Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 8819094, 11 pages
https://doi.org/10.1155/2021/8819094

*Research Article*

# Identifying and Labeling Potentially Risky Driving: A Multistage Process Using Real-World Driving Data

**Charles Marks**,[1] **Arash Jahangiri**,[2] **and Sahar Ghanipoor Machiani**[2]

[1] *Interdisciplinary Research on Substance Use Joint Doctoral Program,*
*San Diego State University and the University of California San Diego, San Diego, CA, USA*
[2] *Department of Civil, Construction, and Environmental Engineering, San Diego State University, San Diego, CA, USA*

Correspondence should be addressed to Arash Jahangiri; ajahangiri@sdsu.edu

Received 24 August 2020; Revised 15 December 2020; Accepted 27 January 2021; Published 15 February 2021

Academic Editor: Ladislav Routil

Every year, over 50 million people are injured and 1.35 million die in traffic accidents. Risky driving behaviors are responsible for over half of all fatal vehicle accidents. Identifying risky driving behaviors within real-world driving (RWD) datasets is a promising avenue to reduce the mortality burden associated with these unsafe behaviors, but numerous technical hurdles must be overcome to do so. Herein, we describe the implementation of a multistage process for classifying unlabeled RWD data as potentially risky or not. In the first stage, data are reformatted and reduced in preparation for classification. In the second stage, subsets of the reformatted data are labeled as potentially risky (or not) using the Iterative-DBSCAN method. In the third stage, the labeled subsets are then used to fit random forest (RF) classification models—RF models were chosen after they were found to be performing better than logistic regression and artificial neural network models. In the final stage, the RF models are used predictively to label the remaining RWD data as potentially risky (or not). The implementation of each stage is described and analyzed for the classification of RWD data from vehicles on public roads in Ann Arbor, Michigan. Overall, we identified 22.7 million observations of potentially risky driving out of 268.2 million observations. This study provides a novel approach for identifying potentially risky driving behaviors within RWD datasets. As such, this study represents an important step in the implementation of protocols designed to address and prevent the harms associated with risky driving.

## 1. Introduction

Each year, globally, traffic accidents result in 1.35 million deaths and 50 million injuries [1]. In 1998 in the United States, the National Highway Traffic Safety Administration (NHSTA) identified that aggressive driving behaviors occur in approximately two-thirds of all fatal car accidents [2]. Since then, multiple studies have corroborated the connection between aggressive driving behaviors and fatal car crashes [3–8]. The AAA Foundation found that, from 2003–2007, over half of fatal accidents were the result of aggressive driving behaviors [9]. In order to reduce the harms of aggressive driving behaviors, novel strategies for identifying such driving behaviors are required.

The concept of "aggressive driving" was formally defined in Meyer Parry's 1968 work, "Aggression on the Road," in which he stated that "the increasing stress involved in motoring nowadays makes the psychological efficiency of the driver a more important factor than the mechanical efficiency of the vehicle he drives" [10]. Looking at several studies on the topic, there is not a formal consensus on the definition of aggressive driving, but it ranges from acts of carelessness and recklessness to "road rage" [11–14]. One definition which captures these varying conceptions of aggressive driving is as follows: "A driving behavior is aggressive if it is deliberate, likely to increase the risk of collision, and is motivated by impatience, annoyance, hostility, and/or an attempt to save time" [15]. Since it is not usually possible to accurately assess the impatience, annoyance, or attitude of drivers at scale, it is generally simpler to focus on the middle of this definition—driving behaviors which increase the risk of collision. Therefore, the term "risky driving" was used in the present study instead of "aggressive driving." However, since aggressive driving has been used in several previous studies, the same terminology was used when referring to those.

While examples of risky driving, such as tailgating, running red lights, and speeding, are easily recognized [15], in practice, identifying real-world risky driving at scale is complicated by a lack of both data and strategies to properly assess said data. A video may catch a car running a red light and a GPS unit may record that its vehicle is speeding, but the steps required to take available data and identify patterns of risky driving behaviors require innovative strategies. This is especially important when dealing with "big data," which is currently limited in the transportation research literature.

With advances in technologies, the ability to collect large quantities of real-world driving data (RWD, such as the speed, acceleration, and heading of a vehicle across entire trips) has greatly increased. The use of machine learning strategies to try to identify and classify aggressive driving behaviors within these large RWD datasets is a field of budding interest. An array of supervised learning methods such as linear regression [16, 17], naïve Bayes classification [18], support vector machines [19], artificial neural networks [19, 20], dynamic time warping with k-nearest neighbors [21], random forests [22], and deep learning approaches [23] has been used to classify RWD data as either aggressive or not. Unsupervised methods such as k-means [24, 25], self-organizing maps (a type of unsupervised neural network) [25], and DBSCAN [26] have been incorporated into aggressive driving classification efforts, as well.

These studies represent important advancements in the efforts to identify aggressive driving from RWD data. Feng et al. used the measurement of longitudinal jerk in order to identify aggressive driving behaviors [16]. Wang et al. created an index to identify jerky driving movements as potential indicators of aggressive driving [17]. Jahangiri et al. identified aggressive driving while negotiating turns by modeling vehicles crossing lane stripes [22]. Several studies used RWD data collected from smartphones [18, 19, 21, 27]. Hong et al. and Johnson et al. used RWD data from smartphones to identify aggressive driving styles [18, 21]. Yu et al. identified the statistical profiles of specific types of aggressive driving (e.g., weaving, slamming the breaks, etc.) and used smartphone RWD data to train models to identify these behaviors [19]. Jeihani et al. leveraged a series of machine learning strategies to identify observations characterized by sudden changes in statistical profiles (i.e., sudden drops in speed and sudden turns) [28].

While these endeavors represent important steps in mitigating the harms of risky driving, for agencies and organizations dedicated to improving traffic safety, these individual studies do not provide a full account of all the necessary steps (such as restructuring RWD data for analysis and accounting for the large size of RWD data via time- and memory-efficient algorithmic choices) to identify risky driving behaviors from RWD data. Providing a guide to the implementation of risky driving classification strategies is necessary to ensure that agencies are empowered to utilize such strategies to improve traffic safety within their jurisdictions.

