

Detecting Brute Force Attacks Using Machine Learning

Amer Ali Hamza ^{1*}, and Rana Jumma surayh Al-Janabi²

¹College of Computer Science and Information Technology, University of Al-Qadisiyah, Iraq

²College of Computer Science and Information Technology, University of Al-Qadisiyah, Iraq

Abstract: The importance of identifying network traffic abnormalities in cybersecurity cannot be emphasized enough, particularly considering the growing frequency and complexity of computer network assaults. With the emergence of new Internet-related technology, there is a corresponding increase in complex assaults. A significant difficulty is dictionary-based brute-force assaults (BFA), which need efficient real-time detection and mitigation techniques. This study explores the detection of SSH and FTP brute-force attacks via the use of the primary objective of our study is to use the machine learning methodology for the identification and detection of SSH and FTP brute-force assaults. Employing many classifiers enables a pretty comprehensive examination of the effectiveness of machine learners in spotting brute force attacks on SSH and FTP. Brute-force attacks are a widely used and perilous technique for acquiring usernames and passwords. Utilizing ethical hacking is a commendable method to assess the impact of a brute-force attack. This article explores several defense tactics and methodologies for using brute-force attacks. The advantages and disadvantages of many defense techniques are listed, along with details on the kind of attack that is most straightforward to recognize. we made use of machine learning (ML) classifiers: Naive Bayes (NB), decision Tree (DT), random forest (RF) Logistic Regression (LG), Quadratic Discriminant Analysis (QDA), Stochastic Gradient Descent (SGD), Linear Discriminant Analysis (LDA), Multi-Layer Perceptron (MLP), we determined that the Random Forest (RF) algorithm achieved the highest level with an accuracy of 99.9%.

1 Introduction

A brute force attack is a method where an attacker tries all possible Iterating through many combinations of passwords or encryption keys until the right one is discovered. It's a common method used to crack passwords. Due to the increasing dependency on digitalization, various security incidents, such as unauthorized access [1], Intrusion detection systems (IDS), firewalls, and antivirus software are just a few of the security measures available.

The early warning system against network assaults of the IDS makes significant contributions. On high-speed networks, encrypted communication makes it difficult to identify these kinds of assaults at the network level. [3], The study provides a thorough evaluation of many machine learning methods utilized in computer security systems.

[4] In this research, we show that just keeping an eye on the network may help us detect malicious activity like BFA. At the network level, identifying dictionary-based SSH and FTP brute-force assaults requires an efficient and high-performing technique.

Section II, machine learning is covered first. The literature review is described in Section III. The relevant works that were examined in the preceding part are covered in part IV. Section V contains the final presentation of the findings. The contribution to this is collecting previous work and identifying the benefits and harms of each research project, the strengths and weaknesses, the methods used, including algorithms and others, the accuracy of each classifier, and the type of data set used for the last five years, and then presenting the method used in my research, which is a very modern method that will achieve very high results.

* Corresponding author: amorali87@gmail.com

Due to the large volume of data communications the majority of which are benign we must carefully examine the traffic flow data in order to identify malicious activity. [5]. the attacker guessing list of usernames and passwords to get combinations to reach to the successful credential as shown in figure 1[6].



Fig. 1 Key steps of Brute Force Attack. [6]

During a dictionary attack, a hacker will use a dictionary file to attempt hundreds or possibly millions of popular passwords one by one until they find the right one. A hacker may access a system and all of its data if they are successful in carrying out a dictionary attack., the figure 2 shown how attacker using dictionary attacks to get logon.[7].

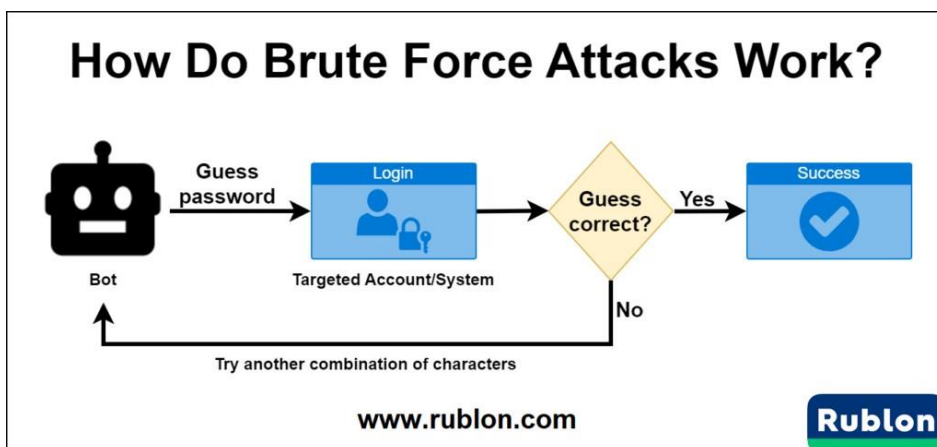


Fig. 2 Work architecture of Brute Force Attack.[7]

We study the performance of Detection of SSH and FTP brute-force attacks, together with the classifiers' classification accuracy.

The subsequent sections of this essay are organized in the following manner: Within Section II, machine learning is covered first. The literature review is described in Section III. The relevant works that were examined in the preceding part are covered in part IV. Section V contains the final presentation of the findings. The contribution to this is collecting previous work and identifying the benefits and harms of each research project, the strengths and weaknesses, the methods used, including algorithms and others, the accuracy of each classifier, and the type of data set used for the last five years, and then presenting the method used in my research, which is a very modern method that will achieve very high results.

1.1 Machine learning

Machine learning (ML) is a subfield of artificial intelligence (AI) and a branch of computer science that focuses on methods for system recognition. It enables computers to reason and learn without the need for explicit programming. ML is used in various computational tasks and aims to train machines using data that is provided to them. [8]

To improve results for the given problem, data can be categorized as either labeled, where input learning systems are paired with output variables, or unlabeled, where input learning systems are used without predicted output variables. Unsupervised learning encompasses algorithms for association mining and cluster analysis. Ultimately, the objective is for computers to learn from their previous experiences. [9]

The majority of the research on brute force assault detection has been on host-level detection. Access logs are examined at the host level, and an alert is generated if the quantity of unsuccessful login attempts during a certain period of time over a predetermined threshold. This study focuses Regarding the identification of brute-force assaults at the network level. When compared to a host-based detection technique, it scales more effectively.

