Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Intrusion Detection System in Vehicular Network System

Winter Semester 2022-23

INFORMATION SECURITY MANAGEMENT

CSE

Project report

project submitted in partial fulfilment of the requirements for the degree of Bachelor of Technology in Computer Science and Engineering

Under guidance of:

RUBY . D

Submitted by:

P. Keertheswar 20BDS0304

V. Aishwarya 20BDS0230

Pidikiti Sai Chandu 20BDS0246

Gaurav Abhinav 20BCT0323

Arpit Sharma 20BCE0848

TABLE OF CONTENT

1. ABSTRACT:

In order to protect the security and privacy of vehicular ad hoc networks (VANETs), a crucial element of the intelligent transportation system, intrusion detection system (IDS) is required. In order to share information on traffic and road conditions, these networks, which are made up of vehicles and infrastructure elements, communicate with one another. Unfortunately, VANETs are susceptible to a number of attacks, such as replay attacks, data spoofing attacks, and denial-of-service attacks.

Researchers have suggested a variety of intrusion detection techniques for VANETs to address these security issues. Based on research we propose a system where various classification models are applied to design in-vehicle IDS. The dataset was extracted from CAN protocol and communication mechanism of various Connected Autonomous Vehicles (CAVs) generated in real-time environment. A Dataset of size 42,000 and target variable of 3 classes of attacks, is trained with various supervised machine learning models like Random Forest, K Nearest Neighbor, SVM, Artificial Neural Network and evaluated. With proper pre-pruning and implementation of models, it yields a maximum of 94.8% accuracy in case of Random Forest.

2. INTRODUCTION

With the emergence of networked devices, from the Internet of Things (IoT) nodes and cellular phones to vehicles connected to the Internet, there has been an ever-growing expansion of attack surfaces in the Internet of Vehicles (IoV). Attacks on IoV may lead to malfunctioning of Electronic Control Unit (ECU), brakes, control steering issues, and door lock issues that can be fatal in CAV. In order to safeguard autonomous vehicles against potential assaults, it is more crucial than ever to implement strong cybersecurity safeguards. Using deep learning algorithms for intrusion detection systems (IDS) made exclusively for autonomous vehicles is one way to tackle this problem.

Since vehicle networks are susceptible to a variety of security risks, intrusion detection systems (IDS) are essential to maintaining network security. Using Random Forest, a machine learning methodology that may efficiently detect intrusions by examining the behaviour of vehicular network traffic, is one efficient method for vehicular IDS. Due to its capacity for handling big datasets, high dimensional feature space, and the ability to deliver precise and quick detection findings, Random Forest is a well-liked option for vehicular IDS.

The CAN communication protocol is a carrier-sense, multiple-access protocol with collision detection and arbitration on message priority (CSMA/CD+AMP). CSMA means that each node on a bus must wait for a prescribed period of inactivity before attempting to send a message. The CAN protocol is a standard created to enable communication between the microcontroller and other devices without the use of a host computer. The broadcast bus type is what distinguishes the CAN protocol from other communication protocols.

2.1 INTRUSION DETECTION SYSTEM

An Intrusion Detection System (IDS) is a system that monitors network traffic for suspicious activity and issues alerts when such activity is discovered. It is a software application that scans a network or a system for harmful activity or policy breaching. Any malicious venture or violation is normally reported either to an administrator or collected centrally using a security information and event management (SIEM) system. A SIEM system integrates outputs from multiple sources and uses alarm filtering techniques to differentiate malicious activity from false alarms. A SIEM system is made to combine outputs from many security-related systems, such as IDS, firewalls, antivirus software, and other such systems. To distinguish between hostile activity and false alerts, it makes use of sophisticated algorithms and filtering methods. To provide a holistic view of the network's security posture, the system connects the events gathered from multiple sources.An IDS's objective is to identify malicious activities and send out prompt notifications to stop network harm. IDS may be network-based or host-based. Whereas network-based IDS keeps an eye on the whole network, host-based IDS is deployed on a single host and only monitors that host's behaviour.

2.2 ANOMALY BASED IDS

As an anomaly-based Intrusion Detection System (IDS), our model operates by analyzing network traffic or system behavior to identify patterns or behaviors that deviate significantly from normal or expected behavior. Here are some justifications for how our model works as an anomaly-based IDS:

Anomaly Detection: Our model is designed to detect anomalies in data. It is trained on a large corpus of historical data, which includes normal behavior patterns. During training, it learns to recognize patterns of normal behavior and builds a model of what is considered "normal" based on this data. When deployed in a live environment, it continuously monitors incoming data and compares it against this learned model of normal behavior. If the incoming data deviates significantly from the learned model, our model raises an alert, indicating that an anomaly has been detected.

Behavior-based Analysis: Our model analyzes the behavior of the system or network, rather than relying on predefined rules or signatures. This allows our model to adapt to changes in the system or network behavior and detect new or previously unseen types of attacks. It can detect anomalies that may not be captured by signature-based IDS, which rely on known patterns or signatures of known attacks.

In summary, our model works as an anomaly-based IDS by leveraging its anomaly detection capabilities, behavior-based analysis, adaptive learning, low false positives, and scalability to continuously monitor and analyze system or network behavior for deviations from normal behavior, and raise alerts when anomalies are detected.

3. <u>LITERATURE SURVEY</u>:

SL. NO	NAME OF THE TRANSACTION/JOURNAL/CONFERENCE WITH YEAR	MAJOR TECHNOLOGIES USED	RESULTS/OUTCOME OF THEIR RESEARCH	DRAWBACKS IF ANY
1.	Intelligent Intrusion Detection System for VANET Using Machine Learning and Deep Learning Approaches	It proposed effective Intrusion Detection System using machine learning and deep learning approaches such as Adaptive Neuro Fuzzy Inference System (ANFIS) and Convolutional Neural Networks (CNN)	The Attack Detection Rate of the malicious attack detection is about 98.5%, Botnet attack is about 98.9% and Brute Force attack is about 93.9%.	.The deep learning algorithm constructed multi-layer of decision i.e. cascade structure than reduce the time of intrusion detection.

