Network Attack Analysis Using Machine Learning

Abstract

Cyber threats result in compromising the confidential information of a particular person or the organization. Hackers may leak necessary information to the unknown which may be hazardous. Cyber-attacks are not only limited to bigger organizations but also individuals. So, the hacker can target anyone for the information. The tendency of large-scale deployment of essential corporate applications in cloud centers has risen dramatically as cloud computing has progressed.

To solve this problem a new concept called Threat Intelligence was introduced by the research developers. Due to its capacity to evaluate threat intelligence data from various online assaults, threat intelligence will become an essential tool for resolving sophisticated network attacks, real-time threat time indications, and assault monitoring. This project's major objective is to employ various machine learning techniques to examine the threat intelligence datasets in predicting network attacks coming either from a corporate network or from a specific cluster.

Introduction

- The network has become the main tool or weapon equipment for some malicious organizations to do evil, and the types of network threats are more complex and changeable.
- Threat Intelligence is a subset of intelligence focused on information security.

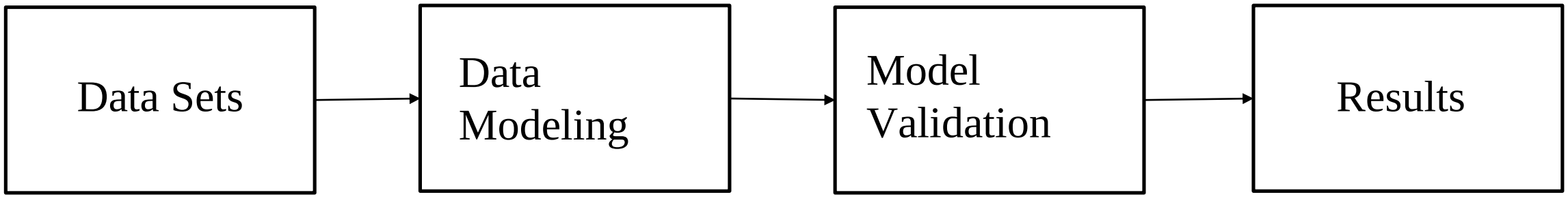


Figure 1: Overview of the proposed model

Cont..

- Security intelligence has become a lightweight important means to solve complex network attacks, threat early warning and security traceability.
- Therefore, it is very important to reasonably use threat intelligence, share security intelligence information in real time, and establish a set of border real-time warning and defense system driven by security intelligence.
- The main objective is to not to identify which algorithm is best, but to review the main dataset publicly available used to train and test security solutions that employ classification/regression algorithm. Hence KDD datasets were taken into considerations.
- Training on infrastructure specific datasets is more efficient from the security point of view than training an old attack signatures.

Literature Survey

Sl. No	TITLE	AUTHOR	YEAR OF PUBLICATION	LIMITATIONS
1	Cyber Threat Intelligence Model: An Evaluation of Taxonomies, Sharing Standards, and Ontologies within Cyber Threat Intelligence	Vasileios Mavroeidis, Siri Bromander	Feb 2021	Hurdles to achieve include a lack of dedicated ontological cyber threat intelligence efforts that can account for the strategic, operational, and tactical levels; ambiguity in defined concepts that prevents ontology integration and adoption
2	Detecting Vulnerabilities in Critical Infrastructures by Classifying Exposed Industrial Control Systems Using Deep Learning	Pablo Blanco-Medina , Eduardo Fidalgo , Enrique Alegre , Roberto A. Vasco-Carofilis , Francisco Jañez-Martino and Victor Fidalgo Villar	2018	Further improving the proposed solution, increasing our training images, and adding new layers on top of the given architectures based on VGG16 results.

Literature Survey

Sl. No	TITLE	AUTHOR	YEAR OF PUBLICATION	LIMITATIONS
3	A Cloud Computing Based Network Monitoring and Threat Detection System for Critical Infrastructures	ZhijiangChen, GuobinXu, VivekMahalingam, LinqiangGe, JamesNguyen, WeiYu∗, ChaoLu	2 November 2015	The researchers did not put the computational cost into consideration in the choice of which machine learning technique to use for spam mail filtering. Their main focus is performance in terms of classification accuracy.
4	Analysis of adversary activities using cloud-based web services to enhance cyber threat intelligence	Hamad Al-Mohannadi1 · Irfan Awan1 · Jassim Al Hamar	2020	Attacker tried to get access to root but failed. In a UNIX-based system, getting access to the root gives attackers full control to the system. Since the authentication was not confirmed, the attacker could not get the access.

Cont..

- Disadvantages
  - Lack of precision in attack detection.
  - It is difficult to control the access of network in real time.
  - Less security and privacy.
- Advantage of Proposed system
  - Predicts the precision about insider and outsider attacks.
  - Improves reliable data access control and can handle huge number of data.

Problem Identification

- With the development of the Internet and the application of information technology, the general trend is opener. Important data and important business applications will face more serious security threats.
- The network has become the main tool or weapon equipment for some malicious organizations to do evil, and the types of network threats are more complex and changeable.
- The existing security solutions are difficult to cope with the high level, continuous, group and weaponization threats.
- However, important industries and key information infrastructure security network units mainly support apt detection and defense equipment with high application threshold and general landing effect.

Objectives

- To analyze the security given from the insider and outsider attacks. Applications employ machine learning techniques to recognize and respond to assaults. Larger data sets of security events can be analyzed to find correlations in harmful activity and help with this. When similar actions are found, ML makes it if the trained ML model can intelligently handle them.
- It is crucial to investigate and use advanced techniques for the detection and reporting of such assaults because there are more and more cyberattacks targeting crucial networked resources that cannot be detected by conventional monitoring tools. Hence the main objective is to improve the accuracy of attack detection.

Proposed System

- As the previous sections explain the existing problems, the proposed system is being developed for analyzing the dataset using a Graphical User Interface. It mainly analyses the insider or outsider attacks to enhance the accuracy of attack detection. It improves reliable data access control and can handle huge amounts of data.
- Threat Intelligence is the process of identifying the threats and analyzing them in other words this feature informs the user about the current threats and the threats which are about to happen in the current scenario of the organization.
- Hence several machine learning algorithms are employed just to check the efficiency of the network attack blocking method. Threat intelligence feeds can help with this approach by detecting real-time indications of the breach and advising preventative measures.
- The results are examined utilizing network attack data sets, demonstrating that it is one of the best competitor data sets for threat intelligence analysis and testing.

Requirement Analysis

- Software Requirements
  - Anaconda
  - Python Programming language
- Software Packages
  - Keras
  - TensorFlow
- Software And Hardware Requirements
  - Operating system used is the Windows 8/10/11.
  - Processor – Ryzen 7@ 3.20 GHz
  - RAM – 8GB and above
  - Hard disk –20 GB is required to store the data set obtained.

Threat Intelligence

1. Threat intelligence, or cyber threat intelligence, is information an organization uses to understand the threats that have, will, or are currently targeting the organization. This info is used to prepare, prevent, and identify cyber threats looking to take advantage of valuable resources.
   - Ensure to stay up to date with the often overwhelming volume of threats, including methods, vulnerabilities, targets and bad actors.
   - Help one to become more proactive about future cybersecurity threats.
   - Keep leaders, stakeholders and users informed about the latest threats and repercussions they could have on the business.

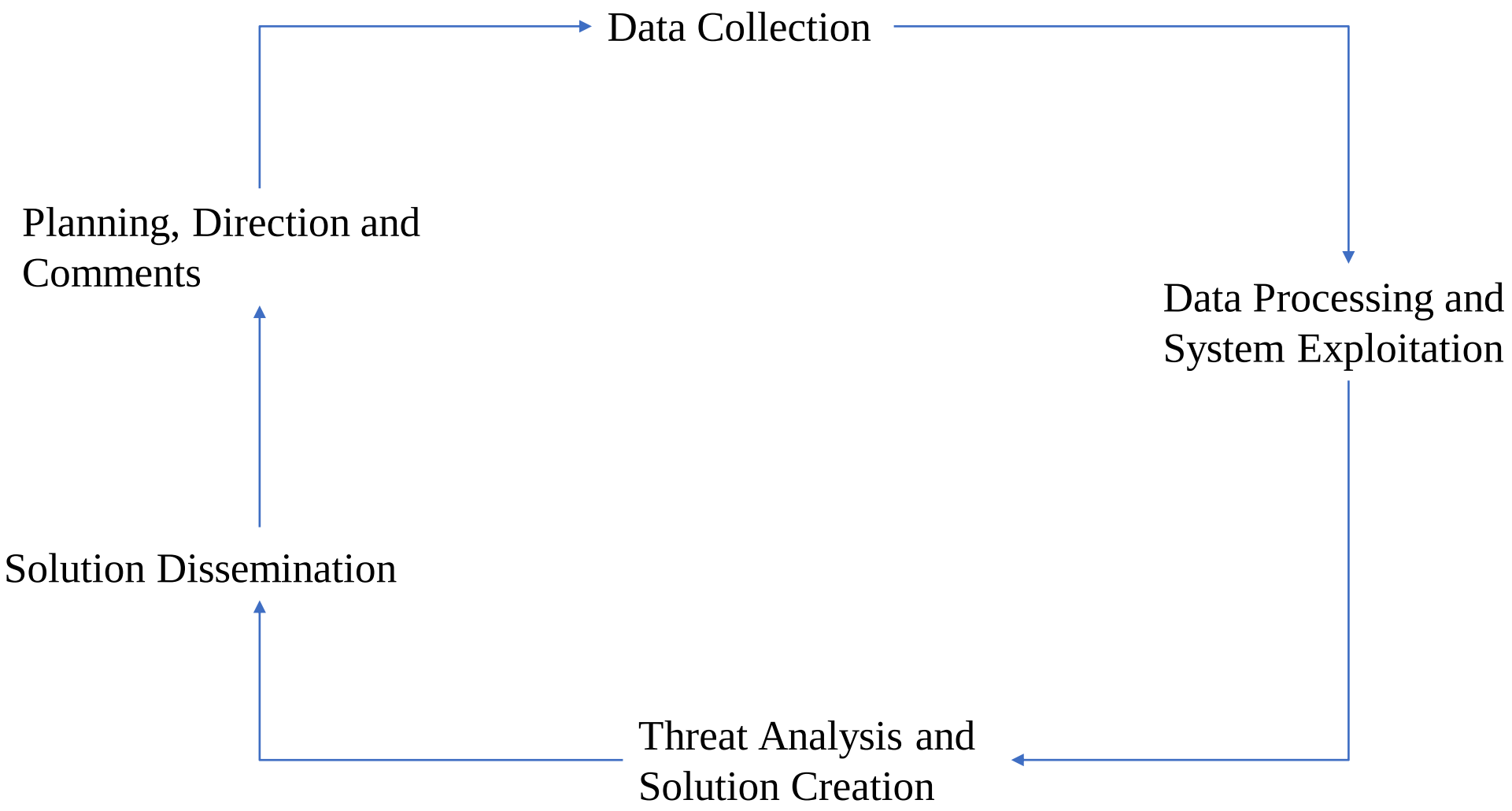


Figure 2: Life cycle of Threat intelligence

System Workflow

- The System Workflow description is briefly explained where the first step in creating a predictive model is to create a dataset. The dataset can be thought of as practically being an MN matrix, where M stands for the columns (features) and N for the rows (samples). One might divide columns into X and Y for example. The method of picking features, whether automatic mode or not, depends on which components are most important to the target variable or desired outcome.
- Algorithms' performance was accurately depicted after being evaluated against the aforementioned criteria. Some experts were just concerned with the algorithm's correctness, some were worried about how well the system performed when tested against every matrix parameter.

