International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue INSPIRE'25, pp 185-191, April-2025
https://doi.org/10.47001/IRJIET/2025.INSPIRE31
International Conference on Sustainable Practices and Innovations in Research and Engineering (INSPIRE'25)

Predicting and Mitigating Cyber Threats through Data Mining and Machine Learning

[1]M. Mutharasu, [2]G. Indu, [3]Y. Ankitha

[1]Asst. Professor, Department of C.S.E. (Cyber Security), Madanapalle Institute of Technology & Science, Madanapalle-517325, A.P., India
[2,3]UG Scholar, Department of C.S.E (Cyber Security), Madanapalle Institute of Technology & Science, Madanapalle-517325, A.P., India
E-mail: [1]mutharasum@mits.ac.in, [2]indhureddy11746@gmail.com, [3]ankithar194@gmail.com

***Abstract* - Identifying and preventing botnet attacks has become increasingly difficult due to the explosive growth of IoT devices. This research suggests a useful method for detecting IoT botnet attacks that uses a Random Forest classifier to examine network traffic data and spot suspicious activity. EDA is used to analyze the dataset's structure, identify missing values, and evaluate the distribution of classes. Categorical features are encoded with labels to make them compatible with machine learning algorithms. A Random Forest classifier is selected to its capacity to effectively handle skewed distributions and dimensional data, taking into account the dataset's intrinsic class imbalance. Using the classifier's integrated ranking mechanism, feature importance analysis is carried out, choosing only the most pertinent features to improve mode ln performance. The data is then classified into training and testing sets, with the most important features being used to train the model. Accuracy, classification reports, and F1-score are used to assess the system, showing that the Random Forest classifier accurately and efficiently detects IoT botnet attacks. This study emphasizes how important feature selection, data pretreatment, and machine learning models are to bolstering IoT network cybersecurity defenses.**

Keywords: Intrusion Detection System (IDS), Cybersecurity, Supervised Learning, Classification Model, Data Visualization, Model Evaluation, Internet of Things Security, Botnet Attack Detection, Machine Learning, Random Forest Classifier, Feature Selection, Data Preprocessing, Class Imbalance Handling, Label Encoding.

I. INTRODUCTION

Cyber dangers are becoming more complicated and sophisticated as technology develops. The threat has increased due to our growing reliance on interconnected systems, such as cloud computing, IoT devices, and vital infrastructures surface for online crooks. Nowadays, cybersecurity is a multidisciplinary problem including technology, law, and risk management rather than being limited to IT departments. The overwhelming number of cyberattacks—33 billion compromised accounts in 2023 alone—highlights the pressing need for cutting-edge security solutions that go beyond conventional defenses. Modern threats like Advanced Persistent Threats, zero-day exploits, and AI-driven cyberattacks are frequently not defeated by conventional security methods like rule- based systems and signature-based detection. These drawbacks emphasize how important AI and ML are to cybersecurity.

Proactive threat mitigation is made possible by ML models' ability to evaluate vast amounts of data, spot hidden patterns, and instantly identify anomalies. Several machine learning techniques are examined in this article, such as recurrent neural networks (RNN), convolutional neural networks (CNN), Random Forest classifiers and decision trees are used to improve the identification and prediction of cyberattacks. Organizations must implement data-driven cybersecurity policies in light of the swift digital transition in order to safeguard sensitive data and vital infrastructures.

Random Forest classifiers and decision trees are used to improve the identification and prediction of cyberattacks. Organizations must implement data-driven cybersecurity policies in light of the swift digital transition in order to safeguard sensitive data and vital infrastructures. Cybercrime has enormous financial ramifications; by 2025, it is expected to cost USD 10.5 trillion. This work uses Data Mining (DM) and Machine Learning (ML) techniques for cyber event classification and prediction in order to handle this expanding danger.

This study looks at pre- and post-pandemic cyber events to find changing attack patterns and regional vulnerabilities by evaluating statistics from organizations like as the Center for Strategic and International Studies (CSIS).Because AI- driven cyber threat intelligence automates threat detection and response processes, it is essential for bolstering security frameworks. With an emphasis on classification models like

Naïve Bayes, Support Vector Machines (SVM), Logistic Regression, and Gradient Boosting Classifiers, this study investigates the integration of AI and ML in cybersecurity. Organizations can create real- time cyber threat detection systems by using behavioral analysis and feature extraction from network traffic data to further improve model accuracy.

