EAI Endorsed Transactions
on Energy Web

Research Article EAI.EU

RoadSpeedSense: Context-Aware Speed Profiling from Smart-phone Sensors

Ratna Mandal[1,*], Pallav Sonowal[1], Manish Kumar[1], Sujoy Saha[1] and Subrata Nandi[1]

[1] National Institute of Technology, Durgapur, India.

Abstract

INTRODUCTION: There are several online mapping systems like Google Maps, Waze, Here, Apple Maps, Bing Maps, etc. which are developed to visualize real-time traffic conditions which rely on crowdsourced GPS trails; obtained from worldwide smartphone users. Such systems still suffer from some limitations like a) inadequate traffic information in suburban cities and rural zones, b) system failure to infer the proper reasons for slow traffic state, c) difficulties in the extraction of raw traffic data for further development of any customized application. Significant spatio-temporal similarity patterns are observed in city traffic behavior unless there are some exceptional events like any disaster, VIP visit, international cricket match or bad weather condition, etc.
OBJECTIVES: Designing a framework and developing a system which enables collection of raw sensor information from users and to identify a model to generate a speed profile of city roads using historical logs as well as to infer the context of slow traffic based on ambient subjective road features and to provide map visualization.
METHODS: We have used road surface quality, density of vehicles, type of neighborhood and road geometry for developing speed profile for a particular road segment. We have carried out the experiments on different classification algorithms like, K-nearest Neighbor(KNN), Decision Tree(DT), Random Forest(RF) and Gradient Boost(GB) with necessary tuning of parameters.
RESULTS: GB outperforms other classification algorithms in estimating the speed class of road segments among all classifier algorithms with highest *F1- Score* of 0.8345. A fair driver rating system which can be derived from our results.
CONCLUSION: The results obtained from the proposed novel framework provide a proof of concept that speed profiles may be successfully derived from ambient road features even when sample space is sparse.

Keywords: Speed Profiling, Honk, WiFi, Road Condition, Intersection Density, Interpretable Machine Learning

Received on 31 August 2019, accepted on 04 January 2020, published on 24 January 2020

doi: 10.4108/eai.7-1-2020.162802

*Corresponding author. Email: ratna.mandal.iem@gmail.com

1. Introduction

"If you make more roads, you will have more traffic"- Jan Gehl

Understanding traffic behavior and estimating traffic-related parameters[1] like speed, volume, flow, density, and modelling their correlation has always been challenging. A precise estimation of traffic parameters and modelling of traffic behavior is the key to developing services like travel time estimation, predicting traffic congestion, route recommendation, and city traffic planning. These days there had been several online mapping systems like Google Maps[2], Waze[3], Here[4], Apple Maps, Bing Maps[5], etc. which are developed to

visualize real-time traffic conditions. Additionally, many of these systems provide services like route recommendations along with driving directions, show rerouting options, give transit information, assist in locating parking lots, etc. Some systems also allow users to report and localize several traffic instances like road accidents, hazards, weather conditions, police warnings, etc. on map.

The systems mentioned above rely on crowdsourced GPS trails; obtained from the smartphone sensors; recorded by their massive worldwide user base. However, such systems still suffer from some limitations, as following. *Firstly*, the traffic information is inadequate and is often completely missing in many areas of smaller suburban cities and rural zones[6], especially along by-lanes, where the density of users using such a system is still very sparse. *Secondly*, the traffic map systems mostly provide visualization of color-coded traffic information[7] [8], e.g., Green (Free Flow - vehicles can move easily at particular speed limit), Orange (Light traffic - slight congestion but drivers can still travel at relative speed), Red (Heavy Traffic - expect significant delays) and Dark red (Traffic is near standstill). It may be noted that these colors encode real-time space-mean speed[9] [10] value for a road segment. Here, the reason for lower speed ranges is always interpreted as traffic congestion. However, studies suggest that vehicle speed is significantly affected by the roughness of the road surface and the intensity of light at night. Vehicles may be forced to travel at a slower speed due to extremely poor road surface conditions[11]. Hence, present systems fail to infer the proper reasons for lower speed levels of traffic. *Thirdly*, although these systems are generally free to use, there is no way (or at least no reasonably easy and convenient way [7]) to extract the raw traffic data or information like current and average traffic speeds, etc. for further development of any customized application. Developers are free to use API services, which are also not available[12] [13]. Open-source software systems like OpenTraffic[14] is designed to process and aggregate a global, freely available traffic speed data, but it too has almost zero coverage in developing cities or towns. Hence, developing novel 3rd party services for such areas is still fancy.

Objective - To overcome the above shortcomings, this paper aims to design a framework and to develop a system which enables collection of raw sensor information from users and to identify a model to estimate the space- mean speed[15] using historical logs as well as to infer the context for lower speed levels and to provide map visualization. Instead of dealing with data sparsity problem, the proposed system will use historical data logs to provide a semi-real time traffic information for a given road segment at a particular time of day. The information will be sufficient enough for developing services as speed profiles generally show significant spatio-temporal similarity patterns unless there occur some exceptions, i.e., events like festive occasions, accidents, sports events, etc.

Issues and Challenges: The primary issue in designing the system is to deal with the sparsity of mobility logs that will be available for speed estimation for a given road segment. Moreover, as driving maneuvers widely vary among drivers, it may happen that for a road segment at a particular time of the day, the GPS mobility logs are available from such a user whose driving pattern is reckless, i.e., exhibiting much higher speed compared to the average driving speed; observed for that road. In such a case, calculating speed from multiple such GPS logs and estimating the overall traffic behavior, might not reflect the exact aggregate behavior (detail in Section 4) of the vehicles passing through the road segment.

It may derive a wrong speed profile, e.g., a moderate traffic state might be inferred as free flow. The reverse may also be true where a user is over-cautious while driving, thereby, generating speed samples of lower range than observed speed range as an aggregate behavior. Dealing with such anomalies is challenging, especially in areas where most user density is sparse. Hence, one should not rely on GPS mobility logs or accelerometer to estimate speed; rather, might

[1] https://www.webpages.uidaho.edu/niatt_labmanual/chapters/trafficflowtheory/theoryandconcepts/TrafficFlowParameters.htm

[2] https://www.google.co.in/maps/dir///@22.5764753,88.4306861,15z?hl=en

[3] https://www.waze.com/

[4] https://wego.here.com/?x=ep&map=20,77,10,normal

[5] https://www.bing.com/maps

[6] https://www.compare.com/auto-insurance/guides/top-3-best-traffic-apps

[7] https://support.google.com/maps/forum/AAAAQuUrST81I07Ih50xdk/?hl=en&gpf=%23!topic%2Fmaps%2F1I07Ih50xdk

[8] https://waze.uservoice.com/forums/59223- waze-suggestion box/suggestions/ 3273686-color- coding-for-traffic-reportsindication

[9] https://nptel.ac.in/courses/105101087/downloads/Lec-31.pdf

[10] https://support.google.com/maps/forum/AAAAQuUrST81I07Ih50xdk/?hl=en&gpf=%23!topic%2Fmaps%2F1I07Ih50xdk

investigate additional features like road surface, traffic condition, and map-related information, which has a huge influence or impact on speed profile for a road segment. Therefore, proper choice of those features, sampling them using smartphone sensors, encoding the raw samples, and finally developing a model to map the feature set into speed profile classes, is non-trivial. Hence, the overall challenge is to derive a speed profile from ambient road features without directly using speed signatures to eliminate anomalies due to specific driving maneuvers, especially when the sample set is sparse.

Existing Works: Prior research has been carried out on intelligent road traffic and transport management systems [3, 23] road traffic congestion measurement [12], travel time estimation [20], Modelling urban mobility pattern, Characterization of city traffic and road, etc. Researchers have worked on off-line or online GPS trace of vehicles, online social media data [22], sensor data from accelerometer and microphone of high-end mobile phone or image or video captured by roadside cameras [13]. GPS trace is analyzed to develop transport and traffic monitoring systems, traffic congestion, or any special event is extracted from online social media. Sensor data are applied to detect bad road surface conditions and traffic congestion.

Some researchers have worked on analyzing the impact of external factors (e.g., weather, event) on city traffic flow [6]. These existing works do not mention the quantitative measures to assess road behaviors like the roughness of road surface, narrow roads, crowdedness during traffic analysis. These subjective factors are still not considered in existing traffic estimation systems. To the best of our knowledge, a study of speed estimation using ambient subjective features and inferring its context has not yet been done.

Contribution: *Firstly*, through an extensive study, we identified features that significantly influence the aggregate speed profile of a road segment and also can be easily derived. We have used road surface quality, the density of vehicles, type of neighborhood, and count of road intersections for a particular road segment and developed a machine learning model. To eliminate the noise due to driving samples, for training purposes, we collected samples in a controlled environment, using a group of 12 volunteers who travelled 1184km city roads at Durgapur, a suburban city in India, on 2-wheelers for five months. These volunteers recorded count of potholes and bumps (as measure of roughness of road surface), honking levels (a measure of vehicle density) WiFi hotspot density (denoting type of neighborhood market place/highway) by using our customized smartphone app, *RoadSpeedSense* for all road segments of uniform length of 200m along a particular route. The above features, coupled with the count of road intersections extracted from Google map for a road segment and time of the day (i.e., morning/noon/evening/night) is used to train the learning model. For each sample of this feature set is associated with an estimated speed range derived from GPS mobility logs of volunteers along a road segment. We have carried out the experiments on different classification algorithms like K-nearest Neighbour(KNN), Decision Tree(DT), Random Forest(RF), and Gradient Boost(GB) with the necessary tuning of parameters which are discussed in Section 4. The experimental results show that GB outperforms other classification algorithms in estimating the speed class of road segments among all classifier algorithms with the highest *F1-Score* of 0.8345. *Secondly*, to understand the context, we used the interpretable LIME model, which is explained in Section 5, to analyze the learning model behavior for a particular feature. *Thirdly*, as a possible application, we illustrated how a fair driver rating system could be derived from our results, as discussed in Section 5. The results obtained from the proposed novel framework provides a proof of concept that speed profiles may be successfully derived from ambient road features even when sample space is sparse. In this paper, we have presented a detail literature study in Section 2, and a comprehensive end-to-end system is explained in Section 3 whereas in Section 4, a detailed analysis of correlation between speed of vehicles and selected road features, the extraction of road features from raw sensor data as well as the machine learning algorithms are well discussed. In Section 5, a thorough analysis of experimental results are presented, Section 6 describes an application of *RoadSpeedSense*, and finally, Section 7 concludes the paper.

