AD-VR: a new approach for Cyber-Attack Detection with Machine Learning and Data Manipulation via the Variation Rate

Zeynep Gürler, Serra Uysal, and Ahmet Yemliha Öner

Abstract— Ensuring the security of the connected vehicles is crucial in terms of protecting human life and keeping personal data private as well as preventing financial losses that can be caused by cyberattacks. The brisk development of technology and the rapid increase in consumption day by day, have significantly increased the security vulnerabilities in connected vehicles. The use of the internet for internal and external communication of connected vehicles has made it possible to access connected vehicles remotely. For this reason, cyberattacks on connected vehicles have grown and increased the importance of conserving the security of connected vehicles. Therefore, in this article, we propose a cyberattack detection framework using machine learning and data manipulation via the variation rate (AD-VR) to detect cyberattacks on connected vehicles. In order to do that, we apply feature reduction to obtain the most important and essential features by selecting them via the variation rate, which is our contribution to already existing cyberattack detection methods. Then, we manipulate our reduced dataset by normalization and train the manipulated and reduced dataset using machine learning classifiers. Finally, we test our model with testing data and detect the incoming operations whether they are normal operations or attacks. We also provide the type of attacks that we detect and compare the accuracy results. Our proposed detection framework outperformed most of the existing methods by having a better accuracy rate in cyberattack detection.

Index Terms—Error analysis, machine learning, feature extraction or construction, security, automotive, autonomous vehicles, hardware/software Protection

1 Introduction

IN the last decade, people have witnessed a great deal of technological developments and improvements that have reflected in the most basic places of our daily life. Not only have we got used to these improvements but also, we have become dependent on them in every aspect of our lives. Considering transportation constitutes a significant place in our lifestyle, it is inevitable to expect the least from this industry to keep up with the advantages that increasing technology brings with it.

Connected vehicles come into the scene at this point. As mentioned in [1], connected vehicles comprise all the applications, services, and technologies that provide the capability to engage with outer devices, networks, applications, and services as well as communicate between other appliances found in the vehicle. In order to secure all these, automotive manufacturers use Automotive Secure Development Lifecycle (ASDL) which functions with STRIDE (Spoofing user identity, Tampering with data, Repudiation, Information disclosure, Denial of service, and Elevation of privilege) scheme [2]. However, developments in connected vehicles have happened too rapidly that manufacturers mainly focused on functionality rather than security so these precautions run short in the meaning time. Latterly, this caused many security-related problems in electronic control units (ECU) with the controller area networks (CAN) of the connected vehicles [3].

When the subject is networks and technological services, the vulnerabilities that they will bring within should always be taken into consideration since these vulnerabilities often arise some crucial security problems in the system. Cyberattacks are very crucial security threats to connected vehicles. [4] estimated that the cost of cybercrime, which was 3 trillion dollars in 2015, will increase to 6 trillion dollars by 2021 and they stated that these attacks have begun to directly target people and the tools they use in daily life such as cars, planes, etc. in a way that threatens their lives, instead of only affecting computers, networks or smartphones. Therefore, even though technological developments ease our life such as remote door locks, flashing headlights, and remote parking [5], they generally, come with the consequence of facing a series of security problems.

To make a system secure, there are a few main methods; using cryptography, increasing network security, malware detection, or software vulnerability detection. In our paper, we focus on vulnerability detection in connected vehicles.

- *Zeynep Gürler is an undergrad student in the Computer Engineering Department, Istanbul Technical University. E-mail: gurler17@itu.edu.tr.*
- *Serra Uysal is an undergrad student in the Computer Engineering Department, Istanbul Technical University. E-mail: uysals17@itu.edu.tr.*
- *Ahmet Yemliha Öner is an undergrad student in the Computer Engineering Department, Istanbul Technical University. E-mail: oner15@itu.edu.tr.*

xxxx-xxxx/0x/$xx.00 © 200x IEEE Published by the IEEE Computer Society

Since machine learning approaches constitute a consequential place as it is one of the most popular subjects in recent years, in this paper, we propose an attack detection framework that uses machine learning and data manipulation approaches for connected vehicles to detect cyberattacks.

Our data set where we test the framework includes a number of different attack types. Yet, one of them is mainly focused in this paper, which is called Denial of Service (DoS) attack. A DoS attack is a cyberattack, which generates a large quantity of network traffic packets to crush an Internet service host by exceeding its network operation capacity [6]. An Internet host that has been DoS attacked can no longer provide its services and do its tasks. This situation might immobilize large-scale routine operations and functions in a system which can result in an economic loss or even loss of someone's life. DoS attacks are adequately common among cyberattack types targeting connected vehicles.

We use a dataset that contains some common cyberattack types among connected vehicles, where DoS attacks constitute a large proportion in the set. Before directly diving into machine learning operations, the data set has to be arranged optimally in a manner that will suit our desired form. Feature extraction via the variation rate and normalization is preferred to achieve this goal, which is the main contribution of this paper that differs the proposed framework from the other machine learning-based cyberattack detection methods. We briefly discuss the differences and pros of our framework in the related parts. In the resulting framework, we calculate the probability of an attack in the process between connected vehicles. As a result, we end up with a reliable attack detection software system called AD-VR. We implement AD-VR systems in a learning-based manner and we propose this system to measure whether additional advanced security layers are needed for a specific process in a system, by detecting the probability of the processes being a cyberattack, or the process is safe to work with.

