Investigating Security Issues in Autonomous Vehicle

Submitted By

Thaker Jay K

20MCEI13

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2022

Investigating Security Issues in Autonomous Vehicle

Major Project

Submitted in partial fulfillment of the requirements

for the degree of

Master of Technology in Computer Science and Engineering

(Information and Network Security)

Submitted By

Thaker Jay K

(20MCEI13)

Guided By

Dr. Sudeep Tanwar

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2022

Certificate

This is to certify that the major project entitled **" Investigating Security Issues in Autonomous Vehicle"** submitted by **Thaker Jay K(20MCEI13)**, towards the partial fulfillment of the requirements for the award of degree of Master of Technology in Computer Science and Engineering of Nirma University, Ahmedabad, is the record of work carried out by him under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for examination. The results embodied in this Major Project Part-II, to the best of my knowledge, haven't been submitted to any other university or institution for award of any degree or diploma.

Dr. Sudeep Tanwar
Internal Guide & Professor
CSE Department
Institute of Technology
Nirma University, Ahmedabad

Dr. Sharada Valiveti
Associate Professor & PG Coordinator (M.Tech-INS)
CSE Department
Institute of Technology
Nirma University, Ahmedabad

Dr. Madhuri Bhavsar
Professor & Head
CSE Department
Institute of Technology
Nirma University, Ahmedabad

Dr. Rajesh N Patel
Director
Institute of Technology
Nirma University, Ahmedabad

Statement of Originality

I, **Thaker Jay K**, **20MCEI13**, give undertaking that the Major Project entitled **"Investigating Security Issues in Autonomous Vehicle"** submitted by me, towards the partial fulfillment of the requirements for the degree of Master of Technology in **Computer Science & Engineering (INS)** of Institute of Technology, Nirma University, Ahmedabad, contains no material that has been awarded for any degree or diploma in any university or school in any territory to the best of my knowledge. It is the original work carried out by me and I give assurance that no attempt of plagiarism has been made.It contains no material that is previously published or written, except where reference has been made. I understand that in the event of any similarity found subsequently with any published work or any dissertation work elsewhere; it will result in severe disciplinary action.

_____

Signature of Student

Date:

Place:

Endorsed by

Dr. Sudeep Tanwar

(Signature of Guide)

Acknowledgements

It gives me immense pleasure in expressing thanks and profound gratitude to **Dr. Sudeep Tanwar**, Professor at Computer Science and Engineering Department, Institute of Technology, Nirma University, Ahmedabad for his valuable guidance and continual encouragement throughout this work. The appreciation and continual support he has imparted has been a great motivation to me in reaching a higher goal. His guidance has triggered and nourished my intellectual maturity that I will benefit from, for a long time to come.

It gives me an immense pleasure to thank **Dr. Madhuri Bhavsar**, Hon'ble Head of Computer Science And Engineering Department, Institute of Technology, Nirma University, Ahmedabad for her kind support and providing basic infrastructure and healthy research environment.

A special thank you is expressed wholeheartedly to **Dr. Rajesh N Patel**, Hon'ble Director, Institute of Technology, Nirma University, Ahmedabad for the unmentionable motivation she has extended throughout course of this work.

It gives me an immense pleasure to thank Mr. Nilesh Jadav and Mr. Rajesh Gupta for their guidance, kind support and providing basic infrastructure and healthy research environment.

I would also thank the Institution, all faculty members of Computer Science and Engineering Department, Nirma University, Ahmedabad for their special attention and suggestions towards the project work.

- Thaker Jay K

20MCEI13

Abstract

Autonomous vehicles (AVs) are a potential technology for improving safety and driving efficiency in intelligent transportation systems (ITSs). AVs are subject to a variety of cyber-attacks, including denial-of-service, spoofing, brute force, and cross-site scripting. To solve the security issues in AV. We proposed an intrusion detection system (IDS) framework for the intelligent classification of malicious and non-malicious attacks. We utilized an ensemble-based machine learning model to efficiently classify attacks. We divided the proposed model as data collection, pre-processing data for the imbalanced tree feature selection, ensemble model, and detection. This model builds the ensemble learning using stacking the model. Finally, we evaluate an ensemble model using different ai matrix accuracy, precision, recall, and f1-score. XGBoost is out-performance in this ensemble model. This proposed model benefits to attain a high detection rate and low computational cost at the same time.

Abbreviations

AV	Autonomous Vehicle
IDS	Intrusion Detection System
IoT	Internet of Things
TCP	Transmission Control Protocol
IP	Internet Protocol
ITS	Intelligent Transportation Systems
DT	Decision Tree
ET	Extra Tree
RF	Random Forest
XGBoost	Extreme Gradient Boosting
DoS	Denial of Service
SAE	The Society of Automotive Engineers
ECU	Electronic Control Unit
SQL	Structured Query Language
XSS	Cross-Site-Scripting
CAN	Controller Area Network
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
KNN	The k-nearest neighbors

List of Figures

List of Tables

Contents

Chapter 1

Introduction

In this chapter, we discuss about an autonomous vehicle (AV) technology. Modern technologies are improvising at a fast speed and moving towards more development of automobile sectors. It has been focused on new technologies which are called AV. These vehicles help to work on the security of travelers and increment the productivity of transportation by interfacing with the outer world through vehicle-to-everyone (V2X) and vehicle-to-vehicle (V2V) interchanges. These technologies are transforming information like speed, position, and current angle which help to predict the location of the vehicle. For, this technology's main motive is driver-less cars and the absence of human operators. The benefit of AV is the accessibility of transportation and reducing accident. Additionally, it also reduces traffic congestion and accident which directly impact society. On the other hand AV has some challenges of privacy, legal, and security. Mainly, security has a big impact on AV. Its direct impact on the confidentiality, integrity, and availability of AV. It has prone to be hacking. Some attackers can hack the car because it has some security miss configuration or lack security standards. An attacker tracks the vehicle and during the communication of vehicles,an attacker modified a message communication which directly leads to risk of AV. Another challenge in AV is fewer job opportunities and non-functional drastic change due to weather and path challenges.

In autonomous vehicle it drives with different six levels of automation. Now here we can look that different levels of automation.

Level 0: No Automation = It is under Manual Control.

Level 1: Driver Assistance= Functions are under control like speed

Level 2: Partial Automation= this human can take control. It is automatically running by software.
Level 3: Conditional Automation= It has environmental detection capability. So based on environment changes automation software run it.
Level 4: High Automation= Automation running under control of driver based on geographic required.
Level 5: Full Automation=It has no human interaction required. Its fully auto matted

1.1 Motivation

Since many years automobile sector has been improving a day by day. In term of AV, the security challenges day by day increases. AV achieving and solving a new challenges from automation zero to automation level five. There are many technologies which have some solution related to challenges of AV. Among them, Machine learning (ML) is kind of technology which helps to more suitable to predict my model. ML provides an algorithm to learn from it and based on that predict a model. ML is wide area for new research and daily updates. They have follow some models based on that they give estimated results[8].

In AV, taking example of innovation from 2004 year to 2022 year. AV achieving and solving a new challenges from automation zero to automation level five. In this graph 4.1, it represents in 2004 institute research for AV. After two years, AV technology comes into the cruise system which is come into the picture related sensor and other transportation way [19]. It has also one new innovation which is related to sea land technology for under water related self-driving vehicles. In the year of 2008 nearly, one new innovation related autonomous has been proposed related sport information and lane system. This technology has very helpful to self-driving car which is facing issues related path and information collecting challenge. After this innovation one challenge is related to collision support and break warning for pedestrian's safety and traffic signal following it by inner connectivity or software. In the year of 2014 nearly, it has found new innovation related to hybrid mode of self-driving car. This is huge demandable and helpful for human and machine interaction. After nearly three years fully automated and advance phase devel-