Furthermore, network-based detection offers some defense for devices without internal (host-level) security and is essential in identifying network-based threats. Examine the use of machine learning for automated, flow-data-based network-level brute force attack detection (on the SSH and FTP protocols). Analyze a range of machine learning algorithms (classifiers) to detect brute force attacks. [10].

1.1.1 Categories of Brute Force Attack (BFA)

Each brute-force attack might utilize a variety of tactics to unearth confidential information. Any of the following common brute force techniques could be used against the intended victim:[33].

- **Basic Traditional Brute Force Attack:** This is a basic kind of brute force attack in which a hacker is given a username or list of usernames and tries to guess passwords until the right combination is discovered, either manually or by using a brute force programming script.
- **Reverse Brute Force Attacks:** This technique involves a hacker using a known password, either via a breach or regularly used, and systematically trying several usernames until a successful combination is discovered. These attacks vary from standard brute force or dictionary attacks in that they operate in reverse, beginning with known passwords rather than known users.
- **Dictionary Attacks:** A sophisticated technique in which a hacker uses a precompiled list of terms derived from extensive study on the target or tiny modifications of commonly used (or possible) passwords, in order to systematically test them against a given login. The selected list is regarded as a "lexicon" of modified or slightly modified words or character combinations.
- **Hybrid assaults** refer to the technique of merging classic brute force assaults with dictionary attacks. The hacker employs a technique known as dictionary attack, where they extract the most frequently used phrases and words from a predefined "dictionary" and systematically generate several password permutations until they successfully discover the correct combination.
- **Credential Stuffing** refers to a technique in which a hacker exploits a known set of login and password combinations to gain unauthorized access to additional accounts, profiles, or systems linked to the same user. This attack is successful because to the common practice of individuals reusing passwords across many accounts.

1.1.2 Identification and mitigation of brute force attacks against FTP and SSH protocols:

There are some ways to detect and prevent a brute force attack, Account lockout: Enforce a protocol that automatically restricts user accounts after a certain threshold of unsuccessful login attempts. This may aid in mitigating brute-force assaults.

Intrusion detection systems (IDS): Use an IDS to monitor network traffic and detect any unusual or repetitive patterns that may indicate a brute force attack in progress. IDS can analyze traffic and identify multiple failed logins attempts from a single or multiple IP addresses.

Two-factor authentication (2FA): Enforce the use of 2FA for FTP and SSH logins. This enhances security by requiring an additional piece of information to be provided, such as a one-time password or biometric verification, in addition to the regular login credentials.

IP blocking: Track repeated failed login attempts from specific IP addresses and automatically block them after a certain threshold is reached. This can be done at the firewall or server level.

Log monitoring and analysis: Regularly monitor the logs of your FTP and SSH servers for suspicious login attempts and analyze patterns to identify potential threats. This can help you proactively detect and respond to brute-force attacks.

Enforce rate-limiting: Deploy methods that limit the number of login attempts allowed within a certain period of time. This can help prevent automated attacks that attempt to deduce passwords by imposing restrictions on the number of permissible trials.

Use strong and inimitable passwords. And Use Web application firewalls (WAFS), and Employ a CAPTCHA.

2 Related works

Many academic papers have been written to propose solutions for reducing BFA attacks. However, despite the vast amount of academic research, BFA attacks continue to be widespread, especially when it comes to SSH and FTP attacks. Despite this, we will now present a summary of the most significant studies on these types of attacks to emphasize The significance of our job.

In the paper [11], John Hancock et al., The issue of brute force assaults in large data and the research on Big Data illustrates the viability of using simple decision tree models with two independent variables to precisely classify SSH and FTP brute force assaults. with dataset CSE-CIC-IDS2018 Accuracy% 0.99, are capable of detecting. Limitations in this research is Models trained on datasets that include just one feature are not dependable.

Stiawan et al. [12] Investigated a methodology for systematically testing various attack patterns in an Internet of Things (IoT) network environment. The brute-force assault was successfully identified. This study examines The FTP server of the IoT network was subjected to a brute force assault (BFA). It utilises a statistical correlation that is dependent on time. technique to discover the assault patterns and visualize them to determine the AC.% 96 percent

Najafabadi et al. [13] Conducted an investigation on detecting Network-level SSH brute-force attacks may be detected by analysing NetFlow data. A dataset specifically designed for attack detection was generated, using machine learning techniques that have shown effectiveness in recognizing brute-force attacks. The researchers investigated distributed SSH brute-force attacks and evaluated an 8-year login dataset including many users. It has been shown that some individual attack detection methods provide difficulties in terms of implementation, as indicated by AUC values over 0.97.

Satoh et al. [14] The researchers conducted an analysis of SSH dictionary attack detection using machine learning. Subsequently, they included two innovative components for detecting such attacks.

Kahara Wanjau et al. [15] The research presents a very effective approach for detecting SSH-brute force network attacks using a supervised deep learning method called Convolutional Neural Network (CNN). The CNN-based model outperforms existing machine learning approaches in detecting SSH brute force assaults. It has an F1-score of 91.8%, an accuracy rate of 94.3%,

An accuracy rate of 92.5% and a recall rate of 97.8% were achieved. The model was tested with the CIC-IDS 2018 dataset, which was pre-processed by converting raw data into images for training and testing. The study results demonstrate that deep neural networks (DNN) exhibit superior performance compared to other intrusion detection systems based on machine learning. The proposed method in the study combines feature selection and a deep learning algorithm for SSH-brute force attack detection.

In paper [16], Stephen Kahara Wanjau et al. Based on the Convolutional Neural Network, a supervised deep learning algorithm, this paper suggests a good way to find SSH brute force network attacks. The model's performance was compared to the outcomes of five well-known Machine learning techniques The mentioned machine learning algorithms include Naive Bayes, Logistic Regression, Decision Tree, k-Nearest Neighbour, and Support Vector Machine. Four often used metrics, namely accuracy, precision, recall, and the F-measure, were

utilized. Our investigation revealed that the CNN-based model outperforms conventional machine learning techniques in detecting SSH brute force assaults. The F1-score was 91.8%, the accuracy rate was 94.3%, the precision rate was 92.5%, and the recall rate was 97.8%.

In paper [17], Liang Zhou et al. Suggest an innovative methodology that leverages Machine learning techniques are used to assist in the classification of cyberattacks. We created a deep neural network (DNN) model and carefully determined the appropriate global parameters to attain outstanding generalization performance. The evaluation result demonstrates that the proposed methodology can effectively identify cyber-attacks in smart grids, with an accuracy rate of up to 96%.