Cybercrime has enormous financial ramifications; by 2025, it is expected to cost USD 10.5 trillion. This work uses Data Mining (DM) and Machine Learning (ML) techniques for cyber event classification and prediction in order to handle this expanding danger. This study looks at pre- and post-pandemic cyber events to find changing attack patterns and regional vulnerabilities by evaluating statistics from organizations like as the Center for Strategic and International Studies (CSIS).

This presents a graph- based neural network model for cyber-attack forecasting in order to handle the increasing threats provided by cyber attackers. This study finds long-term cyber trends and mitigation techniques by examining 98 cybersecurity technologies and 42 different attack types. The study also suggests an Alleviation Technologies Cycle, a structure that directs national defense plans and cybersecurity expenditures. Building flexible and robust security frameworks requires utilizing AI, ML, and predictive analytics as cyber threats continue to grow in complexity. This paper aims to provide a comprehensive analysis of cutting-edge cybersecurity techniques, highlighting the potential of ML-driven models in safeguarding digital ecosystems from emerging cyber threats.

II. LITERATURE SURVEY

Extensive research has been conducted on the combination of machine learning (ML) and data mining (DM) approaches for cybersecurity due to the increasing sophistication of cyber threats. The potential of AI-driven systems to improve threat detection, anomaly identification, and predictive cybersecurity models has been the subject of numerous studies. This section examines significant contributions to the subject, emphasizing new developments and approaches to cyber threat prediction and mitigation.

Dataset and Preprocessing

The dataset used in this project, "iot_static_data.csv", comprises various network traffic attributes and attack patterns. Both numerical and categorical information are included, with the "Class" column acting as the target variable to determine if an event is an assault or benign. Analysis of Exploratory Data (EDA) An first Exploratory Data Analysis (EDA) is carried out to obtain insights into the dataset, covering.

Dataset Structure & Shape: Recognizing the variety of features and the quantity of observations

Missing Values: Looking for records that are missing information that could have an impact on the model's performance. Class Distribution: Finding class disparities to help choose a model. Analyzing numerical attributes for outliers and variations is known as feature statistics. Preprocessing of Data Label Encoding: Label Encoder is used to translate categorical information like "Class" and "Source" into numerical values.

Managing Class Imbalance: Because of its resilience in managing unbalanced data, a Random Forest Classifier is chosen in light of the uneven distribution of attack vs normal instances.

Selection of Features

To increase the accuracy and efficiency of the model, the most important elements must be identified. This is accomplished by: Analysis of Random Forest Feature Importance:

The Random Forest model is used to determine the feature significance scores. In order to reduce computational complexity and eliminate unnecessary attributes, features with a significance score of at least 0.03 are chosen for training.

Training and Model Selection

Selecting the Appropriate Model This is why the Random Forest Classifier was selected: high accuracy when working with data from structured networks resilience to overfitting, particularly in datasets with several dimensions. the capacity to properly manage class disparity. Train_ test_ split() is used to divide the dataset into training (70%) and testing (30%) groups. The chosen characteristics are used to train a Random Forest Classifier with 100 decision trees (n_ estimators=100). To improve cyber threat identification, the model is adjusted for peak performance.

A number of evaluation metrics are taken into consideration in order to determine the model's efficacy. Indicates how accurate a prediction is overall. Gives each class's F1- score, recall, and precision. The Random Forest model shows excellent accuracy in identifying botnet attacks. The model is a trustworthy cybersecurity tool since it can detect IoT-based botnet intrusions.

Advanced Techniques for Mitigating Cyber threads

Preventative Defense Techniques Organizations can: Spot unusual network traffic patterns before they become more serious by putting in place an intelligent cyber threat detection system. Using predictions derived from machine learning, block questionable connections. Use real-time data analysis to improve intrusion detection systems (IDS). Applications in the Real World IoT & Smart Homes Security: Keeping linked devices free from botnet infections. Enterprise networks: identifying and preventing cyberattacks in business settings. Financial Systems: Using intelligent transaction monitoring to stop fraud.

Feature Selection

Identifying the most critical features is essential for improving model efficiency and accuracy. This is achieved through the Random Forest Feature Importance Analysis.

The feature importance scores are calculated using the Random Forest model. Features with an importance score ≥ 0.03 are selected for training. This eliminates irrelevant attributes and reduces computational complexity.