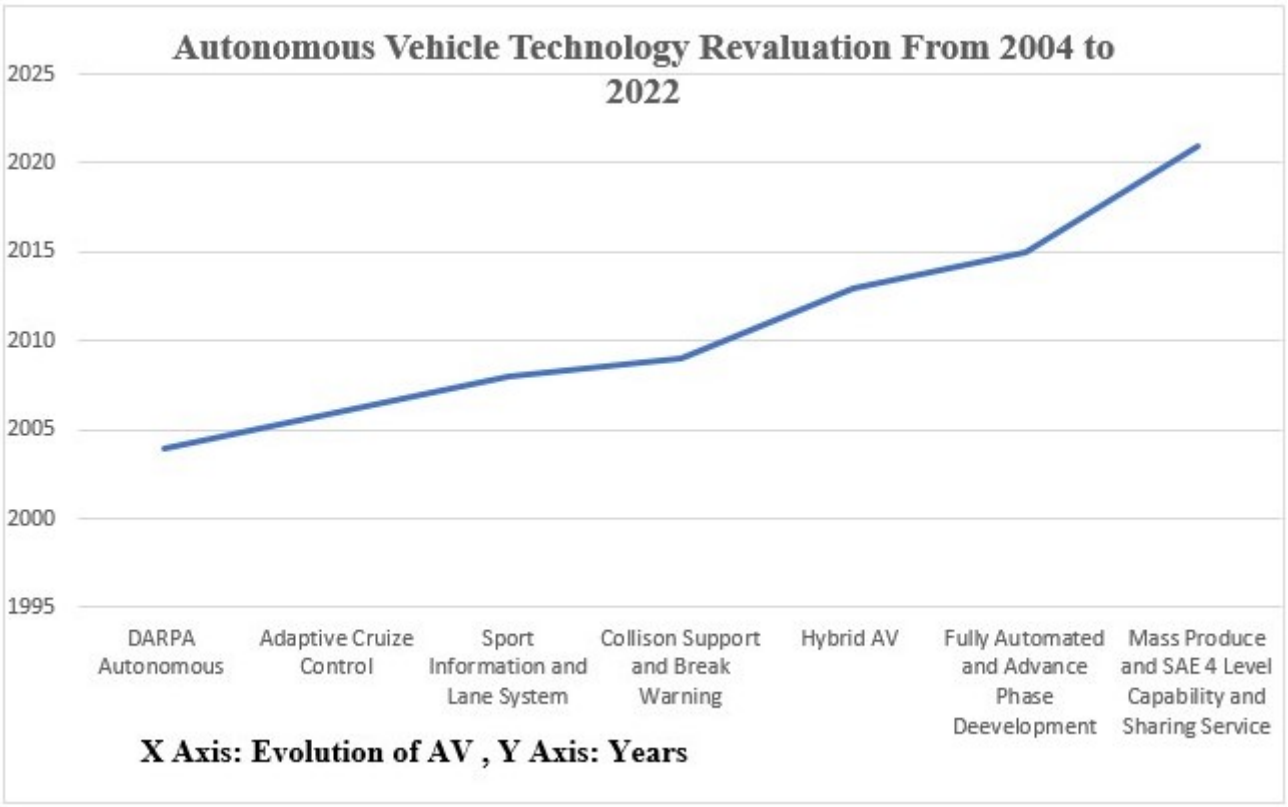

Figure 1.1: Advancement of Autonomous Vehicle

opment comes into the picture of technology of self- driving vehicle software. Currently ongoing research innovation related mass produce and level 4 capability of sharing information service. This research is ongoing for passenger identified and capacity of seating on vehicle without human interaction.

AV technologies enable communication to network devices, and internet of things (IoT) devices using sensors to other hardware. AVs are susceptible to community threats with excessive results due to the fact attacking automobiles on the street direction a dangerous hazard to human lives. Because of that, some potential impact on AV. Denial of Service (DoS) that sends multiple packets at the same time to sensor nodes and takes control of the vehicle. Due to that, an electronic control unit (ECU) has been stopped processing requests and accident occurs in AV [15]. Additionally, attackers perform spoofing attacks then they do masking and provide fake information related to geolocation and due to that communication is disturbed. Additionally, a sniffing attack is a type of attack where the attacker scans an open port and get the confidential information related to vehicle systems. This impact directly leads on ECU about confidentiality, integrity, and availability (CIA). Moreover, intruders try to access server communication by bypass-

ing structured query language (SQL) injection and cross-site scripting(XSS) attacks. In this all attacks can be mitigated using gateway or firewall which provides a security of AV.

This all attacks from external threats communications. There are also inter-vehicle communication attacks. As a communication between ECUs, a bus protocol for controller area network (CAN) is deployed. CAN reduces equipment expense, mass, and redundancy while delivering an error detection approach for stable output. The CAN Bus interconnect all of the ECUs, making them vulnerable to numerous attack if any one of them is compromised. An intruder can inject malicious script into the driving system through CAN bus connectivity, take control of the driving system, and cause an accident. Appropriated assaults can be sent off on the CAN transport in a specific technique they can be launched from outside networks to occupy resources in other ways and deliver harmful information such as driving statistics. Additionally, fuzzy attack where an attacker passes by injecting an arbitrary message and interrupted a vehicle communication.

Chapter 2

Literature Survey

2.1 Related Work

AV is the next upgraded technology that runs on smart networks. AV transforms information from one sensor to another. During this communication, there are some gaps related to security constraints. To solve those gaps some research community has provided solutions across the world. For example, Gao *et al.* [16] whose main objective is sensors and actuators. This research was carried out in the year 2019. The proposed model with flow-based in AV. The advantages of this model, it used low computational resources and required periodic checks because it collects information from node to node. Another research was carried out in the year 2015 using the flow-based method by Taylor *et al.*[20]. This detection method works with message frequency and based on that they proposed one detection method. For this method demerit is it works on periodically signals.

Seo *et al.* [16], this author used generative adversarial nets. In this proposed researcher suggested a pattern of CAN ID So the attacker can detect based on the known attack only. But for this purpose, it required expensive hardware like a detection machine for CAN ID and hardware simulation and all. Li *et al.* [11], this researcher proposed an adaptive network-based fuzzy inference system. The research's main objectives are busload, and message frequency analysis. This proposed system has detect attack type and have a simple solution. However, it works with simple attacks detection and updates each second so that needs a feature database for this detection method.

Table 2.1: Comparison of existing solution

Author	**Year**	**Objectives**	**Pros**	**Cons**	**Method**
Guo *et al.* [16]	2018	Sensor and Actuators	Low computation resource	Required periodically check	Flow Based
Seo *et al.* [16]	2018	Pattern of CAN ID	CAN train itself for known attack	Expensive hardware	Generative adversarial Nets
Li *et al.* [11]	2017	Busload, message frequency analysis	Detect attack type, simple solution	Works for simple attacks, updated each second, needs a feature database	adaptive Network-based Fuzzy Inference System
Muter *et al.* [13]	2011	Entropy of IDs, payload	Does not require much information about traffic data	Very vulnerable to some attacks which include random bits	Entropy-based
Taylor *et al.* [21]	2016	Payload	Does not require pre-knowledge	Does not understand the natural change	Long Short-term Memory
Larson *et al.* [9]	2008	Protocol policy	Less dependency	IDS should be placed at every ECU	Specification based
Stabili *et al.* [18]	2017	Payload	Low computation	Low detection	Hamming Distance
Lee *et al.* [10]	2017	Remote frame timing	Simple efficient algorithm with low-cost hardware	Increased traffic	Offset ratio and time interval
Marchetti *et al.* [12]	2017	Sequence of ID	Low memory and computation requirement,detection of inserted few malicious messages	very vulnerable to attacks which have a similar sequence of normal traffic	Analysis of ID Sequence
Choi *et al.* [4]	2018	Electrical signal	Robust to some attack types, first IDS to differentiate between an error and an attack	High cost and vulnerable to environmental changes	Support Vector Machine and Boosted Decision Tree
Cho *et al.* [3]	2016	Clock skew	Robust to some attack types,	Only works on periodic signals	Recursive Least Squares
Groza *et al.* [6]	2018	Message identifier, payload	Low memory usage for membership testing	Complex algorithm	Bloom Filtering
Hamada *et al.* [7]	2018	Reception cycle period (frequency analysis)	Online learning	Hard to authenticate a non-periodic message	Probability Density Function
Taylor *et al.* [20]	2015	Message frequency	Simple algorithm	Only works on periodic signals	Flow-based
Thaker *et al.*	2022	Inter communication and External Packet length	It detect an abnormal behavior and predict an alert	-	Tree based ensemble Learning

Additionally, some researchers did the research objective based on payloads and proposed an algorithm for the detection of attack in AV. Muter *et al.* [13], has proposed entropy-based learning using ML. An author focused on entropy-based ID attack detection. For this benefit has been no required information on traffic and other data. However, it has possible some random bits are used in the attack and are unable to track those attacks. Moreover, researcher used entropy-based analysis either 0 or 1 based on that predict attacks in AV and for that does not need a lot of data about traffic information. However, his research it possible to crack and pass the system because of random bits.

In the year of 2017, Stabili *et al.* [18] used hamming distance-based model which has used payloads and the benefits are low computation resources required to detect attacks. But it founds that using hamming distance possibility of detection ratio has been less. Moreover, Choi *et al.* [4], this researcher used a method of support vector machine(SVM) and boosted the DT. This researcher used an objective as electrical signals. On the flip side, the high cost will be vulnerable and environmental changes.