2 Attack Detection with Machine Learning

In this section of the paper, we examine the detection and classification of cyberattacks on connected vehicles. Then, we analyze the performance of the method that we suggest by testing it over a dataset and discuss the results. Before jumping into that part, first, we need to briefly explain the inspiration behind this work.

Our literature review about cyberattacks on connected vehicles [7] showed that hackers had to have a physical admission in order to manipulate vulnerabilities of the system where the target of the attack could be only a single-vehicle, jamming attacks can be given as an example of that which will cause detected objects to disappear from the Autopilot System [8]. These kinds of attacks are harmful in terms of protection of civil life and financially cause a loss for individuals along with the manufacturers. However, connectivity in the wireless network manner on vehicles abolished the restriction of being in a certain distance and gave power to hackers to perform malicious attacks via a distance as well as the ability to damage the whole system rather than focusing on an individual vehicle which will increase the cost of loss impressively higher. Some examples of these cyberattacks on connected vehicles can be mentioned as:

Fig. 1. Cyber Vulnerabilities Overview [7]: shows the fundamental attack types targeting the connected vehicles

1. Charlie Miller and Chris Valasek [9], showed that even though manufacturers rely on their implementation and argue for having "unhackable" connected vehicles, in 2015 they released the method that they used for finding IP addresses of the cars and after locating them sending arbitrary CAN messages which resulted causing the cars to drive off-road and crash. They also included a list of vehicle models that they tested.
2. Martyn Wiliams published a news article that points out how hackers mimicked the BMW servers that are managed by the BMW assistance line to help drivers who have been locked out of their cars. According to the news, in order to fix the issue, BMW published software patches to 2.2 million cars with the Connected Drive program [10].
3. Jonathan Gitlin from Arstechnica wrote a news article that tells the story of security researchers that discovered how to use software-defined radio(SDR) to unlock and lock millions of Volkswagens. To do that, they only needed $40 worth of Arduino and SDR [11].
4. Keen Security Lab researchers find a way to control a Tesla car from out of the car, by 12 miles (19km). Hackers were able to open doors, switch lanes while the car was moving. Tesla published an update to fix the issue [12].

In the presence of this news, it can be seen that cyberattacks are causing significant damage both financially and morally, by deaths and injuries caused by incidents, with intergrowth of connected vehicles in our lives and vulnerabilities it contains. That is why the importance of prevention of these attacks has increased. Fig 1. summarizes attack types that target vulnerabilities of attack surfaces on connected vehicles.

According to [13], denial of service (DoS) attacks are the most common attack type among connected vehicles. DoS attacks include a diversity of regular and reactive jamming attacks and various other higher layer attacks [14], [15]. The dataset that we used represents the commonness of DoS attacks among others as well (details will be explained in the related section).

Our literature review on this area shows that the most

TABLE 1
ATTRIBUTES OF ATTACK TYPES IN KDD99 AND CV-KDD

Attack Names	Attack Types	Num in 10%KDD	Num in CV-KDD	Percentage
normal	not attack	97278	9727	19.69107
buffer_overflow	u2r	30	0	0.00607
loadmodule	u2r	9	0	0.00182
perl	u2r	3	0	0.00061
neptune	dos	107201	10721	21.69968
smurf	dos	280790	28080	56.83766
guess_passwd	r21	53	0	0.01073
pod	dos	264	0	0.05344
teardrop	dos	979	98	0.19817
portsweep	prope	1040	104	0.21052
ipsweep	probe	1247	125	0.25242
land	dos	21	0	0.00425
ftp_write	r21	8	0	0.00162
back	dos	2203	220	0.44593
imap	r21	12	0	0.00243
satan	prope	1589	158	0.32165
phf	r21	4	0	0.00081
nmap	prope	231	0	0.04676
multihop	r21	7	0	0.00142
waresmaster	r21	20	0	0.00405
warezclient	r21	1020	102	0.20647
spy	r21	2	0	0.00040
rootkit	u2r	10	0	0.00202

common solution to detect these attacks is to use machine learning methods. Many researchers have developed attack detection frameworks adopting predictive robustness of machine learning methods. For instance, [16] proposed an Intrusion Detection System (IDS) using Support Vector Machine (SVM) to detect network attacks such as DoS attacks and User to Root (U2R) attacks in healthcare systems.

As an addition to the machine learning approaches, [17] combined feature engineering and machine learning to detect Distributed Denial of Service (DDoS) attacks considering that feature reduction boosts the performance of the learner.

Drawing inspiration from previous works, we contemplate the importance of data engineering and perform feature reduction and normalization operations on 10% KDD99 [18] dataset in order to result it a boost in detection accuracy.

