Identifying Important Risk Factors Associated with Vehicle Injuries Using Driving Behavior Data and Predictive Analytics

Michal Monselise
College of Computing and Informatics
Drexel University
Philadelphia, PA
mm4837@drexel.edu

Ou Stella Liang
College of Computing and Informatics
Drexel University
Philadelphia, PA
ol54@drexel.edu

Christopher C. Yang
College of Computing and Informatics
Drexel University
Philadelphia, PA
chris.yang@drexel.edu

Abstract: Road injuries are rated among the top 10 causes of death by the World Health Organization, and the only one that is not a disease. The total economic cost of motor vehicle crashes in the United States was estimated to be $242 billion a year. This study examines multiple factors of accidents simultaneously with a goal of generating an interpretable model that can predict the occurrence of an accident given road conditions and driver behavior. The study compared 4 machine learning and deep learning modeling techniques on a dataset of 7707 trips collected by the Second Strategic Highway Research Program. A gradient boosted model was found to be most accurate and interpretable in accident prediction. This modeling technique also allows us to rank the feature importance of the factors in the model. The study finds that driver behavior, pre-incident maneuvers and secondary task duration are the most important variables in the predictive model. Using these conclusions will allow us to perform more work to infer these accident causes directly from vehicle sensor data in the future.

Keywords: naturalistic driving, driving behavior, secondary tasks, driving disengagement, predictive analytics, vehicle injuries, injury prevention, big data

I. INTRODUCTION

Road traffic accidents remain to be one of the leading causes of death across the world. The World Health Organization rated road injury to be the eighth most deadly (2018), and the only one of the top 10 causes of death that is not a disease. Some parts of the Africa and South America continents experience higher mortality rate than the rest of the world. In the United States where our data source is collected, motor vehicle traffic related deaths amount to 34,439 in 2016 or 11.6 per 100,000 population. Although the total fatal crashes have been slowly decreasing, the fatality rate disproportionately affects the lives of teenage and young adults. Furthermore, motor vehicle related injuries impact 600 or more lives per 100,000 population. The total estimated economic cost of traffic crashes to the society was $242 billion in 2010 [1].

Traditional road injury research utilizes simulator-based observational studies. Advanced driving simulators can detect hands-off-wheel behaviors, track eye gazes, and monitor physical manifestation such as perspiration and breathing/heart rate. The simulated driving scenarios are customizable and repeatable to study human behaviors in various scenarios. While driving simulators are particularly helpful in studying targeted, known factors, such as when evaluating a new driver assistance system, its limitation is inherent in the absence of actual physical dynamics of the vehicles that is the child of complex environmental circumstances and human decisions. Unknown detriments, by definition, cannot be designed into simulated scenarios. Simulator-based driving studies are often limited by recruitment size of participants.

In contrast, another approach to study road injuries involves naturalistic driving by participating subjects in real-world or minimally-modified conditions. It generates data of human driving behaviors and vehicle response kinematics representative of naturally occurring scenarios experienced by everyday commuters. Depending on the onboard data acquisition system (DAS) made available in naturalistic driving studies (NDS), in-car driver behaviors can be captured similar to those in simulator-based driving assessment. Compared to simulator-generated results, NDS is more expensive to organize due to the recruitment of study participants, fitting of study vehicles, and administrative tasks associated with the data procurement process. However, NDS datasets offer an unprecedented opportunity to study a plethora of data attributes indicative of human-machine-road interactions.

Natural driving studies generate rich data that meets the widely-accepted dimensions of “big data” in volume, velocity, variety, variability, and value. NDS usually involve a large cohort of participants and the study period can span over multiple years. Time-series data collected by sensors, such as camera and radar, result in a large volume of data in NDS. A large variety of data attributes related to driver and vehicles are collected from natural driving trips. Depending on the individual drivers and road conditions, NDS data exemplify great variation. NDS data can also assume great velocity during an ongoing trip. Sensor data at 10 Hz or higher can be analyzed for real-time prediction of risks. The value of NDS is evident in generating high-fidelity data that reflects human driving in genuine conditions.

Research efforts around the world in the past decade have been organizing naturalistic driving studies. In the U.S., the Virginia Transportation Technology Institute (VTTI) first pioneered a 100-car study over a 12-month period generating 50,000 hours of naturalistic driving data. Success of the study propelled the Second Strategic Highway Research Program (SHRP 2), an expanded follow-up study recruiting more than 3,000 volunteer drivers whose trips over 4-24 months were recorded. The program installed a comprehensive data acquisition system in participating vehicles that include a head unit with three recording cameras providing four views of driver's face, driver's hands, forward roadway and rear roadway. (Fig. 1) Radar, GPS, and accelerometers-based data was also captured in the DAS [2]. The result was an NDS dataset the largest of its kind that included more than 5.5 million trips and 3,900 data hours of driving. Variables available including time-series sensor data, driver characteristics and road infrastructure information. In Europe, PROLOGUE and UDRIVE are two large-scale NDS implemented to evaluate driver behaviors in different EU regions and types of vehicles. Australia and Japan have undertaken similar efforts in recent years.