In paper [18] by M. D. Hossain et al., The assaults occur the user's text is empty. Dictionary-based brute-force attacks (BFA) are prevalent forms of sophisticated cyber-attacks. This research investigates the identification of SSH and FTP brute-force assaults using the use of the Long Short-Term Memory (LSTM) deep learning technology. Furthermore, we used machine learning classifiers such as J48, naive Bayes, decision table, random forest, and k-nearest neighbour to augment our detection capabilities. We used the highly esteemed annotated dataset CICIDS2017. We evaluated the effectiveness of the LSTM and ML algorithms and performed a comparative study of their performance. The results suggest that the LSTM model outperforms the ML algorithms in terms of performance. achieving a precision level of 99.88%.

In paper [19] by Noura Alotibi et al., There has been an increase in the occurrence of brute-force assaults that specifically target FTP and SSH protocols. As a reaction, we provide a new and clever method that relies Utilising a dataset, the focus is on employing deep learning techniques to detect and classify brute-force attacks on FTP and SSH protocols. The CSE-CIC-IDS2018 approach achieves a remarkable accuracy rate of 99.9%, surpassing previous comparable approaches in identifying brute-force attacks. The model architecture used was LSTM combined with the SMOTE method, resulting accuracy 96.2.

In the paper [20], by Shailesh Singh Panwar et al., The primary emphasis is on seven distinct approaches, such as the brute force attack, achieved via the use of diverse Algorithms that pick features based on subsets. The execution assaults have occurred determined regarding several aspects. The use of these methodologies has led to the identification of the optimal set of qualities for detecting all types of attacks, using relevant classification algorithms. The efficacy of Intrusion Detection Systems (IDS). Performance assessment of brute force assaults using the classifier with a 90% accuracy rate and the CICIDS-2017 dataset for intrusion detection employing WEKA.

In paper [21], Karel Hynek et al. Suggest an innovative method for identifying SSH brute-force assaults on high-speed networks. Instead, then using host-based methods, our approach focuses on analyzing network traces to detect and identify intruders.

the existing resolution. In order to address the problem of elevated false positive rates, we suggest using a machine learning (ML) technique for detection that utilizes specifically expanded IP flows. Recent publications detail the methodology of accurately identifying BF assaults using just NetFlow data, achieving a high level of precision and a low percentage of false positives. Additionally, these articles discuss the structure and design of the detection system. The training dataset was constructed by meticulously examining actual traffic that was recorded, authentic SSH traffic that was manually manipulated to resemble brute force assaults, and ultimately a packet path including Logs of SSH activity from authentic production servers.

in a paper [22] by Joffrey L. Leevy et al., there has been a rise in cyberattacks to match The rapid expansion of computer networks and network applications on a global scale. The survey study we conducted has yielded some significant conclusions. Upon analysis, we found that the performance ratings for each research, when accessible, exhibited unusually high levels of achievement. This phenomenon might perhaps be attributed to the overfitting of large data studies. Furthermore, we found that the documentation about the data cleaning process of CSECIC-IDS2018 was insufficient in all areas, suggesting potential issues with the replicability of studies. Our survey has also found significant research deficiencies.

In the paper [23] by Stephen Kahara Wanjau et al., Brute force assaults are a prominent kind of network attack that presents significant dangers to network security. To stop in order to prevent the recurrence of such assaults, it is necessary to implement certain remedial measures. This research proposes an effective method for detecting SSH brute-force network assaults using a supervised deep learning methodology, namely a convolutional neural network, and a machine learning algorithm. The machine learning methods mentioned include (NB), LR, DT, k-NN and Support SVM, with 94.3%.

in paper [24] According to Maryam M. Najafabadi et al., A brute force assault is a very common network attack that poses a significant danger to machines linked to the network. This study examines the use of machine learning techniques to identify brute force attacks on the SSH protocol at the network level. With accuracy 95%.

in paper [25] Deris Stiawan et al. The File Transfer Protocol (FTP) server located at a data sink or gateway is frequently configured incorrectly. Simultaneously, password cracking and theft are prevalent methods used by attackers to target The Internet of Things (IoT) network. This study aims to provide a deeper understanding into this specific kind of assault, with the primary objective of identifying Possible assault strategies assist the administrator of the Internet of Things (IoT) system in analysing Comparable assaults. This study examines Brute force attacks (BFA) targeting the FTP server of the Internet of Things (IoT) network. It employs a temporally dependent statistical correlation technique to analyse and visualise the assault patterns that were observed may be identified.

TABLE 1. Comparison of SSH and FTP brute force Attacks Detection Schemes using ANN and CNN Architectures

Scheme	Data used	Model architecture	Result in %
[26]	CICIDS 2017, UNSW-NB15, NSL-KDD, Kyoto, WSN-DS	ANN + ReLU activation	Accuracy: 78.5, 95.6, Precision: 81.0, 96.2, Recall: 78.5, 95.6, F1- score: 76.5, 95.7
[27]	ISCX VPN	CNN	accuracy: 99.85
[28]	NSL-KDD	ANN + ReLU activation	Accuracy: 86.35, Precision: 81.86, Recall:77.32, F1-score: 73.89, FAR: 0.1619
[29]	KDD 99	ANN + ReLU activation	Accuracy: 99.01, Recall:99.81, FAR: 0.0047
[30]	Network data Simulated by IoT	ANN + Sigmoid activation	Accuracy: 99

TABLE 2. Comparison of SSH and FTP brute force Attacks LSTM RNN and Other Deep Learning Architectures

Scheme	Data used	Model architecture	Result in %
[31]	NSL-KDD, binary and 5-class classification	RNN	Accuracy: 81.29
[32]	KDD 99	LSTM network	Accuracy: 97.54, Precision: 97.69, Recall:98.95, FAR: 9.98
[33]	CSE-CIC- IDS2018	Broad Learning System	Accuracy: 97.08 F1- score: 77.89 Precision: NA Recall: NA
[34]	CSE-CIC- IDS2018	LSTM+ SMOTE algorithm	Accuracy :96.2 F1- score: NA Precision :96 Recall :96
[35]	CSE-CIC- IDS2018	Spark ML + Conv-AE	Accuracy: 98.20 F1- score: 98 Precision: NA Recall :98

3 Methodology

The methodology section outlines the process followed in developing the Brute Force attack on SSH and FTP protocol detection model using Machine Learning techniques. The chosen algorithm is based on the CSE-CIC-IDS 2018 dataset, specifically the FTP/SSH brute-force attacks, serve as the basis for the model. The entire process is broken down into distinct stages, as detailed below:

Obtain the proposed benchmark dataset: The CSE-CIC-IDS 2018 dataset is acquired, containing eight different attack types. Only FTP/SSH brute-force attacks are used in this study.

