# A Robust Environment-Aware Driver Profiling Framework Using Ensemble Supervised Learning

Abdalla Abdelrahman, *Student Member, IEEE*, Hossam S. Hassanein, *Fellow, IEEE*, and Najah Abu Ali, *Senior Member, IEEE*

***Abstract*— Driver profiling is the real-time process of detecting driving behaviors and computing a driver's expected risk based on detected behaviors. Predicting risk based solely on the inclusion of detected behaviors may not be accurate because this method of predicting ignores the environmental (e.g., weather conditions, traffic density level) context of detected behaviors. Moreover, coupling detected behaviors with their environmental context can be leveraged towards creating personalized risk profiles for drivers in each driving environment. These profiles can be utilized in various ITS applications including personalized safety-based route planning. In this paper, a novel driver profiling environment-aware framework is presented. In the proposed framework, data processing is distributed over three computational layers to enhance the overall reliability of the system. A risk prediction model is hosted on the edge/fog to determine the driving risk while considering the joint effect of the in-vehicle detected behaviors and their environmental context. Risk values along with a driver's compliance to warnings are both utilized to compute the risk profile on the cloud. Using SHRP2 Naturalistic Driving (ND) dataset, the development of a novel risk prediction model is presented herein with the underlying sub-processes of data preprocessing, error analysis, and model selection. Then we analyze both the performance of the developed risk prediction model and the overall performance of the proposed system. Validation results for the developed model indicate a good compromise between bias and variance. Moreover, the results of the overall risk scoring model reflect its robustness and reliability in assigning accurate risk scores.**

***Index Terms*— Driver profiling, intelligent transportation systems (ITS), driving behavior classification, supervised learning, random forests, telematics.**

Manuscript received 7 December 2019; revised 5 March 2021 and 20 August 2021; accepted 17 November 2021. Date of publication 3 December 2021; date of current version 12 September 2022. This work was supported in part by the Roadway Transportation and Traffic Safety Research Center at United Arab Emirates University under Project 31R014-Research Center RTTSRC-4-2013 and in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant STPGP 479248. The Associate Editor for this article was S. S. Nedevschi. *(Corresponding author: Abdalla Abdelrahman.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the General Research Ethics Board (GREB) at Queen's University.

Abdalla Abdelrahman is with the School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON K1N 6N5, Canada (e-mail: aabdelr4@uottawa.ca).

Hossam S. Hassanein is with the School of Computing, Queen's University, Kingston, ON K7L 2N8, Canada (e-mail: hossam@cs.queensu.ca).

Najah Abu Ali is with the College of Information Technology, United Arab Emirates University, Al-Ain, United Arab Emirates (e-mail: najah@uaeu.ac.ae).

Digital Object Identifier 10.1109/TITS.2021.3129506

## I. Introduction

THE recent advancements in vehicular sensing, cellular communications, as well as cloud computing have enabled the deployment of various Intelligent Transportation System (ITS) applications. Given the high vehicle crash rates [1], these ITS applications are promising to lower these rates considerably.

An emerging safety-based ITS application is driver behavior profiling [2]. Driver profiling is the process of acquiring real-time vehicular data using CAN-bus through On-Board Diagnostics II (OBD II) units or mobile-sensed data using inertial smartphone sensors to detect behaviors and warn drivers if risky behaviors are detected. Driver profiling has been widely deployed in different safety-based applications. Pay-How-You-Drive (PHYD) is an example of car telematics insurance scheme in which an insurance premium is rated according to a driver's per-trip driving score [3]. Other emerging driver profiling applications include fleet telematics profiling systems [4], safety-based route planning, and driver self-coaching systems [5].

Most of the literature in the context of driver profiling research has been focused on the detection of certain behaviors which are considered risky. Detected behaviors are then inputted into scoring functions that assign different weights to these detected behaviors based on the expected risk of each [6]. Not only are such scoring functions subjective due to the absence of a valid risk measure (i.e., a risk measure quantified in terms of the actual risky events such as crash and near-crash), but also they ignored the environmental (e.g., weather and road conditions, traffic density level, etc.) effect on risk given the detected behaviors. For instance, an aggressive lane change in a highly dense driving environment could impose more risk than performing the same behavior in less dense traffic conditions. However, current profiling systems would equally penalize the subject driver in both scenarios regardless of where the behavior occurred since these systems only consider the behavior detection process [2], [7]–[10].

Among the wide range of driving data collection methods, naturalistic driving studies (NDSs) have lately predominated in the field of driving behavior analysis [11]–[13]. Unlike controlled experimental approaches, naturalistic driving studies (NDSs) have provided large-scale data about behavioral causes of risky events (i.e., crashes and near-crashes), as well

as the environmental context of such behaviors (e.g., weather and road conditions, traffic density level, etc.). In addition, NDSs provide the same behavioral and environmental information during normal driving episodes, which enables the development of environmental-aware statistically significant risk prediction models [14]. Recently, the Virginia Tech Transportation Institute (VTTI) conducted the largest NDS to date named Strategic Highway Research Program II Naturalistic Driving Studies (SHRP 2 NDS) [14]. This dataset contains the behavioral and environmental contextual information of nearly 9,000 crash and near-crash events and more than 20,000 baseline events captured during normal driving episodes.

The research question this paper addresses is:

*Are driving behavioral habits together with their environmental context good predictors for measuring risk probability?*

To answer this question, the behavioral and environmental details of driving events presented in SHRP2 NDS are utilized to build a risk prediction model that can be incorporated in a complete cloud-based environment-aware driver profiling framework. The research contributions of this paper are:

1) A novel cloud-based environment-aware driver profiling (CEDP) system is presented and discussed. The system provides a view on a "*next generation*" driver profiling system in which drivers are profiled based on the expected risk of their environmentally stamped driving behaviors and their compliance to warnings. The risk notion is mathematically developed and the terms: behavior detection, driving risk probability, driver scoring, and driver profiling, that are used interchangeably in literature, are clearly distinguished and mathematically defined.
2) An ensemble supervised machine learning algorithm based on randomized trees is selected and customized to reflect the predicted driving risk probability while jointly considering the detected behaviors and their environmental context. The model is proven to provide an acceptable compromise between bias and variance. The developed risk prediction model is trained and validated using an unprecedented amount of real driving data from SHRP2 NDS. This enhances the reliability and the practicability of the proposed system which is reflected in the performance results.
3) Given predicted risk probabilities, the performance of the overall risk scoring system is validated. Validation results show the robustness of the proposed system as it consistently provides accurate results over different training and validation samples.

The remainder of this paper is structured as follows. In section II, we provide background information on the existing driver profiling systems and on the driving dataset used in this work. Section III provides a detailed description of the envisioned CEDP system covering its in-vehicle, on edge/fog, and on cloud data processing. In section IV, the adopted preprocessing, error analysis and model selection processes for the risk prediction problem are described. In section V, results are presented and analyzed. An illustrative example of the trip scoring process using the proposed framework is discussed in section VI. Our conclusions are presented in section VIII.

## II. Background and Related Work

### A. Driving Behavior Profiling

In literature, the term "driver behavior profiling" has been used to describe different behavioral characterization processes, which may have caused some confusion. We found that some of the literature has used "driver profiling" interchangeably with "behavior classification or detection". Although behavior classification is the building block in the driver profiling hierarchy, other processes such as risk scoring and profiling are as important as behavior classification. A complete profiling system that includes behavior detection, risk scoring, and profiling is still very primitively presented in the literature to date.

In the context of behavior detection, authors in [15] utilized variations of Recurrent Neural Network (RNN) models to detect seven distinct types of behaviors using smartphone sensors. In [16], authors utilized the CAN Bus data to train a Long Short Term Memory (LSTM) and a one-dimensional Convolutional Neural Network (CNN) classifiers that are able to accurately distinguish between normal and aggressive driving behaviors. In [17], vehicular communication data between connected vehicles was utilized to model the behavior of drivers using unsupervised clustering techniques. Similarly, authors in [18] developed a two-step clustering algorithm applied on a large-scale number of driving events to determine different driving styles. Four distinct driving styles - normal, aggressive, calm, and experienced styles - were inferred from the utilized data. A dynamic Bayesian network was developed in [19] to classify the acceleration, braking and cornering behaviors of drivers using only GPS data. In [20], a risk prediction model was developed using elastic net regularized multinomial logistic regression along with SHRP2 dataset. It was shown that the model can be customized for each individual driver by incorporating driver specific variables. Authors in [10] used static supervised machine learning techniques such as Support Vector Machines (SVMs) and Artificial Neural Networks (ANNs) to detect certain driving maneuvers. Likewise, in [21], authors proposed a sequence modelling HMM-based classifier to classify aggressive and normal driving maneuvers in both forward and lateral directions based on smartphone sensory data with a classification accuracy of 95 %. Similar work is presented in [22] in which five HMM models were trained to infer the fault contribution of the subject driver in two types of driving conflicts. A Semi-supervised machine learning approach was introduced in [23] to detect distraction. The proposed approach utilized unlabeled training data to improve detection performance at a little cost. Computer vision techniques were also explored to detect certain driving behaviors. For instance, authors in [24] developed a 3D convolutional neural network to automatically extract driving behaviors from videos.