The overall purpose of this study is to demonstrate a multistage process for classifying observations in a large RWD dataset as potentially risky or not, using kinematic data only. We present four distinct stages in which the process is divided: formatting the data for analysis; labeling a subset of the data as potentially risky or not using unsupervised learning techniques; training supervised learning models on these labeled datasets; and, finally, using these models to label the remaining RWD data as potentially risky or not. At each step, we provide specific implementation details which can help inform future strategies for identifying potentially risky driving behaviors within RWD data. Thus, our approach first seeks to group observations by driving behavior (i.e., left turns, right turns, accelerating, and merging) and then seeks to identify outlying observations within each group. Further, while researchers and agencies may opt to utilize different specific tools and strategies within each phase of the classification process, the four overarching phases presented herein provide a novel approach for implementing risky driving classification. We note as well that future research should seek to confirm if the process we employ successfully identifies observations related to risky driving outcomes such as car accidents and traffic violations, and we provide recommendations for future steps in the discussion.

## 2. Data Description and Study Site

Data from the Safety Pilot Model Deployment (SPMD) study were obtained through the Research Data Exchange, via the U.S. Federal Highway Administration (and is now available through Data.gov) [29]. Data were collected during the months of October 2012 and April 2013 in Ann Arbor, MI, from nearly 3,000 vehicles. For this study, data from the first week of April 2013 were utilized and were subsetted to only include data within Washtenaw County (which is, conveniently, in the shape of a rectangle).

This study used basic safety messages (BSMs) transmitted by participating vehicles. BSMs were transmitted at a rate of 10 Hz and contain data on vehicle's state of motion (i.e., speed, acceleration, and yaw rate) and location. Specifically, data from the "BsmP1" file corresponding to April 2013 were used. This file is 204 GB with approximately 1.5 billion observations. For this study, a subset of this file was used corresponding to four weekdays and two weekend days in this first week and contained approximately 268 million observations. Data were stored locally on a PostGreSQL database and were accessed and manipulated using the R programming language. For further details about the "BsmP1" file, the metadata files are referenced [30, 31].

## 3. Methodology

The overall goal of this study was to design and present a protocol for identifying potentially risky driving behaviors within large RWD datasets. The primary logic of our approach is that the data profile of potentially risky driving behaviors will look quite similar to the data profile of nonrisky variations of the same behavior (i.e., a risky left turn and a not risky left turn will have similar data profiles) and then that potentially risky behaviors are those which are least normal for its given behavior (i.e., a potentially risky left turn would have a data profile which outlies the average data profile of all left turns in the dataset). As such, the process was divided into four primary stages: reformatting the unlabeled BsmP1 data subsets for analysis (one subset for each day); labeling subsets of the reformatted data as potentially risky or not using the Iterative-DBSCAN (I-DBSCAN) method; using the labeled subsets to train classification models (random forest) for each respective day; and, finally, using the classification models to label the entire day's corresponding data. Random forest was chosen after comparing it with logistic regression and artificial neural networks.

To begin, the BsmP1 data from April 1–7, 2013, were stored in seven different PostGreSQL tables, one for each respective day. Due to a compilation error, the table from Wednesday, April 3, was not included for analysis within this study. As such, the six tables of BsmP1 data corresponding to April 1–2 and 4–7 were utilized. We chose to analyze the data from each day separately for three primary reasons: first, as a matter of feasibility due to the large size of the data files; second, to ensure the reproducibility of the process we employed; and third, because we hypothesize that driving patterns on weekdays versus weekends are likely different (due to work commuting), and thus different types of risky driving behavior may emerge. Regarding the second, we note that consistent reproducibility—while not a reflection of accuracy—is an important feature to establish for any methodological approach. Regarding the third, we generated histograms of observations by time of day for both weekdays and weekends to confirm this hypothesis. Each of these tables (~2–5 GB) was too large to effectively analyze in *R*, and as such, for the first three stages of our process, a random subset of the data (~7–10% of full data) for each of the six days was selected. It was important to ensure that these random samples contained "full driving trips." If we simply pulled random observations, then there would be no guarantee that continuous sequences of observations would be extracted—in the stage one description, the importance of this will be clarified. The BsmP1 data includes unique vehicle IDs and, as such, we randomly selected 100 vehicle IDs for each day (representing ~7–10% of all vehicle IDs) and then extracted all observations corresponding to those vehicle IDs.

*3.1. Stage One: Reformatting Subsets.* Data were reformatted to address two issues: first, to ensure the data were in a format to best identify potentially risky driving; and second, to reduce the size of the data to improve the runtime feasibility of our labeling method in stage two. Regarding the first, the BsmP1 data are a set of observations measured at a rate of 10 Hz. What is readily apparent when considering these data is that the driving behavior of a vehicle cannot be understood by looking at individual *time-point* observations. A single observation does contain information about speed and acceleration and yaw, but it lacks the context of the full event it is contained within. As such, part of our reformatting process was to take continuous sets of 30 BsmP1 data points and merge them into single observations of *monitoring-period* data representing 3-second windows (30 observations of 10 Hz data correspond to 3 seconds). Regarding the second, these *monitoring-period* observations were generated at one-second intervals (1 Hz), meaning that the reformatted datasets contained 10% of the total number of observations as the original subsets. In Figure 1, we provide a visual depiction of how time-point observations (red diamonds) are converted into monitoring-period observations (blue and green rectangles) for a vehicle moving at a constant velocity—as can be seen, each monitoring-period rectangle contains thirty time period diamonds, with a new monitoring period beginning every ten time period diamonds.

The reformatting process for a single subset was as follows. First, the observations were organized by vehicle ID and then by time. We did not want to combine data corresponding to different vehicles, nor different trips from the same vehicle, so we split each vehicle's data by continuous trip. Since we sorted the data by time as well, we identified the start of new trips by jumps in the recorded time between observations. At this point, the data are divided into individual continuous trips. Then, for each of these trips, the *time-point* observations are merged such that at intervals of one second, three second's worth of data (i.e., thirty observations) were merged into a single observation. The *time-point* data measures of speed, acceleration, yaw, and heading were merged to create *monitoring-period* data measures of average, standard deviation, maximum and minimum values of speed, acceleration, and yaw rate, as well as overall change in heading and standard deviation of change in heading. An array of the unique data identifiers for the 30 *time-point* observations merged was generated as well. The reformatted datasets of *monitoring-period* data were used in the next stage.

