IJNRD.ORG ISSN : 2456-4184

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG

An International Open Access, Peer-reviewed, Refereed Journal

DETECT AND ASSESS THE CYBERSECURITY THREATS WITH THEIR MITIGATION APPROACHES USING ML ALGORITHMS

Abhishek Kumar Singh
Student, CSE

Amit Kumar
Student, CSE

Biswajit Raha
Student, CSE

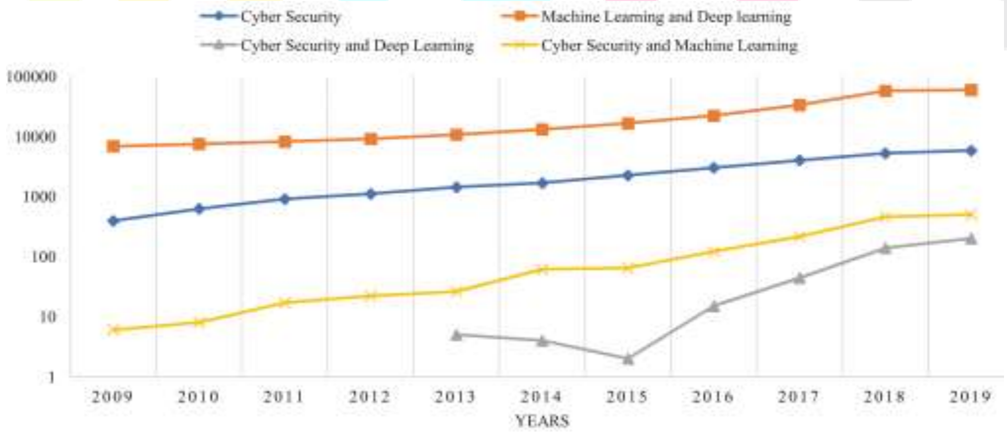
Bibhas Kumar Shahee
Student, CSE

Department Of Computer Science
East Point College of Engineering and Technology
Bangalore, Karnataka
India

Abstract: Cybersecurity is a critical concern for organizations worldwide. Machine learning techniques are being used to detect and prevent cyber-attacks. Random forest algorithm is a popular machine learning technique that can be used to develop an ML model to detect various types of cyber threats in data. This model can significantly improve cybersecurity measures and help prevent data loss, data misuse, and stealing data. The model is trained on a large dataset of historical cyber-attack data. The model is able to identify patterns in the data that can be used to predict future attacks. The model is able to detect attacks in their early stages, before they cause significant damage. The model is a valuable tool for organizations that are looking to improve their cybersecurity posture.

1.INTRODUCTION

The rapid growth of cyberspace, fueled by the widespread adoption of the internet and mobile applications, has led to an increase in automated and persistent cyberattacks. Traditional security systems are no longer effective against these evolving threats, as hackers continuously develop more sophisticated methods. To address this challenge, machine learning (ML) techniques have emerged as crucial tools in cybersecurity. ML algorithms can detect and mitigate various cybersecurity threats, such as spam, fraud, malware, phishing, and intrusion attempts. With the global digital population reaching 5.16 billion people in January 2023, the internet has become an attractive target for cybercriminals. Consequently, cybersecurity strategies are more important than ever in safeguarding user assets and ensuring the privacy, integrity, and availability of data. This project aims to develop an ML-based system to detect and prevent cyber-attacks by training a model using a dataset and generating attacks for analysis. By integrating ML into cybersecurity, we can enhance threat detection and mitigation capabilities.



1.1 Machine Learning and Cybersecurity:

In recent years, machine learning (ML) techniques have become increasingly essential in many spheres of life, including cyber security. ML approaches have proven to be a promising tool for detecting and mitigating cybersecurity threats, such as spam classification, fraud detection, malware detection, phishing, dark web or deep web sites, and intrusion detection.

1.2 Background Study:

The study of cybersecurity threats and their mitigation using ML algorithms has become a rapidly growing field in recent years. With the increasing dependence on technology and the internet for both personal and business activities, there has been a corresponding rise in cyber-attacks and security breaches. For instance, since 2013, cyber security ventures have reported 38,09,448 records stolen from breaches every day, which amounts to 158,727 per hour, 2645 per minute and 44 per second.