Driver profiling is the process of augmenting different driving behaviors, over several driving trips, into a scoring function to measure a driver's overall competence level.

Notable research in this context is the work presented in [2]. Authors in this paper have made a clear distinction between behavior detection and driver profiling. Using a fuzzy logic algorithm hosted by a smartphone application, four unique driving events: harsh braking, aggressive acceleration, speeding, and aggressive steering were accurately detected by acquiring a smartphone's accelerometer, gravity, magnetic, and GPS data. A scoring function was then introduced to reflect the overall driving trip score given the detected behaviors. Despite the proposals and findings of the paper, the scoring function was very primitive, since it did not reflect the statistical correlation between actual risk and detected behaviors. Moreover, it did not show how to find an overall driving profile as a function of many trips. In other words, it did not elaborate on how the individual trip scores will be used towards building a driver's profile.

Recently, a more rigorous work has been presented in [25]. Authors in this paper have presented a data-driven scoring system using SHRP2 NDS. The behavioral information of a very large number of driving events, as well as total driving time were used to predict driving risk using supervised machine learning algorithms. A driving score was then formulated as an additive inverse of the predicted driving risk probability.

To the best of our knowledge, no work in the literature has comprehensively considered a complete and detailed driver behavior profiling system that considers the sub-processes of behavior detection, risk prediction, driver's behavior scoring and profiling, and with consideration given to the driving environment. Although the environmental effect on risk has been comprehensively researched in the literature, the joint effect of driving behaviors and their environmental context on driving risk is presented in very few works, and not in the context of driver profiling [20]. In [26], authors performed a statistical retrospective cohort study on the effect of traffic and road conditions on driving risk using the 100-CAR NDS. Authors in [27] used an NDS containing 1670 near-crash events to study the factors that are proportional to the increase in near-crash risk. They found that the road condition is one of the significant factors that affect driving risk.

In this paper, an envisioned data-driven driver profiling system is introduced and discussed. We specifically targeted the problem of driving risk prediction by utilizing behavioral and environmental data of a large scale NDS (i.e., SHRP2). The development of the risk prediction model is based on an error analysis of different supervised machine learning models to achieve the best bias-variance trade-off. The overall risk scoring system is then validated.

### B. Dataset and Methodology

*1) Participants:* More than 3,000 drivers were recruited in six sites across the United States. Drivers were originally chosen equally across different genders (i.e., males and females) and 16 age groups. Automobiles of recruited drivers were equipped with inconspicuous data acquisition systems (DASs). Among many sensors, DAS mainly comprised four video cameras to capture the road forward and rearward views as well as the driver's face view, accelerometers, Geographic Positioning system (GPS), and an OBD unit to obtain the vehicle network information. Participants were then asked to use their vehicles over extended time periods (at least a year) as they drive in their normal driving routines.

*2) Data Source:* We utilized the SHRP2 event dataset [28], which is the largest NDS to date. Raw data contains the detailed information of more than 29,000 driving events. Detailed information includes behaviors that are apparent within seconds before risky events or during captured normal driving episodes. Behaviors in the context of this work are different than the *in-vehicle* distractions. They are *vehicle-kinematic* observations that can be noticed from outside the vehicle such as aggressive driving and speeding. In addition to driving behaviors, SHRP2 NDS has the environmental contextual information at which these behaviors happened. Environmental information can be categorized into three types:

1) Static: This refers to long-term environmental features, such as road curvature, number of lanes, traffic flow direction, etc.
2) Quasi-Static: Environmental features that slowly change over a course of time. Road lighting is an example.
3) Dynamic: This refers to the environmental features that rapidly change over a course of time. It includes features such as traffic density.

A driving event in SHRP2 is one of the following types which are mentioned in [14]:

1) Crash: Any contact that the subject vehicle makes with an object, a vehicle, a pedestrian, a cyclist, or an animal either moving or fixed. Also includes inadvertent departures of the roadway.
2) Near-Crash: Any driving conflict that requires an evasive action to avoid a crash.
3) Crash-Relevant: Any driving conflict that requires a non-rapid evasive maneuver.
4) Non-subject Conflict: Any risky event, captured on video but does not involve the subject vehicle.
5) Balanced Baseline Events: Epochs of data selected to provide exposure information. They are 21 seconds long and their frequency is proportional to the total driving time for each driver.

*3) Analysis:* In this work, behaviors and the three aforementioned environmental categories are used as predictors to predict driving risk, quantified herein in terms of the probability of crash, near-crash, or crash relevant events. The mathematical definition of the prediction outcome is shown in section III in equations 7, 8, and 9. Different candidate ML algorithms were first selected and tested. A customized random forest algorithm was shown to provide the best prediction performance. Full analysis details are provided in section IV.

## III. Proposed Driver Profiling Framework

In this section, the proposed cloud-based environment-aware driver profiling framework is discussed. We cover the details of the complete driver profiling system, from the in-vehicle data acquisition to the cloud-based profiling. In short, acquired in-vehicle data is utilized to detect different driving behaviors. Detected behaviors are leveraged along with the environmental

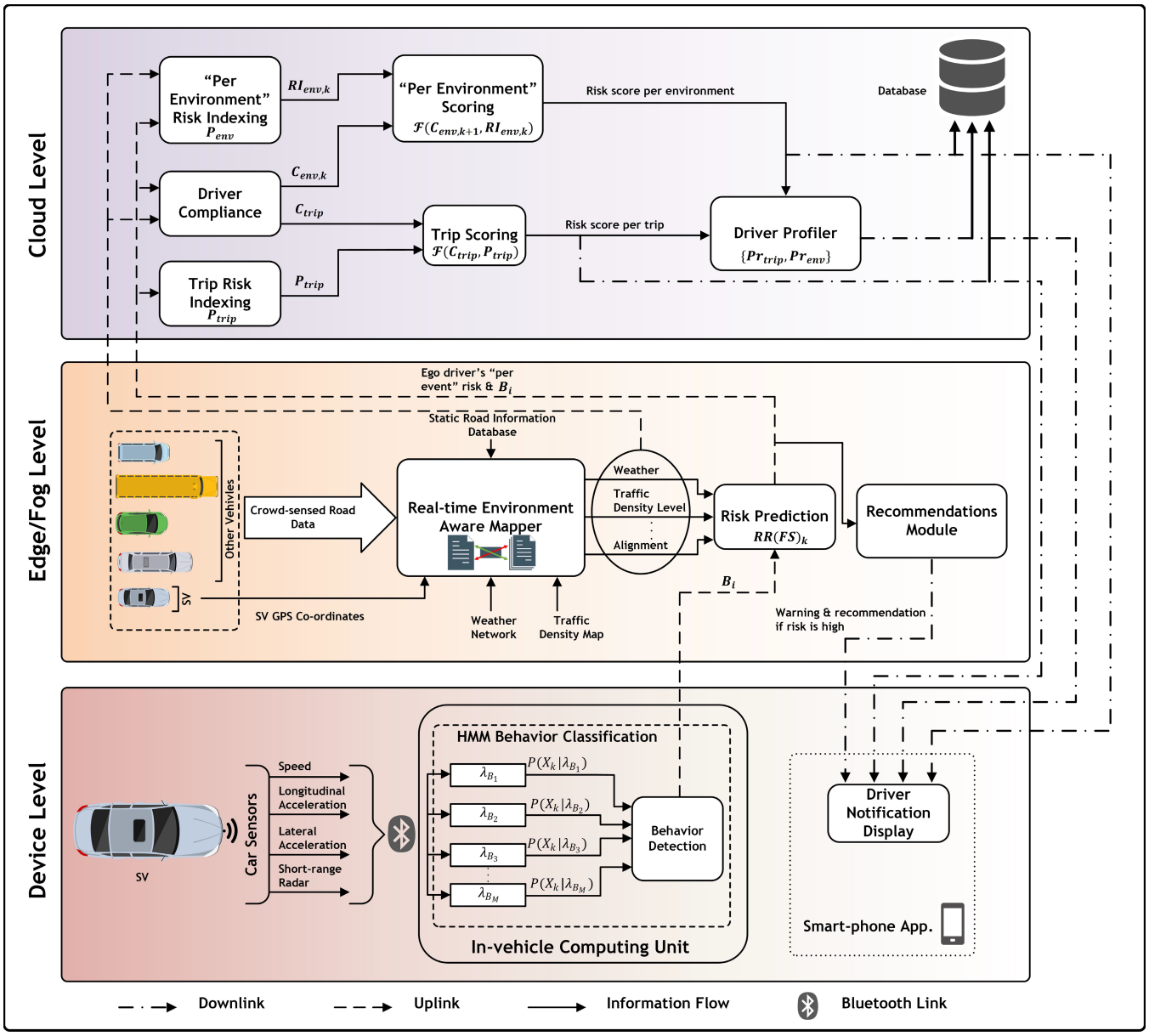



Fig. 1. Proposed cloud-based environment-aware driver profiling framework.

context in which they occurred to predict driving risk through a trained risk prediction model. If a predicted risk is higher than a pre-determined threshold, the subject driver ($sd$) is notified to change the driving behavior. Aggregated risk probabilities and the $sd's$ compliance to warnings throughout a certain driving trip are inputted to a scoring function to calculate the $sd's$ trip score. The $sd's$ risk profile is then calculated as a weighted sum of different trip scores. Unlike other profiling systems, the proposed system is motivated by statistically significant results as will be shown in section V. Figure 1 depicts the framework block diagram.