*3.2. Stage Two: Labeling the Reformatted Data, an Unsupervised Learning Approach.* After reformatting, the data were ready to be labeled as potentially risky or not. This task was completed using an unsupervised learning approach, through two primary steps: first, by utilizing k-means clustering algorithm and change in heading thresholds to subset the data into elementary driving behaviors (EDBs); and, second, by utilizing the density-based spatial clustering of applications with noise (DBSCAN) clustering algorithm in an iterative fashion to identify potentially risky driving [32]. The underlying concept behind this approach is that there is a set of EDBs that occurs (such as accelerating, making a U-turn, merging onto the highway, etc.) and that these EDBs will likely have similar statistical profiles to one another. Potentially risky behaviors, then, were identified as the data points which were the further outliers from their prescribed cluster, as identified by running DBSCAN on each EDB cluster—this is meant to capture abnormal instances of EDMs.

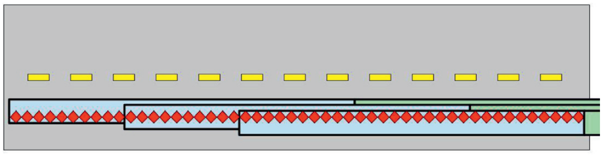

Figure 1: Converting TP data to MP data—using a vehicle moving at a constant velocity as an example. The red diamonds represent TP observations and the blue and green rectangles represent MP observations. Each MP observation contains 30 TP observations and a new MP observation begins every 10 TP observations. Of importance, it is to be noted that there is an overlap in each MP. The fourth, fifth, and sixth monitoring periods are colored green in order to improve visual readability of the figure—the color distinction does not hold further meaning.

The first of these two steps was to identify all EDB clusters within the data. To do this, we first subdivided the data by speed and change in heading. To divide by speed, we ran k-means using only the *average speed* variable to generate three distinct clusters (low, medium, and high speed). The data categorization based on speed has been conducted as a preparatory step in similar previous studies [17, 24]. Then, the data were further subdivided into five different turning classes based on change in heading (left and right turns (change in heading greater than 45 degrees); left and right curves (change in heading between 10 and 45 degrees); and straight (change in heading under 10 degrees)). Subsequently, k-means was run on each of these fifteen subsets, utilizing the sum of squared distances "elbow" method to identify optimal number of clusters (clustering variables were: average, maximum, and standard deviation of speed; average, maximum, minimum, standard deviation, and jerk of acceleration; and, average, maximum, minimum, standard deviation, and jerk of yaw rate). The results of this round of k-means represent the EDB clusters.

For each of the EDB clusters identified, DBSCAN was performed iteratively (I-DBSCAN) [26]. The idea is that, since the data have been clustered into EDBs, the data are dense and that each iteration of DBSCAN will cluster most of the data together. DBSCAN returns *n* clusters and one set of noise (i.e., unclustered data). One iteration of I-DBSCAN is as follows: first, DBSCAN is run on the dataset—the "elbow" method is utilized to determine the optimal epsilon parameter; second, the "normal" cluster is identified as the cluster consisting of at least 90% of the dataset—if no such "normal" cluster exists, I-DBSCAN is terminated and run again from the beginning; third, all data identified as noise are extracted and labeled as potentially risky; fourth, if any additional clusters have been identified, they are extracted and labeled as potentially risky—if no such additional cluster is identified, then it is checked if this is the third such time no additional cluster has been found and, if so, I-DBSCAN is terminated and the results are returned; finally, if not terminated, another I-DBSCAN iteration is undertaken utilizing the "normal" cluster as the dataset. In a sense, this process is like peeling the layers off of an onion, where the furthest outlying data points are "peeled away" and labeled as potentially risky and the dense set of data in the middle is labeled as not potentially risky. After I-DBSCAN has been run on all the generated EDB clusters, the labeled datasets are merged back together. After running I-DBSCAN on all EDB clusters and merging the results, we have labeled the entire dataset.

In order to complete this entire stage, software is needed to be written to streamline and automate the process. Since the "elbow" method utilized within both k-means and DBSCAN cannot be easily automated, an *R* script was written to semiautomate the labeling process as is described. The script written walked the user through the labeling process, prompting the user to input the values for the "elbow" method when necessary and automating all other aspects of the process.

*3.3. Stage Three: Predicting Risky Driving, a Supervised Learning Approach.* With the data labeled, the next stage is to train classification models to identify potentially risky driving behaviors. First, it was necessary to identify the optimal classification model to undertake this task. We opted to compare logistic regression, random forest, and artificial neural networks.

*3.3.1. Logistic Regression.* The logistic regression model is frequently used across the statistical sciences due to both its ease of implementation as well as the ability to extract estimates of causal relationships (in the form of log-odds ratios) [33]. Given a dichotomous outcome $Y$ with possible values of 0 and 1, it is of interest to calculate the probability (as a value from 0 to 1) that an event occurs ($Y = 1$), given a set of known predictors $X = \{x_1, x_2, \ldots, x_n\}$. A typical linear regression model, of which the outcome values range from $-\infty$ to $\infty$, is not appropriate for modeling dichotomous outcomes [33]. As such, the logistic regression model is defined as follows based upon the logistic distribution:

$$E(Y \mid X) = \frac{e^{\beta_0 + \beta_1 z_1 + \cdots + \beta_n z_n}}{1 + e^{\beta_0 + \beta_1 z_1 + \ldots + \beta_n z_n}}, \quad (1)$$

in which $E\{Y|X\}$ can be understood as the expected value of $Y$ given a set of predictors $X$ [33]. A labeled dataset consisting of dichotomous outcome $Y$ and set of predictors $X$ can be used to fit a logistic regression model, utilizing a maximum likelihood estimator, to calculate model coefficients $\beta = \{\beta_0, \beta_1, \ldots, \beta_n\}$. Once a logistic regression model has been fit, the model can be used to label a dataset consisting $m$ observations of predictors $X$. For each set of observations, $X_i = \{x_1, x_2, \ldots, x_n\}$, $E(Y|X_i)$ can be calculated, and this value is then assigned to each observation as the prediction of the probability that $Y_i = 1$ [33].

