ISSN: 2581-3404 (Online) IF: 5.445 (SJIF)
IJIRTM, Volume-7, Issue-3, June-2023.

Detection of Cyber Attacks Using Machine Learning Techniques

Albert.N.Rejy[1], Dr. Sadhana K. Mishra[2]
M.Tech. Research Scholar[1], Prof. & Head[2]
Dept. of CSE, LNCT, Bhopal[1,2]

Abstract: *Professionals in cyber security give risk assessment more consideration and provide ways to mitigate it. Designing effective methods was a goal established for the field of cyber defense. Despite its success in cyber defense, machine learning has also grown to be a significant worry for data privacy. Unprecedented advancements in computing, storage, and computational technologies have led to the fast growth of cloud computing, networking, and evolutionary computation. There is an increasing need for comprehensive and sophisticated information security and privacy problems as the world rapidly digitizes. Additionally, there are increasingly complex strategies for defending against security threats. Worldwide cyber terrorism is growing thanks to various computer flaws. Global computer security issues including malware detection, ransom ware recognition, fraud detection, and spoofing identification were addressed using machine learning techniques. The study examines the use of cyber training to both offensive and defense, offering information on cyber risks based on machine learning techniques and Machine learning methods that describe how machine learning is used for computer defense, such as the discovery and avoidance of attacks, vulnerability scanning and recognition, and public internet risk assessment, are used to analyze the far more prevalent types of cyber security concerns.*

Keywords: Cyber security, Malware detection, Machine learning, cyber threat intelligence. Cyber- attacks.

Introduction

Global safety is also a top issue as the world progressively digitizes. Through frequent publishing of scientific journals and increased openness in the Modern Western world, internet-based communication advancements have made access to public innovations and scientific discoveries quite straightforward [1–3]. Unfortunately, using cutting-edge scientific and technical breakthroughs is equally available to intelligence researchers and cybercriminals with distinct agendas in need of these tools and information. Models and applications to enhance safety methods that can identify possible dangers and effectively deal with them have been developed as a result of analysis and advancement in the field of machine learning [4,16]. Internet barriers may be defined as a technology-based discipline, practices, and initiatives designed to guard against dangers, disruptions, and unauthorized access to technology, computers, services, and records [4,17]. In 2016, there were several developments in the field of machine learning that contributed to safety, including those in the areas of personal assistants, verbal connection, healthcare, and healthcare. From several audit repositories that are utilized to locate intruders, they would be used to find helpful information [3,18].

Cyber assaults have a significant advantage in the cyberwar since they only frequently succeed after several efforts. However, on the opposing side, a 100% survival rate is required for the best defence. A study reveals that many businesses, organisations, people, and applications were targets of cyberattacks in 2017. Private information, accounting records, and secret information were all included in the hacked data [5,15]. The usage of some data can be disastrous, especially if it is made widely available to the public or traded on the black market. Over $3.9 billion in assets have been stolen, and costs for reducing the harm caused by theft have been reported in recent decades. These are just a few statistics on the impact of cyber protection on businesses, organisations, and individuals. Through 2022, there is predicted to be a significant increase in the number of cyber-security employment, with institutions throughout the globe investing at least $20 million annually in data security defence. ¾ According to reports, intruders generate over $1 trillion in income for Ransom each year. Machine learning is a kind of artificial intelligence that allows computers to develop and practise utilising data without ever putting an agreement into practise. The forecasting of training data is dependent on computer simulations gathered through primary collection analysis [3,16]. Machine learning technologies should be used to make decisions based on consumer behaviour, interests, and medical services, and artificial intelligence should be used to forecast pandemics or the likelihood that a person will recover from certain illnesses such as cancer on the basis of their medical history [5]. Artificial intelligence technologies cover a wide range of industries, including e-commerce.

Machine learning algorithms play a significant part in boosting security measures across this intrusion detection and prevention system. There are two categories of ML algorithms: controlled and unattended. They can be separated from the information they are gathering [4]. A community with expertise in labelled teaching may offer simulations as part of a controlled learning strategy in order to clarify the differences between the labels. Uncontrolled learning is a strategy for using knowledge formulae that are difficult to pin down and are meant to inflate courses on their own. The labelled information is frequently quite limited [5].