Prepare the data: Data preprocessing involves correcting issues such as missing values and outliers, ensuring that the dataset is clean and ready for analysis.

Use exploratory analysis: This step involves understanding the dataset's content and selecting the most suitable algorithm for the given problem.

Train the model: The best-performing algorithm from the literature review is used to train the model on the prepared dataset.

Evaluate the model: Evaluation techniques are employed to assess the model's performance and ensure that it meets the desired accuracy and detection standards.

Optimize the model: If the model's performance is unsatisfactory, alternative algorithms are considered or the current model's parameters are adjusted to improve its effectiveness.

3.1 Proposed Model

In order to carry out the algorithm that has been suggested, it is of utmost importance to acquire the CSE-CIC-IDS2018 benchmark dataset. This dataset is structured in a CSV file format, which contains several columns such as FlowID, Destination-IP, Source-Port, and Protocol. Within this dataset, one can find a wealth of information encompassing more than 80 different network traffic characteristics. These characteristics have been carefully chosen to encompass a wide range of attack types, including but not limited to denial-of-service attacks, Heartbleed attacks, web attacks, botnet attacks, and infiltration attacks. as shown in figure 3

The following steps outline the data preprocessing:

Data cleaning and normalization: Convert all required features to nominal values and clean the data, depending on the created features.

Normalization of data: Normalize numeric feature values to a chosen scale, such as the [0, 1] range, to decrease data scale and improve model accuracy and processing time.

Splitting data: Divide the data into three parts: a training set with 60% of the data, a validation set with 20%, and a testing set with the remaining 20%.

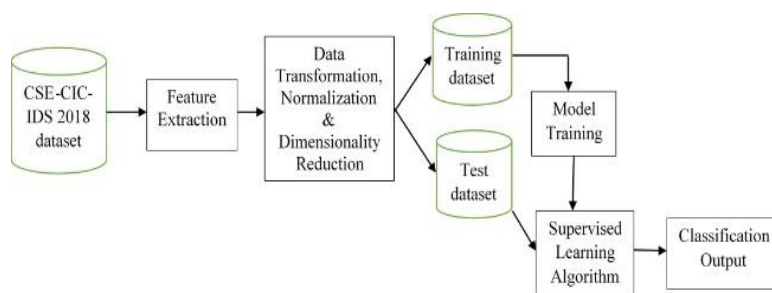


Fig. 3 displays the architecture of the proposed model

Fig 3. Proposed Classifier for SSH- and FTP Brute force attack detection.

- Obtain the proposed benchmark dataset: The CSE-CIC-IDS 2018 dataset is acquired, containing eight different attack types. Only FTP/SSH brute-force attacks are used in this study.
- Prepare the data: Data preprocessing involves correcting issues such as missing values and outliers, ensuring that the dataset is clean and ready for analysis.
- Use exploratory analysis: This step involves understanding the dataset's content and selecting the most suitable algorithm for the given problem.

- Train the model: The best-performing algorithm from the literature review is used to train the model on the prepared dataset. And test model.
- Evaluate the model: Evaluation techniques are employed to assess the model's performance and ensure that it meets the desired accuracy and detection standards

3.2 Dataset

We use data set CSE-CIC-IDS2018 in our search The Communications Security Establishment (CSE) and the Canadian Institute for Cybersecurity (CIC) developed the CSE-CIC-IDS2018 dataset to meet the needs of the attack detection benchmark dataset that represents traffic composition and attack on the current modern network This dataset consists of 80 features, including labels. The dataset is collected from Amazon’s AWS LAN network It includes seven different attack classes features extracted from the captured traffic using CICFlowMeter-V3. The output of the application is in CSV file format with six columns labeled for each flow, namely FlowID, SourceIP, DestinationIP, SourcePort, DestinationPort, and Protocol with more than 80 network traffic features.

3.3 Algorithms and Techniques

This section gives brief overviews of the three ML algorithms used in this work and their graphical presentations.

3.3.1 Random Forest

The random forest (RF) method is comprised of several decision trees. Every one of these trees produces a forecast. Subsequently, the algorithm utilizes these predictions to formulate a judgment by considering the majority of the expected values. RF has many benefits, including as its versatility in solving both classification and regression issues, its independence from scaling requirements, and its capability to effectively manage outliers. The approach has many drawbacks, such as its high processing requirements due to the involvement of numerous decision trees, resulting in lengthier training times for models. as shown in figure 4.

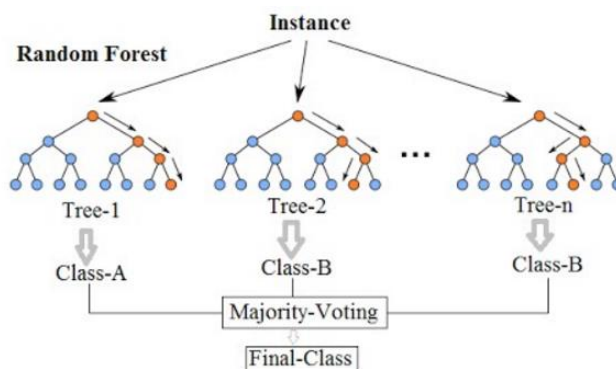


Fig. 4 Random Forest (RF)

3.3.2 Gaussian Naive Bayes

Bayesian classifiers are statistical classifiers. They may predict the likelihood that the supplied model is suitable for a certain class. Gaussian Naive Bayes (GNB) is a variant of the Naïve Bayes algorithm that utilizes Bayes's theorem. The underlying assumption of these algorithms is that, within a certain class, the attribute value is unrelated to the values of other attributes. The name given to this theory is class-conditional independence. It is mostly used for datasets containing continuous data. This approach presupposes that classes adhere to a Gaussian distribution. The Gaussian distribution, often known as the normal distribution, may be examined using the following formula. Some of the benefits of utilizing GNB include that it is a rapid technique to train, is good

for datasets with numerous classes, and is typically used for categorical issues. An inherent drawback of this approach is its tendency to consider each characteristic in isolation, a scenario that does not necessarily reflect real-world conditions. This renders the algorithm less applicable to real-world scenarios, as shown in figure 5.

