Intrusion Detection System And Feature Analysis Of Network Attacks In Vanets

[1]Pavithra T, [2]Nagabhushana B S

[1]VTU Research Scholar, [2]Professor

[1]Department of Electronics and Communication Engineering,

[1]BMS College of Engineering, Bengaluru, India

Abstract - Vehicular Adhoc networks (VANETs) is the most promising research area. Implementation of VANETs needs to address issues on security, privacy and speed. Security in VANETs is very important. Understanding that the attack is happening is very important to address these issues. Our previous paper includes a comprehensive survey on security attacks in VANETs and the impact of attacks on the network. This paper discusses how efficiently the Machine Learning algorithms help identify the attack. Machine Learning algorithms are utmost widely used to make such predictions because of their well-accepted accuracy. This paper discusses DDoS, PortScan and DoS Hulk attack classification using different trained models to see which algorithm is more effective and why. Models are developed using MATLAB. With the help of results, an attempt has been made to explain the reason for misclassification and why certain Machine Learning algorithms have greater classification accuracy.

Index Terms - VANET, Decision Tree, K nearest neighbor, SVM, Neural Networks

I. INTRODUCTION

VANET is a communication network where communication takes place between vehicles or between vehicles and Road Side unit[1]. It is a most promising technology in the automotive field that helps in a significant reduction in the number of accidents, and road traffic and provides infotainment service. Implementation of VANETs helps the public to address traffic issues. To deploy the service, one needs to develop confidence in vehicle owners regarding the safety of their data and privacy. An attacker can perform any attack, Active or Passive. Networks need to defend against such attacks to protect vehicle owners. In our next paper, we discuss comprehensively about the VANET attacks and their classification along with the layer at which an attack occurs. To identify such attacks, the network must be aware of its behaviour during an attack.

To make this happen, proper training of the model by capturing the data during an attack and understanding the pattern of these data is necessary. Then only it can predict if an attack has occurred or not. Machine Learning (ML) algorithms are the best choice while making such predictions.ML algorithms also provide solution for Privacy and trust issues. In this paper, we are going to discuss the results of the training model developed by considering the DDoS attack, DoS attack and PortScan attack. The dataset used is from the CICIDS attack dataset for training the model [2]. MATLAB is used to develop a training model for classification and validation.

Section II covers a literature review on Supervised and Unsupervised training of Machine Learning that helps to understand the significance of developing a training model. In Section III various classification learning algorithms, such as Decision Tree, Naïve Bayes and Support Vector Machine were discussed. Section IV discusses the various classification models used in our experiment, Section V discuss about Simulation Result and the paper ends with Section VI discussing the Conclusion.

II. LITERATURE SURVEY

Machine learning concept, an Application of Artificial Intelligence will help us to train the computer to learn so that a model can be developed, which is capable of providing proper prediction to the query with acceptable accuracy after analyzing a given problem. Machine learning generally works in the following manners. First, it learns the given concept, develop a model by using the training data and validate the developed model using validation dataset. Also, it provides the accuracy of the model based on Data validation. Generally, a dataset will be classified into training dataset and validation data set consisting of 70% and 30% of the total data respectively.

For any Machine Learning Algorithms, use of Training data generally involves the following steps [3].

i. Data Collection
ii. Data Preprocessing
iii. Feature selection and feature extraction

Data collection involves collecting the data or accessing the standard training data developed by someone else.

Data preprocessing has to be done to remove any observation with missing value. Also, data will have numeric features as predictors but all those predictors are not necessarily helpful in performing the classification. Instead, they will increase the computation complexity involved with the algorithm [4]. Therefore, such redundant features can be eliminated using some feature selection method. Feature selection also helps in avoiding the data overfitting, reduces the training time, makes it easy to interpret the data and reduces the data error. Resultant data is then used for training the model using a suitable learning algorithm.Common methods used for feature selection are:

i. Filter Methods
ii. Wrapper Methods
iii. Embedded Methods

In case of Filter Methods, most relevant features are selected based on statistical measures. The statistical parameters include Information gain, chi-square test, Fisher score, correlation coefficient and variance threshold. This method does not depend upon the type of algorithm and requires very less computational time.