[11] https://www.ncbi.nlm.nih.gov/pmc/articles/

[12] https://developers.google.com/maps/documentation/distance-matrix/start

[13] https://stackoverflow.com/questions/43277342/can-we-get-speed-limit-information-from-waze-apis

[14] http://opentraffic.io/#partes

[15] Space-mean speed is the harmonic mean of the spot speed over a specified road segment of unit length. Space mean speed is always lower than the time mean speed as it weights slower vehicles more heavily as they occupy the road stretch for longer duration of time

2. Related Works

This section provides an overview of existing literature that relates to intelligent road traffic monitoring and management system. Our broad focus is to review the parameters used to estimate the mean speed of vehicles along a road and the techniques to measure those parameters. The review will *Firstly,* discuss methods for estimation of parameters like traffic volume, flow, density using sensor infrastructures. *Secondly,* we highlight the works on real-time traffic/congestion monitoring and route planning using GPS data. *Thirdly,* we discuss techniques on detecting traffic states, traffic speed, congestion using several other features other than GPS data like sound, pollution emission, weather data, GSM signal, physical characteristics of the road, social media, etc. Finally, we also mention the systems that use smartphone sensors to detect road surface conditions, which is an important feature that affects vehicle speed.

Study of traffic behaviour using only GPS logs: Several systems use GPS logs of probe vehicles for real-time traffic monitoring and understanding travel behavior. Biagioni et al. [3] extracts routes, stops, and schedule from GPS traces of cars, i.e., cabs, bus. Thiagarajan et al. [20] focus on real-time route planning, hotspot detection, and travel time estimation from GPS traces collected by the war-driving approach. Xin et al. [23] aim at forecasting collector road speeds by utilizing car GPS signals. Mandal et al. characterize the stoppage pattern of public buses from historical GPS logs [9] [10]. Nguyen et al. [12] propose a traffic congestion monitoring system using historical and real-time vehicular data. By modelling traffic flow from historical logs, they have predicted and verified traffic congestion. Works described above only utilizes GPS traces for extraction of a different kind of traffic information. It may be noted that sudden changes in vehicle speed are not always well captured well by GPS, as reported by Zong [25].

Study of traffic behaviour using features in addition to GPS logs: In addition to GPS logs, researchers have worked on multi-modal non-GPS features like video/image, the sound of honks, GSM radio signal, weather data and road network data as well as social media feeds for subjective analysis of traffic behavior. Hoang [6] et al. utilizes GSM radio signal, road network data, and weather data in addition to GPS logs to capture seasonal changes and instantaneous changes. Park et al. [13] have used the GPS data in combination with camera images to propose a traffic risk detection model that automatically detects dangerous situations by monitoring driving behavior. Vij et al. [21] have used microphone data of smartphones in an effective way of identifying different traffic states. Authors show that, by using audio analysis, detection of a particular traffic state becomes faster and easier. Sen et al. [17] estimate the speed of a vehicle from vehicular honks. Moreover, in [16] the author presents an acoustic sensing based technique for real-time congestion monitoring on chaotic roads. Zheng et al. [18] estimate the gas consumption and pollution levels emitted by vehicles traveling on ways to determine the travel speed of road segments. Additionally, using road features, POI, and the global position of the road, they aim to infer the traffic volume from travel speed. In another approach, Wang et al. [22] combine social media data with road features to identify traffic congestion on city roads. They estimate citywide traffic congestion from social media by mining the spatial and temporal correlations of the road segments facing congestion from historical data. The authors extracted the physical features of roads from road network data. They also measured the impact of a social event on nearby road segments. Hence, the above-studied features can be used along with GPS to infer traffic behavior to a certain extent. However, the surface condition of a road segment also plays a vital role in affecting vehicle speed.

Study of road surface condition using GPS smartphone sensors: Even though many of the road features like the number of intersections, road type (highway/arterial), etc. may be extracted from map data mining, gauging the surface condition of roads is nontrivial. Various road sensing techniques have been deployed using data from different smartphone sensors such as accelerometers, microphones belonging to users traveling in vehicles. This sensing technique is applied for the detection of road surface conditions or traffic congestion. From a detail literature study, we have observed that most of the researchers either have combined online data with sensor data or historical logs of traffic data with smartphone-like GPS, accelerometer data, microphone data, and mobile phone signals, etc. Real-time GPS data provides proper traffic information, but modelling the traffic characteristics needs a detail behavior analysis of other sensor data and especially those for road characterization. Eriksson et al. [5] developed a Pothole Patrol system to detect road anomalies by using accelerometers and GPS sensors installed in taxis. Mohan et.al.[11] presents a system to monitor road conditions by detecting potholes, bumps, honking from an accelerometer, microphone, GSM radio signal, and GPS data. Vittorio et al.[1] applied the anomaly detection method on mobile devices based on vertical acceleration and GPS signals. Perttunen et al. [14] proposed a road anomaly detection method using GPS and acceleration signals. Wolverine [2] uses a smartphone sensor to detect road conditions and bumps. The authors have proposed a reorientation using a magnetometer beside accelerometer and GPS sensors.

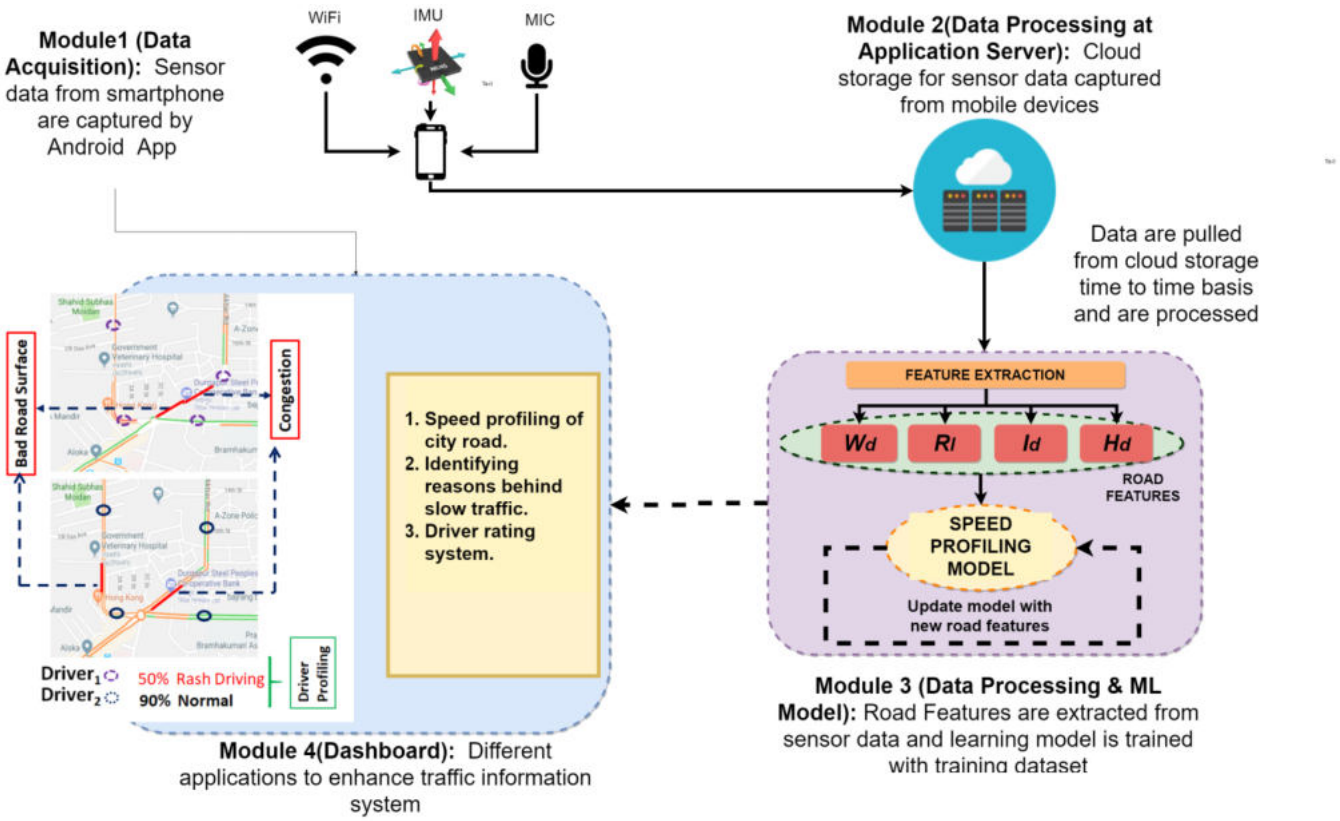


Figure 1. System architecture for Speed Profiling, For a road segment (Seg_i), road features are: (a)Honk duration(H_d), (b)Road Surface Index (R_I), (c)Intersection Density (I_d) & WiFi Density (W_d)

Nuno Silva et al. [19] detects road condition from GPS and accelerometer. Haofu Han et al. [24] estimates speed from an accelerometer, which senses natural driving conditions in urban environments, including making turns, stopping, and passing through uneven road surfaces. They contribute to eliminating the errors in speed estimation caused by accelerations in real-time. A. Chowdhury [4] investigates the noise performance of accelerometers, available in smartphones and finally apply the analysis for estimating the speed of moving vehicle because sudden changes in vehicle speed are not always captured well by GPS. X. Zong [25].

From a detailed study on existing literature, we may observe that people have used offline and online speed traces of vehicles for a traffic monitoring system. Temporal and spatial speed traces along city roads are analyzed to characterize the traffic. But, the speed of vehicles is random as drivers follow a random driving style along a city road, especially at the bad road surface and populated road segment. These aspects place a significant question mark on the applicability of speed traces for chaotic road conditions. Applicability of online social media data is also not suitable where the density of the crowd is sparse, and network connectivity is intermittent. By application of sensor data (i.e., accelerometer and microphone) apart from GPS, people have focused on identifying particular events like bad road surface or traffic congestion, etc. However, to generate a good speed profile for a road segment, it is essential to sense the road characteristics of a particular road segment because the speed of the vehicle gets very much affected by different road characteristics. In summary, the existing literature fails to consider the Spatio-temporal uncertainties of personalized driving behavior. The challenge lies in (a) detecting the actual speed profile from Spatio-temporal anomalies in speed traces of vehicles, and (b) Finding the dependency of vehicles on different road features. (c) Generating context-aware temporal speed profile for city road segments. Such a context-aware speed profile will also trigger in generating the driver rating system, which will help the transport authority to evaluate the driver profile properly. It is concluded that several road factors affect the speed of traffic.Therefore, these important factors should be considered together while ranking the traffic state as *veryfast, fast, slow,* etc. for different road segments. Hence, each road segment should be ranked according to these features, and based on this ranking, the speed of traffic should be classified into different categories. Given a Road Segment Seg_i of the City C_i with features **(a)** Honk Duration **(**H_d**)**, **(b)** Road Surface Index **(**R_I**)**, **(c)** InterSection Density **(**I_d **)**, and **(d)** WiFi Density (W_d), Is it possible to classify the (Seg_i) as *slow, normal, fast* and *veryfast* ?

