Anticipation and Alert System of Congestion and Accidents in VANET Using Big Data Analysis for Intelligent Transportation Systems

Hamzah Al Najada , Imad Mahgoub
Florida Atlantic University, 777 Glades Road, Boca Raton, FL 33431 USA
Email: halnajada2014@fau.edu, mahgoubi@fau.edu

***Abstract*—Vehicular Networks (VN) have a huge potential to increase roadway safety and traffic efficiency. Big Data analysis can be instrumental in realizing this potential and enhancing the Intelligent Transportation Systems (ITSs). We study the causes of road accidents using big real-time accidents data obtained from Florida Department of Transportation (FDOT) - District 4. The ultimate goal is to prevent or at least decrease traffic accidents and congestions. Our approach is based on dividing the roadway into segments, based on the infrastructure availability and the secondary accidents factors. We design a real-time Big Data system that receives online streamed data from vehicles on the road in addition to real-time average speed data from vehicles detectors on the road side to (1) Provide accurate Estimated Time of Arrival (ETA) using a Linear Regression (LR) model (2) Predict accidents and congestions before they happen using Naive Bayes (NB) and Distributed Random Forest (DRF) classifiers (3) Update ETA if an accident or a congestion takes place by predicting accurate clearance time. To make this system fast, accurate, and reliable we have implemented Lambda Architecture (LA) in our framework because of its speed, scalability, and fault tolerance. Furthermore, we have optimized the efficiency, the speed, and the accuracy of the designed model by securely selecting the most relevant and significant set of features required for the analysis.**

***Index Terms*—Big Data, Machine Learning and Data Mining, Vehicular Ad-hoc Network (VANET), Intelligent Transportation Systems (ITSs),Lambda Architecture (LA), Traffic Crashes Prediction, Traffic Congestions Prediction.**

I. Introduction

Today the world is overwhelming with all kinds of data from all domains. People have discussed and studied almost all aspects of Big Data such as their sources, characteristics, complexity, availability, privacy and a lot more. Big Data is the information of extreme size, diversity and complexity. Actually Big Data is not just a realm of voluminous data, it is a systematic way to collect, process, mine, and analyze the data to represent the uncovered trends, relationships, patterns, and insights in the data to make better decisions [1]. A formal definition of Big Data by International Data Corporation (IDC)is: "A new generation of technologies and architectures designed to economically extract value from very large volume of a wide variety of data, by enabling high-velocity capture, discovery and/or analysis"

In the domain of Intelligent Transportation Systems (ITS), there is so many potential solutions for traffic congestion and crashes. Congestion and crashes are lives, money, time, and environment drains [2]. Comprehensive cooperation between traffic engineers and computer scientists should be done to minimize the losses resulting from congestion and crashes.

One of the most important available sources for traffic data is the accidents data which usually has extreme size. Studying and analyzing this data will give us true insight, which will be helpful in designing new rules and policies to improve traffic flow and safety [3].

Increased attention has been directed in recent years towards Vehicular Networks (VN) or Vehicular Ad-hoc Networks (VANETs), because of its huge potentials, which could save lives, time, environment, and money as well. Furthermore, VANET would be one of the most important components of Intelligent Transportation Systems (ITS) [4] [5].

VANET is a special form of Mobile Ad-hoc Networks (MANETs) that formulate the framework of Intelligent Transportation Systems (ITS). In VANET, each vehicle represent a node in MANET. VANETs based on short-range wireless communication (e.g., IEEE 802.11p) between vehicles. The Federal Communications Commission (FCC) has allocated 75 MHz in the 5.9 GHz band for licensed Dedicated Short Range Communication (DSRC) [6], aimed at enhancing bandwidth and reducing latency for vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication. More elaboration about VANET could be found on [4]. Developing a real-time Big Data mining system for VANET to handle real-time streamed Big Data is very important. Such a system will enable countless applications.

In this work, we apply Big Data mining and analysis techniques to real-life accidents data in VN. We propose a system that aims at preventing or at least decreasing traffic congestions as well as crashes. Our method uses DSRC, cellular, wi-fi, and hybrid communication. The data analysis is carried out in the cloud that combines the historical data and the real-time data received from the vehicles operating on the road and queued into the lambda engine [7]. In this work, we have divided the roadway into segments to make the prediction process more precise based on a spatiotemporal technique.

The remainder of this paper is organized as follows. Section II presents the related work. Section III presents the challenges, problem components and the architecture for our proposed

system. Section IV presents our experiments, experimental results, and analysis. Finally, conclusions and future work are presented in Section V.