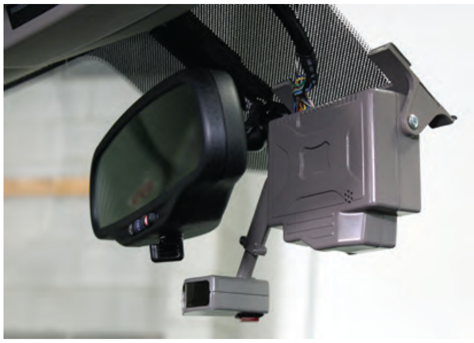


Fig. 1.a. A head unit of the DAS recording four camera views

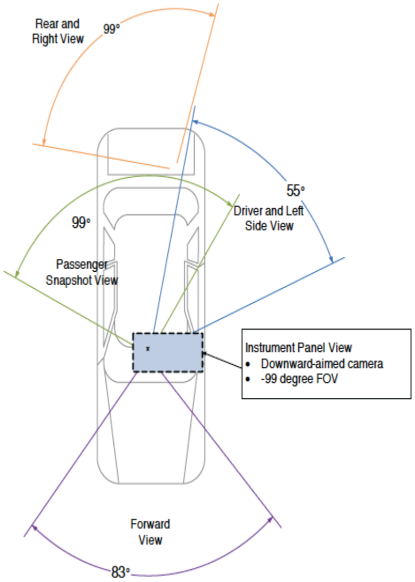


Figure 1.b. Schematic of the four fields of view

Our motivation is to take advantage of the rich data afforded by naturalistic driving datasets and to tackle the immanent challenges of large-scale data processing and computing. In this study, we sought to better understand the associations of driver behaviors and road injuries in natural driving conditions.

II. Literature Review

We provide a brief review of published literature focused on motor vehicle crash prediction. We consider the research on automobile injury prevention highly interdisciplinary spanning across engineering, medicine, public health and ultimately connected by mathematic algorithms. This is evident in the heterogeneity of data attributes commonly studied and often intertwined: driver behavior, vehicle kinematics, and environmental factors. Driver behavior attributes range from driver demographics, psychological evaluation and existing conditions to in-car activities and physiological state when driving. Vehicle kinematics are results of driver maneuvers but represented in physical unit measures. Environmental factors can include both naturally-occurring conditions and vehicle-to-road, vehicle-to-vehicle dynamics.

Many studies have been conducted on the effects of a single factor or a handful of factors on driver behavior and the probability of car accidents. Gershon et al. (2017) found the increased risk of accidents for new drivers by studying 90 adolescents and 131 of their parents.[1] Precht et al. (2017) examined the impact of anger as a cause of accidents. This study looked at 10 minutes of 108 trips from the SHRP 2 dataset and analyzed driver behavior related to anger. Furthermore, Tivesten (2015) and Seo (2004) showed an increased risk of accidents associated with cell phone use in their respective studies [3][4].

Recent studies using naturalistic driving datasets have shown findings that reinforce simulator-based studies. Dingus et al. using the SHRP2 dataset found that driver-related factors, including impairment, error, and distraction, were present in close to 90% of the crash cases; high-emotion driving state, knowledge deficiency, aggressiveness and distraction have strong indication [5]. Wang et al. using the SHRP2 dataset with 324 driving events identified speeding, visual distraction and inclement road condition to be risk factors associated with safety critical events on curvy roadways [6]. Vehicle information was combined with driver in-car activities to predict crash. Victor et al. (2015) concluded that crashes arise from the "perfect storm" of the change rate at which vehicles closing in on each other and the duration of the ego driver's last glance unrelated to driving before an accident [7].

While a large portion of studies on driver behaviors examine a limited number of factors and the impact of those factors on accidents, the main goal of this study is to produce a model that utilizes multiple features captured in the SHRP2 dataset. Using this model, we will be able to examine multiple factors simultaneously and prioritize the features based on their importance in the model.

III. Data Selection

A. The SHRP2 Dataset

This study uses the SHRP2 dataset. The SHRP2 dataset is a relational database containing data collected by VTTI [8]. The study contains information regarding the subjects' preferences and attitudes as well as video and sensor data of all trips driven by the study participants. The data collected consists of four tables. The four tables contain: information about the vehicles included in the study, the trips taken by all drivers in the study, the drivers, and the events (including crashes, near crashes, and baseline trips). The drivers are identified using an anonymous participant ID and the events (crashes or baseline trips) are identified using an event ID. The sample used in this research is of 7707 trips from the SHRP2 dataset. 7167 of those trips were not accidents or near accidents and 540 of those were accidents. The severity of the accidents ranged from a low risk tire strike to a severe accident with injuries. These trips originated from 1100 unique drivers, whose age distribution is shown in Table I. The dataset used in this research contains various pieces of information about the trip and the accident.