Table 1. Information of 3 routes from Durgapur, India used in our case study

Route	No.of Trails	Highway road	Market road	Normal city road	Total Distance
Route 1 (8.7 km)	60	2.2 km	2.8 km	3.7 km	522 km
Route 2 (6.2 km)	60	0 km	4.5 km	1.7 km	372 km
Route 3 (8.3 km)	35	7.2 km	0 km	1.1 km	290 km

3. RoadSpeedSense:System Overview

The overall system architecture along with different modules are shown in Figure 1. It consists of different modules for data acquisition and storage, processing of data in web-server, training speed estimation model using the processed data, generation of context-aware speed profile of city roads followed by development of smart-phone or desktop based application to reveal traffic behaviour based on road characteristics of city roads on digital map. Following is a brief description of each module.

Module 1: Sensor Data Acquisition – To capture the inherent features of city roads which affects the speed of traffic; we have collected smart-phone sensor data for total 1184km of city roads for three different routes at city of Durgapur by a group of volunteers on 2 wheeler vehicles. The detail specifications of routes are given in Table 1. We have developed a simple smart-phone based android application to collect these sensor data from smart-phone; carried by volunteers for almost five months of duration as shown in Figure 2. The application records the WiFi, accelerometer, microphone and GPS sensors data of android smart- phone. A volunteer while starting his/her trip, starts the application for logging the sensors data. The application records, i) GPS data with time. ii) SSID and signal strength of WiFi. iii) Accelerometer data. iv) Microphone sensor data.

Module 2: Data Processing at Application Server– Collected sensor data are stored in main application server. Collected data are pulled from cloud storage time to time and are processed by separating data for different modes of transport (i.e. Bus, MotorCycle, Car) while collecting. Data are also geo-tagged by corresponding GPS trails based on date and time of collection. Newly uploaded data are identified and are verified whether it is from a new route or an existing one. Identification of road segments are also performed in this module.

Module 3: Data Processing and Machine Learning Model– Different road features like road surface condition, road network, available WiFi signal and road network are extracted from collected sensor data for each road segments. The features of one road segment are tagged with particular speed class and time as observed by volunteers. Thus, a robust training dataset is prepared to characterize the city road features with its speed at a particular time. These training dataset is used to train the learning model.

Module 4: Application Development- Different applications are developed from speed estimation of learning model like, speed profiling of city roads, finding reasons behind *slow* traffic and driver rating system.

Figure 2. Android Application for Sensor Data Acquisition.

4. Choice of Features Affecting Speed

To understand the characteristics of a particular road segment, we have selected some features which may have strong correlation with the speed of vehicles on the road. Sound of honks from vehicles well describes a busy road; caused by overloaded traffic, available WiFi signal signifies a high activity area like a market place, stores, commercial complex, tourist spot, stadium, auditorium, or any popular point of interest; where people gather for different purposes. At such high activity areas, during busy hour, traffic movement gets disrupted. Therefore, number of available WiFi signals is the right choice of feature to detect the high presence of crowd on a road segment. The presence of bumps, potholes on the road, or broken road seriously affects the speed of vehicles. Therefore, Road surface condition is an essential feature for traffic speed estimation. Another exciting feature is the structure of the road network, which plays a useful role in creating traffic congestion because traffic gets interrupted when vehicles from intersecting roads enter the main road. Hence, if the number of intersecting roads is high, then there is more chance of *slow* traffic. Therefore, number of intersections is another useful feature to correlate with the speed of the vehicle. We have classified features mentioned above into four categories like following,

– Heavy load of vehicles (i.e., traffic feature).

– High human activity (presence of pedestrians on the road).
– Road Surface Condition (i.e., Presence of potholes, bumps on the road or broken road).
– Road network (Number of crossroads/intersections, junction points along a road segment).

4.1. Honk Duration (H_d)

The amount of honking is an excellent feature to capture the load of vehicles on a road segment [17], [21], [16]. On city-roads, as well as on highways, honks of vehicles are used to alert the surrounding vehicles for safe driving. Honk increases noise level when vehicles get stuck in traffic congestion and try to communicate with other vehicles to clear the road blockage. Therefore, during congestion, the duration of honk sound is much higher than a standard traffic condition. Therefore, honk duration is one of the most available features to detect traffic congestion when speed of traffic is obviously slow. Hence, H_d has a strong correlation with the mean speed of vehicles along a road segment.

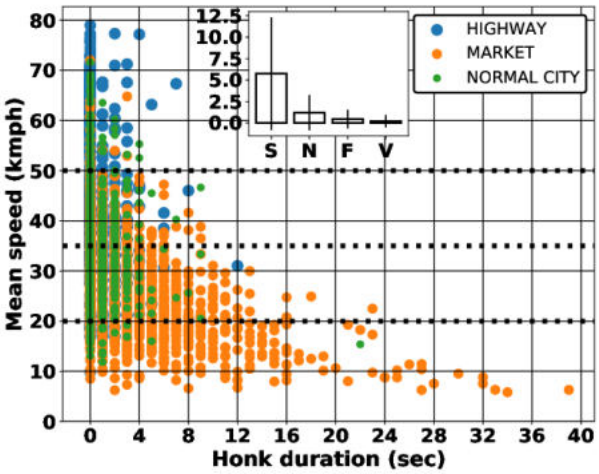


Figure 3. Correlation of H_d and mean speed of vehicles.

Figure 3 shows the correlation H_d and mean speed for different road segments from the market, highway, and normal city zones. It clearly shows that, if H_d increases, then mean speed along a particular road segment decreases. The most exciting observation made from Figure 3 is that almost all the segments, which have a honk duration greater than 4 seconds, are the segments from the market zone. The mean speed values for these segments also range below 20km/hr, which denotes a *slow* speed class as per our observation during data collection. Therefore, H_d can clearly distinguish the market segments and their *slow* speed class. But the values of H_d for highway and normal city road segments are almost the same. This feature of H_d is prominent for traffic congestion, which is an obvious case of the market zone. But, in normal city roads or highways specifically for our city of case study, such congestion is very rare because of more extensive road space and lesser gathering of people than market area. Therefore, this feature is very relevant to a *slow* traffic state. The inset of Figure 3 shows the honk duration in seconds for each speed class. *Slow* speed class shows the highest honk duration of 5.17 sec on average, whereas, *veryfast* speed class has an average H_d of 0.2sec. A detail observation of Figure 3 shows that there are some market segments where mean speed of vehicles has exceeded 70km/hour speed, and honk duration is also either 0 or lesser than 4sec. These may happen for rash driving by any particular driver during the non-busy hour (i.e., early morning or late night)

when market is closed. Even for some segments of normal city roads also, the mean speed is very high i.e., above 70km/hour.

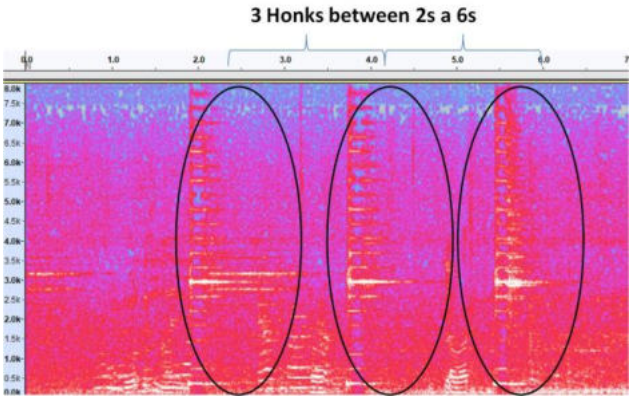


Figure 4. Honks Spectrogram.

Figure 4 depicts the spectrogram of a horn sound i.e., a time versus frequency plot of sound trace for a road segment, with higher sound power depicted by bright colored yellow shades. The frequency harmonics are visible (with a fundamental frequency under 500 Hz), and there is a considerable amount of energy around the 3 kHz band, which is human ear sensitive. Since most of the honk frequency is within the range of 2-4 kHz; therefore, we apply the honk detection method on the sound data for each trail with frequencies within this range. We apply a bandpass filtering method to filter out the noise outside this range. After this step, we sample the filtered sound trace into 1-second windows and calculate the maximum amplitude for each window. It was observed that most honks exceed 80 dB or more than 80dB. Therefore, 80 dB is chosen as threshold value to detect honks. So any 1-second window having amplitude more than this threshold is tagged as a honk. Finally, we have labeled each 1-second window of sound trace, recorded for each trail. In the next step, these sound data is localized with GPS location from the GPS trail. H_d of a particular segment is the sum of those particular 1-second windows where the sound has exceeded the threshold value 80dB. The calculation of H_d is shown in Equation 1.

$$H_d(Seg_i) = \sum_{i=1}^{TravelTime(Sec)} Second_i \quad \text{where } Sound_i \geq 80\text{dB} \quad \ldots(1)$$

4.2. WiFi Density (W_d)

Available WiFi signal is an excellent feature to capture human activity, which varies with location and time. In this smart society, WiFi access points are mounted for people to provide internet facility. Higher WiFi density indeed signifies the high possibility of human presence in an area because WiFi access points are mounted in such a planned place where a large gathering of people is found. WiFi signals are available at market places, stores, commercial complex, tourist spots, stadium, auditorium, or at any popular point of interest. Therefore, higher WiFi signals indicate higher human activity which causes decrease in speed of vehicles along a road segment. W_d is a handy feature to identify a busy area like market area in our case study and speed of vehicle is obviously slow in this area. Therefore, W_d shows a strong correlation with *slow* speed class as shown in Figure 5. But, for highway or normal city area, W_d is not so prominent factor. Inset of Figure 5 shows different speed classes and their average W_d. Our customized android application for data collection continuously captures the WiFi data. The application scans for available WiFi access points every second and records the MAC address, SSID, and signal strength, received by each access point. First, the timestamp is compared in both GPS and WiFi log files, and a new file is generated that combines GPS positions and corresponding WiFi signal with time. For a specific segment, the MAC ID of the distinct WiFi access point is counted to calculate the total number of distinct WiFi access points along a segment for a single trail. W_d is calculated by dividing the number of distinct access points by the length of the segment, as shown in Equation 2 and 3. W_d is the number of distinct access points per unit distance.

$$W = \text{count (unique } MAC_i \text{)} \quad \ldots(2)$$

$$W_d\ (Seg_i) = W / \text{len}(Seg_i) \quad \ldots(3)$$

In Figure 5, W_d and corresponding mean speed of vehicles are displayed for different road segments from market highway and normal city zones.

