

Misbehavior Detection in VANETs Using Tree-Based Machine Learning Techniques

Dariush Ebrahimi[1(✉)], Shreya Bandyopadhyay[2], Ankit Kundal[2], and Fadi Alzhouri[3]

[1] Department of Computer Science and Physics, Wilfrid Laurier University, 75 University Ave W, Waterloo, ON N2L 3C5, Canada
debrahimi@wlu.ca

[2] Department of Computer Science, Lakehead University, Thunder Bay, ON, Canada
{sbandyo1,akundal}@lakeheadu.ca

[3] Department of Electrical and Computer Engineering, Gulf University for Science and Technology (GUST), Mishref, Kuwait
alzhouri.f@gust.edu.kw

Abstract. The Internet of Things (IoT) is rapidly becoming an integral part of our lives. Concurrently, the new dawn of Artificial Intelligence is also witnessed by the world, hence integrating these IoT devices with the power of Machine Learning (ML) aims to elevate our standard of living in a better way. Intelligent Transport System (ITS) is one of such IoT technologies that uses modern information and telecommunications technology to provide smart transportation systems. One of the sectors of ITS is Vehicular Ad-hoc Networks (VANETs), constituting vehicles equipped with wireless devices. One of the critical messages in these networks is broadcast-authenticated periodic one-hop beacons, which are used for safety and traffic efficiency applications, like traffic jam detection and collision avoidance. In such a network environment, it is often habitual for cyber attackers to conduct intrusionsy. For example, Place spoofing, which is an intentional location deception in a vehicle-to-everything (V2X) communication, can cause traffic congestion and other safety hazardsy. The main objective of this paper is to trace such misbehavior conducted by attackers with the help of lightweight ML tree-based models, such as Decision Tree, Random Forest, and Extra Tree Ensemble algorithms, and compare the models to analyze the best approach in terms of computational cost and complexit. This will be achieved by utilizing the Dedicated Short Range Communications (DSRC) Data Set.

Keywords: Vehicular ad hoc networks (VANETs) · intelligent transport systems (ITS) · dedicated short range communications (DSRC) · machine learning · intrusion detection system · internet of things (IoT) security breach

1 Introduction

The IoT is an emerging technology that brings significant changes in various industries such as government, retail, and Wholesale, healthcare, etc. Intercon-

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16768, pp. 70–88, 2027.
https://doi.org/10.1007/978-3-032-30488-9_5

nection of devices over the Internet provides people with immense comfort, such as smart transportation, reducing traffic jams, home automation, e-health, etc. [1,2]. One of the fundamental issues posing a serious threat to the IoT networks is a security breach. For example, remote adversaries might compromise Implantable Medical Devices (IMDs) for patients or smart vehicles, which not only cause tremendous economic damage, but are also life-threatening [3]. Due to the connectivity of IoT devices, remote attacks are feasible, and intruders can easily initiate attacks on a massive scale. In addition to the simplicity of executing efficient attacks, the magnitude and effect of the aftermath provide the attackers with ample chances to initiate further attacks.

In the automobile sector, via the use of Vehicular Ad-Hoc Networks (VANETs), developments have been made towards incorporating autonomous driving vehicles into Intelligent Transport Systems (ITS). VANET architecture consists of different components, such as On-Board Units (OBUs) and Road-Side Units (RSUs) . This type of network intends to promote reliable communication between vehicles (V2V communication) or between vehicles and infrastructure (V2I communication) to enhance the safety of the driver, prevent traffic congestion, and effectively manage transportation networks [4]. VANETs have the ability to greatly increase transportation quality. VANETs enable vehicles to connect with one another and exchange information such as vehicle dynamics and driving intentions, allowing for autonomous driving management. V2I links vehicles to infrastructure networks and even the Internet, allowing for real-time data distribution and traffic administration. Advances in data analysis and control algorithms have made the transportation network more efficient, where vehicles in the VANET can adapt and organise their driving behaviours, such as driving routes and speed [5]. Vehicle Platooning is an application of VANETs that enables vehicles to pursue a leading vehicle very closely by receiving acceleration and steering details via wireless communication [6]. One of the main safety application fields of VANETs is the traffic information system, which offers the most up-to-date obstacle data to a vehicle's satellite navigation system [7]. In the Electronic Brake Warning (EBW) feature, a vehicle sends out an event-based message to alert other vehicles about its braking operations [8]. This allows the driver to react timely manner to the other vehicles braking, even when the view of the brake lights is blocked by other vehicles.

According to the World Health Organization (WHO) [9], every year, nearly 1.35 million people's lives are cut short as a result of a car accident. Non-fatal accidents affect between 20 and 50 million more people, with many of them resulting in disability as a result of their injury. Individuals, their communities, and nations as a whole suffer significant economic damage as a result of road traffic accidents. Many countries lose 3% of their gross domestic product due to traffic accidents. Therefore, VANETs have a great impact on traffic management and avoiding road accidents. However, this path towards restricted human interference in vehicular communication implies that VANETs are vulnerable to a series of attacks, such as incorporating false information, manipulating and replicating messages, etc. [10,11]. This paper focuses on a few such attacks, par-

ticularly the jamming and spoofing attacks. These are the worst degree of Denial of Service (DOS) attacks. In these attacks, the intruder blocks the communication link, and therefore, other devices are unable to connect to the network. There are two possible cases in these types of attacks. In the first possibility, in a specific domain, high frequencies sent by an intruder jam the communication between nodes. Consequently, nodes in that network are not able to send or receive messages. Therefore, services are not accessible in that specific network. It is only possible to transmit or receive messages when a node leaves the target domain. Jamming the communication channel between the nodes and the RSU is the second scenario of attack. The scenario is that, near the RSU, the intruder executes an attack to disrupt the network, causing the network to break down. Nodes and RSUs are also noteworthy for sending or receiving messages from each other, contributing to network inaccessibility [12]. With the assistance of a tree based Machine Learning (ML) models such as Decision Trees (DT), Random Forest (RF), and Extra Tree Ensemble (ETE), this paper aims to track misconduct by attackers and compare the models to analyze the best approach in terms of computational cost and complexity. This paper attempts to leverage a comprehensive analysis of the problem and explore findings and insights to address the problem using lightweight ML tree-based models: DT, RF, and ETE. The contributions of the paper are summarized as follows:

- Exploring the findings of DT, FR, and ETE algorithms. This contribution will be made through devising the results based on two levels: hyperparameters and performance metrics.
- In pursuit of finding the best decision to maintain high-security ANETs, empirical discussions are produced to find the best algorithm in terms of computational cost and complexity.
- Providing significant insights to create more secure VANET systems and reduce potential cyber-attacks. Considering manuscript results increases VANET security.