2.	Towards a Lightweight Intrusion Detection Framework for In-Vehicle Networks	The real time Dataset obtain from CAN insulated cars is fine tuned and encoded to classify the attack label as Normal, Dos, Fuzzing attack, Reconnaissance using DNN model. Models used: Learning rate=0.001, epoch=2, hidden layer=2, input neuron=5 Other models like Random forest, Decision tree, KNN were also implemented.	Performance metric used are: accuracy, precision, recall, f1 score DNN: acc= 98.67% Precision= 96.72% Recall 96.50% F1 Score=0.967 Accuracy of other models are: RF: 97.92% KNN: 94.89% DT: 96.38%	1.Slow for large dataset. 2. Less no.of features used 3. Prone to over fitting
3.	Network intrusion detection system: machine learning approach	KNIME analytics platform. Resilient propagation (RPROP). CFS,PAC,IGR Some models used here are as follows: SVM RProp Decision tree	The detection accuracy of this model varies between 97.0% and 98.0% when it faces unknown attack types. The model had an accuracy score of over 80% and a detection latency of 0.61 ms , 90% using SVM method	Lack of accuracy of the data set This requires very large amount of data sets . It has inadequate infrastructure. Due to which it makes difficult to proceed .
4.	Analyzing Attack Strategies Against Rule-Based Intrusion Detection Systems (2018) [2]	This paper proposed a rule-clustering method to help prevent targeted attacks against a rule-based IDS.	Attackers can compromise an internal machine and launch attacks from that machine, which will evade the detection of 96% of rules.	works only as signature IDS that is it can detect only known attacks.
5.	Distributed collaborative intrusion detection system for vehicular Ad Hoc networks based on invariant (2020) [3]	A distributed collaborative IDS framework is adopted to construct an efficient and accurate detection system. An invariant-based dynamic behavior	Graphical representation is presented i.e., when the attack rises (given from 10 to 40%) the detection rate of DCDIV detection method is always higher than other	For field test of proposed detection method very large number of vehicle, driver and computer operators are needed.

		analysis technology is used to detect malicious behaviors, and the invariant contributes to analyze normal driving behavior.	detection method.	
6.	A game theory based multi layered intrusion detection framework for VANET (2018)	A multi layered game theory based intrusion detection framework is proposed for VANET that uses a set of specification rules and a lightweight neural network based classifier module to detect various type of attacks in VANET. The interaction between IDS and malicious vehicles is projected as two player non cooperative game which helps minimizing IDS traffic in bandwidth without compromising the overall performance of intrusion detection system.	In this IDS framework the detection rate is 99.63% and false alarm rate is only 9.13%. The proposed algorithm ensures the stability of the IDS framework by generating stable vehicular clusters with enhanced connectivity among member vehicles.	The game theory for interaction between IDS and malicious vehicle can produce very high IDS traffic in bandwidth constrained VANET but happens rarely.
7.	Intrusion Detection on the In-Vehicle Network Using Machine Learning (2021)	The iForestASD algorithm is based on the Isolation Forest (iForest) anomaly detection algorithm, which has linear time complexity, low memory requirement, and the ability to build a model with only a small amount of data. So, this paper builds on a work that applies iForest to intrusion detection for the CAN bus with promising results.	The high values obtained for the Area Under Curve (AUC) measure in the two cases, 0.966 and 0.974, indicated the effectiveness of this approach for intrusion detection.	Fully dependable algorithm on dataset so a major challenge for availability of relevant dataset.

8.	Cybersecurity Attacks in Vehicular Sensors Publisher: IEEE	Vehicle dynamics sensors (e.g., Tire Pressure Monitoring Systems (TPMS), magnetic encoders, and inertial sensors) and environment sensors (e.g., Light Detection and Ranging (LiDAR), ultrasonic, camera, Radio Detection and Ranging (Radar) systems, and Global Positioning System (GPS) units).	Using algorithms to reduce coupling between these modules and eventually reduce the response time.	There is a high degree of coupling, cohesiveness, and interactions among vehicle's CPS components (e.g., sensors, devices, systems, systems-of-systems) across sensing, communication, and control layers. Cyber-attacks in the sensing or communication layers can compromise the security of the control layer.
9.	Attacks to Automatous Vehicles: A Deep Learning Algorithm for Cybersecurity Publisher: MDPI	A real automatic vehicle network dataset, including spoofing, flood, replaying attacks, and benign packets. Preprocessing was applied to convert the categorical data into numerical. This dataset was processed by using the convolution neural network (CNN) and a hybrid network combining CNN and long short-term memory (CNN-LSTM) models to identify attack messages.	The results revealed that the model achieved high performance, as evaluated by the metrics of precision, recall, F1 score, and accuracy. The proposed system achieved high accuracy (97.30%).	Along with the empirical demonstration, the proposed system enhanced the detection and classification accuracy compared with the existing systems and was proven to have superior performance for real-time CAN bus security. But still the ouput is very slow and as the dataset increases the usability decreases

10.	A novel Intrusion Detection System for Vehicular Ad Hoc Networks (VANETs) based on differences of traffic flow and position (2019) [1]	An improved growing hierarchical self-organizing map I-GHSOM for IDS in VANET. An extraction algorithm is proposed to ex- tract the differences of traffic flow and of position.	The accuracy proposed IDS when the 40% of vehicles are rogue vehicles is 99.69%.	Limitation in finding more complex attack.
11.	A survey of new orientations in the field of vehicular cybersecurity, applying artificial intelligence-based methods Publisher: Wiley	The data set was analyzed by the K-means clustering and decision tree analysis methods to identify and characterize the generated groups of papers.	The dataset can be improved by taking into account other factors.	A database from 140 articles from the field of automotive security. In the database, we assigned specific attributes to every article (such as Web of Science Impact Factor or the number of citations). But the dataset is not covering the variety required for the same.
12.	Intrusion Detection System Using Deep Neural Network for In-Vehicle Network Security Publisher: Plos one	The parameters building the DNN structure are trained with probability-based feature vectors that are extracted from the in-vehicular network packets. For a given packet, the DNN provides the probability of each class discriminating normal and attack packets, and, thus the sensor can identify any malicious attack to the vehicle.	A proper train using a certain dataset to avoid the system's vulnerabilities being exploited.	As compared to the traditional artificial neural network applied to the IDS, the proposed technique adopts recent advances in deep learning studies such as initializing the parameters through the unsupervised pre-training of deep belief networks (DBN), therefore

