A New Combination of Machine Learning Algorithms using Stacking Approach for Misbehavior Detection in VANETs

Abhilash Sonker[1] and Dr. R K Gupta[2],

[1, 2] Department of CSE&IT, MITS, Gwalior (474020), India

Summary
Road safety, traffic efficiency and passenger comfort are main reasons for the emergence of Vehicular Adhoc Networks (VANETs). The misbehavior in the nodes can be detected with its communication to other nodes. The performance of VANET applications depend on messages and information shared by vehicles. When a message is sent from one node to another node it has some features. With the study of these features it can be found that the message sent is malicious or not. The detection of malicious activity is hence an important component. In this paper, a new combination of machine learning algorithms using stacking approach is built to find the misbehavior in the message log sent by a node in VANETs. Correspondingly, it will be detected that the message sent from the node is malicious or not. A new combination is designed with Random Forest (bagging) and Xgboost (boosting) using stacking to get more accuracy in multiclass classification of attacks. With this new combination of algorithms using stacking 98.44% of accuracy is achieved. This accuracy is evaluated on the test data. For this work, VeReMi dataset (public dataset for the malicious node detection) is used.

Key words:
Misbehavior Detection; Machine Learning; Stacking Algorithm; Vehicular Adhoc Networks

1. Introduction

Vehicular ad hoc network is play vital role for future cooperative intelligent systems [1]. Vehicles in VANETs work as a node and they communicate to each other through message logs. These message logs are short lived and have several features like sending time, sending Id, message Id, position, noise in the position, speed, noise in the speed, etc. VANETs are susceptible to the life threatening situations such as road congestions and accidents [2].

The five types of attacks which are taken into consideration for this work are constant attack, constant offset attack, random attack, random offset attack and eventual attack. The constant attacker transfers fixed, pre-configured position; the constant offset attacker transfers fixed, pre-configured offset added to their actual position; the random attacker sends a random position; the random offset attacker transfers a random position in a pre-configured rectangle around the vehicle; the eventual attacker behaves normally for some time, and then attacks by transmitting the current position repeatedly. For every message a new random sample is taken by random attacks [3].

The current work is to present a new combination of machine learning algorithms for misbehavior detection in VANETs using approach. Stacking models have been overcome the defects of using a single supervised learning method. Stacking models combine different methods to improve classification accuracy. Combined models apply the same algorithm repeatedly through partitioning and weighting of a training data set. Combined models also have been called Ensembles [4]. Ensembles are techniques that create multiple models and then combine them to produce improved results. Ensemble method usually produces more accurate solutions than a single model.

This paper proposes a new combination of classification methods to improve the classification accuracy. In ensemble learning there are different techniques such as Blending, Stacking, Bagging and Boosting. Stacking ensemble machine learning approach is used in this paper because compared to individual techniques the accuracy of this type of classification achieves the highest accuracy and it will perform better for the VeReMi dataset. In stacking multiple different learners are built and an intermediate prediction is obtained, one prediction for each learned model. Then it adds a new model which learns from the intermediate predictions the same target. The final stacked model improves the overall accuracy of the model, which is better than a single algorithm based model. One more benefit stacking approach is that you can improve a learning model with different types of models which are capable to learn some part of the problem, but not the whole space of the problem [5]. We have used VeReMi dataset for this work. The major contribution of our paper is to design a new combination of basic algorithms using stacking for the detection of misbehavior in VANETs using stacking ensemble learning approach.

The studies which are done earlier in the analysis of misbehavior in the VANETs are done with the simulations and a detailed description of few recent works which are done in the advancement of misbehavior detection is discussed below.

