An Ensemble Learning Based Approach to Position Falsification Detection in Internet of Vehicles

Goodness Oluchi Anyanwu, Cosmas Ifeanyi Nwakanma, Jae-Min Lee, and Dong-Seong Kim
Department of IT Convergence Engineering Kumoh National Institute of Technology Gumi, South Korea
(anyanwu.goodnes, cosmas.ifeanyi, ljmpaul, dskim@kumoh.ac.kr)@kumoh.ac.kr

***Abstract*—Misbehaviour detection is seen as an important development to guarantee that vehicles are certified on the IoV network. To solve this, a recent dataset known as the VeReMi dataset was created. In this paper, we presented an Optimized Ensemble Neural Network approach using MATLAB R2019b. To validate the idea in this work, three other ensemble learning algorithms were investigated to show the best performed. The result shows that the proposed optimized scheme (AdaBoostM2) outperformed other state-of-the-art algorithms as well as related works with an accuracy of 99.6%.**

***Index Terms*—IoV, Basic Safety Messages, Ensemble Learning, AdaBoost.**

I. Introduction

Internet of Vehicles (IoV) is an enhanced and scalable web-based system of smartly connected vehicles. IoV is an evolution from the conventional Vehicular Ad Hoc Networks (VANET) [1] with the aim of breeding nodes that use wireless networking technologies as a method of communication, permitting vehicles to exchange information in real time [2].

However, the number of the security threats in IoV has increased overtime as a great deal of vehicle manufacturers are beginning to pay more attention to these threats [3]. In IoV, basic safety-related messages (BSM) are at risk of several threats [4]. This can on the long run modify the safety related data features transmitted on the network resulting in a confusing and false alert warnings on the network. As illustrated in Fig. 1, the vehicle represented as $V4$ on the network sends a misleading and falsified location report represented as $V4_1$ to other vehicles on the network. Consequently, it is essential to look out for abnormal conduct nodes using artificial intelligence methods by analysing their BSMs, identify malicious behavior of vehicles deliberately disseminating inappropriate information in order to warn drivers of intending threats.

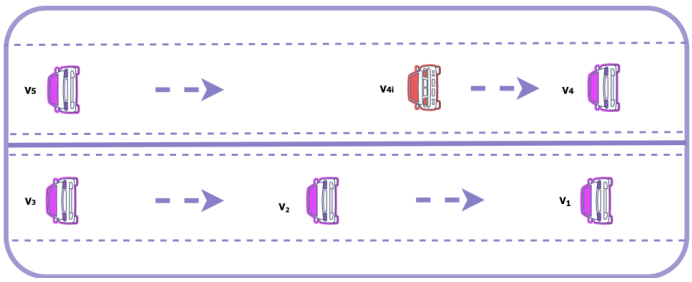

Fig. 1. Location Falsification in IoV

IoV and location falsification challenges has attracted research attention recently. For instance, in addition to a misbehavior detection model, a hash-chain algorithm was proposed in [5] to forward the right safety messages to vehicles in a well-timed manner using the VeReMi dataset. This prevents VANET from falsification attacks and delivers higher throughput with very low delay. Similarly, the authors in [6] adopted a federated machine learning approach, to ensure an enhanced and locally trained model in a distributed manner. However, the requirements for implementing federated learning was not considered. Despite the high accuracy achieved by various AI approaches, the need for combining, skipping or sparsely connecting various schemes by tuning hyper parameters has yielded promising results [7] prompting the interest in ensemble learning. The paper organisation is as follows: In section II, the proposed Artificial Intelligence methodology used in the detection of the misbehaviour in IoV was represented. In section III, result evaluation and discussion was presented while the paper was concluded in section IV.