In the proposed framework, data processing is distributed over three computational layers based on the computational requirements of processes, delay, delivery, and accessibility requirements of processed data, and processed data size. The details are provided next.

### *A. Device Level: In-Vehicle Behavior Detection*

The in-vehicle module contains data acquisition, pre-processing and modeling processes that occur inside the vehicle to detect different driving behaviors. In this module, collected data can be divided into two types:

1) *Type 1*: Data that reflects the longitudinal and lateral behavior of the vehicle. This data is collected through the vehicle's Controller Area Network (CAN) bus and by utilizing the vehicle's On-Board Diagnostic (OBD/OBDII) port.
2) *Type 2*: Data that reflects the relative position of the subject vehicle to the surrounding vehicles and provides driving context awareness. This is gathered using short range radar (SRR) sensors.

Let $\boldsymbol{x}_\tau$ represent the feature vector that contains the collected vehicular data at time instant $\tau$ and

TABLE I
SUMMARY OF NOTATIONS

| Notation | Description |
|---|---|
| $sd$ | Subject driver |
| $sv$ | Subject vehicle |
| $\boldsymbol{x}_\tau$ | In-vehicle feature vector at $t = \tau$ |
| $\boldsymbol{X}$ | In-vehicle feature matrix |
| $R_s$ | Sampling rate of vehicular data |
| $B_i$ | A detected driving behavior |
| $T$ | A single in-vehicle time frame |
| $\lambda_{B_i}$ | The sequence model representing the behavior $B_i$ |
| $F_l$ | Initial feature vector for risk prediction |
| $FS_l$ | Engineered feature vector for risk prediction |
| $P(Risk\|FS_l)_k$ | Predicted risk probability of an event $k$ given $FS_l$ |
| $P(Risk\|F_l)$ | Calculated risk probability given $F_l$ |
| $env_j$ | A vector with extracted environmental attributes |
| $RR(F_l)$ | The relative driving risk of $F_l$ |
| $RI(k)$ | The risk index of an event $k$ |
| $C_{trip}$ | Driver's overall compliance in $trip$ |
| $N$ | Total number of captured risky events per trip |
| $P_{trip}$ | Average risk in $trip$ |
| $Sc_{trip}$ | $sd's$ score in $trip$ |
| $Sc_{env_j}$ | $sd's$ score in $env_j$ |
| $Pr_{trip}$ | $sd's$ risk profile after $trip$. |
| $\xi$ | Weight of EMWA filter |

expressed as:

$$\boldsymbol{x}_\tau = [v_\tau, a_{x,\tau}, a_{y,\tau}, R^F_{x,\tau}, R^F_{y,\tau}, R^R_{x,\tau}, R^R_{y,\tau}] \tag{1}$$

where $v_\tau$ represents the velocity of the subject vehicle ($sv$), $a_{x,\tau}$ and $a_{y,\tau}$ represent the acceleration in the longitudinal and lateral directions of the $sv$, respectively, $R^F_{x,\tau}$ and $R^F_{y,\tau}$ are, respectively, the ranges between the $sv$ and the closest forward object in the longitudinal and lateral directions, $R^R_{x,\tau}$ and $R^R_{y,\tau}$ are, respectively, the ranges between the $sv$ and the closest rearward object in the longitudinal and lateral directions, all at the time instant $t = \tau$. After $\tau_c$ seconds, collected data can be expressed in the following matrix notation:

$$\boldsymbol{X} = \begin{bmatrix} — & \boldsymbol{x}_1 & — \\ — & \boldsymbol{x}_2 & — \\ & \vdots & \\ — & \boldsymbol{x}_{(\tau_c \times R_s)} & — \end{bmatrix} \tag{2}$$

or equivalently:

$$\boldsymbol{X} = \begin{bmatrix} | & | & & | \\ \boldsymbol{x}(1) & \boldsymbol{x}(2) & \dots & \boldsymbol{x}(Le) \\ | & | & & | \end{bmatrix} \tag{3}$$

where $R_s$ stands for the data sampling rate and $Le$ is the length of the feature vector $\boldsymbol{x}_\tau$) (i.e., seven in this case). Data is collected and sent from OBD and radar interfaces to the $sd's$ in-vehicle computing unit (e.g., smartphone) through a Bluetooth link. In the in-vehicle computing unit, the time-series vehicular data ($\boldsymbol{X}$) is acquired over a pre-determined time interval $\tau_c$ and sequence modeling for behavior classification

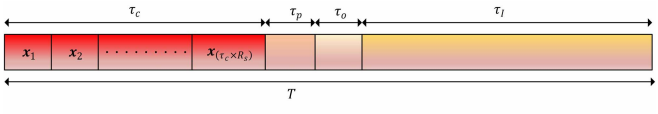



Fig. 2. A single time frame of collecting and offloading data.

(e.g., HMM-based Modeling) is applied. The behavior classification is defined as the process:

$$\mathcal{F} : \{\boldsymbol{x}(1), \dots, \boldsymbol{x}(Le)\} \rightarrow B_i \tag{4}$$

where $B_i$, $i = 1, \dots, M$ represents one of $M$ output behaviors on which the sequence model is trained to detect.

A single time frame in the in-vehicle module is depicted in Figure 2 and can be expressed mathematically as:

$$T = \tau_c + \tau_p + \tau_o + \tau_I \tag{5}$$

where $\tau_p$ is the sequence model's processing time for behavior detection, $\tau_o$ is the time required for off-loading a detected behavior to the edge/fog, and $\tau_I$ is the idle time where no vehicular data is acquired.

After the behavior $B_i$ is detected, it is sent to the edge/fog, along with the GPS co-ordinates of the $sv$ for analysis and processing.

In the proposed framework, behavior detection is performed inside the vehicle to ensure high detection accuracy and to minimize the cost of data off-loading. High levels of accuracy in behavior detection is essential given its importance for predicting risk. With the high rate at which vehicular data are sampled (on the scale of sub-seconds), performing behavior detection inside the vehicle should diminish data loss caused by off-loading data, and hence, should ensure high detection accuracy. Furthermore, transmitting vehicular data to the fog/cloud may incur a transmission cost to drivers. To illustrate, the total amount of traffic data in a 1 hour trip with $\tau_c = 10s$ and $\tau_p + \tau_o + \tau_I = 10s$, and with a data rate of $1KB/s$ will be $1.8MB$ of transmitted cellular data.

Algorithm 1 shows a summary of the explained behavior detection process.

### *B. Edge/Fog Level: Risk Prediction and Recommendation Modules*

At the edge/fog level, driving risk is predicted based on the detected behavior of the $sd$ along with the environmental context in which the behavior was detected. The $sd$ is warned and advised to change their driving behavior through a recommendation module if expected risk exceeds a pre-defined threshold. We pre-assume the existence of a real-time environment aware mapper to which the $sv's$ GPS co-ordinates are inputted and the environmental road segment attributes, which the vehicle was subjected to during detected behavior, are returned. The envisioned mapper has access to the static and dynamic road information databases on the area the designated edge/fog covers. The mapper is hosted in the edge/fog level rather than in the cloud level to minimize the time required for pulling out the environmental information of the desired road segment. To explain, having a centralized road information database

**Algorithm 1:** In-Vehicle Behavior Detection

**Input:** Vehicular data: $\{\boldsymbol{x}\}_{\tau=1}^{\tau=\tau_c \times R_s}$, Data Collection Time: $\tau_c$, Idle Times: $\{I_F, I_T\}$
**Output:** $B_i$

```
repeat
    for τ ← 1 to τ_c × R_s do
        X.append(x_τ)
    for k ← 1 to M do
        Calculate P(X|λ_{B_k})
        P.append(P(X|λ_{B_k}))
    i = arg max{P}
    Offload B_i & location co-ordinates
    if warning = 'FALSE' then
        τ_I = I_F
    else
        τ_I = I_T
until trip = 'FALSE'
```

TABLE II
RISK SEVERITY LEVELS

| Risk Severity | Definition |
|---|---|
| **Severe (4)** | A driving event where $RR(F_l) > 4$. A warning is issued. A non compliance to warnings results in zero compliance score. |
| **Critical (3)** | A driving event where $3 < RR(F_l) \leq 4$. A warning is issued. A non compliant driver receives only a one-quarter compliance score. |
| **High (2)** | A driving event where $2 < RR(F_l) \leq 3$. A warning is issued. A non compliant driver looses half of his/her compliance score. |
| **Normal (1)** | A driving event where $1 < RR(F_l) \leq 2$. A warning is issued. A non compliance to warnings results in a one-quarter reduction in compliance score. |
| **Low (0)** | A driving event where $RR(F_l) \leq 1$. No warning is issued. |

in the cloud that contains the information of a large traffic network would increase the search time needed for extracting the information of a designated road segment, and hence will increase the time needed for predicting risk. Likewise, both risk prediction and recommendation modules are hosted in the edge/fog level to reduce the time required for calculating the expected risk of a captured event and to reduce the time latency between predicting risk and warning a risky driver.