Predictive or pattern recognition techniques are two categories of machine learning techniques. In supervised learning, there is frequently an objective parameter, and the system will forecast using different learning methods [14]. For instance, a machine learning model can forecast the frequency and demand of online applications, as well as if a given IP address was used as part of a DDOS attack's target ip physical layer. Numerous machine learning techniques, such as linear and integrated review, random forest, and productivity index are included in variously regulated programming [1].

Machine learning (ML) and artificial intelligence (AI) have quickly risen to the top of the list of technologies used in cyber security. AI and ML are being utilised to improve the security of companies and people because to the growing volume of data and sophisticated cyber threats. Large volumes of data are analysed using AI and ML to find patterns that could point to the existence of a cyber threat. This makes it possible for businesses to identify risks from the internet more rapidly and precisely than they could before. In this post, we'll look at some of the key present uses of AI and ML in cyber security as well as their potential moving forward. We'll also see how employing AI and machine learning in cyber security has drawbacks.

II. Literature Review

Li et al. [7] described an application that uses the RBF kernel high energy gradient boosting svm classifier in conjunction with the most popular KDD'99 copy data collection to detect pre-default groups like DoS or Search, U2R, and R2L as well as standard transport. In order to verify the method and produce a speedier model, Amiri et al. employed a less-rounded variable categorization approach. Hu and co. [8] In their investigation, a variant of the function help classification was being employed to distinguish anomalies. "Wagner et al." Utilising a single vector

enables classifiers to spot anomalies in analysis and different types of attachments, such as NetBIOS, DoS attacks, POP spammers, and SSH scanners. Kruegel and others [9] identifying TCP/IP data forwarding issues by use of a probabilistic model based on assumptions. Benferhat et al. detected a denial of service intervention detection using the same Bayesian method in their investigation. Koc and co. Using the same naive Bayes classifier, a multi-faceted intrusion detection approach was built. This dissertation devotes several reasons to KNN, another popular strategy for machine learning that establishes a dot's identification by the immediate neighbours. Amomani et al. [10] offer a thorough questionnaire on all major e-mail filters and machine learning (ML) technologies to recognise and comprehend typical phishing emails. The most recent research on specific dangers was given, and all of these strategies were evaluated side by side [11]. Dolly Uppal et al. [12] described a technique for identifying and categorising malware based on the n-gram algorithm. A before the programme was used to collect application data and keep track of test implementation. After building the vector, those that employed a variety of machine learning approaches then got the best results with the SVM classification. A methodology for evaluating organisational and staff operations using scale-hybrid -IDS-AlertNet is proposed by R. Vinaya Kumar et al. in [13]. A concept known as deep networking was created. They created a modular architecture that is based on Apache clusters and large-scale data frameworks.

III. Proposed Work

We only summarized models that were suitable for the particular cyber security concern. Authentication protocol can be resolved by strong strategies of feature discovery and classifiers such as Linear Regression (LR). KNN classifier, Decision Tree Classifier and AdaBoost Classifier can successfully overcome detection techniques To maintain a comprehensive model against malicious software and attain highly accurate results, micro procedures are needed. The choice of a specific design plays an important role in addressing cryptography problems. In our methodology, we first took note of the protection functions' classification per the significance and afterwards developed a simplified authentication scheme centered also on tree based on the selected important features

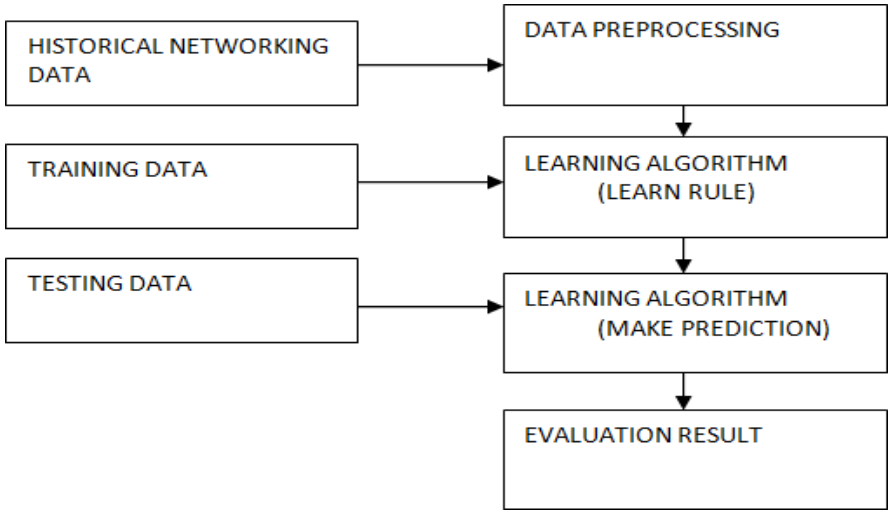


Figure 1: Proposed Block Diagram.