II. System Model

A. Proposed False Location Detection System (FLDS)

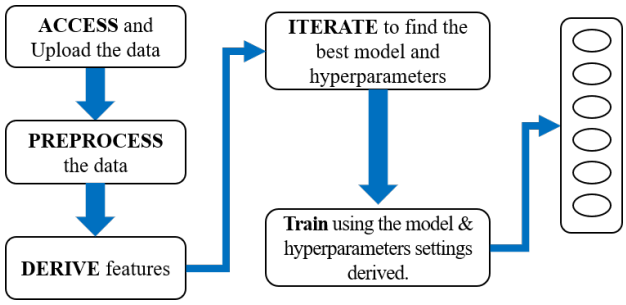

Fig. 2. Optimized False Location Detection System

In this work, the problem is considered as a multi-classification problem. Given BSM information about a vehicle (vehicle's ID, its speed, transmission time, and position e.t.c.), it is possible to predict an attack in terms of falsification of the BSM. The aim of the detection system proposed in this work is to use an optimized ensemble multi-classifier algorithm to detect and accurately classify attack types. Using machine learning techniques, ensemble learning systems have shown a proper effectiveness in the detection and classification of various kinds of attacks [8].

$$\eta = \bigcup(\chi, \gamma, \epsilon, \theta) \tag{1}$$

The equation 1 is a mathematical representation of the ensemble learning model, where: η represents the ensemble learning method, χ represent the matrix of dataset, γ represents the observations or vector of responses, ϵ specifies the aggregation methods, and θ represents various hyperparameter tuning, types of learners and method argument. while the $\bigcup$ summarizes the functional relationship between the variables.

In this work, various ensemble learning methods were investigated using MATLAB R2019b. By default, the app protects against over-fitting by applying cross-validation. The simulation algorithm parameters settings for the proposed optimized algorithm with an expected improvement per second plus acquisition function is represented in table I. The choice of ensemble models are bagging and boosting as they possess characteristics required for multi-classification.

TABLE I
SIMULATION PARAMETER FOR HYPERPARAMETERS TUNING OF ENSEMBLE LEARNING CANDIDATES

HyperParameters	Search Range
Ensemble Methods	Bag, AdaBoost, RUSBoost
No of Learners	10-500
Learning Rate	0.001-1
Maximum No of splits	1-512433
Iteration	30
Optimizer	Bayesian Optimizer

B. Dataset Description

The VeReMi dataset is used for the evaluation of misbehavior detection mechanisms in IoV. It contains $512,434$ rows and 13 predictors and was developed using a simulation platform of the Luxembourg City Vehicle Network. The dataset includes malicious messages intended to trigger incorrect application behavior [9]. The attack types modelled in the dataset are represented as follows: Constant Attacks (1), Constant Offset Attacks (2), Random Attack (4), Random Offset Attacks (8), Eventual Stop (16) and 0 which refers to normal BSM; the fact that the vehicle is not an attacker [9].

III. PERFORMANCE EVALUATION

Apart from the proposed optimized algorithm, state-of-the art ensemble networks were modeled. These include: Bagged trees, RUSBoosted, Subspace K-nearest neighbour (KNN) then a Fine KNN (number of Neighbors as 1). The performance metrics for comparison are accuracy, receiving operating characteristics (ROC) and minimum classification error (MCE). In the end, the proposed optimized algorithm out-performed other algorithms with the least classification error and an accuracy of 99.6% as shown in table II.

From Fig. 3, the proposed AdaBoost ensemble learning shows a false discovery rate of ≤ 1 for the various attack types. This means that the system possesses the capability to accurately predict values beyond 99% as desired for a typical IoV system. Similarly, Fig. 4 shows the ROC of 100% depicting that the MCE value at 553 is substantially negligible considering the volume of the dataset. Using an ensemble algorithm allows for more accommodating and better structures to enhance the performance of a system.

TABLE II
PERFORMANCE METRICS

Algorithm	Accuracy (%)	ROC (%)	Training Time (s)	MCE (#)
Ensemble Bagged Trees	99.5	100	328.5	586
Ensemble RUS Boosted	78.1	81	146.07	28012
Ensemble Subspace KNN	88.3	99	156.61	14994
Fine KNN	94.6	95	1869.8	6863
Optimizable Ensemble Ada Boost	**99.6**	**100**	**5001.8**	**553**