Manuscript received October 5, 2020
Manuscript revised October 20, 2020
https://doi.org/10.22937/IJCSNS.2020.20.10.13

Jyoti Grover et al. in 2011 presented a study on machine learning for multiple misbehavior detection in VANETs. The different algorithms that were used are Naive Bayes, IBK, AdaBoost1, J-48 RF. But Random Forest and J-48 gave the best results. Dataset was consisting of 3101 legitimate and 1427 malicious samples. Results were based on metircs with high values of TPR (0.93), TNR (0.99) and small values of FPR (0.005) and FNR (0.06) [6]. Another study by Jyoti Grover et al. in 2012 presented a concept of Misbehavior Detection Based on Ensemble Learning in VANETs. The algorithms used were Naive Bayes, IBK, AdaBoost1, J-48, RF and Ensemble based learning. Ensemble based learning gave the highest accuracy TPR (0.95), FPR (0.01) and TNR (0.99), FNR (0.03). Dataset was consisting of 3101 legitimate and 1427 malicious samples [7].

Uzma Khan et al. in 2014 presented a study on Detection of Malicious Nodes (DMN) in Vehicular Ad-Hoc Networks. They proposed a novel algorithm called DMN (Detection of Malicious Nodes in VANETs). They performed the simulation in network simulator [8]. Jaskaran Preet Singh, and Rasmeet S. Bali in 2014 presented a concept on Hybrid Backbone Based Clustering algorithm for Vehicular Ad-Hoc networks. In this paper a hybrid backbone based clustering algorithm is proposed. The proposed algorithm uses a backbone known as cluster leadership to decide upon the cluster-head. In future more parameters like distance, density and geographical location of a vehicle can be considered [9].

Alok Kumar et al in 2016 presented a study on Historical Feedback based Misbehavior Detection (HFMD) algorithm in VANET. In this paper Misbehavior Detection Algorithm and Boolean RSU_Verification algorithm these two algorithms were used [10]. S. Sharanya and S. Karthikeyan in 2017 presented a concept on Classifying Malicious Nodes in VANETs Using Support Vector Machines with Modified Fading Memory. In this paper algorithm used were Support Vector Machine (SVM) along with Modified Fading Memory (MFM) a semi-supervised learning alogrithm. SVM-MFM shows the ROC of 98% [11]. Boubakeur Achichi, et al in 2018 presented a concept on Hybrid Approach for Congestion Control in VANETs. In this paper they proposed a congestion control algorithm based on the combination of two approaches event-based and measure-based. This combination provided the better results [12].

Pranav Kumar Singh et al. in 2019 presented a concept on Machine Learning Based Approach to Detect Position Falsification Attack in VANETs. In this paper two algorithms were used SVM and Logistic Regression. SVM performed gave better accuracy than Logistic Regression [13]. K. Nirmala and S. Prasath in 2019 presented a study on Adaptive Boosting Classifier Based Attack Detection for Secured Communication in Vanet. In this paper algorithm used was an ensemble classifier Multi-Objective Reweighted Adaptive Boosting which uses the ANN as a weak learner [14].

Fuad A. Ghaleb et al in 2019 presented a concept on Ensemble-Based Hybrid Context-Aware Misbehavior Detection Model for Vehicular Ad Hoc Network. In this paper algorithms proposed were Ensemble-based Hybrid Context-Aware MDS (EHCA-MDS), Hybrid Context-Aware MDS (HCA-MDS) and Data-Centric Context-Aware MDS (DCA-MDS). Results were EHCA-MDS: 97.01%, HCA-MDS: 93.51%, DCA-MDS: 90.98% [15]. Another study by Fuad A. Ghaleb et al., in 2020 presented a concept on Misbehavior-Aware On-Demand Collaborative Intrusion Detection System Using Distributed Ensemble Learning for VANET. In this paper system proposed was misbehavior-aware on-demand collaborative intrusion detection system (MA-CIDS) based on the concept of distributed. Algorithms with F1 score MA-CIDS (RF): 0.98, MA-CIDS (SVM): 0.90, MA-CIDS (XGBoost): 0.95 [16]. Waleed Ahsan et al in 2020 presented a concept on Optimized Node Clustering in VANETs by Using Meta-Heuristic Algorithms. In this paper algorithm proposed was grasshoppers' optimization-based node clustering. The proposed algorithm reduced network overhead in unpredictable node density scenarios [17].