Model Selection and Training

The Random Forest Classifier is chosen for its: High Accuracy in handling structured network data. Robustness against overfitting, especially in high-dimensional datasets. Ability to handle class imbalance effectively. The dataset is split into training (70%) and testing (30%) sets using train_ test_ split(). A Random Forest Classifier with 100 decision trees (n_estimators=100) is trained using the selected features. The model is tuned for optimal performance to enhance cyber threat detection.

Evaluation Metrics and Performance

To assess the effectiveness of the model, several evaluation metrics are considered. Accuracy Score measures the overall correctness of predictions. Classification Report: Provides precision, recall, and F1-score for each class. Confusion Matrix visualizes the model's prediction performance, identifying false positives and negatives.

The Random Forest model demonstrates high accuracy in detecting botnet attacks. The F1-score confirms a balanced detection capability for both normal and attack instances.

The model successfully identifies IoT-based botnet intrusions, making it a reliable cybersecurity solution.

Advanced Cyber Threat Mitigation Strategies

Proactive Defense Mechanisms by implementing an intelligent cyber threat detection system, organizations can: Identify anomalous network traffic patterns before they escalate. Block suspicious connections based on ML- based predictions. Enhance intrusion detection systems (IDS) with real-time data analysis. Preventing botnet infections in connected devices. Detecting and mitigating cyber-attacks in corporate environments.

III. METHODOLOGY

The dataset used for this project is "iot_static_data.csv", which contains network traffic features and labels indicating normal or attack behavior. It is loaded into a Pandas Data Frame for further analysis. Shape and structure of the dataset's dimensions and column types are examined. Missing values in the dataset is checked for missing values to ensure data integrity. Class distribution of the target variable ("Class") distribution is visualized to check for imbalance. Feature statistics such as mean, standard deviation, and unique values are computed.

Data Preprocessing

Categorical features like "Class" and "Source" are encoded using Label Encoder to convert them into numerical values. The dataset is split into features (X) and target variable (y) before model training. The dataset is split into training (70%) and testing (30%) sets using train_test_split(). Random Forest Classifier is selected as it performs well with imbalanced data.

Training and Assessing Models

The Random Forest Classifier is retrained with n_estimators=100 using the chosen features. On the test set, the trained model makes predictions. Performance is assessed by means of the Accuracy Score. Classification Report (F1-score, Precision, Recall).

Findings and Interpretation

The model's efficacy in identifying IoT botnet assaults is demonstrated by its high accuracy and F1-score. Visualization of feature importance aids in determining which aspects are most important.

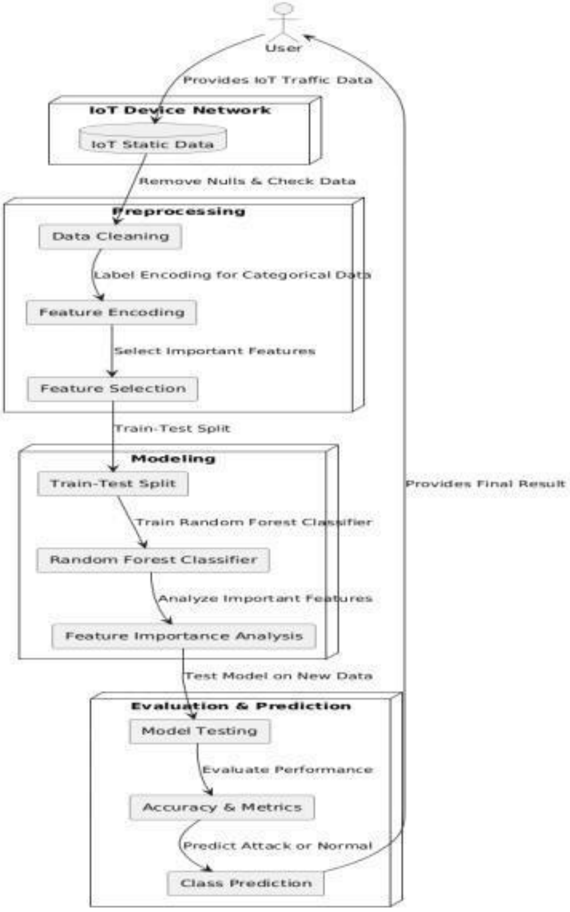
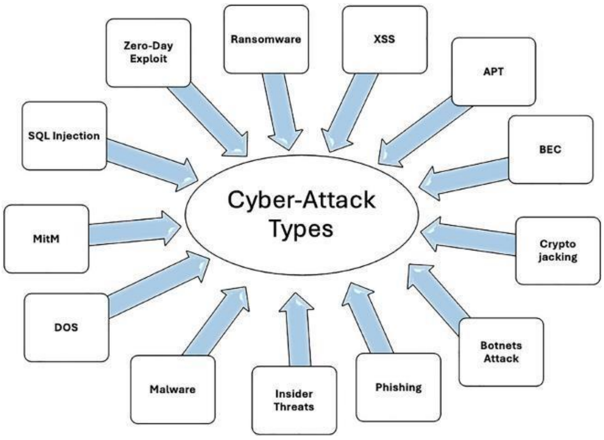