Table I Age Distribution

Age Group	Number of Trips
16-19	3561
35-39	914
40-44	918
45-49	1144
50-54	1170
Total	**7707**

B. Variable Selection

Our analysis focused on variables related to road and weather conditions as well as overall driver behavior, age, and number of years driving. The goal was to use these variables to predict accidents. The main motivation is to develop a model that can help predict in real time a potentially safety-critical event based on driver, vehicle, and road conditions. We included every annotated fields based on the in-cabin video capture that are available to both baseline and crash events as presented in Appendix I. They are 30 categorical variables and 3 selected numeric variables. These variables can be divided into driver characteristics, driver behavior and road characteristics. Driver characteristics are driver predispositions that are not related to one individual trip, such as age and years of driving experience. Driver behaviors include both specific driver maneuvers of vehicle that are considered unsafe as well as driver secondary tasks that are non-driving related and may be distracting. Driver impairment such as high-emotion state and drowsiness are also captured. Notably, the SHRP2 dataset provides a high-fidelity description of driver behaviors that were not available in previous studies. For example, there are over 60 different types of secondary tasks that were manually annotated, 10 of which were related to driver's cell-phone use. Chi-square independence tests and Point-Biserial correlation were used to evaluate relationship of categorical and numeric variables with crash outcomes. All categorical variables except weather, alignment, grade, and construction zone were shown to be associated with crash outcome. Years of driving experience are shown to be negatively correlated with crashes, whereas the durations of secondary tasks are positively correlated with crash. Time series data, such as steering wheel angle, brake and throttle pressure, deceleration, velocity, etc., collected from radar and a variety of vehicle sensors is omitted in this study, but will be investigated in the future.

IV. Preliminary Analysis

One of the goals of this study is to determine the importance of secondary tasks in causing car accidents. Secondary tasks are all tasks not related to driving. The SHRP2 dataset contains three variables that describe what secondary tasks were performed by the driver. The secondary tasks are recorded during the five seconds that preceded the event for a crash or near crash. For the baseline events, the secondary tasks are recorded for the last 6 seconds of the baseline epoch. This includes the last 5 seconds prior to the event end and one second after the event end. The dataset contains a total of 59 secondary tasks. Since this will produce a large group of sparse dummy variables in our model, the 59 secondary tasks have been grouped into 8 similar task groups. The first group is labeled no tasks. Approximately a third of all trips containing an incident had no tasks. The second task group is interaction. This task group includes interacting with other passengers or pets, dancing or singing. The third group is external secondary tasks. This group contains tasks like looking at pedestrians, accidents, construction, or any other distraction that is outside the vehicle. The fourth group of tasks is internal tasks, which includes adjusting the mirrors, reading or writing, inserting a CD, or picking up a dropped object. The fifth group is all phone related tasks including texting, holding or talking on a cellular phone. The sixth group is all tasks related to consuming or holding any food or beverage items. The seventh group of tasks is all tasks related to grooming. These tasks include nail biting, adjusting or removing clothing, adjusting glasses, applying make-up, etc. The final group is all tasks manually coded as unknown. When looking at the entire dataset, we can see that secondary tasks are involved in a large portion of trips, including those that did not end in an accident. In fact, over 55% of trips contained at least one secondary task. We can examine the counts of at least one secondary task appearing in a trip. The most common secondary task overall is interaction while the least common secondary task is consuming food or beverage.

While most secondary tasks are infrequent, the aggregate groups tend to be observed as causing an accident with a similar probability. For example, phone usage has one of the highest rates of contributing to an accident while interaction and grooming are the least likely to contribute to an accident. This is illustrated in Table II.

V. Model Selection

In conducting this research, a number of machine learning models were investigated. The selected models are random forest, deep neural network, gradient boosted classifier, and gradient boosted classifier with grid search. These models have been widely used in predictive analytics of many different health conditions. The models were compared for their performance in accuracy, sensitivity and specificity.

TABLE II COUNT OF SECONDARY TASKS BY ACCIDENT

Has an accident occurred? **Secondary Task Group**	**No Accident**	**Accident**	**Total**
No Task	3184	211	3395
Interaction	1860	132	1992
External	803	60	863
Phone	644	73	717
Internal	539	58	597
Grooming	405	32	437
Unknown	302	37	339
Food and Beverage	257	16	273

A. Random Forest

The first model examined in this analysis is a random forest model. Random forests are predictive models that consist of an ensemble of tree classifiers [9]. The goal behind using an ensemble of tree classifiers is to reduce the variance of our model by aggregating a large number of noisy classifiers. To create a random forest model, we draw multiple bootstrap samples, construct a classifier for each sample and then find the prediction using the majority vote of all classification trees.

Random Forest Algorithm

```
1) For k=1...K
   a) Draw a random bootstrap sample from the sample data
   b) Create a tree from the bootstrapped sample using the following steps
      i. Select a subset of variables from the sample to generate a tree
      ii. Pick the optimal values to split for each of the m variables and split the node on the value
   c) Make a prediction using majority vote of the ensemble of trees
```

Random forests have been widely used in predictive modeling throughout the field of health informatics. Examples include the prediction of in-hospital mortality in emergency department patients with sepsis [10], prediction of breast cancer diagnosis and prognosis [11], and for the detection and prediction of Alzheimer’s disease using MRI imaging [12].