II. Related Work

Bedi and Jindal [8] studied and surveyed the use of Big Data technologies in VANETs. Their paper indicated that most of VANETs characteristics overlap with Big Data characteristics, allowing VANET problems to be treated and solved as Big Data problems. In addition, they motivated the need for building real-time adaptive models to optimize the traffic flow in an efficient and safe way.

As the era is mainly going towards the cloud solutions, data analysis can certainly benefit from this direction. A recent survey has been done on identifying and discussing the ability and the challenges of carrying out the analysis on clouds. Marcos et, al. [9] surveyed the approaches, technologies and the environment for carrying out analytics on clouds for Big Data. Changing the business model in analytics to a cloud based model has a great return on investment despite its complex high-level required resources and demands.

"How Fast is Fast?" . This question has been asked by Mike Barlow in his book [10]. Most of the literature agree that VANET should be treated as a Big Data problem. We believe that having a real-time analytics for VANET is key to solving traffic problems and catastrophes. Predicting accidents and congestions is not an easy task, mainly because "computation speed" is key challenge in this domain. The vehicles are moving fast and the topology is changing fast so, the computation, analysis, and the response should be done very fast. Mike in his book specified the predictive analytics procedure in five phases: Data distillation, model development, validation and deployment, real-time scoring, and model refresh. Claiming that following this procedure will provide the real-time analysis for streamed big data.

Because there is no single tool that runs a single command to give a solution or analyze multiple, diverse, and scattered datasets, Nathan Martz [7] designed Lambda Architecture (LA) to help in running and processing functions on the fly and get results from arbitrary datasets. LA has three layers: speed layer, serving layer, and batch layer. LA proved its scalability, robustness, ability for generalization, extensibility, and fault tolerance, which make it a good fit for real-time solutions in Big Data analytics [7].

A conceptual model for vehicular Big Data analysis has been proposed by Daniel et, al. [11] in which the author claims that such a model will keep pace with the latest trends to the emerging Big Data paradigm. The model aims at efficiently utilizing the huge datasets by near real-time data streaming in vehicular networks environment. This work has not used real-life data to test the model or validate the concept they came with. This work has proposed an algorithm to analyze the vehicle density on the road in real-time, but without any validation for this algorithm.

III. System, Challenges, Components, and Architecture

A. Challenges

Fortunately, the number of accidents is decreasing over the years in the USA. But, the National Safety Council estimates 38,300 people were killed and 4.4 million injured on U.S. roads in 2015, which saw the largest one-year percentage increase in half a century [12]. This assures that scientists should take serious actions to cut down this rise. Solutions should be proposed through incorporating the technology with the available infrastructure, since the available roads and infrastructure cannot be significantly expanded and people safety is the main target, in addition to the economical loss.

We can see that vehicular networks are going to be very important in providing significant solution to this problem. This can be achieved through giving proper timely alerts and directions to the drivers on the roads in advance since the vehicles are communicating interactively together. Taking into consideration that vehicles are moving very fast, and the topology for VANET is changing faster as well. VANET is a massive Big Data generator and we believe strongly that data always has trends and gives useful insights. We don't yet have VANET's data, but we have a valuable data right now in our hand, which is the accidents data that has rich information. Our work is based on a real data that has been obtained from Florida Department of Transportation - District 4 (FDOT-D4). The data has all information for more than 200 miles of the three highways (I-95, I-595, and I-75). The data has information about cameras, speed detectors, road devices, and all accidents information over a period of 7 years. This size of the data is challenging, and this poses additional challenges that should be overcome in order to get the right solution.

Primarily when it comes to mine the vehicles' data in real-time, two main issues raised, and solving them will strongly affect the system's efficiency. The first issue raised, comes from the fact that every car has its own mobile database of information (distributed). The second issue that the mining technique is centralized in one central server. A real-time system will face the problem of increased overhead on the network communication to send and receive alerts, new patterns and updating patterns. We are focusing in this research on getting the most efficient models, with less overhead on the network and less response time to get the maximum potential towards safer roads. In this paper we are improving the prediction results and decreasing the processing time needed to get the useful patterns and check the new patterns that sent out from any vehicle in real-time as well.