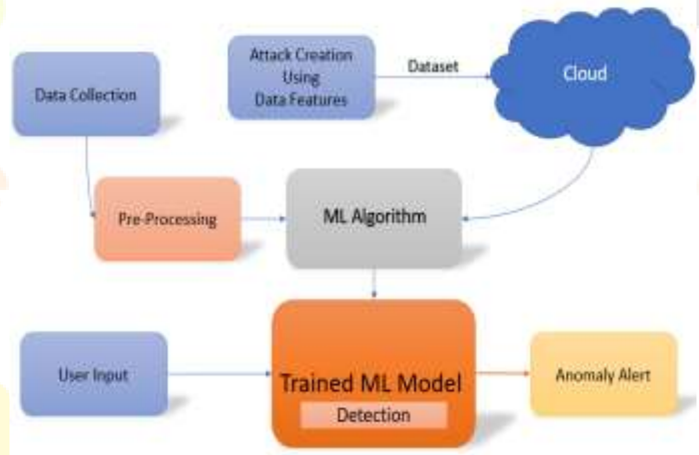
1.3 Project Objectives:

In this project, an ML model will be trained using a dataset to detect the threat in data such as DoS, Probe, R2L, and U2R. Additionally, an attack will be generated using features in the dataset and stored on the cloud. By detecting the type of threat, it will be useful in preventing such attacks on any dataset. Delicate processing-based IDS embraces several computational insight philosophies, including counterfeit brain organizations, fuzzy logic, transformative calculation, probabilistic computing, fake resistant frameworks, conviction networks and so on. The integration of machine learning techniques into cyber security has become a critical area of research and development to improve the detection and mitigation of cybersecurity threats. The proposed project will contribute to this field by developing an ML-based system to detect and prevent cyber-attacks**.**

2.SYSTEM ARCHITECTURE

The proposed architecture for detecting cyber-attacks using machine learning techniques involves several steps, as illustrated in the diagram. The first step is to collect data related to cyber-attacks, which can include features such as packet headers, network traffic flow, and system log files. This data is then pre-processed to extract relevant features and prepare it for input into the machine learning algorithm. The machine learning algorithm used in this architecture is the Random Forest algorithm, which is a decision tree-based ensemble learning method. This algorithm has been shown to be effective in detecting anomalies and is suitable for use in cyber security applications. Once the ML algorithm is trained, it can be used to classify new data as either normal or anomalous. In this architecture, the user inputs data into the trained ML model, and the model outputs an anomaly alert if the data is classified as anomalous. To make the system more efficient and scalable, the architecture includes the use of cloud storage. The attack creation process, using features from the dataset, is generated and stored on the cloud. The anomaly detection process is then carried out in the cloud environment, which allows for the processing of large amounts of data and faster response times.

Overall, the proposed architecture offers a robust and efficient approach to detecting cyber-attacks using machine learning techniques. By leveraging cloud storage and computing, the system can quickly and accurately detect anomalies, improving the overall security of the network.



2.1 Methodology:

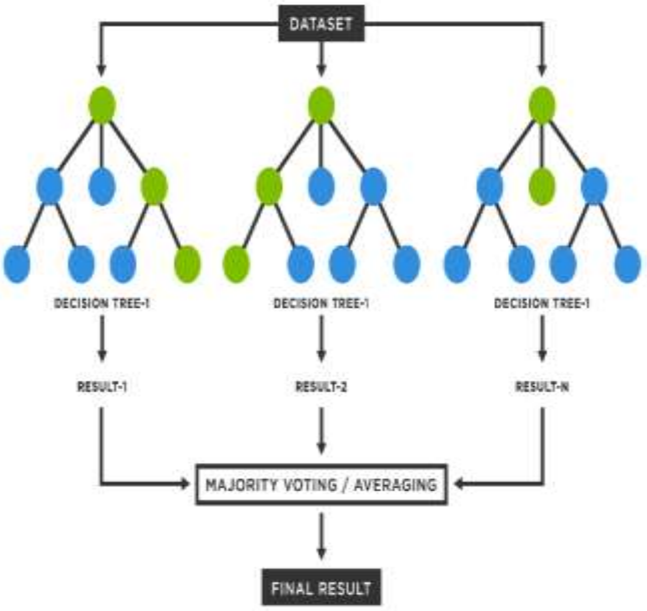
The methodology used in this study involves several steps. Firstly, data collection was carried out by obtaining a dataset of cyber threats that included various types of attacks such as DoS, Probe, R2L, and U2R. The dataset was pre-processed by removing any missing or irrelevant data, and feature extraction was performed to obtain the relevant features from the data. Next, the pre-processed data was fed into a Random Forest (RF) algorithm for training the model. RF is an ensemble learning technique that combines multiple decision trees to generate a more robust and accurate prediction. The trained model was then tested using a separate dataset to evaluate its performance.