ISSN: 2581-3404 (Online) IF: 5.445 (SJIF)
IJIRTM, Volume-7, Issue-3, June-2023.

IV. Implementation & Result Analysis

For this purpose, machine learning algorithm to solve different computer security problems. While most scientists used all the models for computer vision with all four information safety problems, we only summarized models that were suitable for the particular cyber security concern. Authentication protocol can be resolved by strong strategies of feature discovery and classifiers such as Linear Regression (LR). KNN classifier, Decision Tree Classifier and AdaBoost Classifier can successfully overcome detection techniques

To maintain a comprehensive model against malicious software and attain highly accurate results, micro procedures are needed. The choice of a specific design plays an important role in addressing cryptography problems. In our methodology, we first took note of the protection functions' classification per the significance and afterwards developed a simplified authentication scheme centered also on tree based on the selected important features. People have achieved it to enhance the estimation precision of the protection model to unknown unit tests and to reduce computer cost by techniques described in a lower proportion when producing the resulting leaf structure. Finally, a variety of cyber security databases trials

Were conducted to test the feasibility of a DT design. The findings of the DT system have been contrasted with many conventional mainstream methods of master training to assess the efficiency of the corresponding security framework.

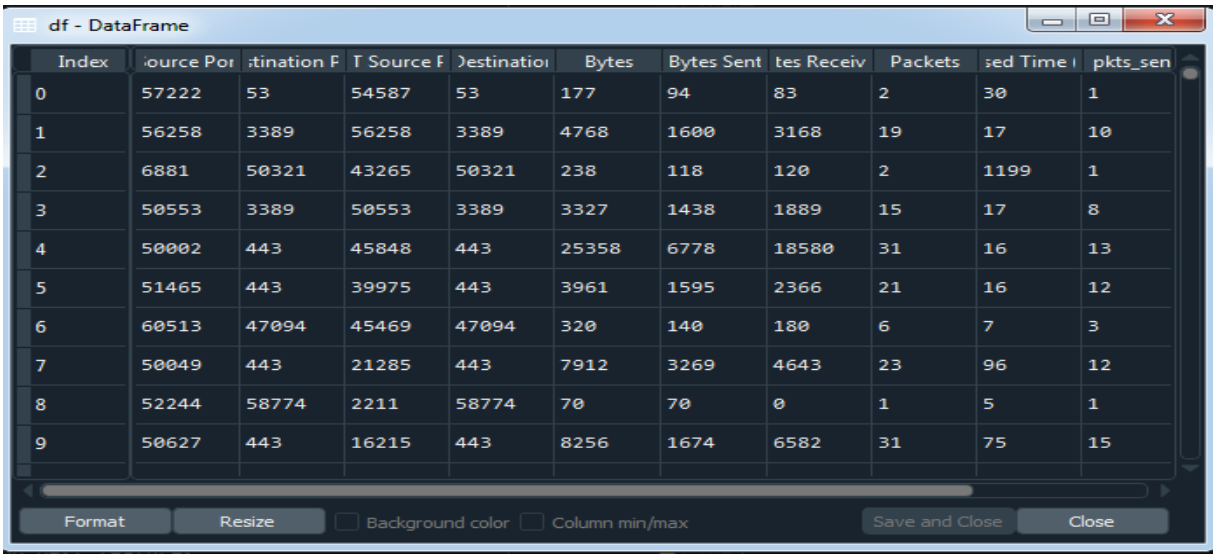
4.1 Datasets and Inputs

Internet Firewall Data Set