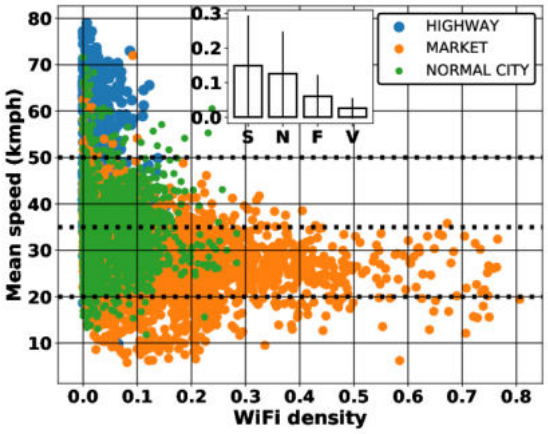


Figure 5. Correlation between W_d and Mean Speed of Vehicles.

Table 2. Four levels of road surface condition based on IRI.

Road Surface Class	IRI	RSI class	Length of the road network (km)	%
Excellent/Good	$0 \leq IRI \leq 2$	RSI1	166.63	47.08
Fair	$2 \leq IRI \leq 4$	RSI2	107.74	30.43
Poor	$4 \leq IRI \leq 9$	RSI3	63.53	17.94
Very poor	$IRI \geq 9$	RSI4	16.12	4.55

It shows that road segments from the market zone are distinguishable by W_d with a value greater than 0.2 to 0.3. But, for both highway and normal city road segments, the W_d values are almost the same (i.e., <0.2). From inset of Figure 5, we may infer that all segments of *slow* speed class have mean W_d value of 0.15, all segments of *normal* speed class have mean W_d of 0.12, all segments of *fast* speed class have mean W_d of 0.06 and finally, all segments of *veryfast* speed class have mean W_d of 0.02. In our case study, we have considered a suburban city of India where we do not observe the sufficient number of WiFi signals in normal city roads and highways because the city is not so highly populated in those areas. Sometimes the normal city roads behave like highways, especially at non-busy hours. Then we get the almost same value of W_d for *fast* and *veryfast* speed class.
Another problem with this feature is the temporal variations of feature values. W_d indicates the active WiFi access points around, which varies with time. It decreases when human activity around a segment drops, and access points are switched off or moves away. For most of the market segments, W_d is observed to be lower during morning hours; than the rest of the day hours because human activity remains high from 9 am to 1 pm and from 4 pm to 8 pm. Due to these temporal variations, at a non-busy hour, W_d becomes a weak feature for speed prediction.

4.3. Road Surface Index (R_I)

The speed of vehicle significantly depends on road surface condition. Although road is free from load of vehicles yet speed will must be slow for the bad road surface. The road roughness can be estimated from the accelerometer sensor data of the smart-phone; carried by travelers. The sampling frequency of the sensors was 50Hz. The vertical acceleration (i.e., along the Z-axis) is required to calculate the IRI-Proxy [8] index. This index is the metric to measure the road surface quality, which has a significant impact on the speed of the vehicle and thus on traffic state.
The International Roughness Index (IRI-proxy) [8] is the roughness index commonly used to examine the road surface condition. We have calculated IRI-proxy for each road segment by multiplying the speed-normalized RMS by 100, as shown in Equation 4. For each road segment, IRI-proxy can be calculated as follows,

$$\text{IRI-proxy} = (n \bullet \frac{R}{\sum_{i=1}^{n} vi}) \bullet 100 \quad \ldots(4)$$

where n is the number of measurements gathered from a road segment, R is the RMS of this road segment, V_i is the real-time speed of the vehicle at the location of the ith acceleration measurement.

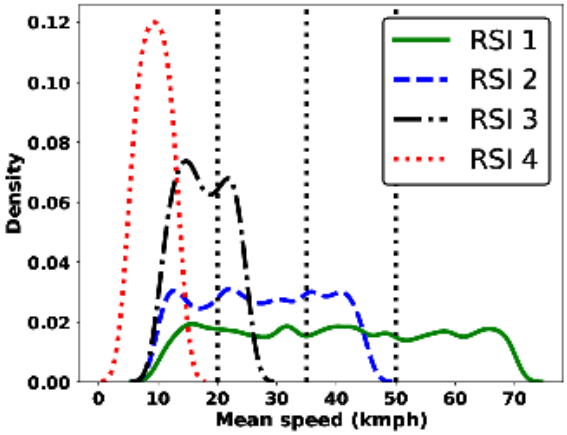


Figure 6. Probability Density Function of Mean Speed for Different Road Surface Index(R_I).

From volunteer's observation, the road roughness condition can be classified into 4 levels (i.e. Excellent, Fair, Poor and Very Poor). These four road condition indices are summarized in Table 2.
Figure 6 shows the probability of mean speed; falling within a particular range of values instead of a single value. We can observe an interesting fact from Figure 6 that, for RSI 4 (i.e. very poor) category of road segments, the mean speed is 10km/hour and the variation of minimum and maximum speed range is very less i.e. -10km/hour to +10km/hour from mean speed. But, for road surface category RSI 1 (i.e., excellent road surface), the mean speed is 40km/hour ,which denotes *veryfast* speed class as per our observation. The variation of minimum and maximum speed range for RSI 1 is -30km/hr to +30km/hour from mean speed. It means that mean speed is higher for excellent road surface than the mean speed of vehicle along inferior road surface but, there is another interesting fact that the standard deviation of the mean speed of vehicles is also higher for excellent road surface than the standard deviation of speed on the poor road surface. It is because, on a good road surface, the normal behavior of speed is *veryfast*, but it may differ for other road factors, which results in high standard deviation. But, for the poor

category of the road surface, the mean speed and standard deviation both are very low. Therefore it is explainable that, for the poor road surface, the speed of the vehicle must be slow, and no other road factors can change the speed of vehicles. Hence, R_I is another useful feature that strongly correlates with *slow* speed class. For the other two categories of RSI, we can conclude the same analysis from Figure 6.

4.4. Intersection Density (I_d)

Road Intersection is a crossover of two or more than two road segments. These are the locations where the normal flow of traffic gets stuck due to the crossing maneuvers of vehicles moving in different directions. Vehicles get slow down at these locations that cause traffic delays. Overall, traffic flow depends on this number of intersections. In our experiment, for all road segments, we have counted all three and four-head intersections for each road segment, and intersection density is calculated by dividing the total number of intersections for a particular segment by the length of that segment as shown in Equation 5 and 6. We have observed that the maximum number of intersection was found on road segments passing through the market area and the normal city area.

$$Intersections\ (Seg_i) = count\ (T_{junction}\ ,\ Y_{junction}\ , Crossroads) \quad \ldots(5)$$

$$I_d\ (seg_i\ ) = Intersections(seg_i\ )/(length(seg_i\ )) \quad \ldots(6)$$

Figure 7 shows boxplot, which reveals interesting facts about the correlation between speed and ID. Figure 7 shows that, for ID value 0, the mean speed varies within

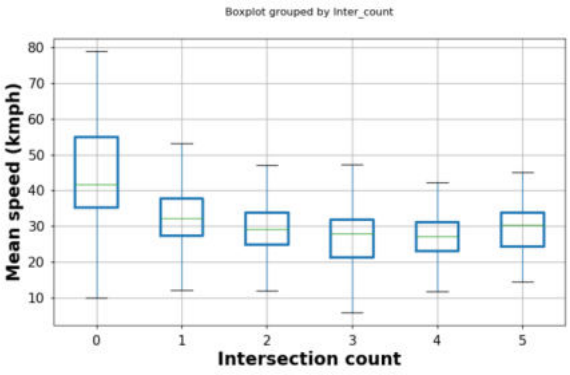


Figure 7. Speed behavior based on Intersection Density (I_d)

the range from 10km/hour to 80km/hour and the 1st quartile is at mean speed value of 35km/hour, the median value is 42km/hour, and the 3rd quartile is at 55km/hour. For ID values 1, 2, 3, and 4, we can observe no significant variation either in range of minimum and maximum value of mean speed or in 1st quartile or 3rd quartile values of mean speed. This analysis represents the fact, that if there are no intersections (i.e., traffic signal) in a road segment, then the mean speed may increase up to 80km/hour, but for a single intersection in between a road segment, maximum mean speed drops within 40km/hour to 50 km/hour. More interestingly, if there is more than one intersection, then also the distributions of mean speed are the same with a single intersection. It means that the impact of single and multiple intersections is almost the same. Therefore, we can notice a clear separation between two scenarios; one is a road segment with intersection and road segment without intersection. The mean speed of vehicles is correlated with these two scenarios.

Significant Observation: From this detail investigation of features, we can conclude that most of the features are most prominent for *slow* speed class among all other speed classes. In a busy hour, all road segments from the market zone depict congested traffic state. Therefore, honk duration is a nice feature for capturing busy market road segments. Similarly, WiFi density is another essential feature to capture the human presence, but it is time-variant. It is also observed that, for the poor road surface, the speed of the vehicle must be slow. Road Surface index is a powerful feature that strongly correlates with *slow* speed class. In the case of intersection density, we notice a clear separation between two scenarios; one is a road segment with intersections and road segment without intersections. The mean speed of vehicles is correlated with these two scenarios. A single feature can clearly distinguish the *slow* speed class from other class ,but for further classification of other speed classes (i.e., *normal*, *fast*, or *veryfast*), we need a combination of all features.