Environmental attributes contain static information about the road characteristics, and the road real-time information such as density level, weather condition, traffic flow, and lighting conditions. In this framework, we utilized the following environmental attributes: weather condition ($W$), traffic density level ($TD$), road lighting conditions ($L$), traffic control ($TF$), road flow ($RF$), and road alignment ($A$). The returned environmental attributes vector $env_j$, where $j \in [1, \ldots, J]$, along with the $sd's$ detected behavior $B_i$ form the initial feature vector $F_l$, $l = 1, \ldots, L$:

$$F_l = [B_i, env_j] \tag{6}$$

Feature extraction and selection is then performed on the initial feature vector. The engineered feature vector ($FS_l$) is then inputted to a trained risk prediction model.

The risk prediction model uses $FS_l$ to predict the driving risk probability $P(Risk|FS_l)_k$, where the subscript $k$ is an integer that represents an event index. The driving risk probability is expressed herein in terms of the crash and near-crash rate:

$$P(Risk|FS_l)_k = P(C|FS_l)_k + P(NC|FS_l)_k \tag{7}$$

where $P(C|FS_l)_k$ and $P(NC|FS_l)_k$ are, respectively, the conditional probabilities of crash and near-crash events (including crash relevant events) given the feature vector $FS_l$ at event $k$. The conditional risk probabilities in different driving environments are calculated as:

$$P(Risk|F_l) = \frac{R_{F_l}}{R_{F_l} + NR_{F_l}} \tag{8}$$

where $R_{F_l}$ and $NR_{F_l}$ are, respectively, the number of risky and non-risky events, given $F_l$. In SHRP2 data-set, a non-risky event is either a non-subject conflict, or a balanced baseline event, as they are previously defined.

Once risk probability is predicted, a warning is issued to the subject driver. The level of warning severity changes according to the level of risk the detected behavior imposes. Since the risk probability is data-set dependent and is characterized by the sampling rate at which normal driving events are captured, the thresholds between risk levels can be set using the following relative driving risk equation:

$$RR(F_l) = \frac{P(Risk|F_l)}{P(Risk|F_l')} \tag{9}$$

where $RR(F_l)$ is the relative driving risk of $F_l$, and $F_l'$ is the complement of $F_l$ (i.e., $[B_i, env_j]'$).

Based on the relative driving risk values, risk severity is assigned and warnings are issued accordingly. In this work, risk severity during a driving event belongs to the set $\{Severe, Critical, High, Normal, Low\}$ or equivalently to the integer set $\{4, 3, 2, 1, 0\}$, as shown in Table II.

As shown in Table II, risk severity levels are assigned depending on the relative driving risk of a captured event. A driving event with a relative driving risk of 1 possesses a risk probability equivalent to the average risk probability of events captured in other driving environments. Consequently, a relative driving risk of 1 was chosen as a threshold between low-risk events and other events. If a captured event imposes some risk, the $sd$ will be notified and advised to change his/her behavior as to reduce risk. The $sd$ receives a complete compliance score unless he/she does not change behavior to normal. If the $sd$ is not compliant, the reduction of his/her compliance score will be directly proportional to the event risk severity.

The $sd's$ compliance to warnings along with a weighted sum of the aggregated risk probabilities over a certain trip are

both used to compute the final trip score $Sc_{trip}$ as will be detailed in the next section.

### C. Cloud Level: Scoring and Profiling Processes

At the cloud level, time-tolerant computationally intensive operations are hosted. On the cloud, the overall risk and compliance of drivers through their driving trips are computed. Risk and compliance are utilized afterwards to calculate trip scores or to update personalized competency levels of drivers in various driving environments. Based on risk and compliance scores, risk profiles of drivers are continuously updated after each driving trip and stored in a centralized database. Drivers are notified about their overall scores and about updates in their driving profiles following the end of each driving trip. Processing such large a amount of data requires a high performance computing (HPC) servers which are available on the cloud level. To highlight the asymptotic time complexity of the on-cloud operations, let's take the computation of driver compliance as an example. In a time slot $t$, computing the compliance to warnings for $M$ drivers in $E$ events will be of the order of $O(M \times E)$. Repeating this process $K$ times will incur a computational cost of $O(M \times E \times K)$. With such high computational cost, it is reasonable to host such operations in the cloud. Next, the logical flow of information on the cloud is detailed.

Following risk prediction of event $k$, predicted risk is offloaded to the cloud and inputted to the "Trip Risk Indexing" module. Based on the predicted risk severity level, the event $k$ is assigned a risk index $RI(k)$ according to Equation 10:

$$RI(k) = 0.25 * sl_k \tag{10}$$

where $sl_k \in \{0, 1, 2, 3, 4\}$ is the risk severity of event $k$ and is one of the risk severity levels shown in Table II. Risk indices for all captured events during a driving trip are computed and stored. The overall trip risk index $P_{trip}$ can be simply calculated as the trip average risk, which is denoted by the following formula:

$$P_{trip} = \frac{1}{N} \sum_{k=1}^{N} RI(k) \tag{11}$$

where $N$ is the total number of captured events in a trip.

The $sd$ compliance to a warning following being involved in a risky behavior during event $k$ is calculated through the "Driver Compliance" module during event $k+1$ (i.e., monitoring the driver behavior after issuing a warning). As shown in Table II, compliance is computed according to the risk severity of $k$. To explain, the $sd$ is given the full compliance score of 1 if the driver is compliant. If the driver is non-compliant, a deduction in compliance score is weighted according to risk severity during the event $k$. Lets define the binary variable $c_{k+1}$ as follows:

$$c_{k+1} = \begin{cases} 1, & \text{if } RI(k+1) > 0 \\ 0, & \text{if } RI(k+1) = 0 \text{ or } B_{i,k} \text{ is normal} \end{cases} \tag{12}$$

Then, compliance to a warning following a risky behavior in event $k$ is expressed as:

$$C(k) = 1 - c_{k+1} * RI(k) \tag{13}$$

Similar to the overall trip risk index $P_{trip}$, the overall trip compliance, $C_{trip}$, is calculated as the average compliance throughout a driving trip. It is expressed mathematically as:

$$C_{trip} = \frac{1}{N-1} \sum_{k=1}^{N-1} C(k) \tag{14}$$

The argument above requires repeating the in-vehicle processes of data collection, behavior detection, and data offloading, as well as the cloud risk prediction process, each time after detecting a risky behavior. This repetition verifies whether the driver was compliant to the warning or not. A simpler and more practical yet less accurate approach is to calculate the $sd's$ compliance based on their compliance probability distribution for events of different severity levels.

Under the assumptions of:

1) Independent $sd$ compliances in different risky events.
2) Equally probable compliance rates in different driving environments and for events with the same risk severity level.

the probability of $l$ compliances in $N_{sl}$ risky events of severity level $sl$ would follow a binomial distribution with parameter $p_{sl}$:

$$P(C_{sl} = l) = \binom{N_{sl}}{l} p_{sl}^{l} (1 - p_{sl})^{N_{sl} - l} \tag{15}$$

The overall compliance per trip $C_{trip}$ would be the probability of being always compliant (i.e., $l = N_{sl}, \forall sl \in \{0, 1, 2, 3, 4\}$). Substituting Equation 15 in Equation 13, $C_{trip}$ can be expressed as follows:

$$C_{trip} = \sum_{sl=0}^{sl=4} (1 + RI_{sl}).P(C_{sl} = N_{sl}) - RI_{sl} \tag{16}$$

This simplified formulation will require only calculating the probability parameters $p_{sl}, \forall sl \in \{0, 1, 2, 3, 4\}$ in a primary training phase, which is more practical in many situations. These probability parameters can be updated regularly to track the changes in a driver's compliance behavior.

The trip score is then computed as a function of the trip weighted sum of the risk index $P_{trip}$, and the driver's per trip compliance value $C_{trip}$:

$$Sc_{trip} = \mathcal{F}(C_{trip}, P_{trip}) \tag{17}$$

Given that $P_{trip} \in [0, 1]$ and $C_{trip} \in [0, 1]$, a normalized $Sc_{trip} \in [0, 1]$ can be written as:

$$Sc_{trip} = \gamma.C_{trip} + \alpha.(1 - P_{trip}) \tag{18}$$

where

$$\gamma + \alpha = 1 \tag{19}$$

The values of $\gamma$ and $\alpha$ determine how much weight is given to $C_{trip}$ and $P_{trip}$. For instance, if $\alpha = 1$, the overall trip score will be determined solely based on the value of $P_{trip}$ (i.e., $\gamma = 0$).

Finally, a subject driver's profile after a certain trip ($Pr_{trip}$) can be computed using an exponentially moving weighted average (EMWA) filter applied on various trip scores to

assign exponentially increasing weights for recent trips. This is expressed as:

$$Pr_{trip} = \begin{cases} Sc_1, & \text{if } trip = 1 \\ \xi.Sc_{trip} + (1-\xi).Pr_{trip-1}, & \text{if } trip > 1 \end{cases} \quad (20)$$

where the value of $\xi$ determines the number of trips which the filter will use to calculate $Pr_{trip}$.

