Know Thy Neighbor - A Data-Driven Approach to Neighborhood Estimation in VANETs

Karsten Roscher[1], Thomas Nitsche[1] and Rudi Knorr[1,2]
[1]Fraunhofer Institute for Embedded Systems and Communication Technologies ESK, Munich, Germany
[2]Chair for Communication Systems, Institute of Computer Science, University of Augsburg, Germany
Email: {karsten.roscher, thomas.nitsche, rudi.knorr}@esk.fraunhofer.de

***Abstract*—Current advances in vehicular ad-hoc networks (VANETs) point out the importance of multi-hop message dissemination. For this type of communication, the selection of neighboring nodes with stable links is vital. In this work, we address the neighbor selection problem with a data-driven approach. To this aim, we apply machine learning techniques to a massive data-set of ETSI ITS message exchange samples, obtained from simulated traffic in the highly detailed *Luxembourg SUMO Traffic (LuST) Scenario*. As a result, we present classification methods that increase neighbor selection accuracy by up to 43% compared to the state of the art.**

I. Introduction

Vehicular networking based on IEEE 802.11p has recently seen a significant increase in popularity with the up-coming legislation of the DSRC standard in the United States. Further, with the standardization of the competing LTE-V2X feature in 3GPP Release 14 [1], the race for the technological enabler of future connected mobility is opened. In this context multi-hop ad-hoc network architectures are being reconsidered. This is even the case for V2X in 5G [2], since ubiquitous cellular coverage is impractical. With ad-hoc networking also being inherent to IEEE 802.11p based technologies as for example ETSI ITS-G5, a central question for dynamic network structures reemerges: *In the fast changing topologies of vehicular networks, which of your neighbors is a reliable communication partner?*

Vehicular network topologies change quickly. Vehicles passing each other on opposite lanes of a highway easily reach relative speeds of up to 300 km/h. In urban scenarios buildings create severe signal blockage towards a receiver which turns around a corner. Thus, proper selection of neighboring nodes, that ensure a reliable link at least for the next upcoming messages is challenging. But, transmitting to a neighboring node that unexpectedly gets out of reach does not only impact multi-hop routing strategies. It also hinders other dissemination aspects. For example geo-messaging according to the ETSI ITS standard buffers messages based on a neighborhood decision [3], and also heterogeneous wireless technology selection can be based on the current set of neighbors. Correct classification of the neighborhood relationship is thus of very high importance for vehicular networking.

Currently, the state of the art for neighbor selection in VANETs is defined by a threshold for the time since the last direct message of a node as specified in the ETSI GeoNetworking standard [3]. In the scientific literature, more complex decision criteria are considered, however neighbor selection is predominantly seen to be part of ad-hoc network routing. Unfortunately, most of this work is not applicable since the distinct dynamics of vehicular networks are omitted. This also accounts for [4] and [5] where neighbor classification is addressed explicitly. Exceptions that assume vehicular node movement are either focusing on vehicle grouping for long term routing stability [6] or leverage information from a modified PHY-layer to estimate link stability [7]. More recently, Hoang et. al [8] consider reliable neighbor selection for their cooperate positioning algorithm but focus on the GPS data fusion aspect.

Contrary to the work mentioned before we explore neighbor selection in vehicular networks as an independent problem, with a data driven approach. By extensive simulation of ETSI ITS communication using the ezCar2X [9] framework in combination with the *Luxembourg SUMO Traffic (LuST) Scenario* [10] we obtain a wide data basis, which we use to analyze the behavior of neighbor classification in complex traffic scenarios. From our data-set, which reflects the success of message exchange with network neighbors, we analyze the potential of several classification features to reliably predict network neighborhood. Then, we evaluate the performance of various classification methods for neighbor selection in vehicular ad-hoc networks, which are based on the selected features. In particular, the contributions of our work are as follows:

- From a wide set of classification features, we determine the most relevant ones for neighbor selection in vehicular networks. Our findings are valid for a broad set of traffic densities and message update rates.
- Further, the performance of several neighborhood classification methods is evaluated on highly detailed VANET simulations based on a realistic city scenario.
- From this we propose improved thresholds for current state of the art for neighborhood decision based on last-message-received thresholds.
- Last, we give a recommendation on which classification method is best used for different vehicular network applications.

This work is structured as follows. Section II describes our simulation setup. Section III evaluates the significance of different classification features, which are used in Section IV

to evaluate the performance of neighbor classification methods. Section V concludes the paper.