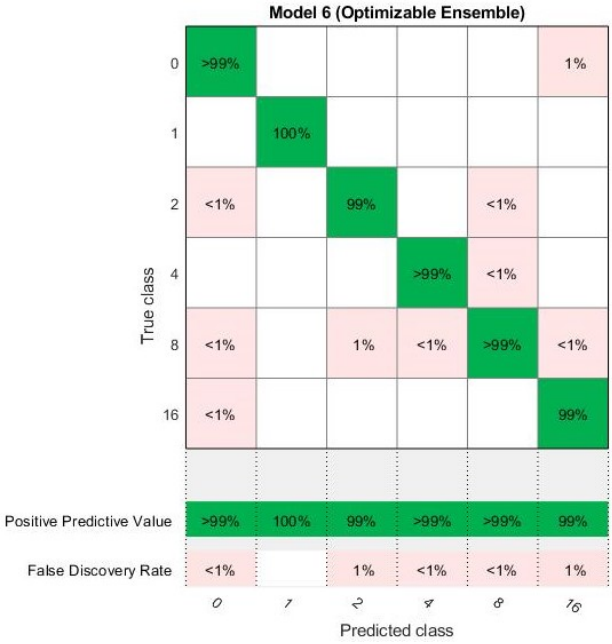

Fig. 3. Confusion matrix of the AdaBoost showing False Discovery Rate and Positive Predictive Value

Using MCE to select the best point, Fig. 5 shows that the AdaBoost Ensemble Algorithm performed optimally at the 26^{th} iteration (epoch) where the Estimated MCE is below the Observed Minimum Classification Error. The goal of ensemble is to boost the performance of the conventional algorithms.

A. Adaptive Boosting for Multi-class Classification

$$\varepsilon_t = \frac{1}{2}\sum_{i=1}^{\chi}\sum_{\kappa \neq y_i} d_{i,\kappa}^{(t)}(1 - h_t(x_i, y_i) + h_t(x_i, \kappa)) \tag{2}$$

Adaptive boosting- AdaBoostM2 is considered best for multi classification [10]. The ensemble learning app thus, selected

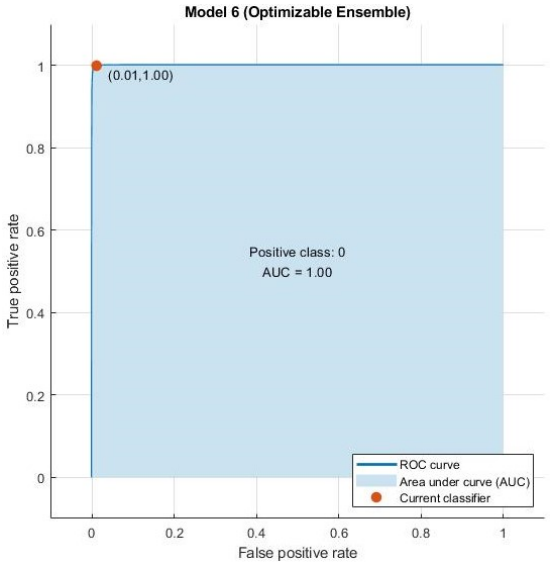

Fig. 4. Receiver Operating Characteristics Curve indicating also the Area Under Coverage of the Optimized Classifier

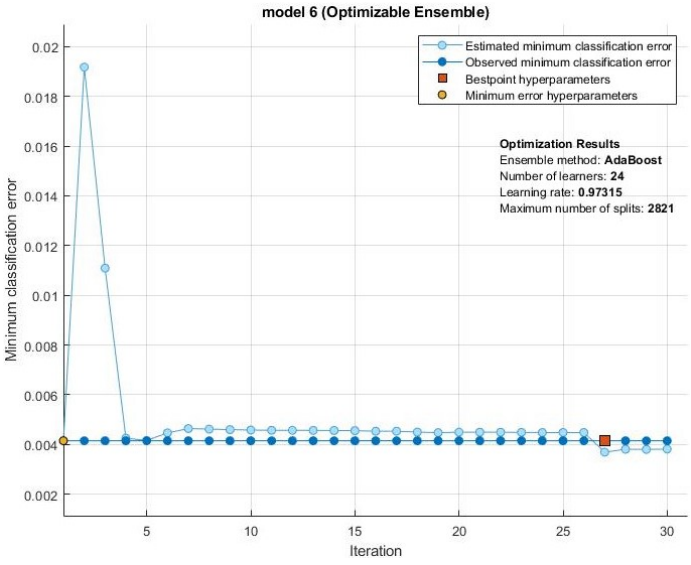

Fig. 5. Minimum Classification Error Plot

it in addition to a relatively low memory consumption. Rather than the weighted classification error, AdaBoostM2 utilizes weighted pseudo-loss for 'χ' number of observations and ψ classes as represented in equation 2, where $h_t(x_i, \kappa)$ represents the confidence of prediction by learner at step t into class κ with range from 0 (not confident) to 1 (highly confident), $d_{i,\kappa}^{(t)}$ are observation weights at step t for class κ, y_i is the true class label taking one of the ψ values. The second sum is the over-all classes other than the true class y_i.