				improving the detection accuracy. But still the outcome is not super fined and the implementation has loopholes
13.	Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization Canadian Institute for Cybersecurity (CIC), University of New Brunswick (UNB), Canada ICISSP 2018	•CICFlowMeter, 2017. Which is a flow-based feature extractor. •DARPA •KDD'99 •LBNL •CAIDA	In order to assess the effectiveness of intrusion detection systems, this study introduces a novel intrusion detection dataset that comprises multiple types of attacks and traffic characterization approaches.	Lack of the feature set and metadata.
14.	Efficient Text Classification Using Tree-structured Multi-linear Principal Component Analysis. University of Southern California, Feb 2018.	•TMPCA Algorithm and PAC. SMS Spam dataset (SMS SPAM). It has "Spam" and "Ham" as two target classes. Sandford Sentiment Treebank (SST). It has "positive" and "negative" as two target classes	Using tree-structured multi-linear principal component analysis, this research suggests a novel approach to classifying texts that both reduces computer complexity and considerably increases classification accuracy.	•Low accuracy in LTSM method. •Latency Is low.
15.	Tree-structured multi-stage principal component analysis (TMPCA): Theory and applications. Expert Systems with Applications Volume 118, March 2019.	•High efficiency. •Low information loss. •Sequential preservation. •Unsupervised learning. •Transparent mathematical properties. SMS Spam (spam) (Almeida, Hidalgo, & Yamakami, 2011). It is a dataset collected for mobile Spam email detection.	The idea of tree-structured multi-stage principal component analysis (TMPCA), a potent technique for evaluating huge datasets with complicated structures, is introduced in this work along with a discussion of its uses in several disciplines.	•PCA is not robust against outliers •Its sensitive to the scale of the features. •It does not take into account any a-priori knowledge, as the parametric algorithms do.

16.	Evaluation of SVM performance in the detection of lung cancer in marked CT scan dataset. Research gate, March 2021	SVM Lung Cancer, CT-Scan, Computer Vision, Datasets. three types of features are employed for training the model. First, the Gabor filter only. second GLCM matrix only, and third a combination of the two mentioned features. SVM was applied using three kernels, linear kernel, RBF kernel and polynomial kernel.	In this study, the effectiveness of the support vector machine (SVM) at identifying lung cancer in a marked CT scan dataset is assessed. The results are encouraging and have the potential for future applications.	•It's not much suitable for large data sets. •Target classes are overlapping.
17.	CICIDS2017 dataset: performance improvements and validation as a robust intrusion detection system testbed. Published Online inder science , August 2021	•Intrusion detection system IDS. •network security. •network attacks. •CICIDS2017, principal component analysis PCA. •Machine learning.	This study shows how the CICIDS2017 dataset has improved in performance and validated, proving its use as a reliable testbed for intrusion detection systems.	• Inability to respond or stop attacks upon detection. •Noise can severely limit capacity to work.
18.	Deep Transfer Learning Based Intrusion Detection System for Electric Vehicular Networks (2021).	•TML •DL •DTL. •CNN	In order to increase the accuracy of intrusion detection, this study suggests a deep transfer learning-based intrusion detection system for electric vehicle networks.	Should Concentrate on improving the performance of the proposed model by optimizing the hyper-parameters.
19.	Secure message propagation protocols for IoV's communication components	•Cryptographic operations •Hash functions •Symmetric encryption. •Symmetric decryption.	For the communication parts of the Internet of Vehicles (IoV), this research suggests secure message propagation methods that enable dependable and secure message transmission in vehicular networks.	Performance analysis should be given in detail. Expected more acuracy.

20.	Integrating security and privacy in software development (2020)	•Software development approach •Privacy Oriented •Software Development (POSD)	In addition to outlining a framework for secure software development that includes security and privacy criteria, design principles, and testing procedures, this paper highlights the significance of incorporating security and privacy into software development.	•No specific drawback can be stated. •Accuracy could be increased and shown Easily.
21.	Survey and Classification of Automotive Security Attacks	•Threat analysis or security testing. •Each step in the vehicle development process can leverage it.	The impact and defences against each form of attack are covered in this study, along with a survey and classification of automobile security assaults, including physical, network, and software attacks.	•No Specific new model is introduced and it's based on treat analysis and •Making existing taxonomy effective.
22.	Survey of Automotive Controller Area Network Intrusion Detection Systems Publisher: IEEE	•IDS. •Special emphasis on techniques for detecting attacks on CAN modules.	The CAN bus, a vital part of automotive networks, is targeted frequently by attacks. This paper provides an overview of CAN intrusion detection systems that can identify and stop such attacks.	•Survey has overall knowledge of all possible attacks. •Attack occurrence rate could be given more clearly.

23.	A Novel Intrusion Detection Method for Intra-Vehicle Networks Using Recurrence Plots and Neural Networks.
	AUTHOR: Omar Y. Al-Jarrah, Karim El Haloui, Mehrdad Dianati and Carsten Maple
	
	Publication: ResearchGate January 2023
		Anomaly based Intrusion detection System is build using "CAN- intrusion dataset" to classify the attacks into DoS, Gear attack, or Fuzzy attack.
		