The rest of this paper is organized as follows: Sect. 2 reviews VANETs' security literature. Section 3 illustrates the system model and defines the problem. Section 4 presents all the proposed Tree-based ML algorithms for misbehaviour detection. Section 5 reports detailed experimental results and analysis. Section 6 concludes the paper and highlights future works.

2 Literature Review

In VANETs, malicious nodes may disrupt the network performance. For example, modifying or inserting fake information in the network, which could incur life-threatening accidents. The security of VANETs is a highly important issue that has been the focus of research during the last few decades. One of the most important challenges in this regard is the existence of rogue vehicles. Several approaches have been reported in the literature to tackle this issue in VANETs. The authors of [13] and [14] have used features like Centroid Localization and

the bloom filter data structure to detect attacks and examine jammers. The authors of [15] have proposed a data mining technique for a real-time detection of jamming attacks in V2V communication for platoons of vehicles. In [11], the authors have used two layers of metric combination, which comprises the physical layer and the application layer, in which the relative speed between the sender and receiver is measured. Physical layer provides the Signal to Interference plus noise Ratio (SINR), Received Signal Strength Indicator (RSSI), and PDR for detecting the jamming attacks. Also, the application layer helps in extracting Global Positioning System (GPS) coordinates for location tracking.

Many studies have provided several approaches to address the security threats in VANETS. These approaches are generally divided into three categories: 1) Confidence or reputation-based system, 2) data-centric detection system for misbehaviour, and 3) Intrusion Detection System (IDS). In [16–18], a decentralized confidence-based infrastructure scheme has been implemented. The authors have built a reputation management mechanism for each vehicle that facilitates rapid adjustment to changing local circumstances and establishes confidence relationships with other vehicles. A centralized framework (an attack-resistant confidence management scheme) has been proposed by [19], which not only identifies and deals with cyber attacks, but also determines the reliability of information and vehicles. The method of data-centred misbehaviour detection has been suggested in [20,21]. It has been used to enhance the reliability of VANETs in shared data.

In [11], the effect of the density of vehicles on network efficiency is defined in order to achieve accurate identification, followed by analysis of jamming effectiveness, based on various mobility patterns and attitudes. In [6], the authors introduced a two-step intrusion detection algorithm that detects both spoofing and jamming attacks in a Connected Autonomous Vehicle (CAV) environment. The proposed method reduces the risk of cyber attacks that cause traffic disruptions and accidents. The identification mechanism of the proposed IDS is centered on the RF, k-Nearest Neighbour (KNN), and One-Class Support Vector Machine (OCSVM) approaches, as well as cross-layer data fusion methods. Specifically, the paper has focused on analysing the effect of jamming based on the behaviour of jammers and mobility trends. The effect of the distribution of vehicles on network efficiency is first established in order to accomplish trustworthy identification. The efficacy of jamming is then analysed by the following various mobility patterns (fixed, irregular, or attacking) and behaviours (steady, arbitrary, and responsive). Their approach demonstrates promising results in IEEE802.11p networks to reliably detect various types of jammers.

It is worth mentioning in this context to recall the debate about the Federated Learning (FL) approach. Recently, many new applications have been seen utilizing FL to improve the quality and privacy of data-driven applications, especially in IoT applications. The unbalanced distribution of samples across local nodes in FL poses a critical issue in terms of varying performance among the local nodes. For more information about the FL, we direct the reader to [22]. It may be unjust to discuss security matters in any area without any reference to the

most widely known technique, namely, blockchain technology. The technology's popularity has soared with the advent of Bitcoin. In building a secure system, it's important to consider security at all layers of the technology, for example, secure communication, data privacy, and whether it is a public or private system. In fact, we can implement blockchain in certain layers to strengthen security, but this technique works well in a high density of clients and data. While our environment assumes hundreds of connected vehicles in a given area, which is very small compared to other applications. For more information about the pros and cons of blockchain security, we forward readers to read [23–25].

The authors in [26] adopted machine learning techniques, Random Forest, on the VeReMi dataset to detect the misbehavior. However, our paper poses tree-based ML algorithms like DT, RF, and ETE approaches to create an IDS. To the best of our knowledge, the ETE approach has not been used in this regard. Although it is better than the other two algorithms in terms of computational cost and complexity. The study compares the approaches and explains how the ETE approach surpasses the other approaches and will give better results.

3 System Model and Problem Definition

An IDS is an integral part of a defense mechanism that helps in detecting attacks or proof of unusual activity in a vehicular communication system. In the network security and intrusion prevention domains, various ML techniques have been extensively used. An ML-IDS is focused on models that help to classify gathered information. The use of ML techniques, especially the tree-based approaches, has shown precision of IDS in the field of network security. In this paper, three tree-based ML methods are collectively compared and tried to determine the best of them. In Fig. 1, a visual representation of the flow of the proposed system is created. When these algorithms are applied independently, each of the proposed classification algorithms provides consistent results. The mixture of these algorithms will therefore lead to an overall efficient IDS. Algorithms such as RF, ETE, and DT are assessed to determine attacks in communication. VANETs use vehicles as mobile nodes to communicate between neighboring vehicles and between vehicles and nearby roadside units; however, due to their characteristics, they tend to be different from other networks. Primarily, while traveling, the nodes(vehicles) are restricted to road topology; therefore, if the road information is available, a vehicle's future location can be predicted. In addition, vehicles can manage critical computation, communication, and sensing capabilities and have persistent transmission power to support these functions independently. Nonetheless, VANETs have some challenging features as well, such as being potentially large-scale networks and having high mobility. As shown in Table 1, VANETs consist of various components such as On Board Units (OBUs), Road Side Units (RSUs), Application Units (AUs), and Trusted Authorities (TAs) [27].