*3.3.2. Random Forest.* The random forest classification model is a powerful method to implement a form of "ensemble learning," in which many classification trees are generated and whose outputs are aggregated to generate classification predictions [34, 35]. Random forest is built upon the concept of "bagging," in which $n$ classification trees are generated independently of one another, each generated using a unique bootstrap sample of the training data set [35]. For binary classification, each of the $n$ trees is considered to have a vote, and the final classification of the observation is determined based on majority vote by the $n$ trees. In a standard classification tree, starting from a root node, each node is split based upon all predictors included in the model, but, in random forest, the split decision at each node is made using a random subsample of the available predictors [35]. As noted by Liaw and Wiener, "this somewhat counterintuitive strategy turns out to perform very well compared to many other classifiers, including discriminant analysis, support vector machines, and neural networks, and is robust against overfitting" [35].

As such, given a dichotomous outcome $Y$ with possible values of 0 or 1 and the training set of $m$ vectors of predictors $X$, $n$ classification trees are generated through the method described above. After being trained, predictions are generated as follows: each of these trees, $f_t$, given a new set of predictors $X'$, returns a value of either 0 or 1, denoted as $f_t(X') = \{0, 1\}$. The result from each individual tree is considered a vote. The result, either 0 or 1, which gets the most votes, $V$, is returned as the predicted value $Y'$ for the set of predictors $X'$. This can be understood mathematically as follows:

$$V = \frac{1}{n}\sum_{i=1}^{n} f_i(X') \longrightarrow Y' = \begin{cases} V < 0.5, & 0 \\ V \geq 0.5, & 1 \end{cases}. \tag{2}$$

*3.3.3. Artificial Neural Network.* Artificial neural networks arose in response to a digital conundrum: computers are able to solve mathematical computations at a rate that far exceed human capacity, but, simultaneously, cannot solve complex problems that humans are able to do so instantaneously [36]. The overarching concept is that the neural architecture of the human brain is well designed for answering complex questions, and as such, an algorithm replicating this architecture can similarly answer. For this project, we considered a feed forward single hidden layer neural network [37]. In such an architecture, there are three layers of neurons: the input layer, hidden layer, and output layer. The input layer corresponds to the input variables (i.e., one neuron for each variable). Each variable in the input layer is connected by a weighted flow, $w$, to each of the hidden layer neurons [37]. We used a grid-search approach to determine the optimal number of hidden layer neurons by ranging from 1 to the number of neurons in the input layer. Each of the hidden layer neurons is connected by a weighted flow, $\beta$, to the single output layer neuron [37]. As such, given $n$ input variables $X = \{x_1, x_2, \ldots, x_n\}$, $m$ hidden neurons, dichotomous outcome $Y$, and linear activation function $g$, the neural network can be defined as follows:

$$G(X) = \sum_{i=1}^{m} \beta_i g(w_i \cdot X + b_i), \tag{3}$$

where $w_i = (w_{i1}, w_{i2}, \ldots, w_{in})^T$ is the vector of flows connecting the $n$ input neurons to the $i^{\text{th}}$ hidden neuron, $\beta_i$ is the flow connecting the $i^{\text{th}}$ hidden neuron to the single output neuron, and $b_{i,}$ is the bias associated with the $i^{\text{th}}$ hidden neuron [37]. Given a sample with $L$ total observations, each with predictor sets $X_i$ and dichotomous outcome $Y_i$, the values of $w_i$, $\beta_i$, and $b_i$ are found by minimizing the distance between the model output and the actual outcome value, as follows [37]:

$$\sum_{i=1}^{L} G(X_i) - Y_i. \tag{4}$$

*3.3.4. Evaluating Best Model Fit.* In order to evaluate which of these three modeling approaches is best suited for predicting potentially risky driving behaviors, we ran 5-fold cross validation on the labeled subsets. In this process, the dataset is split into 5 groups. For each combination of four groups, the selected four groups are used to train the classification model and then we assess how well the model does at identifying potentially risky driving within the fifth group. The true positive rate and false positive rate of each iteration are calculated in order to create our primary evaluation metric, the area under the receiver operating curve (AUC). We repeated these 5-fold validations 25 times for each of the three classification models and extracted the average AUC scores and corresponding receiver operating curves. As a secondary outcome, runtime was extracted as well. As shall be discussed in the results, the random forest classification model outperformed others.

After it was determined that random forest was the best choice of classification model, a random forest model was fit for each of the six days of data (April 1–2 and 4–7).

*3.4. Stage Four: Labeling All the Data.* As the random forest models for each of April 1–2 and 4–7 were trained on subsets of BsmP1 data from each of those days, the random forests models were then used to label all of the data in each of these datasets. To do this, data were extracted from each of these datasets by vehicle ID, converted into monitoring-period data format (using the same procedure described in stage one), and then labeled utilizing the respective random forest model. These labeled datasets were then saved in the database by day. At this point, all of the BsmP1 data, reformatted into *monitoring-period* format, for April 1–2 and 4–7, were labeled as potentially risky or not. Since each *monitoring-period* observation included a reference to the 30 *time-point* observations merged to created it, the option is also then available to label the original BsmP1 observations as potentially risky or not (risky if they appear in any *monitoring-period* observations labeled as risky). As an additional analysis, we labeled each daily dataset with each of the other 5 random forest models (i.e., we labeled the April 1 dataset with each of the April 2 and April 4–7 datasets). We then calculated the proportion of the potentially risky observations observed by the daily model (i.e., the April 1st model labeling the April 1st dataset), which are also identified as risky by each of the other day's models. Finally, to better characterize differences between observations labeled as potentially risky and those that are not, we generated histograms of the distribution of two variables: acceleration jerk (derivative of acceleration) and yaw jerk (derivative of yaw). These values were calculated by comparing the first and last time point of each monitoring period. These variables were chosen because we hypothesize that risky driving behaviors will often be characterized by sudden changes in movement, which may be captured by changes in yaw and acceleration. Given large size of the datasets, we present the histograms with data corresponding to April 1.

## 4. Results

BsmP1 data were subsetted by calendar day, with a total of six subsets corresponding to April 1–2 and 4–7, 2013 (see Table 1 for number of data points in each table and corresponding number of vehicles). For analysis, 100 vehicle IDs were randomly selected from each day and all data corresponding to each vehicle ID and respective day were extracted (see Table 1 for size of 100 vehicle random sample). Due to technical database issue, the data corresponding to April 3 was not used. We had hypothesized, as well, that weekday and weekend driving patterns would be distinct, with weekday driving patterns being defined by peak driving activity during the morning and evening. In Figure 2, we show histograms of weekday and weekend driving observations by time of day, confirming this hypothesis.