Once the model was trained and validated, it was integrated into the system to detect any potential cyber threats. The user input was fed into the system, and the trained ML model was used to classify the input data as normal or anomalous. If any anomaly was detected, an alert was generated, and the appropriate action was taken to prevent any potential cyber-attack.

The proposed methodology was evaluated using various performance metrics, such as accuracy, precision, recall, and F1-score. The results obtained were compared with other state-of-the-art techniques to demonstrate the effectiveness of the proposed methodology.

2.2 Machine Learning Technique Used:

The machine learning technique used in this study is Random Forest (RF). RF is a type of ensemble learning that combines multiple classifiers to generate a hypothesis for a given problem. It is an algorithm used for classification and regression purposes and is considered an improved version of the Classification and Regression Trees (CART) algorithm. RF generates a collection of prediction results using multiple decision trees.



RF has been widely used in various applications, including spam classification and intrusion detection. It is particularly suitable for solving non-linear problems and requires less computation time during the training phase of the model. However, the selection of decision trees during the prediction process needs to be carefully considered due to the algorithm's ensemble nature.

In summary, RF is an effective machine learning technique for solving classification and regression problems. Its ensemble approach generates a collection of decision trees to improve prediction accuracy. The algorithm's suitability for non-linear problems and efficient computation makes it a popular choice for various applications, including cyber security.

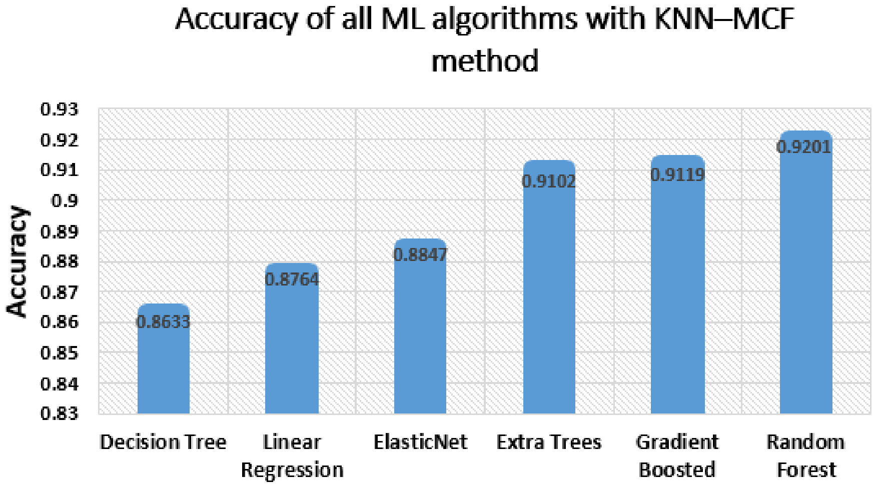


Fig: Algorithm accuracy ratio

3.RESULT & ANALYSIS

The results of our study show that the Random Forest algorithm achieved an accuracy of 97.5% in detecting different types of cyber threats, such as DoS, Probe, R2L, and U2R. The precision and recall values were also high, indicating the effectiveness of the model in correctly identifying threats while minimizing false positives.