This data set was collected from the internet traffic records on a university's firewall. There are 12 features in total. Action feature is used as a class. There are 4 classes in total. These are allow, action, drop and reset-both classes, Attribute Information: Source Port, Destination Port, NAT Source Port, NAT Destination Port, Action, Bytes, Bytes Sent, Bytes Received, Packets, Elapsed Time (sec), pkts_sent, pkts_received.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn import metrics
from sklearn.metrics import accuracy_score, confusion_matrix
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import precision_score
from sklearn.metrics import recall_score
from sklearn.metrics import f1_score
from sklearn import metrics as ms
df=pd.read_csv("log2.csv")
df.shape
df.isnull().any()
df = df.fillna(method='ffill')
df.isnull().sum()
X=df.iloc[:,0:10]
X.hist()
plt.show()
y=df.iloc[:,-1]
df.head()
```

Figure 2: Dataset Loading Library.



Index	ource Por	tination F	T Source F	Destinatio	Bytes	Bytes Sent	tes Receiv	Packets	sed Time	pkts_sen
0	57222	53	54587	53	177	94	83	2	30	1
1	56258	3389	56258	3389	4768	1600	3168	19	17	10
2	6881	50321	43265	50321	238	118	120	2	1199	1
3	50553	3389	50553	3389	3327	1438	1889	15	17	8
4	50002	443	45848	443	25358	6778	18580	31	16	13
5	51465	443	39975	443	3961	1595	2366	21	16	12
6	60513	47094	45469	47094	320	140	180	6	7	3
7	50049	443	21285	443	7912	3269	4643	23	96	12
8	52244	58774	2211	58774	70	70	0	1	5	1
9	50627	443	16215	443	8256	1674	6582	31	75	15

Figure 3: Log2 Dataset.

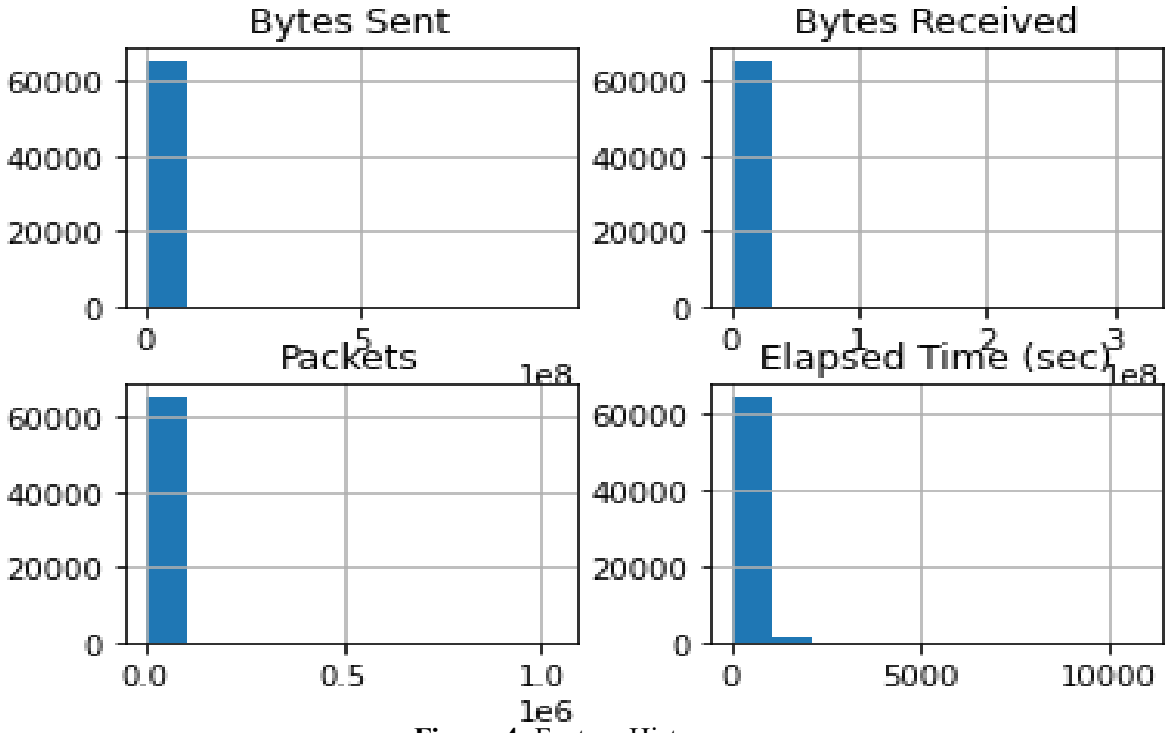


Figure 4: Feature Histogram.

4.2 Model Building
After preprocessing the data, we start building a machine learning model and use a simple linearregression model as a reference model. For this, follow these steps:
Linear Regression ModelStep 1:
Load preprocessed dataStep 2:
Split the data into training and test pairsStep 3:
Train a linear regression model on the training set and get predictions.Step 4:
Get test set predictionsStep 5:
Find Accuracy.