		MODEL USED:
		LSTM Neural network.
		Random Forest,
		Decision Tree.
			Performance metric used are: accuracy, f1 score
			Accuracy for
			
			LSTM = 93.78%
			RF = 87.18%
			DT = 88.19%
				•Random forest acts as a black box to analyze the modeling.
				•Requires enormous time to train the data.
				•LSTM does not work well with highly non-linear data

24.	Classification of Recurrence Plots Distance Matrices with a Convolutional Neural	The proposed method consisted of transforming the original	The resulting distance matrices were stacked together as a single	Slow for larger dataset,

	Network for Activity Recognition Science direct 2018	raw 1D signal for each axis into a 2D image format that captures temporal patterns. The transformation was based on recurrence plots analysis and consisted of computing a distance matrix for each input 1D signal.	image and a CNN was trained propose a method based on recurrence plots' distance matrices and convolutional neural network (CNN) that does not require feature engineering learning rate = 0.001, 50 epochs overall accuracy of proposed method was 0.942, recall=0.804	Complex computational processing
25.	Future Intelligent and Secure Vehicular Network Toward 6G: Machine-Learning Approaches Publisher: IEEE XPLORE	Naturally, employing ML into vehicular communication and network paving the way for the future intelligentization in 6G vehicular networks	The techniques covered are well but they don't provide the best output instead several other can be employed for the same.	A survey on various ML techniques applied to communication, networking, and security parts in vehicular networks and envision the ways of enabling AI toward a future 6G vehicular network.

3.2 RESEARCH GAP

The following points are not consistently addressed in numerous articles, according to a survey of the work done by different research fraternities. The following is a summary of the research gaps from the research publications mentioned above:

- Only a few have used datasets related to vehicles for their analysis.
- Most of them use traditional classifiers that limit performance and creativity.
- The majority of IDS models are not evaluated using real-time datasets.
- The major challenge is the availability of relevant datasets related to vehicular systems.
- Most of the work in the papers used only one performance metric in evaluating the proposed model.

4. SYSTEM DESIGN AND ARCHITECTURE DIAGRAM

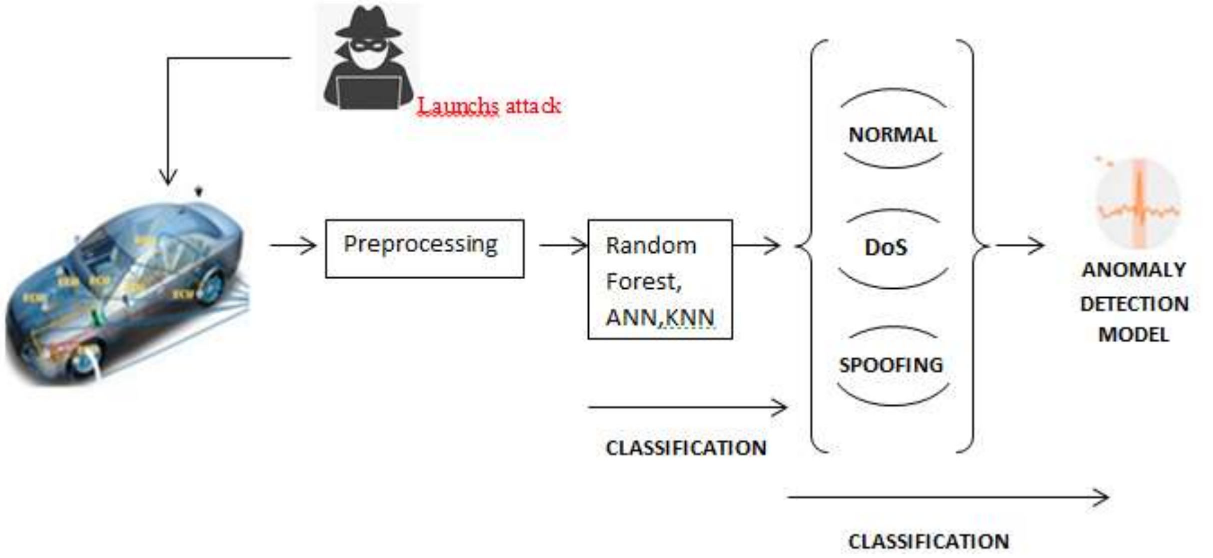


Fig 1: Architecture model of our IDS model

4.1 SUMMARY OF MODULES IN ARCHITECTURE MODEL

i) PREPROCESSING:-

Preprocessing in machine learning refers to the data preparation steps that are performed on raw data to make it suitable for training a machine learning model. Preprocessing involves several techniques that are applied to the data to remove noise, handle missing values, reduce data dimensionality, normalize data, and convert categorical data into numerical data.Some of the common techniques used by us in our model for preprocessing are:

Data cleaning: This involves removing irrelevant or duplicate data, correcting errors, and handling missing data.

Feature scaling: This involves scaling or normalizing the data to ensure that the features are in the same range.

Feature encoding: This involves converting categorical data into numerical data to enable the model to understand the data better.

Data splitting: This involves splitting the data into training, validation, and testing sets to evaluate the performance of the model.

Preprocessing is an essential step in machine learning as it helps in improving the accuracy and efficiency of the model by ensuring that the data is suitable for the model's training.

ii) CLASSIFICATION MODELS

RANDOM FOREST:

Random forest improves on bagging because it decorrelates the trees with the introduction of splitting on a random subset of features. This means that at each split of the tree, the model considers only a small subset of features rather than all of the features of the model.

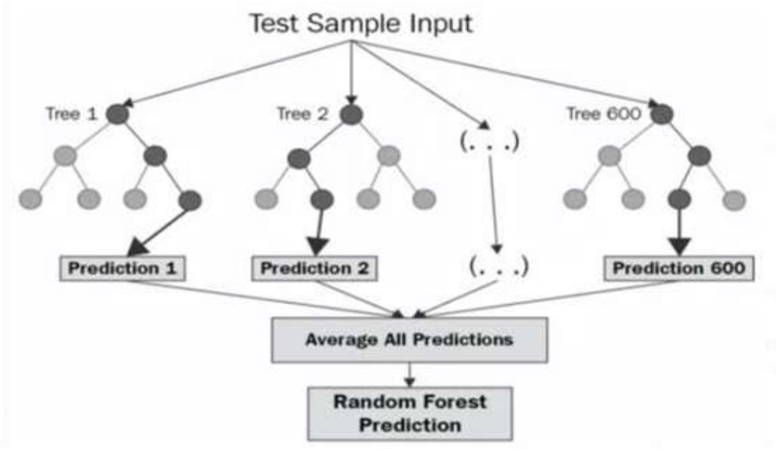


Fig 2: Random Forest

K NEAREST NEIGHBOR (KNN)

K-Nearest Neighbors (KNN) is a simple but powerful machine learning algorithm used for classification and regression tasks. In KNN, the prediction for a new instance is based on the "k" closest instances in the training data, where "k" is a user-defined parameter. The distance metric used to calculate the distance between instances is typically the Euclidean distance or the Manhattan distance.