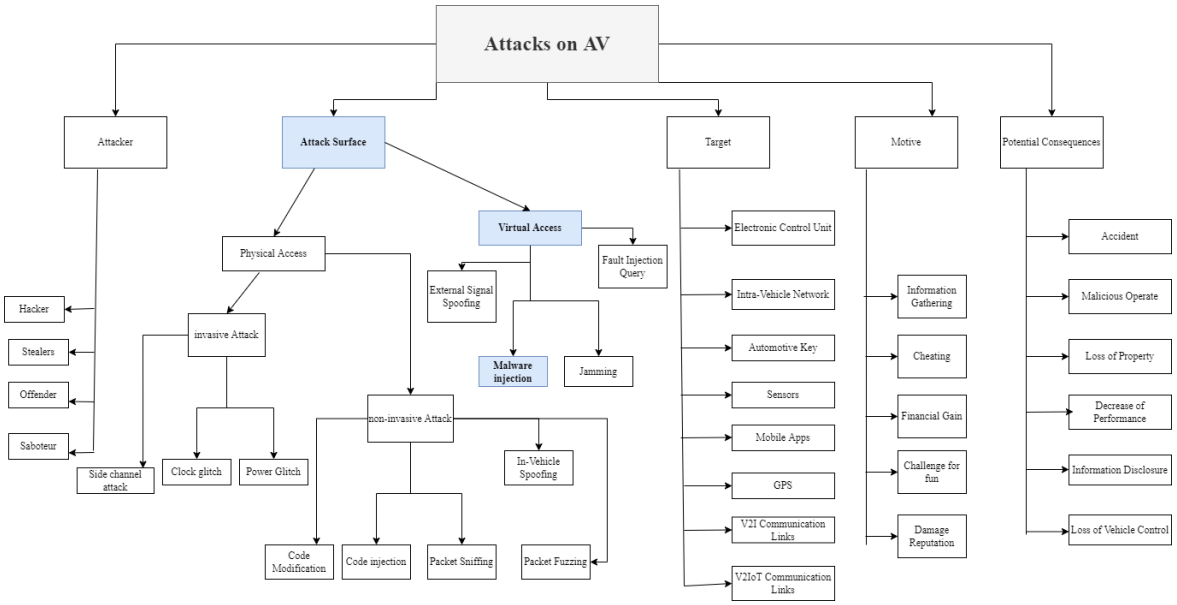

Figure 2.1: Taxonomy of attack in AV

- **In Vehicle Network** - This system is working with a connected vehicle communication connectivity. This system is usually carried out with multiple channel communication through a sensors. In 2007 Hoppe and Dittman is trying to experimental review on CAN bus and trying to do a sniffing or replay attack. As experimental review they shown result of impact is bus communication got chocked and accident or miss configuration found in this system. This system basically kind of channel based attack[1].

- **Malware Attack** - Malware attack is kind of security flow based vulnerability class. In this attacker is exploit vehicle security using some connected instrument of vehicle which point out some malicious operation occurred through transmission of radio or bluetooth. For example, sometimes attacker try to connect device and passes malicious code which will direct connected with vehicle. To impact of that

Table 2.2: Security threats in Autonomous Vehicle

Name	**Vulnerability Class**	**Description**	**Countermeasures Mechanism**
Man-In-The–Middle Attack	Communication Exchange Vulnerability	Insider and Attack Monitoring	Encipherment (Data integrity, Confidentiality)
Malware Attack	Security Flow Based Vulnerability	Using some connected device attacked do a gain access of it. Operate wirelessly malicious things	H-MAC and Digital Signature (Availability and Authentication)
Denial of Service Attack	Security Flow Based Vulnerability	Increase traffic of message and delaying to message transmission which issues with availability on right	Traffic Padding and Digital Signature (Availability,Authentication)
Ransomware Attack	Security Flow Based Vulnerability	To lock a car or hijack a data and ask for smart connectivity	Digital Signature and Encipherment (Data Confidentiality Authenticity)
Spoofing Attack	Security Flow Based Vulnerability	Its kind of linking attack which attacker tempering data.	Digital Signature and Encipherment (Data Confidentiality, Authentication)
Sybil Attack	Sensor Based Attack Vulnerability	Its part of Jammed a network.	Digital Signature (Authentication)
Sensor In-personation	Sensor Based Vulnerability	Hardware flaws of speed control work. It disclose personal information.	Digital Signature(Authentication and Gateway)
Wrong Information	Channel based Vulnerability	Network flooding with wrong message passing it.	H-Mac and Digital Signature (Authentication and Integrity)
Masquerading	Security Based Vulnerability	To Try to gain a unauthorized access by some user identity information	Digital Signature (Authentication)
Blackhole	Channel Based Vulnerability	To try router replay packet instead of discards.	Routing Control (Availability)
Spamming	Security Flow Based Vulnerability	Malicious and Insider	Routing Control (Availability)
Timing Attack	Channel Based Vulnerability	It create a side channel to analysing a time of execution system.	H-Mac and Digital Signature (Data Integrity and Authenticity)
GPS Spoofing	Channel Based Vulnerability	Its inaccurate information passing through channel by attacker	Digital Signature (Authentication)
Wormhole turnelling	Channel Based Vulnerability	Outsider, Malicious and Monitoring Task	Digital Signature and Encipherment,(Authentication, Confidentiality)
Illusion Attack	Sensor Based Vulnerability	Insider and Malicious operation passing to other node.	Digital Signature (Authentication)
Impersonation	Security based flow Vulnerability	Some third un-known person use someone identity to communication with other	Digital Signature (Authentication)

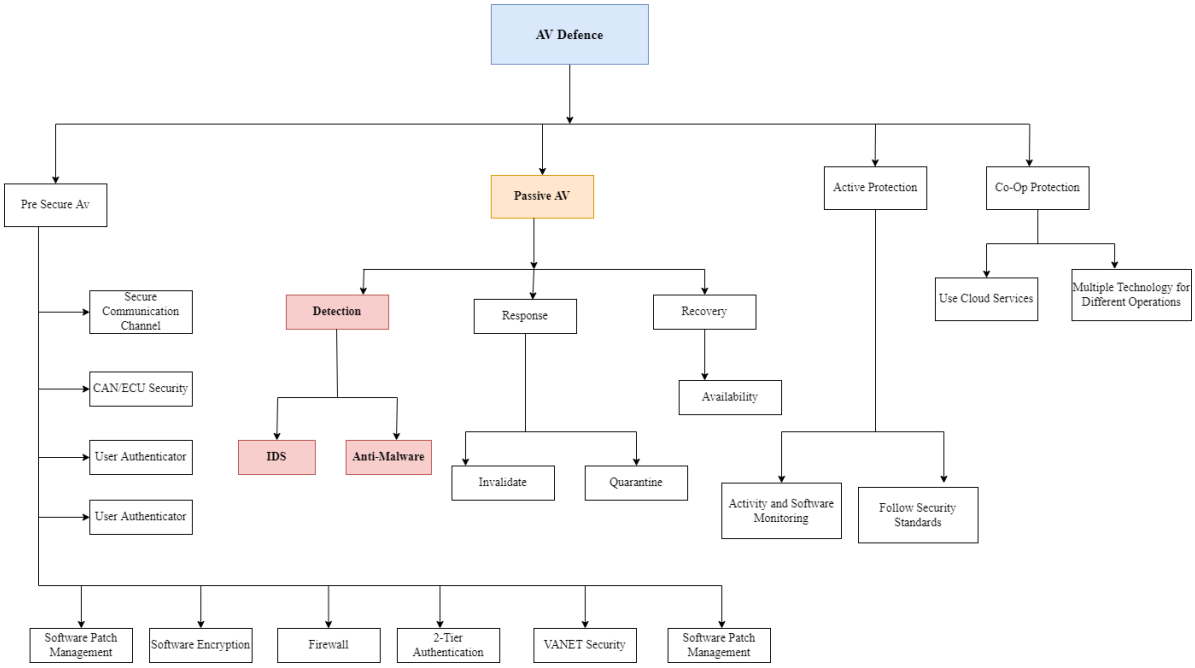

Figure 2.2: Taxonomy of defence in AV

it creates hazards of system to drive AV. To prevent this attack, we required to focus on availability and authentication so that required mechanism are h-mac and digital signature.

- **Man-In-The Middle Attack** - AV is totally connected and communication transferring through wireless network connectivity. Sometimes during message transmission attacker is trying to do insider attack. In this vulnerability class is communication exchange vulnerability. In that need to focus on data integrity and confidentiality. To prevent this attack we required some mechanism of encryption techniques[24].

- **Offensive of Service Attack**- DoS is an attack where attacker start attack over network which stop or unavailable required service so that it feels that wastage of time and due to the timing operation misleading accident or dangerous situations occurred it. This attack is security flow based attack. It required a system which provides an availability and authentication so that digital signature and traffic padding fulfill the system security.

- **Ransomware Attack**- Ransomware attack is demandable ongoing attack where attacker do malicious operation either hijacked machine or encrypt some data and asking for some ransom. Once ransom given then might be possible to decrypt it

or not. So that it is part of security concerns to protect internal function of system in autonomous vehicle technology. This attack vulnerability class is security flow based vulnerability. To protect from this attack required digital signature and encipher techniques for data security so that it is not tampering[8].

- **Spoofing Attack**- Spoofing attack is kind of communication between two cars once communication established between that it start modifying data and sending other information which is not important and forgery of link. So that it is also known as linking attack. This attack vulnerability class is security flow based attack. To protect from attack required digital signature and encryption techniques.

