R.V. COLLEGE OF ENGINEERING

DEPARTMENT OF TELECOMMUNICATION ENGINEERING

2018-2019

Technical Seminar Presentation(phase-2) on

Application Of Machine Learning In Network Security

Under The Guidance of
Mr. Pawan Kumar B
Assistant Professor

Presented By:
Aditya Vikram
1RV15TE003

Introduction:

- Rapid growth and development in wireless networking domain has led to increase in wireless devices like APs and IoT devices.
- This explosive growth and increase in the usage of networks has led to increase in attacks.
- Network Intrusion Detection (NID) systems are used to identify malicious network activity both at the network end point and inside threats.
- Traditional Rules-based Network Intrusion Detection are no longer able manage these unknown attacks.
- Need for automated detection systems which can learn from previous attacks and detect new ones.

- Supervised Machine Learning

 Approach in which models are trained on labeled data set and learns a decision boundary.

- Machine learning based models are generally employed for:

 Network Traffic Classification

Pivotal significance owing to high growth in the number of internet users. It is very crucial for internet service providers (ISPs) to keep an eye on the network traffic.

 Network Intrusion Detection System

It is a system that attempts to detect hacking activities, denial of service attacks etc. and aims to identify patterns in data that do not conform to expected behaviour.

Advantages:

- **Machine Learning based systems can handle the volume** and automates the process of detecting advanced threats.
- **Machine Learning techniques for cybersecurity can learn over time**. AI can identify malicious attacks based on the behaviors of applications and the behavior of the network as a whole.
- **Machine Learning models identifies unknown threats**. Hundreds of millions of malicious attacks are launched every year and rules-based intrusion detection system are not able to detect new threats.

Applications:

- Network Intrusion Detection
- Social Network Spam Detection
- UEBA for internal threat detection
- Machine learning for end-point threat detection
- DDoS attack detection for IoT devices

Literature Review:

- G. Karatas, O. Demir and O. Koray Sahingoz (2018)

 In this paper, it is aimed to survey deep learning based intrusion detection system approach by making a comparative work of the literature and by giving the background knowledge either in deep learning algorithms or in intrusion detection systems.[1]

- R. Patgiri, U. Varshney, T. Akutota and R. Kunde (2018)

 In this paper, the authors have conducted a rigorous experiment on Intrusion Detection System (IDS) that uses machine learning algorithms, namely, Random Forest and Support Vector Machine (SVM). [2]

- H. Azwar, M. Murtaz, M. Siddique and S. Rehman, (2018)

 The security phases of intrusion detection using machine learning approach have been deliberated in this paper. Within this paper a broad assessment of the current datasets by means of data preprocessing is made.[3]

- N. Chaabouni, M. Mosbah, A. Zemmari, C. Sauvignac and P. Faruki(2019)

 In this survey, the main focus is on IoT NIDS deployed via Machine Learning since learning algorithms have a good success rate in security and privacy. The survey provides a comprehensive review of NIDSs deploying Internet of Things.[4]

- X. Bao, T. Xu and H. Hou,(2016)

 The paper first gives an introduction to the basic concept of intrusion detection and the basic principle of the classifier based on support vector machine, then discusses algorithm of support vector machine, and finally forms network intrusion detection system.[5]

Basic Workflow for ML classification:

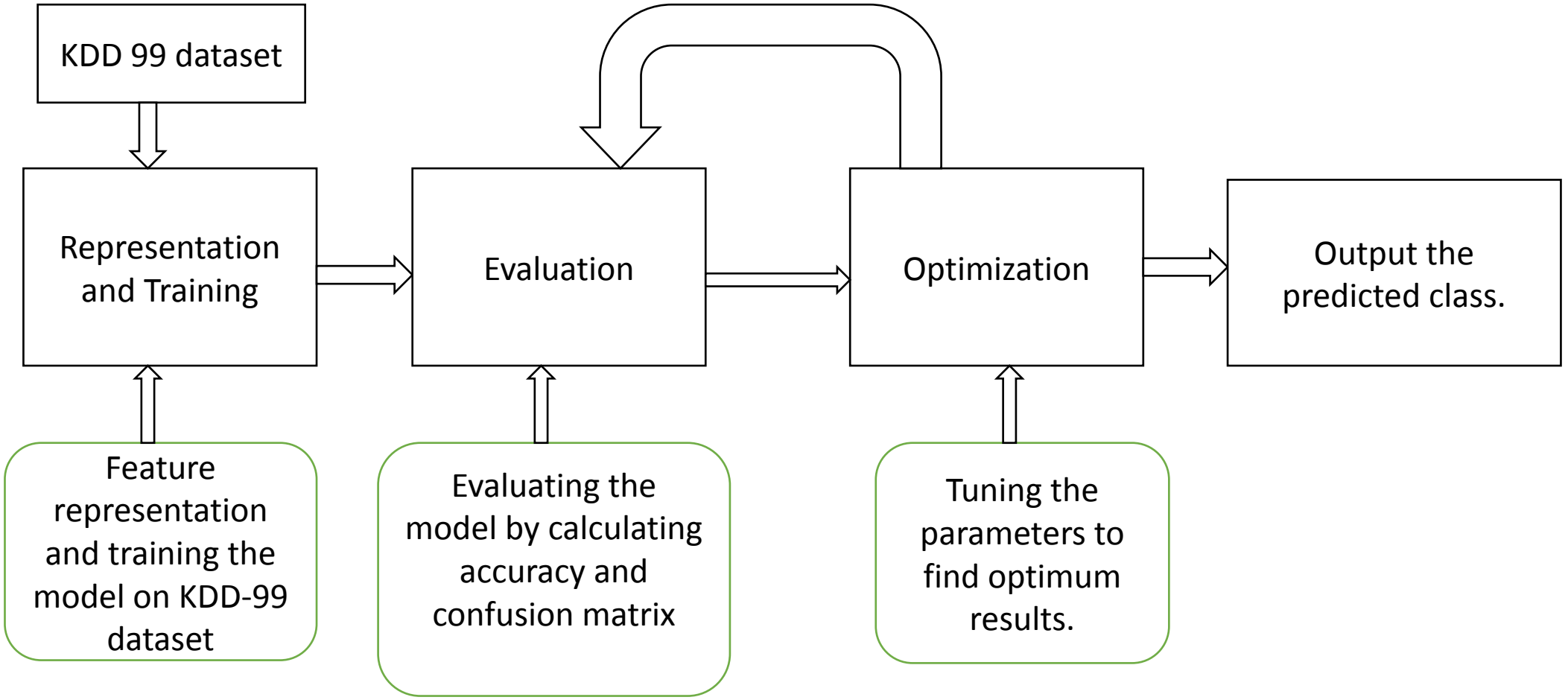
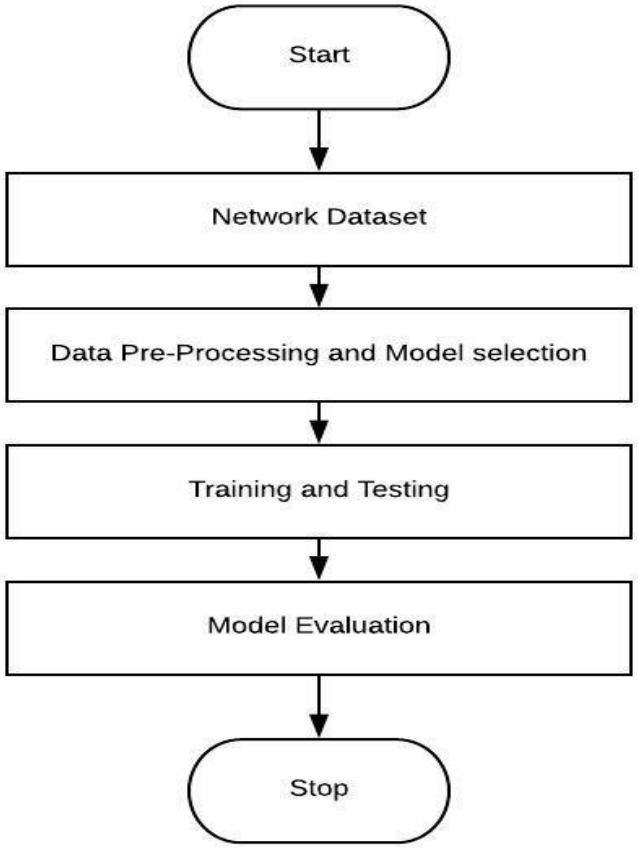


Figure 1: This figure represents a workflow for the given classification of attacks.