Communications in VANETs are concentrated on Vehicle-to-Everything (V2X), which plays a key role in the ITS by providing actual and exceptionally accurate information such as accident alerts, road bottlenecks information,

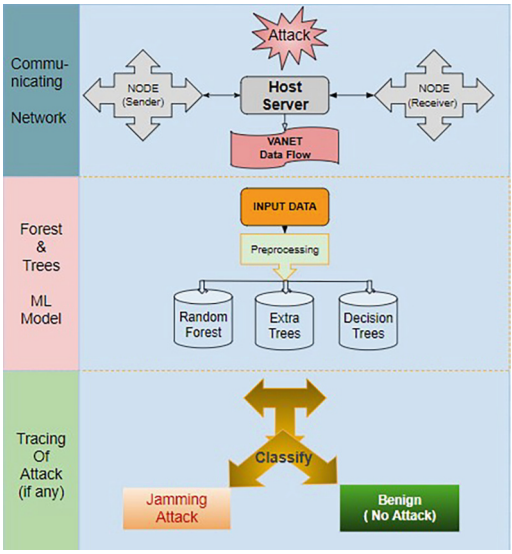

Fig. 1. System Flow Diagram

traffic congestion warnings, emergencies, and other transportation services to enhance traffic quality, traffic protection, and driving experiences. Some of specific important terms in V2X communication are [28]:

- *V2V Group Communications:* Only vehicles that share some of the same characteristics shall engage in this communication, such as vehicles with the same brand or vehicles sharing the same location over time.
- *Vehicle Beacons:* In this method, the vehicle regularly sends beacon messages to all nearby vehicles and RSUs. These messages provide the speed, coordinates, velocity, and acceleration of being sent.
- *Vehicle Warning Infrastructure:* In order to improve the flow of traffic and road safety, warning signals are transmitted from the infrastructure through RSUs to all vehicles in the vicinity of an area, where potential accidents or collisions are detected, in particular at curves, intersections, or narrow roads.

3.1 Attacks in Focus

Our model addresses two kinds of attacks: Jamming and Spoofing attacks.

Spoofing Attacks: To warn neighbouring vehicles of their presence, all the vehicles in a platoon broadcast beacon messages. Each beacon message consists of various fields such as Time Instance, Identifier Node, Mac Address, and Vehicle Current Position. The intruder will change any of these. In order to transmit false GPS position coordinates inside the cameras or sensors, the intruder exploits these fields, misdirecting the platoon of linked vehicles to an incorrect spot. As

Table 1. VANET Components

Components	Description	Function
On Board Units (OBU)	A vehicle-fixed network system that is used to share data with nearby vehicles, OBUs, or RSUs	V2V and V2I communication
Road Side Units (RSU)	Base Station	Connects all vehicular nodes to the Internet and allows V2V and V2I communication
Trusted Authorities (TA)	Law enforcement bodies	It is accountable for the issuing of digital certificates to RSUs and vehicles that uniquely validate the vehicle and RSU in the ad hoc vehicle network
Application Units (AU)	Appliance mounted on vehicles	It deals more with security applications and personal digital assistance (PDA) services offered by the provider

a consequence of wireless communication between transmitter (Tx) and receiver (Rx), this has to shift to a different stage (to the observed RSSI values [6]) to indicate the spoofing attack. The level of the RSSI values changes drastically when the location of the transmitter, which is the spoofer during the spoofing attack, is somewhat different from the legitimate position. The spoofing attack can be indicated by this fact, also demonstrating that the RSSI can be a critical metric for detecting a spoofing attack using an ML approach.

Jamming Attacks: A jammer is a system that does not comply with valid MAC protocols and attempts to interfere with the transmitter or receiver to disrupt communications. Various jammers can be categorized as shown in Table 2.

3.2 System Model

The System model consists of a troop of vehicles fitted with embedded Cooperative Adaptive Cruise Control (CACC) technology, which is a more sophisticated version of Adaptive Cruise Control (ACC). Vehicles in CACC systems can interact with the vehicle in front of them using V2V communication. We have considered the concept of vehicle platooning (or road trains), which is basically a collection of vehicles or nodes in which one leading vehicle, which is manually driven, has following vehicles that are automatically controlled by the leading vehicle. Figure 2 depicts four vehicles moving on a highway. We assume that the first two vehicles, V1 and V2, are forming a platoon, with V1 being the leading vehicle followed by V2, and consequently, they can share information by using V2V communication, and the other two vehicles, V3 and V4, are out of the platoon. Let us also assume that V1 and V2 have local sensors and are installed with computing devices in order to avoid longitudinal instability while forming a platoon. Connected nodes will broadcast beacon or basic safety messages (BSMs)

Table 2. Jamming Techniques

Jamming Techniques	Properties	Merits from the standpoint of the intruder
Constant	It generates a signal made up of random bits. No MAC protocol rules are observed by the signal generator	It will still keep the media occupied
Deceptive	It generates packets where the header is useful, but the payload is of no use	Legitimate nodes will sense a busy channel and hence withdraw themselves due to permanently being in the listening mode
Random	There are two modes: 1) Jammer keeps the channel busy at random time periods; 2) Jammer turns off its transmitter for a random time period	
Reactive	Targets the receiver. It uses enough noise to change as many bits as possible, because the modification of the bits requires only a minimal amount of power. If a checksum is executed at the receiving end of a packet, it is listed as invalid and discarded	It jams communications when it detects one, so that it does not waste energy

over the 5.9 GHz Dedicated Short Range Communication channel (around one kilometer range). These messages contain information such as GPS coordinates, location, speed, velocity, and warnings related to accidents or collisions, in addition to information related to vehicles' route changes. The intruder, which is not a part of the platoon, would use vehicles out of the platoon, like vehicles V3 and V4, to execute jamming and spoofing attacks, respectively (see Fig. 2). The attacker may jam the communication by interfering with any BSMs longitudinally and block communication between vehicle V1 and V2 by transmitting in a reactive manner. The vehicles forming the platoon, i.e., V1 and V2, are connected to the Remote Cloud and Management Center (RCMC) that manages all the real-time statistical data containing values like transmission power, received power, PDR, RSSI, and so on, which have also been collected from other nodes in the area (such as V3 and V4). Each node is directly connected to the internet through LTE technology for sending real-time statistical data to a cloud or server. The RCMC, by using the statistical data received from all nodes, can detect whether there is a jamming attack on nodes in the platoon or not. On the other hand, the RCMC can also detect a also beenspoofing attack by using and monitoring the BSM data sent in V2V communication. Hence, this data can be used for misbehavior detection.