KDD99 is an intrusion detection benchmark dataset, which consists of 41 features and 22 types of attacks with 6 of them being DoS attacks. It is one of the most used datasets for research projects in the cybersecurity field. As an example of such works, [19] processed the KDD99 dataset and generated a modified dataset called CAV-KDD aiming to detect cyberattacks on Connected and Autonomous Vehicles (CAVs). This influenced us to name our reduced dataset as CV-KDD since we work on connected vehicles. Different from their approach, we reduce the features of the dataset by trying two different methods and settle upon the one that has a better outcome, which is the variation rate method. This way we obtain a more meaningful list of features for our detection system. Later on, we normalize this reduced dataset to come by with better fit clusters to work with. Initially, we apply a machine learning approach by using 65% of the dataset for training and 35% of the dataset for testing purposes.

3 Dataset

In this section, we will concisely explain the dataset that is used for the evaluation of our proposed cyberattack detection model. As mentioned earlier, we used the 10% KDD99 [18] intrusion detection benchmark dataset for the sake of simplicity. This dataset includes normal operations and 6 different types of DoS attacks; neptune, smurf, back, teardrop, pod, and land. In addition to DoS attacks, there are a few other types of attacks that constitute less ratio in the dataset. Those attack types are user to root(u2r), root to local(r2l), and probe attacks. The attacks buffer_overflow, loadmodule, and pearl belong to u2r type; guess_passwd, ftp_write, imap, phf, mutihop, waresmaster, warezclient, and spy belong to r2l type; and lastly portsweep, ipsweep, satan, and nmap belong to probe attack type. In the sense that KDD99 consists of 41 features per entry, it possesses being one of the most preferred datasets for cyber-security related study projects by researchers.

We use 10% KDD99 dataset and ignore the types that constitute less than 900 samples from the set in order to balance the dataset. So we end up with attacks; neptune, smurf, teardrop, portsweep, ipsweep, back, satan, and warezclient.

Later on, we randomly choose 10% of the resulting part (1% of the original data) for the sake of simplicity. After these steps, the new dataset (CV-KDD) contains 49334 operations in total where 39449 of these represent the operations with one of the cyberattack types that have been mentioned above. Table [1] shows the percentage of operations in the dataset.

4 Attack Detection with Data Manipulation and Machine Learning for Connected Vehicles

Attack detection is a crucial step in maintaining secure systems. If a cyberattack can be detected beforehand the disservice that it may bring to the system can be prevented. This enhances the reliability of high technology products and encourages consumers to incorporate with these services. This enables manufacturers to warrant millions of dollars. That is why they started to pay more attention to the security aspect of newly developed technological improvements where connected vehicles form one of the subareas for these improvements. Therefore, a proactive security system is a must for a secure network connection for vehicles [20], our designed model will protect connected vehicles from wireless cyberattacks by evaluating the operation before it affects the whole system.

We create a system to detect cyberattacks using data manipulation and machine learning methods. Fig 2. shows the fundamental outline of that system. AD-VD, as it will be explained in more detail in further parts, serves as a model that is used to detect the cyberattacks and determine the type of the attack in order to inform the system to have an additional security layer. This way it will prevent the malfunction of the connected vehicle.

Looking at the implementation phase of a cyberattack detection system, first and foremost, researchers have to understand the data that they work with, so we paid attention to what we have in our hands as a dataset and examined the details of it. The technicalities of the dataset have been explained in the previous section (i.e. section 3 Dataset). As it is stated there, the dataset contains specific information about the process which sometimes can constitute a ground for unnecessary data clutter. In order to overcome this prodigality of data, which can affect the performance of the methods to be used in further steps, one needs to use some data manipulation technique that will provide a more meaningful dataset to be worked. Following this information, we come up with a data manipulation method that provides an advancement in detection of the cyberattacks. Later on, as our literature review suggested we got our eyes on machine learning methods that could help us to achieve our goal. There are plenty of machine learning methods all with some pros and cons depending on the problem that is being examined. As we will explain in detail in further sections, we

chose the algorithm depending on the performance of the accuracy results that we obtain.
The following sections give information in depth about the implementation and working principle of the AD-VR method and key steps of this framework.

4.1 Overview of AD-VR

As an overview, our proposed Cyber-Attack Detection with Machine Learning and Data Manipulation via the Variation Rate (AD-VR) framework, firstly transforms KDD99 dataset into CV-KDD dataset for connected vehicles by reducing features, which is done by eliminating some of the features by looking at their variation rates. Next, AD-VR normalizes these reduced features to maintain a better workspace for classification. The last step is the classification of the dataset to achieve the detection of network operation types in real-life situations.
According to [21], there are four main approaches to cyberattack detection being Bayesian detection with binary hypothesis (i.e., Naive Bayes classifier), weighted least square approaches (i.e, support vector machine), X^2 detector based on Kalman filters, and fault detection and isolation techniques. Our proposed framework uses one of the weighted least-square approaches, support vector machine (SVM). So last, AD-VD trains a part of the modified dataset by SVM classifier and tests the other part to examine the success of the framework. Our main contributions in this paper are articulated as below:

1. AD-VR is a framework for cyberattack detection with machine learning and data manipulation via the variation rate can be integrated into connected vehicles as a software cyberattack detector agent and be utilized in real-life.
2. We propose a feature selection method defining a value called the Variation rate to avoid redundant and unnecessary features.