4.5. Machine Learning Models

Predictive modelling is the concept of building a model that is capable of making predictions. Typically, such a model includes a machine learning algorithm that learns specific properties from a training dataset to make those predictions. Here, we have applied the pattern classification to assign discrete speed class labels to road segments based on particular observations as outcomes of a prediction. In this paper, we are trying to classify the speed profile of a road segment, not only based on its speed but also road surface condition, road network structure, demographic profile and load of traffic, etc. We have labeled dataset, collected by volunteers, to train our learning model. Our problem is a classification problem. We have carried out our experiment using four classification algorithms over the dataset. The algorithms are K Nearest Neighbour(KNN),

Decision Tree(DT), Random Forest(RF), and Gradient Boost(GB). Different machine learning models are parameterized so that their behavior can be tuned, specific to our given problem. These models have several parameters, and finding the best combination of parameters is non-trivial. Tuning is the process of maximizing a model's performance with neither over-fitting nor creating a high variance. The traffic follows a regular pattern for the whole city and a specific time in regular life. The variation in traffic characteristics depends on demographic characteristics, loads of traffic, which vary with time, road surface condition, road network, human mobility, weather condition, etc. So we can assume that there are particular traffic patterns like weekdays patterns, holiday patterns, time level patterns, seasonal patterns, etc. Machine learning algorithms are composed of finding similar past patterns from the data set and, with the help of these patterns, classify the new observation. In supervised machine learning, each data input object is preassigned a class label. The main task of supervised algorithms is to learn a model that ideally produces the same labeling for the provided data and generalizes well on unseen data (i.e., prediction). It is the main objective of classification algorithms.

4.5.1 Dataset Preparation

Data preparation is an essential phase before applying any machine learning algorithms. If there is much irrelevant and redundant information present or noisy and unreliable data, then it causes the algorithm to produce less accuracy. Data pre-processing includes cleaning, instance selection, normalization, transformation, feature extraction, and selection, etc. Generally, we split the dataset with a ratio of 70% for the training data and 30% as test data for cross validation. For our dataset, we split into 80% for training data and 20% for the test data. A sample of our dataset is shown in Table 3.

Table 3. Sample of the dataset (Segment ID (Seg_i), WiFi density (W_d) Road Surface Index(R_I), Honk duration (H_d), Time(T), Intersection Density (I_d)

Seg_i	W_d	R_I	H_d	T	I_d	Class
27	0.012966	1	0.0	3	0.000480	Fast
49	0.008956	1	0.0	2	0.0	VeryFast
25	0.009631	2	0.093750	1	0.001445	Normal
24	0.333728	4	0.176471	4	0.001927	Slow

We have presented a detail description of our training dataset as well as the test data set, which has been used for our experiment following in a tabular form.

Table 4. Description of Training Data

	Number of Segments of Different Class				
Route#	*Fast*	*Normal*	Slow	*Very fast*	Total
Route 1	355	1425	81	5	1866
Route 2	62	52	23	255	392
Route 3	612	853	200	31	1696
New Route	274	459	117	111	961
Total	1303	2789	421	402	4915

In Table 4, we have shown a total number of road segments that have been used to train the learning model for three routes individually. We have presented the number of road segments for each class for each of the three routes. We have enhanced our training dataset by introducing a new route. At Route 1, we have 355 road segments from *fast* class, 1425 number of segments of *normal* class, 81 number of segments from *slow* class, and 5 segments from *veryfast* class. Similarly, for Route 2, we have 62 *fast* class, 52 *normal* class, 23 *slow* class, and 255 *veryfast* class of road segments. From Route 3, we have 612 number of *fast* class, 853 number of *normal* class, 200 number of *slow* class, and 31 *veryfast* class of road segments. From the new Route also, we have collected data for 274 *fast* class, 459 *normal* class, 117 *slow* class, and 111 *veryfast* class of road segments. In total, we have 1303 number of *fast* class road segments, 2789 number of road segments with *normal* class labeling, 421 number of *slow* class segments ,and 402 number of *veryfast* road segments. A total 4915 number of road segments have been used to build a robust training dataset to train our learning model.

Table 5. Description of Test Data

	Number of segments of different class				
Route#	*Fast*	*Normal*	*Slow*	*Very fast*	Total
Route 1	112	429	28	2	571
Route 2	19	18	11	81	129
Route 3	183	300	73	9	565
New Route	78	138	32	40	288
Total	392	885	144	132	1553

In Table 5, we have shown the total number of road segments that have been used to test the learning model for three routes individually. We have also enriched the test dataset by introducing data from a new route. In total, we have 392 number of *fast* class road segments, 885 number of road segments with *normal* class labeling, 144 number of *slow* class segments, and 132 number of *veryfast* road segments. A total of 1553 number of road segments have been used to build a robust test dataset to test our learning model. In Table 1, in addition to the route description, we have mentioned the characteristics features of each route in terms of the type of locality/zone along the route i.e., the spatial variations of the route like highway, market or normal city road. We have also collected trails from each route at different hours of the day. While collecting the data, we have tried to select those routes which have a good combination of different types zones and we have also tried to collect the data for all time zones of the day for each route to develop a well-balanced robust training and test dataset.

4.5.2 Experiment with Classification Algorithms

Table 6. ML configurations for K Nearest Neighbor (KNN), DecisionTree(DT), Random Forest(RF), GradientBoost (GB)

Algorithm	Parameter	Values	Best Fit
KNN	n neighbors	3,5,7,9,11	7
	Weights	uniform, distance	uniform
	distance metric	euclidean, manhattan	manhattan
DT	max depth	2,3,4,5,6,7	5
	min samples leaf	(0.1,0.5]	0.1
RF	max depth	2,3,4,5,6,7	7
	n estimators	10,15,20,25,30	20
	min samples leaf	(0.1,0.5]	0.1
GB	max depth	2,3,4,5,6,7	3
	n estimators	100,120,150,200,250	120
	learning rate	(0,1.0]	0.125
	min samples leaf	(0.1,0.5]	0.1

Parameter Tuning of KNN Selection of k is non-trivial for KNN because the small value of K will have a stronger influence of noise on the outcome and a large value of K makes the algorithm computationally expensive. So, we performed 10 fold cross-validation to find out the best K value. Our whole dataset is divided into training set and validation set. We take one single instance from the training set and use it to train the model. Then we measure the error on the validation set and on that single training instance. The error on the training instance is 0. However, the error on the validation set is huge because the model is trained with a single training instance. As the size of the training set increases, the accuracy of classification for any training instance i.e., training score decreases and accuracy of the validation set i.e., testing score increases. Initially, when the value of K = 1, each sample is using itself as a reference, which is a case of over-fitting. When K increases from 1 to 7, the test score improves subsequently. For the K value of more than 7, the score becomes consistent. Therefore, 7 is the right choice of k for this particular dataset. The co-relation or similarity of features is defined by distance metric (Manhattan or Euclidean) between two data points. For our dataset, we have chosen Manhattan distance because Manhattan distance may be preferable to Euclidean distance for the case of high dimensional data. For our case, we have tested for both distance metric, but Manhattan distance gives a better result.

Parameter Tuning of Decision Tree The more rooted the tree, the more splits it has, and it captures more information about the dataset. We fit a decision tree with depths ranging from 1 to 32. We see an increase in the test score as depth increases till depth = 5 but gradually decreases after 5 as the decision tree model suffers from overfitting. So, depth = 5 was chosen. Additionally, the Min-sample-per-leaf node was set to 1 by default, which would naturally make the tree overfit and learn from all the data points, including outliers. We increase it to 1% of the data points to stop the tree from prematurely classifying these outliers.

Parameter Tuning of Random Forest, Usually the more is the number of decision trees (n estimators), the more accurate classification is possible. But, more number of decision trees can slow down the training process. We have observed that the test score remains consistent as we keep on adding more n estimators. For this model, we have chosen n estimators = 20.

Parameter Tuning of Gradient Boost, Increasing the number of boosting iterations (n estimators) to perform in the Gradient Boosting algorithm, reduces the error on the training set, but setting it too high may lead to over-fitting. So, in our case, we have chosen n estimators = 120. A technique to slow down the learning in the gradient boosting model is to apply a weight factor (learning rate) for the corrections by new trees when added to the model; lower learning rate requires more iterations. Other parameters to tune depth of the tree, is min number of data points allowed in a leaf node.
Table 6 shows the parameters used in the classification algorithms and their optimal parameter values for the best performance.

5. Result Analysis

We have carried out our experiments with four learning algorithms i.e. K Nearest Neighbour

Table 7. Significance of True Positive, False Positive, False Negative and *F1-Score*.

Notations	Definition/Interpretation
True Positive (TP)	All actual instances of speed category $Speed_x$ from volunters GPS trail, that are classified as $Speed_x$.
False Positive (FP)	all non-$Speed_x$ instances that are classified as $Speed_x$.
False Negative (FN)	all $Speed_x$ instances that are not classified as $Speed_x$
Precision (P)	Ratio of True Positive Detections and Total Detections. [TP/(TP+FP)]
Recall (R)	Ratio of True Positive Detections and Total Instances. [TP/(TP+FN)]
F1-Score	Harmonic Mean of Precision and Recall.[2*P*R/(P+R)]

(KNN), Decision Tree(DT), Random Forest(RF), and Gradient Boost(GB) on overall data set; collected from a suburban city of Durgapur.

In the following sections, we have evaluated the performance of different learning models. We have analyzed our experimental results further in detail for each speed category individually as well as for each road type separately. The model is also evaluated for different hours of the day, which also infers some exciting facts about city traffic behavior. By experimental results, we have shown that our learning model not only efficiently predicts the practical speed limit from road features, but it also provides additional information about probable reasons of slow speed nature considering the features as mentioned earlier of road segments. Finally, after characterizing all city road features, we can have a clear view of the effective speed behavior of vehicles across the whole city. Based on such a speed profile of a road, we can better assess the individual driving performance of a driver because driving style is very much subjective to road characteristics and such assessment can be extended as an alarming system to prevent rash driving according to different road conditions.

5.1 Performance Metric

We have used the metric of *F1-Score* to evaluate the accuracy of prediction for unknown test data where True Positive(TP), False Positive(FP), and False Negative are explained following in Table 7.

5.2 Performance Evaluation of Learning Models

The performance of learning model is measured based on the accuracy of prediction of speed class for a particular road segment; compared to the actual speed class for that segment. Obtained *Accuracy* for each of the four classification algorithms is shown in Figure 8, and detail results of Accuracy(%), Recall, Precision, and *F1-Score* of each classifier algorithm are presented in Table 8. It clearly shows that GB outperforms all classification algorithms with the highest *F1-Score* of 0.8345. We have considered four categories of speed classes i.e., *slow*, *normal*, *fast* and *veryfast* depending on different ranges of speed. We have presented four cases of different speed ranges, appropriate for our case study, as shown in Table 9.