As shown in Fig. 2, the data flow is managed by the cloud when a sender node/vehicle is sending data to a receiver node/vehicle via a host server (i.e., when vehicles are connected in a platoon through V2V communication). The

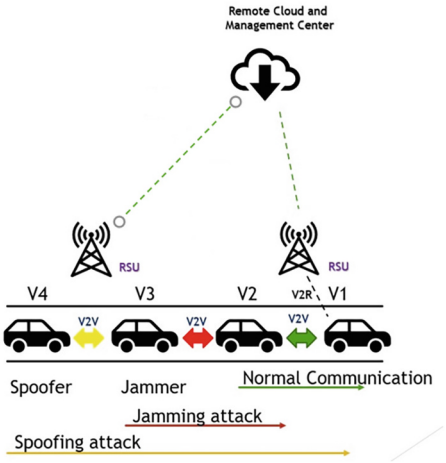

Fig. 2. VANET System Model

received data in the platoon, which contains information like PDR ratio, transmission power, and other values, is preprocessed before being fed into machine learning algorithms to classify whether it is malicious or benign, and eventually predict whether a jamming attack has occurred.

3.3 Jamming and Spoofing Detection System Framework

The purpose of an IDS is to detect evidence of attacks and suspicious operations. In intrusion detection and enhancing the security of the VANET network, various ML approaches can be beneficial. With each day, the use of ML methods in VANETs is gaining more popularity. ML for Jamming and Spoofing Detection Framework is essentially focused on classification models that help to identify the information analyzed with enhanced precision. Using feature extraction and data classification, three of the most effective models that will help distinguish the type of attack are incorporated.

3.4 Dedicated Short Range Communications Data Set

The used dataset for our IDS consists of information regarding Dedicated Short Range Communications (DSRC) [29] between vehicles and roadside units [11]. The two separate sets of normal scenario and attack scenario are combined into one set for the experimental analysis. For running vehicular network simulation, Veins [30] has been used here. The communications in the dataset were set up based on IEEE 802.11p standards at 5.9 Ghz, 10 BSM (Basic Service messages) per second, and are using Control Channel (Ch172), a 10 Mhz channel. The dataset consists of 390 simulations in a normal scenario as well as in the presence of an attacker. The key attributes are Transmitted node ID number, Received

Node Id number, Received Signal Strength in dbm, Packet Error Rate, Received Signal Strength Indicator, Signal to noise ratio, and the Target value, indicating whether the scenario is normal or in the presence of a jammer.

4 Tree-Based Algorithms and Frameworks

The ML-based IDS is focused on categories for classifying the analyzed data. In ML, the most popular and most often used supervised learning approaches are tree-based learning algorithms. They are diversified, easy to understand, and can be extended to a number of cases. This paper proposes three well known Tree-based ML algorithms for misbehaviour detection in vehicle communication.

These proposed algorithms are categorized under the supervised learning technique, and they are DT, RF, and ETE. A DT is a decision-making mechanism that uses a tree-like decision model. It is one way of displaying an algorithm consisting only of conditional statements. DTs are widely used classification techniques, but they are also a prominent means of ML in organizational analysis, especially for evaluating decisions. It is the building block of all tree models, including RF and ETE.

The structure of the RF algorithm is shown in Fig. 3, and it is a supervised learning algorithm focused on decision-tree models that partitions the data with the majority of votes in the groups of individual trees by a subset of training. The training algorithm applies the general technique of bootstrap aggregation, known as bagging. This method of bootstrapping boosts model efficiency as it lowers model variance without increasing the bias. This implies that, while for a single tree, predictions are extremely noise sensitive, the average of several trees is stable as long as trees are not correlated. RF is generally favored over others due to several reasons: Firstly, it has considerably higher precision for both classification and regression tasks. Secondly, it does not allow unnecessary trees in the model if there are more trees. It has the ability to manage a massive, multidimensional data set. Finally, the RF classifier manages the missing values and leaves a significant proportion of the data accurate.

Adding another randomization stage results in extremely random trees, or Extra Trees. While they are an aggregation of individual trees, they are comparable to ordinary random forests, as two key distinctions exist. First, each tree is trained using the whole sample of information and learning (rather than a bootstrap sample), and second, the top-down division of a tree learner is randomized. The random cutoff-point for any feature is chosen instead of determining the local optimum cutoff (based on Information gain or Gini impurity). This value is chosen from a uniform distribution within the analytical spectrum of the feature (in the tree training set). The division that gives the highest score is then chosen to divide the node with all randomly generated splits. The random number of features to be taken into account at any node can be defined, equivalent to ordinary random forests. Thus, it makes the ETE algorithm have less computational cost as compared to RF. The structure of the ETE method is shown in Fig. 4.