4.2 Key Steps of Framework

In this section, we introduce the key steps of our cyberattack detection framework for connected vehicles. We divided the framework into three elemental key steps as a basis depending on the operation procedures and our approach for implementing the system. The nature of the cyberattack detection systems that have been implemented earlier, led us to follow these steps and work our knowledge upon them. We will explain each of these three key steps of our detection framework as shown in Fig. 3.

4.2.1 Feature reduction

Big data, which is generated with the increasing use of technology over time, provides us an opportunity to process and learn from the data. The first step in learning from data is to know your data. Knowing the data leads us to be able to make result-oriented manipulations on the data, and to get rid of unnecessary features.
In high-dimensional data-based studies, as in our case, removing unnecessary and redundant features from the data leads to a more intelligible model [22]. Therefore, we apply feature reduction to our dataset using two different methods: 1) selecting the features that have slight standard deviations (STDs). We execute our first method by computing the STD values of each feature and selecting the ones that have an STD value greater than 0.38, which constitute 19 of the features. We use the result of this method for implementation of AD-STD, which is the attack detection method that we have created by taking the

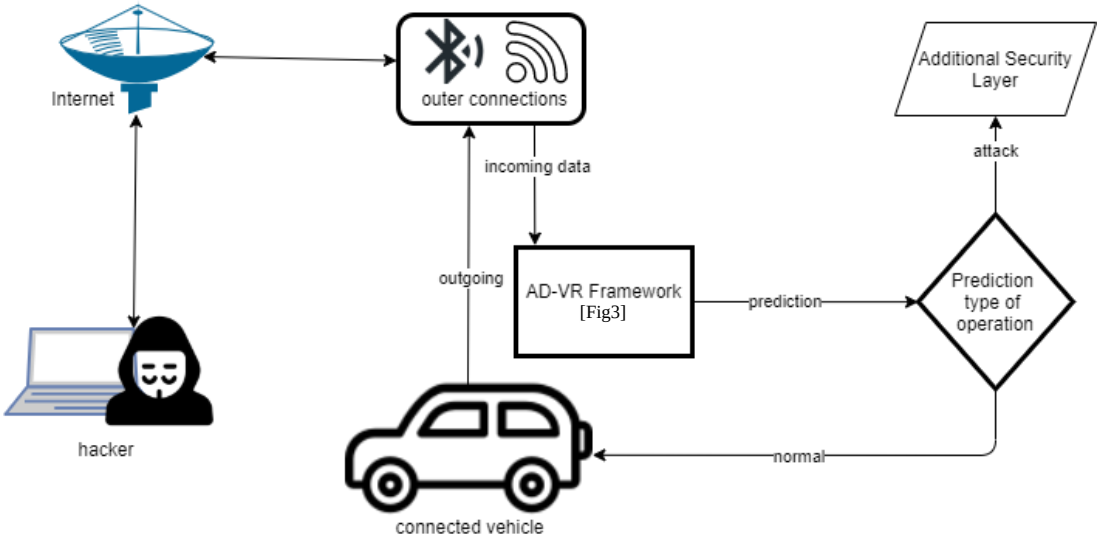

Fig. 2. Data-flow of operation information and planned integration that shows the basic outline of the AD-VD framework which detects the cyberattacks and informs the system according to attack type and additional security layer will be combined with the system.

STD values as the basis for our feature reduction. 2) selecting the features according to their variation rate with respect to the normal operation sample with the closest distance to the population center, which we will call center normal later on in this paper. For our second method, we count the values of each feature that are different from the center normal, which is generated by computing the cumulative distance to the population center of normal operations and selecting the closest one, and convert the counted value to a percentage for each feature to obtain variation rates. This strategy can give us the feature-wise similarity of samples. By avoiding nearly-same features, we highlight the distinguishing features and better discriminate classes. Also, looking at each feature from the perspective of the center normal might give us better discrimination between attacks and normal operations, which leads us to be able to differentiate them and achieving the main goal that is keeping the environment secure. Moreover, this approach can be seen as normalization or registration with respect to the center normal and we can see the calculations as the distance to the center normal. Therefore, we dodge the features that are not able to spread in data space meaning staying close to each other and the center normal. Adapting what we did for AD-STD for this part, we use the result of this method for implementation of AD-VR, which is the attack detection method that we have created by taking the variation rate values as the basis for our feature reduction. Next, we realize our second method by choosing the features with variation rates lower than 80%. Table 2 shows the STD value and variation rate of each feature.

This aspect gives us reduced dimensionality of features and extraction of the most important and effective features, which leads us to have a more robust and reliable model that suits our desired purpose since handling high-dimensionality is a real deal. We use the result of this part as a database for further steps.