### 4.1. Stage One: Reformatting the Data.

Each of the six subsets was converted from *time-point* observations into *monitoring-period* format. This resulted in the size of the datasets being reduced by an order of magnitude (see Table 2 for number of observations in each table before and after conversion, as well as the number of distinct continuous driving trips identified within each sample).

### 4.2. Stage Two: Labeling Subsets with I-DBSCAN.

The clustering protocol described was applied separately to each of the size reformatted datasets to label all points as either potentially risky or not. The proportion of each dataset labeled as potentially risky ranged from 8.25% to 10.0%, indicating that the clustering protocol behaved in a consistent fashion (see Table 3 for the crude number of data points and the proportion of data points labeled as potentially risky in each dataset).

### 4.3. Stage Three: Fitting Random Forest Models.

With the labeled data in hand, we then compared the performance of three different classification models at correctly identifying potentially risky driving points using 5-fold cross validation. Overall, we found that random forest outperformed both logistic regression and artificial neural network (see Figure 3 for AUROC of each model and Table 4 for mean AUC score and runtime of each classification model).

After identifying random forest as the best classification model, we fit distinct random forest models to each of the six labeled datasets. These random forest classification models correspond to each of the six days.

### 4.4. Stage Four: Labeling All the Data.

The six random forest models fitted in the prior stage were then used to label all of the data in the PostGreSQL database corresponding to the same day. Data were extracted by day and by vehicle, reformatted into *monitoring-period* structure, labeled using the corresponding random forest model, and then inserted into a new PostGreSQL table corresponding to the date of the observation. Table 5 shows the size of the original database tables, the size of the new reformatted, labeled tables, and the proportion of the entries labeled as potentially risky. In Figure 4, we present two heat maps corresponding to data from 250 randomly selected vehicles: one of all observations for these vehicles (left) and the other of the observations labeled as potentially risky.

Next, we sought to determine the performance of cross-applying each random forest model on each of the other datasets. In Table 6, we present the proportion of potentially risky driving behaviors that the same-day model originally found that the cross-day model also found. For example, the April 6 random forest model labeled 223,075 of the April 6 observations as potentially risky—the April 5 random forest model also labeled 72.6% of those 223,075 observations as potentially risky. Overall, the cross-day model always labeled at least 46.6% (ranging up to 80.2%) of the observations that the same-day model had labeled as potentially risky. This provides an indication that different potentially risky driving events occur across different days, and thus separate-day model training seems to be capturing those differences. There appears to be substantial variations by model and day, and thus future research efforts should seek to better understand these variations and improve upon them.

Finally, we sought to characterize differences between potentially risky and not potentially risky driving observations. We hypothesized that some risky driving events would be characterized by more sudden changes in motion and, thus, that the change in acceleration (acceleration jerk) and in yaw rate (yaw jerk) would, on average, be greater than that on nonrisky events. To assess this, in Figure 5, we present histograms of the distribution of the logarithm of acceleration and yaw jerk for both potentially risky and not potentially risky observations from April 1. Plots indicate that risky driving observations tended to be characterized by greater yaw and acceleration jerks. Given the hypothesis that risky driving behaviors are often characterized by sudden changes in movement, this provides initial validation that our approach appropriately identified such observations.

## 5. Discussion

Here we have presented a multistage process for taking a large, unlabeled RWD dataset and identifying observations representing potentially risky driving behaviors. Modern technological advancements have made bountiful data accessible to transportation researchers, but approaches and solutions to work with these data are requisite if we are to make meaningful improvements to transportation safety. We have shown how unsupervised learning methods—k-means, DBSCAN, and principal component analysis—and supervised learning

Table 1: Subsetting the BsmP1 data.

| Date | Database size[1] | Number of vehicles | 100-vehicle sample size[1] |
|---|---|---|---|
| Mon, April 1, 2013 | 44.5 | 1,395 | 3.61 |
| Tue, April 2, 2013 | 51.4 | 1,418 | 3.03 |
| Thu, April 4, 2013 | 50.0 | 1,430 | 3.27 |
| Fri, April 5, 2013 | 50.0 | 1,405 | 2.97 |
| Sat, April 6, 2013 | 39.7 | 1,133 | 3.37 |
| Sun, April 7, 2013 | 32.6 | 1,072 | 3.14 |

[1] Number of observations, in millions.

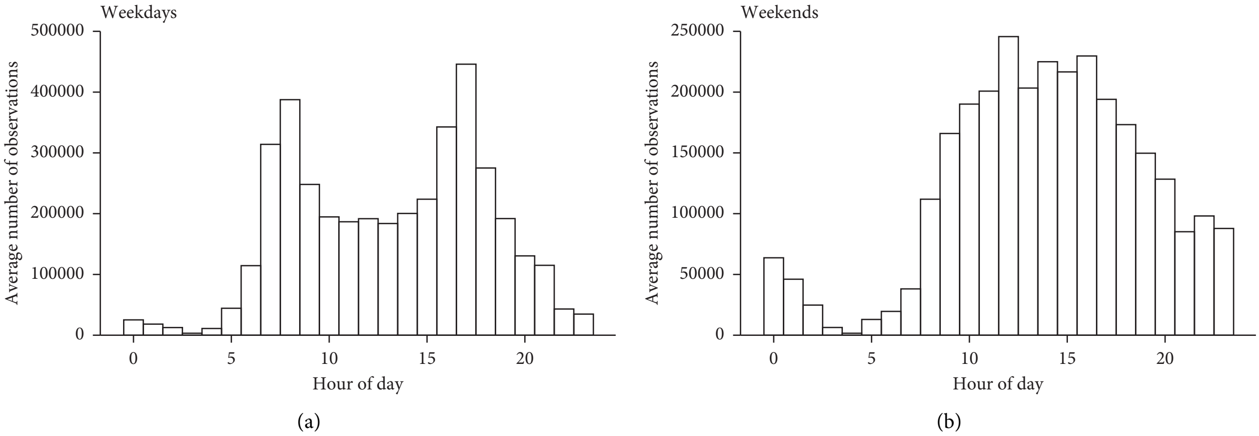

Figure 2: Histograms of observations by time of day for both weekdays (a) and weekends (b).