All transportation data is valuable and important to solve such a problem via Big Data analysis. But, what is the best vehicles' data that can be projected as vehicular networks data. We mentioned before that transportation data is an important source and key solution to our problem. However, not any transportation data will help us solve this problem. To solve the traffic congestion and accidents, we need traffic data. We found that accidents' data has all the attributes of the accidents

that ultimately caused the congestion on the road. So, we made our decision to work on accidents data.
Fortunately, we are able to obtain a big accidents dataset from FDOT - District 4 for this purpose. Preprocessing and cleansing the data was not an easy task, it was a challenge. The data was in relational database format, we had to convert it into CSV and ARFF Big Data readable file formats. We cleaned the data by getting rid of all incomplete records that can cause noise to the data. We analyzed and validated the data by using H2O and R Big Data tools [13] [14].
Another challenge that we have tackled in this work is the processing and computation speed. The prediction system would not be useful without being very prompt in sending and receiving packets. The transmitted packets will be either a prediction request or an alert response. For this purpose we have proposed using LA which we have mentioned earlier in our system to tackle the computation speed problem [7].

B. System Components

Our proposed model consists of a set of devices and software modules. They work collaboratively in a cloud based model to serve as a system for congestion and accidents prediction. The proposed model will provide alerts to the vehicles. The ultimate goal of this system is achieving the Intelligent Transportation System (ITS) goals, mainly the roadway safety and efficiency. Due to the space limitation we will describe the system components briefly and provide the appropriate references if possible.

1) **On-Board Unit (OBU):** Every mobile node is equipped with this unit to facilitate the communication with other moving nodes (vehicles) and fixed stations (Roadside Units) via DSRC and has the capability to communicate by using cellular radio networks such as (GSM, 4G, WiFi and WiMAX).
2) **Road-Side Units (RSUs):** These are base-stations that support VANET's applications and coordinate actions to share and process information as well as disseminate data, provide traffic directories, act as location servers, and connect to the Internet and external centralized or distributed servers [15].
3) **Lambda Architecture in the Cloud (LA):** This is a generic, fault-tolerant, and scalable processing architecture that best accommodates distributed data processing. LA consists of three layers: batch layer, serving layer, and speed layer [7].
4) **Vehicle Detection System:** Our case study is based on the FDOT - District 4 system data and infrastructure. They have a detection system consisting of roadside detectors placed every half mile on I-95, I-75, and I-595 highways. Those detectors provide online traffic data, such as speed, volume and occupancy to assist the Regional Transportation Management Center (RTMC) operators to detect abnormal traffic flows [16]. We use this data to calculate the average speed on each segment to update ETA for each registered trip on the road.

C. Framework Description

Our proposed work is mainly based on partitioning the roadway into segments in order to easily predict and manage the events occurring there. Selecting segment's length is a critical task, since long segments would pose certain challenges such as complexity and inefficiency. On the other hand, selecting short segments will not be helpful enough for the drivers to get early alerts which will make it difficult for them to change their route smoothly, if suggested. Furthermore, to make our segment's length selection decision reasonable and beneficial, secondary accidents causes, circumstances, and attributes are considered in this decision. A secondary accident is defined as an accident that takes place in the same direction within 2 miles or/and 2 hours of the primary accident [17]. So, segmenting the road into 2-mile segments will alleviate the secondary accidents, as well as will give a driver early alert and status about the road he/she is heading to. In our case study, using 2-mile segments gives a driver the choice to freely select at least one alternative route before he/she gets into a congestion, e.g. on I-95, the 2-mile marker will have 2 exits most of the time, which means 2 alternative routes.
The system begins to work when a driver starts the vehicle. By using the OBU the driver signs into the cloud via cellular or DSRC connection if possible (if an RSU is available) to register his/her trip by entering the desired destination. The cloud gets back to the driver with a random ID every time he/she signs in. Random ID is mapped to an IP address in order to identify the car, but the ID does not hold any driver's information in order to maintain the driver's privacy. Based on the navigation map and the road status updated in the cloud the driver will be able to select a route with the computed ETA.
Computing ETA in this model is based on the real-time average speed we get from the vehicle detectors for each segment that are placed every half mile of the highway [18]. Since we have selected the 2 miles marker we are going to use the sum of the expected time of travel (T) in each segment for the whole trip trajectory as described in equations 1 and 2.

$$ETA = \sum_{0}^{n-1} T_i \tag{1}$$

$$T_i = \frac{Distance_i}{Speed_i} \tag{2}$$

Where Ti is the time needed to travel segment i and n is the number of segments.
The aforementioned simply calculates ETA by getting a real-time average speed from the vehicles detectors, since all data is updated in the cloud and indexed by the segment number. We don't need to get all detectors data, we need to receive a detector data every 2 miles. Average speed should be updated whenever the average speed is increased or decreased by 5 miles/hour to update ETA accordingly. The other part in the cloud is the Big Data analysis engine, which has the largest impact on this system, since it is used to predict (1) Accidents