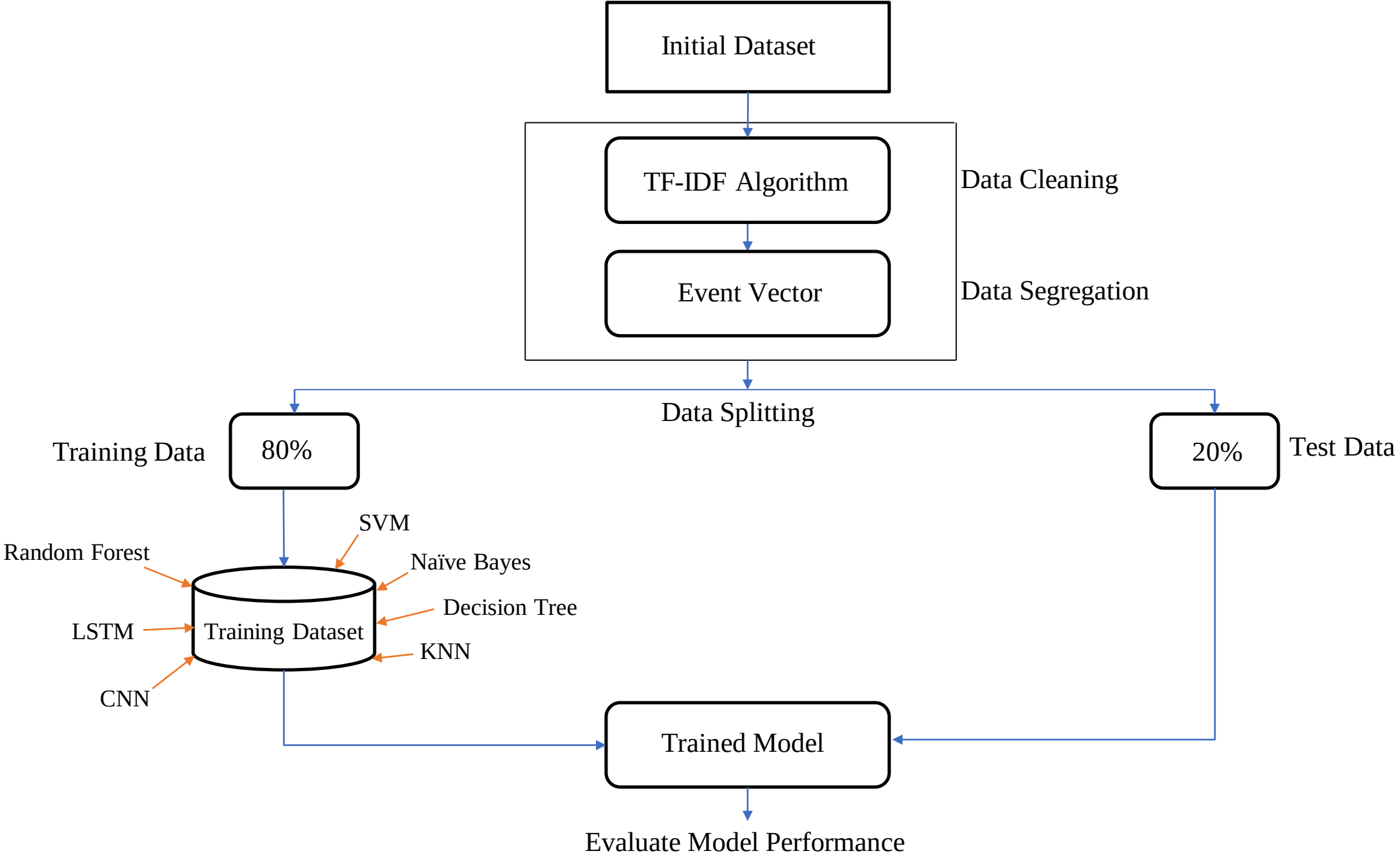


Figure 3: System Workflow of the proposed system

Sequence Diagram

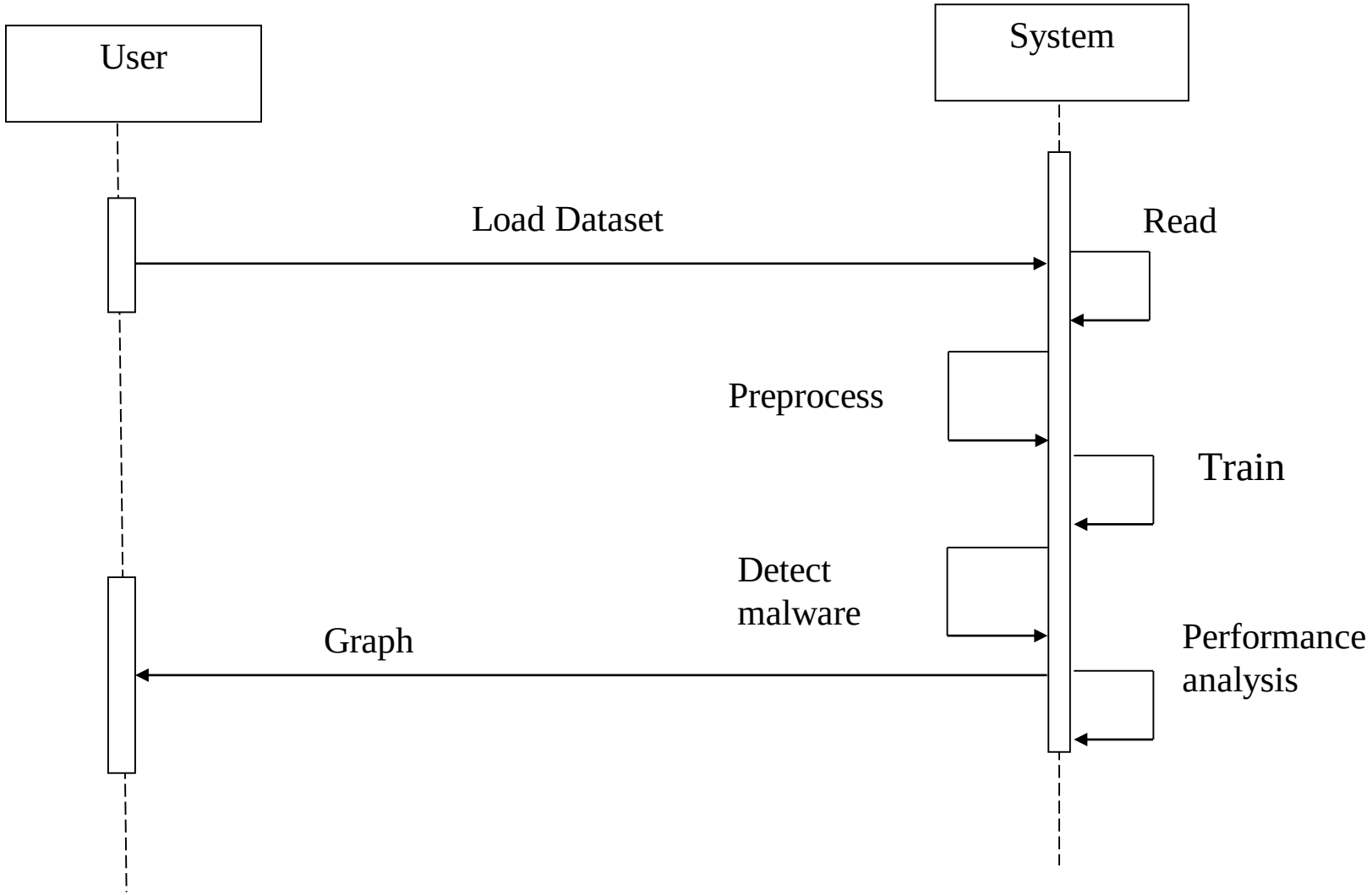


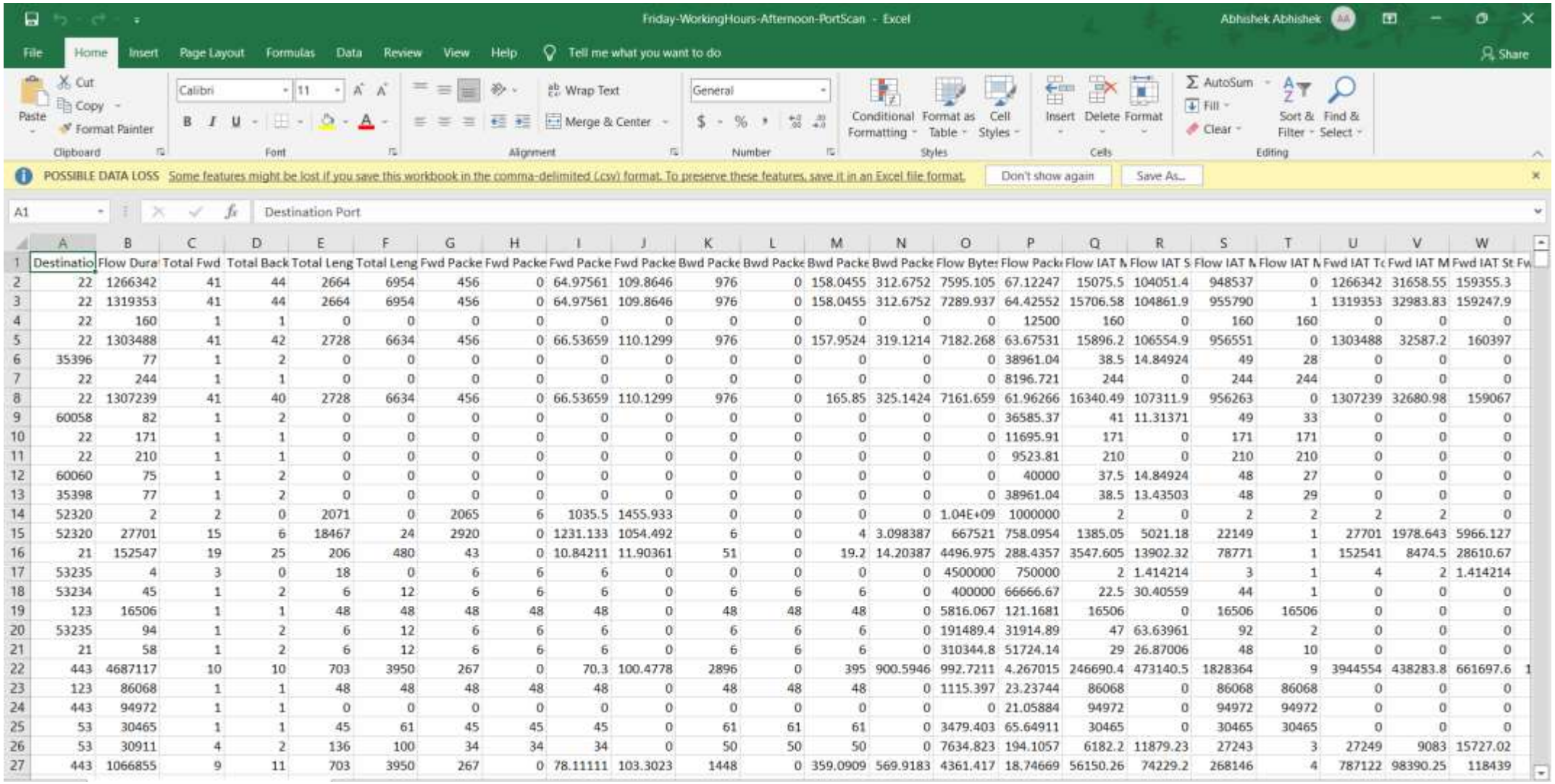
Figure 4: Sequence Diagram

Implementation

Dataset description

- The dataset is collected from the kaggle.com
- The dataset contains day wise packet transaction detail attributes like Destination Port, Flow Duration, Total Fwd Packets, Total Backward Packets, Total Length of Fwd Packets,Total Length of Bwd Packets, Fwd Packet Length Max, Fwd Packet Length Min, Fwd Packet Length Mean, Fwd Packet, Length Std, Bwd Packet Length Max, Bwd Packet Length Min, Bwd Packet Length Mean Bwd, total 76 attributes and 1 label class.