- **Sybil Attack**- It is sensor based attack. In this attack jamming the network communication which is directly proposed to impact of accident or malicious operation occurred. This attack vulnerability is sensor based flow attack. To prevent this attack required strong authentication digital signature.

- **GPS Spoofing Attack**- In this attack attacker is trying to gain access and then passing an inappropriate information of location or routing or map which is leads to put system decision confusion. This attack is channel based attack. To required solve this attack strong routing mechanism and digital signature to verifying trustable source.

- **Impersonation attack**- In this attack someone using their identity to driving a software and sometimes to do malicious operation as another user which put system into difficulty. This attack vulnerability class is security based flow vulnerability. To protect this attack required digital signature so third party software malicious operation identified it.

- **Fake Authentication Attack**- In this attack someone using their fake remote authentication or trying to gain some activity using bypassing their IP and other permissions. In this attack they try to do decrypt the password pattern or by passing fake internet protocol(IP) request system got confused during that they can take over control and do remote code execution. Thus, it impact of malicious operation can be transferred on system host.

- **Network Protocol Attack**- There are multiple protocols are using for interconnection and central node connectivity and for transformation. transmission control protocol (TCP)/ IP and open systems interconnection (OSI) model have also some vulnerability which have already exposed. In between connection of two layers there is possible to data breach by some network protocl (NP) attack for AV.

- **Rogue Updates**- Rogue updates is basically one kind of attack which directly to miss lead of autonomous vehicle system patch. It kind of remote code execution attack which leads to major accident. Rogue basically kind of unknown malicious software auto installation.

- **Password and Key Attack**- In this attack someone cracked or thief your password by doing phishing or passive attack which points to lead miss security configuration in a system. Thus it is necessary to focus on this system.

- **Malicious Email Bomb Attack**- In this attack scenario attacker send a email to track a software or user activity by just clicking this email they send malicious script which execute on host machine and it transfer to information about open ports and other miss lead information. Thus attacker easily gain access and do malicious operation.

- **Blackhole Attack**- In this attack attacker send multiple request packets of different or same size which carried out potential impact of dropping communication to other nodes.

- **Falsified-Information**- In this attack attacker send false request and messages to sensor nodes which carried out to accident and miss configuration in security. Thus, it creates a dangerous situation in auto pilot mode.

- **Timing Attack**- Time synchronization is a critical part of keen associated vehicles. Vehicles move all through networks quickly, which presents the requirement for constant updates and data trade between both ECU and vehicles.

The figure 2.1 defines a taxonomy of attack in AV .Moreover, the figure 2.2 showing a taxonomy of defence in AV. In the AV attacks showing different types of category in the

figure 2.1. Thus, this all security concerns are possible to secure using ML Algorithm. ML can predict a suitable model with accuracy which defines a solution of this AV software. To overcome the aforementioned issues, we have proposed an intrusion detection system (IDS) defined on ensemble learning. We divided the total process into five steps. The initial stage is to gather particular data on the network's two absorbing states: normal and malicious, which are created by various forms of attacks. The data was obtained by packet sniffers, but they have network properties that are ideal for IDS development. Firstly, we collect data, and then secondly apply for the pre-processing task using min-max normalization. Additionally, for class oversampling we used synthetic minority oversampling technique (SMOTE). Then based on the output we proceed with feature selection to reduce an input variable and collect only relevant data for the machine learning model. After the feature selection, we make a classification of the tree for the extra tree(ET),decision tree(DT), Extreme Gradient Boosting (XG Boost), and random forest(RF). After that base model has been input of ensemble model which is using stack as meta classifier. Ensemble model is the second level of the learner. The stack for better performance in IDS. Finally, we received detection result of model where xgboost suitable to our model.

Chapter 3

Problem Formulation

3.1 Problem Formulation

In this section, we formulate the problem of AV by considering a proposed framework. Its primary concern is to secure vehicle communication of AV. There are many facts that the AV system is vulnerable through intra communication and external communication. For instance, there are $(X1, X2, X3....Xn) \in X = AV$. The communication of AV, there are several X who participated in the communication channel. Suppose, when X1 wants to share the information with X2 then they pass information in the form of Data (D) $X1 \xrightarrow{D} X2$. Where,$Xa \notin X$. D can be possibly manipulated by an attacker (Xa). Xa can be modified as data and passed to the CAN Bus which directly leads to the security risk of AV. Due to false information of data like passing wrong geolocation or wrong information of path it affects to wrong information. This false information directly leads to the decrease performance of AV. In the equation 3.1 D is data and based on the equation we are securing maximum data.

$$MAX \sum_{i=0}^{n} Security(D) \tag{3.1}$$

In AV to prevent an attack and improve performance, we need to secure messages during communication from one to another Vehicle. For that purpose, we proposed a ensemble-based IDS approach that plays the role of a security guard.

3.2 Intrusion Detection System Overview and Architecture

In this system, overview IDS offers security for each intra and outer communication, the recommended IDS is applied in many places in the AV. The IDS might be situated at the highest point of the CAN transport to handle each sent message and guarantee the hubs are not compromised to recognize dangers and settle it[14]. This IDS is also configured inside the gateway to ensure complete security from the outside to the inside communication Network Nodes. The following diagram Figure 3.1 depicts the structure of IDS implementation on automotive systems.

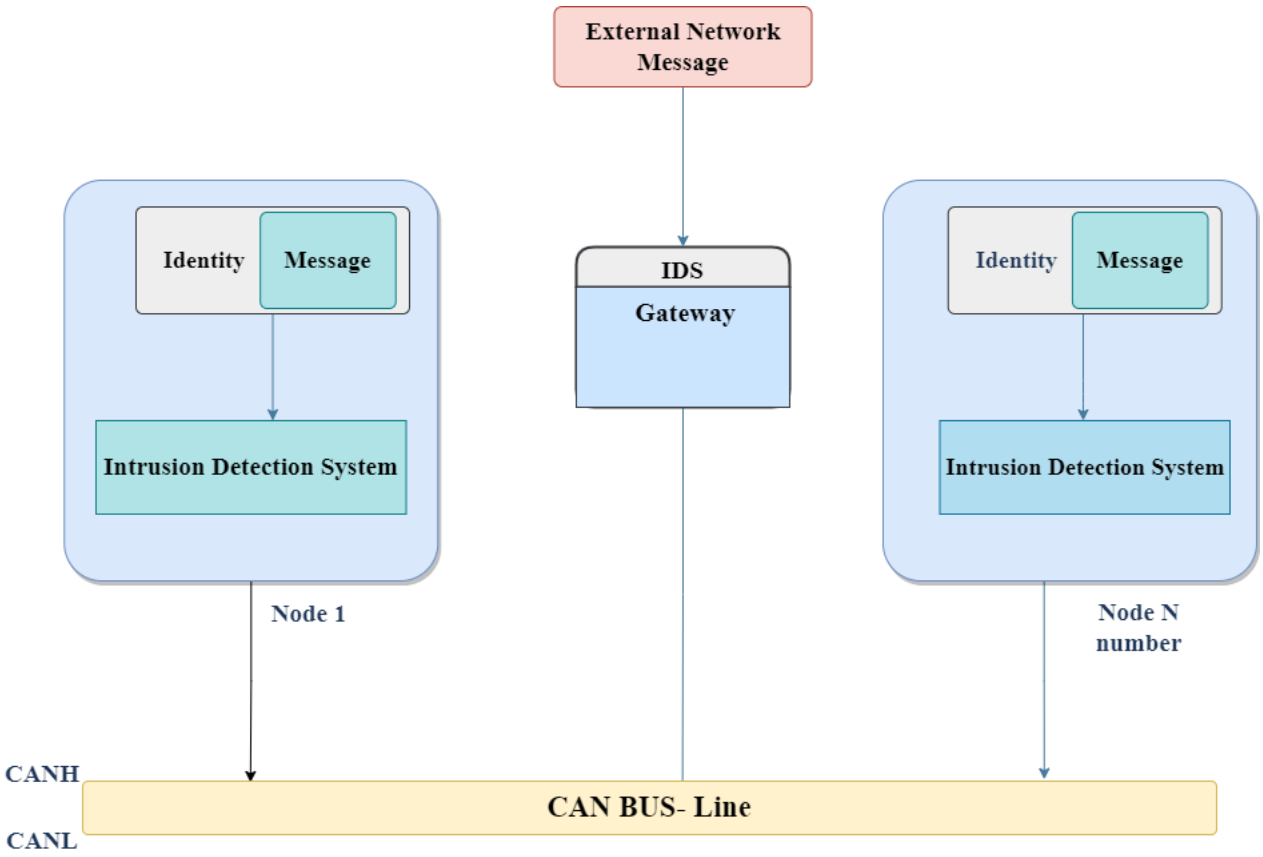

Figure 3.1: IDS in AV

This figure 3.1 IDS diagram shows how IDS can be helpful to protect the communication of intra and external communication in the CAN Bus protocol. In this diagram, Identity and Message is a subset of ID and Data which are passed to IDS and IDS passed to Signal (S) to check when the sign line of the CAN transport changes from CAN

High(CANH) to CAN Low (CANL). CANH and CANL are two nodes of AV. Between two nodes' communication, an external network message is entered using a gateway so it prevents DoS or other types of attacks using this IDS. When a communication is sent from the outside world to the intranet, it is routed through the gateway's IDS and checked.