Flow chart:



- KDD 99 dataset was used in this project and dataset is divided into five categories.
- Data cleaning and pre-processing was performed.
- Train and test split was performed.
- Test set includes 6000 samples and training set includes 29000 samples.
- The model was trained on SVM classifier and tested using accuracy and Detection rate.

Figure 2: This flowchart represents the workflow for the network intrusion detection model.

KDD 99 Dataset:

- Data set can be divided into five categories.
- These categories are:

1.) DOS 2.) Probe 3.) R2L 4.) Normal 5.) U2R

- Contains 41 features to predict the required class.

Type	*Number of samples in training set*	*Number of samples in test set*
Normal	14000	2900
DOS	10287	1200
U2R	52	20
Probing	4107	806
R2L	1126	1074
Total	29572	6000

Figure 3:This table represents the no. of samples for each category.

Dataset Features and their data types:

Number	*Feature name*	*Data type*
1	duration	continuous
2	protocol_type	discrete
3	service	discrete
4	flag	discrete
5	src_bytes	continuous
6	dst_bytes	continuous
7	land	discrete
8	wrong_fragment	continuous
9	urgent	continuous
10	hot	continuous
11	num_failed_logins	continuous

Figure 4:This Table represents a fraction of features in the KDD 99 dataset.

Data Pre-processing

- The non-numeric values and char values were encoded and changed into numeric values.
- Data cleaning was performed to remove missing data.
- Feature scaling was done to get various features in comparable range.
- It was determined that different samples had different influence on output.
- This was done using Min-Max normalization:

$$X_i = \frac{V_i - \min(V_i)}{\max(V_i) - \min(v_i)}$$

where, vi is the actual value of the feature, and the maximum and minimum are taken over all values of the feature.

Support Vector Machines:

- Supervised machine learning classification algorithm.
- The algorithm outputs an optimal hyperplane which categorizes new examples based on labeled training examples.
- In LSVM, it is a linear decision boundary which divides the 2D plane in two parts.
- LSVM is maximum margin classifier which means it selects the decision boundary with maximum margin from both classes.
- This ability makes SVMs very accurate and stable.

Working Principle of Support Vector Machine:

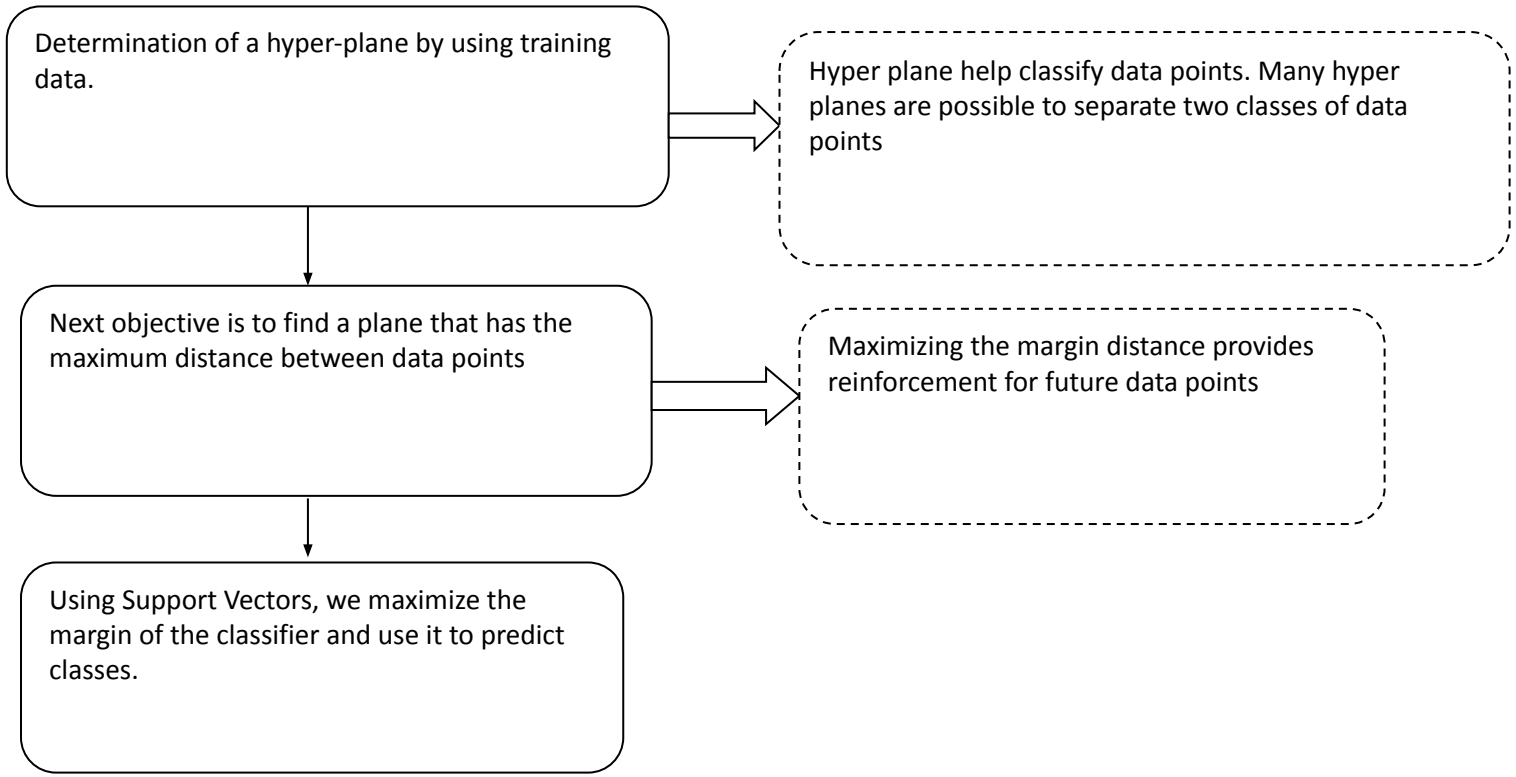
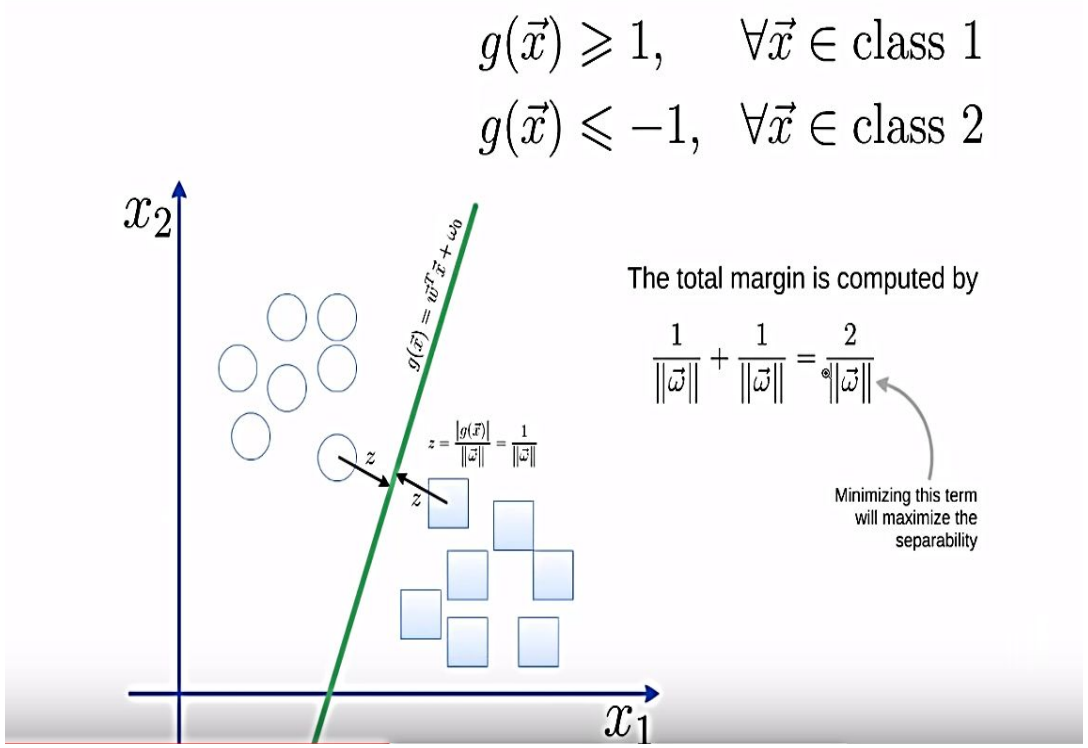