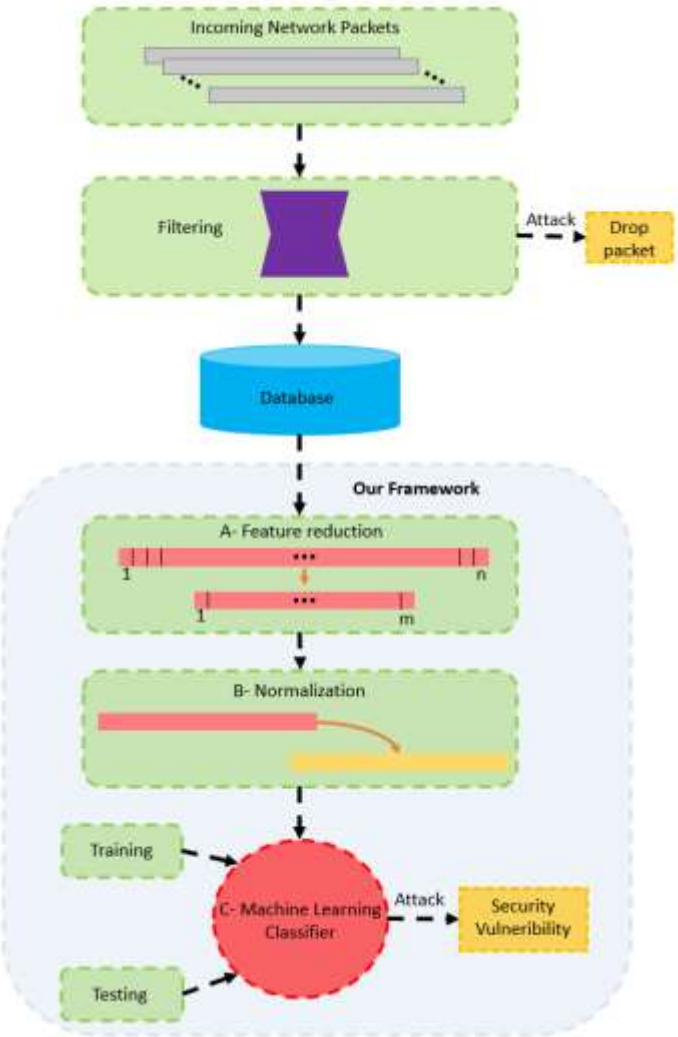

Fig. 3. A scheme to visualize our proposed framework. Subcategories of the framework are given as A- Feature reduction, B- Normalization and C- Machine Learning Classifier. Details of these categories are explained in the related sections (4.2.1,4.2.2, and 4.2.3 respectively)

4.2.2 Normalization of reduced features

Data Normalization is one of the most influential pre-processing steps for data analysis. By transforming the data into precise order with linear transformation, raw data is homogenized to create more suitable clusters and better the precision of algorithms [23]. Normalization can be also realized and explained as a distance measurement between all data points and a fixed point, which will transform the data into a different coordinate space. This utilized distances and new space can mean so many things and offer so many opportunities in case of learning and making inferences from data. Since it is a common and effective approach, we normalize the features that have been reduced in the previous step by dividing each with the Frobenius normalized selves. As a result of this step, we obtain normalized and reduced features that can be called as manipulated features in total, which will be used in the next step.

TABLE 2
STD VALUES AND VARIATION RATE OF FEATURES

Ind	Feature Name	VR	Std	Used
1	duration	0.975	720.3773	False
2	protocol type	0.384	0.96071	True
3	service	0.130	5.06073	True
4	flag	0.766	0.81840	True
5	src bytes	0.000	58502.637	True
6	dst bytes	0.000	38009.016	True
7	land	0.999	0.00900	False
8	wrong fragment	0.997	0.13608	False
9	urgent	0.999	0.00000	False
10	hot	0.993	0.75984	False
11	num failed logins	0.999	0.01102	False
12	logged in	0.148	0.35521	True
13	num compromised	0.995	0.09348	False
14	root shell	0.999	0.00450	False
15	su attempted	0.999	0.00450	False
16	num root	0.998	0.22250	False
17	num file creations	0.999	0.10279	False
18	num shells	0.999	0.00636	False
19	num access files	1.000	0.02880	False
20	num outbound cmds	1.000	0.00000	False
21	is host login	0.998	0.00000	False
22	is guest login	0.005	0.03841	True
23	count	0.016	213.198	True
24	srv count	0.819	246.305	False
25	serror rate	0.821	0.38078	False
26	srv serror rate	0.941	0.38098	False
27	rerror rate	0.939	0.23160	False
28	srv rerror rate	0.773	0.23223	True
29	same srv rate	0.773	0.38817	True
30	diff srv rate	0.929	0.08208	False
31	srv diff host rate	0.001	0.13893	True
32	dst host count	0.010	64.653	True
33	dst host srv count	0.704	106.120	True
34	dst host same srv rate	0.702	0.41115	True
35	dst host diff srv rate	0.001	0.10920	True
36	dst host same src port	0.894	0.48121	False
37	dst host srv diff host	0.809	0.04223	False
38	dst host serror rate	0.811	0.38063	False
39	dst host srv rerror rate	0.928	0.38084	False
40	dst host rerror rate	0.930	0.23068	False
41	dst host srv rerror rate	0.196	0.23048	True

4.2.3 Machine Learning Classifier

Machine Learning methods are one of the most popular subjects these days and commonly used in cyberattack detection. In a recently done research [24], a supervised machine learning algorithm, SVM (Support Vector Machines) is used for DoS attack detection along with Naive Bayesian Network and Decision Tree classification where the results of the SVM classification model was better than those other methods.
SVM is linked with learning algorithms, which examine data for categorization and regression analysis. These results led us to use the SVM model in our framework. In this step, we train 65% of our manipulated CV-KDD dataset as our training set by SVM classifier to classify each attack type and normal operation. Then we predict the classes of testing samples that are being the 35% of the manipulated CV-KDD dataset as seen in figure 2′ part C.