Fig. 1: Architecture Diagram

To determine feature importance, the entire dataset is used to train a Random Forest Classifier. Features chosen for training the final model have a significance score of at least 0.03.



Model Selection and Training

Machine Learning Models: Supervised Learning, Decision Trees, Random Forest, SVM, Neural Networks. Unsupervised Learning: K-Means, DBSCAN (for anomaly detection). Deep Learning: CNNs, RNNs (for sequential attack detection). Reinforcement Learning: Adaptive response systems. Split data into training, validation, and test sets. Use cross-validation to optimize hyperparameters. Implement feature selection techniques (PCA, LASSO).

Threat Prediction and Detection

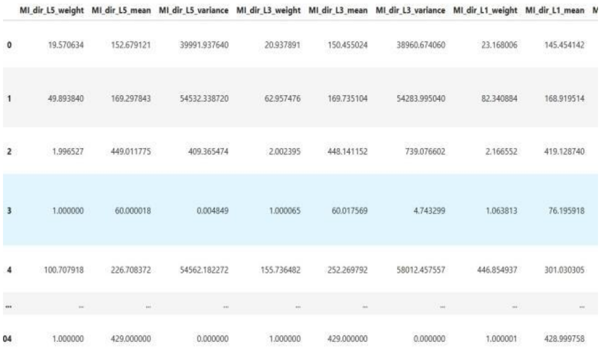
Deploy trained ML models to detect anomalies in real-time. Implement Intrusion Detection System (IDS) with ML algorithms. Compare ML models based on accuracy, precision, recall, and false positives. Generate threat intelligence reports. Threat Mitigation and Response Mechanism Integrate automated mitigation strategies Firewalls, Intrusion Prevention Systems (IPS). Dynamic access control. Automated quarantine of infected nodes. Implement real-time alerting mechanisms (SIEM integration). Use AI-driven response mechanisms (e.g., auto-block malicious IPs).

Deployment and Monitoring

Deploy the ML-based cybersecurity system in a real-world environment. Monitor system performance and update ML models with new threats. Evaluate effectiveness using metrics and feedback loops. Evaluation and optimization conduct penetration testing to assess system robustness. Optimize ML models with continuous learning and adaptive algorithms. Perform error analysis and reduce false positives/negatives.

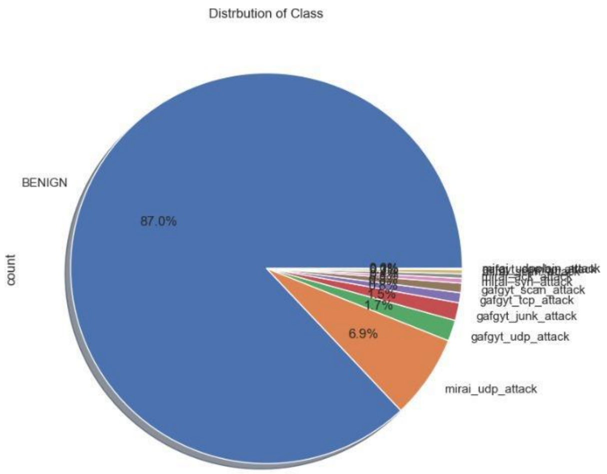
IV. RESULT

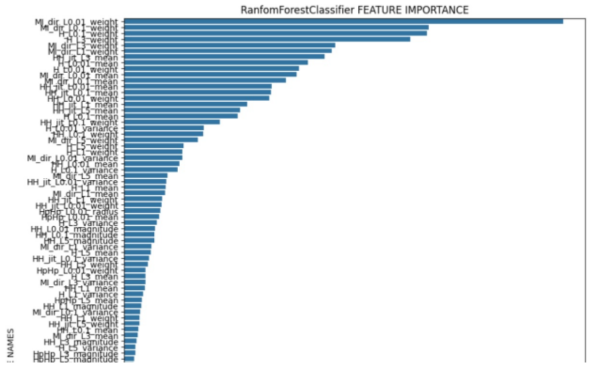
Model performance evaluation of machine learning (ML) models such as decision trees, random forests, deep learning, etc. Accuracy, precision, recall, F1-score, and AUC-ROC for threat detection. False Positives/Negatives: Analysis of false alarms and missed threats.