```
#linear regression classifier
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
regressor = LinearRegression()
regressor.fit(X_train, y_train)
print(regressor.coef_)
y_pred = regressor.predict(X_test)
df = pd.DataFrame({'Actual': y_test, 'Predicted': y_pred})
df1 = df.head(10)
plt.scatter(y_test,y_pred,  color='gray')
plt.plot(y_test, y_pred, color='red', linewidth=1)
plt.show()

print('Mean Squared Error:', metrics.mean_squared_error(y_test, y_pred))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
```

Figure 5: Linear Regression Classifier.

Linear Regression Model
Table 4.2.1:- Linear Regression Classification Report

	precision	recall	f1-score	support
accuracy			0.68	14200
macro avg	0.01	0.01	0.01	14200
weighted avg	0.63	0.68	0.68	14200

KNeighbors Classifier
Table 4.2.2 :- KNN Classification Report

	precision	recall	f1-score	support
accuracy			0.74	13107
macro avg	0.03	0.03	0.03	13107
weighted avg	0.71	0.74	0.72	13107

Adaboost Classifier

Table 4.2.3 :- Adaboost Classification Report

	precision	recall	f1-score	support
accuracy			0.75	16383
macro avg	0.01	0.01	0.01	16383
weighted avg	0.73	0.75	0.73	16383

Decision Tree Classifier

Table 4.2.4 :- Decision Tree Classification Report

	precision	recall	f1-score	support
accuracy			0.94	16383
macro avg	0.01	0.01	0.01	16383
weighted avg	0.91	0.94	0.94	16383

4.3 Comparison of models

Effectiveness comparison in terms of precision, recall, f1 score, and accuracy of different machine learning based security models.

EVALUATION CRITERIA

Accuracy: The precision of a forecast is just part of the success of the prototype. Precision has become one of the measurements used to test classification techniques.

True positive (TP) = the number of cases correctly. False positive (FP) = the number of cases incorrectly. True negative (TN) = the number of cases correctly. False negative (FN) = the number of cases incorrectly.

Accuracy: The accuracy of a test is its ability to differentiate the patient and healthy cases correctly. To estimate the accuracy of a test, we should calculate the proportion of true positive and true negative in all evaluated cases. Mathematically, this can be stated as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Sensitivity: The sensitivity of a test is its ability to determine the patient cases correctly. To estimate it, we should calculate the proportion of true positive in patient cases. Mathematically, thiscan be stated as:

Specificity: The specificity of a test is its ability to determine the healthy cases correctly. To estimate it, we should calculate the proportion of true negative in healthy cases. Mathematically, this can be stated as:

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