II. Simulating A Realistic City Scenario

Our data-driven approach requires a sufficient amount of input data covering a variety of situations to avoid biasing the classification algorithms towards special cases. The *Luxembourg SUMO Traffic (LuST) Scenario* [10] fits the requirement well. It provides 24 hours of mobility simulation in the city of Luxembourg and the surrounding highways. The number of active cars ranges from a few dozens during night to more than 5000 in rush hours. We selected three different time periods over day to cover varying traffic densities as summarized in Table I.

TABLE I
Traffic Densities

Density	Time Period	Vehicles
low	11:05am - 11:20am	1700
medium	1:20pm - 1:35pm	3200
high	8:15am - 8:30am	5000

In total we captured 17 features for each transmitted packet. The data includes information about the sender: position, velocity, channel busy ratio measurement and size of the neighbor table; and about a potential neighbor: position, velocity, update history and received signal power. Furthermore, we derived relative metrics like distance, relative speed and heading. We intentionally limited the features to data that is directly available or can be collected without additional communication overhead to ease integration into existing protocols. The status of neighbors is derived from periodic GeoNetworking updates (beacons [3] or Cooperative Awareness Messages (CAMs) [11]) and therefore reflects the state of the last successfully received packet. Local data is assumed to be sampled from a positioning device with a fixed frequency of 10 Hz.

Unicast and broadcast transmissions on the medium access (MAC) layer behave differently. While the former uses acknowledgements and a retransmission scheme to improve reliability, the latter is simply transmitted once. Since we did not want to limit our investigation to a specific usage scenario, we investigated both modes separately. A similar approach was applied to different beaconing schemes for status updates, see Table II: GeoNetworking beacons [3] and Cooperative Awareness Messages (CAMs) with adaptive [11] and 10 Hz fixed frequency. For each combination of scenario, beaconing scheme and transmission mode we collected approximately 1.5 million samples of which 75% were used as training data and the remaining 25% for testing.

The simulation environment consists of the network simulator ns-3 [12], the traffic simulator SUMO [13] and the ezCar2X framework [9] implementing the ETSI ITS protocol stack. The main simulation models and parameters are summarized in Table III. Data analysis and statistical inference are based on Scikit-learn [14].

TABLE II
Beaconing Schemes

Scheme	f_{min}	f_{max}
GeoNetworking [3]	0.27 Hz	0.33 Hz
CAM adaptive [11]	1 Hz	10 Hz
CAM 10 Hz	10 Hz	10 Hz

TABLE III
Simulation Parameters

Parameter	Value
Standard	802.11p
Tx Power, Tx Rate	23 dBm, 6 Mbps
Propagation Model	Cheng et al. [15] with [16]
Penetration Rate	20%
Neighbor Timeout	20 s [3]
Packet Size	300 Bytes
Simulation Runs	32
Duration per Run	1000 s

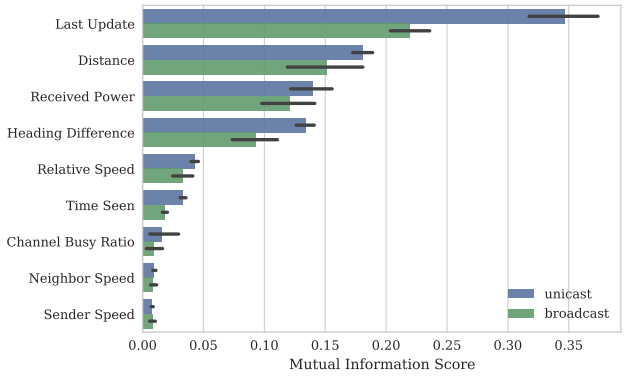

Fig. 1. Feature relevance based on mutual information estimates averaged over all traffic densities and beaconing schemes.

III. Feature Selection

In this section we identify the most predictive features for the classification of neighbor relationships in VANETs. Using fewer features can increase the generalization performance of a classifier and also reduce its complexity. We ranked all features based on an estimate of the mutual information [17] between the feature and a successful reception of the packet. Figure 1 summarizes the results for the most relevant predictors with a distinction between unicast and broadcast transmissions. In both cases, the four main features are: time since the last received update, distance between the nodes, signal strength of the last received packet and the difference between the headings (driving direction).

In case of unicast all features appear to carry more information about the current status of a potential neighbor. This can be explained by the retransmission scheme that compensates for some of the random error, e.g. introduced by fast fading, that cannot be deduced from the collected data. The time since the last update alone appears to be a very powerful predictor for the current connectivity - especially for unicast.