Using the same analogy of updating the $sd$ per trip profile, the $sd's$ per environment profile is updated. The "Per Environment risk indexing" module calculates $RI_{env_j}(k)$ which is the risk index for event $k$ taken into consideration the environmental context of the event, calculated for each $env_j$. $RI_{env_j}(k)$ is utilized to reflect the driving competency level of the $sd$ in the driving environment $env_j$ along with the compliance $C_{env_j}(k)$. Consequently, the score of the $sd$ in $env_j$ at event $k$ is:

$$Sc_{env_j}(k) = \gamma.C_{env_j}(k) + \alpha.(1 - RI_{env_j}(k)) \quad (21)$$

An $sd$ profile in $env_j$ ($Pr_{env_j}$) can then be updated after each event captured in $env_j$. Similar to the per trip profile, $Pr_{env_j}$ can be computed using an EMWA filter to assign exponentially increasing weights for recent captured events in $env_j$.

An important feature of the presented framework is the prediction of driving risk probabilities given the behavioral and environmental attributes. Non-accurate values of these probabilities can result in missed or false warnings as well as unreliable driving scores. The rest of the paper contains the necessary steps for the development of the driving risk prediction model. Moreover, the effect of risk prediction results on the overall scoring performance is analyzed using SHRP2 naturalistic driving data.

## IV. Data Pre-Processing and Model Selection

Raw data contains the information of $\sim$ 29,000 driving events, each with a certain severity level. In the original dataset, event severity levels are exclusively contained in the following set: *Severity* $\in$ *{Crash, Near-Crash and Crash-Relevant, Non-Subject Conflict, Balanced Baseline}*. An event $k$ in the dataset is represented by a vector that contains the captured driving behavior of the subject driver prior to a risky event (or during a normal driving event) ($B_i$), the environmental context in which these behaviors happened ($env_j$), and the event severity ($Severity$):

$$k = [[B_i, env_j] \xrightarrow{yields\ to} Severity] \quad (22)$$

Since we are concerned with classifying the risk level of an event given the behavior of the driver and the environmental context, the notion of risk is developed as shown in Equations 7-9. The initial feature matrix is transformed from the original event-based matrix to the following matrix (23), shown at the bottom of the page.

### A. Data Pre-Processing

*1) Data Merging: Crash, Near-Crash* and *Crash-Relevant* severity levels are put under the common severity level of *Risky*. *Non-Subject Conflict* and *Balanced Baseline* events are used to represent the *Normal* level. Under each environmental category, similar features are merged to increase their importance in order to enhance the prediction model performance (e.g., under road alignment category, curved to the right and curved to the left features are considered the same). Similarly, we used 13 behaviors that were previously identified in [29]. For the sake of completion, the set of the previously identified behaviors are displayed again in table III. Identified environmental features are shown in Table IV.

*2) Data Filtering:* Rows in the feature matrix are filtered out if their relative driving risk values ($RR(F_l)$) are not statistically significant. The $p - value$ is utilized to signify statistical significance. Rows which possess a $p - value > 0.1$ are filtered out. The filtered feature matrix has $L'$ rows. With the Contingency table shown in Table V, the $p - value$ is calculated for each row $l$ using Fisher's exact ratio as:

$$p_l = \frac{\binom{a+c}{a}\binom{b+d}{b}}{\binom{n}{a+b}} \quad (24)$$

where $n = a + b + c + d$.

*3) Data Encoding:* After data merging, the behavioral and environmental categorical variables are encoded to integers. To calculate risk probability, events with the same behavioral and environmental features are combined and the corresponding risk probability for each is calculated. To represent data in a meaningful way for the machine learning algorithms, the *one-hot encoding* technique is utilized.

### B. Model Selection

The encoded data is then divided into training and development sets according to the ratio of 70% and 30%, respectively. Using the mean absolute error ($MAE$) as a performance metric, an error analysis for a simple multiple linear regression model indicated a high bias (i.e., low training set performance). More complex structured SVM-based models, on the other hand, were able to model training data accurately, but were not capable of generalizing on the development set (i.e., high variance). To achieve a good bias-variance trade-off, a customized random forest model was selected. In random forests [30], multiple decision trees are built, each from a

$$\left[\begin{array}{c|cc|c} \text{Feature Vector } (F_l) & B_i & env_j & \text{Outcome } (RR(F_l)) \\ \hline F_1 & B_1 & env_1 & RR(F_1) \\ F_2 & B_1 & env_2 & RR(F_2) \\ \vdots & \vdots & \vdots & \vdots \\ F_L & B_\mu & env_J & RR(F_L) \end{array}\right] \quad (23)$$

TABLE III
SUMMARY OF DRIVING BEHAVIORS [29]

| Behavior index | Behavior | Description |
|---|---|---|
| 1 | Excessive speeding | Exceeding safe speed/speed limit |
| 2 | Inexperience or unfamiliarity | Apparent general inexperience driving, unfamiliarity with a vehicle or a roadway |
| 3 | Avoiding an object | Avoiding a vehicle, pedestrian, or an object |
| 4 | Sudden braking | Sudden or improper stopping on a roadway |
| 5 | Right-of-way error | Right-of-way error due to decision or recognition failures, or an unknown cause |
| 6 | Driving slow | Driving slowly in relation to other traffic or below speed limit |
| 7 | Improper reversing | Improper backing up due to inattentiveness or other causes |
| 8 | Illegal or unsafe lane change or turn | Any improper or illegal lane change or turn |
| 9 | Aggressive driving | Such as aggressive acceleration or aggressive lane changing |
| 10 | Signal or sign violation | Violation action at traffic signs or signals |
| 11 | Normal | No evidence/presence of risky behavior |
| 12 | Fatigue | Drowsiness, sleepiness, and fatigue |
| 13 | Negligence | Includes improper or failure to signal, and driving past dusk without lights |

TABLE IV
SUMMARY OF ENVIRONMENTAL CONDITIONS

| Traffic Flow | Traffic Density | Traffic Control | Weather Conditions | Lighting Conditions | Road Alignment |
|---|---|---|---|---|---|
| Divided | Stable | Yes | No Adverse Conditions | Dark | Straight |
| Not Divided | Stable With Flow Restrictions | No | Foggy | Lighted | Curved |
| No Lanes | Unstable | - | Rainy | - | - |
| - | - | - | Snowy | - | - |

TABLE V
CONTINGENCY TABLE FOR THE NUMBER OF RISKY AND NON-RISKY EVENTS

| | Risk | No risk |
|---|---|---|
| $F_l$ | a | c |
| $F_l'$ | b | d |

TABLE VI
HYPER-PARAMETERS OF RF MODEL

| Hyper-parameter | Classification | Regression |
|---|---|---|
| *Number of Trees* | 100 | 100 |
| *Split Criterion* | Entropy | $MSE$ |
| *Max No. of Features per Tree* | $\sqrt{N_{tot}}$ | $N_{tot}$ |

sample of the training set. The best split in each tree is based on a random subset of the input features rather than the whole feature set. The average performance of the various trees is then used to reflect the forest performance. Although this approach theoretically causes a slight degradation in the training set performance, it reduces over-fitting due to the averaging process. Our results indicate that a customized random forest model resulted in the best bias-variance performance. Figure 3 depicts a comparison between customized random forests (RF), linear, and SVM regressors. The figure shows a 5-fold cross-validation for $MAE$, mean squared error ($MSE$), and the coefficient of determination ($R^2$). The RF regressor outperforms the linear and SVM regressors in all three evaluation metrics. The RF regressor has a mean $MAE$ of 0.5 as opposed to 0.61 and 0.79 for linear and SVM resgressors, respectively. Moreover, It has a high mean $R^2$ value of 0.65 as opposed to 0.59 and 0.27 for linear and SVM regressors, respectively. It is also clear that the RF regressor performance is more consistent when trained over different training set folds. Such consistency is inferred from the dispersion profile of the RF regressor box plots.

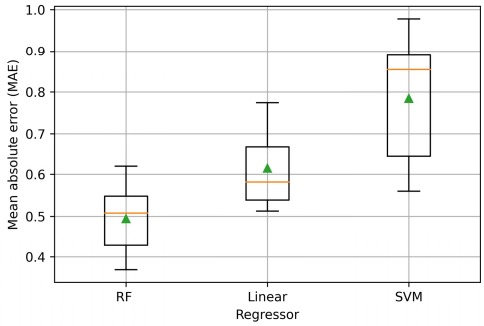
The adopted hyper-parameters of the selected model are shown in Table VI, where $N_{tot}$ represents the number of all behavioral and environmental features, and $MSE$ is the mean square error.

## V. PERFORMANCE EVALUATION AND DISCUSSION

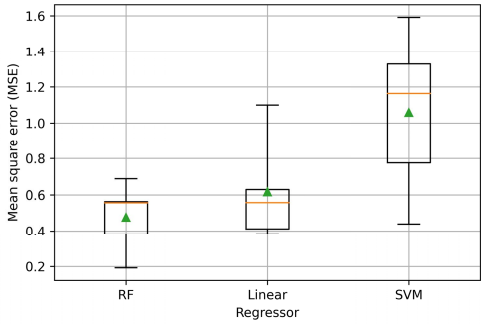
We investigate the performance of the Random Forest risk prediction model presented in section IV. The model was implemented in Spyder (Python 3.6) integrated development environment (IDE) using the Scikit-Learn Library for Machine Learning and Data Mining. Results in the regression context are discussed along with the relevant risk index $RI$ and the overall risk scoring results. Reported results are those obtained from the customized RF model after trying different random seeds. They represent the best obtained results.