Cont..



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	Destinatio	Flow Dura	Total Fwd	Total Back	Total Leng	Total Leng	Fwd Packe	Fwd Packe	Fwd Packe	Fwd Packe	Bwd Packe	Bwd Packe	Bwd Packe	Bwd Packe	Flow Bytes	Flow Packe	Flow IAT M	Flow IAT S	Flow IAT M	Flow IAT M	Fwd IAT To	Fwd IAT M	Fwd IAT St	Fw
2	22	1266342	41	44	2664	6954	456	0	64.97561	109.8646	976	0	158.0455	312.6752	7595.105	67.12247	15075.5	104051.4	948537	0	1266342	31658.55	159355.3	
3	22	1319353	41	44	2664	6954	456	0	64.97561	109.8646	976	0	158.0455	312.6752	7289.937	64.42552	15706.58	104861.9	955790	1	1319353	32983.83	159247.9	
4	22	160	1	1	0	0	0	0	0	0	0	0	0	0	0	12500	160	0	160	160	0	0	0	
5	22	1303488	41	42	2728	6634	456	0	66.53659	110.1299	976	0	157.9524	319.1214	7182.268	63.67531	15896.2	106554.9	956551	0	1303488	32587.2	160397	
6	35396	77	1	2	0	0	0	0	0	0	0	0	0	0	0	38961.04	38.5	14.84924	49	28	0	0	0	
7	22	244	1	1	0	0	0	0	0	0	0	0	0	0	0	8196.721	244	0	244	244	0	0	0	
8	22	1307239	41	40	2728	6634	456	0	66.53659	110.1299	976	0	165.85	325.1424	7161.659	61.96266	16340.49	107311.9	956263	0	1307239	32680.98	159067	
9	60058	82	1	2	0	0	0	0	0	0	0	0	0	0	0	36585.37	41	11.31371	49	33	0	0	0	
10	22	171	1	1	0	0	0	0	0	0	0	0	0	0	0	11695.91	171	0	171	171	0	0	0	
11	22	210	1	1	0	0	0	0	0	0	0	0	0	0	0	9523.81	210	0	210	210	0	0	0	
12	60060	75	1	2	0	0	0	0	0	0	0	0	0	0	0	40000	37.5	14.84924	48	27	0	0	0	
13	35398	77	1	2	0	0	0	0	0	0	0	0	0	0	0	38961.04	38.5	13.43503	48	29	0	0	0	
14	52320	2	2	0	2071	0	2065	6	1035.5	1455.933	0	0	0	0	1.04E+09	1000000	2	0	2	2	2	2	0	
15	52320	27701	15	6	18467	24	2920	0	1231.133	1054.492	6	0	4	3.098387	667521	758.0954	1385.05	5021.18	22149	1	27701	1978.643	5966.127	
16	21	152547	19	25	206	480	43	0	10.84211	11.90361	51	0	19.2	14.20387	4496.975	288.4357	3547.605	13902.32	78771	1	152541	8474.5	28610.67	
17	53235	4	3	0	18	0	6	6	6	0	0	0	0	0	4500000	750000	2	1.414214	3	1	4	2	1.414214	
18	53234	45	1	2	6	12	6	6	6	0	6	6	6	0	400000	66666.67	22.5	30.40559	44	1	0	0	0	
19	123	16506	1	1	48	48	48	48	48	0	48	48	48	0	5816.067	121.1681	16506	0	16506	16506	0	0	0	
20	53235	94	1	2	6	12	6	6	6	0	6	6	6	0	191489.4	31914.89	47	63.63961	92	2	0	0	0	
21	21	58	1	2	6	12	6	6	6	0	6	6	6	0	310344.8	51724.14	29	26.87006	48	10	0	0	0	
22	443	4687117	10	10	703	3950	267	0	70.3	100.4778	2896	0	395	900.5946	992.7211	4.267015	246690.4	473140.5	1828364	9	3944554	438283.8	661697.6	1
23	123	86068	1	1	48	48	48	48	48	0	48	48	48	0	1115.397	23.23744	86068	0	86068	86068	0	0	0	
24	443	94972	1	1	0	0	0	0	0	0	0	0	0	0	0	21.05884	94972	0	94972	94972	0	0	0	
25	53	30465	1	1	45	61	45	45	45	0	61	61	61	0	3479.403	65.64911	30465	0	30465	30465	0	0	0	
26	53	30911	4	2	136	100	34	34	34	0	50	50	50	0	7634.823	194.1057	6182.2	11879.23	27243	3	27249	9083	15727.02	
27	443	1066855	9	11	703	3950	267	0	78.11111	103.3023	1448	0	359.0909	569.9183	4361.417	18.74669	56150.26	74229.2	268146	4	787122	98390.25	118439	

Figure 5: Snapshot of Dataset

Cont..

- **Destination Port**: Destination of the packets in the network.
- **Flow Duration**: Time taken by the packet to reach from source to destination
- **Total Fwd Packets**: Total number of packets that flowed from source to destination
- **Total Backward Packets**: Total number of packets that flowed from destination to source
- **Total Length of Fwd Packets**: Total length of the Forwarded packet.
- **Total Length of Bwd Packets**: Total length of the backward packets
- **Fwd Packet Length Max**: Max length of the forward packet
- **Fwd Packet Length Min**: Min length of the forward packets
- **Fwd Packet Length Mean**: Mean of the forward packets which specifies the mean value
- **Fwd Packet, Length Std, Bwd Packet Length Max, Bwd Packet Length Min, Bwd Packet Length Mean Bwd, total 76 attributes and 1 label class.**

Preprocessing

- Getting the dataset.
- Importing libraries.
- Importing datasets.
- Finding Missing Data.
- Removing the duplicate data.
- Fixing Structural errors.

Algorithms used

- **Random forests** are ensemble learning methods that are used for either classification or regression purposes. Random forests are composed of several decision trees that are combined together to make a unanimous decision or classification. Random forest are better than just regular decision trees because they do not cause overfitting of the data.

Algorithm 1

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

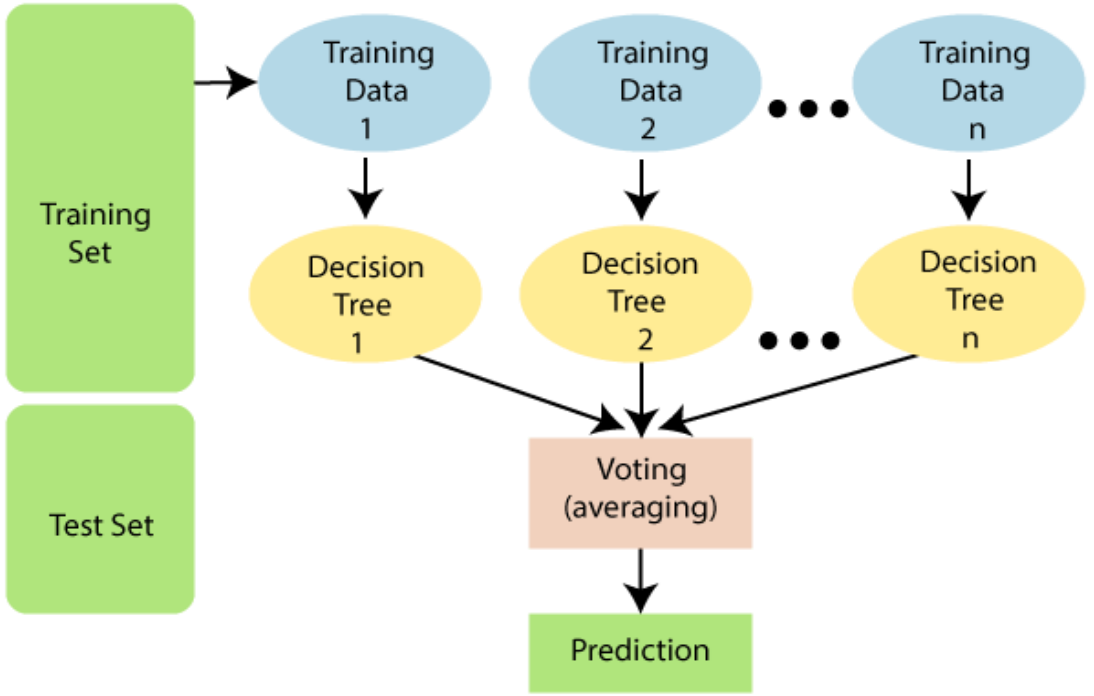


Figure 6: Random forest tree

Cont..

- **KNN(k-nearest neighbor)** is a lazy learning, non-parametric algorithm. It uses data with several classes to predict the classification of the new sample point. KNN is non-parametric since it doesn't make any assumptions on the data being studied, i.e., the model is distributed from the data.

Algorithm 2

Step 1 − For implementing any algorithm, we need dataset. So during the first step of KNN is to load the training as well as test data.

Step 2 − Next, we need to choose the value of K i.e. the nearest data points. K can be any integer.

Step 3 − Calculate the distance between test data and each row of training data with the help of any of the method namely: Euclidean, Manhattan or Hamming distance. The most commonly used method to calculate distance is Euclidean.

Step 4 − End

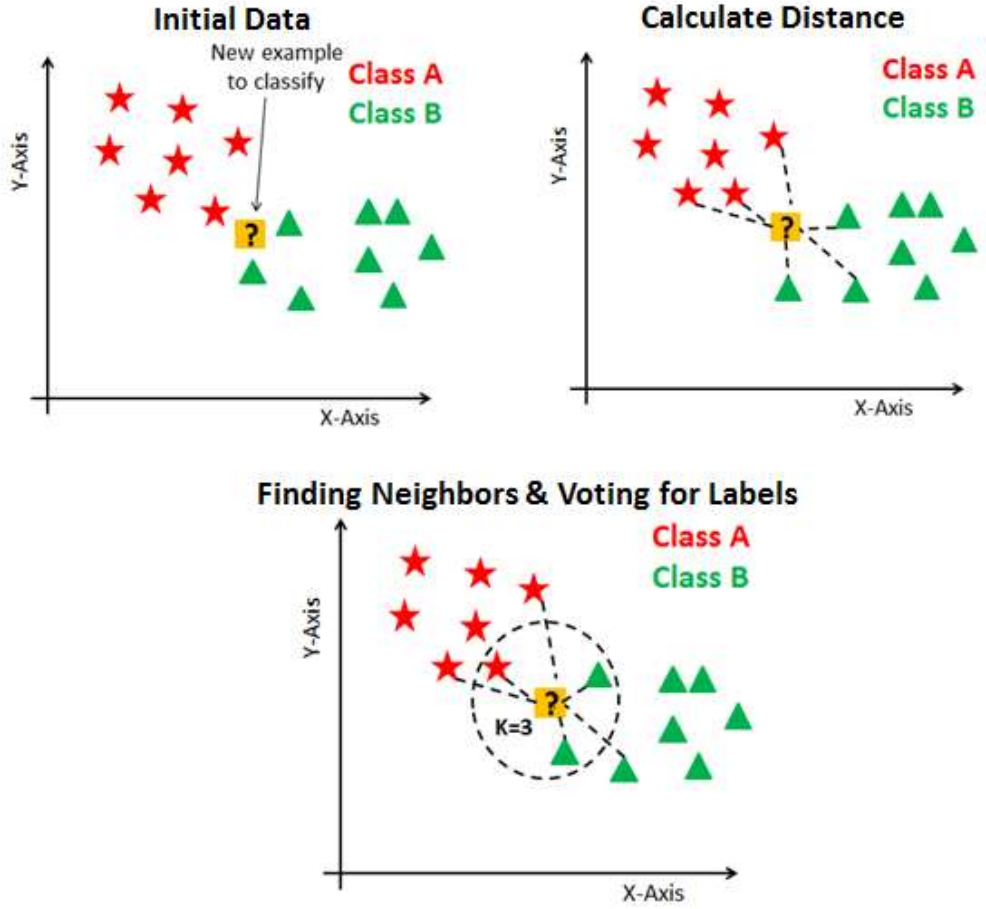


Figure 7: KNN Algorithm

Cont..