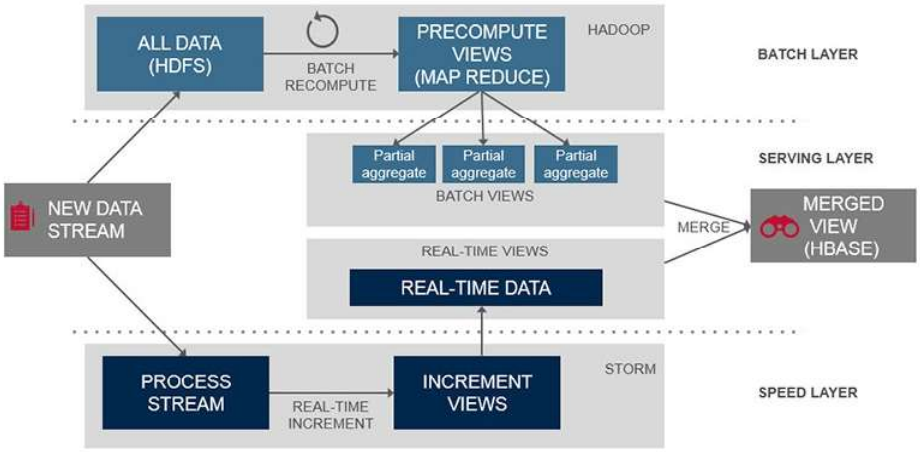

Fig. 1. Lambda Architecture framework [7]