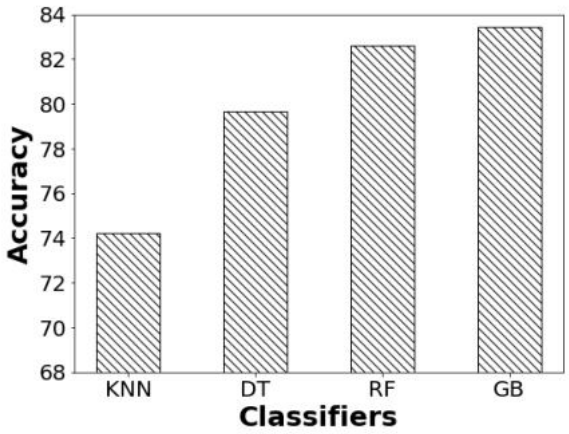


Figure 8. Obtained Accuracy for all classification algorithms.

Table 8. Comparison of KNN, DT, RF and GB algorithms on test data set in terms of Accuracy, Precision, Recall and *F1-Score*.

Classifiers	Accuracy (%)	Recall	Precision	*F1-Score*
KNN	74.46	0.7551	0.7438	0.7447
DT	79.86	0.8046	0.7930	0.7875
RF	82.18	0.8378	0.8171	0.8143
GB	**83.40**	0.8401	0.8382	**0.8345**

Table 9. Speed Ranges of Multiple Cases for Different Classifications.

	Slow km/hr	*Normal* km/hr	*Fast* km/hr	*Veryfast* Km/hr
Case 1	0-26	26-32	32-39	>39
Case 2	0-20	20-35	35-50	>50
Case 3	0-15	15-30	30-45	>45
Case 4	0-25	25-40	40-55	>55

Table 10. Precision, Recall and *F1-Score* of GB algorithm for multiple instances of speed classes

Speed	Case 1			Case 2			Case 3			Case 4		
	P	R	*F1-Score*	P	R	*F1-Score*	P	R	*F1-Score*	P	R	*F1-Score*
Slow	0.78	0.74	0.76	0.85	0.75	0.8	0.89	0.72	0.79	0.85	0.76	0.8
Normal	0.79	0.73	0.76	0.86	0.91	0.88	0.86	0.83	0.84	0.85	0.90	0.87
Fast	0.64	0.66	0.65	0.75	0.67	0.71	0.8	0.84	0.82	0.74	0.58	0.65
Veryfast	0.75	0.82	0.78	0.68	0.71	0.70	0.75	0.73	0.74	0.62	0.76	0.68

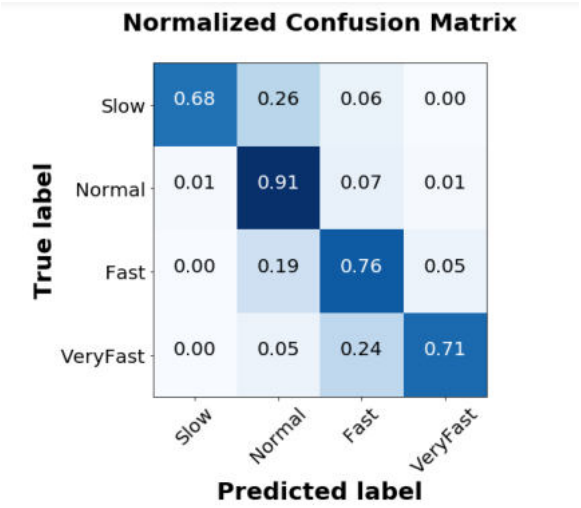


Figure 9. Confusion matrix for GB classification algorithm.

We have classified the speed categories of vehicles, based on our observations. We have applied a sharp threshold on speed ranges to distinguish two adjacent speed classes. But avoidance of human perception in choice of these thresholds is non-trivial. Therefore, We have excellent volunteer experience, and based on their expert observation, we have considered four instances where we have slightly changed the boundary thresholds to define different speed classes. Due to minimal deviation in speed at boundary conditions, a *slow* may be predicted as *normal* or vice versa. It may happen for each adjacent classes, which reduces the accuracy of our prediction model. To justify such fuzziness in boundary conditions of each speed classes, we have carried out our experiment for four sets of different speed ranges, which shows that little changes in speed ranges do not affect the accuracy drastically, as shown in Table 10.

It shows that accuracy of the classifier algorithm lies within a range of 0.8 to 0.88 for 37.5% of cases, 0.7 to 0.79 for 37% of times, and lies within a range of 0.65 to 0.68 for 18.75% of times.

A class wise analysis shows that the *F1-Score* varies from 0.76 to 0.8 for *slow* speed class, from 0.76 to 0.88 for *normal* speed class, from 0.65 to 0.82 for *fast* speed class and finally from 0.68 to 0.78 for *veryfast* speed class. The lesser amount of variation is observed in *F1-Score* for *slow* and for *veryfast* speed class because these two classes are of extreme category whereas *normal* and *fast* classes are intermediate and are very hard to be hard classified. This indicates that a fuzzy classification, especially for *normal* and *fast* speed class, may improve the result. The performance of the GB classifier is presented in a confusion matrix as shown in Figure 9, which shows the confusion in the prediction of adjacent speed classes. We have used 70% of the dataset as a training dataset, and the remaining 30% of data have been used as test dataset for cross validation. The confusion matrix shows the very high accuracy of prediction for the case of *normal* speed class and shows the least accuracy for *slow* speed class because we have more instances for *normal* speed class than *slow* speed class instances. This is obvious because traffic behavior remains *normal* for maximum city road segments, and *slow* instances of speed classes are observed in very few road patches.

5.3 Impact of zone on Speed

We have analyzed the performance of best classification algorithm i.e., GB for different zones, which we have observed for the city of our case study. We have divided the dataset according to the specific zone, and we have trained the learning model with 70% of the dataset for a particular zone. The residual 30% data for a zone have been used as test data. Table 11 shows the *F1-Score* values for each zone for the GB classification algorithm, where accuracy of prediction of speed class at the market zone is satisfactorily high (i.e., 0.8976) and accuracy for highway zone is comparatively less (i.e.,

0.6535). Normal city road shows a moderate accuracy (i.e., 0.8001). Table 12 presents the zone-wise prediction and their percentage. We have obtained a good accuracy for market area because the road features which we have considered to train our learning model, are highly co-related with almost all types of congestion factors.

Table 11. Zonewise Accuracy of GB Classifier.

Zone	*F1-Score*
Highway	0.6535
Market	0.8976
Normal City	0.8001

In a suburban city of a developing country, there is no separate market complex. The shops are built on both roadsides basically on pavements. Moreover, there are no separate lanes for different modes of transport as well as for pedestrians. All types of vehicles and paratransit like a scooter, cycle, motor-bike, auto-rickshaw, bus, and pedestrians travel along the same lane of the same road segment. Sometimes the upstream and downstream traffic moves on the same lane.

Table 12. Zonewise Prediction and Percentage for GB classifier.

Actual and Predicted	Zone	Count	%
Slow (0.68)	Highway	4	4.08
	Market	88	89.8
	Normal city	6	6.12
Normal (0.91)	Highway	29	3.59
	Market	579	71.75
	Normal city	199	24.66
Fast (0.76)	Highway	69	23.08
	Market	85	28.43
	Normal city	145	48.49
Veryfast (0.70)	Highway	79	84.95
	Market	85	28.43
	Normal city	145	48.49

Due to these reasons, a market area of such a city can be characterized by all types of congestion factors; likewise, a) a huge load of vehicles. b) a high number of intersections of other roads. c) high human activity and d) bad road surface (i.e., due to improper maintenance). Therefore, the features like honk duration (co-related with loads of vehicles), number of road intersections (co-related with number of intersecting roads), WiFi (co-related with high human activity) and road surface index (co-related with road condition) become prominent for market area and captures all congestion factors efficiently. But in highway segments, such congestion factors are missing. The highway road segments are wide enough.

Although a large number of vehicles move along the highway, those do not get stuck in the traffic jam and do not make the sound of horns. The highways in a city are constructed for congestion-free fast communication at long distances. Therefore, number of intersecting city roads on highway segments is very less. The number of pedestrians is very less on the highway, and the road surface is also well-maintained by the highway authority. Therefore, sometimes, the features like honk duration, intersection density, WiFi density, and road surface shows similar behavior with normal city roads. By observing each feature value of highway segments, we have found that WiFi density is the most overlapping feature between normal city roads and highway. In a suburban city of a developing country, a number of smartphone and internet users are still very less, and the facility of WiFi is also very limited.

5.4 Experiment on Timelevel and Zone

We have analyzed the performance of the GB classifier algorithm for different hours of the day and also for different zones. The time levels are labeled as early morning, mid-day, evening, and night time. Figure 10 shows the plot of accuracy obtained on the basis of time levels. The accuracy thus obtained is 73.5% for early morning, 77.66% for mid-day, 83.46% for the evening, and 76.97% for night time. We find that the accuracy obtained during the evening is satisfactorily high, with the least being during the early morning. Moreover, all features are prominent enough for *slow* speed class and in the evening mostly *slow* speed class is observed. But in the early morning, the traffic remains either *normal* or *fast* and *veryfast*.

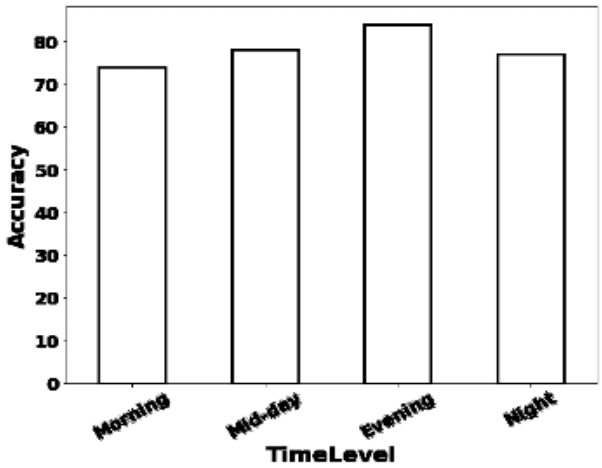


Figure 10. Accuracy of GB classifier algorithm at different times.

Table 13. Understanding Model Behavior by LIME.

Seg ID	Class	Rank of features	Inference
54	*Slow*	Honk Duration-6, Intersections-3,	Huge load of vehicles, more crossing and high
59	*Slow*	Intersections-3, WiFi-24	More crossing and very high human activity.
60	*Normal*	WiFi-69, Intersection-1	High human activity.
50	*Normal*	Intersection-5, RSI-3	More crossing and poor road surface condition.
72	*Fast*	RSI-1, WiFi-2	Good road surface condition and moderate human
46	*Fast*	Intersections-0, Honk Duration-0,	No crossing, No congestion of vehicles and good
67	*Veryfast*	RSI-1,Intersections-0	Good road condition and no crossing.
77	*Veryfast*	RSI-1, Honk Duration-0	Good road condition and no congestion of vehicles.