- **Naive Bayes** classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

Algorithm 3

1.Convert the given dataset into frequency tables.

2.Generate Likelihood table by finding the probabilities of given features.

3.Now, use Bayes theorem to calculate the posterior probability.

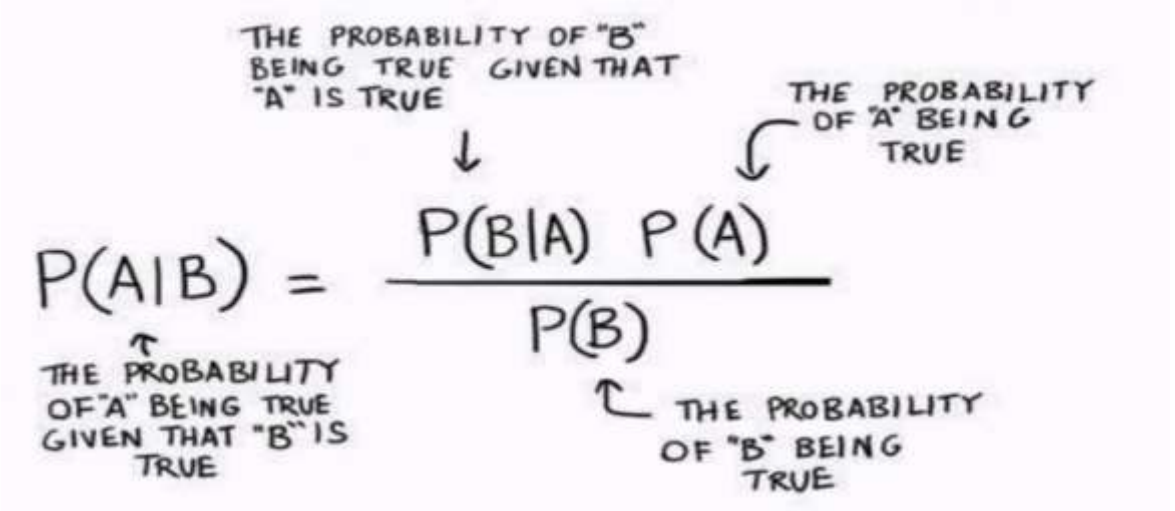


Figure 8: Naïve Bayes Algorithm

Cont..

- **Decision tree** are simple to implement and equally easy to interpret.
- Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features

Algorithm 3

Step-1: Begin the tree with the root node, says S, which contains the complete dataset.

Step-2: Find the best attribute in the dataset using **Attribute Selection Measure (ASM).**

Step-3: Divide the S into subsets that contains possible values for the best attributes.

Step-4: Generate the decision tree node, which contains the best attribute.

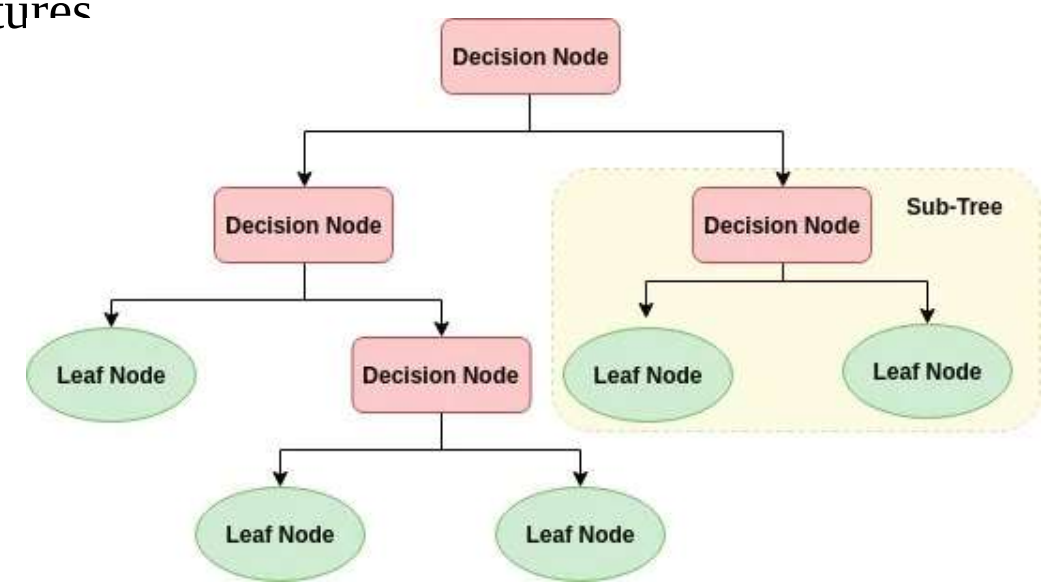


Figure 9: Decision Tree Algorithm

Cont..

- **SVM(Support Vector Machine)** is a supervised machine learning algorithm used for both classification and regression. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features. The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points.

- Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes.

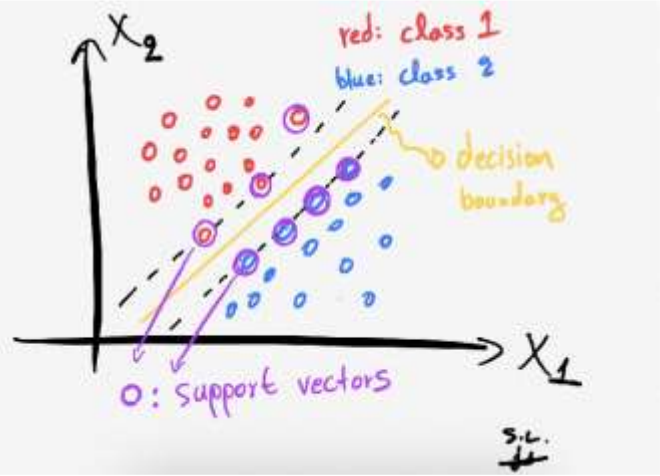


Figure 10: SVM Algorithm

Cont..

Neural Networks also known as Artificial Neural Networks (ANNs) or Simulated Neural Networks (SNNs), are a subset of machine learning and are at the heart of deep learning algorithms. Their name and structure are inspired by the human brain, mimicking the way that biological neurons signal to one another.