### A. Risk Prediction

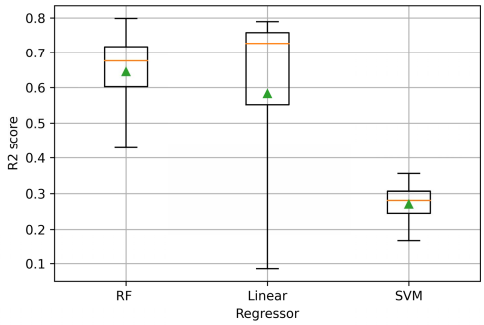
The developed RF model is trained to predict the relative driving risk of a specific event given the driver's behavior and the environmental context. The model was trained and validated according to the splitting ratio of 70 % and 30 %,




(a) Mean absolute error performance.




(b) Mean square error performance.




(c) The coefficient of determination ($R^2$) performance.

Fig. 3. Performance comparison between customized RF, linear, and SVM regressors for the prediction of relative driving risk.

respectively. The $10-fold$ cross validation was performed to reflect the average performance of the model over different training samples. The normalized absolute error histograms of the model for both training and validation sets are depicted in Figures 4 and 5, respectively.

The normalized absolute error ($NAE$) percentage of a feature vector $F_l$ is calculated according to Equation 25:

$$NAE(F_l)\% = \frac{|RR_{act}(F_l) - RR_{pred}(F_l)|}{\max(RR_{act}) - \min(RR_{act})} \tag{25}$$

where $RR_{act}(F_l)$ and $RR_{pred}(F_l)$ are, respectively, the actual and predicted relative driving risk values for the feature vector $F_l$, and $RR_{act}$ is the vector that contains the actual relative driving risk values for all the feature vectors in the data-set. Figure 4 shows that the sample count is exponentially decreasing as the $NAE$ increases, with a maximum $NAE$ of 27%.

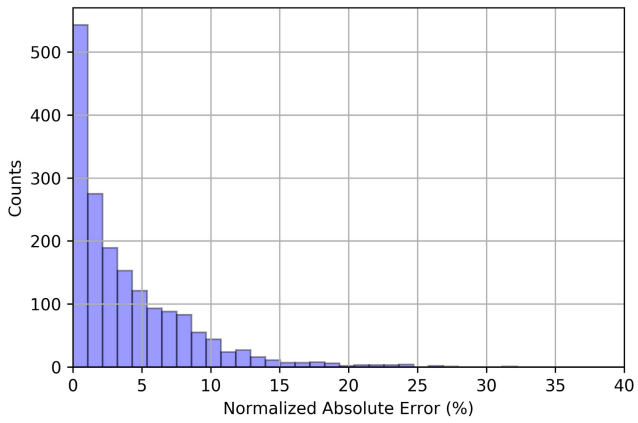



Fig. 4. The normalized absolute error histogram for the training set using the developed RF risk prediction model.

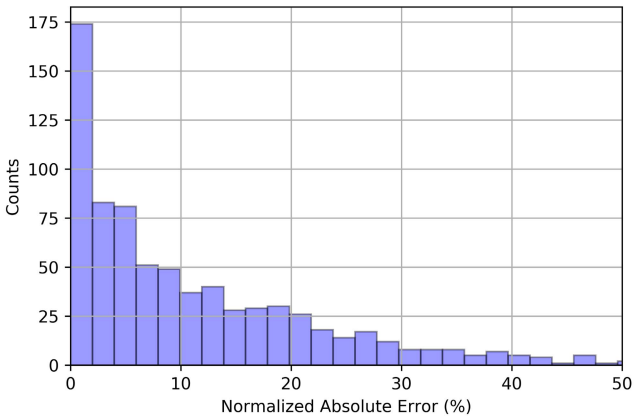



Fig. 5. The normalized absolute error histogram for the validation set using the developed RF risk prediction model.

Similarly, the validation set $NAE$ performance resembles an exponential distribution but with a higher normalized mean absolute error $NMAE$ as shown in Figure 5. The summary of the model $NMAE$ and $R^2$ results is shown in Table VII. The validation set results show the high performance standards the developed model can achieve with an average $NMAE$ of only 10.7% and with an ability to explain most of the variability in the data output as shown from the coefficient of determination value (e.g., 0.66). Moreover, the training set performance indicates that the developed model has a very small bias with an $NMAE$ value of 4.25% and $R^2$ value of 0.95. Despite the good performance results for both training and validation sets, the validation set performance shows a 6.45% degradation in the $NMAE$ performance when compared to the training set. Furthermore, a 0.29 difference in the $R^2$ value is noticed - an indication the developed model is slightly over-fitted. Although this bias-variance combination was the best achieved, over-fitting was unavoidable which may be attributed to the data-set sample size.

Using Equation 10, the actual and predicted risk indices are respectively computed from the actual and predicted relative driving risk values. The mean absolute error ($MAE$) metric

TABLE VII
SUMMARY OF THE RF MODEL RESULTS

| Performance Measure | Training Set (%) | Development Set (%) |
|---|---|---|
| *NMAE (%)* | 4.25 | 10.7 |
| *Adjusted* $R^2$ | 0.95 | 0.66 |

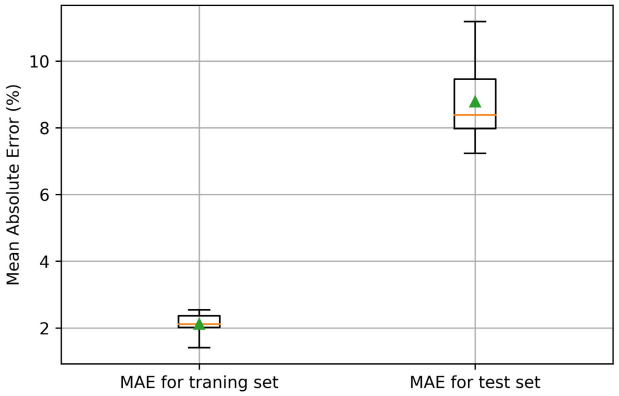



Fig. 6. Whisker plot for the $MAE$ performance of $RI$ using $10-fold$ cross-validation.

is utilized to signify the performance. Figure 6 depicts the Whisker plot of the $MAE$ for the risk index $RI$ using $10-fold$ cross validation. The average $MAE$ for the training and validation sets is, respectively, 2% and 8.7%. Such negligible average errors highlight the accurate $RI$ results, and hence, the accurate scoring results as will be shown in section V-B.

### B. Driver Scoring

We derive and empirically calculate the expected value of the deviation in an event score given the SHRP2 event-based dataset. As shown in section III-C, the performance of a driver in an event $k$ is calculated based on the risk index $RI(k)$ and the compliance $C(k)$. The absolute error in the score of a driver given the feature vector, $F_k$, is defined as the absolute difference between the actual and predicted scores of a driver in an event $k$. It is denoted by $Sc_{error}(k)$ and can be expressed mathematically as:

$$|Sc_{actual}(k) - Sc_{pred}(k)| = \alpha.|RI_{act}(k) - RI_{pred}(k)| + \gamma.|C_{act}(k) - C_{pred}(k)| \quad (26)$$

where $Sc_{act}(k)$ and $Sc_{pred}(k)$ are, respectively, the actual and predicted risk scores of a driver in event $k$. The expected value of the absolute error $Sc_{error}$ can then be expressed as:

$$\mathbb{E}(Sc_{error}) = \alpha.\,\mathbb{E}(|RI_{act} - RI_{pred}|) + \gamma\,.\,\mathbb{E}(|C_{act} - C_{pred}|) \quad (27)$$

where $\mathbb{E}(|RI_{act} - RI_{pred}|)$ and $\mathbb{E}(|C_{act} - C_{pred}|)$ are, respectively, the mean absolute errors for the risk index and the compliance scores. Let $F_l$ and $F_j$ denote the features vectors at two consecutive events, where $F_j$ is the feature vector of

TABLE VIII
CONFUSION MATRIX FOR TRAINING SET COMPLIANCE CLASSIFICATION

| Actual \ Predicted | Complaint | Non-compliant |
|---|---|---|
| Compliant | 168 | 29 |
| Non-compliant | 4 | 1551 |

TABLE IX
CONFUSION MATRIX FOR VALIDATION SET COMPLIANCE CLASSIFICATION

| Actual \ Predicted | Complaint | Non-compliant |
|---|---|---|
| Compliant | 67 | 28 |
| Non-compliant | 11 | 654 |

the following event $\mathbb{E}(Sc_{error})$ can be written as:

$$\mathbb{E}(Sc_{error}) = \alpha.\sum_{l=1}^{L} P(F_l).|RI_{act}(l) - RI_{pred}(l)| + \gamma.\sum_{j=1}^{L'}\sum_{l=1}^{L} P(F_j|F_l).|C_{act}(l) - C_{pred}(l)| \quad (28)$$

where $L$ is the total number of all possible combinations of behaviors and environments, $P(F_l)$ and $P(F_j|F_l)$ are the probability of $F_l$ and the conditional probability of $F_j$ given $F_l$, respectively.

