Mid-High Speed=Mid ==> HR=Mid-Low	0.90	Distraction may imply Tension (High Blood Pressure) [Distraction to Health]
3	GazeDirectionLeftZ=High ==> BVP=Mid-High	0.92	Distraction may imply Tension (High Blood Pressure) [Distraction to Health]
4	GazeDirectionRightZ=High ==> BVP=Mid-High	0.92	Distraction may imply Tension (High Blood Pressure) [Distraction to Health]
5	GazeDirectionLeftZ=High GazePointY=Mid PupilPositionRightZ=Mid-High ==> EDA=High	0.91	Distraction may imply Emotion (Anomalous EDA) [Distraction to Health]

6	GazeDirectionRightZ=High GazePointY=Mid PupilPositionRightZ=Mid-High ==> EDA=High	0.91	Distraction may imply Emotion (Anomalous EDA) [Distraction to Health]
7	GazePoint3DY=Low GazePointY=Mid PupilPositionRightZ=Mid-High 179 ==> EDA=High	0.91	Distraction may imply Emotion (Anomalous EDA) [Distraction to Health]

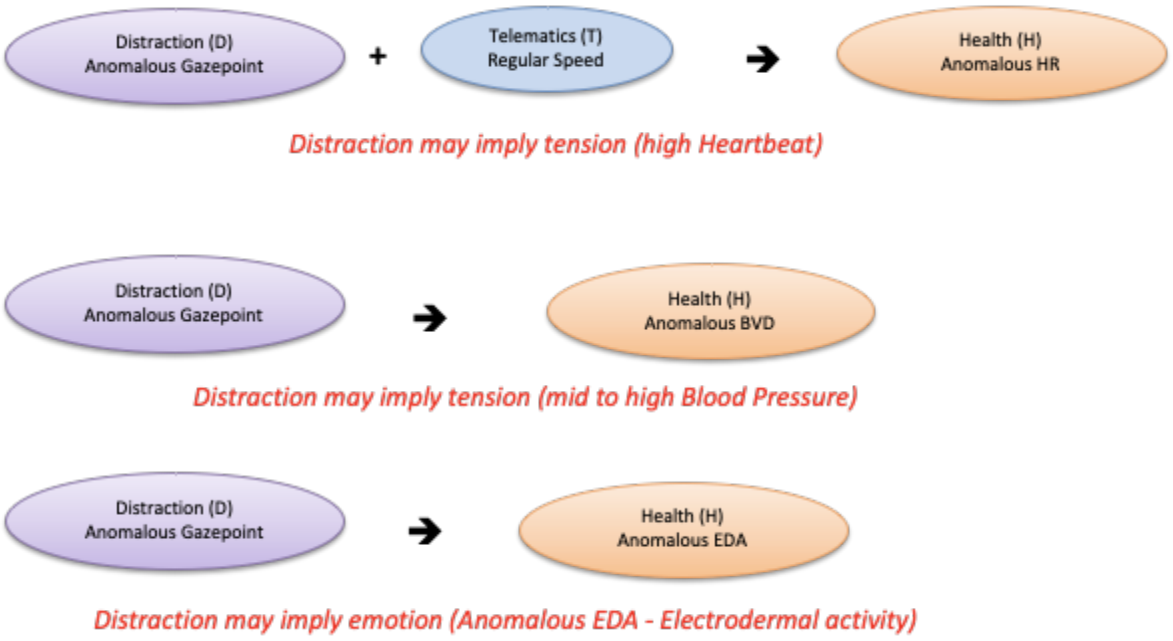


Figure 35: ARM on Normalized Matrix (MP) Profile Data

Likewise, we see the following rule on MP data on heterogeneous data stream as shown in Table 18 and Figure 36:

- Distraction & Tension may or may not have impact on Acceleration [Distraction & Health to Telematics]
- Distraction & Tension may impact Speed [Distraction & Health to Telematics]

Table 18: ARM on Normalized Matrix Profile (MP) Data

Index	*Rules*	*Confidence*	*Result*
1	GYRX=Mid-Low GazeDirectionLeftZ=High GazeDirectionRightZ=High GazePointX=Mid Z=Mid-High ==> AccelerationSensor=Mid	0.91	Distraction & Tension may or may not have impact on Acceleration [Distraction & Health to Telematics]
2	GYRX=Mid-Low GazeDirectionLeftZ=High GazePoint3DY=Low GazePointX=Mid Z=Mid-High ==> AccelerationSensor=Mid	0.90	Distraction & Tension may or may not have impact on Acceleration [Distraction & Health to Telematics]
3	GYRX=Mid-Low GazeDirectionLeftZ=High GazePoint3DZ=Low GazePointX=Mid Z=Mid-High ==> AccelerationSensor=Mid	0.90	Distraction & Tension may or may not have impact on Acceleration [Distraction & Health to Telematics]
4	GazeDirectionLeftZ=High GazePoint3DY=Low X=Mid Y=Mid-High HR=High ==> Speed=High	0.98	Distraction & Tension may impact Speed [Distraction & Health to Telematics]
5	GazeDirectionLeftZ=High GazePoint3DZ=Low X=Mid Y=Mid-High HR=High ==> Speed=High	0.98	Distraction & Tension may impact Speed [Distraction & Health to Telematics]
6	GazeDirectionRightZ=High GazePoint3DY=Low X=Mid Y=Mid-High HR=High ==> Speed=High	0.98	Distraction & Tension may impact Speed [Distraction & Health to Telematics]

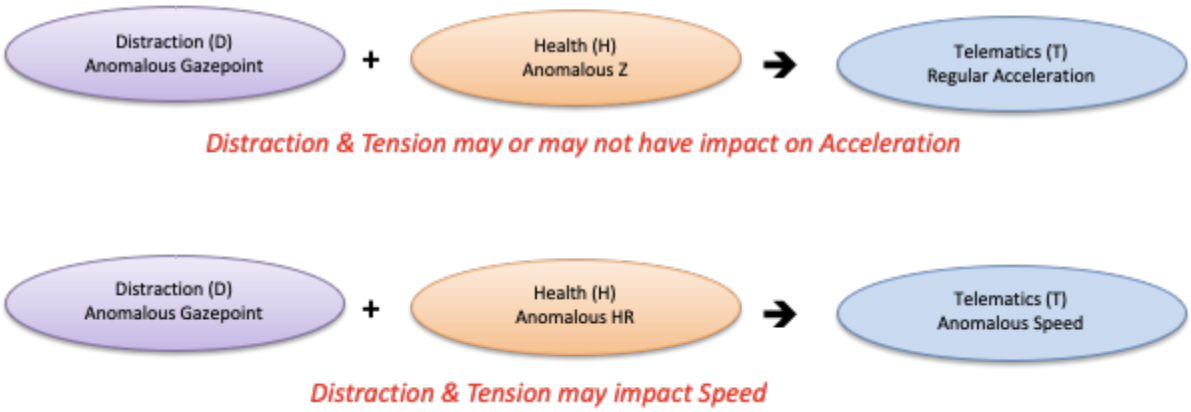


Figure 36: ARM on Normalized Matrix (MP) Profile Data

4.2.2.4 ARM on Matrix Profile (MP) Data with (+-)1 Standard Deviation as Anomaly

We run Association Rule Mining (ARM) with Classification based Association on the heterogeneous matric profile (MP) data stream with labeled anomalies that are one standard deviation away and validate the relationships between the data attributes from three different data streams. We use 1 standard deviation for labeling the data attributes 'anomalous' because of the small scale of anomaly mainly on distraction data. A small and short span of distraction while driving may create a dangerous situation as the driver may lose control and derail on the road. For instance, as depicted in Table 19 and Figure 37, we validate the following rules:

- Distraction may imply Tension (High Heartbeat) [Distraction to Health]
- Distraction may imply Tension (Anomalous Blood Pressure) [Distraction to Health]
- Distraction may imply Emotion (Anomalous EDA) [Distraction to Health]

Table 19: ARM on Matrix Profile (MP) Data with (+-)1 Standard Deviation

Index	*Rules*	*Confidence*	*Result*
1	PupilDiameterLeft=Anomaly-H ==> HR=Anomaly-H	1.0	Distraction may imply Tension (High Heartbeat) [Distraction to Health]
2	PupilDiameterLeft=Anomaly-H Speed=Normal ==> HR=Anomaly-H	1.0	Distraction may imply Tension (High Heartbeat) [Distraction to Health]
3	ACCZ=Anomaly-L GazeDirectionRightX=Anomaly-L ==> BVP=Anomaly-L	1.0	Distraction may imply Tension (Anomalous Blood Pressure) [Distraction to Health]
4	ACCZ=Anomaly-H HR=Anomaly-H ==> BVP=Anomaly-H	1.0	Distraction may imply Tension (Anomalous Blood Pressure) [Distraction to Health]
5	GazeDirectionRightX=Anomaly-L HR=Anomaly-H ==> BVP=Anomaly-L	1.0	Distraction may imply Tension (Anomalous Blood Pressure) [Distraction to Health]
6	ACCZ=Anomaly-L PupilDiameterLeft=Anomaly-H Speed=Normal ==> EDA=Anomaly-L	1.0	Distraction may imply Emotion (Anomalous EDA) [Distraction to Health]
7	ACCZ=Anomaly-L PupilDiameterLeft=Anomaly-H HR=Normal Speed=Normal ==> EDA=Anomaly-L	1.0	Distraction may imply Emotion (Anomalous EDA) [Distraction to Health]

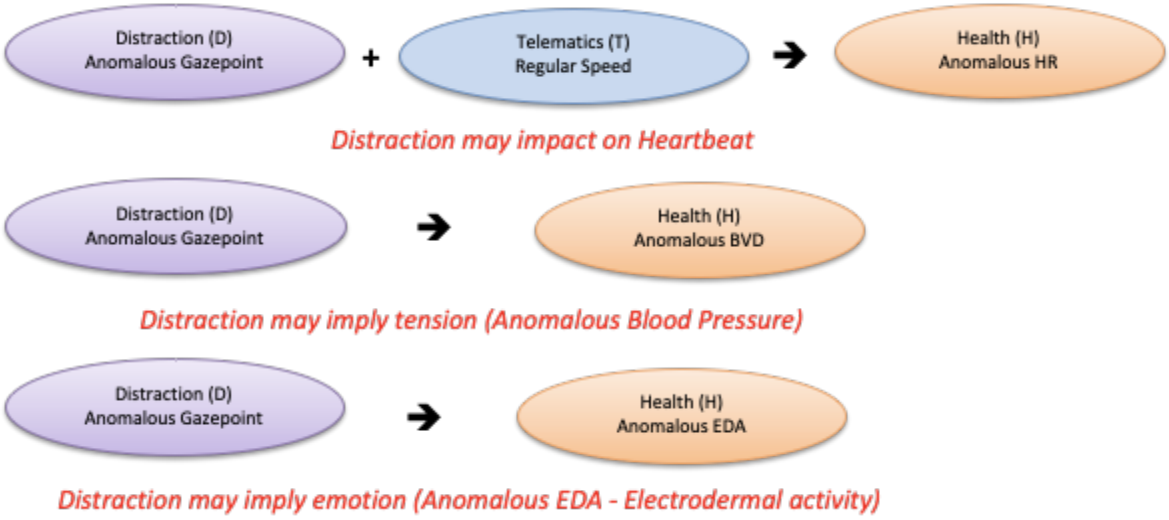


Figure 37: ARM on MP Data with +- STD DEV

Likewise, we validate the following rules as shown in Table 20 and Figure 38:

- Distraction & Tension may impact Acceleration [Distraction & Health to Telematics]
- Distraction may impact Speed [Distraction to Telematics]

Table 20: ARM on Matrix Profile (MP) Data with (+-)1 Standard Deviation

Index	*Rules*	*Confidence*	*Result*
1	PupilDiameterLeft=Anomaly-H EDA=Anomaly-L ==> AccelerationSensor=Anomaly-H	1.0	Distraction & Tension may impact Acceleration [Distraction & Health to Telematics]
2	ACCZ=Anomaly-L GazeDirectionLeftZ=Normal PupilDiameterLeft=Anomaly-H ==> AccelerationSensor=Anomaly-H	1.0	Distraction & Tension may impact Acceleration [Distraction & Health to Telematics]
3	PupilDiameterLeft=Anomaly-H HR=Normal ==> Speed=Anomaly-H	1.0	Distraction may impact Speed [Distraction to Telematics]

4	ACCZ=Anomaly-L GazePointY=Normal PupilDiameterLeft=Anomaly-H HR=Normal ==> Speed=Anomaly-H	1.0	Distraction may impact Speed [Distraction to Telematics]
5	GazePoint3DZ=Anomaly-L ==> Speed=Anomaly-H	1.0	Distraction may impact Speed [Distraction to Telematics]
6	GazePoint3DZ=Anomaly-L PupilDiameterLeft=Anomaly-H ==> Speed=Anomaly-H	1.0	Distraction may impact Speed [Distraction to Telematics]

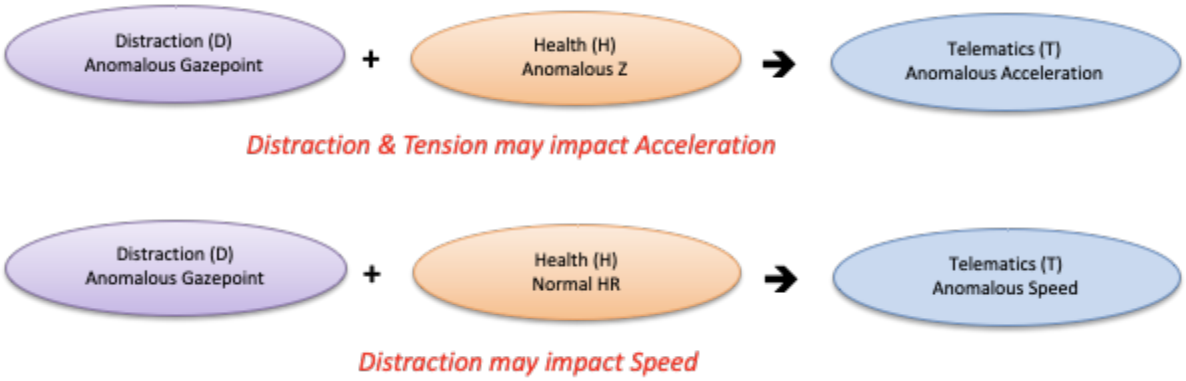


Figure 38: ARM on MP Data with +- 1 STD DEV

So far we have run Association Rule Mining (ARM) with Classification based Association to identify and validate the relationships among the key attributes from multi-scale data streams. For example, we identify how distraction can impact health and how distraction can impact speed of the vehicle. From the subject, we notice there is not much of an impact between health and telematics without distraction. We validate the same findings from normalized signal data, normalized Matrix Profile data and also with the data that are 1 (one) standard deviation away from the mean.

4.2.2.5 Classification on labelled Data

We use the same multi scale data stream to build a tree of rules that identifies anomalies on the driving behavior. We first take the normalized datasets and label the data attributes by Normal, High (anomalous) and Low (anomalous) based on the close proximity to the mean (centroid), 1 standard deviation higher and 1 standard deviation lower, respectively. Because of the high impact data attributes on the distraction data domain, one standard deviation away from the main is anomalous and labeled as A (Anomalous) and data close to the mean is labeled as NA (Not Anomalous) We validate the labelling of the heterogeneous signal data of NA (Not Anomalous) and A (Anomalous) based on the actual distraction. We identified the pockets of distractions on a time series by looking at the video recording of the subjects (drivers). For instance, the driver was preoccupied with other things like looking around, looking at the passenger seat, tuning the radio, looking at the phone, etc. during 2 min 4 seconds through 2 minute 31 seconds and again got distracted during 4 min 38 seconds through 5 minutes, from 6 minute 17 seconds to 6 minutes 31 seconds and lastly from 9 min 05 seconds to 9 minute 29 seconds. We perform the same annotated labelling to the matrix profile data. As the matrix profile data gives us the change in the heterogenous data sets and shows a clear indication of how gradual the changes are happening over time, we have better results with MP data than signal (raw) data. We label the datasets based on distraction than health and telematics because distraction is found to be more extensive and dominant than the other two datasets as discovered in the earlier experiments shown in the above sections.

We run J48 based decision trees on this heterogeneous labeled matrix profile data that validates the same results as we have derived from ARM. For example, we see the following results from the decision tree with high level of accuracy as shown in Figure 39 & 40:

Decision Tree shows that when Eye Gaze is high, and BVD (blood flow) is normal, AirfuelRatio and EDA could be both low (anomalous). Basically, distracted drivers may or may not show anomalous health situation, however, may show telematics anomaly.

High Eye Gaze Data -> Normal BVD -> Low AirFuelRatio -> Low EDA

Likewise, when Eye Gaze is high, BVD (blood flow) is normal and AirFuelRatio is normal, the speed could be high (anomalous).

High Eye Gaze Data -> Normal BVD -> Normal AirFuelRatio -> High Speed

Similarly, when Eye Gaze is high, BVD (blood flow) is normal and AirFuelRatio is normal, the speed could be Low (anomalous).

High Eye Gaze Data -> Normal BVD -> Normal AirFuelRatio -> Low Speed

Lastly, when Eye Gaze is high, BVD (blood flow) is normal, AirFuelRatio is normal, and the speed is normal, the EAD could be High (anomalous).

High Eye Gaze Data -> Normal BVD -> Normal AirFuelRatio -> Normal Speed -> High EAD

We can have numerous rules of different attributes by these decision trees. Looking at the rules from the decision tree, it is apparent that distraction is the most prevalent attribute for detecting anomalous driving condition. Figure 39 shows a J48 Pruned tree that classifies the rules with 85% accuracy. From the Confusion Matrix, we see the Precision (Positive Predictive Value) of 73%, which is a good predictor. It is rather difficult to simulate the real-world scenario with the subjects that are anomalous with noticeable distractions, reckless driving behavior and anomalous health impacts which could be very unsafe situation on the road. Considering the limitations, our result shows a good validation on anomalous driving behavior.

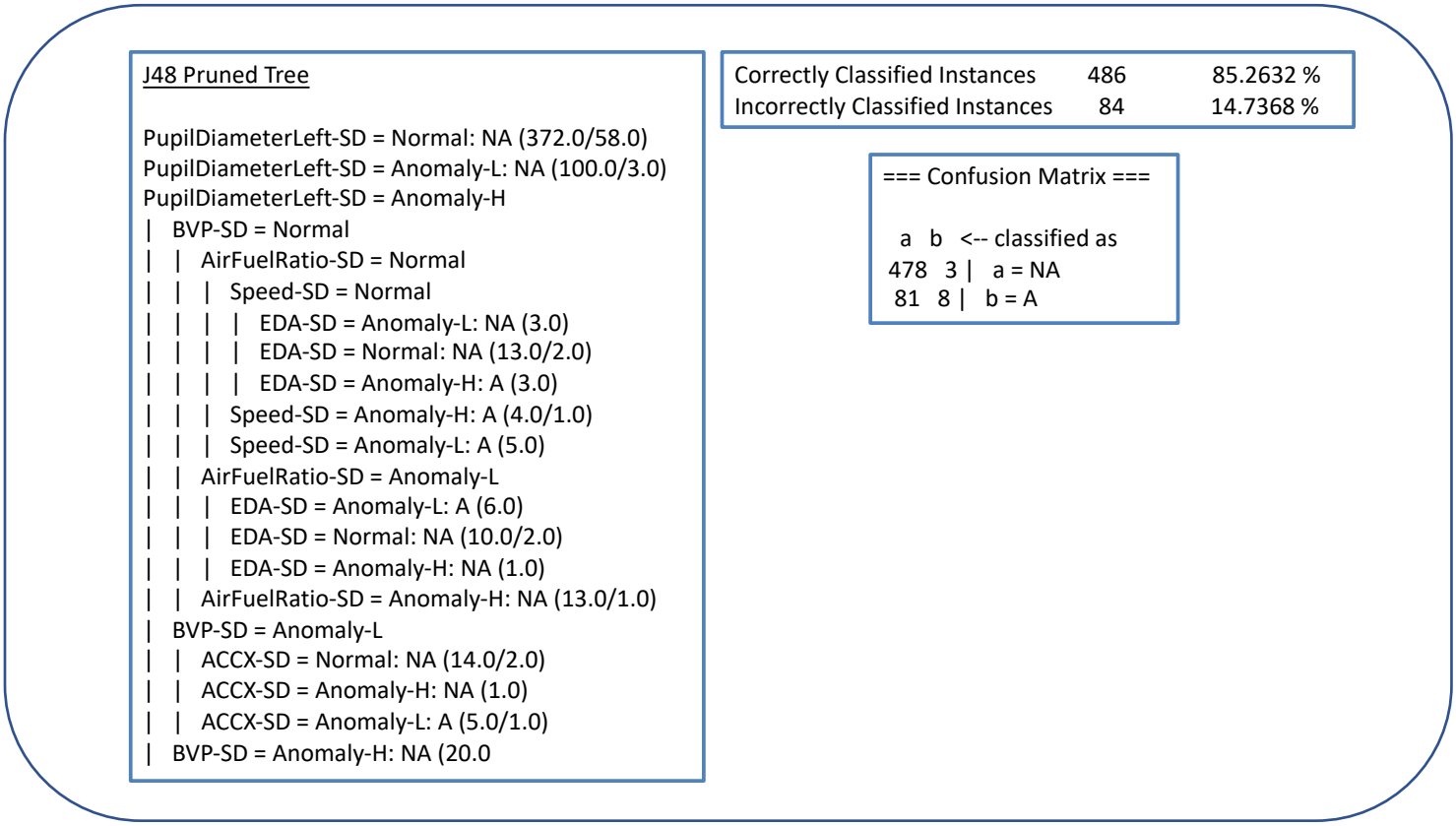


Figure 39: J48 Pruned Tree on MP Data

Following Decision Tree in Figure 40 shows the paths to different anomalous scenarios. We have many other paths corresponds to both anomalous and normal states. We show 5 key anomalous scenarios from our subjects.