Figure 5:This represents the work flow in SVM algorithm used for supervised classification.

Working Of SVM:



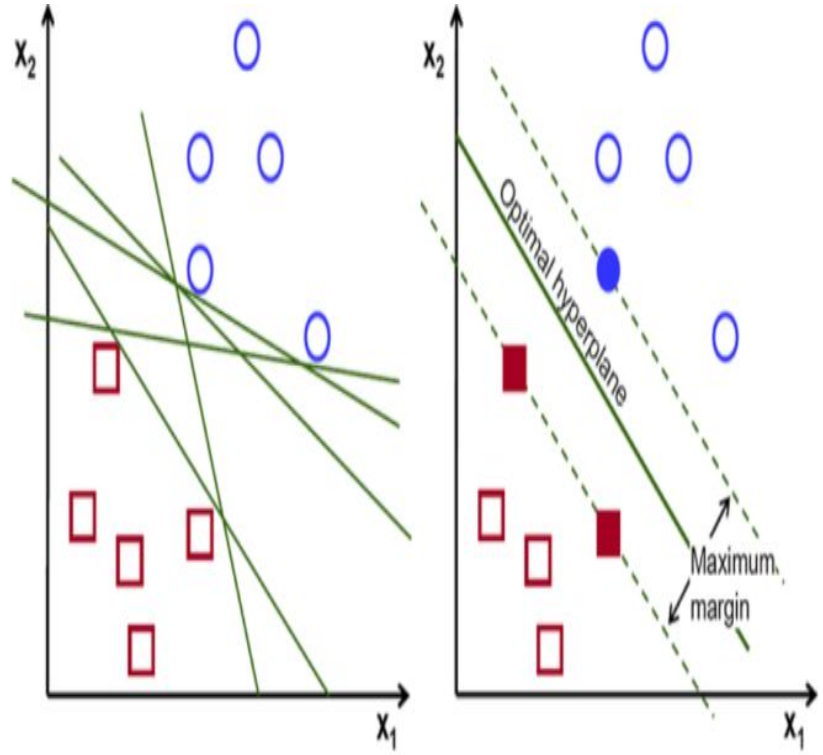


Figure 6 & 7: This figure represents selection of optimal hyperplane in SVM using maximizing of margin.

Design:

- SVM classifier comes up with linear decision boundary.
- Maximising the margin:

$\frac{1}{||w||} + \frac{1}{||w||} = \frac{2}{||w||}$ minimizing ||w|| will maximize the margin.

- Mathematical Formulation:

$$g(x)=wx(i)+b$$

Where, x is a training example and w is the weight and b is the bias term.

- For $g(x) \geq 1$ predict class 1.
- For $g(x) \leq -1$ predict class 2.

Advantages:

- Easy and simple to train.
- Fast prediction
- Scales well in large data sets.
- High accuracy as compared to other models.
- Works well with sparse data.