With the study of such all the works, it come to notice that there is no much significant information is available in the direction of study of misbehavior detection in the stacking machine learning approach in VANETs. This paper is going to provide a discussion of the stacking algorithm in the detection of misbehavior in the VANETs. The new combination is going to increase the detection of attacks than the individual accuracies.

2. Research Method

2.1 About the Dataset

In this paper, VeReMi dataset is used for the misbehavior detection. The dataset contains several files of individual attacks. Some of the files from each class are taken to make a combined dataset of all the classes. The dataset contains mainly five types of attacks with 3 kinds of densities. The normal class is represented with 0 in the target variable that is class type and rest all classes represents attacks as 1, 2, 4, 8, 16 for the type1(constant attack), type2 (constant offset attack), type4 (random attack), type8 (random offset attack) and type16 (eventual attack) attacks respectively. The total number of instances in the combined dataset is 48,575.

2.2 Proposed Research Architecture

The whole experiment is done on the basis for the detection of attacks in message logs. The message logs have five different kinds of attacks hence the multiclass classification is used to detect the attacks. The algorithms which are used in the new combination are Random Forest [18] and Xgboost [19]. The detailed architecture of the research methodology is given in figure ("**Fig 1**") below.

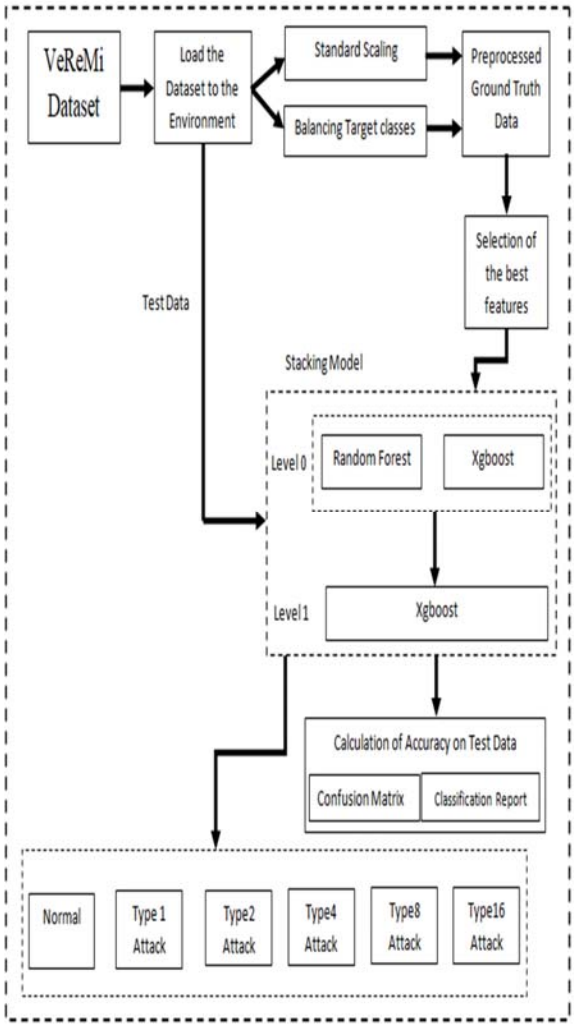

Fig. 1 Proposed Methodology for the misbehavior detection in VANETs using Stacking Machine Learning approach

The whole procedure is divided into following parts:

1. Loading the data
2. Preprocessing of Data
3. Selecting the best features
4. Fitting the stack model
5. Calculation of accuracy
6. Detection of attacks

2.2.1 Loading the data

The dataset is loaded to the programming environment and then features and target class is separated.

2.2.2 Preprocessing and Selecting best features from Data

After loading the data to the environment following main operations are performed for the preprocessing of data. From dataset all the irrelevant features are removed, standard scaling is done to normalize the values, and target variables are balanced in terms of frequency.

2.2.3 Fitting the model

For the training, a model is created using the Random Forest and Xgboost. The data is passed to both the algorithms to fit the models. The detailed description of each algorithm is discussed.