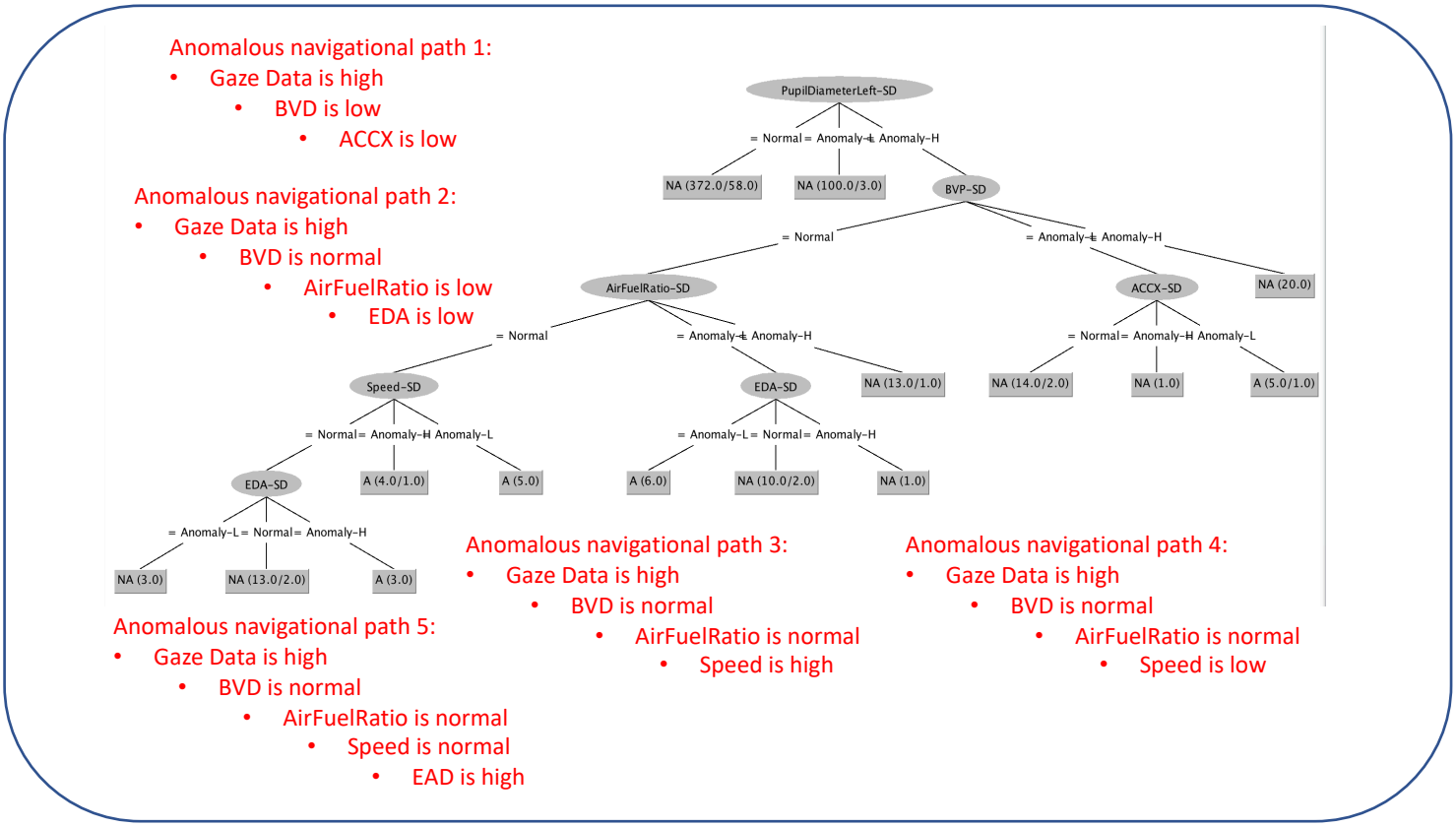


Figure 40: Decision Tree on MP Data Stream

4.2.2.6 Random Forrest and ROC Curve on labelled Data

Finally, we run Random Forrest to get the most precise decision trees and the area under the curve from ROC (Receiver Operating Characteristic) to validate the precision of the result. As mentioned earlier, the result from the matrix profile data is actually more accurate than signal data since MP shows a concise gradual change on a subsequent time based data stream. For instance, for Telematics attribute Speed, we notice the precision for Normal, Low Anomalous values and High Anomalous values are high 78%, 75% and 82% respectively as shown in figure 41. The closer the curve gets to the Y axis, the better the result is. Hence, these curves show high precision of our results are. In other words, the higher the area under the curve is, the more precise our results are,

for example, our labeled data of normal and anomalous speed are showing relatively a high level of precision as the area under the curve is high.

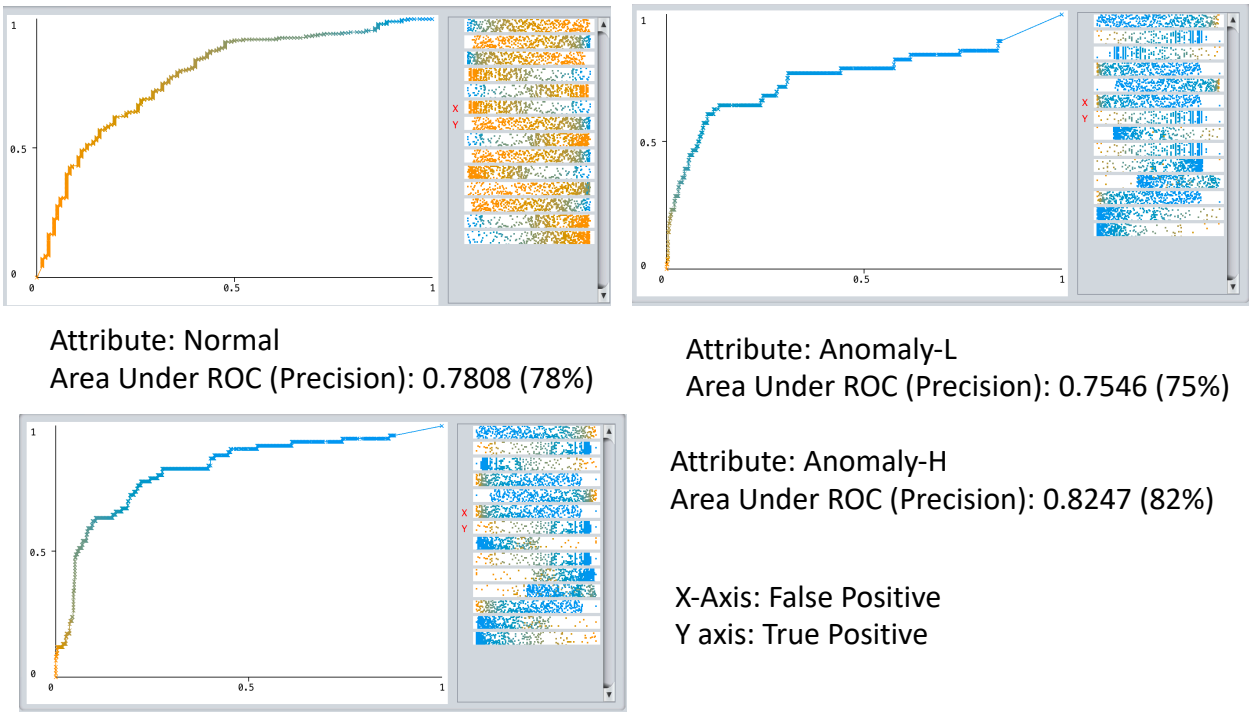


Figure 41: Random Forrest & ROC Curve for Speed (Signal Data)

Similarly, the following ROC curve in figure 42 depicts that anomalous data attributes on MP data that are very close to the Y axis and hence shows high level of precision with 83% True Positive.

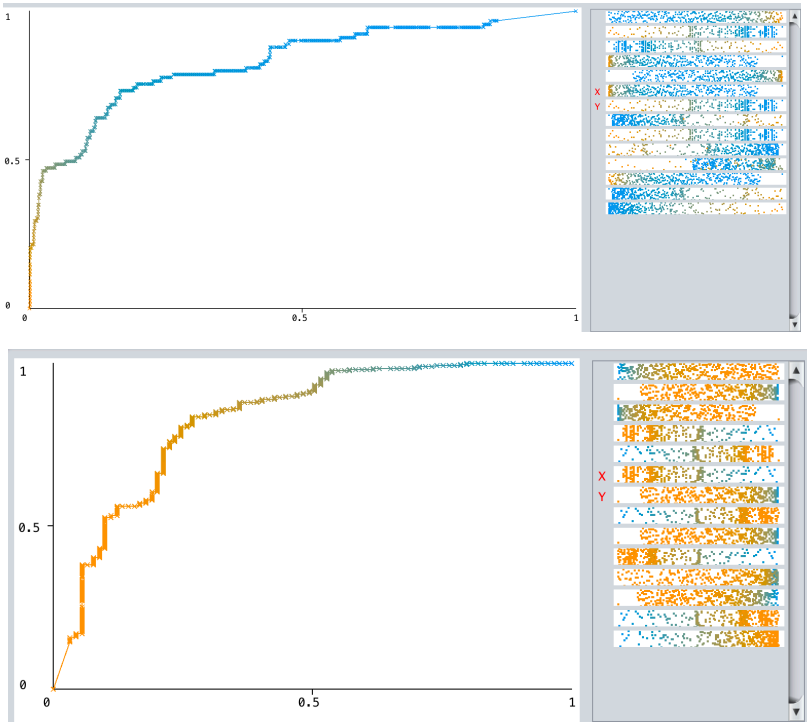
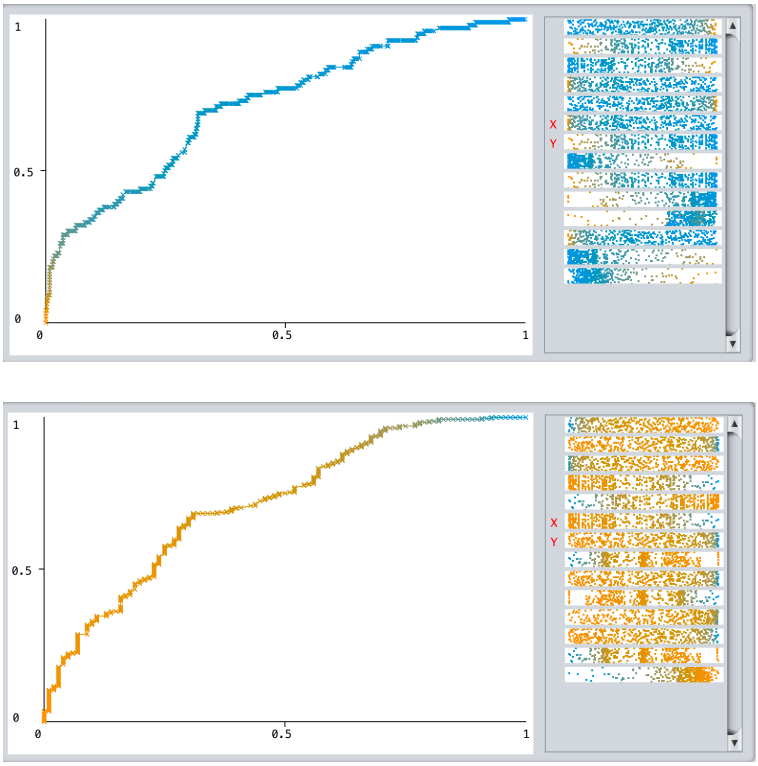