Wrapper method depends on classifier for its operation and hence best set of features will be selected based on classifier output. Therefore, the wrapper methods are considered to be computationally expensive but they are more accurate than filter methods. Some wrapper methods include Recursive feature selection algorithm, sequential feature selection algorithm and Genetic algorithms.

Embedded methods will make use of the combination of learning and hybrid learning methods and hence they offer more accuracy than other two methods. Most commonly used method Embedded method is Random Forest [5].

Training and developing a Model can now be implemented by choosing a proper Machine learning method. However, choosing a method for modelling depends upon the kind of data for which the model has to be developed.

Depending upon the type of data used, Machine Learning algorithms can be classified under 3 groups [6].

i. Supervised Learning
ii. Unsupervised Learning
iii. Reinforcement Learning

Supervised learning involves developing an algorithm using the labelled data. Features provided in the data consists of Predictors and a Response Variable. Training model, developed to classify a problem using the predictors and Predicted Response forms the output of trained Model.

Supervised learning is commonly used for two kinds of problems: Classification and Regression. If the labelled output for any input is discrete valued, then it forms a classification problem and if the output takes continuous values, it leads to regression of the input.

Classification involves multi-layer perceptron approach, Instance Based learning approach and Support Vector Machine [5]. Few examples include: Decision tree, Nearest Neighbor, Naive Bayes and Neural Network algorithms. Regression problem involves Decision Tree, Linear regression and neural networks.

In [6], author discussed about performance of few Machine learning based classifiers and compared the various performance parameters. Paper includes the discussion on our experimental results obtained through simulation using MATLAB.

In **Unsupervised learning**, dataset will not include labels and they require minimum human supervision. Approaches used in unsupervised learning methods include clustering, Anomaly detection, Neural networks and latent variable models [7]. Few algorithms used in unsupervised learning are: Deep learning, K-Means clustering, Fuzzy C means, Neural network and Gaussian mixture [8].

Reinforcement learning is the process of learning through interaction. During this learning phase, there will be both success and failures. Every success will receive a reward and agent tries to collect maximum rewards, thereby optimizing the learning process. Generally, Markov Decision Process and Bellman optimality equations form the core in the formulation of reinforcement learning problems [9].

Reinforcement learning finds it's applications in Natural Language Processing, Robotics, Healthcare electric power systems, Finance, Transportation Systems, Marketing and games.
Few Reinforcement learning algorithm include:
Tabular Method that represents the exact value function and exact method in table.
Approximation solution method where the action value function is represented as a parameterized function approximator
Monte Carlo method and Policy based RL which is effective in continuous and high dimensional spaces but has a drawback of converging to local minima instead of global minima.

VANET Attacks

In this paper, majorly 3 types of attack were discussed: Port Scan attack [10], DDoS attack [11] and Dos Hulk attack. Along with these 3 attacks, Benign data is also considered and tried developing the model for multiclassification.

III. CLASSIFICATION MODELS

In our Research paper, we are using the labelled data and hence made use of Supervised learning algorithm. In this section, a brief discussion on Decision Tree, Neural Networks, K-Nearest Neighbor and Support Vector Machines is done.

Decision Tree

Decision tree is a flowchart like tree structured supervised learning algorithm. It starts with a Root node and branches further to generate leaf nodes. A root node and leaf nodes represent an individual features or attributes from given set of predictors. Later, decision tree can be easily converted into set of classification rules. Decision tree gives similar or better performance compared to other classification algorithms [12]. Iterative Dichotomiser 3 (ID3) is a simple algorithm based on Decision Tree. ID3 is a top-down, greedy search approach to test each attribute at every node. Generally, Decision tree uses Entropy and Information Gain of each feature to decide about root node and leaf nodes at each level of the tree. Decision trees are helpful in generating the visualization of probabilistic business model and they are widely used in intrusion detection system to automatically generate the rules for intrusion detection. Decision Tree also finds its application in image processing, E-commerce and Medical research [13].