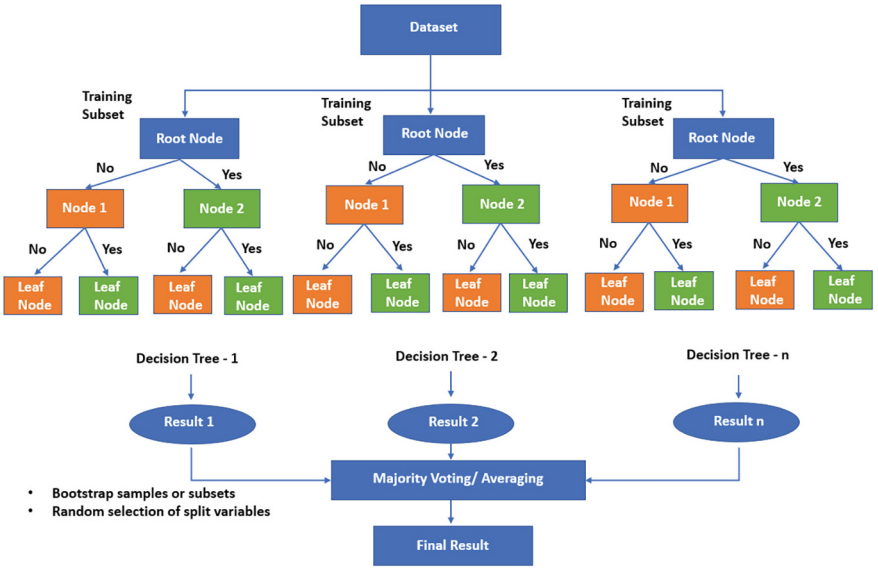

Fig. 3. Structure of Random Forest

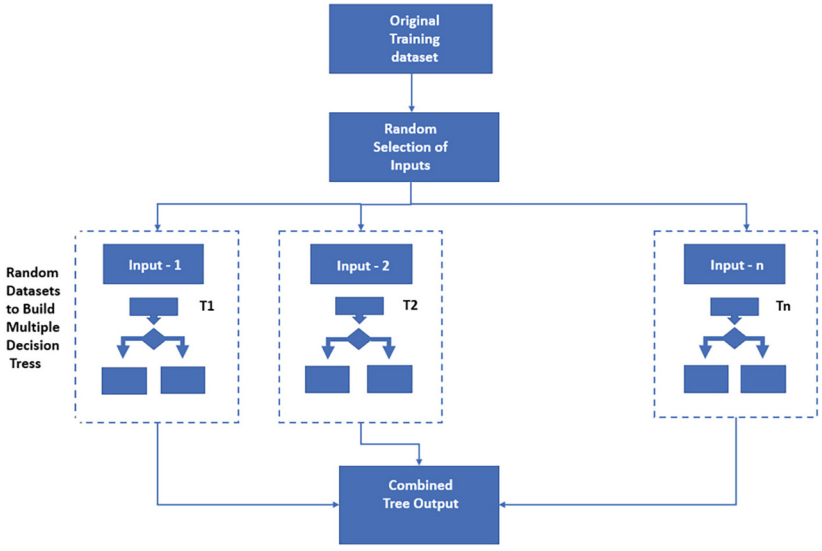

Fig. 4. Structure of Extra Tree

Table 3. Key Notations

Notation	Description
r	Pearson's Correlation Coefficient
x_i	values of the x-feature
$\bar{x}$	mean of the values of the x-feature
y_i	values of the y-feature
$\bar{y}$	mean of the values of the y-feature
AUC	the Area Under the Curve. It is used to measure the ability of a classifier to distinguish between classes
ROC	a graph showing the performance of a classification model across all rating thresholds

5 Experimental Analysis and Results

An ML approach has some principal steps that are applied in any kind of problem-solving scenario. These steps include data accumulation, data preprocessing, training, hyperparameter tuning, and prediction. For ML technique implementation, Jupyter Notebook is used. In addition, we will use Table 3 notation for the rest of the paper.

5.1 Data Preprocessing

The first and foremost step in any ML process is data preprocessing. The outlier features have been removed using the standard deviation. Standard Deviation is a measure that determines how often individual data points deviate from the mean. The features with zero Standard Deviation imply that they do not have any predictive capability. Hence, including such features may lead to overfitting, the scenario where the model gives good performance while training, but fails to perform well with new data.

Data can be available in various formats such as text files, audio files, video files, images, etc. In our case, the DSRC dataset is an Excel table with various numeric values in each feature (Table 4 shows the features of our dataset), the target has two classes: Benign (for normal scenario) and Malign (in the presence of an attacker). A machine cannot read such data since it can only understand binary information. Data Preprocessing is a step in which data is cleaned and encoded so that it can be easily interpreted by the machine. The target values are encoded, '1' representing the malignant data and '0' representing benign data.

To get a meaningful outcome from the experiment, it is important to understand the data as much as possible. The concept of correlation has been used here, and a Pearson's correlation matrix for each feature has been plotted. Correlation is a statistical measure that determines the dependency of features on

Table 4. Features used for modeling

Model Features Used	Description
RSU-BTx-Car	Roadside unit to car bidirectional transmission
RSU-PSR (%)	Road Side unit- soft routing protocol in percentage
Car P-Received	Number of packets received by car
Car-PDR (%)	Car packet delivery ratio in percentage
RSU PDSR (%)	Roadside unit - Packet Delivery / send ratio in percentage
Car Received power (dBm)(%)	The received power for a car in decibel milliwatts
Car-Btx-RSU	Car to Road Side Unit bidirectional transmission
RSU P-Received	Roadside unit packets received
RSU-PDR (%)	Roadside unit packet Delivery ratio in percentage
RSU Received power (dBm)	Power received by Roadside unit in decibel milliwatts
OBU	Type of On-Board units: 1OBU, 10OBU, 20OBU, 40OBU, 60OBU, 80OBU, 100OBU

each other. In other words, it tells us how much one feature is different from another. When a dependent variable and an independent variable have a high correlation value, it implies that the independent variable is very important in deciding the output. It is important to identify correlations among features and discard irrelevant ones to build a high-precision model.