Figure 42: Random Forrest & ROC Curve for Annotated Label on Speed MP Data

Following ROC curves show the precision for Signal (Raw) data from all three heterogeneous data streams. The result is showing less precise than that of matrix profile data. Here the precision for True Positive is 72% for signal data as oppose to the matrix profile data where the True Positivity is 83% as shown in figure 43.



Attribute: Anomaly
Area Under ROC (Precision):
0.7181 (72%)

MP data gave us more
accurate result (83%)

Attribute: Non-Anomaly
Area Under ROC (Precision):
0.7181 (72%)

X-Axis: False Positive
Y axis: True Positive

Curve is closer to
the upper left corner
which is good

Figure 43: Random Forrest & ROC Curve for Annotated Label on Signal Data

4.2.2.7 Collective Anomalies from our Experiments

We apply the collected real-world data along with few more examples to plug them into our proposed rules to show the collective anomaly. We categorize the heterogeneous multiscale data streams into different levels to collective anomalies to visualize the safety of the driving conditions.

Table 21 shows the anomaly spread of our experimental and test data. In our experimental run, we have overlapping anomalies across all three data streams.

Table 21: Anomaly Spread

a_t^M	a_t^H	a_t^G	DA_Q	Anomaly Spread	Comment
15	10	12	1	Collective Spread Impact	all three data streams have at least one overlapping anomaly, 10 intersectioned anomalies among the 3 data streams
15	10	12	0.67	Limited Collective Spread Impact	M intersects with G and not with H, then the Count is 2
10	0	0	0	Non-overlapping Stream Impact	No Anomaly with Health and Gaze data streams
13	4	2	1	Collective Spread Impact	all three data streams have at least one overlapping anomaly, 2 intersectioned anomalies among the 3 data streams
10	7	5	0	Non-overlapping Stream Impact	No intersection
0	0	3	0	Non-overlapping Stream Impact	No intersection
0	0	0	0	Normal State	No Anomaly
193	65	154	1	Collective Spread Impact	Real World

Therefore, we have Collective Spread impact (3/3=1) for our real-world dataset. In other words, we have a good spread of anomalies among all three data streams. We cross-check our result with recorded video and notice that we have frequent overlapping anomalies across all three domains. This means that the driving state for this driver is **unsafe**.

Table 22 shows the Temporal overlaps of our experimental and test data. In our experimental run, we have 65 common anomalies over time across all three data streams.

Table 22: Temporal Overlaps

a_t^M	a_t^H	a_t^G	OA_Q	Temporal Overlaps	Comment
15	10	12	3.33	Collective Temporal Overlaps	10 intersectioned anomalies among the 3 data streams
15	10	12	4.00	Collective Temporal Overlaps	M intersects with G and not with H, then the Count is 12
10	0	0	0	Non-overlapping anomaly	No Anomaly with Health and Gaze data streams
13	4	2	0.67	Limited Collective Temporal Overlaps	Common intersection with 2 anomalies
10	7	5	0	Non-overlapping anomaly	No intersection
0	0	3	0	Non-overlapping anomaly	Driver distracted multiple times, however, no anomaly on Telematics & Health.
0	0	0	0	Normal State	No Anomaly
193	65	154	21.67	Collective Temporal Overlaps	Real World

Therefore, we see Collective Temporal Overlaps (65/3 = 21.67) for our real world dataset. We have all three data sets having overlapping anomalies for a given timespan. Therefore the driving condition of this driver is **unsafe** as discovered on the actual reading from the recording.

Table 23 shows the Intensity of anomaly of our experimental and test data. Intensity gives the maximum number of anomalies among the three datasets over time. In our experimental run, we have 193 Telematics anomalies as the max over time t (600 secs).

Table 23: Intensity of Anomaly

a_t^M	a_t^H	a_t^G	$\|A_\Omega\|$	Intensity of Anomaly	Comment
15	10	12	0.03	Limited Collective Intensity	t=600 sec (10 min)
15	10	12	0.03	Limited Collective Intensity	M intersects with G and not with H
10	0	0	0.02	Limited Collective Intensity	No Anomaly with Health and Gaze data streams
13	4	2	0.02	Limited Collective Intensity	Common intersection with 2 anomalies
10	7	5	0.02	Limited Collective Intensity	No intersection
0	0	3	0.005	Limited Collective Intensity	Driver distracted multiple times, however, no anomaly on Telematics & Health.
30	16	100	0.17	Collective Intensity	Greater than 0.05
0	0	0		Normal State	No Anomaly
97	120	154	0.27	Collective Intensity	Real world
193	65	154	0.33	Collective Intensity	Real world

Therefore, we see Collective Intensity of anomaly (193/600 = 0.33) for our real-world dataset. We have a high collective intensity (27% & 33%) on our data sets for a given timespan. Consequently the driving behavior of this driver is **unsafe**. This is reflected on the actual recorded video.

Table 24 shows the Weighted Anomaly Spread of our experimental and test data. Weighted Anomaly Spread gives us the relative intensity of the anomaly spready across multidomain. In our experimental run, we have 1 anomaly spread with 0.32 Intensity of anomaly.

Table 24: Weighted Anomaly Spread

a_t^M	a_t^H	a_t^G	DA_Ω	IA_Ω	WDA_Ω	**Weighted Anomaly Spread**
15	10	12	1	0.03	0.03	Limited Collective Weighted Spread
15	10	12	0.67	0.025	0.02	Limited Collective Weighted Spread
10	0	0	0	0.02	0.00	Limited Collective Weighted Spread
13	4	2	1	0.02	0.02	Limited Collective Weighted Spread
10	7	5	0	0.02	0.00	No spread
0	0	3	0	0.005	0.00	No spread
30	16	100	0.67	0.17	0.11	Collective Weighted Spread
0	0	0	0	0	0.00	Normal State
97	120	154	1.00	0.26	0.26	Collective Weighted Spread
193	65	154	1.00	0.32	0.32	Collective Weighted Spread

Therefore, we see Collective Weighted Spread for our real-world dataset. We have a high weighted spread (26% & 32%) on our data sets for a given timespan. Therefore, the driving condition for this driver is **unsafe**. Our recorded dataset validates the same.

Table 25 below shows the Weighted Temporal Overlaps of our experimental and test data. Weighted Temporal Overlaps measures the relative overlaps for anomalous states across different domains. In our experimental run, we have 21.67 Temporal Overlaps with 0.11 Intensity of anomaly.

Table 25: Weighted Temporal Overlaps

a_t^M	a_t^H	a_t^G	OA_Ω	IA_Ω	WOA_Ω	**Weighted Temporal Overlaps**
15	10	12	3.33	0.03	0.08	Collective Weighted Temporal Overlaps
15	10	12	4.00	0.03	0.10	Collective Weighted Temporal Overlaps
10	0	0	0	0.02	0.00	Non-overlapping Weighted Temporal Overlaps
13	4	2	0.67	0.02	0.01	Limited Collective Weighted Spread
10	7	5	0	0.02	0.00	Non-overlapping Weighted Temporal Overlaps
0	0	3	0	0.005	0.00	Non-overlapping Weighted Temporal Overlaps
0	0	0	0		0.00	Normal State
193	65	154	21.67	0.11	2.43	Collective Weighted Temporal Overlaps

Therefore, we see Collective Weighted Temporal Overlaps (2.43) for our real-world dataset. We have a high weighted Temporal Overlaps on our data sets for a given timespan. Consequently, the state of the driver behavior is **unsafe**, that is reflected on the actual drive run.

Table 26 below shows the relative presence of a specific domain from our experimental and test data. Relative presence measures the dominating anomalous domain among all other domains. In our experimental run, we have 32% Telematics, 11% Health and 26% Gaze anomalies.

Table 26: Relative Presence of Domain Anomaly

a_t^M	δ_M	a_t^H	δ_H	a_t^G	δ_G	Comment
15	3%	10	2%	12	2%	Most prevakent anomaly is in Telematics
15	3%	10	2%	12	2%	Most prevakent anomaly is in Telematics
10	2%	0	0%	0	0%	Most prevakent anomaly is in Telematics
13	2%	4	1%	2	0%	Most prevakent anomaly is in Telematics
10	2%	7	1%	5	1%	Most prevakent anomaly is in Telematics
0	0%	0	0%	3	1%	Most prevakent anomaly is in Distraction
0	0%	0	0%	0	0%	No Anomaly
193	32%	65	11%	154	26%	Most prevakent anomaly is in Telematics

Therefore, most prevalent (32%) anomaly as shown in Table 26 is coming from Telematics data stream in real world data.

Table 27 shows the weighted intensity of a specific domain from our experimental and test data. Weighted intensity measures the relative intensity of each domain among all other domains. In our experimental run, we have 6.43% of relative intensity for Telematics, 3% of relative intensity for Health and 13% relative intensity for Gaze anomalies.