2.2.3.1 Random Forest

Random Forest is an ensemble technique work on the principle of bagging. Bagging is an ensemble technique. Bagging is short of bootstrap aggregation. The base algorithm used is decision tree. The decision trees are built using the information gain method as shown in Eq. 1 and Eq. 2.

For calculating the entropy of the sample data

$$E(s) = \sum -p_i \log_2 p_i \quad (1)$$

After calculating the entropy, the information gain is calculated for each attribute to get decide the decision node.

$$Gain(S, A) = Entropy(S) - \sum_{v \in values(A)} |S_v| / |S| Entropy(S_v) \quad (2)$$

In the first step, bootstrapping of the different decision trees is done and then aggregation of the algorithms is done using majority voting. It is a parallel computing model on the different base models and then aggregates them to predict a highly accurate result [20]. The Random Forest description is given in figure ("**Fig 2**") below.

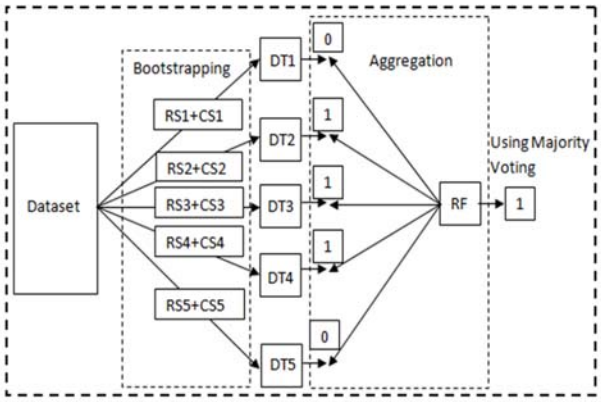

Fig. 2 Working of Random Forest

In bootstrapping the dataset is selected by row sampling and column sampling with replacement, then the data are passed to the different base models and predictions are obtained from the base model. These predictions are then aggregated using majority voting method to predict the final class [21].

2.2.3.2 Xgboost

Xgboost is an ensemble technique uses the tree as the base model [22]. It works as a scalable machine learning system for tree boosting. The reason that Xgboost is widely used in many problems is that it has beauty of scalability and it offer parallel and distributed computing to give memory efficient use [23].

Xgboost works on the sequential model of learning. The base models get trained one after the other and then combined to predict the result. The first base model created on the dataset produces some error and the instances which produces error is passed to next sequential model to minimize the error and the system continues till no error is received. This is done in gradient boosting. In Xtreme Gradient Boosting (Xgboost) reducing the error, regularization, auto pruning of the tree in base models and convergence of base models is added [24].

2.2.3.3 Stacking

Stacking or Stacked Generalization is an ensemble technique. It uses meta data computed by meta learning from the two or more base learning algorithms. The benefit of stacking is that it harnesses the capabilities of individual classifiers and makes predictions better than individual classifiers. It involves combining the predictions of multiple algorithms on single dataset. The architecture of stacking classifier includes two or more classifiers on base level 0 and a meta model to combine the predictions on level 1.

Level 0 Models (Base Models): The base models fit on the training data whose predictions are compiled.

Level 1 Model (Meta Models): The meta models best combine the predictions of base models. The meta model is trained on the predictions made by base models on out of sample data [25].

Out of fold predictions are used as the basis for training dataset for the meta model. The training data for the meta model may also include the inputs to the base models. Once the dataset is created, the meta model is trained in isolation. Stacking model is appropriate when different models with different skills are combined. In this paper, two different ensemble techniques with different skills are combined. One base learning classifier is of bagging ensemble learning and other is a kind boosting ensemble learning. Hence, this paper combines the bagging and boosting technique using stacking technique. The improvement of accuracy in the stacking method is not definite in every case. To achieve an improvement in the final predictions depend on the complexity of the problems and sufficient representation of the training data so that there is more learning by combination. The improvement also depends on the choice of base models and whether they are skillful and sufficiently uncorrelated in their results [26].