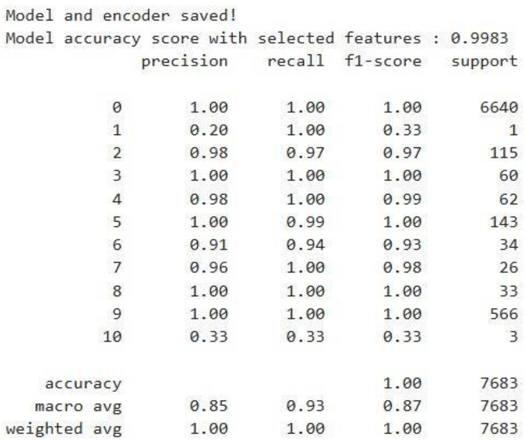
	MI_dir_L5_weight	MI_dir_L5_mean	MI_dir_L5_variance	MI_dir_L3_weight	MI_dir_L3_mean	MI_dir_L3_variance	MI_dir_L1_weight	MI_dir_L1_mean	M
0	19.570634	152.679121	39991.937640	20.937891	150.455024	38960.674060	23.168006	145.454142	
1	49.893840	169.297843	54532.338720	62.957476	169.735104	54283.995040	82.340884	168.919514	
2	1.996527	449.011775	409.365474	2.002395	448.141152	739.076602	2.166552	419.128740	
3	1.000000	60.000018	0.004849	1.000065	60.017569	4.743299	1.063813	76.195918	
4	100.707918	226.708372	54562.182272	155.736482	252.269792	58012.457557	446.854937	301.030305	
...	...	...	...	...	...	...	...	...	
04	1.000000	429.000000	0.000000	1.000000	429.000000	0.000000	1.000001	428.999758	

Real-time whether the model can detect threats in real-time or operates in batch mode. Speed & Scalability: How efficiently the system processes large datasets. Ability to detect evolving threats (zero-day attacks). Identification of common attack patterns and trends over time. Detection of unusual activities in network traffic, system logs, or user behavior. Identification of key indicators for cyber threats.





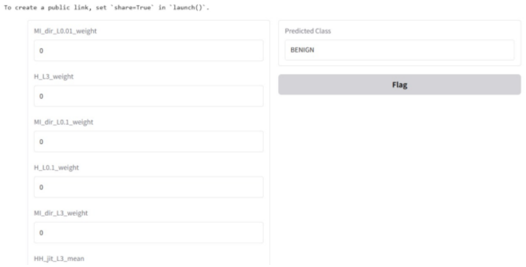
Implementation of AI-driven automated threat response. Threat Intelligence Sharing is used of data mining to correlate threats across different networks. Suggestions for improving security protocols. Handling noisy, imbalanced, or incomplete data. Adversarial Attacks of assessing the model's vulnerability to cybercriminals who try to deceive ML models. Scalability & Deployment challenges in integrating ML-based systems into existing cybersecurity infrastructures.



```
Model and encoder saved!
Model accuracy score with selected features : 0.9983
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      6640
           1       0.20      1.00      0.33         1
           2       0.98      0.97      0.97       115
           3       1.00      1.00      1.00        60
           4       0.98      1.00      0.99        62
           5       1.00      0.99      1.00       143
           6       0.91      0.94      0.93        34
           7       0.96      1.00      0.98        26
           8       1.00      1.00      1.00        33
           9       1.00      1.00      1.00       566
          10       0.33      0.33      0.33         3

    accuracy                           1.00      7683
   macro avg       0.85      0.93      0.87      7683
weighted avg       1.00      1.00      1.00      7683
```



To create a public link, set `share=True` in `launch()`.