ARTIFICIAL NEURAL NETWORK (ANN)

ANN consists of multiple layers of interconnected artificial neurons, where each neuron performs a simple computation based on its input and activation function. The input layer of an ANN receives the raw input data, and the output layer produces the final prediction. The layers in between are called hidden layers, and their purpose is to transform the input data into a form that is more suitable for the prediction task.

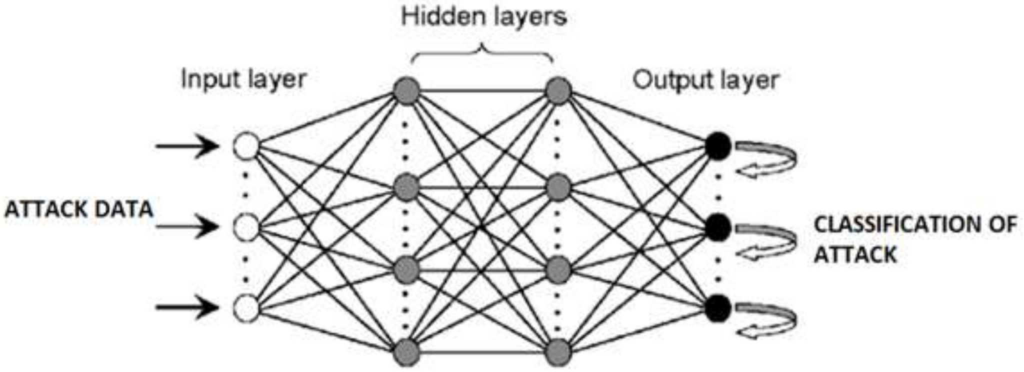


Fig 3: Neural network layers

4.2 CLASSIFICATION OF ATTACKS

i. DoS: DoS attacks can take different forms, such as flooding the module with a large number of requests, injecting malicious data packets, or exploiting vulnerabilities in the module's firmware or software. The goal of such attacks is to prevent the module from functioning properly, which can lead to a loss of control over the vehicle or cause safety hazards for the passengers and other road users.

ii. SPOOFING: Spoofing attacks in vehicles refer to the malicious act of impersonating a legitimate entity or device to gain unauthorized access to the vehicle's electronic systems. Spoofing attacks can take different forms, such as falsifying messages from the vehicle's key fob, GPS device, or tire pressure monitoring system (TPMS). The goal of such attacks is to trick the vehicle into granting access to unauthorized individuals or to cause safety hazards for the passengers and other road users.

4.3 IMPACT OF THESE ATTACKS ON VEHICLES IN REAL LIFE

In a DOS (Denial of Service) spoofing gear attack, the attacker manipulates the communication network between several automobile system components in an effort to impair the vehicle's normal operation. This kind of attack may result in serious implications, such as the loss of essential safety features and potential injury to passengers or bystanders.

1. Unauthorised Access: Either physically or remotely, the attacker gains access to the communication network of the car. This can be accomplished via taking advantage of software flaws in the car, hacking into ECUs, or listening in on wireless connections.
2. Spoofing Gear Deployment: To tamper with communication between various parts of the car's system, the attacker uses spoofing gear, which can be specialised equipment or gadgets.
3. Fake Message Generation: With the aid of spoofing equipment, the attacker creates false signals or messages. These signals may be intended to trick the car's systems into acting inappropriately or choosing the wrong course of action.
4. Message Flooding: The attacker sends a large number of phoney messages across the car's communication network. A denial of service may arise from this overwhelming the system and causing it to malfunction.
5. Timing/Sequencing Manipulation: To prevent the systems in the car from functioning normally, the attacker changes the timing or sequencing of messages. Critical messages, such as those pertaining to braking or throttle control, may be delayed or altered in order to do this.
6. System Disruption in the Car: The timing/sequencing manipulation and bogus signals can cause system disruption in the car. This may cause the car's safety features to malfunction, lose control, or operate improperly, which could cause accidents or inconvenience for the passengers.
7. Attack Disguised: To avoid being discovered by security measures, the attacker may try to hide their attack by erasing logs or hiding their footprints.

It's crucial to keep in mind that this is a simplified overview of a DOS spoofing gear assault on automobile car systems, and the actual attack may require more complex tactics and variants depending on the precise vulnerabilities and systems involved. To prevent such assaults and guarantee the security of motor vehicles, comprehensive cybersecurity solutions, such as tight access controls, encryption, authentication, intrusion detection/prevention systems, and routine security updates, are necessary.

5. **DATASET**

The dataset was first generated by Guillame Dupont [1] in a real-time scenario with two cars, namely, a Renault Clio and an Opel Astrand with another obtained by prototyping using Ardino boards and CAN bus shields. A further study of Intrusion Detection System was proceeded by Dheeraj Basavaraj and Shahab Tayeb in the paper, Lightweight Intrusion Detection System [2].

In case of cars, normal data are captured by driving the cars in urban environment assisted by CAN utility tools. Data manipulations are carried out on CAN bus data to hack automobile units to display false readings of fuel level and speedometer and to prevent activation of critical modules like brake and airbag signals. These types of attacks can be as hazardous as it can lead to dead of passengers.

The Dataset used in this project was taken from the paper Lightweight Intrusion Detection System [2]. The original dataset had about 3 million instance for each attack type. As huge dataset causes computational and storage complexities, the dataset is reduced to **42,374** entries.