The stacking classifier used in this work is mentioned in the figure ("**Fig 3**").

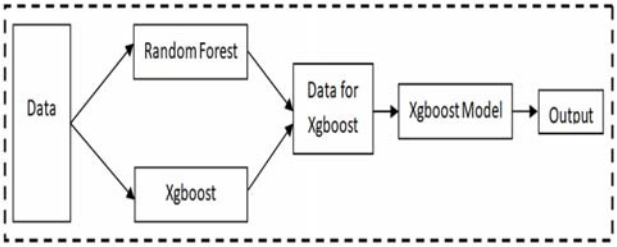

Fig. 3 Stacking algorithm

2.2.3.4 Calculation of accuracy

The calculation of the accuracy of the model is done with the confusion matrix. The confusion matrix represents the relation between the actual value and predicted value [27]. The accuracy is calculated using (3) from the confusion matrix.

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FN)+(FP+TN)} \quad (3)$$

Where TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

The classification report is also given for algorithm containing the precision, recall, f1 score and support. The formula for the different results in the confusion matrix is given by Eq. 4, 5, and 6.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (4)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (5)$$

$$\text{F1 score} = \frac{2*(Precision*Recall)}{(Precision+Recall)} \quad (6)$$

ROC curve is another method of evaluation of accuracy. Since, the work is done in the multi classification hence

ROC curve is not possible. The ROC curve although can be drawn for the multi label classification.

2.2.4 Detection of Attacks

After calculating the accuracy of model when satisfactory results are obtained, it is said that now model is ready to test on the unknown data. So whenever an unknown message log is communicated between the nodes the model based on the features is able to classify the attacks. Once when the attacks are identified corresponding node id also be detected and hence the detection of malicious nodes is done.

3. Results and Discussion

The whole experiment is performed based on the steps in the research procedure. After loading the data to the environment standard scaling is performed across the standard deviation and variance. Then the target variable is balanced in terms of frequency. Number of records for each type of attack was different before balancing and this frequency was huge. This is shown in ("**Fig 4**").

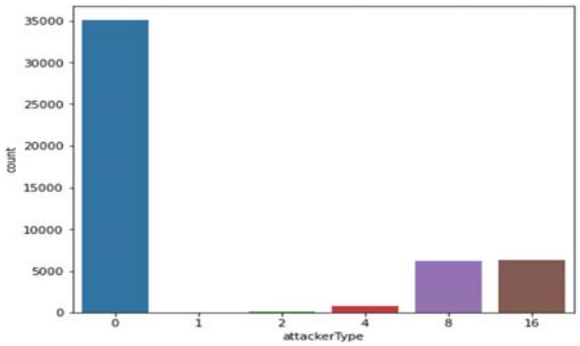

Fig. 4 Attacker type frequency before balancing dataset

If the same dataset is considered, model may become biased to give the output for the majority. For converting the imbalance dataset to balance dataset imblearn library is used. After conversion of the dataset, frequency of each attack is shown in ("**Fig. 5**") below.

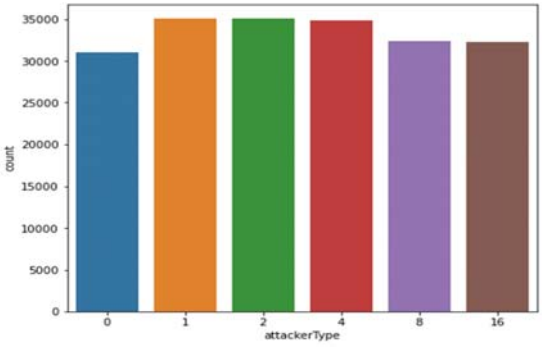

Fig. 5 Attacker type frequency after balancing dataset

After conversion of data, ensemble technique stacking is used for creation of new combination algorithm. Random Forest and Xgboost stacked with Xgboost is the new combination of bagging and boosting along with boosting to create model. Random Forest and Xgboost are used as the base algorithms. Pipeline is used for automating the work flow of the Random Forest and Xgboost. After creating the model the data is fitted to the model and accuracy is calculated.