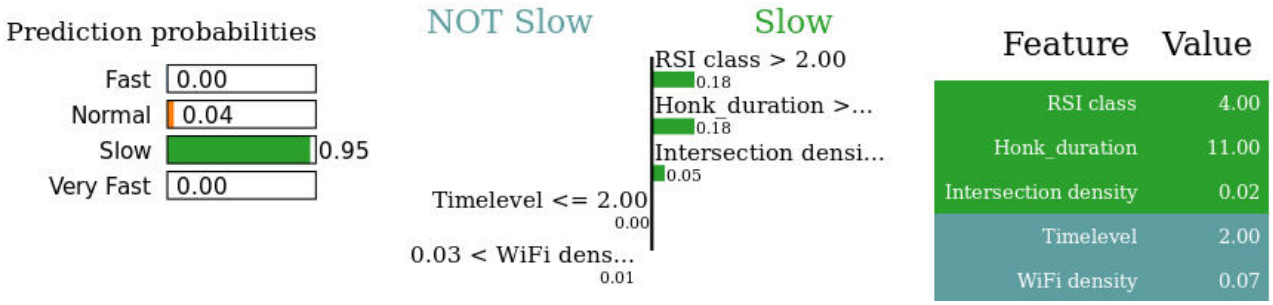


Figure 11. Feature Behaviour Analysis using LIME

5.5 Interpreting Machine Learning Results using LIME

From the above discussion, we have already achieved 83.45% accuracy in prediction on the unknown test dataset by the GB classifier algorithm. This accuracy ensures the strong correlation of road features and speed of vehicle along a particular city road segments as well as this accuracy confirms the worthiness of all selected road features for a right prediction of traffic state. In this particular section, we have contributed to the extraction of additional information as more as possible. Till now, we were satisfied with only information about the traffic state of city roads, but we have tried to gain insights about the reason behind a particular classification of speed class by classifier model, especially *slow* speed class. The actual reason for *the slow* traffic state will help a traveler to plan a route according to her mode of transport, comfort, and punctuality.

5.5.1 Overview of LIME

To understand the individual classification of a road segment by a classifier model, we have used an open-source framework LIME [15]. LIME perfectly outlines the basic ideas behind its explanation methods. Model-agnosticism LIME does not make any assumptions about the learning model whose prediction is explained. It treats the learning model as a black-box, so the only way that it has to understand the behavior of the model. It can be applied to any machine learning model. The technique attempts to understand the model from input of data samples and to understand how the predictions change by analyzing the internal components and how they interact.

LIME provides local model interpretability. Explanations must be easy to understand by users, which is not necessarily true for the feature space used by the model because it may use too many input variables. LIMEs explanations use a data representation (called interpretable representation) that is different from the original feature space. LIME modifies a single data sample by tweaking the feature values and observes the resulting impact on the output. The output of LIME is a list of explanations, reflecting the contribution of each feature to the prediction of a data sample. It provides local interpretability, and it also allows us to determine which feature changes will have the most impact on the prediction. LIME explains a single prediction(i.e., local behavior of single instance). The features are ranked based on their contribution to a particular prediction. From the features of high rank, we can illuminate the proper reason of a prediction.

Table 13 shows a few particular instances of prediction for our case study, where a clear inference is possible from feature ranking by LIME. Figure 11 shows the explanation of LIME for one prediction instance of our case study. It is explaining a *slow* instance with all feature values which have a major impact on this prediction. The probability of this prediction is 0.95. There are 3 features RSI, Honk duration, and Intersection density which have major contribution to this prediction. RSI or Road Surface Index is 4, which interprets a very bad road surface; Honk duration is also high (11 sec.), which indicates traffic congestion and a load of vehicles, and Intersection density has a value of 0.02 which means a moderate number of crossing along the segment. Among these features, road

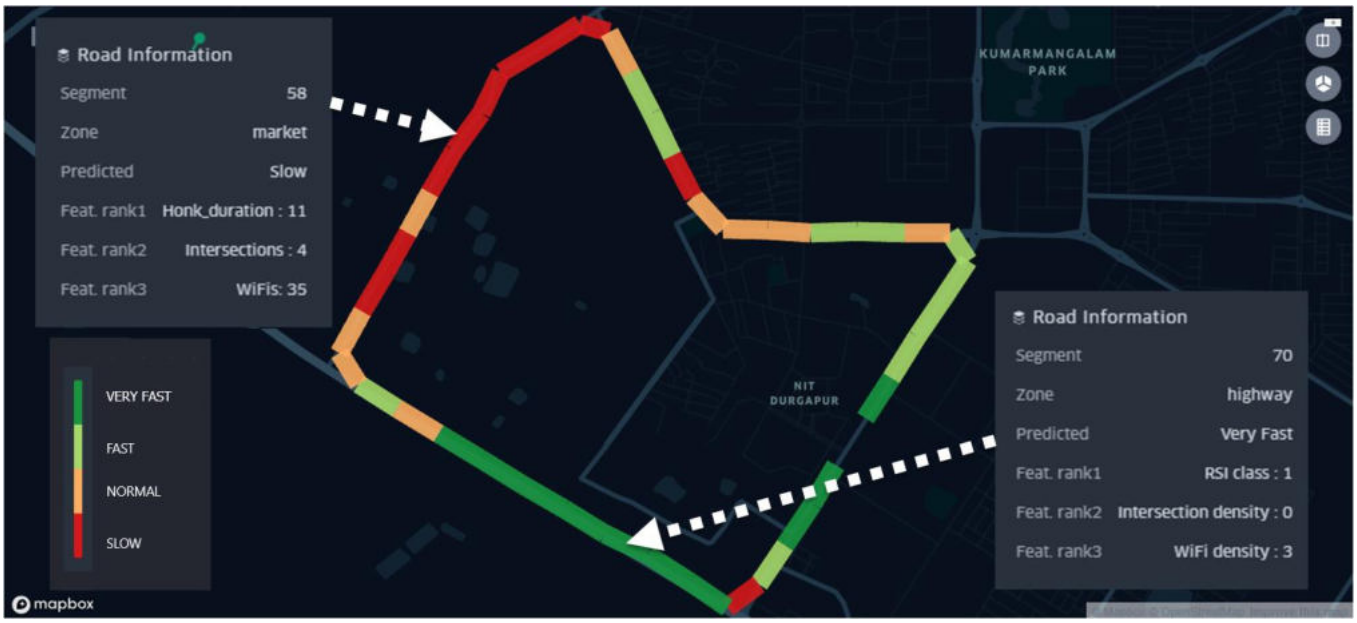
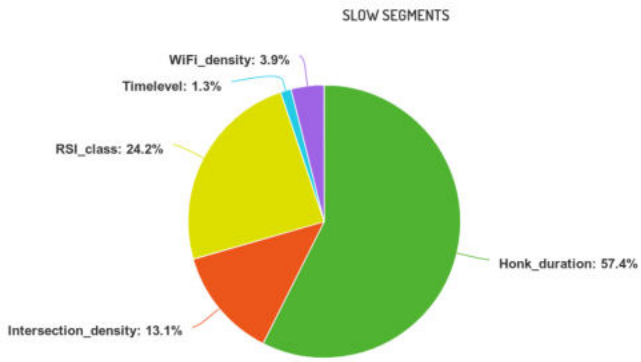


Figure 12. Visualization on OpenStreetMap

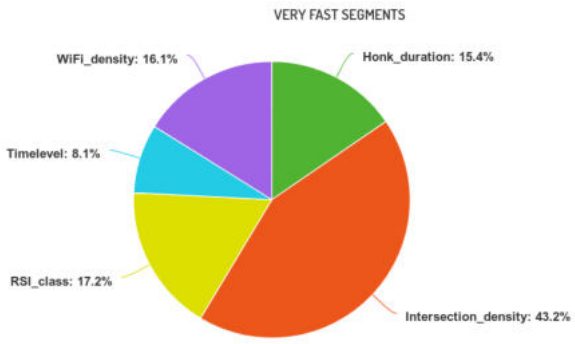
surface condition is the most prominent reason for *slow* speed class, and then traffic congestion and number of signals are also contributing to this prediction. According to the fitness of the features in prediction, the features are ranked by its individual weight, calculated by LIME by fitting a linear model on feature values and their target class. Weights of the linear model are used as an explanation of the decision. The weight of RSI, Honk duration, and Intersection density are 2.00, 0.18, and 0.05, respectively.

5.6 Map Visualisation of LIME

A web interface and Android application are developed to facilitate travelers by providing a traffic information service with proper reasoning and all road relevant information, which may be inferred from feature values. Figure 12 shows a snap of a trail that is already analyzed in Table 13. Our developed system provides not only the traffic information but also displaying the zone and most dominant factors behind predicted traffic state. Figure 13(a) shows that for 57.4% of trails, Honk duration is the prominent feature for prediction as *slow* speed class. For 24.2% trails RSI and for 13.1% trails, Intersection density work as the main factor for *slow* speed class prediction. For rest of the trails, either WiFi density or Timelevel work. These results show that Honk duration is most worthy for *slow* speed class prediction in our case study. Figure 13(b) shows that for 43.2% of trails, Intersection density works most for prediction of *veryfast* speed class. For 17.2% of trails, RSI works, and for 15.4% of trails, Honk duration works as the most prominent factor for *veryfast* prediction of speed class.



(a) Contribution of all features for *Slow* prediction.



(b) Contribution of all features for *veryfast* prediction.

Figure 13. Contribution of all features for *slow* and *veryfast* prediction.

6. Application of RoadSpeedSene: Assessment of Driving Behavior

The above study helps to understand the effective speed according to road characteristics for a particular road segment of the suburban cities in developing countries. We have carried out a detailed experiment for each zone separately at different hours of the day. Therefore, in this section, we have assessed the driving style of a person where the effective speed limit of the particular segment is already known. We have tested the model on speed signatures of 4 different drivers along with the same road segments at the same time of the day.