Different kinds of Decision trees are used based on the nature of applications. Namely,

(1) Classification tree

(2) Regression tree

(3) Decision tree forest

(4) Classification and regression tree

(5) K-Means clustering

Advantages of Decision tree include: No data preprocessing is required, no assumption on distribution of data is required and it can handle collinearity more efficiently. Also, DT provides better understandable explanation over the prediction.

The major disadvantages include data overfitting, which can be avoided using tree pruning.

Neural Networks

Artificial Neural Networks abbreviated as ANN, as the name says are the man- made neural networks. The motivation for ANN is the capability of human brain that plays a major role in taking decision for the complex problems. ANN is robust to noise and it can be used when the target output is real, discrete and is a vector of real or discrete values. Even though the training time for ANN is lengthy, its evaluation time is very less and better accuracy from the ANN models made this very popular.

An ANN model consists of 3 layers: Input layer, Hidden layer and an output layer [14]. There are single layers at input and output stage but the hidden layers can be in more than one stage. Also, the number of nodes in each layer depends upon the type of problem for which the model has to be developed. ANNs are based on units called Perceptron. Each perceptron will have multiple weighted inputs and a single output branch. This

output will in turn be connected as input to multiple nodes of the successive layer. From the obtained output, error will be measured and feedback is then used to modify the weight values for the input. This entire process is generally based on Backpropagation algorithm.

In our work, Medium neural network and Bilayer Neural network were used. Medium Neural network has 1 fully connected network with first layer size being 25. Bilayer Neural network has 2 fully connected networks with each first and second layer size equal to 10.

K-Nearest Neighbor

K-Nearest Neighbor (KNN) is an Instance Based learning method which can be used to approximate a real valued or discrete valued function. This method is also called as lazy learning method because, this method simply stores the given training data and when a new query is encountered [15], set of similar examples are retrieved from the memory and new query is classified based on the retrieved data. Therefore, the cost of classifying the new instance is very high. Another associated disadvantage is that, KNN uses all attributes to perform classification and these irrelevant attributes might reduce the classification accuracy.

KNN algorithm considers all instances as points in the n-dimensional space. Nearest neighbor for any point is calculated using the Euclidean distance

$$d(xi, xj) = \sqrt{\sum_{r=1}^{n}(ar(xi) - ar(xj))^2} \quad (1)$$

In KNN training algorithm, one just need to add training example to the database. In KNN classification algorithm, find the nearest neighbors to the query given and return the most common value among the neighbors. The number of neighbors depends upon the value of N. In our work the k value of 1 and 10 are used.

Support Vector Machines

Support vector machines are the supervised learning models that will analyze data for both classification and regression, but mostly in classification problems [16]. It creates a hyperplane for a linearly separable data in a n dimensional space where n represents the number of features. When the data cannot be separated by a linear plane, SVM classifier uses a technique called Kernel and convert the data to a linearly separable data. Linear SVM kernel is generally used when a huge number of features (>1000) are used because it is more likely that the data is linearly separable in high dimensional space [17].

Some of the advantages associated with SVM are, it has clear margin of separation and it uses a subset or part of training points called support vectors and hence it is memory efficient

The associated disadvantages are that SVM performance degrades when the dataset is large and noisy.