4.3 Evaluation

Python is one of the most used programming languages when it comes to machine learning and data science-based applications [25]. Python provides easy-to-use and diversified environments with its up to date machine learning and data handling modules and libraries such as sklearn and numpy. Therefore, we used python to evaluate our proposed cyberattack detection with machine learning and data manipulation via variation rate (AD-VR) framework. Also, we used Frobenius distance for the normalization of the reduced features.

5 Analysis of Proposed Solution

We benchmarked our Cyber-Attack Detection with Machine Learning and Data Manipulation via the Variation Rate (AD-VR) framework against two comparison methods: 1) Cyber-Attack Detection with Machine Learning and Data Manipulation via STD (AD-STD), which is also our contribution with a change between variation rate and STD in feature reduction step. 2) Attack detection framework for CAVs using J48 decision tree algorithm and CAV-KDD dataset from [19]. [19] modified the same dataset like us however with a different data manipulation method which results in a distinct dataset distribution. Therefore, some of the cyberattacks are not included in both frameworks. So, we show the accuracy rates of joint cyberattacks in the attack detection framework with the J48 decision tree algorithm. Table 3 displays the attack-wise accuracy of the methods.
As we observed, our proposed framework AD-VR outperformed AD-STD in total with an accuracy of 99.6641 while AD-STD being 99.6236. Moreover, we affirmed that normalization worked well since the

accuracy rate was 89.454482 without normalization which is significantly lower than previously obtained results. Also, we can say that AD-VR is more generalized than AD-STD among all types, even the accuracy rate of teardrop detection, which have a slightly lower ratio than other attacks, is 9.68% in AD-STD and boosted to 38.7% in AD-VR despite the dataset being slightly imbalanced. Therewithal, results of the attacks that hold less quantity, meaning less samples in the dataset, in CV-KDD, almost always have a higher accuracy in AD-VR compared to AD-STD.

Also, regarding the joint outcomes between AD-VR and the detection framework with J48, AD-VR outperformed the detection framework with J48 except for the teardrop attack detection. Overall, our proposed solution for this task, AD-VR framework, achieved the best accuracy compared to both already existing models and our other attempts that we tried throughout our implementation phase. As a result, we showed that feature selection with respect to the normal operation population center and normalization is indeed a successful strategy outperforming the comparison methods.

TABLE 3
STD VALUES AND VARIATION RATE OF FEATURES

Types	AD-STD	AD-VR	J48*
total	99.6236	99.6641	-
normal	99.8555	99.7688	99.7
neptune	100	99.9200	99.1
smurf	99.9795	99.9795	99.6
teardrop	9.6774	38.7097	100
portsweep	78.5714	88.0952	-
ipsweep	90.6977	100	96.1
back	98.8372	100	-
satan	83.3333	87.0370	-
warezclient	81.0811	62.1622	-

*Accuracy results for J48 are from [17]

6 Conclusion and Future work

The connectivity in vehicles causes systems to have vulnerabilities that malicious people may take advantage of. This has caused some financial loss for manufacturers in the past and it may cause much more significant losses if no precaution is taken. On 17 March 2016, the Federal Bureau of Investigation together with the U.S Department of Transportation and National Highway Traffic Safety Administration (NHTSA) publicly announced that remote exploits due to the vulnerabilities of motor vehicles have significantly increased [26]. Unlocking doors from a distance, engine breakdown, and disqualifying brakes may be given as examples for these remote exploits. In a newly published article [27], writers describe the connected vehicles as "driving with sharks" referring to the hackers. To secure the systems, one of the main precaution steps is the detection of cyberattacks.

If a cyberattack can be precisely detected, required actions can be taken in order to prevent the malicious effects of the attacks. [28] states that, due to the innovative signal processing and communication capabilities, the importance of cyberattack detection has significantly increased. Therefore, cyberattack detection is a crucial step for establishing secure connected vehicles.

Our literature review on this topic showed that one of the most important tasks of detecting and preventing cyberattacks on connected vehicles is feature extraction. [29] states that data should be efficient and effective at the same time so feature extraction is crucial to obtain these properties. Since there are loads of features in the dataset that we used, we needed to minimize it by extracting the valuable features. After searching for feature extraction methods, we tried some well-known methods like calculating STD values for comparison however we did not get the results that we wanted with that method so we ended up creating our own feature extraction approach, which is feature selection depending on the variation rate. We define the variation rate as the difference rate of the features with the normal operation population center representee.