5.1 DATA ATTRIBUTES

ATTRIBUTE	DESCRIPTION	DATA TYPE	COUNT
DLC	Number of data bytes from 0 to 8	Int64	42374
DATA[0-2]	Data value (byte)	object	42374
DATA[3-7]	Data value (byte)	object	42350
ATTACK_TYPE	Class labels of attacks	object	32374

Table 1: Description of data attributes

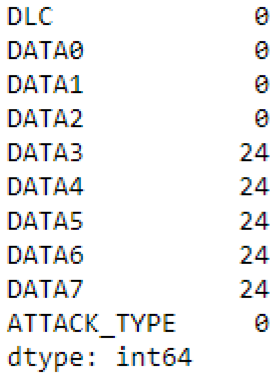
5.2 TARGET CLASSIFIERS

CLASS LABEL	DESCRIPTION	COUNT
Normal	Data captured in real-time driving car	36887
DoS	Data captured in CAN bus after launching DoS attacks from compromised ECU module	2990
Spoofing in the Drive gear	Data captured in CAN bus after launching spoofing attack in the gear module from a compromised ECU	2473

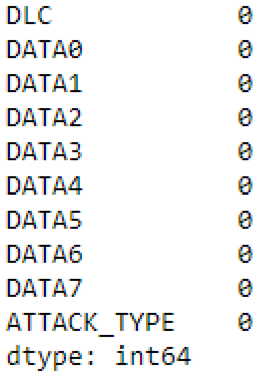
Table 2: Overview of class labels

6. PREPROCESSING

1. As the check for missing values performed, overall missing values were found to be 125, 24 in each attribute DATA[3-7]. 125 instances among 42374 records is far less than the dataset size, which won't affect the quality of dataset. Hence they are dropped.



```
DLC             0
DATA0           0
DATA1           0
DATA2           0
DATA3          24
DATA4          24
DATA5          24
DATA6          24
DATA7          24
ATTACK_TYPE     0
dtype: int64
```

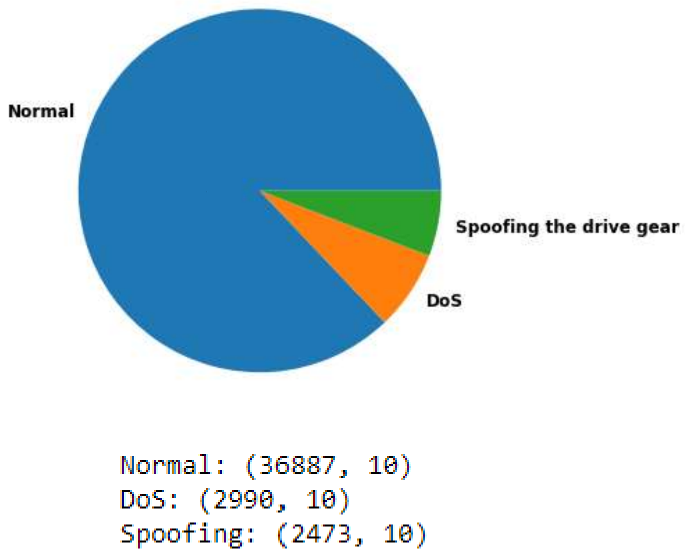


```
DLC             0
DATA0           0
DATA1           0
DATA2           0
DATA3           0
DATA4           0
DATA5           0
DATA6           0
DATA7           0
ATTACK_TYPE     0
dtype: int64
```

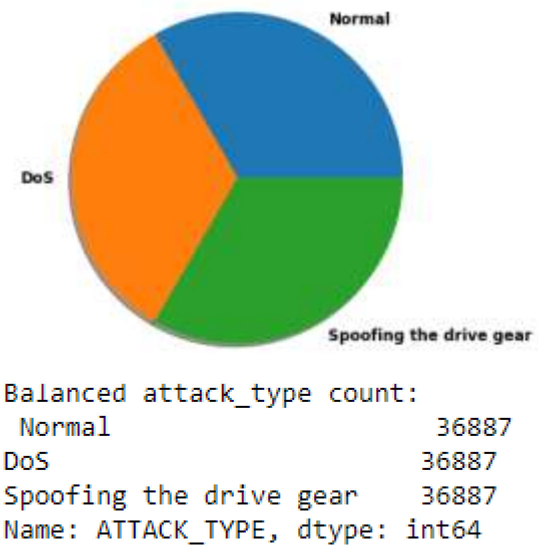
Fig 1: Count of missing values in each column

Fig 2: Count of missing value after preprocessing

2. Analysis of distribution of attack type is performed and visualized using a pie chart where the normal cases where around 36,899 and Dos is 2990, Spoofing were 2789. This shows it is highly imbalanced. So over sampling performed over DoS and Spoofing.



```
Normal: (36887, 10)
DoS: (2990, 10)
Spoofing: (2473, 10)
```



```
Balanced attack_type count:
 Normal                        36887
DoS                           36887
Spoofing the drive gear       36887
Name: ATTACK_TYPE, dtype: int64
```

Fig 3: Unbalanced distribution of target class

Fig 4: Distribution of class label after over sampling

3. Further the categorical attributes were encoded using LabelEncoder as Normal = 0, DoS=1, Spoofing = 2

7. MODEL IMPLEMENTATION

After data are pre-processed, the next step is to implement machine learning models to classify the entities. A comparative study of 3 supervised models is performed to study which model is best suited for Intrusion detection system.

1. **Random Forest** : Random Forest is an ensemble learning method that combines the predictions of multiple decision trees to make a final prediction. Random Forest reduces the chances of overfitting, leading to better generalization performance on unseen data. Random Forest often provides high accuracy in various machine learning tasks.