IV. COMPARISON OF CLASSIFICATION MODELS

Table 1 shows the comparison of various classification models that we used in our experiment

Table1. Comparison of Decision Tree, ANN, KNN and SVM

	Classification algorithms			
	Decision Tree	ANN	KNN	SVM
Model used	Discriminative	Discriminative	Discriminative	Discriminative
Type of problems solved	Classification and Regression	Classification and Regression	Classification and Regression	Classification and Regression
Type of soutions supported	Nonlinear	Both linear and nonlinear	Both linear and nonlinear	Both linear and Nonlinear
Complexity and Speed of the algorithm	Comparatively simpler and Faster	Training time is longer but evaluation time is very less	Relatively slow when it comes to prediction but quick to setup the model	Computationally complex and Slower
Peformance	Better for categorical data	It can handle very large	Less compared to	Better when the data is

	Classification algorithms			
	Decision Tree	ANN	KNN	SVM
	and handles the colinearilty better than SVM	dataset and provide good accuracy	SVM and ANN	small and total number of features is more
Interpretability	Easy	Hard	Hard	Easy for linear and hard for other kernel types
Type of Predictors	Numeric, Categorical, some categorical and some numeric	Numeric, Categorical	Mostly Categorical	Numeric, Categorical

V. PERFORMANCE METRICS

Performance metrics such as True Positive, False Positive, True Negative, False negative, Confusion matrix, Accuracy, Sensitivity, Specificity and Misclassification can be considered for the measurement of system performance [18][19][20].

(1) True Positive: When the actual classification of an instance is positive and the developed model classifies that instance as positive, it is called True Positive.

(2) True Negative: When the actual classification of an instance is negative and the developed model classifies that instance as negative, it is called True Negative.

(3) False Positive: When the actual classification of an instance is negative and the developed model classifies that instance as positive, it is called False Positive. It is even called as Type I error

(4) False negative: When the actual classification of an instance is positive and the developed model classifies that instance as negative, it is called False Negative. It is even called as Type II error

(5) Confusion matrix: It is a matrix listing the number of True Positive, False Positive, True Negative and False negative for the categorical values.

(6) Accuracy: It is the percentage of predictions our model has got right out of all predictions. It is given by

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (2)$$

(7) Sensitivity or recall: It is used to calculate the model's ability to correctly predict the positive values

$$Recall = \frac{True\ positive}{actual\ positive} = \frac{TP}{(TP+FN)} \quad (3)$$

(8) Specificity: It is used to calculate the model's ability to predict the negative values.

$$Specificity = \frac{True\ Negative}{actual\ Negative} = \frac{TN}{(TN+FP)} \quad (4)$$

(9) Misclassification: It is used to find the percentage of instances that are wrongly classified.

$$Misclassification = \frac{(FP+FN)}{(TP+TN+FP+FN)} \quad (5)$$

(10) Precision: It is used to calculate the model's ability to classify positive values correctly.

$$Precision = \frac{True\ Positive}{Predicted\ positive} = \frac{TP}{(TP+FP)} \quad (6)$$

VI. Discussion on Dataset

Dataset

For the data analysis, data used is part of CICIDS dataset 2017 downloaded from https://www.unb.ca.cic/datasets/ids-2017.html. This dataset has 2,830,743 number of rows found by considering 14 various types of attack and a normal traffic named as Benign. This data represents very closely to the real-world network traffic data. In the dataset, as discussed earlier, DoS Hulk, Dos attack and Port Scan attack along with data for Benign case were considered. The impact of data size on model's performance using various classification algorithms were studied. Dataset has totally 85 features in it out of which 75 features/attributes have been extracted after data preprocessing. Various models have been developed with proper training and validation using MATLAB. Details of various model performance is mentioned in further sections.

The dataset has been extracted for 2 different sets. One is the Biased dataset and other is the Unbiased Dataset. Training phase always makes use of Biased data but in most of the cases it is not possible to obtain equal amount of data for all the cases. Hence, the impact of unbiased data is studied while developing the model.