This observation is also in line with the common approach of using timeout values to determine if another node is a direct neighbor or not [3].

The relevance of most features is consistent across scenarios and beaconing schemes, with one exception: channel busy ratio (CBR). In most of our scenarios it carries little information since we see average busy ratios ranging from 0.05% to 3.8% with rare hotspots of up to 60%. However, with an increase in generated messages and road traffic density its relevance for the prediction increases as well. We expect CBR as a feature to be even more significant in dense and crowded situations with a constantly saturated channel.

In addition, we ranked the features using the feature relevance derived from a Random Forest [18] classifier as well as a forward step-wise selection as described by Hastie et al. [19]. Both were consistent in identifying the main features but showed diverging results for the less significant variables.

IV. Neighbor Classification

We model the estimation of the current relationship with a remote node r as a binary classification problem, where an input vector $\vec{x}_r$ consisting of p features is used to make predictions of an output y_r denoted by $\hat{y}_r$ with $\hat{y}_r \in [0, 1]$. r is considered to be a neighbor if $\hat{y}_r > 0.5$ [19]. Our goal is to find a suitable approximation function $\hat{y}_r = \hat{f}(\vec{x}_r)$.

A. Classification Methods

In the following we describe the classification methods that were evaluated on the datasets. We differentiate between the state of the art threshold for last-message-received times and machine learning concepts.

Time Threshold. Our baseline classification function uses a threshold t_{up} for the time $x_{r,t}$ since the last status update from remote node r:

$$\hat{f}_t(\vec{x}_r) = \begin{cases} 1, & \text{if } x_{r,t} \leq t_{up} \\ 0, & \text{if } x_{r,t} > t_{up} \end{cases} \tag{1}$$

This rule is widely applied in different routing protocols and also used in the ETSI GeoNetworking standard where the default threshold $t_{up,GN}$ is defined to be 20 s [3].

However, 20 s appear to be very long especially compared to the short beaconing intervals between 0.1 s (CAM with 10 Hz) and 3.75 s (GeoNetworking beacon). We therefore used the training data to compute the optimal threshold $t_{up,opt}$ for the different data sets. We intend to maximize the classification accuracy, thus we minimized the zero-one loss function:

$$L(y, \hat{y}) = \frac{1}{n} \sum_{1}^{n} I(y \neq \hat{y}) \tag{2}$$

where I is the indicator notation and n is the number of training samples.

Figure 2 shows the optimal thresholds for the individual scenarios, beaconing schemes and transmission modes. It is obvious that the value highly depends on the update frequency. Higher update rates lead to lower thresholds. There is also a significant difference between unicast and broadcast. While

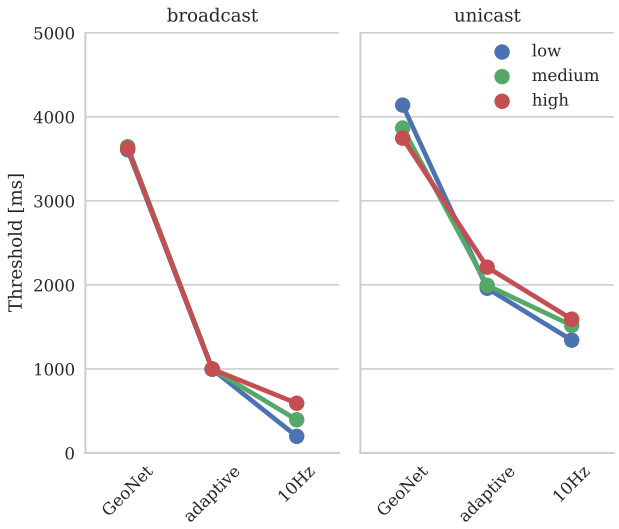

Fig. 2. Optimal update thresholds for low, medium and high traffic densities with varying beaconing schemes.

broadcast needs to cope with more uncertainty due to random fading and collisions, unicast allows for higher threshold values due to the its increased reliability. However, the road traffic scenario appears to have very little influence. We also determined the optimal threshold for a combined data set with samples equally selected from all three traffic scenarios which lead to values almost identical to those for medium density.

Machine Learning. We applied several approaches from the machine learning domain taking all features or a subset thereof into account: k-nearest neighbors, decision trees, random forests and the multilayer perceptron (MLP) [19]. With proper parameter tuning they all provide a similar classification performance with less than 2% of deviation. We therefore focus on the MLP in this paper because once trained it has the lowest memory requirements (weight matrices only) and its output can be computed efficiently with basic linear algebra. This makes it the most suitable candidate for integration into embedded devices and networking protocols.