Fig. 5. Gaussian Naive Bayes (GNB)

3.3.3 Logistic Regression (LR)

Logistic regression (LR) is a statistical technique used for binary classification, which estimates the chance of an event occurring by using a probability function. The probability is calculated using the following formula. Some benefits of this method include its ability to quickly classify data and its ease of extension to handle multi-class problems. An inherent limitation of LR is its inability to address nonlinear problems, as shown in figure 6.

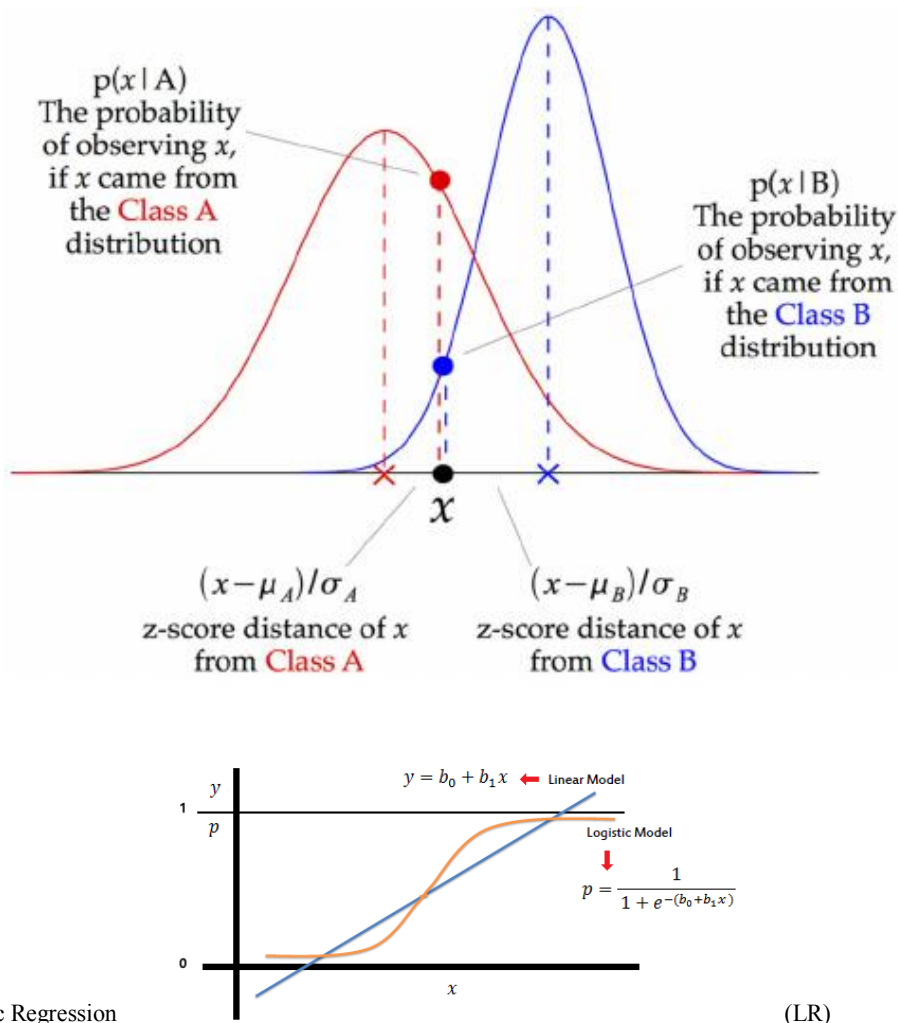


Fig.6. Logistic Regression

(LR)

4 Result

Our programming language of choice is Python, and we use Keras with TensorFlow as the underlying framework. The execution of all the tests was completed. The system specifications include an Intel Core i7 CPU running at a clock speed of 2.20 GHz, 16 GB of RAM, Windows 10 operating system (64-bit), and an NVIDIA GeForce GTX 1050 graphics card. for the detection of brute-force attacks on SSH and FTP protocols. By using machine learning classifiers, it was determined that the Random Forest (RF) algorithm achieved the highest level of accuracy. With 99.9%

4.1 Confusion Matrix

A confusion matrix is a powerful tool for the investigation of incorrect classifications. Unlike other evaluation techniques, the confusion matrix highlights the number of correct and incorrect predictions. It reveals classes that the model considers similar or distinguishes well. In the obtained figures, we can analyze the true values versus the predicted values. An ideal model shows a confusion matrix with most of the predictions in a diagonal line from top left to bottom right. This means that the prediction made by the model correlates with the actual answer. Each number in the diagonal compared to the numbers in the corresponding horizontal line is of interest. This allows the analysis of the predictions per actual class. The number on the diagonal shows the correct classifications, while the other numbers on the left and right denote the numbers of incorrect predictions per class.

The confusion matrix of the random forest (RF) model shows its high accuracy as most numbers are diagonal. This means that the RF model correctly predicted most of the attacks. On the other hand, other models were only able to detect benign attacks or one type of attack. The results of confusion matrices are critical because they highlight what is predicted by each model for every given sample.

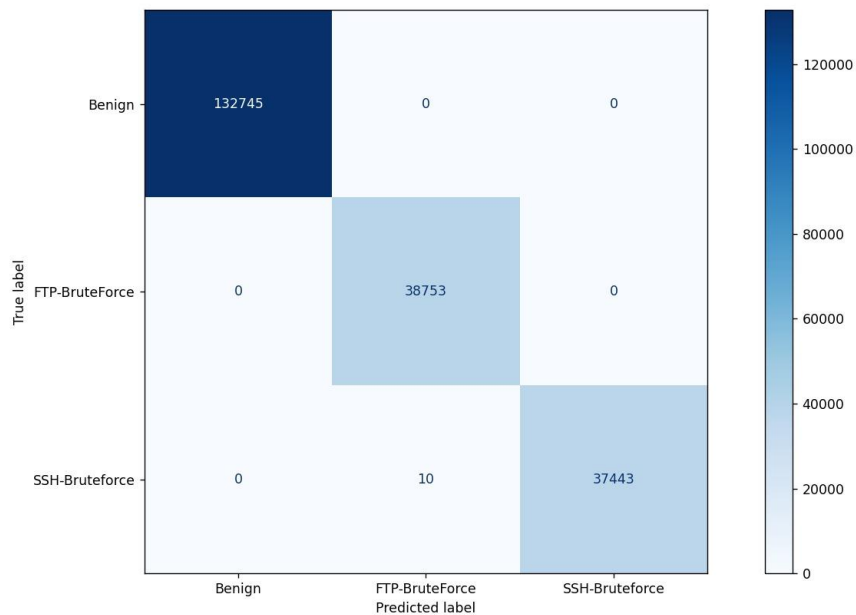


Fig.7. Confusion Matrix of Random Forest (RF)