Chapter 4

Proposed IDS Approach

4.1 Proposed IDS Framework

In this proposed framework, we divided it into five stages. CAN Bus which transforms information from one vehicle to another. It has five stages of this proposed IDS. The first need is to collect data related to intra-external network dataset. After the collected dataset next to pre-processing data. In the second stage, it has been if the class is balanced then moving towards an oversampling using smote. After the processing of data, the feature selection used for relevant input data collected and getting rid of noise data. Once the feature selection has been completed next step of building a model. It has used classification for tree-based learning and ensemble the model used to increase accuracy of models using the stacking method. This model helps to solve classification problems in ML.

The initial step in creating an IDS is to gather particular data on a network of two absorbing states: normal and abnormal, which are created by various forms of threats. The information was gotten by parcel sniffers, however they have network properties that are great for IDS advancement. To begin, an IDS for CAN theft must be implemented before the main CAN bus inject message and data are collected and attacks are made on CAN IDs and that information at any point field of edges.

In terms of external communication with a network, it is required for a general network and needs various regular network vulnerabilities to collect data with more effective IDS to detect different types of attacks.The most successive organization attributes ought

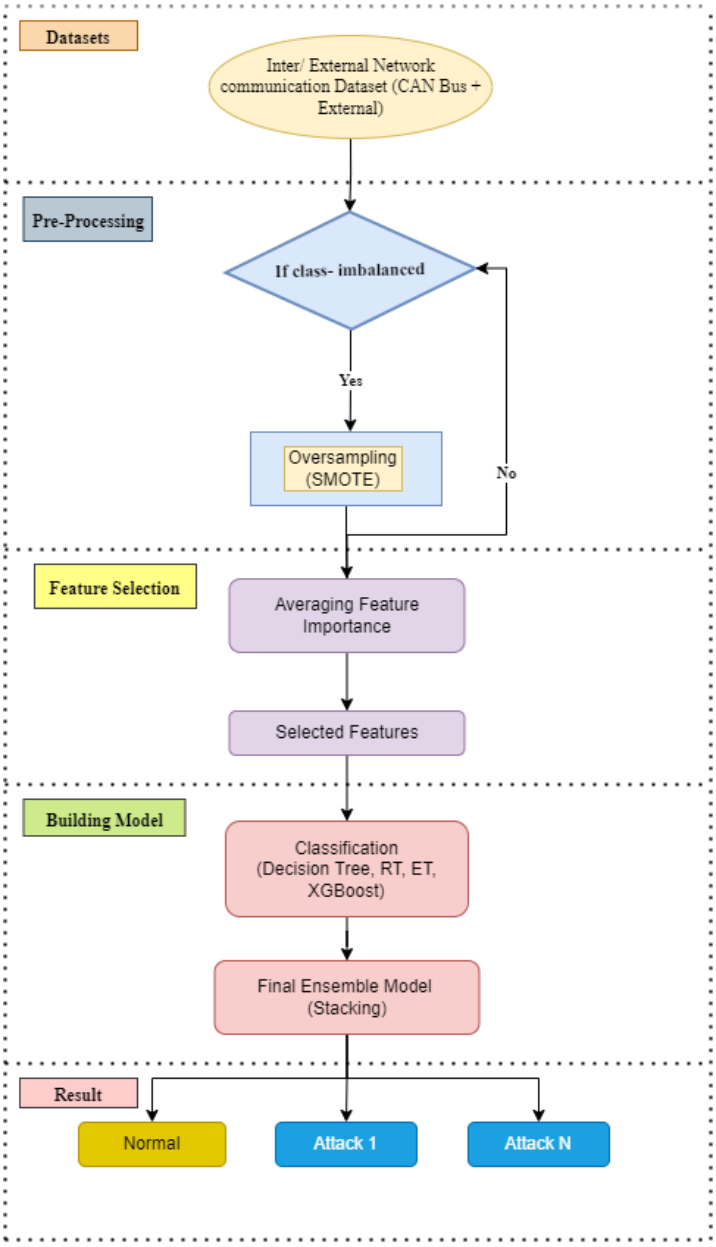

Figure 4.1: Phases of machine learning model for AV

to be thought of, including data length, information move rate, throughput, between appearance time, TCP and counts, section size, and dynamic/inactive period. However, due to the high dimensionality of the data, the computational intricacy of the proposed IDS might increment. As a result, further feature analysis for the external network data is required. After a few steps, the data would be processed to make it more suitable for IDS creation. First of all, the information can be encoded utilizing a one-hot-vector since it has a limit that guides in the separation of ordinary and odd information. But,normalized data is frequently more efficient for ML training. As a result, the range of numerical values for each feature is set to zero to one, and each value after normalization Zn is written as:

$$Zn = \frac{Z–A}{(B-A)'} \tag{4.1}$$

In the above equation, Z is the original value, and B and A are the maximum and minimum values of each feature. Furthermore, network data is frequently class imbalanced since, in actual life, networks are mostly in a normal condition, and attack-label examples are frequently insufficient. Irregular oversampling and the SMOTE can be utilized to create additional information in minority classes where there isn't sufficient to cure the issue of class-imbalanced information, which normally prompts a low inconsistency discovery rate [9]. The fundamental technique for irregular oversampling is to just copy the examples to build the quantity of tests in minority classes.

Since the information acquired would be generally particular as opposed to nonexclusive, the arbitrary oversampling strategy can undoubtedly prompt over-fitting. The SMOTE technique, then again, looks at minority classes and makes new examples that help them in light of K's nearest neighbors (KNN). Subsequently, SMOTE can deliver top notch tests and is utilized in the recommended framework for the minority classes.

4.1.1 The Proposed Machine Learning Methodology

Developing the IDS in the proposed system is a multi-classification challenge to identify multiple cyber-attacks, and ML methods are generally used to address such arrangement issues[17][2]. RF, XGBoost, DT, and ET are among the ML algorithms upheld by a tree structure.

The divide and conquer approach is used by DT to classify their members. A DT is comprised of choice hubs and leaf hubs, which mirror a decision test more than one in every one of the highlights, and consequently the outcome class.

RF is a gathering learning classifier in view of the mass democratic rule, which chooses the category with the maximum number of votes from DT as the classification result. ET, a similar ensemble model, supported a number of randomised DT produced by handling various subsets of the informational collection. XGBoost, on the other hand, might be an ensemble learning algorithm that uses the gradient descent approach to combine several DTs to increase speed and performance.

Other models for classification issues, such as KNN and SVM are prevalent in addition to

the given techniques[1]. For model selection, the computational complexity of common supervised ML algorithms is calculated. Assuming the number of coaching instances is X, the amount of features is Y, and therefore the number of trees is Z, we've got the subsequent approximations. The complexity of DT is O(X^2Y) while the complexity of RF is O(X^2YZ). Additionally, ET and XGBoost have an analogous complexity of O(XYZ). On the opposite hand, KNN's complexity is O(XY), and therefore the SVM algorithm's complexity is O(X^2Y) [5].The suggested models, RF,DT, XGBoost, and ET, enable multi-threading to prevent wasting training time, unlike KNN and SVM. The temporal complexity of RF, ET, DT, and XGBoost drops to O(X^2YZL), O($XYZL$), O(X^2YL), and O($XYZL$), respectively [23]. When the almost number of participating threads of a computer is L. As a result, while the initial time complexity of the algorithms evaluated is identical, the four tree structure techniques take less time to compute because of multi-threading, that is a reason to key argument for selecting these algorithms. The secondary reason behind select algorithms is that most tree structure machine learning models use ensemble learning, which means they commonly outperform single models like KNN. They need the capacity to deal with non-direct and high layered information that the proposed network information has a place with it. The component significance estimations are finished during the structure cycle of these models, which is helpful while performing highlight choice.

It is critical that there are some hyper-boundaries of the proposed algorithm that require be tuned to accomplish better execution. For the DT calculation, the split measure work is set to be Gini index, and furthermore the characterization and relapse trees (CART) model is then constructed, which shows preferred execution over utilizing data gain hypothesis to make an ID3 tree. Accepting that A means the arrangement of all sub-trees, CART chooses the tree in A that limits.

$$T(A) = M'n(A) + \alpha|A| \tag{4.2}$$

In the above equation Where $|A|$ is the cardinality of the tree, is a constant and $M'n(A)$ is the empirical risk using the tree S. Since the further tree has more sub-trees, tree profundity H is a significant boundary of the CART calculation.The amount of decision trees available. Since RF and ET's discoveries depend on the greater part vote of numerous

choice trees, Z is another basic boundary that impacts their presentation. Z can likewise be changed in XGBoost, which depends on a tree group. The regularized objective capacity is limited involving XGBoost specifically[22].