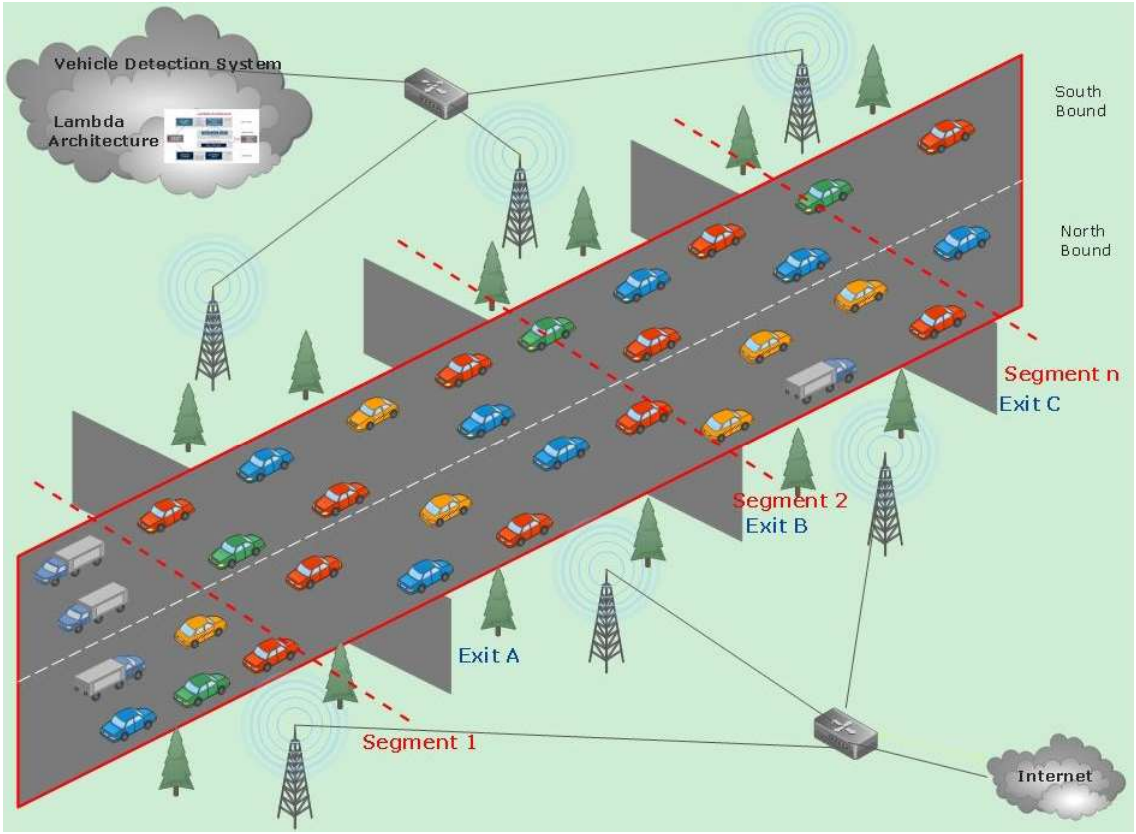

Fig. 2. The proposed system architecture and components

TABLE I
FAU TO FDOT - D4 TRIP ILLUSTRATION EXAMPLE

Seg. #	Seg. Len.	Speed	ETA	Entry	Departure
1	1.4 mile	25 mph	3.4	10:00:00	10:03:24
2	1.3 mile	45 mph	1.7	10:03:24	10:05:06
3	2 mile	65 mph	1.8	10:05:06	10:06:54
4	2 mile	65 mph	1.8	10:06:54	10:08:42
5	2 mile	65 mph	1.8	10:08:42	10:10:30
6	2 mile	65 mph	1.8	10:10:30	10:12:18
7	2 mile	65 mph	1.8	10:12:18	10:14:06
8	2 mile	65 mph	1.8	10:14:06	10:15:54
9	1 mile	65 mph	0.9	10:15:54	10:16:48
10	3.1 mile	45 mph	4.1	10:16:48	10:20:54

(2) Congestion (3) Clearance time when an accident happens. This engine has historical data and a customized prediction model. As well, it receives streams of real-time data from vehicles operating on the roads to provide the analysis and predict the road status. Next section will have more details about this analysis engine. The Big Data analysis part in this proposed system is the most important one because by getting the data and providing the predicted response in a timely manner we could save a lot of lives, money, time, and save the environment. Our proposed model will calculate in addition to ETA, the time of arrival and departure to and from each segment in the planned trip.

For example, in Fig. 3 we can see a planned trip From FAU to FDOT-District 4 at 10:00 a.m. It is an 18.8 miles trip. Inside the city ETA will be calculated for each segment as the length of each segment divided by the speed limit assigned to that segment. On the highway, the speed will be the actual real-time average speed reported by the vehicle detectors. In our example we use the highway's speed limit of 65 mph and a trip of 10 segments for the entire trip as illustrated in Table I and Fig. 3. In addition, we are going to have the entry and departure time to and from each segment in order to use this data in our prediction model to predict the accidents early in a spatiotemporal scenario to prevent them and avoid the congestion.

Our proposed system in the cloud will have the spatiotemporal entries for the anticipated vehicles once they register their trips. The system will stream the data for each segment along with other features such as number of vehicles, weather condition, road surface status, day light, day of the week ... etc. to predict how likely an accident will happen. This process is discussed later in the following section. According to the results, ETA and current route either remain as they are if no accident is predicted or get updated if an accident is predicted.

D. Big Data Analysis (Lambda Architecture)

As we have mentioned before, our model is based on real-time Big Data analysis for road segments. We are totally aware that this critical part would not succeed without being computationally prompt and very fast. For this purpose, first, we have segmented the roads into 2-mile segments