Precision: The optimistic scientific power is precision. Precision The quantity of true positive arguments in comparison to the total of positive claimed by the study is measured.

The precision score is a useful measure of the success of prediction when the classes are very imbalanced. Mathematically, it represents the ratio of true positive to the sum of true positive and false positive.
Precision Score = TP / (FP + TP)
Recall: The reminder is called the real positive rate, meaning the level of potential statements in the template similar to the real variety of positive outcomes in the data.
Recall Score = TP / (FN + TP)

F1 Score: F1 may calculate the output of a template as well. That is an average mean accuracy as well as a template alert.

$$F_1 = \frac{2}{\frac{1}{\text{recall}} \times \frac{1}{\text{precision}}} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$
$$= \frac{\text{tp}}{\text{tp} + \frac{1}{2}(\text{fp} + \text{fn})}$$

Table 4.3: Effectiveness comparison in terms of precision, recall, f1 score, and accuracy of different machine learning based security models.

S.No	Evaluationmetrics	LR	KNN	Ada boost	DT
1	Precision	0.63	0.71	0.73	0.91
2	Recall	0.68	0.74	0.75	0.94
3	F1 Score	0.68	0.74	0.75	0.94
4	Accuracy	0.68	0.74	0.75	0.94

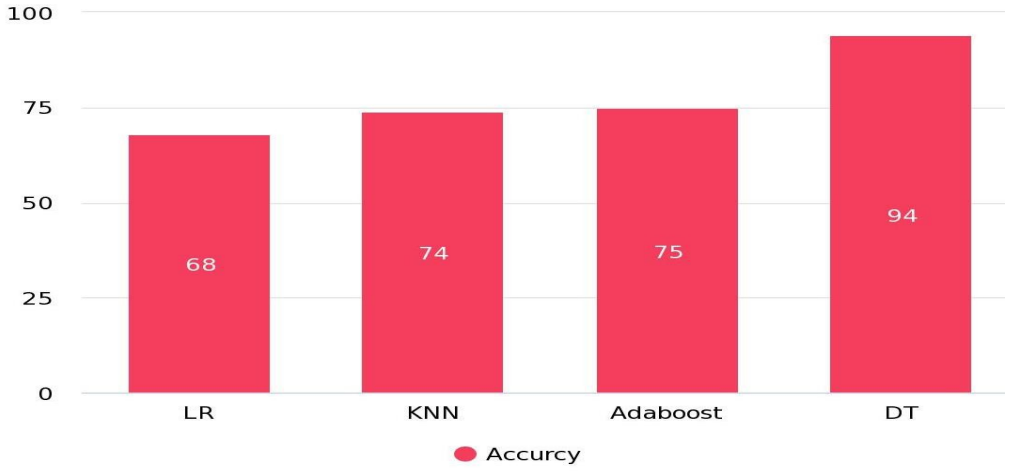


Figure 6: Comparison of models in terms of accuracy.

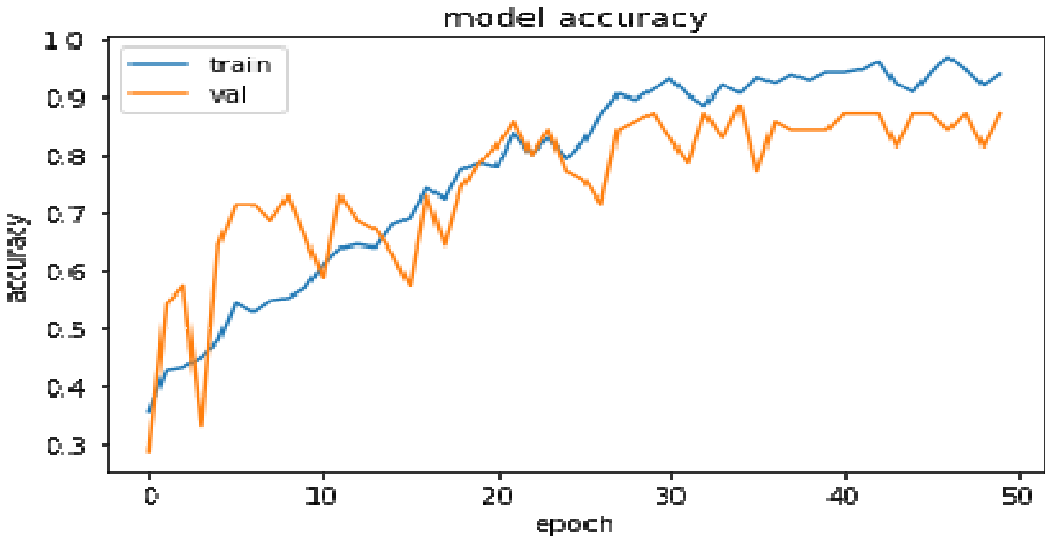


Figure 7: Model Accuracy.

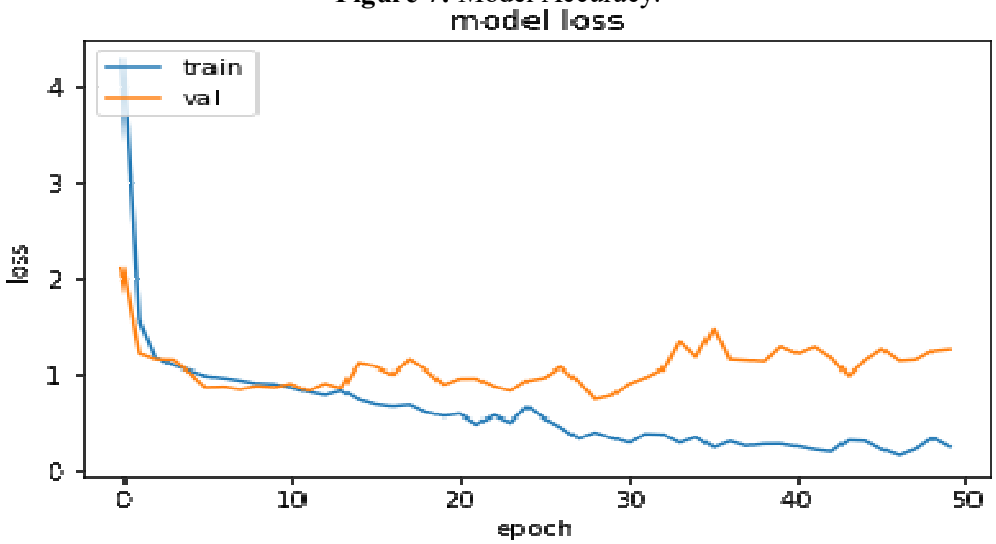


Figure 8: Model Loss.

V. Conclusion

Machine Learning methods are widespread for addressing different kinds of cyber security. Progress in artificial intelligence and critical thinking offers exciting options for network security crises. However, it is necessary to define which algorithm is sufficient for which purpose. To maintain a comprehensive model against malicious software and attain highly accurate results, micro procedures are needed. The choice of a specific design plays an important role in addressing cryptography problems. In our methodology, we first took note of the protection functions' classification per their significance and afterwards developed a simplified authentication scheme centered also on tree based on the selected important features. People have achieved it to enhance the estimation precision of the protection model to unknown unit tests and to reduce computer cost by techniques described in a lower proportion when producing the resulting leaf

structure. Finally, Effectiveness comparison in terms of precision, recall, f1 score, and accuracy of different machine learning based security models. We achieve 94% accuracy using DT Classifier were conducted to test the feasibility of a DT design. The findings of the DT system have been contrasted with many conventional mainstream methods of master training to assess the efficacy of the corresponding security framework.