Figure 5 shows the correlation matrix of features of our dataset. The Pearson's correlation coefficient is computed as:

$$\mathrm{r} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 (y_i - \overline{y})^2}} \tag{1}$$

5.2 Hyperparameter Tuning

Hyperparameters are parameters that cannot be learnt from the data and need to be defined before training. They indicate the ability of a model to train data and the complexity of the model. Hyperparameter tuning is the adjustment of the value of hyperparameters by the trial-and-error method, such that the model obtains an optimal solution. In DT, the maximum depth of the tree is the hyperparameter. In RF and ETE, the maximum depth and number of features considered by each tree while splitting are the hyperparameters. The Area Under

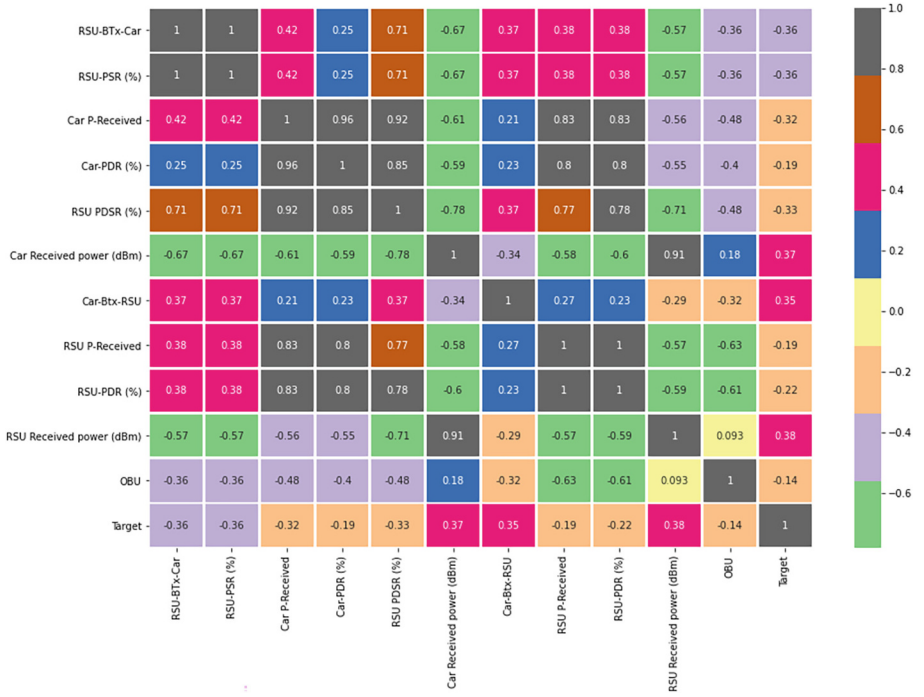

Fig. 5. Correlation Matrix of DSRC Dataset features

the Curve (AUC) graph for the RF model is plotted between AUC scores and the number of Estimators, and for other models, it is plotted between AUC scores and maximum depth. The AUC graph indicates the capacity of a model to distinguish between classes, in our data, malign and benign. If the AUC score is 1, this indicates that the model perfectly classifies positive and negative values, whereas when the AUC score is 0, it implies that the model incorrectly predicts positive values as negative and vice versa.

In DT, the graph becomes constant at the maximum depth of 20, implying that the optimal value of maximum depth for DT is 20. Figure 6 shows the AUC curve for DT, where Fig. 7 shows the AUC curve for the number of estimators. Even though the train AUC (blue line) becomes constant at 20 estimators, the test AUC is still varying. We decided to take 100 as the optimal number of estimators because increasing the number of estimators further will make the model more complex, resulting in poor performance. Figure 8 shows the AUC curve for maximum depth for Random Forest. For the ETE model, shown in Fig. 9, although the AUC curve becomes steady for some time starting at a maximum depth of 6 to 9, there has been a slight increase at the maximum depth of 10, implying that the optimal value of maximum depth for ETE is 10.

Furthermore, to prevent overfitting, we have applied Stratified K-Fold Cross Validation on RF and ETE classifiers, keeping the value of K as 10. In cross-

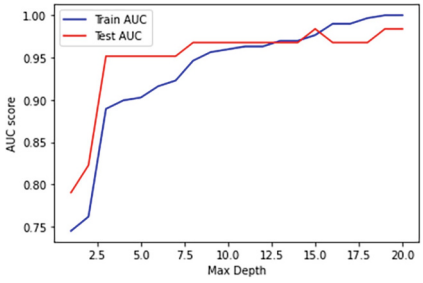

Fig. 6. AUC curve for Decision Tree

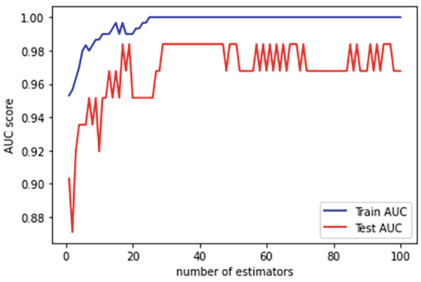

Fig. 7. AUC curve for Number of Estimators in Random Forest

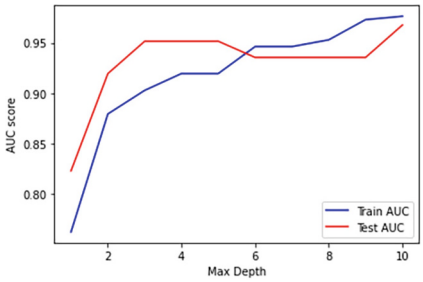

Fig. 8. AUC curve for Maximum Depth in Random Forest

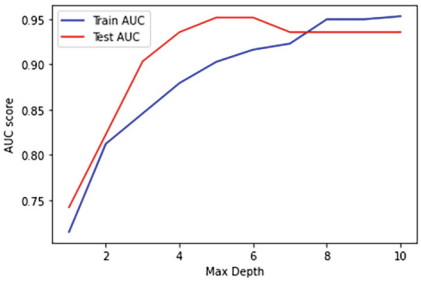

Fig. 9. AUC curve for Maximum Depth in Extra Tree Classifier

validation, data is split into K equal samples and the model is fit K times, with K-1 folds for training data and the Kth fold for validation data. The validation fold keeps changing with each iteration, implying that by the end of the evaluation, the entire data is utilized. The divisions are chosen in stratified K-fold cross-validation such that the average solution value is exactly proportional across all divisions.