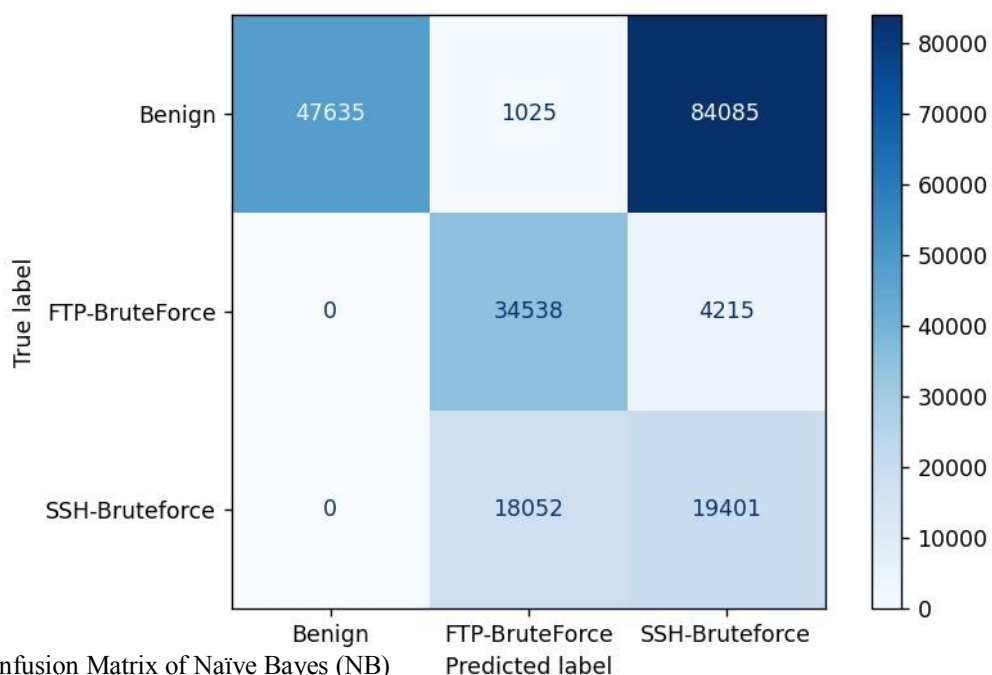


Fig.8 Confusion Matrix of Naïve Bayes (NB)

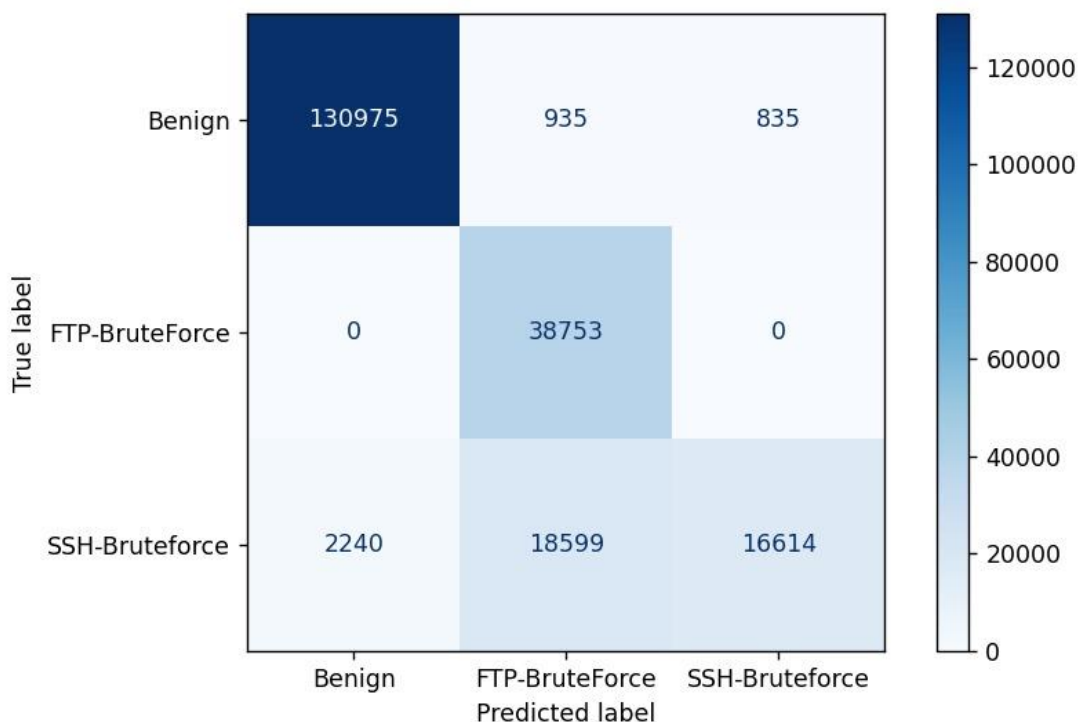


Fig.9 Confusion Matrix of Naïve Bayes (NB)

4.2 Classification Report

Accuracy of Random Forest (RF) : 99.99521418897254
 F1-Score of Random Forest (RF) : 99.99521417817371
 Precision of Random Forest (RF) : 99.995215422360634
 Recall of Random Forest (RF) : 99.99521418897254

Accuracy of Naïve Bayes (NB) : 48.611396930380806
 F1-Score of Naïve Bayes (NB) : 52.214919357857134
 Precision of Naïve Bayes (NB): 78.70544304565682
 Recall of Naïve Bayes (NB) : 48.611396930380806

Accuracy of Logistic Regression (LR) : 89.17975984800263
 F1-Score of Logistic Regression (LR): 88.23268325355058
 Precision of Logistic Regression (LR) :91.8584544599045
 Recall of Logistic Regression (LR): 89.17975984800263

In the section dedicated to the Findings and Analysis, we shall proceed to divulge the performance metrics of our meticulously developed model and conduct a comprehensive comparison with the most current research findings. Our ground breaking model has demonstrated unparalleled excellence by attaining a staggering accuracy rate surpassing the threshold of 99.9%, effectively outshining the existing literature in this field. Moreover, the inclusion of Figure 4, which meticulously depicts the learning curves observed during the arduous training process, serves as a visual testament to the remarkable performance exhibited by our proposed model.