2. **K Nearest Neighbour**: K-Nearest Neighbours (KNN) is a popular algorithm for both classification and regression tasks in machine learning. It is a simple yet effective algorithm that can be used for a wide range of applications. This experimentation worked well with high accuracy rate when K value is equal to 1

3. **Artifial Neural Network**: The accuracy of an ANN model can vary depending on various factors such as the architecture of the network (e.g., number of layers, number of neurons in each layer), activation functions, learning rate, batch size, and quality of the training data. We are using the 'sgd' optimiser with a learning rate 0.01, maximum iteration of 200. Number of hidden layer are 3 with 20 neural size. We used alpha value as 1e-5 which helps to decrease the over fitting.

8. RESULTS AND DISCUSSION

The overall performance of the model is evaluated on the basis of performance metrics, i.e., accuracy, precision, recall, and f-1 score. Comparison is carried out on different datasets with different models which are published by the research team. The metric used for comparison and what it specifies is given below.

- **Accuracy:** It shows the percentage of correctly predicted values compared to overall predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where, TP = actually true, predicted true
TN = actually false, predicted false
FP = actually false, predicted true
FN = actually true, predicted false

- **Percision:** It refers to the ratio of correctly predicted positive observations to overall positive predicted observations.
- **Recall:** refers to sensitivity. It is the ratio of correctly predicted positive observations to overall observations in the output label class.
- **F1 score:** It refers to a weighted average of precision and recall. It is used when there is an uneven distribution of classes.

MODEL	ACCURACY(%)	PERCISION(%)	RECALL(%)	F1 SCORE
Random forest	94.12	94.33	98	0.95
KNN	90.40	93	95	0.94
ANN	61.41	54	63	0.57

Table 3: Overall performance metrics of our model

From the above table, it is evident that Random forest gives highest accuracy rate the CAN dataset. The accuracy obtained in the Random Forest model is 94.12, which is phenomenally critical in the case of the IDS system. Other metrics such as precision and recall show that our model correctly predicted values.

Further analysis of accuracy over different K values for K Nearest Neighbor is also performed.

K VALUE	ACCURACY(%)	PERCISION	RECALL	F1 SCORE
K=1	90.40	0.93	0.91	0.89
K=2	88.32	0.88	0.86	0.85
K=3	89.0	0.80	0.82	0.87
K=4	86.75	0.82	0.79	0.84

Table 4: Performance of KNN for different k values

9. CONCLUSION

Modern automobiles are equipped with various communication networks, multiple sensors, actuators, cameras, radars, and communication devices to improve performance, efficiency, intelligent services, and safety of passengers. This increased complexity and connectivity make vehicles vulnerable to cyber-attacks. Both in-vehicle and VANET networks are exposed to such attacks. IDSs are used to identify cyber-attacks in vehicle networks. Malicious attacks are constantly varying, which makes security sustaining a challenging practice. The main impact of this research is emerging a machine learning approach for intrusion detection. The Internet of Vehicles (IoV) is gaining popularity as it enables vehicles to communicate with traffic externally and communicate with emergency modules internally. As this vehicular communication becomes popular, the CAN communication protocols become vulnerable to cyberattacks. As a result, there has been a rapid increase in research on automotive security.

The use of embedded and portal devices in vehicles provides communication outside and inside the vehicle's environment. As this vehicular communication gains popularity, the vulnerabilities of the communication protocol become critical. As a result, there has been a rapid increase in research on automotive cybersecurity.

The previous methods used different models for the benchmarked datasets, namely, KDD99, CIDDS, and CAN hacking data sets. In our paper, we used a dataset generated from the real-time environment of two cars (Opel Astra and Renault Clio) and of another one obtained by prototyping (consisting of a VW instrument cluster, two Arduino boards with CAN bus shields, and a joystick); these are addressed in [12]. We have encoded the data using an embedded column that improves with training, and a better performance can be achieved due to the dynamic nature of the embedding column encoding technique.

The future scope for Intrusion Detection Systems (IDS) in vehicular networks using machine learning is promising and offers several potential areas of development and advancement. Some of the future directions for IDS in vehicular networks using machine learning may include:

Algorithms for advanced machine learning: Algorithms for advanced machine learning are constantly being developed, and future research may result in the creation of even more sophisticated algorithms that are customised specifically for vehicular networks. This could include ensemble approaches, deep learning models, and reinforcement learning algorithms, which can all improve the IDS's sensitivity and robustness when it comes to spotting intrusions in automotive networks.

Privacy-preserving IDS: In automobile networks, safeguarding the privacy of passengers is of utmost importance. To make sure that the intrusion detection 14 procedure does not compromise the privacy of the data created by cars, future IDS may concentrate on creating privacy-preserving approaches like differential privacy, federated learning, and secure multi-party computation.

Response and mitigation in real time are essential because intrusions into vehicle networks can have catastrophic repercussions. The development of real-time mitigation measures, such as automated incident response, dynamic access restriction, and threat intelligence integration, may be the main emphasis of future IDS in order to effectively respond to security problems in vehicle networks and stop additional harm.