B. Deep Neural Networks

A deep neural network model (also known as a deep learning model) is a prediction model [13] comprised of a number of layers. At each layer, we perform a transformation followed by an activation function that acts as the decision maker for the layer. The output of each layer is passed in as the input to the next layer. Deep neural networks have improved our prediction capabilities in many important applications like image detection and speech recognition. In this paper, we will be using the Keras API to implement a deep learning model to our data.

C. Gradient Boosted Classifier

Gradient boosted classifiers are another type of ensemble machine learning model. This methodology has been introduced by Friedman [14] and proposes an improvement to ensemble models. Typically, ensemble methods combine classifiers by taking an average of the results of all classifiers. Gradient boosted classifiers combine “weak” learners using a technique called boosting. The “weak” learners are classifiers that produce a prediction slightly better than a random guess. Typically, each classifier contains a set of randomly selected vectors ($Y = y_1, y_2, \dots, y_m$) and randomly selected observations ($X = x_1, x_2, \dots, x_n$) and produces a prediction using the sampled data. We combine the classifiers using a boosting algorithm. This means that we aggregate the learners by assigning a weight to each one. We iteratively optimize the weights until we achieve an optimal classification algorithm. As shown in Friedman (2001), our goal is to find an approximation F(x) of the function F*(x) that minimized the loss function L over all values of x and y.

$$F^* = \underset{F}{\operatorname{argmin}}\, E_{y,X} L\big(y, F(x)\big) \tag{1}$$

To do this, we generate an additive model that is a weighted combination of our classifiers (or “weak” learners). This will produce a boosted model.

$$F(x;\{\beta_m,\boldsymbol{a}_m\}_1^M)=\sum_{m=1}^{M}\beta_m h(x;\boldsymbol{a}_m) \tag{2}$$

The function $h(x;\boldsymbol{a})$ is typically a learner with parameters $a_1, \ldots, a_{\mathrm{M}}$. The vector $\beta_1, \ldots, \beta_{\mathrm{M}}$ contains the weights that are generated by the boosting algorithm.

D. Gradient Boosted Classification with Grid Search

We can further improve our gradient boosted model by optimizing the model parameters with grid search [15]. Grid search is a form of hyperparameter tuning. The gradient boosted model has a number of parameters and we examine all possible combinations of these parameters. The model selected is one that optimizes all metrics.

E. Model Selection

After examining all four models, a gradient boosted model with grid search was chosen since it has the highest sensitivity. Since the data contains more baseline observations than accidents, it is crucial to produce a model that performs well in classifying both accidents and baseline observations. Although the deep neural network achieved equal sensitivity, the gradient boosted model has the advantage of interpretability in that we are able to evaluate feature importance. Gradient boosting builds classification trees iteratively. Each tree relies on the classification error from the previous tree. Therefore, the algorithm continues to adjust iteratively to reduce the error. As a result, there is a significant improvement in the correct classification of accidents. An attempt was made to classify the data using a random forest algorithm and a neural network. Random forests have been explored in the study of accidents [16] [17]. However, our study has found them to perform less optimally than the gradient boosted model. Similarly, the deep learning model also did not perform as well as the gradient boosted model specifically in its sensitivity. While the non-accidents were classified correctly at a rate of over 95% for all models, accidents were classified correctly between 69-77% of the time for random forest, deep learning and the gradient boosted model without gird search. It seems that both random forests and neural networks were picking up on the general trend that most of the data was from the non-accident baseline. Despite numerous attempts to optimize the algorithms, we still could not beat the correct classification rate for accidents in gradient boosting. We can see this result in Table III.

VI. Analysis

A. The model

In this study, a gradient boosted model was fitted to the data. Since approximately 93% of the observations in our data are non-accidents, creating the model with the entire dataset may produce a model that will predict all observations to be non-accidents and still have a very high accuracy score. Therefore, the sample has been down-boosted. The proportion of non-accidents in the sample used to generate the model is 76.84%. This model produces a much-improved prediction rate for accidents. The data is then split into training and test datasets. The model is produced using the training data and all metrics are generated using the test data. Additionally, the parameters of gradient boosted algorithm are optimized using a grid search algorithm. Selecting the max features per tree and the minimum samples per leaf using a grid search further optimizes the performance of the algorithm.

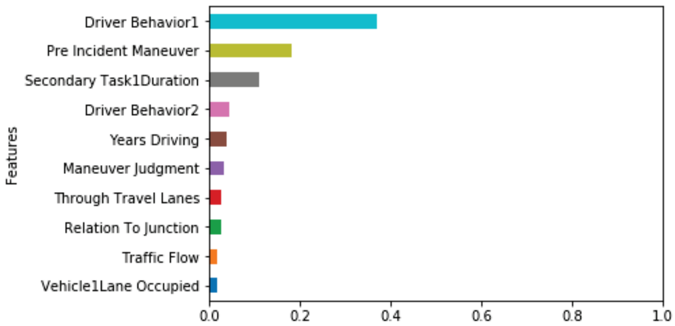


Fig. 2. Feature importance in the gradient boosted model

B. Feature Importance

A useful piece of information that we can extract from the model is the feature importance. The feature importance is a score between zero and one that tells us how useful the feature was in constructing the model. In random forests and gradient boosted tree models, we measure the importance by the improvement in the model at each split in each tree, the improvement is aggregated over all trees in the model [9]. In the gradient boosted model, we observe that the most important features are driver behavior, pre-incident maneuvers, secondary task duration, and number of years driving. (Fig. 2)