After we reduced the features using the variation rate approach over the dataset, we used the normalization method to create our own set of data, called CV-KDD. This new dataset allowed us to do the operations in a more effective way and obtain a better result for the further steps.

We saw that in order to detect cyberattacks, researchers mostly preferred using some machine learning techniques. To train and test the model, we searched the accuracies of machine learning classification methods. Our research on this area showed that [30], the most famous machine learning methods for cyberattack detection are SVM, ANN (Artificial Neural Network), and some Decision Tree methods. However, our tests on the dataset revealed that the most optimal method for our desired solution is the support vector machine (SVM) method with AD-VR feature extraction. This combination resulted in 99.66% accuracy in our implementation.

What we have achieved so far, motivated us for more research and implementation in this area. In our future work, we will dive into the field of imbalanced data handling which is a wide topic by itself. The main structure of our future work will be working on data augmentation to balance out our CV-KDD dataset that we have proposed for this work. The reason behind this is that we have seen the unfavorable side of the imbalance through our implementation. We expect that balancing out the dataset will result in a better outcome. We hope to work in a more powerful environment in the future in order to be able to do the calculations over a larger dataset which will give us a more reliable outcome for our propositions. Our current environment limited our capability of handling the number of data that is used for our training and testing purposes. Also, we will work more on the normalization with respect to a fixed point strategy as feature reduction via variation rate being our main contribution

in this paper. Later on, data space transformation, feature extraction, or dimensionality reduction might be in our vision, too.

REFERENCES

[1] "Connected Vehicles - IEEE Connected Vehicles", *Site.ieee.org*. [Online]. Available: https://site.ieee.org/connected-vehicles/ieee-connected-vechicles/connected-vehicles/. [Accessed 25 December 2020].

[2] A. Oyler and H. Saiedian, "Security in automotive telematics: a survey of threats and risk mitigation strategies to counter the existing and emerging attack vectors", *Security and Communication Networks*, vol. 9, no. 17, pp. 4330-4340, 2016. Available: 10.1002/sec.1610 [Accessed 28 December 2020].

[3] S. Woo, H. J. Jo and D. H. Lee, "A Practical Wireless Attack on the Connected Car and Security Protocol for In-Vehicle CAN," in IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 2, pp. 993-1006, April 2015, doi: 10.1109/TITS.2014.2351612.

[4] Steve Morgan. 2016. Hackerpocalypse: A Cybercrime Revelation. Technical Report. Cybersecurity Ventures. 1–24.

[5] J. TAKAHASHI, "An Overview of Cyber Security for Connected Vehicles", *IEICE Transactions on Information and Systems*, vol. 101, no. 11, pp. 2561-2575, 2018. Available: 10.1587/transinf.2017ici0001 Accessed on: Dec. 28, 2020.

[6] Z. Baig, S. Sanguanpong, S. Firdous, V. Vo, T. Nguyen and C. So-In, "Averaged dependence estimators for DoS attack detection in IoT networks", *Future Generation Computer Systems*, vol. 102, pp. 198-209, 2020. Available: 10.1016/j.future.2019.08.007 [Accessed 29 December 2020].

[7] B. Sheehan, F. Murphy, M. Mullins and C. Ryan, "Connected and autonomous vehicles: A cyber-risk classification framework", *Transportation Research Part A: Policy and Practice*, vol. 124, pp. 523-536, 2019, doi: 10.1016/j.tra.2018.06.033

[8] C. Yan, W. Xu, J. Liu, Can You Trust Autonomous Vehicles: Contactless Attacks against Sensors of Self-driving Vehicle. in DEF CON 24 Hacking Conference, 2016. Available: https://infocon.org/cons/DEF%20CON/DEF%20CON%2024/DEF%20CON%2024%20presentations/DEF%20CON%2024%20-%20Liu-Yan-Xu-Can-You-Trust-Autonomous-Vehicles-WP.pdf

[9] C. Miller and C. Valasek, "Remote Exploitation of an Unaltered Passenger Vehicle", *Countermeasure.ca*, 2013. [Online]. Available: https://www.countermeasure.ca/wp-content/uploads/2013/11/documents_2015_presentations_Chris-Valasek.pdf. Accessed on: Dec. 25, 2020.

[10] M. Williams, "BMW cars found vulnerable in Connected Drive hack", *Computerworld*, 2015. [Online]. Available: https://www.computerworld.com/article/2878424/bmw-cars-found-vulnerable-in-connected-drive-hack.html. [Accessed on: Dec. 25, 2020.

[11] J. Gitlin, "Almost every Volkswagen sold since 1995 can be unlocked with an Arduino", *Ars Technica*, 2016. [Online]. Available: https://arstechnica.com/cars/2016/08/hackers-use-arduino-to-unlock-100-million-volkswagens/. Accessed on: Dec. 25, 2020.

[12] D. Lee, "Tesla updates software after car hack", *BBC News*, 2016. [Online]. Available: https://www.bbc.com/news/technology-37426442. Accessed on: Dec. 25, 2020.