10. REFERENCE

1) Li, H.; Zhao, L.; Juliato, M.; Ahmed, S.; Sastry, M.R.; Yang, L.L. POSTER: Intrusion Detection System for In-vehicle Networks using Sensor Correlation and Integration. In *ACM SIGSAC Conference on Computer and Communications Security (CCS '17)*; Association for Computing Machinery: New York, NY, USA, 2017; pp. 2531–2533. [Google Scholar]
2) He, Q.; Meng, X.; Qu, R.; Xi, R. Machine Learning-Based Detection for Cyber Security Attacks on Connected and Autonomous Vehicles. *Mathematics* **2020**, *8*, 1311. [Google Scholar] [CrossRef]
3) Thapa, N.; Liu, Z.; Kc, D.B.; Gokaraju, B.; Roy, K. Comparison of Machine Learning and Deep Learning Models for Network Intrusion Detection Systems. *Future Internet* **2020**, *l2*, 167. [Google Scholar]
4) Tayeb, S.; Pirouz, M.; Latifi, S. A Raspberry-Pi Prototype of Smart Transportation. In Proceedings of the 2017 25th International Conference on Systems Engineering (ICSEng), Las Vegas, NV, USA, 22 August 2017; pp. 176–182. [Google Scholar] [CrossRef]
5) Trueblood, F.; Gill, S.; Wong, R.; Tayeb, S.; Pirouz, M. A Data-Centric Approach to Taming the Message Dissemination on the Internet of Vehicles. In Proceedings of the 10th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 6 January 2020; pp. 207–214. [Google Scholar] [CrossRef]
6) Gill, S.; Wong, R.; Tayeb, S.; Trueblood, F.; Pirouz, M. Optimizing Connectivity for the Internet of Vehicles. In Proceedings of the Future Technologies Conference (FTC) 2020; Arai, K., Kapoor, S., Bhatia, R., Eds.; Advances in Intelligent Systems and Computing. Springer: Cham, Switzerland, 2021; Volume 3. [Google Scholar] [CrossRef]
7) Hamada, Y.; Inoue, M.; Adachi, N.; Ueda, H.; Miyashita, Y.; Hata, Y. Intrusion Detection System for In-Vehicle Networks. *SEI Tech. Rev.* **2019**, *88*, 76–81. [Google Scholar]
8) Pan, L.; Zheng, X.; Chen, H.X.; Luan, T.; Bootwala, H.; Batten, L. Cyber security attacks to modern vehicular systems. *J. Inf. Secur. Appl.* **2017**, *36*, 90–100. [Google Scholar] [CrossRef]
9) Davis, A.; Gill, S.; Wong, R.; Tayeb, S. Feature Selection for Deep Neural Networks in Cyber Security Applications. In Proceedings of the 2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Vancouver, BC, Canada, 9–12 September 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–7. [Google Scholar] [CrossRef]
10) Denning, D.E. An Intrusion-Detection Model. *IEEE Trans. Softw. Eng.* **1987**, *13*, 222–232. [Google Scholar] [CrossRef]
11) Mukherjee, B.; Heberlein, L.T.; Levitt, K.N. Network intrusion detection. *IEEE Netw.* **1994**, *8*, 26–41. [Google Scholar] [CrossRef]
12) Dupont, G.; Lekidis, A.; den Hartog, J.; Etalle, S. Automotive Controller Area Network (CAN) Bus Intrusion Dataset v2. 4TU.ResearchData. *4TU.ResearchData*. 2019. Available online: **https://data.4tu.nl/articles/dataset/Automotive_Controller_Area_Network_CAN_Bus_Intrusion_Dataset/12696950/2** (accessed on 30 December 2021).
13) Aloqaily, M.; Otoum, S.; Ridhawi, I.A.; Jararweh, Y. An intrusion detection system for connected vehicles in smart cities. *Ad Hocnetworks* **2019**, *90*, 101842. [Google Scholar] [CrossRef]
14) Barletta, V.S.; Caivano, D.; Nannavecchia, A.; Scalera, M. Intrusion Detection for in-Vehicle Communication Networks: An Unsupervised Kohonen SOM Approach. *Future Internet* **2020**, *l2*, 119. [Google Scholar] [CrossRef]
15) Ali, M.H.; Al Mohammed, B.A.D.; Ismail, A.; Zolkipli, M.F. A new intrusion detection system based on fast learning network and particle swarm optimization. *IEEE Access* **2018**, *6*, 20255–20261. [Google Scholar] [CrossRef]
16) Tao, P.; Sun, Z.; Sun, Z. An improved intrusion detection algorithm based on GA and SVM. *IEEE Access* **2018**, *6*, 13624–13631. [Google Scholar] [CrossRef]
17) Brown, J.; Anwar, M.; Dozier, G. An Evolutionary General Regression Neural Network Classifier for Intrusion Detection. In Proceedings of the 25th International Conference on Computer Communication and Networks (ICCCN), Waikoloa, HI, USA, 1 August 2016; Volume 6, pp. 1–5. [Google Scholar]
18) Gautam, S.K.; Om, H. Computational neural network regression model for host-based intrusion detection system. *Perspect. Sci.* **2016**, *8*, 93–95. [Google Scholar] [CrossRef][Green Version]
19) Masarat, S.; Taheri, H.; Sharifian, S. A novel framework, based on fuzzy ensemble of classifiers for intrusion detection systems. In Proceedings of the 4th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, Iran, 29–30 October 29; Volume 8, pp. 165–170. [CrossRef]
20) Oh, D.; Kim, D.; Ro, W.W. A malicious pattern detection engine for embedded security systems on the Internet of Things. *Sensors* **2014**, *14*, 24188–24211. [Google Scholar] [CrossRef] [PubMed]
21) A Ghaleb, F.; Saeed, F.; Al-Sarem, M.; Ali Saleh Al-rimy, B.; Boulila, W.; Eljialy, A.E.M.; Aloufi, K.; Alazab, M. Misbehavior-Aware On-Demand Collaborative Intrusion Detection System Using Distributed Ensemble Learning for VANET. *Electronics* **2020**, *9*, 1411. [Google Scholar] [CrossRef]
22) Mrugnayana, S.; Savekar; Sandeep, A. Identifying Impersonation Attack in VANET using k-NN and SVM Approach. *Int. J. Future Gener. Commun. Netw.* **2020**, *13*, 1266–1274. [Google Scholar]
23) Song, H.M.; Kim, H.R.; Kim, H.K. Intrusion detection system based on the analysis of time intervals of CAN messages for in-vehicle network. In Proceedings of the International Conference on Information Networking (ICOIN), Kota Kinabalu, Malaysia, 13 January 2016; pp. 63–68. [Google Scholar] [CrossRef]

24) Khan, Z.; Chowdhury, M.; Islam, M.; Huang, C.-Y.; Rahman, M. Long Short-Term Memory Neural Networks for False Information Attack Detection in Software-Defined In-Vehicle Network. *arXiv* **2019**, arXiv:1906.10203. [**Google Scholar**]

25) Li, L.; Yu, Y.; Bai, S.; Hou, Y.; Chen, X. An effective two-step intrusion detection approach based on binary classification and -NN. *IEEE Access* **2018**, *6*, 12060–12073. [**Google Scholar**] [**CrossRef**]