MLPs according to [20] are multi layered networks of nodes that apply a non-linear activation function $\sigma(\upsilon)$ to map their input to an output value:

$$z_{k,l} = \sigma(\beta_{k,l} + \vec{\alpha}_l^T \vec{z}_{l-1}) \tag{3}$$

where $z_{k,l}$ is the output of the k-th node in the l-th layer with $l \in [1, m]$, $\beta_{k,l}$ is a static bias term, and $\vec{\alpha}_l$ are the weights applied to the outputs $\vec{z}_{l-1} = \{z_{0,l-1}, ..., z_{i,l-1}\}$ of the i nodes of the previous layer $l-1$. The input is applied as $\vec{z}_0 = \vec{x}_r$. For binary classification the output layer consists of a single node with its own activation function σ_{out} to calculate $\hat{f}$ applying eq. 3 for each intermediate node:

$$\hat{f}_{MLP}(\vec{x_r}) = \sigma_{out}(\beta_{1,m} + \vec{\alpha}_m^T \vec{z}_{m-1}) \tag{4}$$

In our evaluation all input variables were standardized, i.e. scaled to have mean $\mu = 0$ and standard deviation $\sigma = 1$, before being fed into the network. Training of the MLP is

based on stochastic gradient descent as described in [21]. Parameter tuning was performed using a grid search [19] with cross-validation [22] on the training data. We found a MLP consisting of two hidden layers - with 50 nodes in the first and 20 nodes in the second - to perform best if all features are included. With fewer features smaller networks were considered as well. Further, we selected the rectified linear unit function $\sigma(\upsilon) = max(0, \upsilon)$ for the activation of hidden layer nodes while the output node uses a logistic activation $\sigma_{out}(\upsilon) = (1 + e^{-\upsilon})^{-1}$.

For the following results we trained the classifier on a combined data set with data from all three traffic scenarios. However, we evaluated the performance on each test set individually to identify potential bias issues with respect to the traffic densities. Besides accuracy, precision and recall we also calculated the Brier score [23] on the test data:

$$B = \frac{1}{n}\sum_{1}^{n}(\hat{y} - y)^2 \tag{5}$$

where n is the number of test samples and $B \in [0, 1]$. A low Brier score indicates that a classifier is capable of estimating not only the class but also the probability of that class with high accuracy.

B. Results: Unicast

We first investigated the classification performance on unicast transmissions comparing two last-message-received threshold classifiers (standard and optimal) as well as two MLPs, one with all features (MLP_{full}) and one where only the four most significant features according to Section III are considered (MLP_4).

Figure 3 shows the classification accuracy for each beaconing scheme averaged over all traffic scenarios. Overall, performance of classifiers increases with higher update frequencies due to more recent information from remote nodes. Furthermore, missing updates can be detected more quickly leading to shorter periods of uncertainty.

Comparing the individual classifiers reveals the standard neighbor timeout defined in [3] to perform only slightly better than random guessing. In contrast, the other three approaches perform significantly better with a slight advantage for the MLPs over the optimal last-message-received threshold. While the latter improves by up to 32% over the state of the art, the MLP based approaches achieve additional accuracy in the range of 2%. The false positive rate is around 11% for all three approaches with the MLPs offering better recall. The performance is consistent across traffic scenarios indicating good generalization properties. Further, the Brier score displayed in Fig. 5 shows that the MLP is much better in estimating the probability compared to the simple thresholding.

If the improvement offered by the MLPs is worth the added complexity depends on the application. We assume that the thresholding approach should suffice for most scenarios were only binary classification is required. However, if algorithms depend on an estimate for the class probability, MLPs are the better choice, e.g. in an advanced platooning algorithm

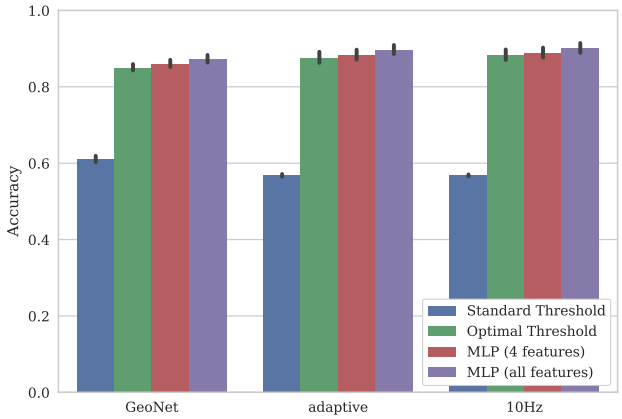

Fig. 3. Classification accuracy for unicast transmissions averaged over the traffic densities.

the controller could adapt the inter-vehicle gap based on the estimated probability to compensate for potential packet loss.