Biased Dataset

For the biased data set, 10 different cases were considered as follows:

Table 2 Biased Dataset details

	DATA SIZE			
	DDoS	PortScan	DoS Hulk	Benign
Data1	10000	10000	10000	5000
Data2	10000	10000	10000	3000
Data3	5000	10000	10000	10000
Data4	3000	10000	10000	10000
Data5	3000	10000	10000	3000
Data6	3000	3000	10000	3000
Data7	10000	10000	10000	NIL
Data8	NIL	10000	10000	10000
Data9	10000	10000	NIL	10000
Data10	10000	NIL	10000	10000

Unbiased Dataset

It has data of all 3 types of attacks and Benign data in equal size. In order to study the model efficiency with varying data size: data size of 2K each, 5K each ,7K each and 10K each were considered.

1. RESULT AND DISCUSSIONS

Result for Unbiased Data

The result analysis presented here for both Unbiased and the biased data. Unbiased data has a very good model efficiency as shown below.

Table 3 Validation Accuracy in Unbiased data (in percentage)

	Type Of Attack			
Type of ML Algorithm	Benign	DDoS	DoS Hulk	PortScan
Bi-layered NN	98.4	99.6	99.6	99.6
Decision Tree	87.3	99.3	94.8	96.3
KNN(N=1)	98.3	99.7	99.8	99.7

KNN(N=10)	95.9	99.1	99.2	99
Medium NN	98.6	99.7	99.8	99.7
SVM	96.8	99.7	99.9	99

Table 3 shows the validation accuracy of different ML algorithms for Unbiased data. From the result, it is clear that Decision Tree is the worst performer whereas Support Vector Machine and Neural networks are the best performers. When the models were developed by giving test data of different size, it was observed that Decision Tree algorithm performs worst where as SVM and Neural networks are the best performers. Graph in Figure 1 shows the classification accuracy of the test data (along Y axis) plotted against different data size (along X axis) for various types of attacks (one graph for each attack) by considering various ML models. From the graph, it can even be observed that the classification accuracy in DoS hulk attack is less compared to any other attacks. Same observation is done even for Biased data which shall be discussed in the later part of this section.

Table 4 Precision in multiclassification ML models (in percentage)

	Type Of Attack			
Type Of ML Algorithm	Benign	DDoS	DoS Hulk	PortScan
Bi-layered NN	99	99	99	98
Decision Tree	96	99	98	92
KNN(N=1)	99	99	99	99
KNN(N=10)	97	98	99	98
Medium NN	99	99	99	98
SVM	98	98	99	99