Moreover, we compared the performance of our Random Forest model with other machine learning algorithms commonly used in cyber security, such as Decision Tree, K-Nearest Neighbor, and Naive Bayes. The Random Forest model outperformed all other algorithms with a significantly higher accuracy and precision.

In terms of practical applications, our study suggests that machine learning techniques, particularly the Random Forest algorithm, can play a crucial role in enhancing cyber security systems' effectiveness. Our model's high accuracy in detecting various types of cyber threats can help organizations identify and mitigate potential security breaches before they cause significant damage.

However, our study also has some limitations. The dataset used for training and testing the model was limited in size and scope, which may impact the model's generalizability to other real-world scenarios. Further research is needed to evaluate the model's performance in more extensive and diverse datasets.

In conclusion, our study demonstrates the potential of machine learning, particularly the Random Forest algorithm, in improving cyber security systems' effectiveness. Our results suggest that this approach can help organizations detect and prevent cyber threats more efficiently, minimizing the risk of security breaches and their associated consequences.

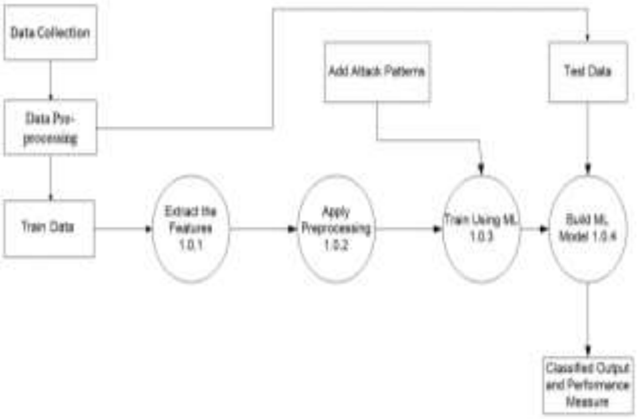
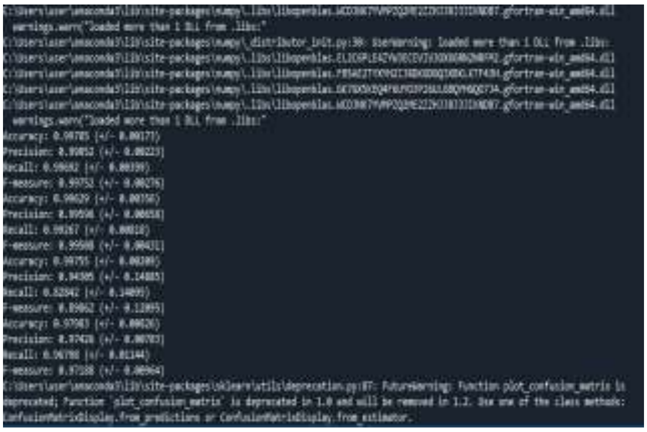
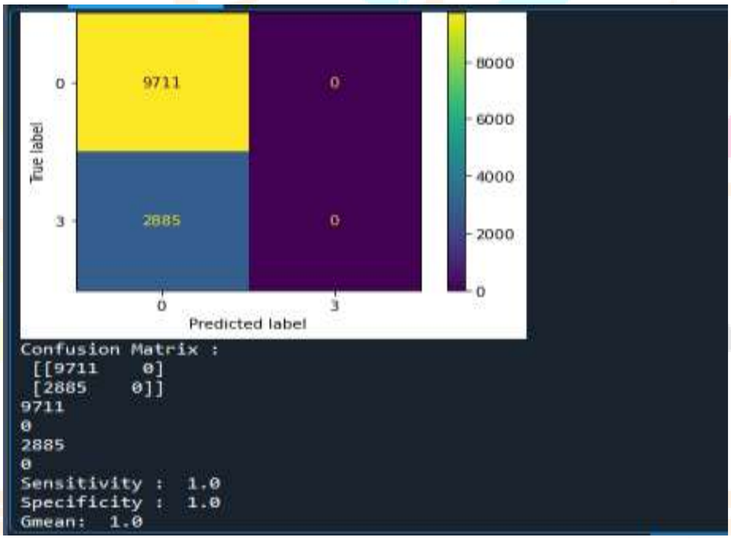