Table 2: Reformatting the data.

| Date | Dataset size prior to conversion | Dataset size after conversion | Distinct vehicle trips |
|---|---|---|---|
| April 1, 2013 | 3.61 million | 291,155 | 1,383 |
| April 2, 2013 | 3.03 million | 257,752 | 1,350 |
| April 4, 2013 | 3.27 million | 277,634 | 3,085 |
| April 5, 2013 | 2.97 million | 250,467 | 1,225 |
| April 6, 2013 | 3.37 million | 203,073 | 1,773 |
| April 7, 2013 | 3.14 million | 212,488 | 811 |

Table 3: Labeling risky driving data points.

| Date | Potentially risky data points | Proportion of dataset (%) |
|---|---|---|
| April 1, 2013 | 24,021 | 8.25 |
| April 2, 2013 | 23,063 | 8.95 |
| April 4, 2013 | 26,296 | 9.5 |
| April 5, 2013 | 25,227 | 10.0 |
| April 6, 2013 | 19,672 | 9.69 |
| April 7, 2013 | 19,666 | 9.26 |

methods—logistic regression, random forests, and artificial neural network—may be applied in a systematic fashion to identify potentially risky driving behaviors within RWD data. While not all RWD datasets will be structured identically, the four stages and details of their implementation provide transportation researchers and professionals the framework necessary to replicate this process and identify potentially risky driving within their own datasets.

While the process defined provides a procedure to identify potentially risky driving behaviors, there are immediate barriers to implementation that must be addressed if such a method is to be made more universally available. In order to undertake the stages as defined, our research team developed software tools in R. DBSCAN, principal component analysis, and k-means all require human interface to identify function parameters (via the "elbow" method), and given that these algorithms needed to be run many times, software which streamlined this process for our team aided in completing this project. As such, there is a need for software solutions which streamline the risky driving identification process. The steps outlined in this paper provide a

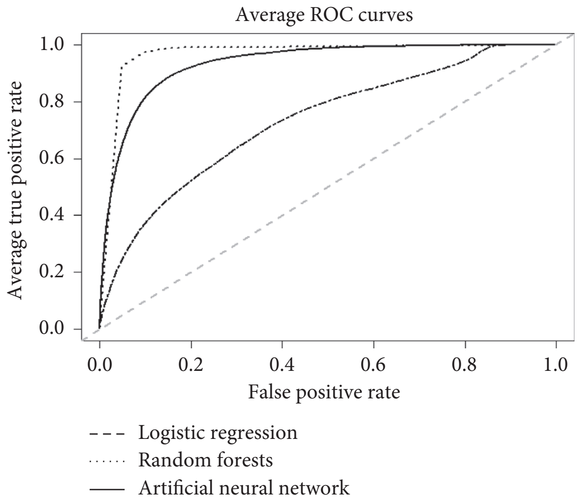

Figure 3: Mean ROC curves for 5-fold cross validation using logistic regression, random forest, and artificial neural network.

Table 4: Mean AUC score and runtime.

| Model | Mean area under ROC curve (AUC) | Runtime for single 5-fold iteration (s) |
|---|---|---|
| Logistic regression | 0.731 | 7.3 |
| Random forest | 0.982 | 87.6 |
| Artificial neural network | 0.927 | 483.0 |

Table 5: Risky driving data propositions.

| Date | Original database size | Labeled, reformatted database size | Proportion labeled potentially risky (%) |
|---|---|---|---|
| April 1, 2013 | 44.5 million | 3.92 million | 7.10 |
| April 2, 2013 | 51.4 million | 4.32 million | 7.54 |
| April 4, 2013 | 50.0 million | 4.60 million | 7.93 |
| April 5, 2013 | 50.0 million | 4.47 million | 8.90 |
| April 6, 2013 | 39.7 million | 2.92 million | 7.62 |
| April 7, 2013 | 32.6 million | 2.43 million | 6.89 |

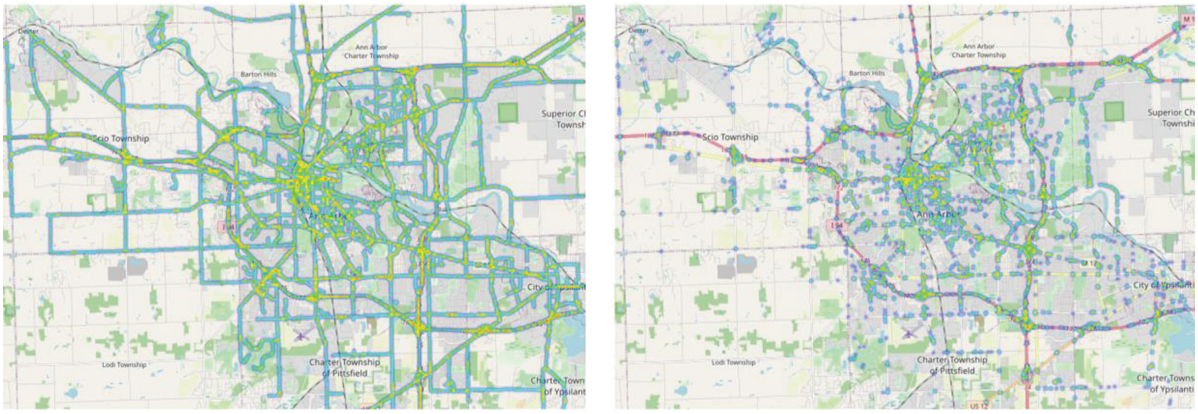

Figure 4: (a) Heatmap of all observations for 250 randomly selected vehicles. (b) Heatmap of all of these observations that were labeled as potentially risky.

novel approach for the implementation of such software solutions.

The applications of this method are immediate. By identifying potentially risky driving behaviors in RWD data, we can identify when and where potentially risky driving behaviors are most concentrated. This will provide transportation agencies real-time, actionable information to improve traffic safety within their given jurisdictions. It also

Table 6: Cross-classifying potentially risky driving behaviors *.

| | | Dataset labeled | | | | | |
|---|---|---|---|---|---|---|---|
| | | April 1 (%) | April 2 (%) | April 4 (%) | April 5 (%) | April 6 (%) | April 7 (%) |
| Random forest models | April 1, 2013 | | 49.1 | 65.7 | 47.1 | 46.6 | 52.8 |
| | April 2, 2013 | 52.0 | | 51.8 | 69.2 | 67.6 | 72.9 |
| | April 4, 2013 | 59.3 | 49.2 | | 47.7 | 50.0 | 57.0 |
| | April 5, 2013 | 56.3 | 73.6 | 56.4 | | 72.6 | 80.2 |
| | April 6, 2013 | 50.6 | 69.0 | 54.4 | 69.4 | | 73.4 |
| | April 7, 2013 | 50.4 | 65.7 | 53.7 | 68.8 | 66.8 | |

*Percentages represent the proportion of the originally labeled observations (by the same-day model) that the cross-day model also identified. We note that all cross-classifications labeled a similar proportion of each dataset as potentially risky (~5–10%).