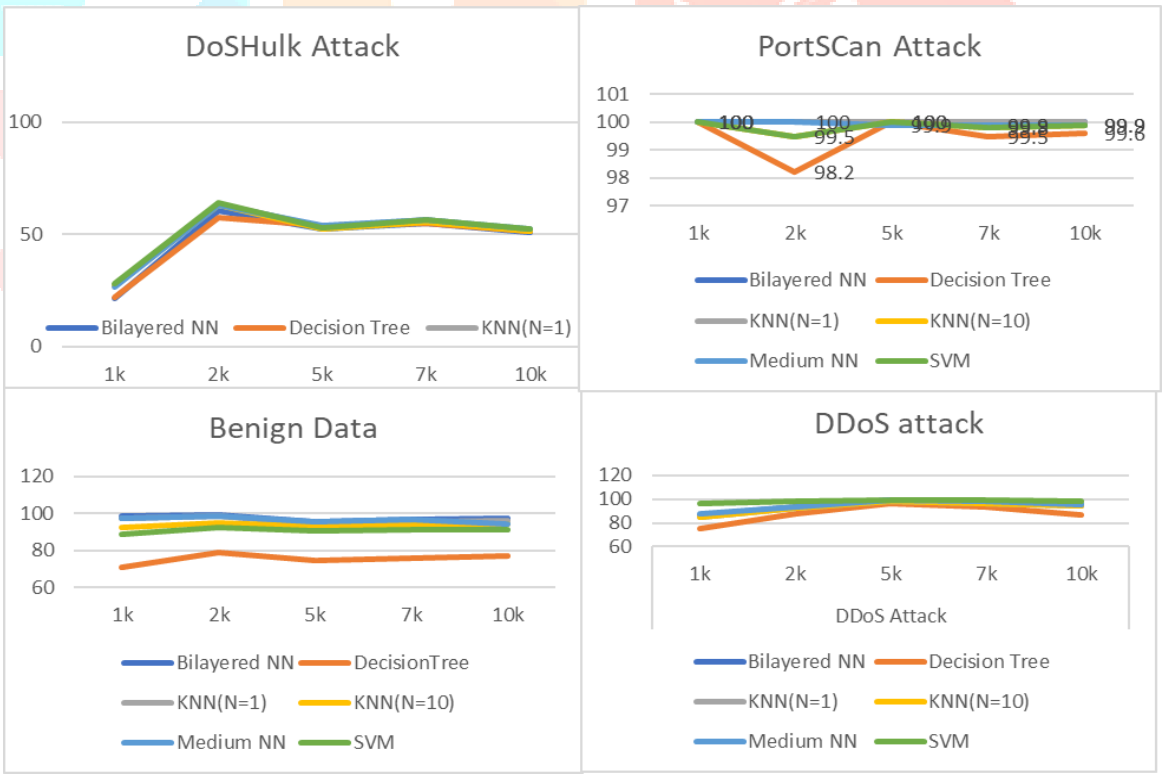


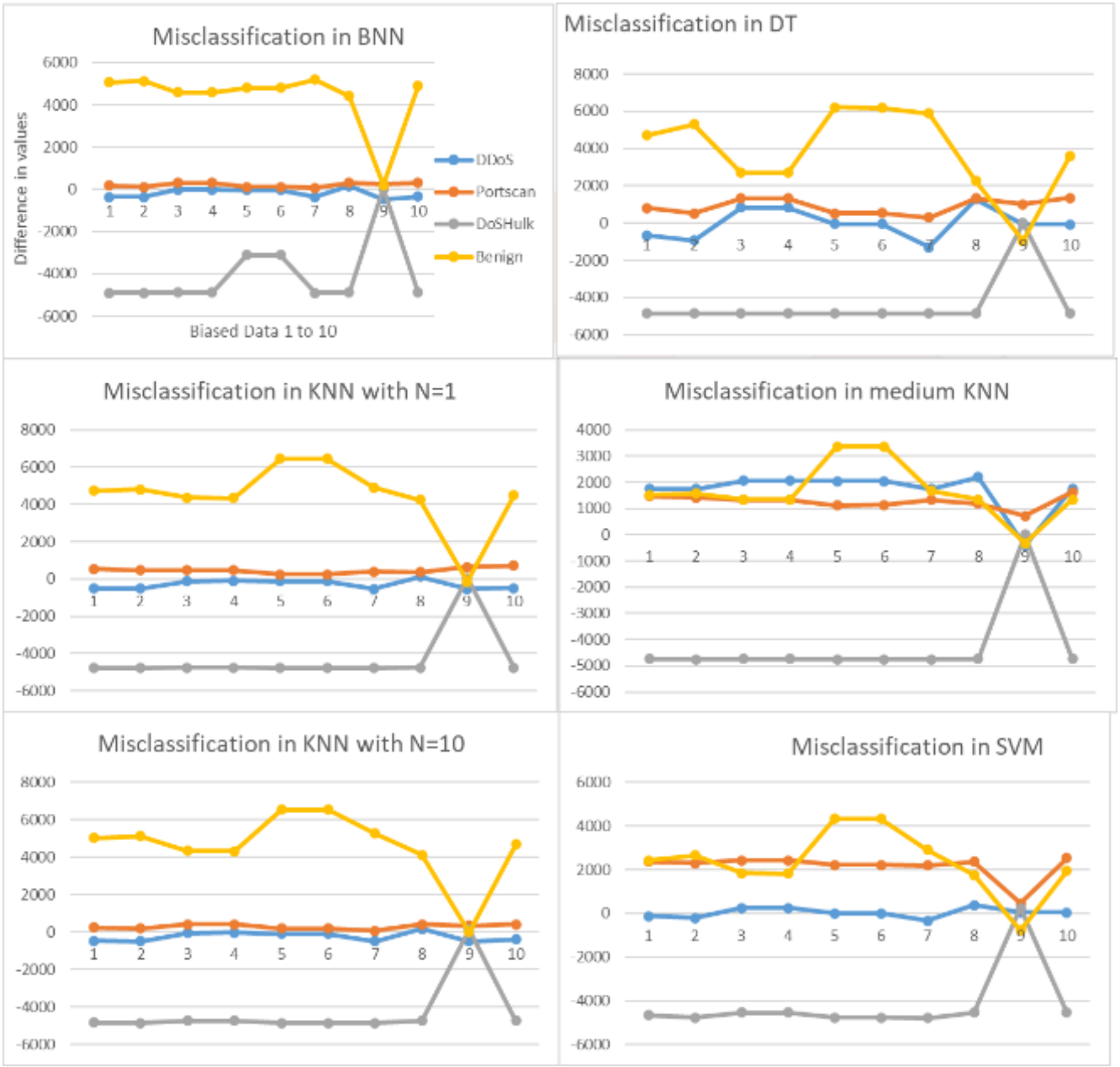
Fig. 1 Classification accuracy in different Machine Learning Models