Table 27: Weighted Intensity of Domain

a_t^M	a_t^H	a_t^G	δ_M	δ_H	δ_G	∂_M	∂_H	∂_G	WIA_a^M	WIA_a^H	WIA_a^G	Comment
15	10	12	2.5%	1.7%	2.0%	0.2	0.3	0.50	0.50%	0.50%	1.00%	Highest weighted intensity is Gaze
15	10	12	2.5%	1.7%	2.0%	0.2	0.3	0.50	0.50%	0.50%	1.00%	Highest weighted intensity is Gaze
10	0	2	1.7%	0.0%	0.3%	0.2	0.3	0.50	0.33%	0.00%	0.17%	Highest weighted intensity is Gaze
13	4	2	2.2%	0.7%	0.3%	0.2	0.3	0.50	0.43%	0.20%	0.17%	Highest weighted intensity is Gaze
10	7	5	1.7%	1.2%	0.8%	0.2	0.3	0.50	0.33%	0.35%	0.42%	Highest weighted intensity is Gaze
0	0	3	0.0%	0.0%	0.5%	0.2	0.3	0.5	0.00%	0.00%	0.25%	Highest weighted intensity is Gaze
0	0	0	0.0%	0.0%	0.0%	0.2	0.3	0.5	0.00%	0.00%	0.00%	No Anomaly
193	65	154	32.2%	10.8%	25.7%	0.2	0.3	0.50	6.43%	3%	13%	Highest weighted intensity is Gaze, second highest is Health and lowest intensity is Telematics

Assuming $\partial_G = 0.5$, $\partial_H = 0.3$, $\partial_M = 0.2$ based on literature and domain knowledge, highest weighted intensity is gaze (Distraction) data stream, which is in our real-world study 13%. Even though the number of Telematics anomaly is higher than that of Gaze anomaly, Gaze (distraction) has more intensity as the impact of distraction is higher. Therefore, if we have Gaze anomaly, in

other words, if the driver is distracted, the intensity of the unsafe situation is worse than having Health anomaly and Telematics anomaly.

In a nutshell, we gather signal data from three data streams – Distraction, Health and Telematics. After we feature extract the data sets, we normalize the data. We then run matrix profile on the data to identify significant changes to the subsequent data attributes which would signify the anomalies. We run Association Rule Mining with Classification based Association to find the relationships between the data attributes from heterogeneous scaled datasets. We run ARM on Signal data, then compare the overlaps on MP data as well as on labelled data. All the key finding and relationships are validated by J48 Decision Trees and its Random Forrest Classifier and ROC with labelled data. We then apply the computational models with the collected driver data to map the intensity of the anomalies with Safe vs. Unsafe conditions. This is how the ground truth of anomalous driving state from the observation and recorded video of subject drivers are fully validated by the data analysis.

4.3 Summary of Results

Following Table 28 summarizes the overall results:

Table 28: Result Summary

Index	Methods	Results
1	Cluster	• Telematics data can be distinctly clustered as Unsafe, Safer, Safe & Safest driving state • Heterogeneous data streams – Distraction, Health and Telematics may not be distinctly clustered as Unsafe, Safer, Safe & Safest driving state

2	ARM with Normalized Signal Data	• Distraction may imply Tension (Anomalous Blood Pressure) [Distraction may cause health anomaly] • Distraction may or may not have an impact on Acceleration [Distraction often cause Telematics to go outside of normal range, not always] • Distraction may impact Speed [Distraction may cause the Telematics to go low or high]
3	ARM with Signal Data with (+-)1 Standard Deviation as Anomaly	• Distraction may imply Tension (High Heartbeat) [Distraction may cause Health anomaly] • Distraction may imply Tension (Anomalous Blood Pressure) [Distraction to Health]
4	ARM with Normalized Matrix Profile Data	• Distraction may imply Tension (High Blood Pressure) [Distraction to Health] • Distraction may imply Emotion (Anomalous EDA) [Distraction to Health] • Distraction & Tension may or may not have impact on Acceleration [Distraction & Health may not cause Telematics to go out of norm]

		• Distraction & Tension may impact Speed [Distraction & Health to Telematics]
5	ARM with Matrix Profile Data with (+-)1 Standard Deviation as Anomaly	• Distraction may imply Tension (High Heartbeat) [Distraction to Health] • Distraction may imply Tension (Anomalous Blood Pressure) [Distraction to Health] • Distraction may imply Emotion (Anomalous EDA) [Distraction to Health] • Distraction & Tension may impact Acceleration [Distraction & Health to Telematics] • Distraction may impact Speed [Distraction to Telematics]
6	Classification on Labelled Data J48 Decision Tree & Random Forrest	• High Eye Gaze Data -> Normal BVD -> Low AirFuelRatio -> Low EDA High Eye Gaze Data -> Normal BVD -> Normal AirFuelRatio -> High Speed [Highly distracted driver may drive fast which is not safe] • High Eye Gaze Data -> Normal BVD -> Normal AirFuelRatio -> Low Speed [Highly distracted driver may also drive very slow which may not be safe either]

		• High Eye Gaze Data -> Normal BVD -> Normal AirFuelRatio -> Normal Speed -> High EAD [High distraction can cause the tension to go high]
7	Random Forrest and ROC	• For Telematics attribute Speed, we notice the accuracy on MP data for Normal, Low Anomalous values and High Anomalous values are high 78%, 75% and 82% respectively • For Annotated Label on MP Data, area under the curve shows high level of accuracy with 83% True Positive • On the contrary, the Annotated Label on Signal Data, area under the curve shows moderate level of accuracy with 72% True Positive • Therefore, our study shows more accurate results on MP data than signal data
8	Heterogeneous Data Streams Analysis	• Combinations of Telematics data stream (T), Eye tracking data stream (D), Driver vital health (H) help detect anomalies
9	Prevalent Anomalous Data Stream Analysis	• Among three data streams, Distraction data stream shows and detects the most

		prevalent and impactful anomalies that indicates the most unsafe driving condition
10	Scale of Anomaly Analysis	• Among the three data streams, Distraction with minimal and small duration may have bigger impact than moderate and longer duration of anomalies on Health and telematics

In summary, we develop a methodology to discover driving patterns using vehicular data and identified anomalous driving states. We utilize telematics data, Eye Gaze distraction data and health vital data to detect anomalies. We discover deviations from baseline driver behavior to detect anomalies and attribute the anomalous behavior to distraction, driver health, and vehicular telematic states. Even though the three data streams may individually help detect anomalous driving behavior, distraction data shows the most effective result for anomalies. The scale of anomaly is also very critical where a very short span of distracted driving could have a huge impact on safety than a moderate anomaly in both health and telematics.

CHAPTER 5

SUMMARY

Our study shows an anomaly detection mechanism for human behavior on heterogeneous temporal data streams. We focus on driver behavior to study anomalies on smart cars to identify unsafe circumstances. Following section talks about the conclusion of our research and its open challenges, direct impact of this research in real world and finally the future work of this study.

5.1 Conclusion and Open Challenges

In our research on driver behavior, we have identified a model to detect anomalous condition of the vehicle and also detect driver's anomalous behavior in near real time. We have applied clusters to baseline driver's driving condition as a whole on Telematics data and compared the baseline with driver's new driving data. We have also learned association rules between the key attributes of the vehicle and its drivers using Apriori Association Rule mining and done predictive analytics based on the rules. We combined few key heterogeneous data streams such as Distraction Data, Health Data, and Telematics Data, and identified the relationships and the similarities among them when it comes to driving anomalies. To validate the results, we compared the anomalies between signal data and matric profile data. We perform clustering on the heterogeneous data streams and found out that Clustering is not the best method of identifying safe and unsafe drivers. Then we applied machine learning methodologies like ARM, Decision Trees, ROC Curve to validate the key rules among different attributes from different data streams. These rules would apply to drivers in near real time to avoid any anomalous situation on the road.

In our study, we have used a limited number of drivers and their real-world driving data. We hoped for more driver population to study the model. We faced several challenges to capture real

world driving data. The biggest challenge was to have the subject drivers drive on the real road while we monitor and capture the driver data. First of all, it was very intimidating for the drivers to wear all the monitoring gadgets like the Tobii eyeglasses with the wire connected to the laptop in the back of the car. Moreover, wearing the E4 wristband to monitor health data as well as connecting their automobiles with OBD2 device as well as running the mobile application to capture speed, turning behavior, break, acceleration, etc. It was rather difficult to capture the real driving behavior as all our subject drivers became very cautious within the situation. Few drivers backed out with the experiments in the middle of the study. With all these limitations, we still wanted to capture the real-world data with very realistic scenarios. Therefore, we did not use any kind of simulators which could be very expensive and yet artificial. In the future, we can minimize the intimidation of the drivers by using an external camera to monitor the eye movements instead of wearing the eyeglasses with wires. Moreover, we have very light weighted fit beats or actual smart watches that mostly have the same data capture as E4.

This research has given us a strong platform to alert drivers from the possibility of harms from cyber threats, engine malfunction and/or driver's reckless and/or distracting driving behavior. The timeliness of this alert could be an issue we need to work on. Timing is of essence and it is critical to identify anomalous behavior in real time, so we can protect the monitoring driver as well as his/her surroundings on the road.

Our research on driver behavior focuses on a time series of a day. However, we have other key attributes that may directly impact the driver behavior on the road, such as age & gender of the driver, road condition, types of car, etc. This research gives us a model to identify anomaly on human behavior based on few key attributes. Extensive model will be derived based on this base model when we add other impactful attributes as mentioned above.

We have a common theme of studying human behavior for driver behavior anomaly detection. As the weakest link, we need to make sure human can identify the anomaly on time and safeguard automobile from any kinds of life-threatening incidents.

Following Table 29 shows the Summary of our findings:

Table 29: Summary of Findings & Contribution

Index	Findings & Contribution
1	Overlapping anomalies across multi-scale heterogeneous data streams lead to a very unsafe driving condition
2	Even a single domain anomaly may also cause unsafe driving condition if the impact of the anomalies for that domain is high.
3	Distraction & Telematics by itself are good indicators of anomalous driving behavior
4	Vital Health by itself is not a good indicator of anomalous driving behavior
5	Telematics data can be distinctly clustered as Unsafe, Safer, Safe & Safest driving state
6	Multi-source data streams – Distraction, Health and Telematics together provide good indicators of anomalous driving state
7	Distraction is a key identifier of anomaly
8	Distraction → Health (Distraction may imply anomalous vital health)
9	Distraction → Telematics (Distraction may imply anomalous Telematics state)
10	Distraction + Health → Telematics (Distraction & Health may imply anomalous telematics state of the vehicle)
11	Health → Telematics (weak rule - Relationships between Health and Telematics could be very weak)

12	Scale of anomaly for distraction is very small and yet very impactful
13	Tolerance of (Anomalous Telematics > Anomalous Health > Anomalous Gaze) Tolerance of ($A_T > A_H > A_G$)
14	Impact of (Change of Anomaly for Gaze > Change of Anomaly for Health > Change of Anomaly for Telematics)
15	Intensity of ($\boldsymbol{\partial}_G > \boldsymbol{\partial}_H > \boldsymbol{\partial}_T$)
16	Clusters on multi-scale data streams are not good indicators of anomaly detection unlike the clusters on Telematics data stream

Our results from the subject drivers show both weak and strong rules that give us a perspective of relativity on different anomalous attributes. We notice anomalous vital health impacting telematics to be a weak rule as ill-health may not usually impact mechanical state of the automobile. On the other hand, Distraction is a dominating factor to impact both telematics and vital health and hence shows a strong rule for identifying the driver behavior anomaly.