The absolute deviation in compliance ($|C_{act}(l) - C_{pred}(l)|$) is calculated for four cases:

1) The model predicts that the driver is compliant given that the driver is actually complaint. In this case, $|C_{act}(l) - C_{pred}(l)| = 0$.
2) The model predicts that the driver is non-compliant while the driver is actually complaint. In this case, $|C_{act}(l) - C_{pred}(l)| = RI_{pred}(l)$.
3) The model predicts that the driver is compliant while the driver is actually non-complaint. The absolute deviation in compliance, in this case, is $RI_{act}(l)$.
4) The model predicts that the driver is non-compliant while the driver is actually non-complaint. The absolute deviation in compliance in this case is $|C_{act}(l) - C_{pred}(l)| = |RI_{act}(l) - RI_{pred}(l)|$.

Under the assumption of independent occurrences of $F_l$, $\forall l \in [1, L]$ and according to the four cases shown above, $Sc_{error}$ can be written as:

$$\begin{aligned} Sc_{error} = \alpha.\frac{1}{L}\sum_{l=1}^{L}|RI_{act}(l) - RI_{pred}(l)| \\ +\gamma.(P(NonC|C).\frac{1}{L}\sum_{l=1}^{L} RI_{pred}(l) \\ +P(C|NonC).\frac{1}{L}\sum_{l=1}^{L} RI_{act}(l) \\ +P(NonC|NonC).\frac{1}{L}\sum_{l=1}^{L}|RI_{act}(l) - RI_{pred}(l)|) \end{aligned} \quad (29)$$

TABLE X
AN ILLUSTRATIVE EXAMPLE OF TRIP SCORING FOR AN *sd* USING PROPOSED RISK SCORING SYSTEM

| Event Index | Behavior | Traffic Flow | Traffic Density | Traffic Control | Weather | Lighting | Road Alignment | Actual Score | Predicted Score |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Normal | Divided | Stable | No | No Adverse Conditions | Dark | Straight | 1 | 1 |
| 2 | Illegal or unsafe lane change or turn | Divided | Stable | No | No Adverse Conditions | Dark | Straight | 0.625 | 0.625 |
| 3 | Normal | Divided | Stable With Restrictions | No | No Adverse Conditions | Dark | Curved | 1 | 1 |
| 4 | Excessive Speeding | Divided | Stable | No | No Adverse Conditions | Lighted | Curved | 0.75 | 0.5 |
| 5 | Excessive Speeding | Divided | Stable | No | No Adverse Conditions | Lighted | Curved | 0.875 | 0.75 |
| 6 | Driving Slow | Not Divided | Stable | No | No Adverse Conditions | Lighted | Straight | 1 | 1 |
| 7 | Aggressive Driving | Divided | Stable With Restrictions | No | No Adverse Conditions | Lighted | Straight | 0.25 | 0.25 |
| 8 | Aggressive Driving | Divided | Stable With Restrictions | No | No Adverse Conditions | Lighted | Straight | 0.625 | 0.625 |
| 9 | Normal | Divided | Stable With Restrictions | No | No Adverse Conditions | Lighted | Straight | 1 | 1 |
| | | | | | | | **Trip Score =** | **7.9/10** | **7.5/10** |

where $P(NonC|C)$, $P(C|NonC)$, and $P(NonC|NonC)$ are, respectively, the probability of the driver being classified as non-compliant given that the driver is actually compliant, the probability of the driver being classified as compliant given that the driver is actually non-compliant, and the probability of the driver being classified as non-compliant given that the driver is actually non-compliant. The mean of those probabilities are empirically calculated from the data-set for the training and validation sets using the confusion matrices shown in Tables VIII and IX, respectively.

The expected value of the absolute score error is then computed using Equation 29. Figure 7 depicts the $10-fold$ cross-validation Whisker plot for $Sc_{error}$, where $\alpha$ and $\gamma$ are set to 0.5. The Figure shows that the average $Sc_{error}$ for the validation set is 9.5%, which means that, on average, the risk score of a driver in a captured event will be deviated from the true value by 9.5%.

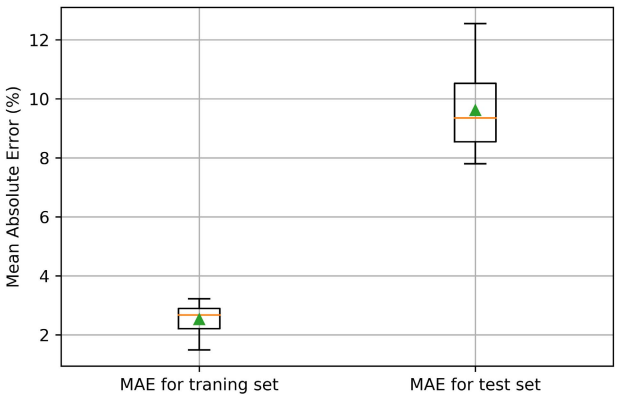



Fig. 7. Whisker plot for the mean absolute event score error using $10-fold$ cross-validation.

## VI. ILLUSTRATIVE EXAMPLE

In this section, an explanation of the trip scoring process for a subject driver using the proposed risk scoring system is provided through an explanatory example. Table X displays the details of an imaginary driving trip composed of nine captured driving events. The table shows the behavioral and environmental features (i.e., $F_l$) that are used as predictors to driving risk. For each row, the actual and predicted scores respectively represent the per-event actual and predicted risk scores. Actual and predicted scores are computed using a weighted average of the actual and predicted risk indexes (equation 10) and compliance to warnings (equation 13). For the actual risk score, $sl_k$ in equation 10 is computed directly from the data-set, and is predicted using $F_l$ for the predicted risk score. The overall trip score is calculated according to equation 18. The two weighing factors $\alpha$ and $\gamma$ in equation 18 are both set in this example to 0.5, which means that for a captured event, the risk score of the *sd* will be calculated by equally considering the risk index of the event and the driver's compliance to a warning.

For the first event, the driver's behavior is classified as "normal" and the driver is consequently assigned the full score of 1. The driver's score in the second event is calculated based on the event's risk index (i.e., $RI(k)$) and the driver's compliance observed in the third event. The driver receives

the full compliance score since he/she changed behavior to "normal". However, given the high risk imposed by the driver's behavior in the second event (i.e., $RI(k) = 0.75$), the overall score is calculated as: $\alpha.(1 - 0.75) + \gamma.1 = 0.5 \times 0.25 + 0.5 = 0.625$. The predicted score in this case coincides with the actual score with no error. During the fourth event, the driver was excessively speeding. In this case, there was a 25% deviation from the actual score given that the actual and predicted risk indices are 0.25 and 0.5, respectively. The driver was not compliant in this case since he/she did not change behavior to "normal" nor the risk index was zero during the following event. Consequently, the score was calculated solely based on the event risk index. The overall absolute deviation in the $sd's$ score during this trip is: $|0.79 - 0.75| \times 100\% = 4\%$.

## VII. Future Work

Some practical considerations need to be addressed for the proposed driver profiling framework to be implementable.

On the device level, the complexity of the hardware required to detect driving behaviors presented in table III should be sufficiently studied. Although the majority of these behaviors can be directly inferred from low-cost sensor platforms (e.g., smartphones), some behaviors might need complex computer vision techniques (e.g., signal or sign violation) or the use of radar sensors for accurate detection.

Another important practical consideration is the choice of the communication protocol between the edge devices and the fog or the cloud. Messaging has to be reliable and suitable for such a bandwidth-limited application. Also handling failed network path scenarios and how scores will be calculated in such cases need to be appropriately studied.

Finally, it is crucial to ensure that data of drivers is protected across the three computational levels (i.e., device, fog/edge, and cloud levels). For instance, tackling security attacks, in which risk scores and profiles of drivers are altered, is pivotal to ensure a robust and a trustworthy profiling system.

## VIII. Conclusion

In this paper, a novel driver risk profiling framework is presented and discussed. The information flow among three different computational layers (i.e., the device, edge/fog, and cloud layers) in the proposed profiling system is investigated. The risk, scoring, and profiling notions are mathematically defined and explained. The paper addressed the risk prediction problem by utilizing the behavioral and environmental contextual information of $29,000$ driving events, using the SHRP2 NDS. Data pre-processing and model selection processes are performed to achieve the best possible prediction performance. By analyzing the mean absolute error of different models, a customized randomized trees model appears to give the best bias-variance trade-off. Results confirm that behavioral and environmental data are together good predictors of driving risk, which is measured in this paper in terms of crash, near-crash and crash-relevant events. The developed model is then utilized to calculate the average error between predicted and actual risk indices and the average overall risk score error. An explanatory example of the risk scoring process using the proposed framework is provided. The results clearly show the robustness and effectiveness of the proposed profiling system in assigning accurate and representative risk scores for drivers.