Table 4 shows the Precision of each ML model. As discussed earlier, Precision tells the model's ability to correctly saying the attack as attack. In the vehicle security, Precision plays a major role and model must process a Precision close to 1. Otherwise, attack will not be detected and the purpose of developing the IDS will fail. However, it was observed that KNN, Neural Networks and SVM are giving the higher Precision value and Decision Tree has the lowest value of Precision. Therefore, SVM and NN models were preferred for IDS.

Result for Biased Data

In our experiment, 10 data sets were considered for biased data and each of these test data were tested for 6 different ML models. So, the results of Unbiased data are available for 60 different cases. 10 cases are plotted in one graph by considering data 1 to data 10 along the X axis (can be referred in table 2) and the corresponding misclassification is plotted along Y axis.

Firstly, we can observe here is that, the models which were giving such a high accuracy when trained with the balanced or unbiased data are performing very poor when the biased data is used to train the model. The attacks were misclassified for a greater number of cases. From all those graphs, it is very clear that Dos Hulk has got the worst classification accuracy and also it was observed that Dos Hulk has been misclassified as Benign by almost all ML models except medium K Nearest Neighbor. The reason for Dos Hulk to be misclassified as Benign. It is because Benign and DoS Hulk have somewhat similar pattern and the biased data with lesser data for Benign case is leading the model to miss out the patterns of data in Benign case and hence it is treating the Dos Hulk as Benign.



VII. CONCLUSION

VIII. In conclusion, our analysis of the CICIDS dataset 2017 has provided valuable insights into the performance of various machine learning models in intrusion detection systems (IDS). By utilizing both biased and unbiased datasets, aimed to understand the impact of data distribution on model efficiency. Our findings indicate that while biased datasets may simplify training procedures, they often lead to decreased accuracy and misclassification, particularly evident in the misclassification of DoS Hulk attacks as benign traffic. This underscores the importance of balanced datasets in capturing the nuances of different attack types. Furthermore, our results highlight the superior performance of support vector machines (SVM) and neural networks (NN) in terms of precision, essential for effective threat detection. Moving forward, these observations emphasize the critical role of dataset selection and model choice in developing robust IDS capable of accurately identifying diverse cyber threats in real-world network environments. Future work could explore techniques to mitigate data bias and improve model generalizability across different network Results and Discussion