Algorithm 1 An algorithm of proposed model.

Input: dataset $\in D$
Output: Clasification of normal and malicious data

procedure
$D \leftarrow$ accumulates randomly sample instances from majority classes
$dfs = dfsortindex()$
if D has empty value using Min-Max Normalization **then**
$D \leftarrow df = df.fillna(0)$
else
no change in D
end if
if $C_0 << C_1$ **then**
processed D $\leftarrow SMOTE(K, C)$
else
no change in D

Feature selection ($fs = \frac{alltreefeatures}{4}$) where fs is average features

$(ImpFeat = 0; ImpFeat <= 0.9; ImpFeat++)$
Classification $= xgb.XGBClassifier().fit(xtrain, ytrain)$
Result= classification (Accuracy, Precision,Recall,F1-Score)

To determine the best values, the method of grid search is used. For accurate, the models are trained using a limited trees and a shallow profundity. Then, with accuracy tested, these two qualities are gradually upgrading until overfitting occurs, which is showed by a decline in inaccuracy. Finally, the trees Z is set to 200 and the tree profundity D is adjusted to 8. Similarly, the least sample split and least sample leaf are set to 8 and 3 respectively after being modified from 1 to 10. Other optimization approaches can be used to further optimize the parameter tuning process.

4.1.2 Stacking and Feature Selection

Stacking, an ensemble learning technique, is used to improve accuracy even more.Stacking is a typical outfit technique wherein the primary layer has a couple of prepared base indi-

cators, the result of which is utilized as the contribution of a meta-learner in the second layer to create a strong classifier[16]. The defined algorithms act as premise models in the primary layer of the stacking outfit approach, and the one of a kind algorithm with the most noteworthy precision among the four base models is picked as the meta-classifier in the subsequent layer.

The stacking work on certainty of the chosen highlights, a gathering highlight determination feature selection procedure is used by calculating the normal element significance records produced by the ML models. It required picked for highlight determination since tree-based calculations compute the significance of each element in light of every tree, and afterward normal the conclusion of the trees to make it more reliable. Moreover, unique conventional component determination strategies, for example, distrubte values are used for framework by setting various boundaries in tree-based techniques to create the persuading highlight significance.

The amount of is 1.0 for the complete component significance. The selected highlights, the elements are positioned with their significance and each component is inserted to the element list from maximum significance to minimum significance till the amount of significance comes to reaches at 0.9. Different highlights with amount of significance under 0.1 will be disposed of to decrease the computational expenses.

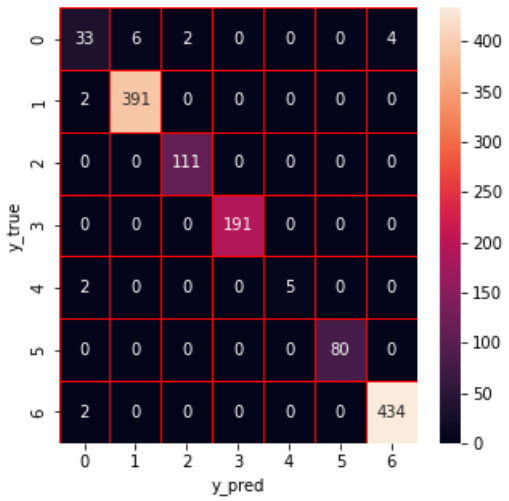

Figure 4.2: Stacking confusion matrix

4.1.3 Validation Metrics

Each researched informational collection is divided into five subgroups, and 5-overlap cross-approval is performed to break down the proposed models. There are some key estimations used to assess the recommended method, and their equations are provided in our methodology. The precision is the level of accurately ordered information. Be that as it may, there might be a class awkwardness in the informational collections, bringing about high typical information grouping precision however low assault discovery rates. Thus, the location rate is determined for assessment, which is the proportion of recognized assault information to add up to strange information.

The selected IDS should have a high detection rate and a low false alarm rate to ensure that the AV does not misbehavior data in order to attain a higher detection rate. Additionally, we defined a overall performance of defined IDS and it also includes their rate of correctness. The execution time, particularly the model training time, is frequently used to provide data on computing performance.

Chapter 5

Result and Performance Evaluation

5.0.1 Simulation System

In this experimental evaluation, we used jupiter notebook and other libraries required to install for running ML approach. In this experimental analysis, we used a core i3 system with 8 GB random access memory. We set up an environment with a notebook system and predicted a model with an implementation.

5.0.2 Dataset Description

In this proposed IDS to evaluate this work, we used a dataset for inter and external communication in AV. In the first, we collected a CAN-intrusion dataset which is used for car hacking proposed new IDS development on CAN bus [24]. We used CICIDS2017 Dataset is considered work that has a label of different attacks. It has a class label as follows DoS, BENIGN, Port-Scan, Brute-Force, Web-Attack, Botnet, Infiltration. In this dataset, there are 78 columns and 56661 rows which include information on forwarding Packet, total backward packet length and flow bytes, flow packets, and other attributes of the dataset which help to identify some security attacks and generate alerts with IDS. Table 5.1 describes a class label of attacks that we can use in our experiments.

Table 5.1: Description of class label in CICICDS2017 Dataset

Class label	Number of instances
DoS	19035
BENIGN	22731
Port-Scan	7946
Brute-Force	2767
Web-Attack	2180
Botnet	1966
Infiltration	36

5.0.3 IDS Performance Analysis

There are some additional libraries installed for ML models. The proposed IDS result of testing using the different algorithms on the CICIDS2017 dataset is already shown in Tables 5.2. In Table 5.2, we generated a sample trained model accuracy on feature selection and we are showing a performance-based analysis of the IDS dataset using ensemble model likewise, DT, RF, ET and XGboost training and prediction. Based on this decision tree and others we found prediction accuracy and trained the model for IDS using the CICIDS2017 dataset.

Table 5.2: Performance analysis result of IDS

Method Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-score
Decision Tree	98.25	98.19	98.25	0.9818
Random Forest	98.41	98.38	98.41	0.9837
Extra trees	98.09	98.04	98.09	0.9803
XGboost	98.57	98.52	98.57	0.9852
KNN [8]	97.40	96.26	96.30	0.9675
SVM [8]	96.50	95.30	95.35	0.9780
Stacking	98.25	98.19	98.25	0.9818

The proposed IDS has higher accuracy compared to SVM and KNN. It has consumed less time and given better results using the ensemble learning IDS approach. After combining all four tree-based models and building a stacking model we received accuracy. Compared to DT, RF has more accurate results Because DT has been following a single-based tree. After training the model all algorithms received a higher result accuracy. So that we can say that instead of KNN and SVM. Ensemble learning algorithms give good results for IDS. It founds that XGBoost received the highest accuracy for the detection

of attacks in this model. Because XGBoost always gives more importance to functional space to reduce the cost of the model.

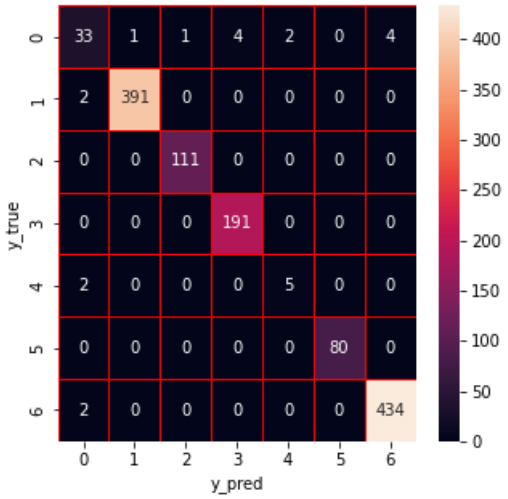

Figure 5.1: Decision tree confusion matrix after feature selection

The figure 5.1 shows a confusion matrix of ai proposed model. It shown the result of True positive class of XGBoost. On the Y-axis confusion matrix has the actual values whereas X-axis has been predicted values of class. This confusion matrix is a way of number of tabulating the number of missclassifications.

The figure 5.2 for shows the different class classifier in term of accuracy, precision, recall and f1-score. From the graph it can be seen that XGBoost has outperforms AI model. This is because XGBoost has gives more accurate result and focus on importance of functional space. In our experiment train model multiple time it gives highest result of classification using XGBoost.During ensemble learning XGBoost gives a good result for classification of this attacks.

Precision

The precision which meaning state which being correct classified it. The ratio of True Positive(TP) class with respect to true corrected class(TP) and false labeled class(FP).This precision shows that how much correctness of our experimental analysis. Precision helps to identifying reliability of ML model to classify the positive.