Table III Model Performance Metrics

	Accuracy	Sensitivity	Specificity
Random Forest	0.906	0.6923	0.9835
Deep Neural Network	0.9657	0.7692	0.978
Gradient Boosting	0.9188	0.7692	0.956
Gradient Boosting with Grid Search	0.9274	0.8076	0.9505

Driver behavior is a categorical variable that describes the driver's overall demeanor while driving. The driver could be driving safely, or act distracted. The driver could be performing improper turns or driving too fast for road conditions. These are all captured in this variable. Secondary tasks are all tasks the driver performs that are not directly related to driving. The most common tasks are interaction with other passengers in the vehicle and cell phone usage. The SHRP2 data records up to 3 secondary tasks and up to 3 driver behaviors.

Secondary tasks are a type of distracted driving. Therefore, they are captured both in general in the driver behavior variables and in the secondary task variables. While this may point to a relationship between variables, this does not pose a problem when using ensemble methods like gradient boosting or random forests. These techniques are more robust, and their results are not affected by correlated variables. This will simply cause the model to have redundant information but will not affect the results.

C. *Model Accuracy*

When plotting the ROC curve for this model, we see that the area under the curve is 0.9433. (Fig. 3) This is a good estimate for a relatively small sample of data. We also withhold a validation set to ensure there is no overfitting.

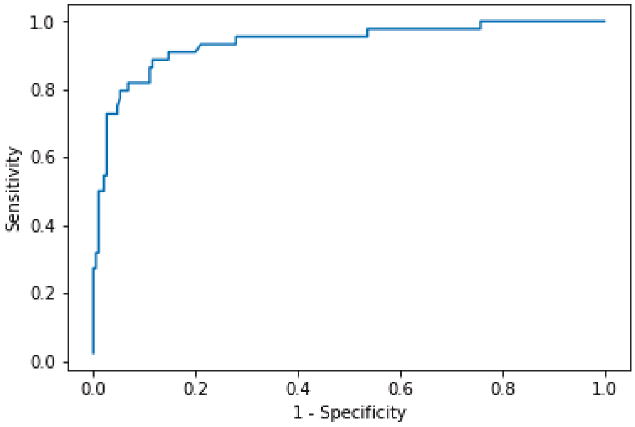


Fig. 3 The ROC curve of the gradient boosted model

VII. Discussion of Features

Our model allows us to uncover interesting insights regarding the importance of certain features in determining the probability of a car accident.

A. Driver Behavior

According to the model, the most important factor in causing car accidents is driver behavior. According to the SHRP2 data dictionary, driver behavior is described as actions "that include what the driver did to cause or contribute to the crash or near-crash" [8]. There is a total of 3 variables that describe driver behavior. Each of the three is a categorical variable and all three have the identical categories. Since the data is manually coded, the person coding the data can select up to three behaviors that a driver exhibits during the trip. The behaviors exhibited by the driver are typically apparent in the video. Examples of driver behavior are aggressive driving, drowsiness, exceeding the speed limit, etc. These behaviors are recorded in the non-accident baseline trips as well. Not only is driver behavior the most important factor in determining whether an accident will occur, we also see that the 4th most important factor is the second driver behavior. This means that displaying more than one behavior increases our ability to predict an accident even further. There are several studies that support this finding and show that driver behavior is a significant factor in causing accidents. One example is the study by Dingus et al. showing the prevalence of driver related factors in accidents [18].

B. *Pre-Incident Maneuver*

This variable describes the type of action or maneuver that the driver was engaged in prior to the accident. If this is a baseline trip, then this variable describes the maneuver prior to the end of the recorded period [8]. There are 21 possible values in this categorical variable. These values describe whether the driver was driving straight or turning, whether they were reversing out of a parking spot, or making a U-turn or another maneuver. This variable is coded by a human watching the video of the trip. A study by Mitra et al. [19] supports this finding. Though this study focuses primarily on accidents in intersections, it shows that the type of maneuver has an impact on the occurrence of a crash. Our study looked at crashes in all driving situations including in intersections. Another study by Box [20] indicates that entering a parking position is a significant cause of accidents. Table IV shows that the most common pre-incident maneuver that ended in an accident is a right turn followed by entering a parking position and going straight at a constant speed. What this table shows is that some maneuvers seem to appear proportionately in both the accident and non-accident trips (like going straight, constant speed) while others are more highly represented in the accident group (like turning right or entering a parking position).

C. Secondary Task Duration

Secondary tasks are defined in the SHRP2 data dictionary as any distractions that are not driving related such as talking, singing, eating, or looking at one's cellular phone [8].