On evaluating the performance of algorithm the accuracy obtained was 98.44%. The accuracy was calculated using confusion matrix shown in ("**Fig. 6**").

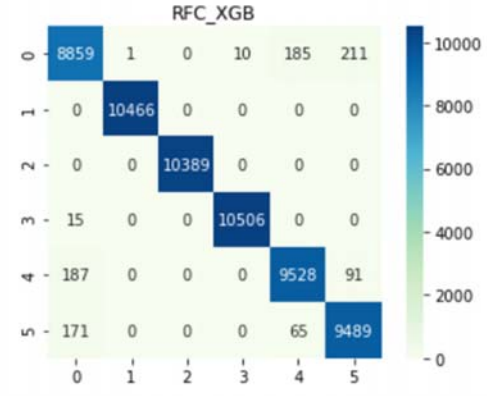

Fig. 6 Confusion Matrix for the combined stacking algorithm

The accuracy is calculated using Eq. 3:

$$\text{Accuracy} = \frac{(8859 + 10466 + 10389 + 10506 + 9528 + 9489)}{(9266 + 10466 + 10389 + 10521 + 9806 + 9725)}$$

$$\text{Accuracy} = \frac{59237}{60173} = 0.9844$$

Hence, the accuracy calculated is 98.44%.

The classification report is shown in ("**Fig 7**") below containing the precision, recall, f1-score and support. The classification report gives the detailed analysis of the algorithm used.

RFC_XG

	precision	recall	f1-score	support
0	0.96	0.96	0.96	9266
1	1.00	1.00	1.00	10466
2	1.00	1.00	1.00	10389
4	1.00	1.00	1.00	10521
8	0.97	0.97	0.97	9806
16	0.97	0.98	0.97	9725
accuracy			0.98	60173
macro avg	0.98	0.98	0.98	60173
weighted avg	0.98	0.98	0.98	60173

Fig. 7 Classification Report for the combined stacking algorithm

The support represents the actual number of class in the specified dataset. The support does not change during

the training of the models instead it remains same and helps in the evaluation process of the model. ("**Fig. 8**") shows the different support levels for each class type attacks.

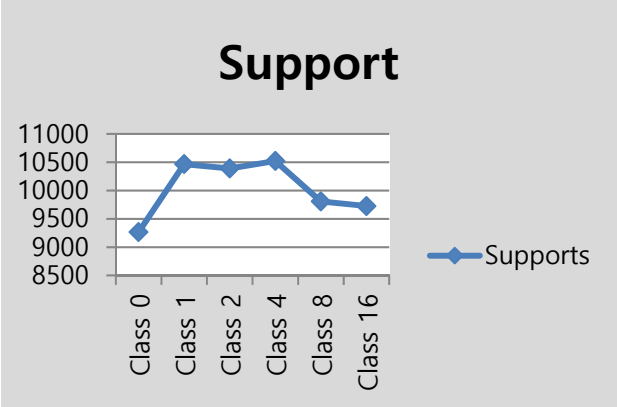

Fig. 8 Support

The ability of a classifier not to label an instance positive that is actually negative is called as precision. It tells the correctness of the classification of positive or negative class. The ability of a classifier to find all the instances is known as recall. ("**Fig. 9**") represents the values of precision and recall on different class labels.