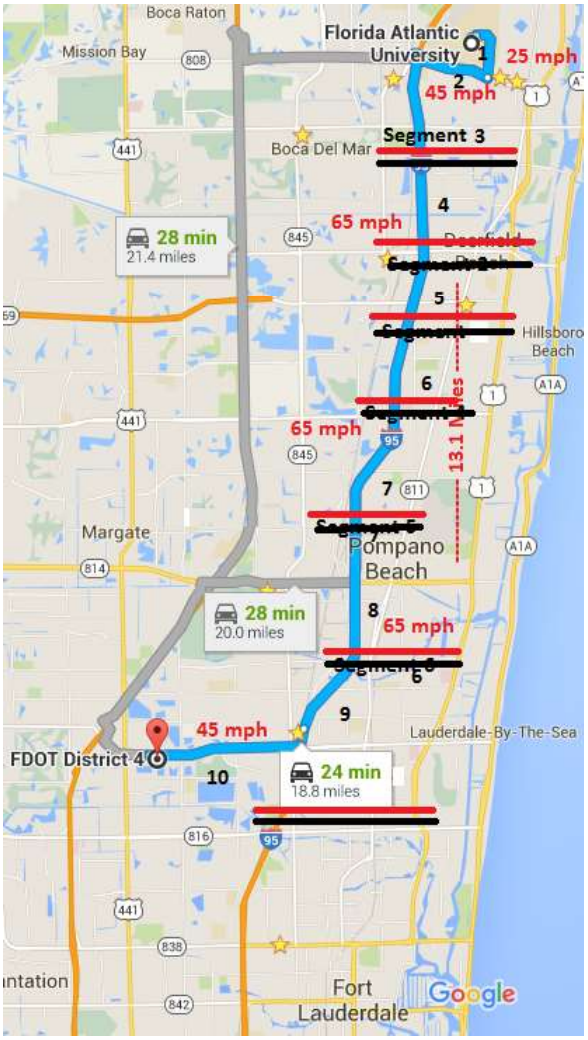

Fig. 3. Example of a trip to illustrate the road segmentation concept

regardless of the available number of lanes. Secondly, we have utilized the Lambda Architecture [see Fig. 1] in our proposed model since it has been proved as a fast, scalable, and fault-tolerant scheme for data processing [7], [19], [20]. We are proposing to use our own cloud resources (storage, processing, sending, and receiving) and our own-built cloud applications to connect with external resources. Our master dataset (accidents dataset) will be stored and clustered on our cloud within LA in the batch layer. In the batch layer the data will be divided into number of batches "batch views", since each batch has the most closely related examples and the batching process continues as long as it receives new data.

The speed layer in LA is the real-time layer, which computes views from the data it receives. The serving layer will query the real-time views from the speed layer quickly. Furthermore, the serving layer has all batch views from the batch layer indexed to provide low latency query answers. Practically, (1) A new example is streamed into the speed layer in real-time (from the vehicle) (2) At the same time the same examples will be stored in the main dataset (3) The new example,

which has the required features moves to the serving layer to query (compare) this batch with the previously generated batches (4) The serving layer gets the solution (answer) back to the query (i.e. predict accident, predict congestion, predict clearance time or predict new ETA).

IV. Experiments, Results and Analysis

In addition to LA, the cloud model will have the previously built models for prediction as well. From our previous analysis [2], we have selected two prediction models to implement in our framework. For the reason why we have selected those two prediction models here, please refer to our previous study [2], which has used accidents data. First, we have selected the Linear Regression model [21] to predict the clearance time after an accident takes place. This aids in updating ETA along with providing the alternative route which is assigned to each segment. Secondly, for congestion and accidents prediction we have selected Naive Bayes (NB), since it is very fast and reliable. In addition, we are using Distributed Random Forest (DRF) because it runs efficiently and unbiased on Big Data [21].

A. Linear Regression Model

The equation below is the linear regression equation we have computed to optimally predict the clearance time required after feeding the system with the required parameters. The Mean Squared Error (MSE) [see equation 3] obtained by using this model with full parameters is equal to 78.46 and the model building took 11.26 seconds. In this regard we need to minimize the time needed to build the model as well as decreasing MSE. To solve this and decrease the computation time we have done a Correlation-based Feature Selection (CFS) [22]. This decreased the number of features to 4 features instead of 21. This decreased MSE to 67.51 and the computation time to 3.37 seconds. Equation 4 shows the linear regression equation before applying CFS. Equation 5 shows the linear regression equation after applying CFS on the full dataset. Clearly, using fewer features improves the speed of sending the features packet from the vehicles to the cloud, as well as decreases the computation time on the cloud. The purpose of calculating the clearance time is to: (1) Update ETA for registered trips if the driver choses to use the same route (2) Give the most accurate time depending on the accident severity to update the digital signs and traffic status accordingly, and (3) Measure the efficiency of the accident's responders in clearing and closing the accident.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 \tag{3}$$

Where n is the total number of examples, $\hat{Y}$ is the predicted value and Y is the actual value.

From Fig. 4 we can see that most of the accidents have been cleared and closed within less than 30 minutes. Our prediction results support this fact as well, which means that predicting the clearance time will strongly assist in providing more accurate ETA.

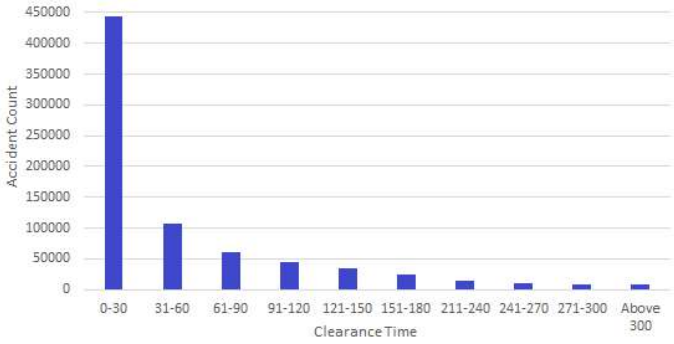

Fig. 4. Actual Clearance Time computed from the accidents dataset

Actual Clearance Time =
(67.1377*is_cloned) +
(-8.1484*eventstatus_id) +
(0.0051*ownedby_org_id) +
(-0.0206*notifier_agency_id) +
(-0.0087*notifier_contact_id) +
(-0.4955*eventtype_id) +
(21.8524*is_hazmat) +
(15.9843*is_fire) +
(18.9766*is_rollover) +
(-7.8191*has_noapplicable_cctv) +
(-0*primary_event_id) +
(10.0435*injurytype_id) +
(0*point_latitude) +
(0*point_longitude) +
(21.1547*can_publish) +
(-0.5688*atis_severity_level) +
(-0.748*day_of_the_week) +
(0.0482*event_month) +
(42.723) (4)