VI. Future Scope

AI and ML are becoming increasingly important in the field of cyber security, as we have seen above. These technologies are being used to strengthen the security of organizations and individuals by automating repetitive tasks, detecting and classifying malware, analyzing network traffic, and identifying potential threats.

References

[1] S. Dolev and S. Lodha, In Proceedings of the First International Conference, CSCML 2017, Beer-Sheva, Israel, June 29-30, (2017).

[2] G. A. Wang, M. Chau, and H. Chen., Proceedings. Cham, Switzerland: Springer,May 23, (2017).

[3] J. Cano, ISACA Journal, 5, 1-5 (2016).

[4] C. Hollingsworth, ISACA Journal, 5, 1-6 (2016).

[5] X. Li, J. Wang, X. Zhang, J. Future Internet, (2017).

[6] M. Nalini and A. Chakram, International Journal of Innovative Technology and Exploring Engineering, 8, 197-201(2019).

[7] Y. Li, J. Xia, S. Zhang, J. Yan, X. Ai, K. Dai, Expert Syst. 39, 424–430, (2012).

[8] W. Hu, Y. Liao,V. R. Vemuri, In Proceedings of the International Conference on Machine Learning and Applications—ICMLA 2003, Los Angeles, CA, USA, 23–24,168–174 (2003).

[9] C. Kruegel, D. Mutz, W. Robertson, F. Valeur, In Proceedings of the 19th Annual Computer Security Applications Conference, Las Vegas, NV, USA, 14–23 (2003).

[10] Deepak Kumar Rathore, Dr. Praveen Kumar Mannepalli, “A Review of Machine Learning Techniques and Applications for Health Care “, International Conference on Advances in Technology, Management & Education, 2021, IEEE proceeding, 978-1-7281-8586-6/21.

[11] A. Almomani, B. Gupta, S. Atawneh, A. Meulenberg, E.Almomani IEEE Communications Surveys and Tutorials, 15 (2013).

[12] V. Padmanaban and M.Nalini, Proceedings of the 2019 international IEEE Conference on Innovations in Information and Communication Technology, (2019).

[13] D. Uppal, V. Jain, R. Sinha and V. Mehra, IEEE, (2014).

[14] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, A. N. Venkatraman, 7, (2019).

[15] R Dubey, D Rathore, D Kushwaha, JP Maurya, "An empirical study of intrusion detection system using feature reduction based on evolutionary algorithms and swarm intelligence methods", International Journal of Applied Engineering Research 12 (19), 2017. pp. 8884-8889.

[16] J. Gardiner, S. Nagaraja, ACM Comput Surv 49 1–59 (2016).

[17] M. Nalini and S. Anbu, International Journal of Applied Engineering Research, 9 (2014).

[18] Dipankar Dasgupta. Immunity-based intrusion detection system: A general frame- work. In Proceedings of the 22nd National Information Systems Security Confer- ence (NISSC). Arlington, Virginia, USA, 1999.