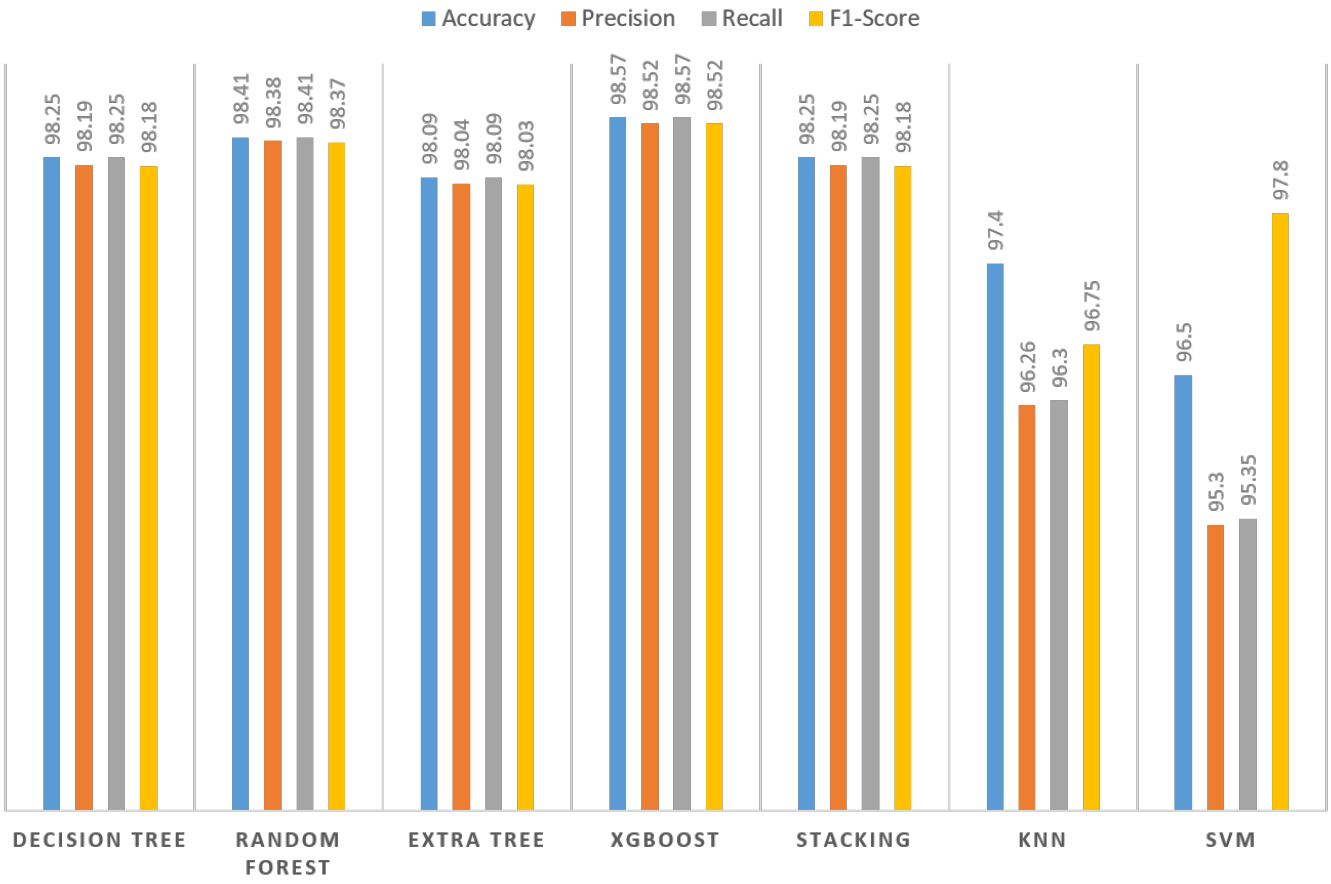

Figure 5.2: Comparison of different algorithms

Precision is also called Positive Predictive Value.

Precision and Recall are performance matrix used in classification and recognition of pattern.

$$Precision = \frac{TP}{(TP + FP)} \tag{5.1}$$

Recall

The recall is defined as the ratio of true corrected class with respect to true collected class(TP) and false negative class(FN). Recall is defined as R.

$$Recall = \frac{TP}{(TP + FN)} \tag{5.2}$$

Accuracy

Accuracy is showing how much the algorithm is accurate with our desired output as a true positive. Accuracy helps to identifying a most suitable model based on the input or training data. Accuracy counts as ratio of correct predictions with respect to all prediction class.

$$Accuracy = \frac{TP + TN}{(TP + FN + TN + FP)} \tag{5.3}$$

F1 Score

F1 score which defines as the harmonic mean of precision and recall of the objects. The F1 score is used to compare the performance of two classifiers. This value range from 0(bad) to 1 (good).

$$F1Score = \frac{2 \times Precision \times Recall}{(Precision + Recall)} \tag{5.4}$$

Table 5.2 describes the result of the prediction and training model result where it founds that accuracy and Precision, Recall, and F-1 Score in this model. The XGBoost founds an accuracy of 0.9857, precision is 0.9852, recall is 0.9857 and F1 score is 0.9852. As a result, XGBoost were chosen for the proposed stacking approach, with XGBoost serving as the stacking model's meta-classifier. Because, according to the experimental data, XGBoost produces a more accurate outcome in the stacking model. Stacking takes longer time to executes but it gives a more accurate result.

5.0.4 Feature Analysis

In this step, feature analysis has been start after a prediction and training model, we need to do a selection method based on test subsets for each of the attacks. In the feature selection, we did important features and respectively their weight of the attacks provided in Table 5.3. We described a result of the weight list of feature analysis. To calculate the

average features (Avg. Feature).

$$AverageFeature = \frac{DTFeature + RFFeature + ETFeature + XGBFeature}{4} \quad (5.5)$$

Table 5.3: Features and Packet weight of attack Method

Attack Method Name	**Attack Features**	**Packet Weight**
Port-Scan	Length of Forward Packet	0.3020
	Normal Weight of Packet	0.1034
	Client Server Flag Count	0.1019
Brute-force	Receiver Destination Port	0.3725
	Minimum Forward Packet Length	0.1020
	Variation in packet length	0.0859
Web-Attack	Init Win bytes backward	0.2463
	Packet Size Average	0.1650
	Destination Port	0.0610
Botnet	Port of Destination	0.2340
	Bwd Packet Length Mean	0.1230
	Avg Bwd Segment Size	0.1140
Infiltration	Forward Packet Length in Total	0.2295
	Forward Bytes of Subflow	0.1340
	Port of Destination	0.1150
DoS	Return Packet Length Std	0.1750
	Average Packet Dimensions	0.1230
	Port of Destination	0.0783

The figure 5.3, it has shown a comparison of macro average and weighted average. It has compared precision, recall, and f1 score weight and the macro average for all algorithms. It has a clear vision that the XGBoost algorithm gives a good an overall result. To ensemble model we used XGBoost. XGBoost has a used a reducing the cost of the model thus it prefect suitable after feature selection and we received highest macro average and weighted average in this model.

Table 5.3, describes an attack name, features of the attacks method, and weight. In this threats, the weight of the packet is another way of important in the intrusion detection method. For example, the average packet size defines a distributed attacks. In