B. Mathematical Background of Receiving Operating Characteristics

In classification problems, ROC curve is considered the best metric for choosing the best point where classification is optimized. This is the most optimized point in terms of trade off when comparing the costs of failing to detect positives against the costs of raising false alarms. These trade offs are represented by equation 3 where α = false positive (cost of false alarm) β = false negative (cost of missing a point) λ = proportion of positive cases. Thus, the average expected cost of classification at point x,y in the ROC space is

$$\psi = (1 - \lambda)\alpha_x + \rho\beta(1 - y) \tag{3}$$

IV. Conclusion

In this work, an optimized ensemble learning model is presented for securing the BSM of IoV. The focus in the future will be to further tune the parameters of the ensemble learning algorithms such that a range of 99.9% and beyond is achieved since IoV will require highest point of accuracy being a mission critical system and considering the rising interest and need for achieving optimal performances.

Acknowledgment

This research work was supported by Priority Research Centers Program through NRF funded by MEST (2018R1A6A1A03024003) and the Grand Information Technology Research Center support program (IITP-2021-2020-0-01612) supervised by the IITP by MSIT, Korea.

References

[1] T.-T. Ngo, T. Huynh-The, and D.-S. Kim, "A Novel VANETs-Based Traffic Light Scheduling Scheme for Greener Planet and Safer Road Intersections," *IEEE Access*, vol. 7, pp. 22 175–22 185, 2019.

[2] B. Ji, X. Zhang, S. Mumtaz, C. Han, C. Li, H. Wen, and D. Wang, "Survey on the Internet of Vehicles: Network Architectures and Applications," *IEEE Communications Standards Magazine*, vol. 4, no. 1, pp. 34–41, 2020.

[3] Y. Sun, L. Wu, S. Wu, S. Li, T. Zhang, L. Zhang, J. Xu, Y. Xiong, and X. Cui, "Attacks and Countermeasures in the Internet of Vehicles," *Annals of Telecommunications*, vol. 72, 11 2016.

[4] J. Liu and A. J. Khattak, "Delivering Improved Alerts, Warnings, and Control Assistance using Basic Safety Messages Transmitted between Connected Vehicles," *Transportation Research Part C: Emerging Technologies*, vol. 68, pp. 83–100, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X1630002X

[5] D. B. Rawat, B. B. Bista, and G. Yan, "Securing vehicular ad-hoc networks from data falsification attacks," in *2016 IEEE Region 10 Conference (TENCON)*, 2016, pp. 99–102.

[6] A. Uprety, D. B. Rawat, and J. Li, "Privacy Preserving Misbehavior Detection in IoV Using Federated Machine Learning," in *2021 IEEE 18th Annual Consumer Communications Networking Conference (CCNC)*, 2021, pp. 1–6.

[7] G. B. Tunze, T. Huynh-The, J.-M. Lee, and D.-S. Kim, "Sparsely Connected CNN for Efficient Automatic Modulation Recognition," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 15 557–15 568, 2020.

[8] B. Zhang, J. Yin, J. Hao, D. Zhang, and S. Wang, ""Malicious Codes Detection Based on Ensemble Learning"," in *Autonomic and Trusted Computing*, B. Xiao, L. T. Yang, J. Ma, C. Muller-Schloer, and Y. Hua, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 468–477.

[9] J. Kamel, M. Wolf, R. W. Van Der Heijden, A. Kaiser, P. Urien, and F. Kargl, "VeReMi Extension: A Dataset for Comparable Evaluation of Misbehavior Detection in VANETs," in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, 2020, pp. 1–6.

[10] J. Torres-Sospedra, C. Hernández-Espinosa, and M. Fernández-Redondo, "Adaptive Boosting: Dividing the Learning Set to Increase the Diversity and Performance of the Ensemble," in *Neural Information Processing*, I. King, J. Wang, L.-W. Chan, and D. Wang, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 688–697.