C. Results: Broadcast

Similar to the unicast results, Figure 4 shows the classification accuracy for broadcast transmissions. Broadcasting is relevant for routing algorithms with receiver-based next hop selection, e.g. Contention Based Forwarding in [3], and multicast or broadcast dissemination in general.

In contrast to the unicast observations all classifiers perform worse with increasing beacon rates since without retransmissions broadcast packets are much more susceptible to frame collisions. However, these cannot be predicted from the collected features leading also to an overall decrease in accuracy compared to unicast transmissions. Also, precision drops with higher network load, e.g. the optimal threshold classifier drops from 77% to 57% if CAMs with 10 Hz are used instead of GeoNetworking beacons.

Despite its high predictive value, the significance of *time since the last update* for broadcast packets is less compared to unicast. This explains why, in contrast to unicast communication, the MLPs with more than one input variable offer a substantial performance gain over the optimal last-message-received threshold. While the latter offers up to 38% improvement over the state of the art, the MLP_{full} achieves 4-5% additional accuracy and up to 8% more precision than the threshold approach. MLP_4 performs very equally, with exception of the 10 Hz scenario were the inclusion of channel busy ratio in the MLP_{full} leads to a slightly larger gap. Similar to unicast, results are consistent across scenarios indicating good generalization properties. The Brier score depicted in Fig. 5 also shows the advanced MLPs' probabilities estimation compared to the last-message-received threshold.

V. Conclusion

In this work, we addressed the problem of neighbor selection in VANETS with a data-driven approach. By methods from the field of statistical learning we evaluated the informative value of several neighbor classification features, which

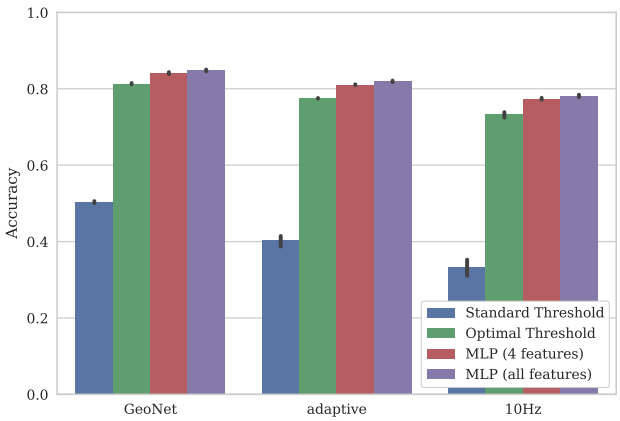

Fig. 4. Classification accuracy for broadcast transmissions averaged over the traffic densities.

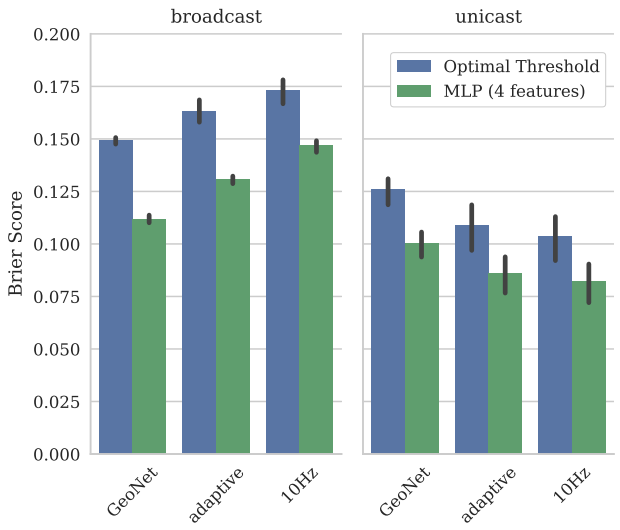

Fig. 5. Brier score for the optimal threshold and MLP with four features averaged over the traffic densities.

are obtainable in practice without additional communication overhead. We highlighted link-age, vehicle distance, received power and heading difference as the significant indicators and identified channel-busy-ratio as an additional promising candidate under high network load. From theses findings we analyzed the performance of a binary multilayer perceptron (MLP) classifier and reevaluated link-age-based neighbor selection as the current state of the art. We show that with a link-age threshold optimized for the beaconing rate of the communication system significant classification improvements above 30% can be achieved. Further, we show that our MLP based alternate schemes outperform state of the art selection accuracy by up to 43% and present recommendations, which classification methods fit varying application requirements.