5.2 Direct Impact of this Research

We would like to show a direct application of our research on Driver Behavior and vehicles. We have developed a mobile application called "School Bus Connect" for School buses and children who ride the bus. K-12 bus riders' safety depends on bus drivers' driving behavior and the condition of the bus. We monitor driver's behavior in real time and any deviation from the associated speed limit, sharp turns or abrupt acceleration, attention of the road as well as deviation from sound health conditions, we raise alerts to the bus driver on the tablet application as well as send messages to the school administrator's workstation. This mobile platform also connects to the OBD2 device via Bluetooth and monitor vehicle's mechanical state like the tire pressure,

Engine temperature, RPM etc. in real time so the system can raise alerts for any anomalous state on the vehicle. We monitor driver behavior by focusing on their speed, sharp turns, abrupt acceleration, abrupt deceleration, distraction, health status and weather condition that may impact driver's ability to drive safely. We raise alerts to the bus driver and the school transpiration in case of anomalous behavior. We raise mechanical alerts for tire pressure, in general, we monitor the operator's behavior and the condition of the vehicle. We are in the process of adding two more attributes that we have studied here to make it more comprehensive and safer. We are adding health vital monitoring wristbands to the drivers to read their health state in real time and raise alert in case we notice anomalous states that are unsafe to continue driving. We are also adding a distraction monitoring device (camera) that tracks drivers' pupil and determine if they are looking on the road or distracted with something else. The school transportation authority would monitor the drivers in real time from their central workstations in the back office and tackle anomalous situations accordingly.

Another good application of our research would be to install this system to public transportation like public busses or even trains to monitor the operator's behavior and the condition of the vehicles to safeguard the citizens from any sort of terrorist activities and communicate to the homeland security and the law enforcement authority instantaneously.

Several auto insurance companies use telematics data to determine the driving behavior of the drivers on a regular basis to determine their risk factors for insurance. For instance, they monitor drivers' speed, abrupt acceleration, abrupt deceleration, braking patterns, etc. to determine the driver behavior. However, insurance companies are missing two other major factors that impact driver behavior which are distraction and health condition that we have added in our study here. Driving behavior and driving state depend on driver's attention and focus on the road as well as

his/her health condition. A distracted driver would not drive as smooth as an attentive driver on the road. Likewise, a tensed and ill driver would not be able to drive as safe as a healthy driver on the road. Consequently, our research brings in a more holistic approach for identifying drivers' anomalous behavior.

5.3 Future Work

We have proposed a model to detect driver's adverse behavior along with the vehicle's anomalous condition. This base model considers the time of the day the driver is driving along with the driver's health condition and distractions. Driver behavior also depends on the age, gender, mood, the experience of the driver, the road condition as well as the traffic condition, the environment like the weather condition and the type of the car. We like to bring in these impactful attributes into our model of anomaly detection, and safeguard drivers and the passengers from any kind of cyber threat, and/or adverse situation. We, the consumers, love connected smart cars. The smart car industry will be worth $43 billion by 2023 [53]. With all the sensors and intelligent systems around the car have the computing power of 20 personal computers with 100 million lines of computer code and process 25 gigabytes of data per hour. All these sophistications bring in the vulnerabilities of car control hacks, smart alarm hacks, as well as insecure apps embedded in our cars [53]. If the car becomes a victim of these cyber-attacks, the driving state would automatically shift to an anomalous one and our model of identifying anomalies in driver behavior and driver state would assist detecting the unsafe state and help take proper actions in a timely manner. We will leverage the common theme of time series, binning, Clusters, ARM, Matrix Profile, Decision Trees and ROC to correlate the proposed models for automobile related threat detection with the addition of driver's mood, age, weather, road condition and traffic condition. Considering these extra attributes will make the anomaly detection complete and driving condition extremely safe.

For single stream cluster that works well to find the anomalies for that particular domain only, we will utilize Silhouette Coefficient to measure the quality of the cluster.

We need to make sure the driver data collection tools and sensors are less invasive or less intimidating for the drivers. In our study, several drivers were reluctant to wear Tobii Eyeglasses with wires connected to the laptop. In our future work, we can capture the same distraction data with more sophisticated micro camera installed on the dashboard and/or on the windshield of the automobile monitoring the eye movements for the driver for distraction data. Likewise, instead of the E4 wristband, the subject can wear the regular iWatch (Apple Watch) that would capture the same types of data attributes. OBD2 reading for Telematics data could also be done in the future with built-in telematics device connecting to the software with a Bluetooth connection without any wire. These seamless data collection strategies would not only make the subjects much more accepting to do the actual experiments but will also make it much easier in the real world for the drivers to adopt with this technology.

This human behavioral anomaly detection model can be applied to any heterogeneous data streams outside of these attributes we used in our research. We can also bring in the predictive analytics aspect for human behavior to predict any potential unusual activity. The result has to be quick and almost real time to alert and safeguard the drivers from any kinds of anomalous situation. Our model of detecting the anomalous driver behavior and anomalous driving state is not always real time and hence, as future work, we need to optimize/enhance and come up a with a real time anomaly detection system using our model to immediately take action and safeguard the drivers from any sort of unsafe driving condition. This model can also predict and/or protect from any kind of cyber-attack on the transportation. Lastly, the research can be modeled for not only the

private transportation but also for public transportation like buses, trucks, trains and/or any kinds of goods transferring vehicles for both short and long distance.

REFERENCES

[1] Academy pro on Genesis Framework. 50+ Car Accident Statistics in the U.S. & Worldwide. The Wandering RV. Source: https://www.thewanderingrv.com/car-accident-statistics/ October 17, 2019.

[2] Adams J. Williams, et al. Collective Views of the NSA/CSS Cyber Defense Exercise on Curricula and Learning Objectives. 2009. Available from http://static.usenix.org/event/cset09/tech/full_papers/adams.pdf.

[3] Administration, N.H.T.S. Traffic Safety Facts: Distracted Driving 2014. Available online: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812260 (accessed on 2 July 2016).

[4] Aggarwal, C. C. Editor. Social Network Data Analytics. Springer, 2011.

[5] Al Mamun A., Al Mamun M. A., Shikfa A. (2018). "Challenges and Mitigation of Cyber Threat in Automated Vehicle: An Integrated Approach," in *Proceedings of the 2018 International Conference of Electrical and Electronic Technologies for Automotive*, (Milan: IEEE;). [Google Scholar]

[6] Albert Bifet. Mining Big Data in Real Time. Informatica 37, 15–20. 2013.

[7] Alperovitch, D. (2011). Revealed: Operation Shady RAT. McAfee. Retrieved from http://www.mcafee.com/us/resources/white-papers/wp-operation-shady-rat.pdf.

[8] Anbarci N, Escaleras M, Register C. Traffic fatalities and public sector corruption. Kkylos. 2006;59(3):327–344.

[9] Anderson L. Catherine, Agarwal, Ritu. Practicing Safe Computing: A Multimethod Empirical Examination of Home Computer User Security Behavioral Intentions. September 2010. MIS Quarterly, vol. 34, No. 3, pp. 613-643.

[10] A. Walker, U.S. traffic death increase caused by speeding, says new study [online], available at https://www.curbed.com/2017/7/28/16051780/us-traffic-death-speeding-statistics-speeding, July 28, 2017

[11] Ball, K.; Rebok, G. Evaluating the driving ability of older adults. J. Appl. Gerontol. 1994, 13, 20–38.

[12] Binde E. Beth, McRee Russ, O'Connor Terrance. Assessing Outbound Traffic to Uncover Advanced Persistent Threat. May 2011. Available from http://www.sans.edu/student-files/projects/JWP-Binde-McRee-OConnor.pdf.

[13] Blasco, J. (2013, Mar 21). New Sykipot Developments. Alient Vault Labs. Retrieved from http://labs.alienvault.com/labs/index.php/2013/new-sykipot-developments.

[14] Bloomberg Jason. Cybersecurity: A Human Problem. August 14, 2012. Zapthink. Available at http://www.zapthink.com/2012/08/14/cybersecurity-a-human-problem/.

[15] Braitman KA, McCartt AT. National reported patterns of driver cell phone use in the United States. Traffic Inj Prev. 2010;11:543Y548.

[16] B. Sheehan, F. Murphy, C. Ryan, M. Mullins, H.Y. LiuSemi-autonomous vehicle motor insurance: a Bayesian Network risk transfer approach Transport. Res. Part C: Emerging Technol., 82 (2017), pp. 124-137

[17] Budhaditya, S., Pham, D., Lazarescu, M., and Venkatesh, S. Effective Anomaly Detection in Sensor Networks Data Streams. Department of Computing, Curtin University of Technology, Perth, Western Australia 2009 Ninth IEEE International Conference on Data Mining.

[18] Burzio G., Cordella G. F., Colajanni M., Marchetti M., Stabili D. (2018). "Cybersecurity of Connected Autonomous Vehicles: A ranking based approach," in *Proceedings of the 2018*

International Conference of Electrical and Electronic Technologies for Automotive, (Milan: IEEE;). [Google Scholar]

[19] Caird J. A meta-analysis of the effects of texting on driving. Accident Analysis and Prevention. 2014;71:311–318.

[20] Carmona, J.; García, F.; Martín, D.; Escalera, A.D.L.; Armingol, J.M. Data fusion for driver behaviour analysis. Sensors 2015, 15, 25968–25991.

[21] Carsten, O.; Merat, N. Protective or not? (visual distraction). In Proceedings of the 2015 4th International Conference on Driver Distraction and Inattention, Sydney, Australia, 9–11 November 2015.

[22] Centers for Disease Prevention and Control. Mobile device use while driving — United States and seven European countries, 2011. Morbidity and Mortality Weekly Report (MMWR). 2013;62(10);177–182 (http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6210a1.htm?s_cid=mm6210a1_w, accessed 15 September 2015).

[23] Chandola, V., Banerjee, A., and Kumar, V. Anomaly Detection: A Survey, University of Minnesota, ACM Computing Surveys, Vol. 41, No. 3, Article 15, Publication date: July 2009.

[24] Chan, M, Global Status Report on Road Safety, 2015. World Health Organization. ISBN 978 92 4 156506 6.

[25] Chen, M., Han, J., and Yu, S. P. Data Mining: An Overview from a Database Perspective. IEEE Transactions on Knowledge and Data Engineering. Vol 8, No. 6. December 1996.