5.3 Performance Metrics

To assess the performance of the ML algorithms, we employ Precision, Recall, Accuracy, and F1 Score. DT, RF, and ETE models are used to assess the efficiency of the established IDS. To validate our approach, we have to conduct the three targeted ML algorithms, including DT, on the same data set based on the same features. Then we compare the results of each algorithm under an identical experimental setup. Table 5 shows comparisons of the performance metrics results of the tree-based ML algorithms. It is evident that DT has 97% better test accuracy performance compared to 94% for RF and ETE, the reason being that our dataset is really small, having 390 instances and 15 features, and DT performs well with simpler data. Moreover, RF and ETE give 89% accuracy in 10 fold Cross Validation compared to 87% for DT, which implies that, overall,

Table 5. Results of the three models

Model	Test Accuracy %	Train Accuracy %	10 Fold Cross Validation Accuracy %	Test AUC	Train AUC	Precision	Recall	F1-Score
Extra Trees	94	95	89	0.93	0.94	1.0	0.87	0.93
Random Forest	94	95	89	0.93	0.94	1.0	0.87	0.93
Decision Tree	97	95	87	0.96	0.94	1.0	0.93	0.96

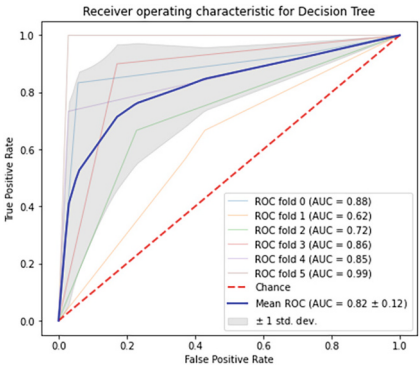

Fig. 10. ROC Curve for Decision Tree. (Color figure online)

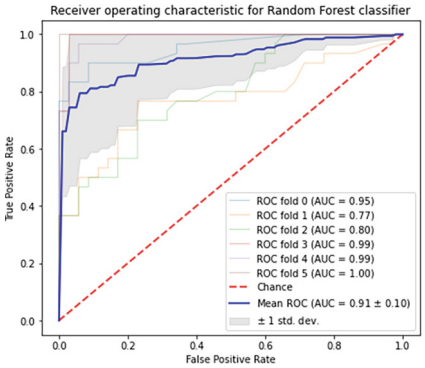

Fig. 11. ROC Curve for Random Forest

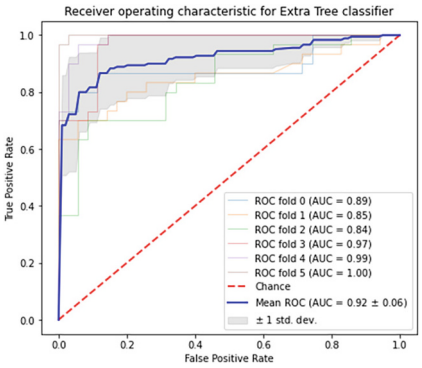

Fig. 12. ROC Curve for Extra Tree

RF and ETE are performing better than DT. In the future, when more data is added to this dataset or a more complex dataset is used, ETE shall surpass RF and DT in performance as it is computationally inexpensive due to random splitting of features at each node.

The results have been plotted on Receiver Operating Characteristic (ROC) plots individually for the three ML techniques. The ROC graph shows how well the classification model is performing by comparing the TPR (Recall) and FPR. An AUC curve is basically a synopsis of the ROC curve. We have plotted an ROC curve for each of the three ML approaches over a 6-fold Stratified Cross Validation data split.

Figure 10 shows the ROC curve for DT. The red dotted line represents the decision boundary or the chance line, and any point on that line will represent equal TPR and FPR. If the curve is above the red line, it denotes a higher TPR. The more the curve is bent above the red line, the better the model performs. If the curve is bent below the line, FPR is high, i.e., the model is performing poorly. The gray region denotes the Standard Deviation (measures how spread out the data points are) between the 6-foldlower and upper true positive rates. The individual 6 folds for DT have AUC scores of 88%, 62%, 72%, 86%, 85%, 95%, which means the scores range between 60% and 95%. The average ROC score and Standard Deviation for DT are 82% and 12%, respectively.

Figure 11 illustrates the ROC curve for RF. The individual 6-fold AUC scores for RF are 95%, 77%, 80%, 99%, 99%, and 100%, indicating a range of 75% to 100%. The average ROC score and Standard Deviation for RF are 91% and 10% respectively. Figure 12 depicts the ROC curve for ETE. The individual 6-fold ETE AUC scores are 95%, 77%, 80%, 99%, 99%, and 100%, indicating a range of 75% to 100%. The average ROC score and Standard Deviation for ETE are 92% and 0.06%, respectively.

From the three ROC graphs, it is clear that ETE has a higher mean AUC-ROC score and low Standard Deviation among all three methods, which implies that the data points are vastly clustered and there is a sparse chance of having outliers, thus preventing overfitting. This concludes that even though the DT model is performing better for the current dataset, ETE shall surpass both DT and RF when experiments are performed with more complex datasets in the future.

6 Conclusion and Future Work

In this paper, an IDS system using tree-based ML techniques, namely, decision tree, random forest, and extra tree ensemble algorithms, for detecting both spoofing and jamming attacks in a VANET network is proposed. The three techniques have been compared to analyze the best approach in terms of computational cost and complexity. For analysis purposes, the dedicated short-range communication DSRC dataset has been used. The simulation results showed that the decision tree in terms of test accuracy outperforms the other two approaches; however, in terms of cross-validation, it scores lower than them. The extra tree ensemble approach shall exceed the other two when tests are conducted with larger and more complex datasets. In future work, we will extend our approach to augmenting features extracted from the behavioral history of each vehicle in the platoon, as well as experiments on a wide range of different datasets.