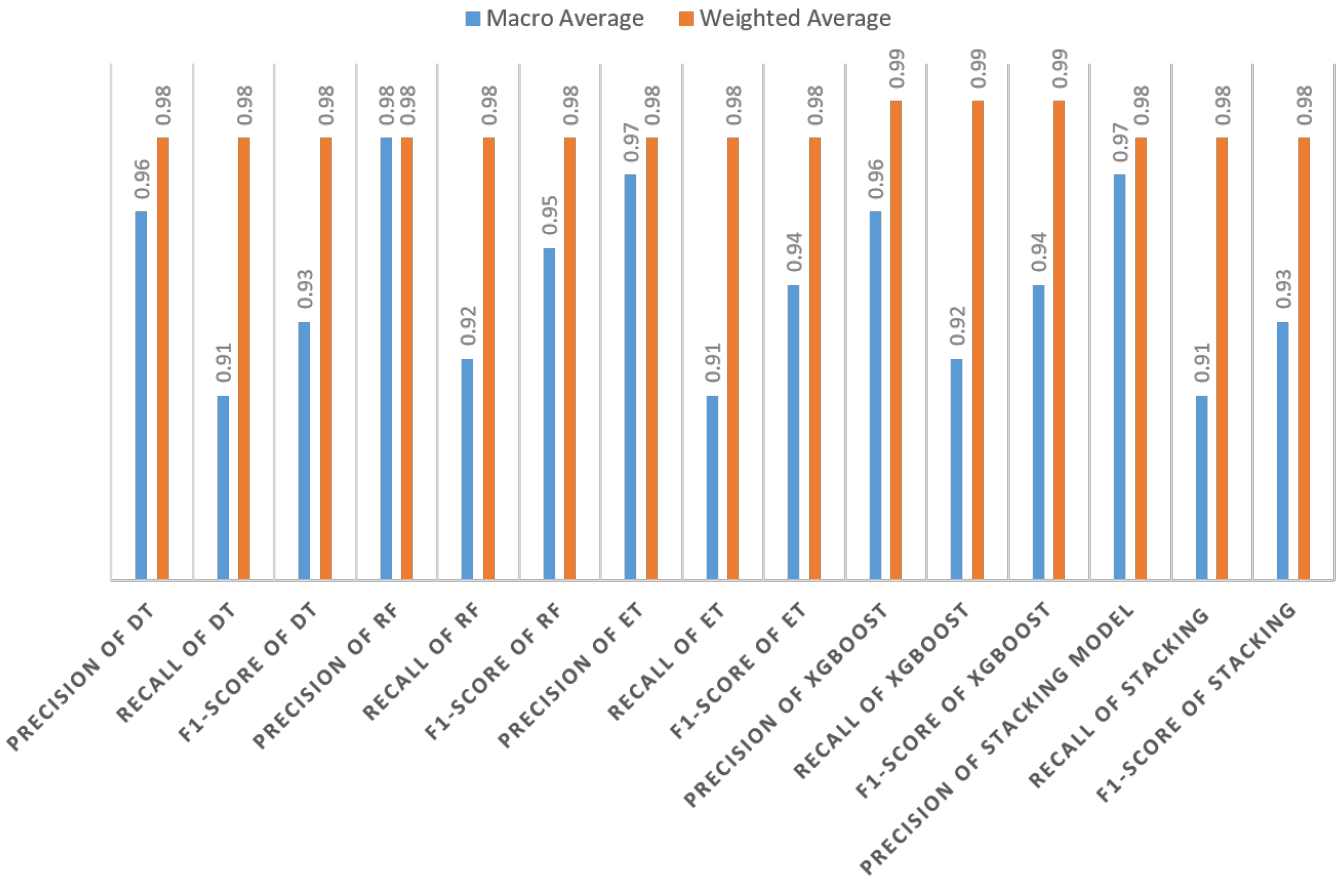

Figure 5.3: Comparison of macro average and weighted average

DoS attack, the packet size is the same input to the AV system and the output will be the same as per the following terms. In the method of port scan length of the forward packet is much higher one time after the receiving of acknowledgment from open ports details from ECU it will be average size is normal while the establishment of a network to intra vehicle and out network.The average fluctuation weight of the bundles in both sending and receiving bearings reflects the brute drive attack. The IDS with other goals, such as developing a dedicated framework for recognising a specific type of assault, can be constructed by picking the significant features based on the list after acquiring the highlight significance list of each assault. The most advanced features can be defined as critical attributes that network supervisors should keep an eye on. The attacks can be immediately noticed if certain attributes alter abnormally.

Chapter 6

Conclusion

In this research we proposed an IDS framework for the intelligent classification of malicious and non-malicious attacks. XGBoost is our final classifier for malicious and non-malicious data. XGBoost could be an ensemble learning algorithm that uses the gradient descent approach to combine multiple DTs to boost speed and performance. In this, we proposed the IDS framework which has divided into five steps. First, collect data from inter and external communication. Based on the dataset preprocessing the data using SMOTE. The next step is feature selection of data to reduce noise data and use necessary relevant data. On the feature selection, we build the model using a classification of a tree. The classification make ensemble learning using the staking model. As part of the result, we can say that if any changes happen in packets or networks based on features then the network controller expert can take it to respond and secure AV. If any attribute changes significantly as abnormal behavior is detected that attack can be detected as part of IDS. In the future, we will add more attacks to this model, and based on that we can predict a new model using ML.

Chapter 7

Future Work

In the future, the security level of the proposed system can be enhanced by adding a more attacks and adding a hybrid tree based algorithm in IDS. We can also focus on malware detection. we can add deep learning methods to add one more security level to our proposed architecture.

Bibliography

[1] Khattab M Ali Alheeti and Klaus McDonald-Maier. Intelligent intrusion detection in external communication systems for autonomous vehicles. *Systems Science & Control Engineering*, 6(1):48–56, 2018.

[2] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[3] Kyong-Tak Cho and Kang G Shin. Fingerprinting electronic control units for vehicle intrusion detection. In *25th USENIX Security Symposium (USENIX Security 16)*, pages 911–927, 2016.

[4] Wonsuk Choi, Kyungho Joo, Hyo Jin Jo, Moon Chan Park, and Dong Hoon Lee. Voltageids: Low-level communication characteristics for automotive intrusion detection system. *IEEE Transactions on Information Forensics and Security*, 13(8):2114–2129, 2018.

[5] George Eason, Benjamin Noble, and Ian Naismith Sneddon. On certain integrals of lipschitz-hankel type involving products of bessel functions. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 247(935):529–551, 1955.

[6] Bogdan Groza and Pal-Stefan Murvay. Efficient intrusion detection with bloom filtering in controller area networks. *IEEE Transactions on Information Forensics and Security*, 14(4):1037–1051, 2018.

[7] Yoshihiro Hamada, Masayuki Inoue, Hiroshi Ueda, Yukihiro Miyashita, and Yoichi Hata. Anomaly-based intrusion detection using the density estimation of reception

cycle periods for in-vehicle networks. *SAE International Journal of Transportation Cybersecurity and Privacy*, 1(11-01-01-0003):39–56, 2018.

[8] Kazuki Iehira, Hiroyuki Inoue, and Kenji Ishida. Spoofing attack using bus-off attacks against a specific ecu of the can bus. In *2018 15th IEEE Annual Consumer Communications Networking Conference (CCNC)*, pages 1–4, 2018.

[9] Ulf E Larson, Dennis K Nilsson, and Erland Jonsson. An approach to specification-based attack detection for in-vehicle networks. In *2008 IEEE Intelligent Vehicles Symposium*, pages 220–225. IEEE, 2008.

[10] Hyunsung Lee, Seong Hoon Jeong, and Huy Kang Kim. Otids: A novel intrusion detection system for in-vehicle network by using remote frame. In *2017 15th Annual Conference on Privacy, Security and Trust (PST)*, pages 57–5709. IEEE, 2017.

[11] Fang Li, Lifang Wang, and Yan Wu. Research on can network security aspects and intrusion detection design. Technical report, SAE Technical Paper, 2017.

[12] Mirco Marchetti and Dario Stabili. Anomaly detection of can bus messages through analysis of id sequences. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 1577–1583. IEEE, 2017.

[13] Michael Müter and Naim Asaj. Entropy-based anomaly detection for in-vehicle networks. In *2011 IEEE Intelligent Vehicles Symposium (IV)*, pages 1110–1115. IEEE, 2011.

[14] Ondrej Pribyl and Michal Lom. Impact of autonomous vehicles in cities: User perception. In *2019 Smart City Symposium Prague (SCSP)*, pages 1–6. IEEE, 2019.

[15] Kui Ren, Qian Wang, Cong Wang, Zhan Qin, and Xiaodong Lin. The security of autonomous driving: Threats, defenses, and future directions. *Proceedings of the IEEE*, 108(2):357–372, 2020.

[16] Eunbi Seo, Hyun Min Song, and Huy Kang Kim. Gids: Gan based intrusion detection system for in-vehicle network. In *2018 16th Annual Conference on Privacy, Security and Trust (PST)*, pages 1–6. IEEE, 2018.

[17] Ramin Shabanpour, Seyedeh Niloufar Dousti Mousavi, Nima Golshani, Joshua Auld, and Abolfazl Mohammadian. Consumer preferences of electric and automated vehicles. In *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, pages 716–720. IEEE, 2017.

[18] Dario Stabili, Mirco Marchetti, and Michele Colajanni. Detecting attacks to internal vehicle networks through hamming distance. In *2017 AEIT International Annual Conference*, pages 1–6. IEEE, 2017.

[19] Aifen Sui and Gordon Muehl. Security for autonomous vehicle networks. In *2020 IEEE 3rd International Conference on Electronic Information and Communication Technology (ICEICT)*, pages 67–69, 2020.

[20] Adrian Taylor, Nathalie Japkowicz, and Sylvain Leblanc. Frequency-based anomaly detection for the automotive can bus. In *2015 World Congress on Industrial Control Systems Security (WCICSS)*, pages 45–49. IEEE, 2015.

[21] Adrian Taylor, Sylvain Leblanc, and Nathalie Japkowicz. Anomaly detection in automobile control network data with long short-term memory networks. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 130–139. IEEE, 2016.

[22] Fangchun Yang, Shangguang Wang, Jinglin Li, Zhihan Liu, and Qibo Sun. An overview of internet of vehicles. *China communications*, 11(10):1–15, 2014.

[23] Li Yang, Abdallah Moubayed, Ismail Hamieh, and Abdallah Shami. Tree-based intelligent intrusion detection system in internet of vehicles. In *2019 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6, 2019.

[24] Abidullah Zarghoon, Irfan Awan, Jules Pagna Disso, and Richard Dennis. Evaluation of av systems against modern malware. In *2017 12th International Conference for Internet Technology and Secured Transactions (ICITST)*, pages 269–273, 2017.