[26] Cority, Improving Safety Culture: Where to focus your attention for maximum impact. Minimizing Motor Vehicle Crashes: How Telematics is the New Driver Safety Measure, September 9, 2019. Available at https://ohsonline.com/Articles/2019/09/09/Minimizing-Motor-

Vehicle-Crashes-How-Telematics-is-the-New-Driver-Safety-Measure.aspx?admgarea=news&m=1&Page=1

[27] C. Patsakis, K. Dellios, M. Bouroche, Towards a distributed secure in-vehicle communication architecture for modern vehicles, Computers & Security. 40 (2014) 60–74

[28] Crancer A, Jr, McMurray L. Accident and violation rates of Washington's medically restricted drivers. JAMA. 1968;205:272–6. [Google Scholar]

[29] Craye, C.; Karray, F. Multi-distributions particle filter for eye tracking inside a vehicle. In Image Analysis and Recognition; Springer: Berlin/Heidelberg, Germany, 2013; pp. 407–416.

[30] Dahdah S, McMahon K. The true cost of road crashes: valuing life and the cost of a serious injury. Washington: International Road Assessment Programme, World Bank Global Road Safety Facility; 2008.

[31] Department of Transportation. Regulations. Available at: http://distraction.gov/content/dot-action/regulations. Accessed July 14, 2012.

[32] Department of Transportation. State laws. Available at: http://distraction.gov/get-the-facts/state-laws. Accessed on July 14, 2012.

[33] Dibaei, Mahdi, et al. Attacks and defences on intelligent connected vehicles: a survey. April 2020. Digital Communications and Networks.

[34] Dischinger, P., Ho, S., and Kufera, J. "Medical Conditions and Car Crashes". Association for the Advancement of Automotive Medicine. Annu Proc Assoc Adv Automot Med. 2000; 44: 335–348.

[35] Dragutinovic N, Twisk D. Use of mobile phones while driving – effects on road safety. Leidschendam, Netherlands: SWOV Institute for Road Safety Research; 2005.

[36] Drews FA, Yazdani H, Godfrey CN, Cooper JM, Strayer DL. Text messaging during simulated driving. Hum Factors. 2009;51:762Y770.

[37] Eyben, F.; Wöllmer, M.; Poitschke, T.; Schuller, B.; Blaschke, C.; Färber, B.; Nguyen-Thien, N. Emotion on the road—Necessity, acceptance, and feasibility of affective computing in the car. Adv. Hum. Comput. Interact. 2010, 2010, 263593.

[38] Fernández, Alberto. Driver Distraction Using Visual-Based Sensors and Algorithms. October 2016. Sensors MDPI.

[39] Flores, M.J.; Armingol, J.M.; de la Escalera, A. Driver drowsiness warning system using visual information for both diurnal and nocturnal illumination conditions. EURASIP J. Adv. Signal Process. 2010, 2010, 438205.

[40] Foster Ian, Prudhomme Andrew, Koscher Karl and Savage Stefan. Fast and Vulnerable: A Story of Telematic Failures. Department of Computer Science and Engineering University of California, San Diego.

[41] Geng, L., and Hamilton, J. H. Interestingness Measures for Data Mining: A Survey. ACM Computing Surveys, Vol. 38, No. 3, Article 9. September 2006.

[42] Global health estimates. Geneva: World Health Organization; 2014 (http://www.who.int/healthinfo/global_burden_disease/estimates/en/index1.html, http://www.who.int/healthinfo/global_burden_disease/projections/en, accessed 15 September 2015).

[43] Global status report on road safety: time for action. Geneva, World Health Organization, 2009 (www.who.int/violence_injury_prevention/road_safety_status/2009/en/index.html, accessed 15 September 2015).

[44] Governors Highway Safety Association. Cell Phone Driving Laws. February, 2010.

[45] Governors Highway Safety Association. Cell phone and texting laws. Available at: http://www.ghsa.org/html/stateinfor/laws/cellphone_laws. Accessed on July 14, 2012.

[46] Greenberg, A. Hackers reveal nasty new car attacks–with me behind the wheel (video) [Online]. Available, https://www.forbes.com/sites/andygreenberg/2013/07/24/hackers-reveal-nasty-new-car-attacks-with-me-behind-the-wheel-video/#54dcbff1228c. 2013

[47] Greitzer L. Frank, Hohimer E. Ryan. Modeling Human Behavior to Anticipate Insider Attacks. Summer 2011. In Journal of Strategic Security: Strategic Security in Cyber Age. Volume 4, Number 2, Summer 2011.

[48] Greitzer L. Frank, Moore P. Andrew, Cappelli M. Dawn, Andrew H. Dee, Carroll A Lynn, Hull D. Thomas. Combating the Insider Cyber Threat. January/February 2008. In IEEE Security & Privacy.

[49] Guibert R, Duarte-Franco E, et al. Medical conditions and the risk of motor vehicle crashes in men. Arch Fam Med. 1998;7(6):554–8. [PubMed] [Google Scholar]

[50] Guo, Chonghui, et al. An Improved Piecewise Aggregate Approximation Based on Statistical Features for Time Series Mining. Springer-Verlag Berlin Heidelberg 2010.

[51] Guyon, Isabelle and Elisseeff Andr´e. An Introduction to Feature Extraction. March 2003. Springer Link.

[52] Hansotia P, Broste SK. The effect of epilepsy or diabetes mellitus on the risk of automobile accidents. N Engl J Med. 1991;324(1):22–26. [PubMed] [Google Scholar]

[53] Hitachi Systems Security Inc, Smart Car Security Threats: Is the Connected Car a Good Idea? HITACHI Inspire the Next, March 26, 2019, Retrieve from https://www.hitachi-systems-security.com/blog/smart-car-security-threats-is-the-connected-car-a-good-idea/

[54] Hoff J, Grell J, Lohrman N, et al. Distracted driving and implications for injury prevention in adults. J Trauma Nurs. 2013;20(1):31Y34.

[55] Holland C, RathodV. Influence of personal mobile phone ringing and usual intention to answer on driver error. Accid Anal Prev. 2013;50:793Y800.

[56] Hosking SG, Young KL, Regan MA. The effect of text messaging on young drivers. Hum Factors. 2009;51:582Y592.

[57] Hubaux Jean-Pierre, Apkun C Srdjan and Luo Jun. The Security and Privacy of Smart Vehicles. May-June 2004. In IEEE Security & Privacy.

[58] Insurance Institute for Highway Safety. Texting bans. Status Report. 2010; 45(10):1Y3.

[59] Icosystem. Cyber Security: Human Behavior Matters. Icosystem Corporation. 10 Fawcett St., Cambridge, MA 02138. Voice: (617) 520 1000. Available at http://www.icosystem.com/cyber-security-human-behavior-matters/.

[60] I. Studnia, V. Nicomette, E. Alata, Y. Deswarte, M. Kaâniche, Y. LaarouchiSurvey on security threats and protection mechanisms in embedded automotive networks
Dependable Systems and Networks Workshop (DSN-W), 2013 43rd Annual IEEE/IFIP Conference on, IEEE (2013), pp. 1-12

[61] Jacobs G, Aeron-Thomas A, Astrop A. Estimating global road fatalities. Crowthorne, Transport Research Laboratory, 2000 (TRL Report 445).

[62] Janeja, V. P., & Palanisamy, R. (2013). Multi-domain anomaly detection in spatial datasets. Knowledge and information systems, 36(3), 749-788.

[63] Jeon, M.;Walker, B.N.; Gable, T.M. Anger effects on driver situation awareness and driving performance. Presence Teleoper. Virtual Environ. 2014, 23, 71–89.

[64] Klauer, S.G.; Neale, V.L.; Dingus, T.A.; Ramsey, D.; Sudweeks, J. Driver inattention: A contributing factor to crashes and near-crashes. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Orlando, FL, USA, 26–30 September 2005; SAGE Publications: Thousand Oaks, CA, USA; Volume 49, pp. 1922–1926.

[65] Koepsell TD, Wolf ME, McCloskey L, et al. Medical conditions and motor vehicle collision injuries in older adults. J Am Geriatr Soc. 1994;42(7):695–700. [PubMed] [Google Scholar]

[66] Koornstra M et al. SUNflower: a comparative study of the development of road safety in Sweden, the United Kingdom, and the Netherlands. Leidschendam: SWOV; 2002.

[67] Kortschot, S. W., Sovilj, D., Jamieson, G. A., Sanner, S., Carrasco, C., and Soh, H. (2018). Measuring and mitigating the costs of attentional switches in active network monitoring for cybersecurity. Hum. Factors 60, 962–977. doi: 10.1177/0018720818784107

[68] Koscher K, Czeskis A, Roesner F, Patel S, Kohno T, Checkoway S, McCoy D, Kantor B, Anderson D, Shacham H, and Savage S. ”Experimental security analysis of a modern automobile”. Evans D and Vigna G, editors, IEEE Symposium on Security and Privacy. IEEE Computer Society, May 2010.

[69] Lee, J.D.; Young, K.L.; Regan, M.A. Defining Driver Distraction; CRC Press: Boca Raton, FL, USA, 2008.

[69] Lee, W., Stolfo, J. S., Chan, K. P., Eskin, E., Fan, W., Miller, M., Hershkop, S., and Zhang, J. Real Time Data Mining-based Intrusion Detection. DARPA Information Survivability Conference & Exposition II, 2001. DISCEX '01. June 2001.

[70] Lesch MF, Hancock PA. Driving performance during concurrent cellphone use: are drivers aware of their performance decrements? Accid Anal Prev. 2004;36:471Y480.

[71] Leyden, J. (2012, Mar 29). NSA's top spook blames China for RSA Hack. The Register. Retrieved from http://www.theregister.co.uk/2012/03/29/nsa_blames_china_rsa_hack/.

[72] Linkov V., Zamecnik P., Havlickova D., Pai C. "Human Factors in the Cybersecurity of Autonomous Vehicles: Trends in Current Research" in Frontiers in Psychology, 3 May 2019, Retrieve from https://doi.org/10.3389/fpsyg.2019.00995

[73] Llerena, L., Aronow, K., Macleod, L., Bard, M., Salzman, S., Greene, W., Haider, A., and Schupper, A. An evidence-based review: Distracted driver. Tampa, Florida. Guidelines, 2015 Wolters Kluwer Health, Inc. All rights reserved.

[74] Macias A. Injury prevention legislation. Bull Am Coll Surg. 2011;96:51Y52.

[75] Mandiant. APT1: Exposing One of China's Cyber Espionage Units. 2013. Available from http://intelreport.mandiant.com/Mandiant_APT1_Report.pdf.

[76] M. Beltov, Security Issues Identified in the MirrorLink Smart Car Standard, Security News, 2017, Retrieve from https://bestsecuritysearch.com/security-issues-identified-mirrorlink-smart-car-standard/

[77] McAfee Labs. (2011). Global Energy Cyberattacks: "Night Dragon". Retrieved from http://www.mcafee.com/us/resources/white-papers/wp-global-energy-cyberattacks-night-dragon.pdf.