Acknowledgment

The research leading to these results has received funding from the Bavarian Ministry of Economic Affairs and Media, Energy and Technology and the European Union in the Horizon 2020 project TIMON, Grant Agreement No. 636220.

References

[1] 3GPP, "3GPP TS 23.285 V14.0.0 (2016-09) Architecture Enhancements for V2X Services," 2016.

[2] Qualcomm, "Leading the world to 5G: Cellular Vehicle-to-Everything (C-V2X) technologies," 2016. [Online]. Available: https://tinyurl.com/l6t8f5rg

[3] ETSI, "ETSI EN 302 636-4-1 V1.2.1 - Intelligent Transport Systems (ITS); Vehicular communications; GeoNetworking; Part 4: Geographical addressing and forwarding for point-to-point and point-to-multipoint communications; Sub-part 1: Media-Independent Functionality," 2014.

[4] M. Gerharz, C. de Waal, M. Frank, and P. Martini, "Link stability in mobile wireless ad hoc networks," in *27th Annual IEEE Conference on Local Computer Networks, LCN 2002*, Nov 2002, pp. 30–39.

[5] M. Heissenbüttel and T. Braun, "Optimizing neighbor table accuracy of position-based routing algorithms," in *IEEE INFOCOM*, 2005, p. 1.

[6] T. Taleb, E. Sakhaee, A. Jamalipour, K. Hashimoto, N. Kato, and Y. Nemoto, "A Stable Routing Protocol to Support ITS Services in VANET Networks," *IEEE Transactions on Vehicular Technology*, vol. 56, no. 6, pp. 3337–3347, Nov 2007.

[7] N. Sofra and K. K. Leung, "Link classification and residual time estimation through adaptive modeling for vanets," in *VTC Spring 2009 - IEEE 69th Vehicular Technology Conference*, April 2009, pp. 1–5.

[8] G. M. Hoang, B. Denis, J. Hrri, and D. T. M. Slock, "Select thy neighbors: Low complexity link selection for high precision cooperative vehicular localization," in *2015 IEEE Vehicular Networking Conference (VNC)*, Dec 2015, pp. 36–43.

[9] K. Roscher, S. Bittl, A. A. Gonzalez, M. Myrtus, and J. Jiru, "ezCar2X. Rapid-Prototyping of Communication Technologies and Cooperative ITS Applications on Real Targets and Inside Simulation Environments," in *Wireless Communication and Information. Dig. Gesellschaft*, 2014.

[10] L. Codeca, R. Frank, and T. Engel, "Luxembourg SUMO Traffic (LuST) Scenario: 24 hours of mobility for vehicular networking research," in *2015 IEEE Vehicular Networking Conference (VNC)*, 2015, pp. 1–8.

[11] ETSI, "ETSI EN 302 637-2 V1.3.2 - Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Part 2: Specification of Cooperative Awareness Basic Service," nov 2014.

[12] "ns-3 (3.25)," 2016. [Online]. Available: http://www.nsnam.org

[13] D. Krajzewicz, J. Erdmann, M. Behrisch, and L. Bieker, "Recent Development and Applications of SUMO - Simulation of Urban MObility," *International Journal On Advances in Systems and Measurements*, vol. 5, no. 3&4, pp. 128–138, dec 2012.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[15] L. Cheng, B. E. Henty, D. D. Stancil, F. Bai, and P. Mudalige, "Mobile vehicle-to-vehicle narrow-band channel measurement and characterization of the 5.9 GHz Dedicated Short Range Communication (DSRC) frequency band," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 8, pp. 1501–1516, 2007.

[16] C. Sommer, D. Eckhoff, R. German, and F. Dressler, "A computationally inexpensive empirical model of IEEE 802.11p radio shadowing in urban environments," in *8th International Conference on Wireless On-Demand Network Systems and Services, WONS 2011*, 2011, pp. 84–90.

[17] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, p. 066138, jun 2004.

[18] Tin Kam Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1. IEEE Comput. Soc. Press, 1995, pp. 278–282.

[19] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. New York, NY: Springer New York, 2009.

[20] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, oct 1986.

[21] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *3rd International Conference for Learning Representations*, dec 2015.

[22] M. Stone, "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 36, no. 2, pp. 111–147, 1974.

[23] G. W. Brier, "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, jan 1950.