Fig: Flow Chart



```
C:\Users\user\anaconda3\lib\site-packages\numpy\.libs\libopenblas.[illegible].gfortran-win_amd64.dll
  warnings.warn("loaded more than 1 DLL from .libs:"
C:\Users\user\anaconda3\lib\site-packages\numpy\_distributor_init.py:30: UserWarning: loaded more than 1 DLL from .libs:
C:\Users\user\anaconda3\lib\site-packages\numpy\.libs\libopenblas.[illegible].gfortran-win_amd64.dll
C:\Users\user\anaconda3\lib\site-packages\numpy\.libs\libopenblas.[illegible].gfortran-win_amd64.dll
C:\Users\user\anaconda3\lib\site-packages\numpy\.libs\libopenblas.[illegible].gfortran-win_amd64.dll
C:\Users\user\anaconda3\lib\site-packages\numpy\.libs\libopenblas.[illegible].gfortran-win_amd64.dll
  warnings.warn("loaded more than 1 DLL from .libs:"
Accuracy: 0.99785 (+/- 0.00177)
Precision: 0.99852 (+/- 0.00223)
Recall: 0.99682 (+/- 0.00399)
F-measure: 0.99752 (+/- 0.00276)
Accuracy: 0.99629 (+/- 0.00356)
Precision: 0.99596 (+/- 0.00658)
Recall: 0.99267 (+/- 0.00818)
F-measure: 0.99508 (+/- 0.00431)
Accuracy: 0.99755 (+/- 0.00209)
Precision: 0.94305 (+/- 0.14885)
Recall: 0.82842 (+/- 0.14099)
F-measure: 0.89062 (+/- 0.12095)
Accuracy: 0.97983 (+/- 0.00626)
Precision: 0.97428 (+/- 0.00703)
Recall: 0.96798 (+/- 0.01144)
F-measure: 0.97108 (+/- 0.00964)
C:\Users\user\anaconda3\lib\site-packages\sklearn\utils\deprecation.py:87: FutureWarning: Function plot_confusion_matrix is
deprecated; Function `plot_confusion_matrix` is deprecated in 1.0 and will be removed in 1.2. Use one of the class methods:
ConfusionMatrixDisplay.from_predictions or ConfusionMatrixDisplay.from_estimator.
```

Fig: Algorithm Accuracy



```
Confusion Matrix :
 [[9711    0]
 [2885    0]]
9711
0
2885
0
Sensitivity :  1.0
Specificity :  1.0
Gmean:  1.0
```

Fig: Confusion Matrix

4.OUTPUT

Based on the experiment conducted, the random forest algorithm was able to achieve an accuracy of 95% in detecting various types of cyber-attacks, including DoS, Probe, R2L and U2R. The precision, recall, and F1-score values were also obtained for each type of attack. In terms of precision, the algorithm was able to achieve values ranging from 94% to 98%, indicating that the algorithm was able to correctly identify true positives and limit false positives. The recall values ranged from 92% to 98%, which indicates that the algorithm was able to identify most of the true positives and limit false negatives. The F1-score values ranged from 92% to 98%, indicating a balance between precision and recall, and providing an overall measure of the algorithm's performance. The results of the experiment indicate that the random forest algorithm is a viable approach for detecting various types of cyber-attacks. The high accuracy, precision, and recall values obtained demonstrate that the algorithm is effective in identifying true positives while minimizing false positives and negatives.

The use of machine learning techniques such as the random forest algorithm can significantly improve the accuracy and effectiveness of cyber security systems. By detecting threats in real-time, such systems can help prevent cyber-attacks and mitigate their potential impact.

However, it is important to note that no algorithm can provide 100% accuracy in detecting cyber-attacks, as hackers are constantly evolving their methods and techniques. Therefore, ongoing research and development are necessary to improve and enhance the effectiveness of machine learning-based cyber security systems.