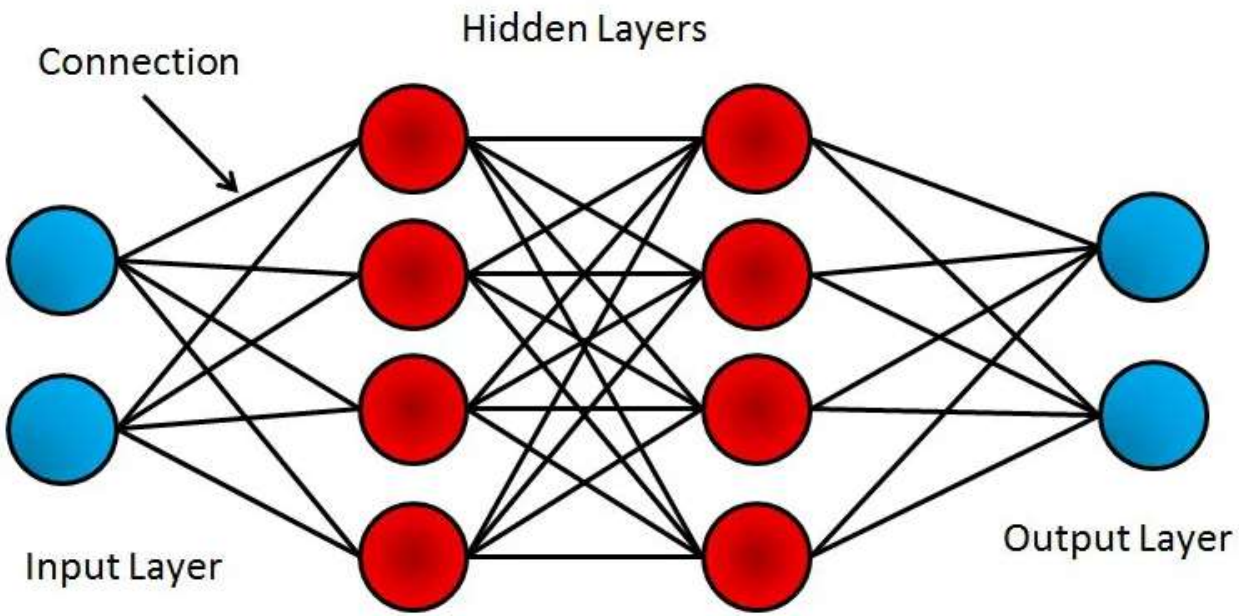


Figure 11: Neural Network

Results and Analysis

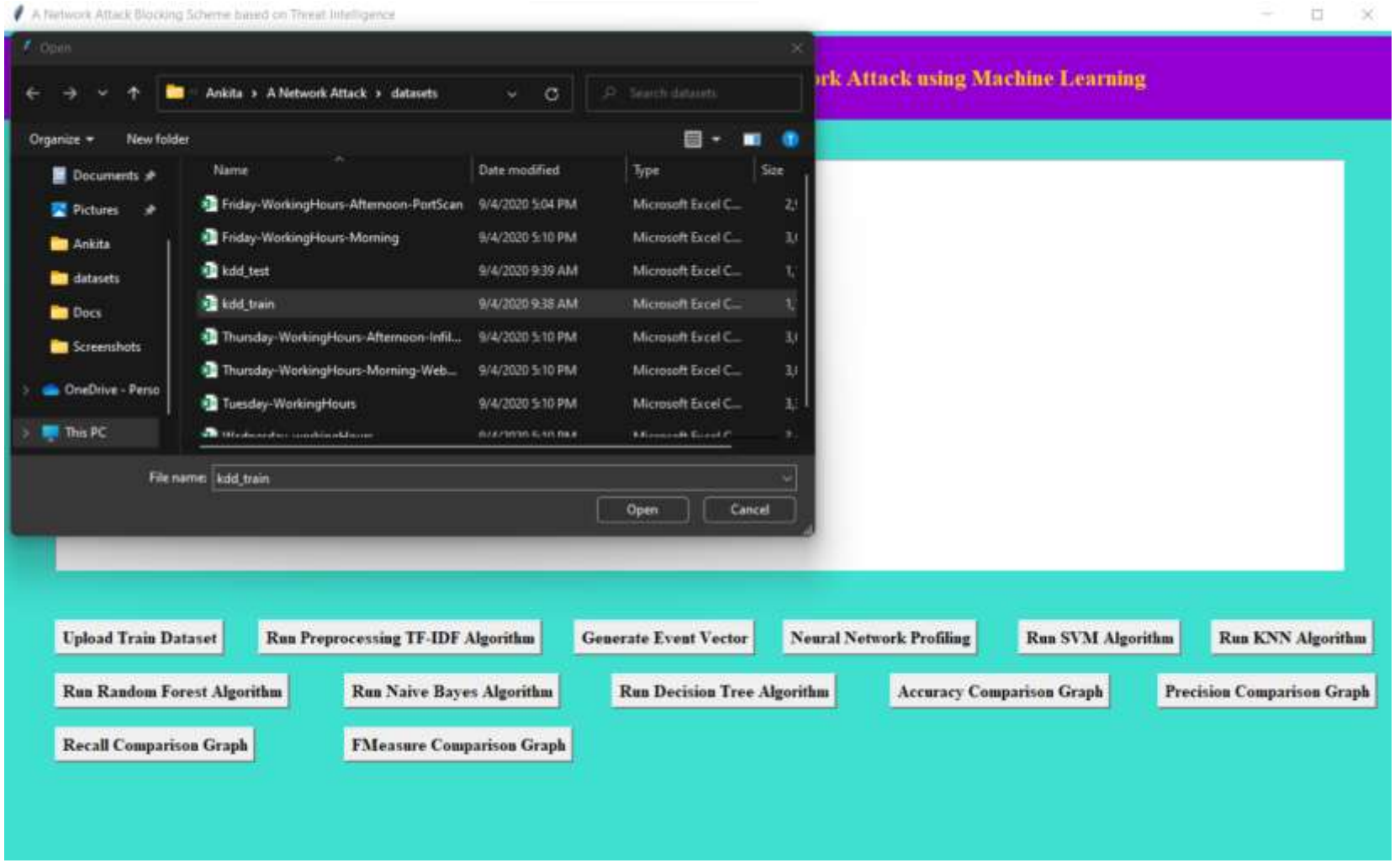


Figure 12: Uploading the dataset

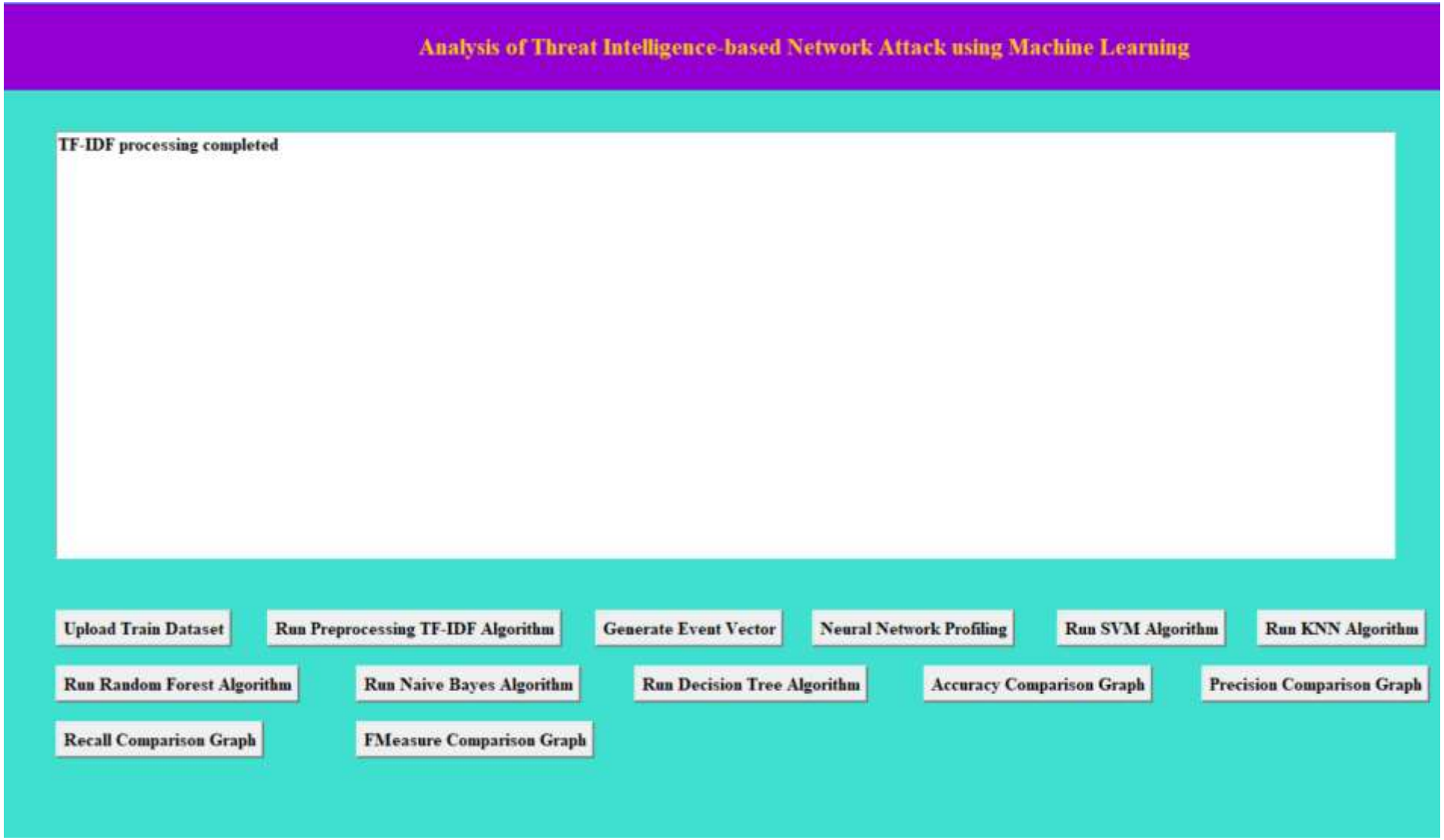


Figure 13: Running the TF-IDF Algorithm

Figure 14: Generating the Event Vector

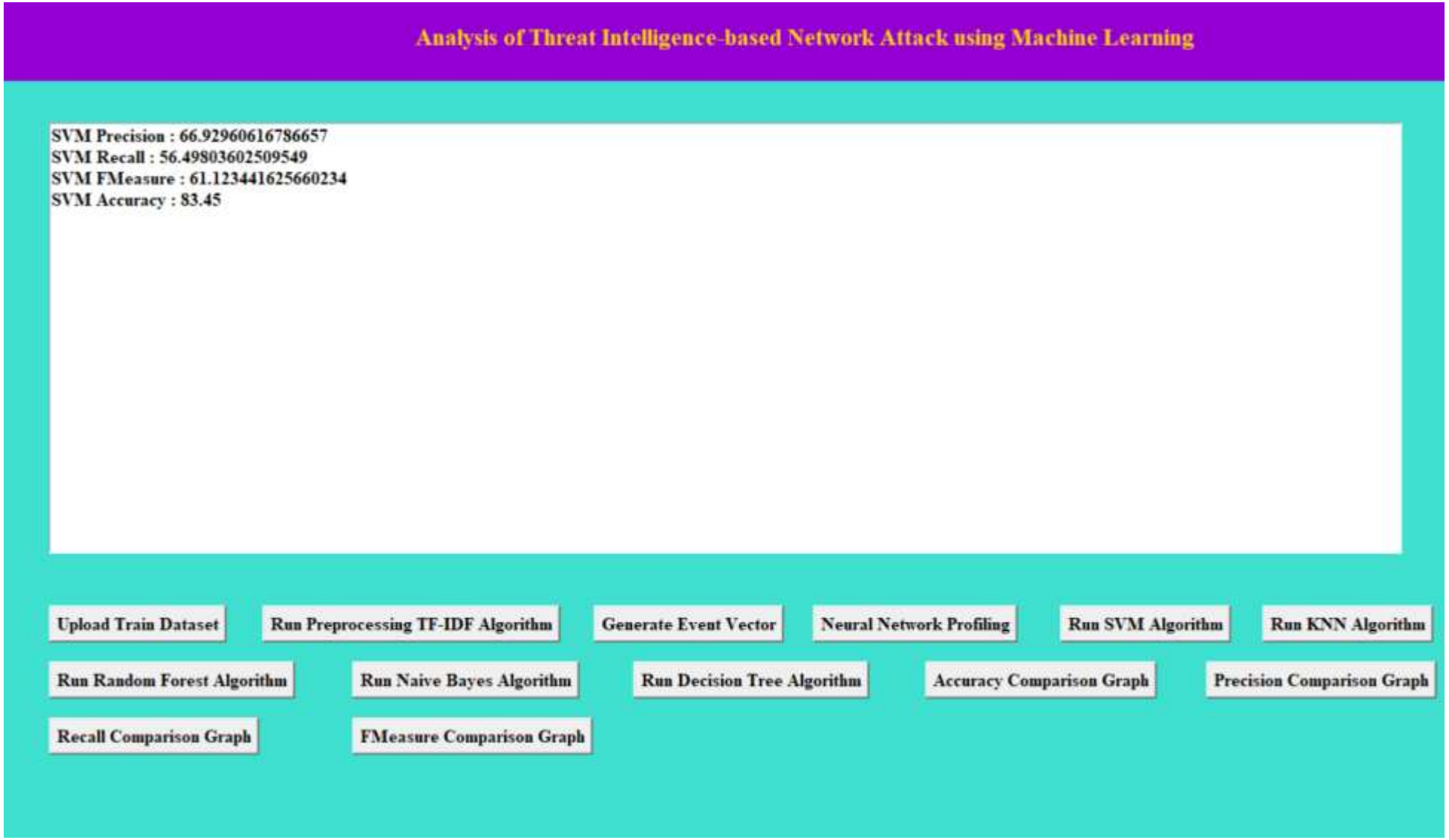


Figure 15: Support Vector Machine Algorithm Prediction Results

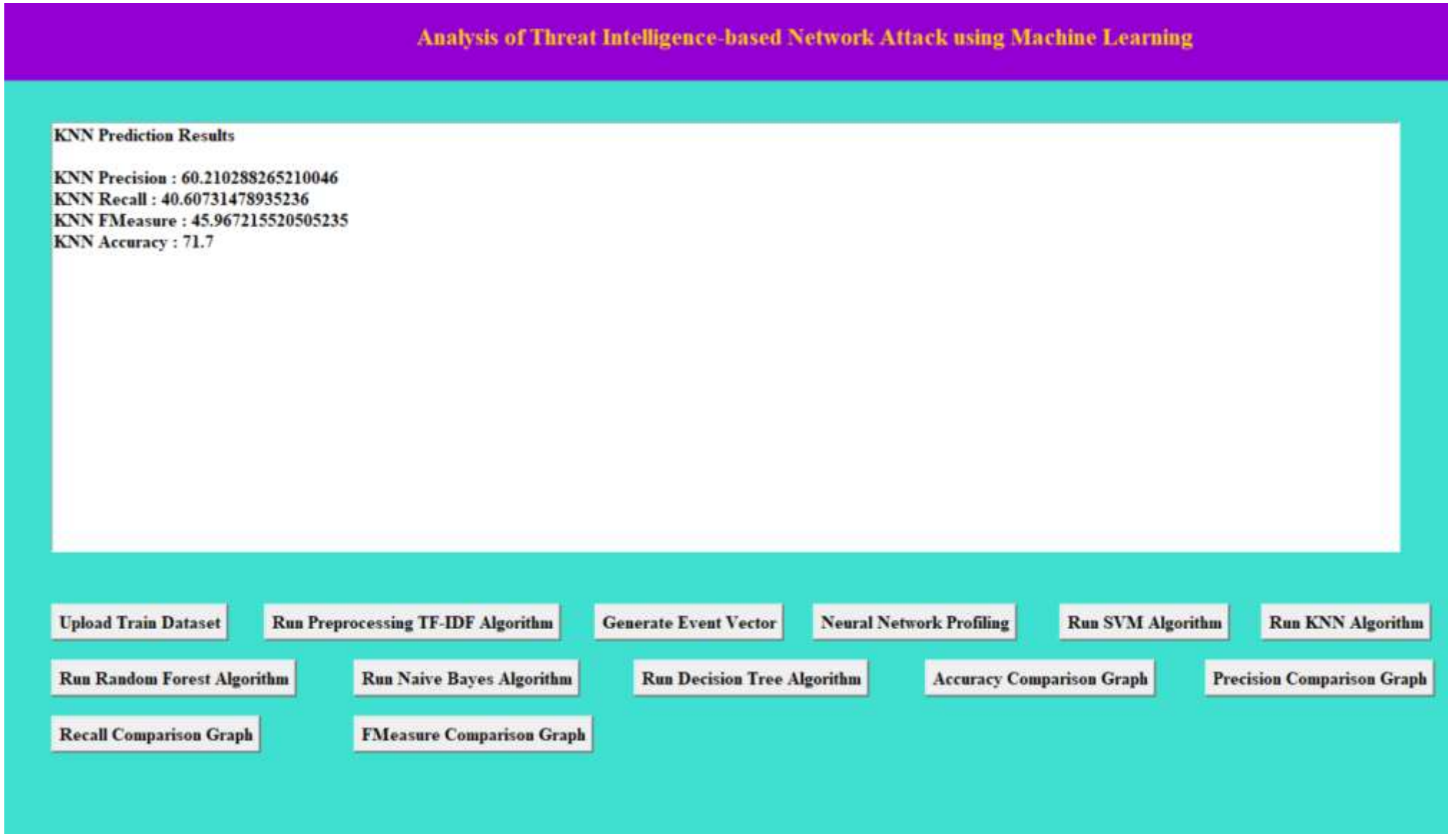


Figure 16: KNN Algorithm Prediction Results

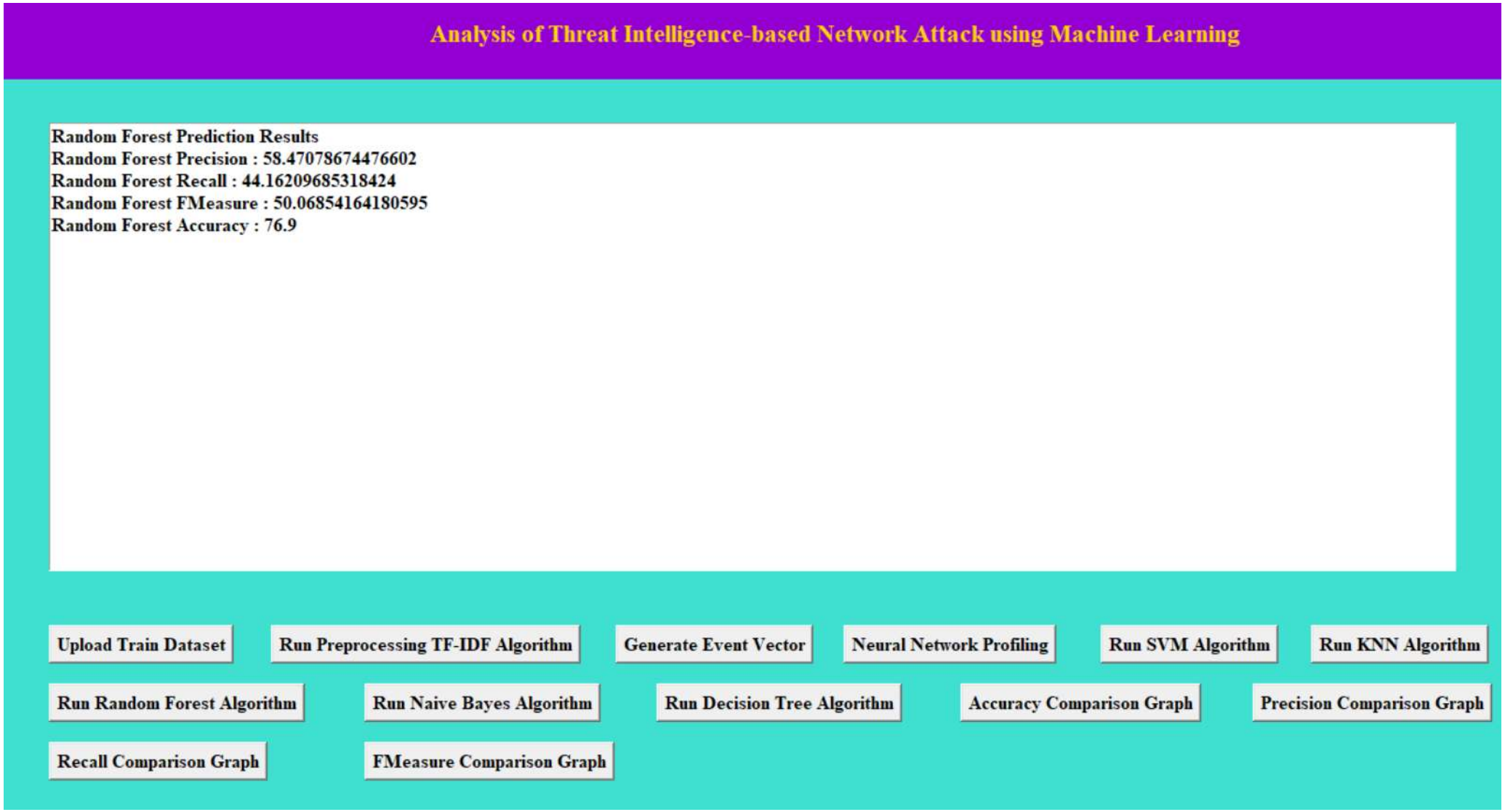


Figure 17: Random Forest Algorithm Prediction Results

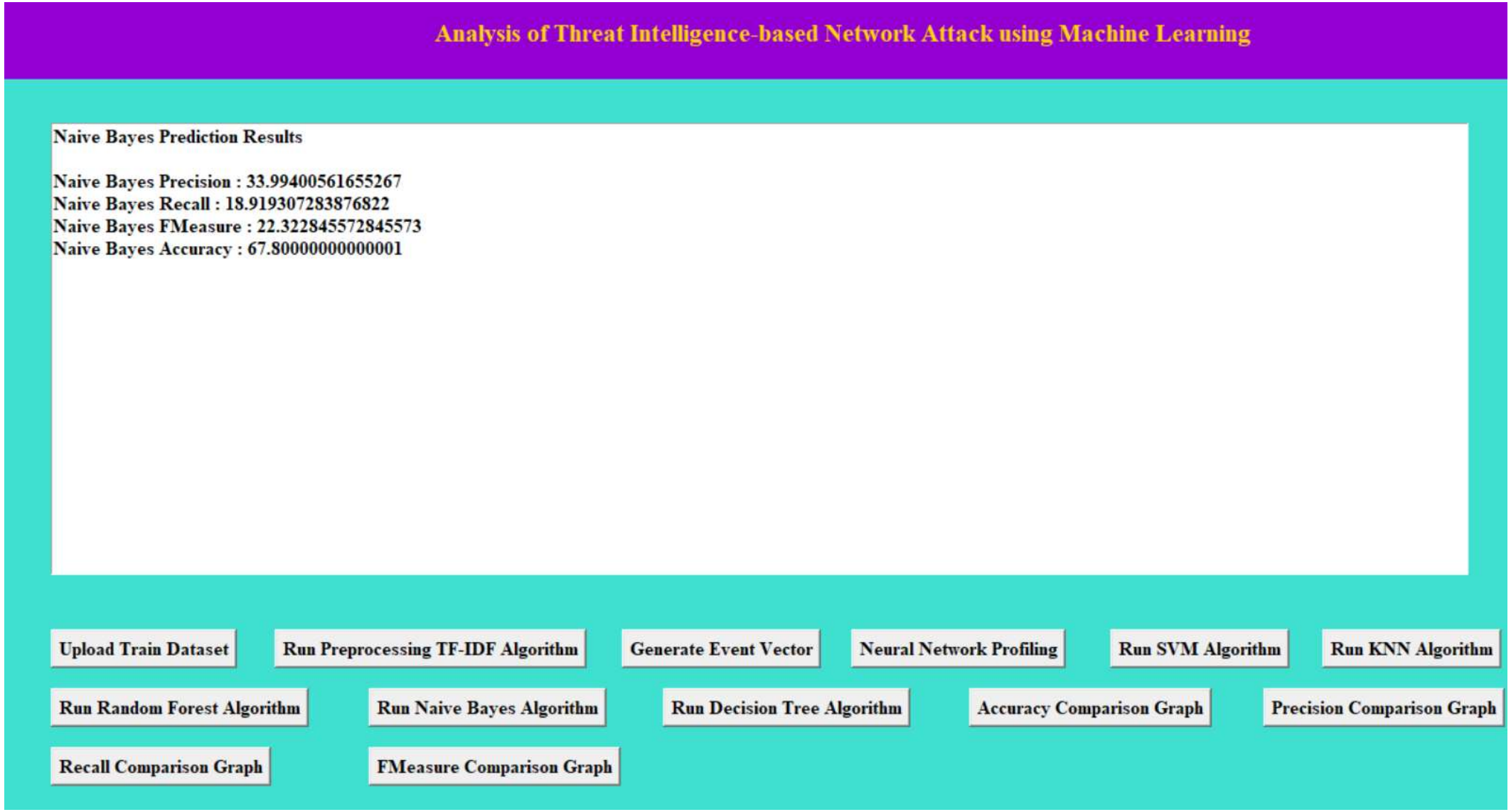


Figure 18: Naïve Bayes Algorithm Prediction Results

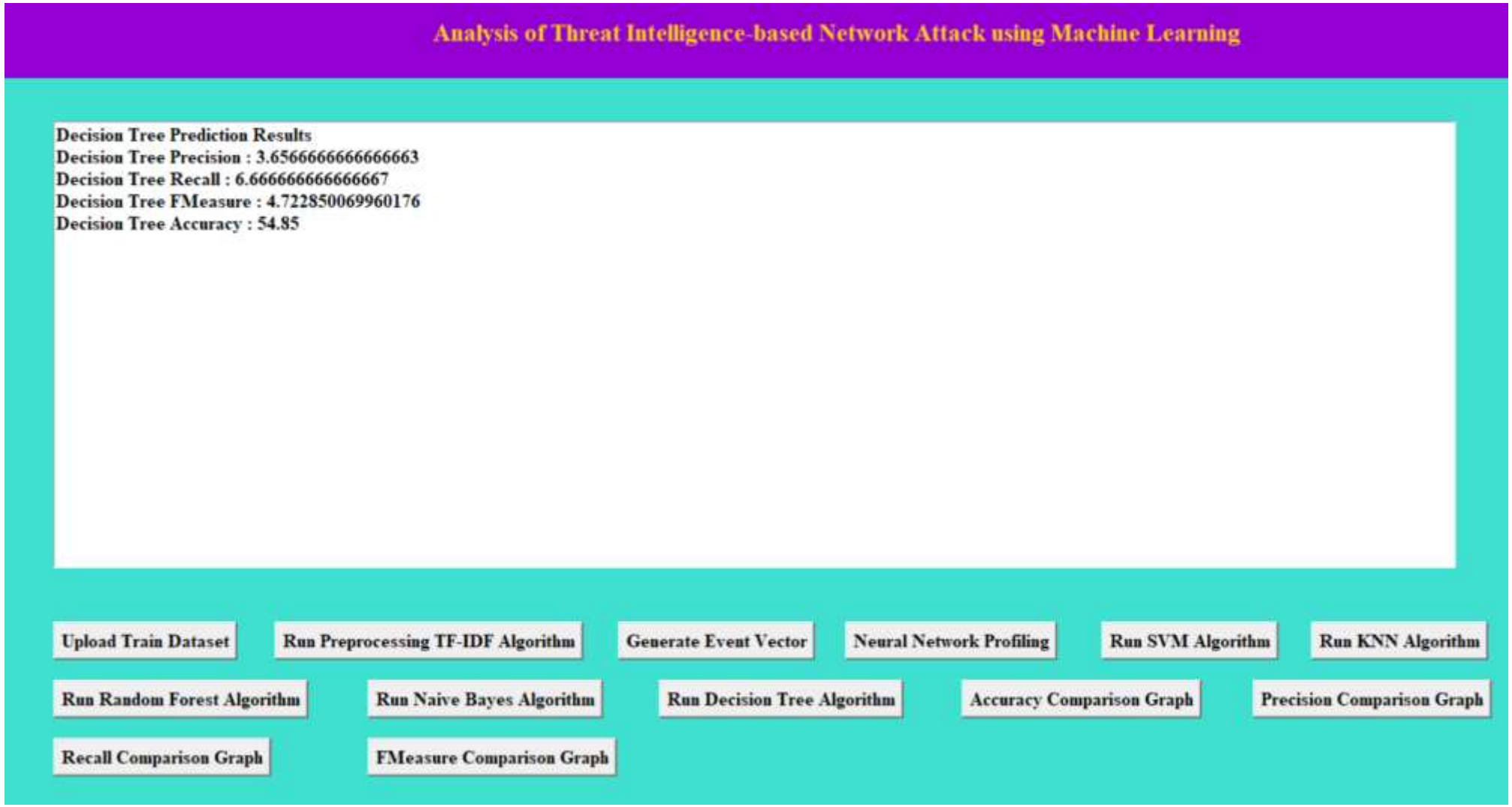


Figure 19: Decision Tree Algorithm Prediction Results