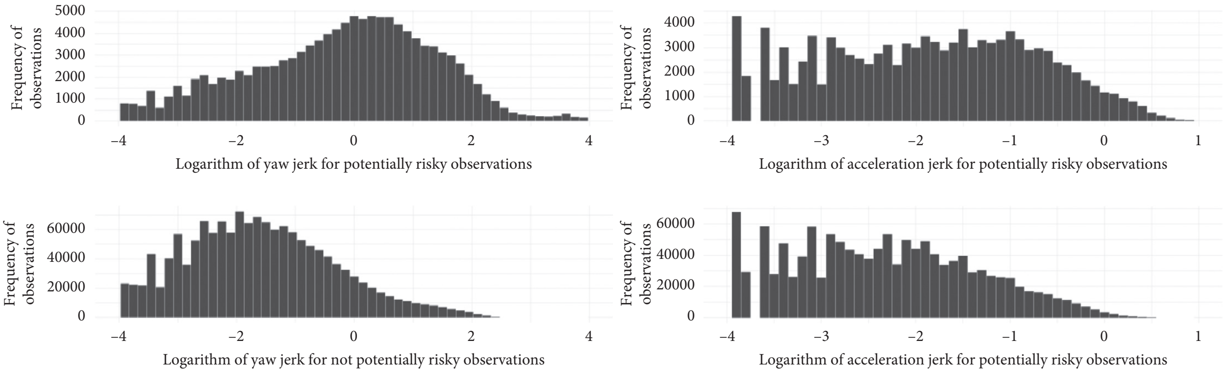

Figure 5: Histograms of the logarithm of yaw jerk (left) and acceleration jerk (right) for both potentially risky labeled observations (top) and not potentially risky labeled observations (bottom) for all observations recorded on April 1.

provides a way to measure the effectiveness of safety countermeasures (i.e., how much risky driving has been reduced after implementation of a desired countermeasure).

A primary limitation of this work is in regard to whether we have truly identified risky driving behaviors or not. The general idea is that, through k-means, we have identified clusters of each elementary driving behavior (EDB) and that potential risky driving points, identified using DBSCAN, are those observations which outlie their given cluster. We have assumed that risky driving behaviors will appear similar to their nonrisky counterparts (i.e., the macro-profile of a nonrisky left turn and a risky left will be very similar), but that when comparing observations of the same EDB, those risky driving behaviors will be identifiable by outlying statistics (i.e., a risky left turn may be identified by a greater acceleration than the average left turn). Future research steps should be taken to assess the external validity of the findings of this method. While we displayed that on average potentially risky driving observations labeled by our approach were characterized by higher yaw and acceleration jerk, future research should also seek to characterize individual EDB to better understand how the statistical profiles of potentially risky data points differ from those not labeled as such. Another limitation of the study was that the models developed were dependent on specific days. Separate-day models were trained, and it was shown that a model trained using a specific day can capture a minimum of 46.6% (up to 80.2% depending on the day) of potentially risky driving events on a different day. This raises a practical consideration in real-world use cases. Future work could focus on developing models for specific days (e.g., Mondays) across different weeks and investigate if, for example, a Monday model could consistently identify different potentially risky events if tested on a different Monday. A hypothesis to explore is that risky driving events are different (to some degree) across different days (i.e., Monday vs Friday) of week but very similar across same days of different weeks (Monday week 1 vs Monday week 2).

## 6. Conclusion

Overall, this study provides multiple contributions to the advancement of risky driving classification. The overarching steps outlined provide a novel approach by which RWD data can be formatted for and how unsupervised and supervised machine learning methods can be applied to the identification of potentially risky driving behaviors. Further, we have shown specifically how k-means, DBSCAN, and random forests may be applied in this endeavor. We evaluated the predictivity of random forests (in addition to logistic regression and artificial neural network), finding it to be highly sensitive and specific in predicting potentially risky driving behaviors. In sum, we have provided a meaningful process for the implementation of a risky driving classification program, a necessary tool in the efforts to improve traffic safety globally.

## Data Availability

The data used to support the findings of this study are publicly available at https://catalog.data.gov/dataset/safety-pilot-model-deployment-data.

## Disclosure

The contents of this paper reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein.

## Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

## Acknowledgments

This research was funded by the Safety through Disruption (Safe-D) National University Transportation Center (UTC), a grant from the U.S. Department of Transportation's University Transportation Centers Program (Federal Grant Number: 69A3551747115).

## References

[1] World Health Organization, *Global Status Report on Road Safety 2018*, World Health Organization, Geneva, Switzerland, 2018.

[2] National Highway Traffic Safety Administration, *Aggressive Drivers View Traffic Different Capital Beltway Focus Groups Find*, National Highway Traffic Safety Administration, Washington, DC, USA, 1998.

[3] C. Ma, W. Hao, W. Xiang, and W. Yan, "The impact of aggressive driving behavior on driver-injury severity at highway-rail grade crossings accidents," *Journal of Advanced Transportation*, vol. 2018, pp. 1–10, 2018.

[4] T. E. Boyce and E. S. Geller, "An instrumented vehicle assessment of problem behavior and driving style:," *Accident Analysis & Prevention*, vol. 34, no. 1, pp. 51–64, 2002.

[5] B. G. Simons-Morton, Z. Zhang, J. C. Jackson, and P. S. Albert, "Do elevated gravitational-force events while driving predict crashes and near crashes?," *American Journal of Epidemiology*, vol. 175, no. 10, pp. 1075–1079, 2012.

[6] S. Klauer, T. Dingus, V. Neale, J. Sudweeks, and D. Ramsey, *Comparing Real-World Behaviors of Drivers with High versus Low Rates of Crashes and Near Crashes*, National Highway Traffic Safety Administration, Washington, DC, USA, 2009.