[13] X. Zhang, H. Du, J. Wei, Z. Jia, S. Jia and G. Ma, "High Gain Observer Design for DoS Attack Detection in CACC Platoon," 2020 International Seminar on Intelligent Technology and Its Applications (ISITIA), Surabaya, Indonesia, 2020, pp. 254-259, doi: 10.1109/ISITIA49792.2020.9163674.

[14] I. Oancea and E. Simion, "Challenges in Automotive Security," 2018 10th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), Iasi, Romania, 2018, pp. 1-6, doi: 10.1109/ECAI.2018.8679052.

[15] R. van der Heijden, T. Lukaseder and F. Kargl, "Analyzing attacks on cooperative adaptive cruise control (CACC)," 2017 IEEE Vehicular Networking Conference (VNC), Torino, 2017, pp. 45-52, doi: 10.1109/VNC.2017.8275598.

[16] M. Begli, F. Derakhshan and H. Karimipour, "A Layered Intrusion Detection System for Critical Infrastructure Using Machine Learning," *2019 IEEE 7th International Conference on Smart Energy Grid Engineering (SEGE)*, Oshawa, ON, Canada, 2019, pp. 120-124, doi: 10.1109/SEGE.2019.8859950.

[17] M. Aamir and S. Zaidi, "DDoS attack detection with feature engineering and machine learning: the framework and performance evaluation", International Journal of Information Security, 2019, pp. 761-785, doi: 10.1007/s10207-019-00434-1

[18] UCI kdd cup 1999 Data Data Set, 1999. Available online: https://archive.ics.uci.edu/ml/datasets.

[19] Q. He, X. Meng, R. Qu and R. Xi, "Machine Learning-Based Detection for Cyber Security Attacks on Connected and Autonomous Vehicles", *Mathematics*, vol. 8, no. 8, p. 1311, 2020. Available: 10.3390/math8081311.

[20] M. Dibaei et al., "Attacks and defences on intelligent connected vehicles: a survey", Digital Communications and Networks, vol. 6, no. 4, pp. 399-421, 2020. Available: 10.1016/j.dcan.2020.04.007 [Accessed 29 December 2020].

[21] M. Mahmoud, M. Hamdan and U. Baroudi, "Modeling and control of Cyber-Physical Systems subject to cyber attacks: A survey of recent advances and challenges", *Neurocomputing*, vol. 338, pp. 101-115, 2019. Available: 10.1016/j.neucom.2019.01.099 [Accessed 29 December 2020]

[22] S. Mukherjee and N. Sharma, "Intrusion Detection using Naive Bayes Classifier with Feature Reduction", *Procedia Technology*, vol. 4, pp. 119-128, 2012. doi: 10.1016/j.protcy.2012.05.017.

[23] V.R. Patel and R.G. Mehta, "Impact of Outlier Removal and Normalization Approach in Modified k-Means Clustering Algorithm", *IJCSI International Journal of Computer Science* vol. 8, no. 2, pp. 331-336 Sep. 2011. Accessed on: Dec. 27, 2020 [Online] Available: www.ijcsi.org/papers/IJCSI-8-5-2-331-336.pdf.

[24] W. Zhe, C. Wei and L. Chunlin, "DoS attack detection model of smart grid based on machine learning method," *2020 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS)*, Shenyang, China, 2020, pp. 735-738, doi: 10.1109/ICPICS50287.2020.9202401.

[25] S. Raschka, J. Patterson and C. Nolet, "Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence", *Information*, vol. 11, no. 4, p. 193, 2020. Available: 10.3390/info11040193 [Accessed 29 December 2020].

[26] Federal Bureau of Investigation. (2016, Mar. 17). Motor vehicles increasingly vulnerable to remote exploits. [Online]. Available: http://www.ic3.gov/media/2016/160317.aspx

[27] M. Hashem Eiza and Q. Ni, "Driving with Sharks: Rethinking Connected Vehicles with Vehicle Cybersecurity," in IEEE Ve-

hicular Technology Magazine, vol. 12, no. 2, pp. 45-51, June 2017, doi: 10.1109/MVT.2017.2669348.

[28] M. N. Kurt, Y. Yılmaz and X. Wang, "Real-Time Detection of Hybrid and Stealthy Cyber-Attacks in Smart Grid," in *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 2, pp. 498-513, Feb. 2019, doi: 10.1109/TIFS.2018.2854745.

[29] C. Torrano-Gimenez, H. Nguyen, G. Alvarez and K. Franke, "Combining expert knowledge with automatic feature extraction for reliable web attack detection", *Security and Communication Networks*, vol. 8, no. 16, pp. 2750-2767, 2012. Available: 10.1002/sec.603 [Accessed 28 December 2020].

[30] M. Tayyab, B. Belaton and M. Anbar, "ICMPv6-Based DoS and DDoS Attacks Detection Using Machine Learning Techniques, Open Challenges, and Blockchain Applicability: A Review," in *IEEE Access*, vol. 8, pp. 170529-170547, 2020, doi: 10.1109/ACCESS.2020.3022963.