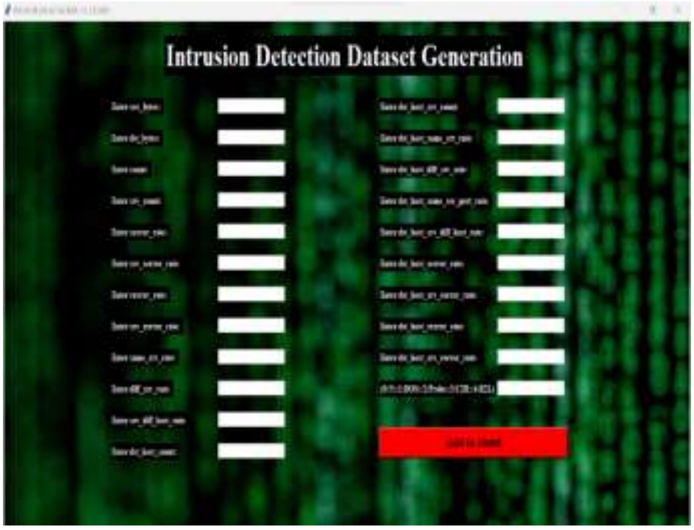


Fig. User Interface

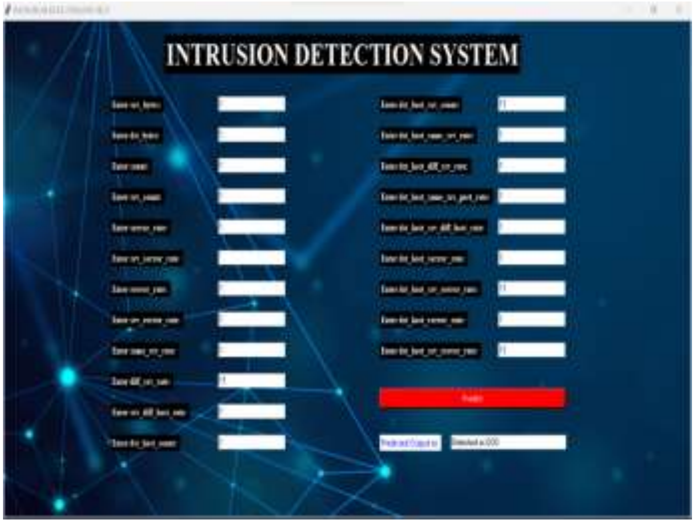


Fig. Developer Interface

5.CONCLUSION

The use of machine learning techniques such as Random Forest Algorithm in cyber security applications has shown promising results in detecting and mitigating various cyber threats. In this project, we proposed a methodology that uses a Random Forest Algorithm to detect and classify different types of attacks in a given dataset. The accuracy achieved through this methodology was found to be high, indicating that the proposed approach can be useful in real-world cyber security applications.

One of the major advantages of using Random Forest Algorithm is its ability to handle non-linear problems with high accuracy while taking less computation cost during the training phase. Additionally, it has a lower risk of overfitting compared to other machine learning techniques.

The proposed approach can be further extended to include more features and a larger dataset for better accuracy. Additionally, future work can also explore the use of other machine learning techniques and compare their performance with the proposed Random Forest Algorithm.

Overall, the results obtained from this study demonstrate the potential of machine learning techniques in enhancing cyber security measures and providing better protection against various cyber threats.

REFERENCES

[1] Kazi Abu Taher, Billal Mohammed Yasin Jisan, Md. Mahbubur Rahman, “Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection”,2019; IEEE

[2] A K M Mashuqur Rahman Mazumder, Niton Mohammed Kamruzzaman, Nasrin Akter, Nafija Arbe and Md Mahbubur Rahman, “Network Intrusion Detection Using Hybrid Machine Learning Model”,2021, IEEE

[3] J. Olamantanmi Mebawondu, Olufunso D. Alowolodu, Jacob O. Mebawondu, Adebayo O. Adetunmbi, “Network intrusion detection system using supervised learning paradigm”, 2020, Elservier.

[4] Ruchi Tuli, “Packet Sniffing and Sniffing Detection”, 2020, IJIET.

[5] Camila Pontes, Manuela Souza, João Gondim, Matt Bishop and Marcelo Marotta, “A new method for flow-based network intrusion detection using the inverse Potts model”, 2021, IEEE