4.1.2 Classification Report

Method	Accuracy%	F1-Score%	Precision%	Recall%
Random Forest (RF)	99.9952	99.9952	99.9952	99.9952
Naïve Bayes (NB)	48.6113	52.2149	78.7054	48.6113
Logistic Regression (LR)	89.1797	88.2326	91.8584	89.1797

In the section dedicated to the Findings and Analysis, we shall proceed to divulge the performance metrics of our meticulously developed model and conduct a comprehensive comparison with the most current research findings. Our ground breaking model has demonstrated unparalleled excellence by attaining a staggering accuracy rate surpassing the threshold of 99.9%, effectively outshining the existing literature in this field. Moreover, the inclusion of Figure10, which meticulously depicts the learning curves observed during the arduous training process, serves as a visual testament to the remarkable performance exhibited by our proposed model.

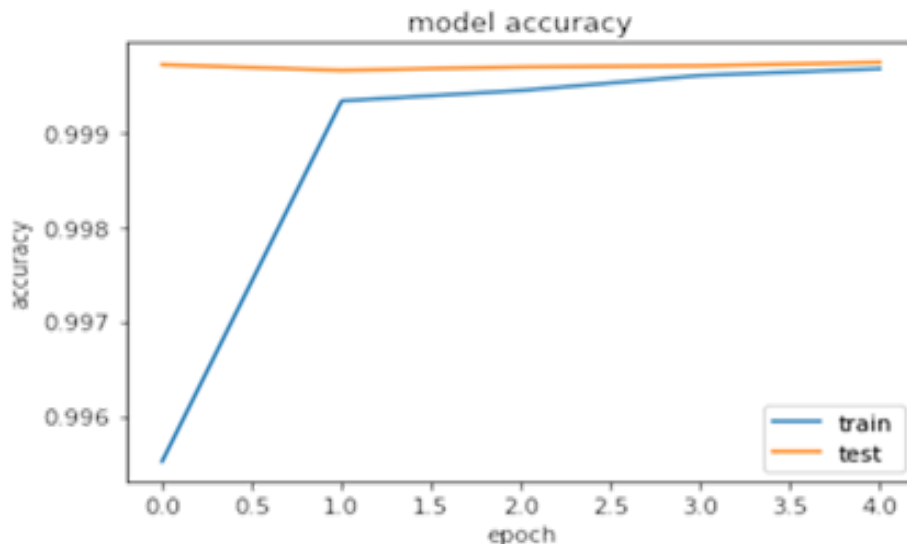


Fig. 10. The accuracy of the proposed model.

Table 3 compares our proposed model with current research based on accuracy, precision, recall, and F-score. Our model produced superior results, with 99.9% accuracy, 99.9% F-score, 99.9% precision, and 99.9% recall.

Table 3. Comparison between the Proposed Model and Current research

Scheme	Dataset	Model Architecture	Result in %
[36]	CSE-CIC- IDS2018	Broad Learning System	Accuracy: 97.08 F1- score: 77.89 Precision: NA Recall: NA
[37]	CSE-CIC- IDS2018	LSTM+ SMOTE algorithm	Accuracy :96.2 F1-score: NA Precision :96 Recall :96
[38]	CSE-CIC- IDS2018	Spark ML + Conv-AE	Accuracy: 98.20 F1- score: 98 Precision: NA Recall :98
Proposed Model	CSE-CIC- IDS2018	RF	Accuracy: 99.9 F1- score: 99.9 Precision: 99.9 Recall: 99.9

These results show that the proposed Random Forest model is significantly more accurate than other models, particularly in terms of precision. The high precision of the proposed model indicates that it is particularly effective at correctly identifying true positives (i.e., correctly detecting attack instances) and minimizing false positives (i.e., misclassifying benign instances as attacks). This is an essential aspect of an

intrusion detection system, as it ensures that genuine threats are identified while minimizing the risk of false alarms.

The recall rate of 99.9% for the proposed model, is still noteworthy. This indicates that the model can successfully identify a high proportion of true positive instances from the total number of actual positive instances. In the context of intrusion detection, this means that the proposed model is effective at detecting most of the attacks in the dataset.

5 Discussion

Based on the assessment findings, machine learning algorithms exhibited a high level of accuracy. However, this is due to the fact that the number of benign occurrences in the dataset is larger than the number of instances of SSH and FTP assaults. Therefore, In the performance analysis, we evaluated precision, recall, F1-score, and AUC-ROC. During the confusion matrix analysis, we noted that the overall weighted classifier accuracy, precision, recall, and F1-score were all high. The F1-score should be increased. When it comes to deploying these types of models in real-time settings. Based on the AUC-ROC curve, we saw that the area under the curve approached 1.0, indicating strong evidence that (RF) model is effective in identifying FTP and SSH brute-force assaults. The (RF) Intrusion Detection System (IDS) we present is very efficient in identifying and detecting FTP and SSH brute-force assaults on network systems.

It is feasible to deploy a model in real-time to identify brute force assaults. The performance of (RF) was subpar in our observations. To identify web brute-force assaults. Our next study will focus on enhancing the detection rate and F1-score of these assaults by using other deep learning models.

6 Conclusion

In this paper we proposed efficient strategies to promptly identify and counteract such brute-force assaults instantaneously. Multiple machine learning (ML) classifiers are used to analyze and identify Secure Shell (SSH) and File Transfer Protocol (FTP) brute-force assault Identification. The following machine learning algorithms include Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), Logistic Regression (LG), Quadratic Discriminant Analysis (QDA), Stochastic Gradient Descent (SGD), Linear Discriminant Analysis (LDA), and Multi-Layer Perceptron (MLP).

The objective of detecting brute-force attacks is to identify and prevent unauthorized intrusion into computer systems or networks via the process of detection repeated, rapid attempts to guess passwords or encryption keys. By detecting these types of attacks, security measures can be implemented to block or minimize the impact of such malicious activities, safeguarding data and ensuring system integrity. This paper Provides a thorough examination of the many methods used in carrying out brute-force assaults, along with recommendations for safeguarding oneself against such attacks. It becomes evident that a combination of multiple protective measures is imperative, thereby enhancing security. It has been demonstrated that the execution of a brute-force attack is comparatively straightforward yet highly extensive, resulting in considerable harm to the targeted users. The benefits and drawbacks of distinct protective mechanisms are underscored, facilitating users in selecting the most appropriate combination of protection methods. End users should adopt brute-force attack defense strategies on a widespread basis because they have proven to be incredibly effective in real-world situations. Future endeavors will concentrate on monitoring novel brute-force attack methodologies and conducting vulnerability analyses of targeted systems while adhering to the principles of ethical hacking to identify optimal defense strategies.