REFERENCES

[1] https://www.sciencedirect.com/topics/computer-science/vehicular-ad-hoc-network

[2] https://www.unb.ca/cic/datasets/ids-2017.html

[3] B. Venkatesh, J. Anuradha, “A Review of Feature Selection and Its Methods”, CYBERNETICS AND INFORMATION TECHNOLOGIES, Volume 19, No 1,

[4] T. R. N and R. Gupta, "Feature Selection Techniques and its Importance in Machine Learning: A Survey," IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), Bhopal, India, pp. 1-6

[5] Dasgupta, Ariruna & Nath, Asoke. (2016). “Classification of Machine Learning Algorithms”, 10.6084/M9.FIGSHARE.3504194.

[6] Hammoudeh, Ahmad., "A Concise Introduction to Reinforcement Learning", 10.13140/RG.2.2.31027.53285.

[7] Siadati, Saman. , “What is UnSupervised Learning”, 10.13140/RG.2.2.33325.10720.

[8] M. Usama et al., "Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges," in IEEE Access, vol. 7, pp. 65579-65615

[9] M. Usama et al., "Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges," in IEEE Access, vol. 7, pp. 65579-65615

[10] J. Gadge and A. A. Patil, "Port scan detection," 2008 16th IEEE International Conference on Networks, New Delhi, India, 2008, pp. 1-6, Doi: 10.1109/ICON.2008.4772622. keywords: {Reconnaissance; Phase detection; Operating systems; Fingerprint recognition; Computer networks; Access protocols; Software systems; Application software, Internet, Computer security},

[11] Mukhopadhyay, Debajyoti & Oh, Byung-Jun & Shim, Sang-Heon & Kim, Young-Chon. (2010). A Study on Recent Approaches in Handling DDoS Attacks.

[12] Mohd Shahrizan Abd Rahman, Nor Azliana Akmal Jamaludin, Zuraini Zainol, Tengku Mohd Tengku Sembok, The Application of Decision Tree Classification Algorithm on Decision-Making for Upstream Business, International Journal of Advanced Computer Science and Applications, Vol. 14, No. 8

[13] Hajjej, Fahima & Alohali, Manal & Alazzam, Malik & Rahman, Md. A Comparison of Decision Tree Algorithms in the Assessment of Biomedical Data. BioMed Research International..

[14] Vasileiadis, Alexandos & Alexandrou, Eirini & Paschalidou, Lydia & Chrysanthou, Maria & Hadjichristoforou, Maria. Artificial Neural Network and Its Applications..

[15] Munazhif, Nanda & Yanris, Gomal & Hasibuan, Mila. Implementation of the K-Nearest Neighbor (kNN) Method to Determine Outstanding Student Classes. SinkrOn. 8. 719-732. 10.33395/sinkron.v8i2.12227.

[16] Rodríguez-Pérez, R., Bajorath, J. Evolution of Support Vector Machine and Regression Modeling in Chemoinformatics and Drug Discovery. J Comput Aided Mol Des 36, 355–362 . https://doi.org/10.1007/s10822-022-00442-9

[17] Q. Wang, "Support Vector Machine Algorithm in Machine Learning," IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian, China, pp. 750-756.

[18] Bradley J. Erickson , Felipe Kitamura, Performance Metrics for Machine Learning Models

[19] Orozco-Arias, S.; Piña, J.S.; Tabares-Soto, R.; Castillo-Ossa, L.F.; Guyot, R.; Isaza, G. Measuring Performance Metrics of Machine Learning Algorithms for Detecting and Classifying Transposable Elements. *Processes* *8*, 638. https://doi.org/10.3390/pr8060638

[20] Rainio, O., Teuho, J. & Klén, R. Evaluation metrics and statistical tests for machine learning. Sci Rep 14, 6086 https://doi.org/10.1038/s41598-024-56706-x