## Acknowledgment

This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof. The findings and conclusions of this paper are those of the authors and do not necessarily represent the views of the Virginia Tech Transportation Institute, SHRP 2, the Transportation Research Board, or the National Academy of Sciences.

## References

[1] *WHO | Global Status Report on Road Safety 2015*. Accessed: Dec. 7, 2019. [Online]. Available: http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/

[2] G. Castignani, T. Derrmann, R. Frank, and T. Engel, "Driver behavior profiling using smartphones: A low-cost platform for driver monitoring," *IEEE Intell. Transp. Syst. Mag.*, vol. 7, no. 1, pp. 91–102, Spring 2015.

[3] D. I. Tselentis, G. Yannis, and E. I. Vlahogianni, "Innovative insurance schemes: Pay as/how you drive," *Transp. Res. Proc.*, vol. 14, pp. 362–371, Jan. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2352146516300898

[4] M. J. Davidson and J. A. Olsen, III, "Systems and methods for utilizing telematics data to improve fleet management operations," U.S. Patent 8 416 067 B2, Apr. 2013. Accessed: Nov. 22, 2021. [Online]. Available: https://patents.google.com/patent/US8416067B2/en

[5] K. Takeda *et al.*, "Self-coaching system based on recorded driving data: Learning from one's experiences," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 4, pp. 1821–1831, Dec. 2012.

[6] P. Händel *et al.*, "Insurance telematics: Opportunities and challenges with the smartphone solution," *IEEE Intell. Transp. Syst. Mag.*, vol. 6, no. 4, pp. 57–70, Winter 2014.

[7] M. Witt, K. Kompaß, L. Wang, R. Kates, M. Mai, and G. Prokop, "Driver profiling—Data-based identification of driver behavior dimensions and affecting driver characteristics for multi-agent traffic simulation," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 64, pp. 361–376, Jul. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1369847818308131

[8] A. A. Rahman, W. Saleem, and V. V. Iyer, "Driving behavior profiling and prediction in KSA using smart phone sensors and MLAs," in *Proc. IEEE Jordan Int. Joint Conf. Electr. Eng. Inf. Technol. (JEEIT)*, Apr. 2019, pp. 34–39.

[9] B. He, D. Zhang, S. Liu, H. Liu, D. Han, and L. M. Ni, "Profiling driver behavior for personalized insurance pricing and maximal profit," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 1387–1396.

[10] J. Ferreira *et al.*, "Driver behavior profiling: An investigation with different smartphone sensors and machine learning," *PLoS ONE*, vol. 12, no. 4, Apr. 2017, Art. no. e0174959. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0174959

[11] V. L. Neale, S. G. Klauer, T. A. Dingus, J. D. Sudweeks, and M. J. Goodman, "An overview of the 100-car naturalistic study and findings," in *Proc. Int. Tech. Conf. Enhanced Saf. Vehicles*, 2005, pp. 1–10.

[12] K. L. Campbell, "The SHRP 2 naturalistic driving study," Transp. Res. Board Nat. Acad., Washington, DC, USA, Tech. Rep., 2012, p. 8.

[13] R. Eenink, Y. Barnard, M. Baumann, X. Augros, and F. Utesch, "UDRIVE: The European naturalistic driving study," in *Proc. Transp. Res. Arena*, 2014, p. 10.

[14] *SHRP2 NDS Data Access*. Accessed: Dec. 7, 2019. [Online]. Available: https://insight.shrp2nds.us/home

[15] E. Carvalho *et al.*, "Exploiting the use of recurrent neural networks for driver behavior profiling," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 3016–3021.

[16] A. Cura, H. Küçük, E. Ergen, and I. B. Öksüzoglu, "Driver profiling using long short term memory (LSTM) and convolutional neural network (CNN) methods," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 10, pp. 6572–6582, Oct. 2021.

[17] B. Leblanc, S. Ercan, and C. D. Runz, "C-ITS data completion to improve unsupervised driving profile detection," in *Proc. IEEE 91st Veh. Technol. Conf. (VTC-Spring)*, May 2020, pp. 1–5.
[18] L. Wu, H. Li, H. Ding, and L. Zhang, "Who has better driving style: Let data tell us," in *Intelligent Systems and Applications* (Advances in Intelligent Systems and Computing), Y. Bi, R. Bhatia, and S. Kapoor, Eds. Cham, Switzerland: Springer, 2020, pp. 467–485.
[19] J. L. Obuhuma, "Driver behaviour profiling using dynamic Bayesian network," *Int. J. Mod. Educ. Comput. Sci.*, vol. 10, pp. 50–59, Apr. 2018.
[20] N. Arbabzadeh and M. Jafari, "A data-driven approach for driving safety risk prediction using driver behavior and roadway information data," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 2, pp. 446–460, Feb. 2018.
[21] S. Daptardar, V. Lakshminarayanan, S. Reddy, S. Nair, S. Sahoo, and P. Sinha, "Hidden Markov model based driving event detection and driver profiling from mobile inertial sensor data," in *Proc. SENSORS*, Nov. 2015, pp. 1–4.
[22] A. Abdelrahman, N. Abu-Ali, and H. S. Hassanein, "Driver behavior classification in crash and near-crash events using 100-CAR naturalistic data set," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2017, pp. 1–6.
[23] T. Liu, Y. Yang, G. B. Huang, Y. K. Yeo, and Z. Lin, "Driver distraction detection using semi-supervised machine learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 4, pp. 1108–1120, Apr. 2016.
[24] H. Miao, S. Zhang, and C. Flannagan, "Driver behavior extraction from videos in naturalistic driving datasets with 3D ConvNets," Nov. 2020, *arXiv:2011.14922*.
[25] A. E. Abdelrahman, H. S. Hassanein, and N. Abu-Ali, "Robust data-driven framework for driver behavior profiling using supervised machine learning," *IEEE Trans. Intell. Transp. Syst.*, early access, Nov. 17, 2020, doi: 10.1109/TITS.2020.3035700.
[26] A. Abdelrahman, N. Abu-Ali, and H. S. Hassanein, "On the effect of traffic and road conditions on the drivers' behavior: A statistical analysis," in *Proc. 14th Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, Jun. 2018, pp. 892–897.
[27] H. A. H. Naji, N. Lyu, C. Wu, and H. Zhang, "Examining contributing factors on driving risk of naturalistic driving using $K$-means clustering and ordered logit regression," in *Proc. 4th Int. Conf. Transp. Inf. Saf. (ICTIS)*, Aug. 2017, pp. 1189–1195.
[28] J. M. Hankey, M. A. Perez, and J. A. McClafferty. (Apr. 2016). *Description of the SHRP 2 Naturalistic Database and the Crash, Near-Crash, and Baseline Data Sets*. [Online]. Available: https://vtechworks.lib.vt.edu/handle/10919/70850
[29] A. Abdelrahman, H. S. Hassanein, and N. Abu-Ali, "Data-driven robust scoring approach for driver profiling applications," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–6.
[30] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

**Abdalla Abdelrahman** (Student Member, IEEE) received the B.Sc. degree in control and instrumentation systems engineering and the M.Sc. degree in electrical engineering from the King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia, in 2010 and 2013, respectively, and the Ph.D. degree from Queen's University, Kingston, ON, Canada, in 2019. He is currently working as a Post-Doctoral Fellow with the School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON, Canada. His work has been published in IEEE flagship conferences and top-tier journals. His research interests include driver behavioral modeling and profiling, intelligent vehicular systems, machine learning, and deep learning. He serves as a TPC member and a reviewer for IEEE flagship conferences and journals.

**Hossam S. Hassanein** (Fellow, IEEE) is currently the Founder and the Director of the Telecommunications Research Laboratory (TRL), Queen's University School of Computing, with extensive international academic and industrial collaborations. He is a leading authority in the areas of broadband, wireless and mobile networks architecture, protocols, and control and performance evaluation. His record spans more than 500 publications in journals, conferences, and book chapters, in addition to numerous keynotes and plenary talks in flagship venues. He has received several recognition and best papers awards at top international conferences. He is the Former Chair of the IEEE Communication Society Technical Committee on Ad Hoc and Sensor Networks (TC AHSN). He is an IEEE Communications Society Distinguished Speaker (a Distinguished Lecturer from 2008 to 2010).

**Najah Abu Ali** (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from The University of Jordan and the Ph.D. degree from the Department of Electrical and Computer Engineering, Queen's University, Kingston, Canada, specializing in resource management in computer networks. She is currently an Associate Professor at the Faculty of Information Technology, United Arab Emirates University (UAEU). Her general research interests include modeling wireless communications, resource management in wired and wireless networks, and reducing the energy requirements in wireless sensor networks. More recently, she has strengthened her focus on the Internet of Things, particularly at the nano-scale communications level, in addition to vehicle-to-vehicle networking. Her work has been consistently published in key publications venues for journals and conference. She has further coauthored a Wiley book on 4G and beyond cellular communication networks. She has also delivered various seminars and tutorials at both esteemed institutions and flagship gatherings. She has also been awarded several research fund grants, particularly from the Emirates Foundation, ADEC, NRF/UAEU funds, and the Qatar National Research Foundation.