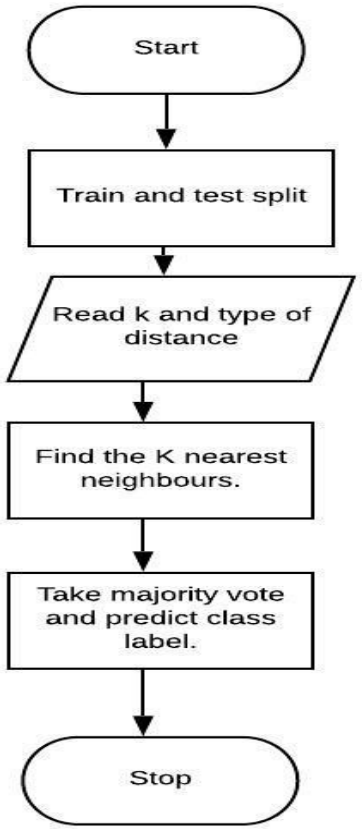
kNN Classifier:

- kNN classifier is based on a distance function that measures the difference or similarity between two instances.
- kNearest Neighbor classification (kNN) algorithm was used prominently in the NIDS even though some drawbacks due to its better detection rates.
- kNN is a lazy learner and takes much computational time as well ashigh storage cost.
- The standard Euclidean distance d(x, y) between two instances x and y is defined in as:

$$d(x,y)=\sqrt{\left(\sum_{i=1}^{n}(x(i)-y(i))\right)^2}$$

where,x(i) is the ith featured element of the instance ,y(i) is the ith featured element of the instance and n is the total number of features in the data set.

Flowchart For kNN:



Evaluation Metrics:

- The accuracy can be calculated as:

  Accuracy=No. of correct predictions /(Total no. of predictions)

- Detection rate is calculated to evaluate the model.

  Detection rate = $TN/(TN + FP)$ (1)

- TP represents the normal sample is correctly predicted, FP represents that abnormal samples are predicted as normal,FN is when normal samples are predicted as abnormal and TN represents that abnormal behavior is predicted correctly.

Results for kNN:

Accuracy			
k Value	3	5	10
U2R	0.9996	0.9995	0.9994
R2L	0.9988	0.9983	0.9976
Probe	0.9981	0.9976	0.9966
DOS	0.9994	0.9991	0.9985

This table represents the accuracy for various k values model.

Results:

Number of Features	Accuracy	Detection Rate
9	97.80 %	93.66 %
7	98.47 %	95.61 %
5	96.95 %	91.21 %
4	98.39 %	95.37 %

Figure 6: This table represents the accuracy and detection rate obtained for various features model.

Conclusion:

- The model developed can predict network intrusion detection for different classes of attacks.
- SVM and kNN were analysed as the required classifiers.
- Different no. of features give different accuracy and detection rates.
- Highest accuracy and detection rates was achieved for 7 features model for SVM model.
- kNN showed higher higher accuracy for smaller k values.

References:

[1] G. Karatas, O. Demir and O. Koray Sahingoz, "Deep Learning in Intrusion Detection Systems," *2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT)*, ANKARA, Turkey, 2018.

[2] R. Patgiri, U. Varshney, T. Akutota and R. Kunde, "An Investigation on Intrusion Detection System Using Machine Learning," *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, Bangalore, India, 2018.

[3] H. Azwar, M. Murtaz, M. Siddique and S. Rehman, "Intrusion Detection in secure network for Cybersecurity systems using Machine Learning," *2018 IEEE 5th International Conference on Engineering Technologies and Applied Sciences ,*Bangkok, Thailand, 2018.

[4] N. Chaabouni, M. Mosbah, A. Zemmari, C. Sauvignac and P. Faruki, "Network Intrusion Detection for IoT Security based on Learning Techniques," in *IEEE Communications Surveys & Tutorials*,2019.

[5] X. Bao, T. Xu and H. Hou, "Network Intrusion Detection Based on Support Vector Machine," *2016 International Conference on Management and Service Science*, Wuhan, 2016.

[6] Y. Chang, W. Li and Z. Yang, "Network Intrusion Detection Based on Random Forest and Support Vector Machine," *IEEE International Conference on Computational Science and Engineering (CSE)*, Guangzhou, 2017.

[7] R. Z. A. Mohd, M. F. Zuhairi, A. Z. A. Shadil and Hassan Dao, "Anomaly-based NIDS: A review of machine learning methods on malware detection," *2017 International Conference on Information and Communication Technology (ICICTM)*, Kuala Lumpur, 2017.

[8] Brao, Bobba & Swathi, K. “Fast kNN Classifiers for Network Intrusion Detection System*”, Indian Journal of Science and Technology*. 2017.