[78] McAfee Labs (2010). Protecting Your Critical Assets. Retrieved from http://www.mcafee.com/us/resources/white-papers/wp-protecting-critical-assets.pdf.

[79] McEvoy SP, Stevenson MR,Woodward M. The prevalence of, and factors associated with, serious crashes involving a distracting activity. Accid Anal Prev. 2007;39:475Y482.

[80] McGwin G, Jr, Sims RV, et al. Diabetes and automobile crashes in the elderly. A population-based case-control study. Diabetes Care. 1999;22(2):220–7. [PubMed] [Google Scholar]

[81] Miller Charlie, Valasek Chris. “Remote Exploitation of an Unaltered Passenger Vehicle”. August 2015. Black Hat USA.

[82] Moisan, F., and Gonzalez, C. (2017). Security under uncertainty: adaptive attackers are more challenging to human defenders than random attackers. Front. Psychol. 8:982. doi: 10.3389/fpsyg.2017.00982

[83] Monticello, M. & Shenhar, G (March 2016). CR Consumer Report. “Must-Have Car Features, and Those You Can Skip. Advice to help with optioning your next car“. Retrieved from https://www.consumerreports.org/cro/news/2015/03/must-have-car-features-and-those-you-can-skip/index.htm

[84] Namayanja, J. & Janeja, V (2013, June 4-7). Discovery of Persistent Threat Structures through Temporal and Geo-Spatial Characterization in Evolving Networks. ISI 2013, Seattle, Washington.

[85] Nantulya VM, Reich MR. The neglected epidemic: road traffic injuries in developing countries. British Medical Journal. 2002;324 (7346):1139–41.

[86] National Conference of State Legislatures. Spotlight: distracted driving. Available at: http://www.ncsl.org/issues-research/transportation. Accessed July 14, 2012.

[87] National Highway Traffic Safety Administration. An examination of driver distraction as recorded in NHTSA databases. Traffic Safety Facts. 2009 DOT HS811 216. Available at: www.nhtsa.gov. Date accessed December 15, 2010.

[88] National Highway Traffic Safety Administration. Distracted Driving 2009. Traffic Safety Facts. 2010 DOT HS 811 379. Available at: www.nhtsa.gov. Accessed date December 15, 2010.

[89] National Highway Traffic Safety Administration. Driver Distraction Program. 2010 DOT HS 811 299. Available at: www.distraction.gov.

[90] Neyens DM, Boyle LN. The influence of driver distraction on the severity of injuries sustained by teenage drivers and their passenger. Accid Anal Prev. 2008;40:254Y259.

[91] NPDN National Plant Diagnostic Network. Operational Security - Social Engineering. 2003-2020. Available at https://www.npdn.org/infosec_social_engineering.

[92] NPDN National Plant Diagnostic Network. Operational Security – Software. 2003-2020. Available at http://npdn-d7.ceris.purdue.edu/infosec_software.

[93] O'Gorman, G. & McDonald, G. (2012). The Elderwood Project. Symantec. Retrieved from https://www.c-cure.dk/Files/Datasheet/Symantec/Symantec%20the-elderwood-project.pdf.

[94] Peden M et al., editors. World report on road traffic injury prevention. Geneva, World Health Organization,2004. (www.who.int/violence_injury_prevention/publications/road_traffic/world_report/en/index.html, accessed 15 September 2015).

[95] Pfleging, B.; Fekety, D.K.; Schmidt, A.; Kun, A.L. A Model Relating Pupil Diameter to Mental Workload and Lighting Conditions. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, 7–12 May 2016; pp. 5776–5788.

[96] Ranney, T.A.; Mazzae, E.; Garrott, R.; Goodman, M.J. NHTSA driver distraction research: Past, present, and future. In Driver Distraction Internet Forum; National Highway Traffic Safety Administration: Washington, DC, USA, 2000; Volume 2000.

[97] Rezaei, M.; Klette, R. Simultaneous analysis of driver behaviour and road condition for driver distraction detection. Int. J. Image Data Fusion 2011, 2, 217–236.

[98] R. Lennon, R. Rentfro, B. O'Leary, Social Marketing and Distracted Driving Behaviors among Young Adults: The Effectiveness of Fear Appeals. Academy of Marketing Studies Journal, Volume 14, Number 2, 2010

[99] Roddick, J., Spiliopoulou, M.: A Survey of Temporal Knowledge Discovery Paradigms and Methods. In IEEE Transactions of Knowledge and Data Engineering, vol. 13, 2001.

[100] Sahayadhas, A.; Sundaraj, K.; Murugappan, M. Detecting driver drowsiness based on sensors: A review. Sensors 2012, 12, 16937–16953.

[101] Sawyer, B. D., and Hancock, P. A. (2018). Hacking the human: the prevalence paradox in cybersecurity. Hum. Factors 60, 597–609. doi: 10.1177/0018720818780472

[102] Schmidt, Jeff. The weakest link. Enterprise communications networks often are victimized by those who use them. January 12, 2012. Operations Cybersecurity: Urgent Communications. (pp. 14–17).

[103] Shannon V. 15 years of text messages, a 'cultural phenomenon'. New York Times. 2007. Available at: http:///www.nytimes.com/2007/12/05/techology/05iht-sms.4.8603150.html. Date accessed August 20, 2010.

[104] Shukla, Suraksha and Janeja P. Vandana. Multi-Domain Anomalous Temporal Association (Multi-DATA). MILETS19, KDD, August 2019, Anchorage, Alaska

[105] Sims RV, Owsley C, et al. A preliminary assessment of the medical and functional factors associated with vehicle crashes by older adults. J Am Geriatr Soc. 1998;46(5):556–61. [PubMed] [Google Scholar]

[106] Sonalker et al. TEMPORAL ANOMALY DETECTION ON AUTOMOTIVE NETWORKS. September 2018. United States Patent. Patent No . : US 10 , 083 , 071 B2.

[107] Speed management. Paris, France: Organisation for Economic Cooperation and Development; 2006 (www.internationaltransportforum.org/Pub/pdf/06Speed.pdf, accessed 15 September 2015).

[108] Straub J., McMillan J., Yaniero B., Schumacher M., Almosalami A., Boatey K., et al. (2017). "CyberSecurity considerations for an interconnected self-driving car system of systems," in *Proceedings of the 2017 12th System of Systems Engineering Conference (SoSE)*, (Waikoloa, HI: IEEE;). [Google Scholar]

[109] Stravrinos D, Jones JL, Garner AA, et al. Impact of distracted driving on safety and traffic flow. Accid Anal Prev. 2013;61:63Y70.

[110] Sun, Y., and Han, J. Mining Heterogeneous Information Networks: A Structural Analysis Approach. SIGKDD Explorations, Volume 14, Issue 2. December 2012.

[111] SWOV fact sheet: Use of the mobile phone while driving. Leidschendam, the Netherlands: Institute for Road Safety Research (SWOV); 2012 (http://www.swov.nl/rapport/Factsheets/UK/FS_Mobile_phones.pdf, accessed 15 September 2015).

[112] Talbot, R.; Fagerlind, H. Exploring inattention and distraction in the SafetyNet accident causation database. Accid. Anal. Prev. 2009, 60, 445–455.

[113] Thackray, H., McAlaney, J., Dogan, H., Taylor, J., and Richardson, C. (2016). "Social psychology: An under-used tool in cybersecurity," in Proceedings of the 30th International BCS Human Computer Interaction Conference, HCI '16, Poole.

[114] The SecDev Group. (2009). Tracking GhostNet: Investigating a Cyber Espionage Network. Retrieved from http://www.nartv.org/mirror/ghostnet.pdf.

[115] Toole, L.M. Crash Risk and Mobile Device Use Based on Fatigue and Drowsiness Factors in Truck Drivers. Ph.D. Thesis, Virginia Tech, Blacksburg, VA, USA, October 2013.

[116] Toroyan T et al, editors. Helmets: a road safety manual for decision-makers and practitioners. Geneva: World Health Organization; 2006 (http://www.who.int/roadsafety/projects/manuals/helmet_manual/en/, accessed 15 September 2015).

[117] Towards zero: ambitious road safety targets and the safe system approach. Paris, France: Organisation for Economic Co-operation and Development; 2008 (http://www.internationaltransportforum.org/Pub/pdf/08TowardsZeroE.pdf, accessed 15 September 2015).

[118] Traffic law enforcement: a review of the literature. Monash University Accident Research Centre: Report 53, 1994.

[119] Trend Micro Incorporate. Detecting APT Activity with Network Traffic Analysis. 2012. Available from http://www.trendmicro.com/cloud-content/us/pdfs/security-intelligence/white-papers/wp-detecting-apt-activity-with-network-traffic-analysis.pdf.

[120] T. Zhang, H. Antunes, S. Aggarwal, Defending connected vehicles against malware: challenges and a solution framework IEEE Internet Things J., 1 (2014), pp. 10-21

[121] Villeneuve, N. & Sancho, D. (2011). The "Lurid" Downloader. Trend Micro. Retrieved from http://www.trendmicro.com/cloud-content/us/pdfs/security-intelligence/white-papers/wp_dissecting-lurid-apt.pdf.

[122] Waller J. A. Health status and motor vehicle crashes [editorial; comment] N Engl J Med. 1991;324(1):54–5. [PubMed] [Google Scholar]

[123] Websense. Advanced Persistent Threat and Advanced Attacks: Threat Analysis and Defense Strategies for SMB, Mid-Size, and Enterprise Organizations Rev 2. 2011.

[124] Wegman F, Aarts L, editors. Advancing sustainable safety: national road safety outlook for 2005–2020. SWOV Institute for Road Safety Research, 2006 (http://www.swov.nl/rapport/dmdv/Advancing_sustainable_safety.pdf, accessed 16 September 2015).

[125] West Point, United States Military Academy. Network Diagram (Pre-CDX 2009). Data Capture from National Security Agency (NSA). 2009. Available from http://www.westpoint.edu/crc/SitePages/DataSets.aspx

[126] White, J. "In My Daughter's Defense: A Perspective on Driver Distraction and Telematics", Available at https://www.nhtsa.gov/DOT/NHTSA/NRD/Multimedia/PDFs/Crash%20Avoidance/Driver%20Distraction/WhiteJ_doc.pdf

[127] Wilson FA, Stimpson JP. Trends in fatalities from distracted driving in the United States, 1999 to 2008. Am J Public Health. 2010;100(11):2213Y2219.

[128] Wybourne N. Martin, Austin F. Martha, Palmer C. Charles. National Cyber Security. Research and Development Challenges. Related to Economics, Physical Infrastructure and Human Behavior. 2009. I3P: Institute for Information Infrastructure Protection.

[129] Yang, Sang-Chin, Wang, Yi-Lu. System Dynamics Based Insider Threat Modeling. 2011. International Journal of Network Security & Its Applications (IJNSA). Vol. 3, No. 3, May 2011.