References

1. H. Sarker, A. S. M. Kayes, S. Badsha, H. Alqahtani, P. Watters, and A. Ng, "Cybersecurity data science: an overview from machine learning perspective," *J Big Data*, vol. 7, pp. 1–29, 2020.
2. M. Ring, S. Wunderlich, D. Scheuring, D. Landes, and A. Hotho, "A survey of network-based intrusion detection data sets," *Comput Secur*, vol. 86, pp. 147–167, 2019.
3. K. Kim, M. E. Aminanto, and H. C. Tanuwidjaja, *Network intrusion detection using deep learning: a feature learning approach*. Springer, 2018.

4. G. Apruzzese, M. Colajanni, L. Ferretti, A. Guido, and M. Marchetti, "On the effectiveness of machine and deep learning for cyber security," in 2018 10th international conference on cyber Conflict (CyCon), IEEE, 2018, pp 371–390.
5. N. Bakhareva, A. Shukhman, A. Matveev, P. Polezhaev, Y. Ushakov, and L. Legashev, "Attack detection in enterprise networks by machine learning methods," in 2019 international Russian automation conference (RusAutoCon), IEEE, 2019, pp. 1–6.
6. <https://www.spiceworks.com/it-security/cyber-risk-management/articles/what-is-brute-force-attack/>
7. <https://rublon.com/blog/brute-force-dictionary-attack-difference/>
8. B. K. Bhavitha, A. P. Rodrigues, and N. N. Chiplunkar, "Comparative study of machine learning techniques in sentimental analysis," in 2017 International conference on inventive communication and computational technologies (ICICCT), IEEE, 2017, pp. 216–221.
9. S. Das, A. Dey, A. Pal, and N. Roy, "Applications of artificial intelligence in machine learning: review and prospect," *Int J Comput Appl*, vol. 115, no. 9, 2015.
10. M. M. Najafabadi, T. M. Khoshgoftaar, C. Kemp, N. Seliya, and R. Zuech, "Machine learning for detecting brute force attacks at the network level," in 2014 IEEE International Conference on Bioinformatics and Bioengineering, IEEE, 2014, pp. 379–385.
11. J. Hancock, T. M. Khoshgoftaar, and J. L. Leevy, "Detecting SSH and FTP Brute Force Attacks in Big Data," in Proceedings - 20th IEEE International Conference on Machine Learning and Applications, ICMLA 2021, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 760–765. doi: 10.1109/ICMLA52953.2021.00126.
12. D. Stiawan, M. Idris, R. F. Malik, S. Nurmaini, N. Alsharif, and R. Budiarto, "Investigating brute force attack patterns in IoT network," *Journal of Electrical and Computer Engineering*, vol. 2019, 2019.
13. M. M. Najafabadi, T. M. Khoshgoftaar, C. Kemp, N. Seliya, and R. Zuech, "Machine learning for detecting brute force attacks at the network level," in 2014 IEEE International Conference on Bioinformatics and Bioengineering, IEEE2014, pp. 379–385.,
14. A. Satoh, Y. Nakamura, and T. Ikenaga, "SSH dictionary attack detection based on flow analysis," in 2012 IEEE/IPSJ 12th International Symposium on Applications and the Internet, IEEE, 2012, pp
15. S. Kahara Wanjau, G. M. Wambugu, and G. Ndung'u Kamau, "SSH-Brute Force Attack Detection Model based on Deep Learning," 2021. [Online]. Available: www.ijcat.com
16. S. K. Wanjau, G. M. Wambugu, and G. N. Kamau, "SSH-brute force attack detection model based on deep learning," 2021.
17. L. Zhou, X. Ouyang, H. Ying, L. Han, Y. Cheng, and T. Zhang, "Cyber-attack classification in smart grid via deep neural network," in Proceedings of the 2nd international conference on computer science and application engineering, 2018, pp. 1–5.
18. M. D. Hossain, H. Ochiai, F. Doudou, and Y. Kadobayashi, "ssh and ftp brute-force attacks detection in computer networks: Lstm and machine learning approaches," in 2020 5th international conference on computer and communication systems (ICCCS), IEEE, 2020, pp. 491–497
19. N. Alotibi and M. Alshammari, "Deep Learning-based Intrusion Detection: A Novel Approach for Identifying Brute-Force A [20] Panwar, S. S., Negi, P. S., Panwar, L. S., & Raiwani, Y. P. (2019). Implementation of machine learning algorithms on cids-2017 dataset for intrusion detection using WEKA. *International Journal of Recent Technology and Engineering*, 8(3), 2195–2207tacks on FTP and SSH Protocol." [Online]. Available: www.ijacsa.thesai.org
20. Panwar, S. S., Negi, P. S., Panwar, L. S., & Raiwani, Y. P. (2019). Implementation of machine learning algorithms on cids-2017 dataset for intrusion detection using WEKA. *International Journal of Recent Technology and Engineering*, 8(3), 2195–2207
21. Hynek, K., Beneš, T., Čejka, T., & Kubátová, H. (2020). Refined detection of SSH brute-force attackers using machine learning. *ICT Systems Security and Privacy Protection: 35th IFIP TC 11 International Conference, SEC 2020, Maribor, Slovenia, September 21–23, 2020, Proceedings 35*, 49–63
22. J. L. Leevy and T. M. Khoshgoftaar, "A survey and analysis of intrusion detection models based on CSE-CIC-IDS2018 Big Data. *J. Big Data* 7, 104 (2020)."
23. S. Kahara Wanjau, G. M. Wambugu, and G. Ndung'u Kamau, "SSH-Brute Force Attack Detection Model based on Deep Learning," 2021. [Online]. Available: www.ijcat.com
24. M. M. Najafabadi, T. M. Khoshgoftaar, C. Kemp, N. Seliya, and R. Zuech, "Machine learning for detecting brute force attacks at the network level," in 2014 IEEE International Conference on Bioinformatics and Bioengineering, IEEE, 2014, pp. 379–385

25. D. Stiawan, M. Idris, R. F. Malik, S. Nurmaini, N. Alsharif, and R. Budiarto, "Investigating brute force attack patterns in IoT network," *Journal of Electrical and Computer Engineering*, vol. 2019,2019