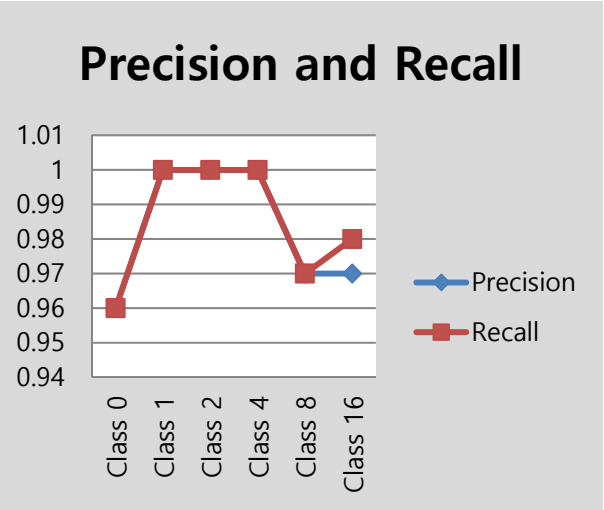

Fig. 9 Precision and Recall.

4. Conclusion

The paper is written with the aim of creating a new combination of algorithms using stacking to achieve a better accuracy in case of misbehavior detection in VANETs. This new combination is achieved with mixing the two different types of ensemble techniques that is bagging and boosting. The bagging algorithm used is Random Forest and boosting algorithm used is Xgboost. The new combination is able to achieve an accuracy of 98.44%.

The misbehavior detection has a very wide field of study. In this paper, a little portion is discussed. Considering the other different types of attacks and areas of intrusions a better study can be proposed. The algorithms can be chosen with the combination of optimization of the hyper parameters of the algorithms to get more accurate results.

Refrences

[1] Wenshuang Liang, Zhourang Li, Hongyang Zhamg, Shenling Wang, and Rongfang Bie, "Vehicular Adhoc Networks: Architectures, Research Issues, Methodologies, Challenges and Trends", International Journal of Distributed Sensor Networks, Volume 11, Issue 8, 2015

[2] Muhammad Sameer Seikh, Jun Liang and Wensong Wang, "Security and Privacy in Vehicular Adhoc Networks and Vehicle Cloud Computing: A Survey", Wireless Communications and Mobile Computing, Volume 2020, article id 5129620

[3] Rens W. van der Heijden, Thomas Lukaseder, and Frank Kargl, "VeReMi: A dataset for Comparable Evaluation of Misbehavior detection in VANETs", International Conference on Security and Privacy in Communication Systems, SecureComm 2018, pp 318-337

[4] Vijay Kotu, Bala Deshpande, "Chapter 2- Data Mining Process", Predictive Analytics and Data Mining, Pages 17-36, 2015

[5] Issam Mahmoudi, Joseph Kamel, Ines Ben-Jemaa, Arnaud Kaiser, Pascal Urien, "Towards a Reliable Machine Learning Based Global Misbehavior Detection in C-ITS: Model Evaluation Approach", International Workshop on Vehicular Adhoc Networks for Smart Cities (IWVSC'2019)

[6] Grover, J., Prajapati, N., K., Vijay, L., Singh, G., M.(2011). "Machine Learning Approach for Multiple Misbehavior Detection in VANET". Conference Paper of Communications in Computer and Information Science (CCIS). https://doi.org/10.1007/978-3-642-22720-2_68

[7] Grover, J., Vijay, L., Singh, G., M.(2012). "Misbehavior Detection Based on Ensemble Learning in VANET". International Conference on Advanced Computing, Networking and Security (ADCONS 2011). https://doi.org/10.1007/978-3-642-29280-4_70

[8] Khan, U., Agrawal, S., Silakari., S.(2014). "Detection of Malicious Nodes (DMN) in Vehicular Ad-Hoc Networks". International Conference on Information and Communication Technologies (ICICT). https://doi.org/10.1016/j.procs.2015.01.006

[9] Singh, J.P., Balib, R., S.(2014). "Hybrid Backbone Based Clustering algorithm for Vehicular Ad-Hoc networks". International Conference on Information and Communication Technologies (ICICT 2014). doi: 10.1016/j.procs.2015.01.011

[10] Kumar, A., Singh, J.R., Singh, D., Dewang, R.k.(2016). "A Historical Feedback based Misbehavior Detection (HFMD) algorithm in VANET". International Conference on Computational Intelligence and Networks(2016). DOI: 10.1109/CINE.2016.11