The Regression equation after doing CFS is as following:

Actual Clearance Time =
(-0.4826*eventtype_id) +
(0*primary_event_id) +
(10.9234*injurytype_id) +
(0*point_latitude) +
(7.4681) (5)
(Table II provides the attributes description)

B. Classification Models

We have evaluated our model for predicting the accidents using the full set of features. We apply FS and then compare two different classifiers namely: NB and DRF. All results have been obtained by doing 10-fold cross-validation on our accidents dataset. The class attribute in this model is the accident severity. Accident severity in the original dataset has 4 different values: minor, intermediate, major, and NULL. The first three severities imply an accident, the last one implies a

TABLE II
DESCRIPTION OF THE USED ATTRIBUTES

#	Attribute	Description
1	Is cloned	if congestion happens after the accident takes place
2	Event Status	Active, Closed, or Unresolved event
3	Owned by org ID	the agency in charge and responsible to manage the accident
4	Notifier	the first individual or agency that notifies about the accident
5	Event type	crash, road work, flood .... etc
6	HAZMAT	if there is hazardous material
7	Rollover	if a vehicles rollover happens
8	Fire	if a fire happens in the accident scene
9	Latitude and Longitude	to show the location of the accident
10	Severity	slight, moderate, severe, and fatal

TABLE III
DRF AND NB CLASSIFICATION RESULTS

Classifier	Before FS				After FS		
	Acc	AUC	F		ACC	AUC	F
NB	49.5	76.9	81.1		95.3	93.2	98.4
DRF	95.18	97.6	98.5		95.3	96.5	98.6

false alarm, which means this is not an accident. Furthermore, we have added false alarm accidents to our dataset marked as no-accident. The full dataset has 775271 examples. The NB classifier had the lowest time of computation. Table III shows the summarized results of classification for all classifiers before and after doing the feature selection. From the results obtained in Table III we can see that applying feature selection reduced the NB computation time. For DRF even-though the classification results are better that the NB results, it took more time. Feature selection didn't affect DRF results because DRF has a built-in feature selection functionality since it uses pruning.

In VANET especially for real-time mode, having less features to collect and using NB would be good in providing most likely the right decision quickly. The alert will be sent to the participating vehicle and the driver have the choice either to accept it or reject it.

V. CONCLUSION AND FUTURE WORK

VANET is a huge Big Data generator. It seems that the mechanism used in this work is easy and straightforward, indeed it is not. The idea of projecting the transportation data as the futuristic connected vehicles data is novel. Recent research found that the properties of VANET intersect and clearly map to the Big Data properties and characteristics. In order to get the maximum benefits from VANET, applications should be designed and developed in robust and scalable modes. Scalable because vehicles are increasing everyday which means data is growing bigger and bigger. The best robust systems are those developed based on real-time data rather than simulated data. In this paper, we propose a real-time Big Data analysis and prediction system, which directly assists in decreasing accidents on the roads and saves lives. Our proposed model is based on real-time accidents' data obtained from FDOT - District 4 for research purposes. Our study has taken into consideration primary and secondary accidents as well as the availability of infrastructure.

The proposed system has a component that receives real-time average speed values from the vehicle detectors to provide the most accurate ETA. The system is cloud based, where LA is proposed in the cloud to foster the Big Data analysis when data is streamed from vehicles. The proposed system has two main functions one is before the accident happens and the other one is after it happens. The former is to prevent the accident from taking place by prediction, which will prevent congestion as well. The latter, will assist in giving the time needed by both drivers and responders who are in charge of clearing the accident to clear the accident. In addition, the system can help in predicting the severity of the accident, which is important to get the needed resources efficiently. Our experiments and results show that applying such a system will decrease accidents significantly, since our classification results have high accuracy and low latency. Our future direction is to design a human factor impacts module to be added to this system based on and inferred from our real-time data. Implementing this system on a real testbed is an ongoing research in our lab by utilizing our designed cloud on our own clusters. Our focus will be on efficiently handling big growing heterogeneous data from all possible sources with heterogeneous behaviors.

ACKNOWLEDGMENT

This work is part of the Smart Drive initiative at Tecore Networks lab, at Florida Atlantic University.