[7] L. Evans, *Traffic Safety*, Science Serving Society, Bloomfield Hills, MI, USA, 2004.

[8] R. Paleti, N. Eluru, and C. R. Bhat, "Examining the influence of aggressive driving behavior on driver injury severity in traffic crashes," *Accident Analysis & Prevention*, vol. 42, no. 6, pp. 1839–1854, 2010.

[9] AAA Foundation for Traffic Safety, *Aggressive Driving: Research Update*, AAA Foundation for Traffic Safety, Washington, DC, USA, 2009.

[10] M. H. Parry, *Aggression on the Road: A Pilot Study of Behaviour in the Driving Situation*, Tavistock Publications, London, UK, 1968.

[11] L. Mizell, M. Joint, and D. Connel, *Aggressive Driving: Three Studies*, AAA Foundation for Traffic Safety, Washington, DC, USA, 1997.

[12] D. Shinar, "Aggressive driving: the contribution of the drivers and the situation," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 1, no. 2, pp. 137–160, 1998.

[13] K. H. Beck, M. Q. Wang, and M. M. Mitchell, "Concerns, dispositions and behaviors of aggressive drivers: what do self-identified aggressive drivers believe about traffic safety?," *Journal of Safety Research*, vol. 37, no. 2, pp. 159–165, 2006.

[14] S. K. Balogun, N. A. Shenge, and S. E. Oladipo, "Psychosocial factors influencing aggressive driving among commercial and private automobile drivers in Lagos metropolis," *The Social Science Journal*, vol. 49, no. 1, pp. 83–89, 2012.

[15] L. Tasca, "A review of the literature on aggressive driving research," in *Proceedings of the First Global Web Conference on Aggressive Driving*, Ontario, Canada, 2000.

[16] F. Feng, S. Bao, J. R. Sayer, C. Flannagan, M. Manser, and R. Wunderlich, "Can vehicle longitudinal jerk be used to identify aggressive drivers? An examination using naturalistic driving data," *Accident Analysis & Prevention*, vol. 104, pp. 125–136, 2017.

[17] X. Wang, A. J. Khattak, J. Liu, G. Masghati-Amoli, and S. Son, "What is the level of volatility in instantaneous driving decisions?," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 413–427, 2015.

[18] J.-H. Hong, B. Margines, and A. K. Dey, "A smartphone-based sensing platform to model aggressive driving behaviors," in *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, pp. 4047–4056, Denver, CO, USA, 2014.

[19] J. Yu, Z. Chen, Y. Zhu, Y. Chen, L. Kong, and M. Li, "Fine-grained abnormal driving behaviors detection and identification with smartphones," *IEEE Transactions on Mobile Computing*, vol. 16, no. 8, pp. 2198–2212, 2017.

[20] M. Shahverdy, M. Fathy, R. Berangi, and M. Sabokrou, "Driver behavior detection and classification using deep convolutional neural networks," *Expert Systems with Applications*, vol. 149, p. 113240, 2020.

[21] D. Johnson and M. M. Trivedi, "Driving style recognition using a smartphone as a sensor platform," in *Proceedings of the 2011 14th International IEEE Conference on Intelligent Transportation Systems*, pp. 1609–1615, Washington, DC, USA, 2011.

[22] A. Jahangiri, V. J. Berardi, and S. Ghanipoor Machiani, "Application of real field connected vehicle data for aggressive driving identification on horizontal curves," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 7, pp. 2316–2324, 2018.

[23] M. H. Alkinani, W. Z. Khan, and Q. Arshad, "Detecting human driver inattentive and aggressive driving behavior using deep learning: recent advances, requirements and open challenges," *IEEE Access*, vol. 8, pp. 105008–105030, 2020.

[24] A. Jahangiri, S. G. Machiani, and V. Balali, "Big data exploration to examine aggressive driving behavior in the era of smart cities," in *Data Analytics For Smart Cities*, pp. 163–182, CRC Press, Taylor & Francis Group, Boca Raton, FL, USA, 2019.

[25] J. Lee and K. Jang, "A framework for evaluating aggressive driving behaviors based on in-vehicle driving records," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 65, 2017.

[26] C. Marks, A. Jahangiri, and S. Ghanipoor Machiani, "Iterative DBSCAN (I-DBSCAN) to identify aggressive driving

behaviors within unlabeled real-world driving data," in *Proceedings of the 22nd Intelligent Transportation Systems Conference*, Auckland, NZ, USA, 2019.

[27] R. Akikawa et al., "Smartphone-based risky traffic situation detection and classification," in *Proceedings of the 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pp. 1–6, Austin, TX, USA, 2020.

[28] M. Jeihani, A. H. Pour, and A. Ardeshiri, *Machine Learning Model for Driving Distraction Detection*, Morgan State University, Baltimore, MD, USA, 2020.

[29] US Department of Transportation, *Safety Pilot Model Deployment Data*U.S. Department of Transportation, Washington, DC, USA, 2018.

[30] US Department of Transportation, *Safety Pilot Model Deployment–Sample Data, from Ann Arbor, Michigan, Version 1*, U.S. Department of Transportation, Washington, DC, USA, 2014.

[31] US Department of Transportation, *Safety Pilot Model Deployment Sample—Data Environment Data Handbook, Version 1.3*, U.S. Department of Transportation, Washington, DC, USA, 2015.

[32] M. Ester and H.-P. Kriegel, *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*, AAAI Press, New Orleans, LA, USA, 1996.

[33] D. Hosmer, S. Lemeshow, and R. Sturdivant, *Applied Logistic Regression*, Wiley, Hoboken, NJ, USA, 3rd edition, 2013.

[34] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[35] A. Liaw and M. Wiener, "Classification and regression by random forest," *R News*, vol. 2-3, 2002.

[36] A. K. Jain, J. Jianchang Mao, and K. M. Mohiuddin, "Artificial neural networks: a tutorial," *Computer*, vol. 29, no. 3, pp. 31–44, 1996.

[37] F. Lolli, R. Gamberini, A. Regattieri, E. Balugani, T. Gatos, and S. Gucci, "Single-hidden layer neural networks for forecasting intermittent demand," *International Journal of Production Economics*, vol. 183, pp. 116–128, 2017.