[11] S, Sharanya., S, Karthikeyan.(2017). "Classifying Malicious Nodes In Vanets Using Support Vector Machines With Modified Fading Memory". ARPN Journal of Engineering and Applied Sciences. https://www.researchgate.net/publication/313606417

[12] Achichi, B., Semchendine, F., Derdouri, L.(2018). "Hybrid Approach for Congestion Control in VANETs". International Conference on Software Engineering and New Technologies(ICSENT 2018). https://doi.org/10.1145/3330089.3330100

[13] Singh, P.K., Gupta, S., V, Vashistha, R., Nandi, S.k., Nandi, S.(2019). "Machine Learning Based Approach to Detect Position Falsification Attack in VANETs". ISEA-ISAP 2018, CCIS 939, pp. 166–178, 2019. https://doi.org/10.1007/978-981-13-7561-3_13

[14] K, Nirmala., S, Prasath.(2019). "Adaptive Boosting Classifier Based Attack Detection For Secured Communication In Vanet". International Journal of Advanced Science and Technology.

[15] Ghaleb, F.A., Mohd, A.M., Zainal, A., Bander, A.S., Alsaeedi, A., Boulila, W.(2019). "Ensemble-Based Hybrid Context-Aware Misbehavior Detection Model for Vehicular Ad Hoc Network". https://doi.org/10.3390/rs11232852

[16] Ghaleb, F.A., Saeed, F., Mohammad, A.S., Bander A.S., Boulila, W., A.E.M., Eljialy., Aloufi, K., Alaza. "Misbehavior-Aware On-Demand Collaborative Intrusion Detection System Using Distributed Ensemble Learning for VANET". DOI: 10.3390/electronics9091411

[17] Ahsan, W., Khan, M.F., Farhan, A., Maqsood, M., Staish, A., Nam, Y., Rho, S.(2020). "Optimized Node Clustering in VANETs by Using Meta-Heuristic Algorithms". doi:10.3390/electronics9030394

[18] Jehad Ali, Rehanullah khan, Nasir Ahmad, and Imran Maqsood, "Random Forests and Decision Trees", IJCSI International Journal of Computer Science Issues, Vol. 9, issue 5, No. 3, September 2012

[19] Tianqi Chen and Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System", arXivLabs: experimental projects with community collaborators

[20] Anne- Laure Boulesteix, Silke janitza, Jochen Kruppa, and Inke R. Konig, "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics", WIREs Data Mining and Knowledge Discovery, Volume 2, Issue 6

[21] Andy Liaw and Mattew Wiener, "Classification and Regression by Random Forest", R news 2.3 (2002): 18-22.

[22] Sahin, Emrehan Kutlug. "Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest." *SN Applied Sciences* 2.7 (2020): 1-17

[23] Rajliwall, Nitten S., Rachel Davey, and Girija Chetty. "Cardiovascular Risk Prediction Based on XGBoost." *2018 5th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE).* IEEE, 2018.

[24] Xia, Yufei, et al. "A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring." *Expert Systems with Applications* 78 (2017): 225-241.

[25] Khraisat, Ansam, et al. "Hybrid intrusion detection system based on the stacking ensemble of c5 decision tree classifier and one class support vector machine." *Electronics* 9.1 (2020): 173.

[26] Knutti, Reto, et al. "Challenges in combining projections from multiple climate models." *Journal of Climate* 23.10 (2010): 2739-2758

Abhilash Sonker received his Bachelor degree (B.E.) from SATI, Vidisha in 2006,Master degree (M.Tech) from MANIT,Bhopal in 2009. He is currently Assistant Professor in Information Technology Department at Madhav Institute of Technology & Science, Gwalior, India. His current research interests include Mobile Adhoc Ne twork and Network Security.

Dr R.K Gupta received his Ph.D. Degree from IIITM Gwalior. He is currently Professor in Computer Science Engineering Department at Madhav Institute of Technology & Science, Gwalior, India. His current research interests include Data Mining, Image Processing & Mobile Adhoc Network.