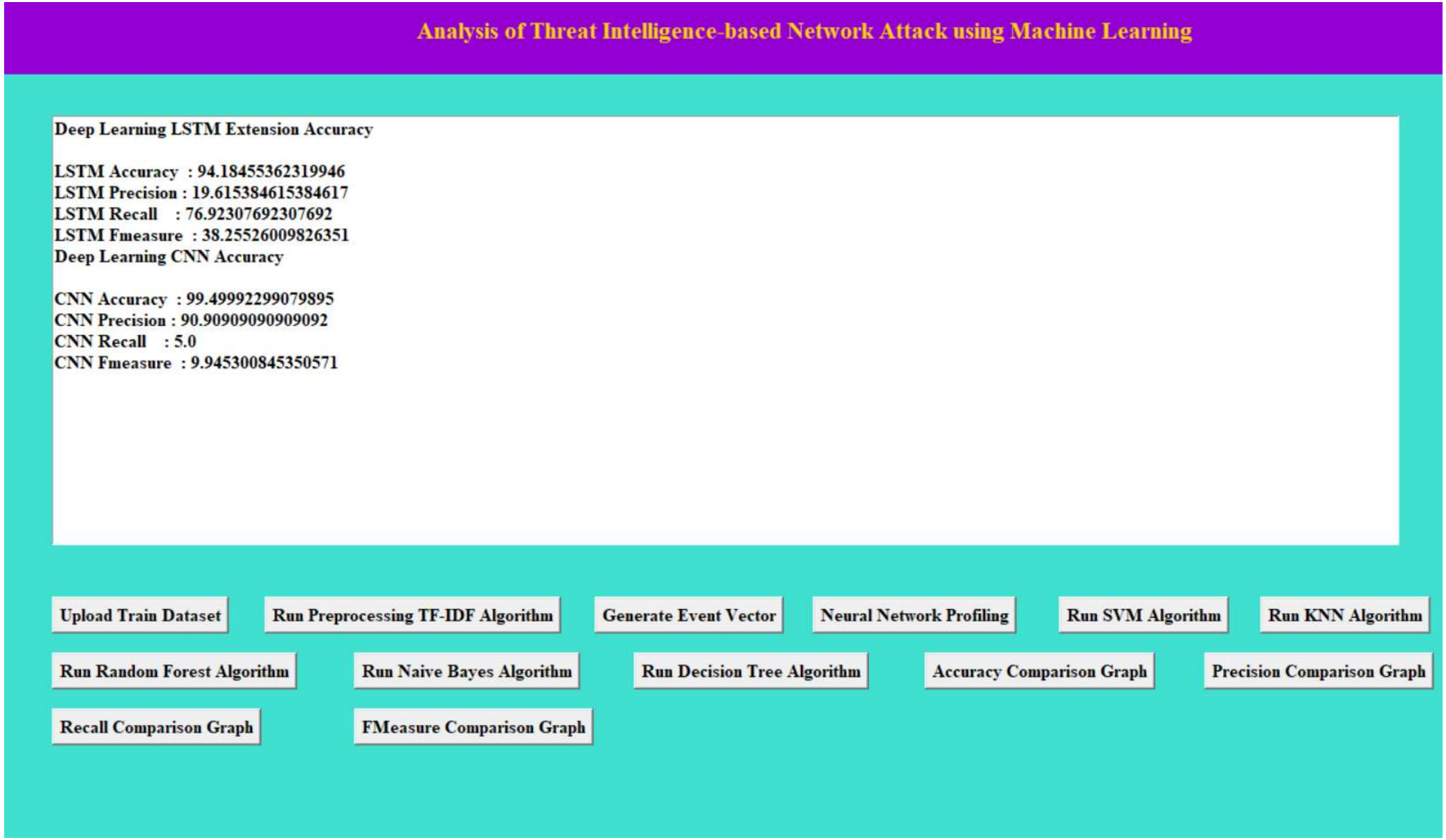


Figure 20: Neural Network Profiling Prediction Results

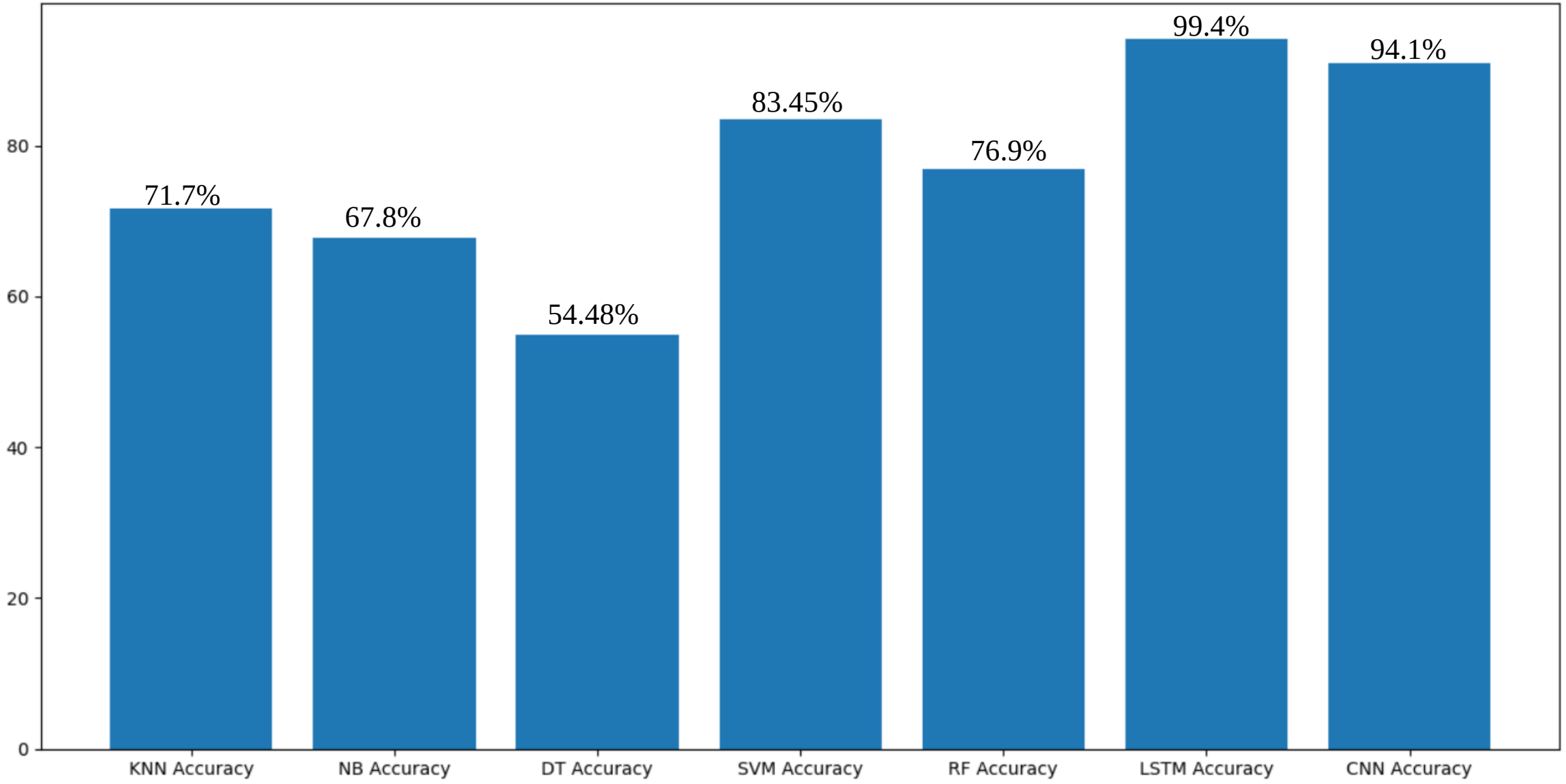


Figure 21: Comparison Graph of all Algorithm

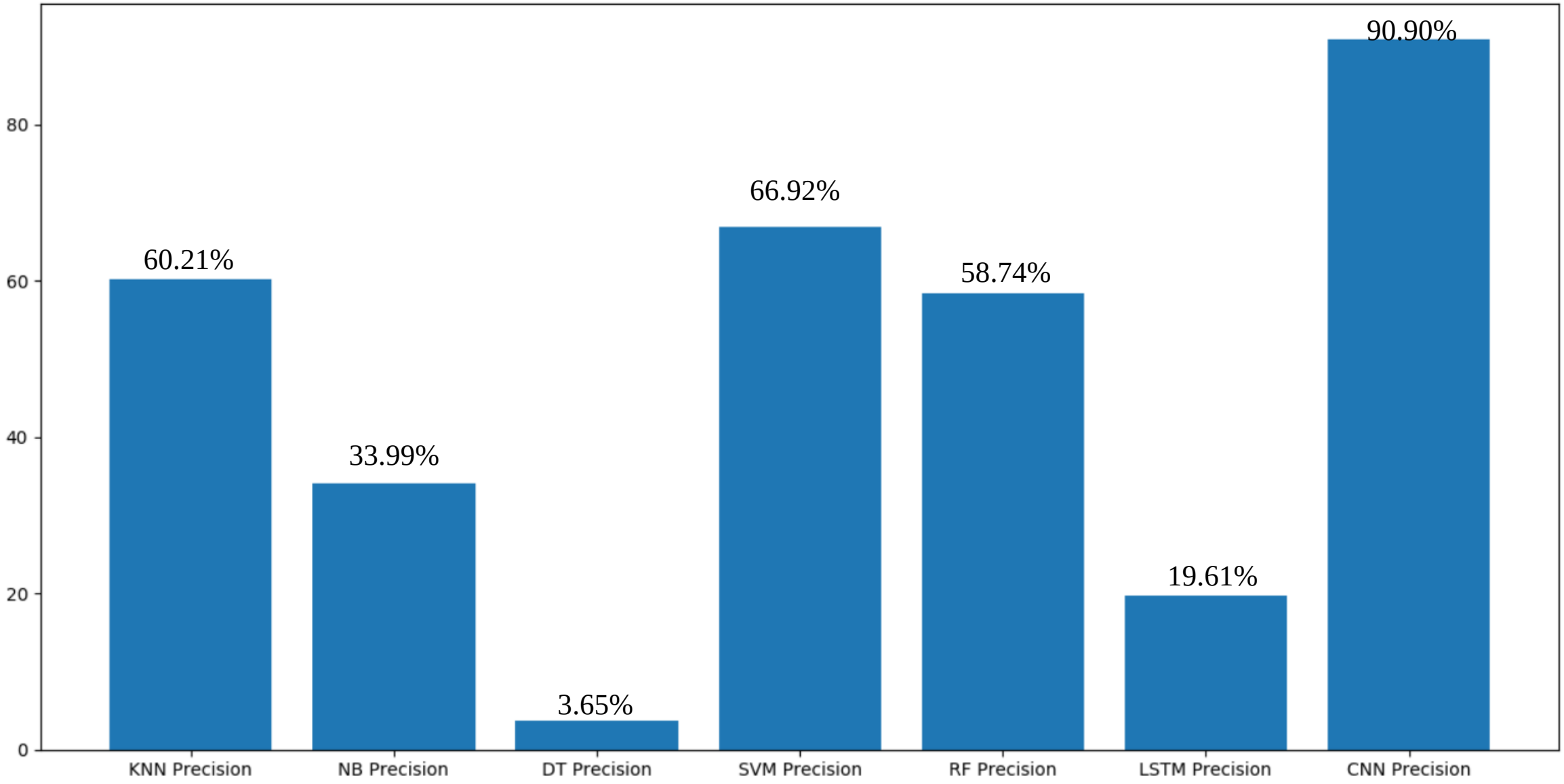


Figure 22: Precision Graph of all Algorithm

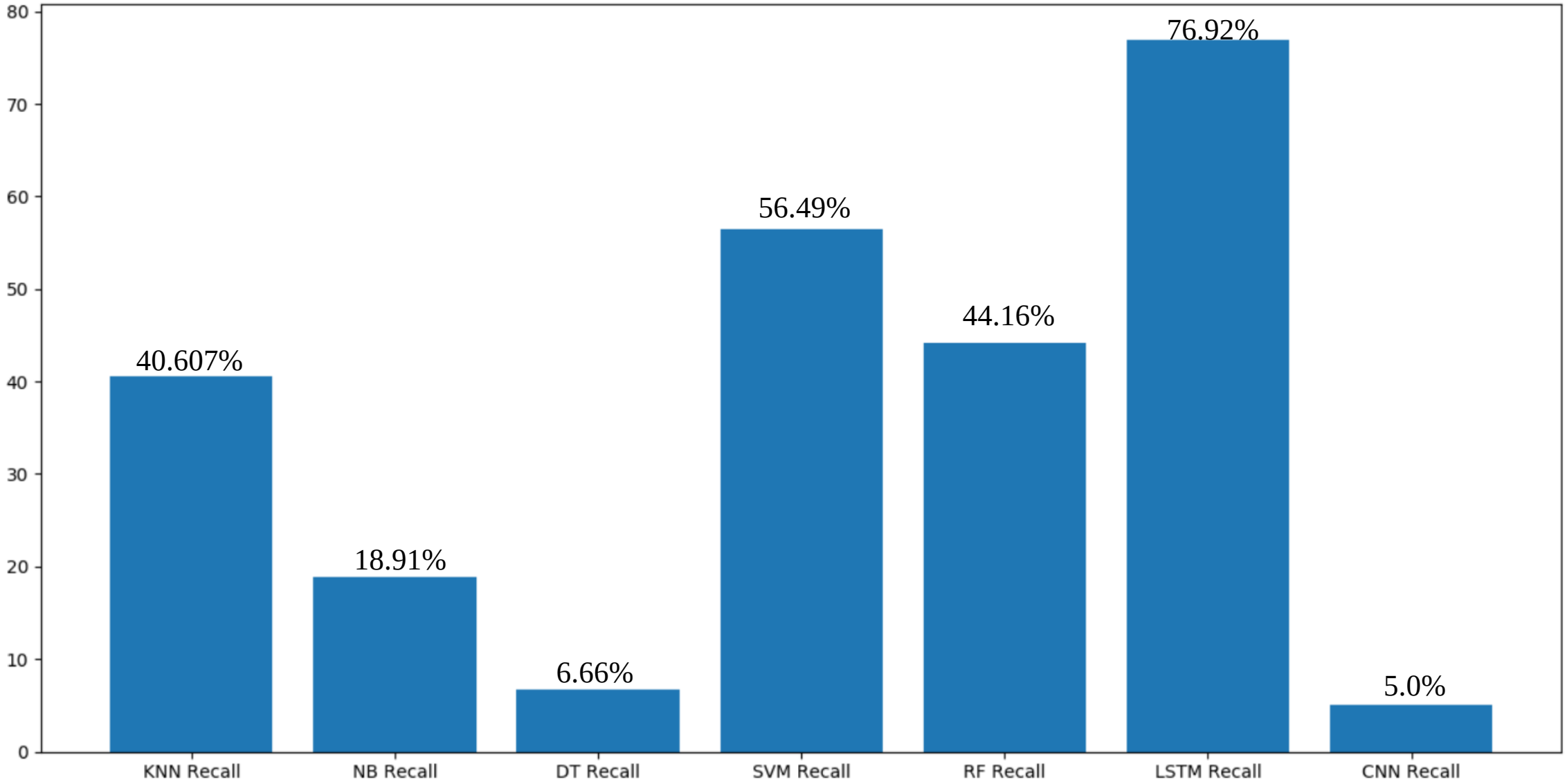


Figure 23: Recall Graph of all Algorithm

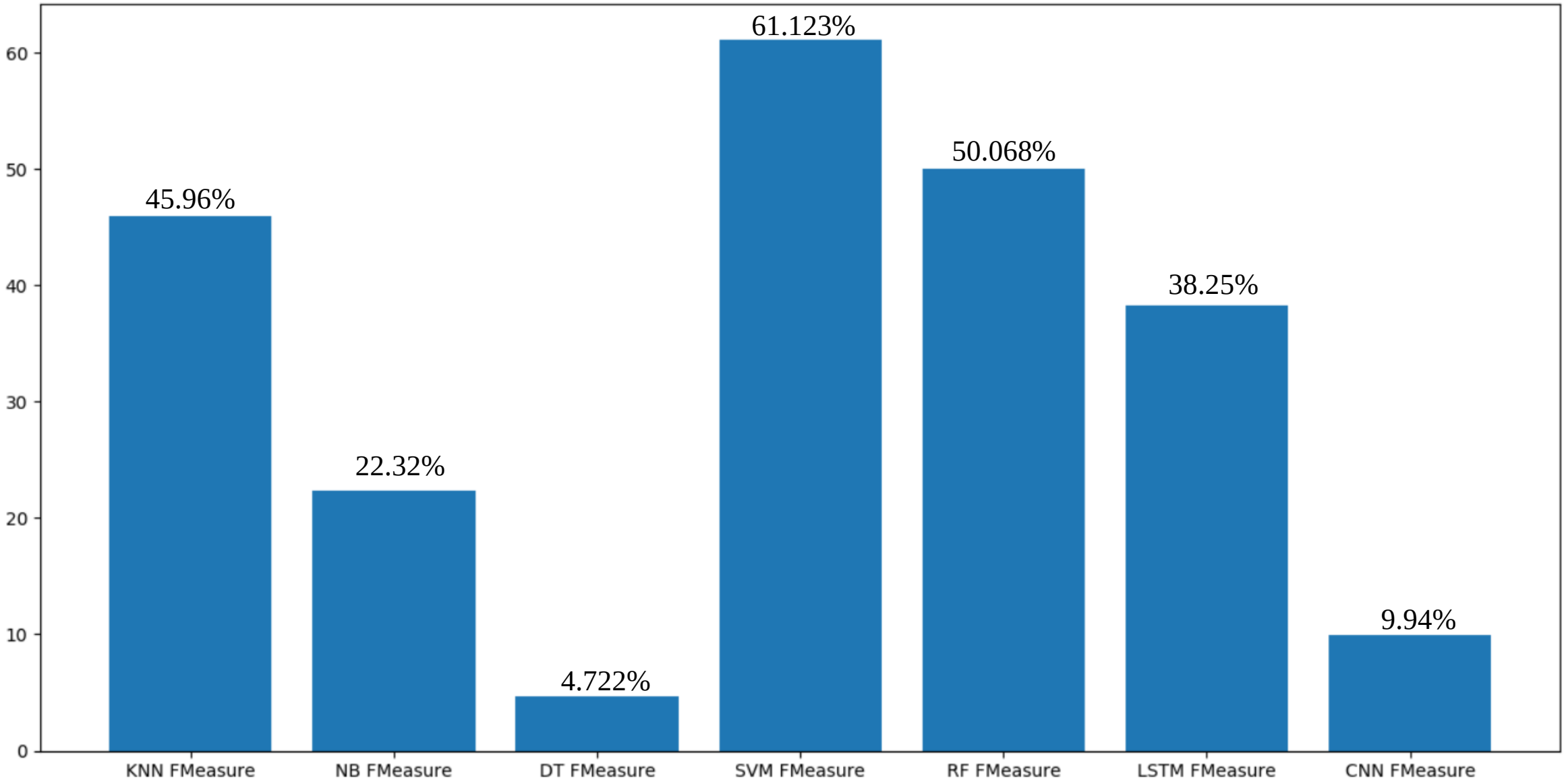


Figure 24: F1 measure Graph of all Algorithm

Conclusion

- The results are examined utilizing network attack data sets, demonstrating that it is one of the best competitor data sets for threat intelligence analysis and testing. The TI model improves both the accuracy and the detection time.