We thank Florida Department of Transportation - District 4 for providing 7 years of real-life incidents database.

REFERENCES

[1] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques: concepts and techniques*. Elsevier, 2011.

[2] H. Al-Najada and I. Mahgoub, "Big vehicular traffic data mining: Towards accident and congestion prevention," in *IWCMC 2016 Smart Cities & Platforms (IWCMC 2016 Smart Cities & Platforms)*, Paphos, Cyprus, Sep. 2016.

[3] M. A. Abdel-Aty and A. E. Radwan, "Modeling traffic accident occurrence and involvement," *Accident Analysis & Prevention*, vol. 32, no. 5, pp. 633–642, 2000.

[4] C. Campolo, A. Molinaro, and R. Scopigno, "Vehicular ad hoc networks,"

[5] M. Gerla and L. Kleinrock, "Vehicular networks and the future of the mobile internet," *Computer Networks*, vol. 55, no. 2, pp. 457–469, 2011.

[6] U. D. of Transportation, *Connected Vehicles Dedicated Short Range Communications*, Updated August 17, 2015. [Online]. Available: http://www.its.dot.gov/DSRC/dsrc_faq.htm

[7] N. Marz and J. Warren, *Big Data: Principles and best practices of scalable realtime data systems*. Manning Publications Co., 2015.

[8] P. Bedi and V. Jindal, "Use of big data technology in vehicular ad-hoc networks," in *Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on*. IEEE, 2014, pp. 1677–1683.

[9] M. D. Assunção, R. N. Calheiros, S. Bianchi, M. A. Netto, and R. Buyya, "Big data computing and clouds: Trends and future directions," *Journal of Parallel and Distributed Computing*, vol. 79, pp. 3–15, 2015.

[10] M. Barlow, *Real-time big data analytics: emerging architecture*. " O'Reilly Media, Inc.", 2013.

[11] A. Daniel, A. Paul, and A. Ahmad, "Near real-time big data analysis on vehicular networks," in *Soft-Computing and Networks Security (ICSNS), 2015 International Conference on*, Feb 2015, pp. 1–7.

[12] S. Z. Newsweek.com, *U.S. TRAFFIC DEATHS, INJURIES AND RELATED COSTS UP IN 2015*, 2015. [Online]. Available: http://www.newsweek.com/us-traffic-deaths-injuries-and-related-costs-2015-363602

[13] H2O, *Fast Scalabale ML API for Smarter Applications*, 2015. [Online]. Available: http://h2o.ai

[14] M. OpenCourseWare, *R for machine learning*, 2015. [Online]. Available: http://ocw.mit.edu/courses/sloan-school-of-management/15-097-prediction-machine-learning-and-statistics-spring-2012/lecture-notes/a

[15] K. Mershad, H. Artail, and M. Gerla, "Roamer: Roadside units as message routers in {VANETs}," *Ad Hoc Networks*, vol. 10, no. 3, pp. 479 – 496, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1570870511001922

[16] R. F. D. 4, *SMART SunGuide Regional Transportation Management Center (RTMC)*, 2008. [Online]. Available: http://www.smartsunguide.com/SunGuideTMC.aspx

[17] A. Khattak, X. Wang, and H. Zhang, "Are incident durations and secondary incidents interdependent?" *Transportation Research Record: Journal of the Transportation Research Board*, no. 2099, pp. 39–49, 2009.

[18] D. V. Arnold, B. R. Jarrett, T. W. Karlinsey, J. L. Waite, B. Giles, and R. S. John, "Detecting roadway targets across beams," Jan. 19 2016, uS Patent 9,240,125.

[19] L. Magnoni, U. Suthakar, C. Cordeiro, M. Georgiou, J. Andreeva, A. Khan, and D. Smith, "Monitoring wlcg with lambda-architecture: a new scalable data store and analytics platform for monitoring at petabyte scale." in *Journal of Physics: Conference Series*, vol. 664, no. 5. IOP Publishing, 2015, p. 052023.

[20] M. Quartulli, J. Lozano, and I. G. Olaizola, "Beyond the lambda architecture: Effective scheduling for large scale eo information mining and interactive thematic mapping," in *Geoscience and Remote Sensing Symposium (IGARSS), 2015 IEEE International*. IEEE, 2015, pp. 1492–1495.

[21] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[22] M. A. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *Proceedings of the Seventeenth International Conference on Machine Learning*, ser. ICML '00. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000, pp. 359–366. [Online]. Available: http://dl.acm.org/citation.cfm?id=645529.657793