Table IV Counts of the top 5 pre-incident maneuvers for accident and baseline trips

Has an accident occurred? **Pre-Incident Maneuver**	**Accident**	**No Accident**
Turning right	143	149
Entering a parking position	68	66
Going straight, constant speed	61	3680
Negotiating a curve	55	825
Decelerating in traffic lane	46	1211

There are three secondary tasks categorical variables that all contain the same categories. Due to the nature of the data, the person coding the data could record up to three tasks per driver, though most drivers engaging in a task only performed one task. About two thirds of drivers involved in an accident performed at least one task. Secondary tasks have been studied as a factor causing accidents in a number of studies. One example is a study by Young et al. examining the involvement of secondary tasks in crashes [21]. During the manual coding, a determination was made whether the secondary task was involved in causing the accident. We can see that some tasks are more likely to cause accidents than others. For example, grooming is very unlikely to cause accidents. On the other hand, in more than half of accidents that involved cellular phone usage, the phone was determined to contribute to the accident. Here we see that specifically the duration of the secondary task has an impact on the model's decision of whether the trip should be classified as an accident. In some secondary tasks, the duration of the task is much longer when an accident occurred. In food and beverage related tasks as well as phone related tasks, the mean duration of secondary task in an accident is significantly longer. However, in other tasks, the length is very similar. In internal related tasks, we see a similar secondary task duration. This data illustrates that accident prediction requires a combination of variables. These results are described in Table V. One limitation of this variable is that the maximum value of secondary task duration for all non-accident trips were coded to be six seconds, whereas trips with accidents may have secondary tasks that are coded to be longer than six seconds.

D. Number of Years Driving

The third most important factor is the number of years the driver has had a license. This is a factor that has been researched in the past by many others. One such study is the study by Gershon et al. that examined teen drivers [22]. We can also intuitively assume that inexperienced drivers are more prone to accidents.

In summary, we can group the features in our model into 3 main groups: environmental factors, driver behavior related factors, and driver related factors. Five of the top 10 features ranked by importance are related to driver behavior. These features are Driver Behavior 1, Pre-Incident Maneuver, Secondary Task 1 Duration, Driver Behavior 2, and Maneuver Judgement. This leads us to conclude that while environmental factors and road conditions have some impact, the most important factors in predicting an accident are related to the driver's behavior and judgement.

E. Future Work

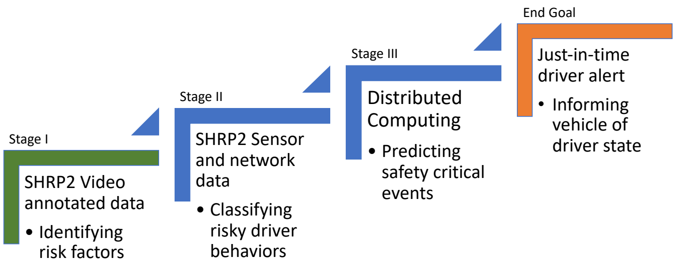
This work highlights the importance of identifying risky driver behaviors in crash predication and motivates direction of our further work, which is to classify risky driver behaviors using on-board sensor data. The SHRP 2 allows identification of these behaviors by video annotation. However, manual video annotation is impractical for real-life risk detection; computer vision algorithms, while improving by day, are highly situation-specific and not appropriate for handling scenarios such as aggressive driving or high-emotion driver state. In real-world driving, video cameras are not standard on-board technology due to manufacturing and privacy concerns. Therefore, we seek opportunities to link risky driver maneuver and in-cabin activities with built-in sensor data from the vehicle in the absence of video capability. For example, vehicle acceleration and yaw rate can be used to detect severe maneuvers and steering controls. Several signals may be linked to driver state of control and distraction, such as positions of accelerator pedal, brake pedal and steering wheel, turn signals, and driver head rotations. Activation of anti-lock braking system (ABS) and electronic stability control (ESC) are indicative of emergency braking and roadway departure. GPS-based signals as well as vehicle-based machine vision provide information on road sections. Vehicle-to-vehicle dynamics can be extracted from radar sensor data.

TABLE V THE PROBABILITY OF A SECONDARY TASK BEING INVOLVED IN AN ACCIDENT AS WELL AS MEAN SECONDARY TASK DURATION

Was the accident caused by the task? **Secondary Task Group**	**No**	**Not applicable or Unknown**	**Yes**	**Mean Secondary Task Duration for Baseline (in seconds)**	**Mean Secondary Task Duration for Accidents (in seconds)**
No Task	0	1	0	0	0
Interaction	0.76	0.08	0.17	5.34	7.07
External	0.51	0	0.49	2.15	2.79
Phone	0.45	0.03	0.52	5.67	7.59
Internal	0.74	0	0.26	3.1	3.89
Grooming	0.86	0	0.13	4.6	5.53
Food and Beverage	0.62	0.08	0.31	5.64	7.99
Unknown	0.56	0.19	0.26	1.98	2.92