- After experimenting with various numbers of traits, it was discovered that 22 of the 76 attributes have a categorization accuracy of roughly 99 percent. And, of the 22 features, 11 are more important to total classification accuracy, while the remaining 9 are less important.
- The classification accuracy is 99.4 percent when these 13 out of 22 features are used.

Future Work

- Can reduce the number false positive rates so that the performance increases by this.
- To develop a system model with a higher precision rate while using fewer characteristics to improve overall accuracy.
- Datasets like NSL-KDD can be considered in future research.

References

[1] Mavroeidis, Vasileios, and Siri Bromander. "Cyber threat intelligence model: an evaluation of taxonomies, sharing standards, and ontologies within cyber threat intelligence." *2017 European Intelligence and Security Informatics Conference (EISIC)*. IEEE, 2017.

[2] Blanco-Medina, Pablo, et al. "Detecting vulnerabilities in critical infrastructures by classifying exposed industrial control systems using deep learning." *Applied Sciences* 11.1 (2021): 367.

[3] Chen, Zhijiang, et al. "A cloud computing based network monitoring and threat detection system for critical infrastructures." *Big Data Research* 3 (2016): 10-23.

[4] Al-Mohannadi, Hamad, Irfan Awan, and Jassim Al Hamar. "Analysis of adversary activities using cloud-based web services to enhance cyber threat intelligence." *Service Oriented Computing and Applications* 14.3 (2020): 175-187.

THANK YOU