References

1. Rocha Filho, G.P., et al.: Enhancing intelligence in traffic management systems to aid in vehicle traffic congestion problems in smart cities. Ad Hoc Networks **107** (2020)
2. Brennand, C.A.R.L., Filho, G.P.R., Maia, G., Cunha, F., Guidoni, D.L., Villas, L.A.: Towards a fog-enabled intelligent transportation system to reduce traffic jam. Sensors **19** (2019)
3. Zhou, W., Jia, Y., Peng, A., Zhang, Y., Liu, P.: The effect of iot new features on security and privacy: New threats, existing solutions, and challenges yet to be solved. IEEE Internet Things J. **6**(2), 1606–1616 (2018)
4. Mihai, S., Dokuz, N., Ali, M.S., Shah, P., Trestian, R.: Security aspects of communications in vanets. In: 2020 13th International Conference on Communications (COMM), pp. 277–282. IEEE, 2020
5. Englund, C., Chen, L., Vinel, A., Lin, S.Y.: Future applications of VANETs. In: Campolo, C., Molinaro, A., Scopigno, R. (eds.) Vehicular ad hoc Networks, LNCS, pp. 525–544. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-15497-8_18
6. Kosmanos, D., et al.: Intrusion detection system for platooning connected autonomous vehicles. In: 2019 4th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM), pp. 1–9. IEEE, 2019
7. Sommer, C., Tonguz, O.K., Dressler, F.: Traffic information systems: efficient message dissemination via adaptive beaconing. IEEE Commun. Mag. **49**(5), 173–179 (2011)
8. Mousannif, H., Khalil, I., Al Moatassime, H.: Cooperation as a service in vanets. J. UCS **17**(8), 1202–1218 (2011)
9. W. H. O. (WHO), "Road Traffic Injuries", 2021. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries. Accessed 21 June 2021
10. Vijayabharathi, V., Malarchelvi, P.S.K.: Implementing hmac in expedite message authentication protocol for vanet. In: International Conference on Information Communication and Embedded Systems (ICICES2014), pp. 1–5. IEEE, 2014
11. Malebary, S., Huang, C.: Jamming mobility in 802.11 p networks: modeling, evaluation, and detection. In: 2016 IEEE 35th International Performance Computing and Communications Conference (IPCCC), pp. 1–7. IEEE, 2016
12. Raghuwanshi, V., Jain, S.: Denial of service attack in vanet: a survey. Int. J. Eng. Trends Technol. **28**(1), 15–20 (2015)
13. Murshed, M., Wickramasurendra, N., Jubaida, A., De Grande, R.E., Carvalho, G.H.: Novel rf-jamming ids for vehicular networks: Lstm, xgboost, and meta-models approaches. Int. J. Crit. Infrastruct. Prot. 100830 (2026)
14. Alam, M.S., Oluoch, J., Kim, J.: A mechanism to localize, detect, and prevent jamming in connected and autonomous vehicles (cavs). IEEE Trans. Intell. Transp. Syst. **25**(2), 1215–1224 (2023)
15. Ahmad, J., et al.: Machine learning and blockchain technologies for cybersecurity in connected vehicles. Wiley Interdiscip. Rev. Data Min. Knowl. Discovery **14**(1), e1515 (2024)
16. Awan, K.A., Din, I., Syed, T., Cengiz, K., Almogren, A.: Trustworthy clustering for wireless energy sharing in vehicular systems for consumer applications. IEEE Trans. Consum. Electron. **71**(4), 12 236–12 243 (2025)

17. Boutahala, R., Fouchal, H., Ayaida, M., Mao, S.: Trust verification in connected vehicles using unsupervised variational autoencoder. In: 2025 12th International Conference on Wireless Networks and Mobile Communications (WINCOM), pp. 1–6. IEEE, 2025
18. Cheong, C., Song, Y., Cao, Y., Zhang, Y., Cai, B., Ni, Q.: Multidimensional trust evidence fusion and path-backtracking mechanism for trust management in vanets. IEEE Internet Things J. **11**(10), 18 619–18 634 (2024)
19. Sun, Y., et al.: PPDR: a privacy-preserving dual reputation management scheme in vehicle platoon. IEEE Trans. Dependable Secure Comput. (2025)
20. Xu, X., Wang, Y., Wang, P.: Comprehensive review on misbehavior detection for vehicular ad hoc networks. J. Adv. Transp. **2022**(1), 4725805 (2022)
21. Nobahari, A., Bakhshayeshi Avval, D., Akhbari, A., Nobahary, S.: Investigation of different mechanisms to detect misbehaving nodes in vehicle ad-hoc networks (vanets). Secur. Commun. Netw. **2023**(1), 4020275
22. Banabilah, S., Aloqaily, M., Alsayed, E., Malik, N., Jararweh, Y.: Federated learning review: fundamentals, enabling technologies, and future applications. Inf. Process. Manag. **59**(6) (2022)
23. Berdik, D., Otoum, S., Schmidt, N., Porter, D., Jararweh, Y.: A survey on blockchain for information systems management and security. Inf. Process. Manag. **58**(1) (2021)
24. Ridhawi, I.A., Aloqaily, M., Karray, F.: Intelligent blockchain-enabled communication and services: solutions for moving internet of things devices. IEEE Robot. Autom. Mag. **29**(2), 10–20 (2022)
25. Makani, S., Pittala, R., Alsayed, E., Aloqaily, M., Jararweh, Y.: A survey of blockchain applications in sustainable and smart cities. Clust. Comput. **25**(6), 3915–3936 (2022)
26. Singh, P.K., Gupta, S., Vashistha, R., Nandi, S.K., Nandi, S.: Machine learning based approach to detect position falsification attack in VANETs. In: Nandi, S., Jinwala, D., Singh, V., Laxmi, V., Gaur, M., Faruki, P. (eds.) Security and Privacy. ISEA-ISAP 2019. CCIS, vol. 939, pp. 166–178. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-7561-3_13
27. Zaidi, T., Faisal, S.: An overview: various attacks in vanet. In: 2018 4th International Conference on Computing Communication and Automation (ICCCA), pp. 1–6. IEEE, 2018
28. Sheikh, M.S., Liang, J.: A comprehensive survey on vanet security services in traffic management system. Wirel. Commun. Mob. Comput. **2019** (2019)
29. Malebary, S.: Dsrc vehicle communications dataset, center for machine learning and intelligent systems, 2017 https://archive.ics.uci.edu/ml/datasets/DSRC+Vehicle+Communications
30. Sommer, C., German, R., Dressler, F.: Bidirectionally coupled network and road traffic simulation for improved ivc analysis. IEEE Trans. Mob. Comput. **10**(1), 3–15 (2010)