A roadmap of our current and future work is seen in Fig. 4. Our current work of identifying high-risk driver behaviors based on video-annotated data is discussed in this paper. We dedicate ongoing efforts to Stage II and Stage III, which deal with sensor-based time series data of significant size and variability. In Stage II, we generate hypotheses and learning models that are able to classify risky driver behaviors by detecting hidden layers of on-board sensor data. In Stage III, we augment the learned models on the entire dataset of SHRP 2 (approximately 5.5 million observations). Analytic tasks such as determining similarity of time-series fragments and graph mining of heterogeneous networks require significant computing resources. We establish an experimental cluster based on Spark and hardware optimization to achieve data parallelism and distributed computing. Our aspiration is to implement the

Fig. 4. Our analytics roadmap to reduce road injuries


Stage I
SHRP2 Video annotated data
• Identifying risk factors
Stage II
SHRP2 Sensor and network data
• Classifying risky driver behaviors
Stage III
Distributed Computing
• Predicting safety critical events
End Goal
Just-in-time driver alert
• Informing vehicle of driver state

optimized algorithm as part of the on-board advanced driver assistance systems so that vehicles can be informed of driver state and provide just-in-time driver support.

VIII. Conclusion

The main goal of this study is to uncover the primary causes of car accidents. In this study, a gradient boosted decision tree model has been generated using decision trees as "weak" learners. This methodology outperforms other models in the accuracy of predicting both crashes and non-crash trips. The machine learning method selected for this study produces a model that allows us to rank the importance of each factor in causing car accidents.

Our results prove that this modeling technique can predict a general accident risk probability based on road conditions and some information regarding the driver and their behavior during the trip. These are pieces of information that can be collected and used to help drivers drive more carefully. For example, this information can be used to calibrate the sensitivity of driver assistance systems in newer vehicles.

The analysis in this study contains many variables that are hand coded. While the technology to infer this information from the video using machine learning currently exists, our goal in subsequent research is to uncover the relationship between these hand coded variables and the sensor data emitted from the vehicle. This will provide us with a solution for integrating with on-board signals and detecting risky behaviors in real time that are more likely to cause an accident. Borrowing from the Internet of Things (IoT) framework, the algorithms will enable machine awareness of its driver. The algorithms will provide utility not only to drivers in the form of alerts, but also to advanced driver assistance systems as signals to activate protection mechanisms.

Acknowledgement

This work was supported in part by the National Science Foundation under the Grant NSF-1741306, IIS-1650531, and DIBBS -1443019. Any opinions, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] L. J. Blincoe, T. R. Miller, E. Zaloshnja, and B. A. Lawrence, "The Economic and Societal Impact of Motor Vehicle Crashes, 2010 (Revised)," National Highway Traffic Safety Administration, Washington, DC, DOT HS 812 013, May 2015.

[2] K. L. Campbell, "The SHRP 2 Naturalistic Driving Study," *TR News*, vol. 282, pp. 30–35, Sep-2012.

[3] E. Tivesten and M. Dozza, "Driving context influences drivers' decision to engage in visual–manual phone tasks: Evidence from a naturalistic driving study," *J. Safety Res.*, vol. 53, pp. 87–96, Jun. 2015.

[4] D.-C. Seo and M. R. Torabi, "The impact of in-vehicle cell-phone use on accidents or near-accidents among college students," *J. Am. Coll. Health J ACH*, vol. 53, no. 3, pp. 101–107, Dec. 2004.

[5] T. A. Dingus *et al.*, "Driver crash risk factors and prevalence evaluation using naturalistic driving data," *Proc. Natl. Acad. Sci.*, vol. 113, no. 10, pp. 2636–2641, Mar. 2016.

[6] B. Wang, S. Hallmark, P. Savolainen, and J. Dong, "Crashes and near-crashes on horizontal curves along rural two-lane highways: Analysis of naturalistic driving data," *J. Safety Res.*, vol. 63, pp. 163–169, Dec. 2017.

[7] T. Victor *et al.*, "Analysis of Naturalistic Driving Study Data: Safer Glances, Driver Inattention, and Crash Risk," *SHRP 2 Rep.*, no. Report S2-S08A-RW-1, 2015.

[8] J. M. Hankey, M. A. Perez, and J. A. McClafferty, "Description of the SHRP 2 naturalistic database and the crash, near-crash, and baseline data sets task report," Virginia Tech Transportation Institute, Blacksburg, VA, Apr. 2016.

[9] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, 2nd edition. New York, NY: Springer, 2016.

[10] R. A. Taylor *et al.*, "Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data-Driven, Machine Learning Approach," *Acad. Emerg. Med.*, vol. 23, no. 3, pp. 269–278, Mar. 2016.

[11] C. Nguyen, Y. Wang, and H. N. Nguyen, "Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic," *J. Biomed. Sci. Eng.*, vol. 06, no. 05, pp. 551–560, 2013.

[12] A. V. Lebedev *et al.*, "Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness," *NeuroImage Clin.*, vol. 6, pp. 115–125, Jan. 2014.

[13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[14] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[15] M. Claesen and B. De Moor, “Hyperparameter Search in Machine Learning,” *ArXiv150202127 Cs Stat*, Feb. 2015.

[16] R. Harb, X. Yan, E. Radwan, and X. Su, “Exploring precrash maneuvers using classification trees and random forests,” *Accid. Anal. Prev.*, vol. 41, no. 1, pp. 98–107, Jan. 2009.