First, we have estimated the effective speed limit of each segment by our learning model and have counted the number segments which come under the *fast* and *veryfast* speed class. After this step, we analyze the driving signature for every driver for each road segment separately. Among the total segments, the only segments of the *fast* and *veryfast* category which match with individual driver's performance are calculated for each driver separately. We also calculate the number of segments where driver has run with *fast* and *veryfast* speed, exceeding the estimated effective speed limit by learning model. Basically, only *fast* or *veryfast* driving does not always imply rush driving because, if road characteristics are in favor of high-speed driving, then it should not be considered as rush driving. The only segment where driver has run with overspeed should be considered as rush driving. If the total number of segments along a route is n and number of segments (S_i) with effective speed limit ($Speed_i$) of *fast* or *veryfast* category which, also matches the speed of driver (DS_i) is denoted as H as shown in Equation 7. The number of segments where a driver runs at *fast* or *veryfast* speed, exceeding the estimated speed limit, is denoted as L, as shown in Equation 8, and performance metric of a driver is denoted as *Driverscore* which can be expressed as Equation 9.

$$H = \sum_{i=1}^{n} Seg_i \quad \forall\, Speed_i > 40km/hr \;\&\; Speed_i = DS_i \qquad \ldots(7)$$

$$L = \sum_{i=1}^{n} Seg_i \quad \forall\, DS_i > 40km/hr \;\&\; DS_i > Speed_i \qquad \ldots(8)$$

$$Driver_{score} = \frac{H}{L} \qquad \ldots(9)$$

Figure 14 shows both the percentage of total road segments where driver's speed is similar to the estimated speed i.e., either *normal* or *fast* and also the road segments where driver has exceeded the estimated speed Overspeed for each of six drivers on same road segments at the same time. In summary, to measure a driver's performance, we need to consider the ratio of over-speed to *fast* or *veryfast* speed, as shown in Figure 14. Here, a driver's score will be maximized if he drives all road segments with estimated speed by our learning model, and it will decrease with over-speed driving. Table 14 shows the accuracy of prediction of the classification model i.e., GB and the recorded speed for each driver individually. The accuracy is calculated again with metric of *F1-Score*. From Table 14, minimum accuracy of 0.77 is observed, whereas the highest accuracy is 0.93.

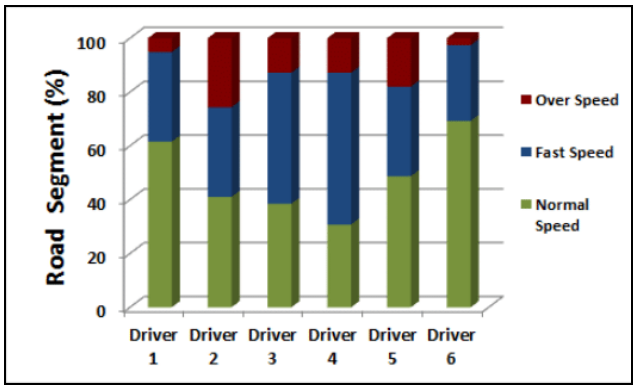


Figure 14. Driver Performance

Table 14. Accuracy of prediction by classification model and individual driving speed.

Driver ID	Accuracy
Driver-1	93.2
Driver-2	77
Driver-3	89.7
Driver-4	84.6

7. Conclusion and Future Scope

RoadSpeedSense analyses smartphone sensor data and extract the significant fetures which contribute to estimating the practical speed limit for a road segment. We selected some features that are derived from different smartphone sensors that can be available via crowdsourcing and used them in a supervised traffic speed estimation approach to estimate the traffic state. We validated the estimations in a controlled environment against volunteer provided labels and evaluated the accuracy of the approach. In the best case, we achieved 83.4% accuracy in estimating the traffic speed by the Gradient boosting approach. We are also analysis the reason for *slow* traffic estimation by popular interpretable machine learning models like LIME. Our *RoadSpeedSense* system also measures an excellent driver rating with accuracy ranging between 0.77 to 0.93 according to road characteristics. Our model can estimate the percentage of road patches along which a driver has over speed for a particular trip. Though in this paper, we

do not address the common issues of crowdsourcing like incentive strategies for data providers, it instead focuses on making interesting estimations of traffic state from the crowdsourced data when the low volume of data is available to model that can enrich the existing map services.

References

[1] Astarita, V., Vaiana, R., Caruso, M.V. , Iuele, T. , Giofr, V., Masi F. (2014) Automated sensing system for mon- itoring of road surface quality by mobile devices. 16th Meeting of the EURO Working Group on Transportation 111, 242–251

[2] Bhoraskar, R., Vankadhara, N., Raman, B., Kulkarni, P. (2012) Wolverine: Traffic and road condition estimation us- ing smartphone sensors. 4th International Conference on Communication Systems and Networks (COMSNETS 2012), 1–6

[3] Biagioni, J., Gerlich, T., Merrifield, T., Eriksson, J. (2011) EasyTracker: Automatic Transit Tracking, Mapping, and Arrival Time Prediction Using Smartphones, in Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems. SenSys' 1 1 (ACM, New York, NY, USA, 2011), pp. 68–81

[4] Chowdhury, A., Chakravarty, T., Balamuralidhar, P. (2016) A novel approach to improve vehicle speed estimation using smartphones ins/gps sensors. 8th International Conference on Sensing Technology, 441–446

[5] Eriksson, J., Girod, L., Hull, B., Newton, R., Madden, S., Balakrishnan, H. (2008) The pothole patrol: Using a mobile sensor network for road surface monitoring. Proceed- ings of the 6th International Conference on Mobile Systems, Applications, and Services (ACM MobiSys), 29–39

[6] Hoang, M.X., Zheng, Y., Singh, A.K. (2016) FCCF: Forecasting Citywide Crowd Flows Based on Big Data, in Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. SIGSPACIAL'16 (ACM, New York, NY, USA, 2016), pp. 6–1610

[7] Izabel A., Tostes, J. , Duarte-Figueiredo, F. , Mar- tins, A., Salles, J., Loureiro, A.(2013) From data to knowledge: City-wide traffic flows analysis and prediction using bing maps. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 12–1128

[8] Li, X., Goldberg, D.W. (2018) Toward a mobile crowdsens- ing system for road surface assessment. Computers, Environment and Urban Systems 69, 51–62

[9] Mandal, R., Agarwal, N., Nandi, S., Das, P., Anvit, A., Sanyal, S., Saha, S. (2015) Stoppage pattern analysis of public bus GPS traces in developing regions, in IEEE International Conferenceon Pervasive Computing and Communication Workshops (PerCom Workshops), pp. 276–279

[10] Mandal, R., Agarwal, N., Das, P., Pathak, S., Rathi, H., Nandi, S., Saha, S. (2014) A System for Stoppage Pattern Extraction from Public Bus GPS Traces in Developing Regions, in Proceedings of the Third ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems, MobiGIS'14.

[11] Mohan, P., Padmanabhan, V.N., Ramjee, R.(2008) Nericell: Rich Monitoring of Road and Traffic Conditions Using Mobile Smartphones. in SenSys '08, (ACM, New York, NY, USA, 2008), pp. 323–336

[12] Nguyen, D.B., Dow, C., Hwang, S. (2018) An efficient traffic congestion monitoring system on internet of vehicles. Wireless Communications and Mobile Computing, 9136813–1913681317

[13] Park, S., Han, H., Kim, B.S., Noh, J.H., Chi, J., Choi, M. J. (2018) Real-time traffic risk detection model using smart mobile device. Sensors 18(11), 3686

[14] Perttunen, M., Mazhelis, O., Cong, F., Kauppila, M., Leppanen, T., Kantola, J., Collin, J., Pirttikangas, S., Haverinen, J., Ristaniemi, T., Riekki, J. (2011) Distributed Road Surface Condition Monitoring Using Mobile Phones, in Proceedings of the 8th International Conference on Ubiquitous Intelligence and Computing. UIC'11 (Springer, Berlin, Heidelberg, 2011), pp. 64–78. ISBN 978-3-642-23640-2

[15] Ribeiro, M.T., Singh, S., Guestrin, C. (2016) "Why Should I Trust You?": Explaining the Predictions of Any Classifier, in Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16 (ACM, New York, NY, USA, 2016), pp. 1135–1144. ISBN 978-1-4503-4232-2

[16] Sen, R., Siriah, P., Raman, B. (2011) RoadSoundSense: Acoustic sensing based road congestion monitoring in developing regions, 8th Annual IEEE Com- munications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks, pp. 125–133

[17] Sen, R., Raman, B., Sharma, P. (2010) Horn-ok-please, in Proceedings of the 8th International

Conference on Mobile Systems, Applications, and Services. MobiSys'10 (ACM, New York, NY, USA, 2010), pp. 137–150. ISBN 978-1-60558-985-5

[18] Shang, J., Zheng, Y., Tong, W., Chang, E., Yu, Y. (2014) Inferring Gas Consumption and Pollution Emission of Vehicles Throughout a City, in Proceedings of the 20th ACM SIGKDD International Conference on Knowl- edge Discovery and Data Mining. KDD '14 (ACM, New York, NY, USA), pp. 1027–1036

[19] Silva, N., Shah, V., Soares, J., Rodrigues, H. (2018) Road anomalies detection system evaluation. Sensors 18(7)

[20] Thiagarajan, A., Ravindranath, L., LaCurts, K., Madden, S., Balakrishnan, H., Toledo, S., Eriksson, J. (2009) VTrack: Accurate, Energy-aware Road Traffic De- lay Estimation Using Mobile Phones, in Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems. SenSys '09 (ACM, New York, NY, USA, 2009), pp. 85–98. ISBN 978-1-60558-519-2

[21] Vij, D., Aggarwal, N.(2018) Smartphone based traffic state detection using acoustic analysis and crowdsourcing. Applied Acoustics 138, 80–91

[22] Wang, S., Zhang, X., Cao, J., He, L., Stenneth, L., Yu, P.S., Li, Z., Huang, Z. (2017) Computing urban traffic con- gestions by incorporating sparse gps probe data and social media data. ACM Trans. Inf. Syst. 35(4), 40–14030

[23] Xin, X., Lu, C., Wang, Y., Huang, H. (2015) Forecasting Col- lector Road Speeds Under High Percentage of Missing Data, in Proceedings of the Twenty-Ninth AAAI Con- ference on Artificial Intelligence, 2015, pp. 1917–1923

[24] Yu, J., Zhu, H., Han, H., Chen, Y.J., Yang, J., Zhu, Y., Chen, Z., Xue, G., Li, M. (2016) Senspeed: Sensing driving conditions to estimate vehicle speed in urban envi- ronments 15, 202–216

[25] Zong, X., Wen, X.(2015) A new approach to estimate real- time traveling speed with accelerometer. Interna- tional Journal of Distributed Sensor Networks , 1–16