[17] S. Krishnaveni and M. Hemalatha, “A Perspective Analysis of Traffic Accident using Data Mining Techniques,” *Int. J. Comput. Appl.*, vol. 23, no. 7, pp. 40–48, Jun. 2011.

[18] T. A. Dingus *et al.*, “Driver crash risk factors and prevalence evaluation using naturalistic driving data,” *Proc. Natl. Acad. Sci.*, vol. 113, no. 10, pp. 2636–2641, Mar. 2016.

[19] S. Mitra, H. C. Chin, and M. A. Quddus, “Study of Intersection Accidents by Maneuver Type,” *Transp. Res. Rec. J. Transp. Res. Board*, vol. 1784, no. 1, pp. 43–50, Jan. 2002.

[20] P. C. Box, “Curb-Parking Problems: Overview,” *J. Transporation Enginnering*, vol. 130, no. 1, pp. 1–5, Jan. 2004.

[21] R. Young, “Removing Biases from Crash Odds Ratio Estimates of Secondary Tasks: A New Analysis of the SHRP 2 Naturalistic Driving Study Data,” presented at the WCX^{TM} 17: SAE World Congress Experience, 2017.

[22] P. Gershon *et al.*, “Crash Risk and Risky Driving Behavior Among Adolescents During Learner and Independent Driving Periods,” *J. Adolesc. Health*, vol. 63, no. 5, pp. 568–574, Nov. 2018.

APPENDIX I LIST OF MODEL VARIABLES AND THEIR DESCRIPTION

Categorial Variable	**Description**	**Source**	**Chi-square Statistics**	**p-value**
ageGroup	Driver age groups	Driver Table	89.4107	0
alignment	Roadway curvature at the time of event capturing	Event Table	0.553	0.7584
constructionZone	Whether vehicle was in or approaching a construction zone at the time of event capturing	Event Table	2.8348	0.5858
contigTravelLanes	Total number of contiguous travel lanes	Event Table	277.8812	0
driverBehavior1	Driver behaviors that contributed or could contribute to an accident in the order of criticalness.	Event Table	2863.0955	0
driverBehavior2	Driver behaviors that contributed or could contribute to an accident in the order of criticalness.	Event Table	976.1454	0
driverBehavior3	Driver behaviors that contributed or could contribute to an accident in the order of criticalness.	Event Table	282.2853	0
driverImpairments	Possible causes that impair driver judgment or driving ability	Event Table	163.3685	0
driverSeatbeltUse	Driver's use of seatbelt	Event Table	59.1721	0
grade	Roadway uphill/downhill	Event Table	3.6249	0.4591
handsOnTheWheel	Number and side of hands placed on the steering wheel	Event Table	32.5187	0.0002
intersectionInfluence	Influence of intersections on subject vehicle's movement	Event Table	433.7971	0
lighting	Lighting condition at the time of event capturing	Event Table	16.2976	0.0026
locality	Type of vehicle surroundings such as business/school/residential	Event Table	122.5916	0
maneuverJudgment	Whether driver maneuver of vehicle is safe and legal	Event Table	327.0328	0
preIncidentManeuver	Driver's last maneuver of vehicle before an event capturing	Event Table	2043.8206	0
relationToJunction	Relation of subject vehicle to a junction where 2+ roadways meet	Event Table	329.8020	0
secondaryTask1	Driver's in-cabin activity not related to driving in chronological order	Event Table	186.5126	0
secondaryTask2	Driver's in-cabin activity not related to driving in chronological order	Event Table	135.7452	0
secondaryTask3	Driver's in-cabin activity not related to driving in chronological order	Event Table	69.6335	0.0006
secondaryGroup1	Driver's first secondary task grouped into 8 types	Derived from Event Table	24.5994	0.0009
secondaryGroup2	Driver's second secondary task grouped into 8 types	Derived from Event Table	49.9279	0
secondaryGroup3	Driver's third secondary task grouped into 8 types	Derived from Event Table	20.2711	0.005
surfaceCondition	Roadway surface condition affecting vehicle friction at the time of event capturing	Event Table	44.6632	0
throughTravelLanes	Number of through travel lanes in the subject vehicle direction	Event Table	350.7616	0
trafficControl	Type of traffic control at the time of event capturing	Event Table	118.4419	0
trafficDensity	Level of traffic density based on manual analysis	Event Table	62.1712	0
trafficFlow	Roadway design	Event Table	298.8629	0
vehicle1LaneOccupied	Lane in which the subject vehicle occupied at the time of event capturing	Event Table	357.3775	0
weather	Weather condition	Event Table	7.4527	0.3833
Numeric Variable	**Description**	**Description**	**Correlation Coefficient**	**p-value**
years_driving	Number of years driving	Driver Table	-0.1043	0
secondaryTask1Duration	Duration of the first secondary task	Derived from Event Table	0.0886	0
secondaryTask2Duration	Duration of the second secondary task	Derived from Event Table	0.0872	0
secondaryTask3Duration	Duration of the third secondary task	Derived from Event Table	0.0357	0.0017