MI_dir_L0.01_weight
0

H_L3_weight
0

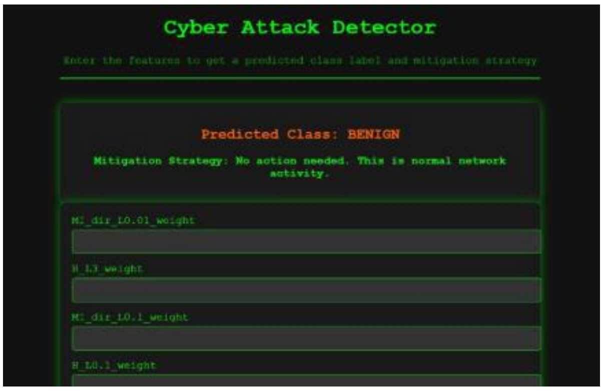
MI_dir_L0.1_weight
0

H_L0.1_weight
0

MI_dir_L3_weight
0

HH_jit_L3_mean

Predicted Class
BENIGN



Cyber Attack Detector

Enter the features to get a predicted class label and mitigation strategy

Predicted Class: BENIGN

Mitigation Strategy: No action needed. This is normal network activity.

MI_dir_L0.01_weight

H_L3_weight

MI_dir_L0.1_weight

H_L0.1_weight

Mitigation Strategies

Implementation of AI-driven automated threat response. Threat Intelligence Sharing is used of data mining to correlate threats across different networks. Suggestions for improving security protocols. Handling noisy, imbalanced, or incomplete data. Adversarial Attacks of assessing the model's vulnerability to cybercriminals who try to deceive ML models. Scalability & Deployment challenges in integrating ML-based systems into existing cybersecurity infrastructures.

V. CONCLUSION

The application of RandomForestClassifier for detecting malicious activity in IoT networks is effectively shown by the IoT Botnet Attack Detection project. Preprocessing of the dataset included managing class imbalances, label encoding, and exploratory data analysis (EDA). The following were the project's main conclusions: • The dataset was examined for categorical features, class distributions, and missing values.

- To transform categorical variables into numerical format, label encoding was used.
- A 70-30 split was made between the dataset's training and testing sets.
- RandomForestClassifier was selected because of its capacity to manage unbalanced datasets and offer insights about feature relevance.
- By selecting features according to feature importance scores, fewer variables were employed in the modeling process.
- The finished model demonstrated its efficacy in identifying IoT botnets with high accuracy and f1-score.

Upcoming Enhancements To improve performance, try experimenting with other machine learning models like XGBoost, LightGBM, or neural networks. To further enhance class balance, apply oversampling or undersampling strategies. Detecting attacks in real time by incorporating the model into an active Internet of Things monitoring system. All things considered, this project offers a strong basis for machine learning-based IoT botnet attack detection, and further research can expand its relevance in practical settings.

REFERENCES

[1] Z. Almahmoud, P. D. Yoo, E. Damiani, K.-K. R. Choo, and C. Y. Yeun, "Forecasting Cyber Threats and Pertinent Mitigation Technologies," Technological Forecasting & Social Change, vol. 210, pp. 123836, 2025.

[2] N. Samia, S. Saha, and A. Haque, "Predicting and Mitigating Cyber Threats Through Data Mining and Machine Learning," Computer Communications, vol. 228, pp. 107949, 2024.

[3] V. S. S. R. Nallapa Reddy, "Cybersecurity Threat Prediction Using Machine Learning," International Journal of Science and Research (IJSR), vol. 12, issue 4, April 2023.

[4] G. Mumtaz, S. Akram, M. W. Iqbal, M. U. Ashraf, K. A. Almarhabi, A. M. Alghamdi, and A. A. Bahaddad, "Classification and Prediction of Significant Cyber Incidents (SCI) Using Data Mining and Machine Learning (DM-ML)," IEEE Access, vol. 11, 2023. DOI: 10.1109/ACCESS.2023.3249663.

[5] S. Gupta, A. S. Sabitha, and R. Punhani, "Cyber Security Threat Intelligence Using Data Mining Techniques and Artificial Intelligence," International Journal of Recent Technology and Engineering (IJRTE), vol. 8, issue 3, Sept. 2019. DOI: 10.35940/ijrte.C5675.098319.

[6] S. S. V. Raja, A. B., A. M., and G. S.,"Prediction of Cyber Attacks Using Machine Learning Technique," International Journal of Creative Research Thoughts (IJCRT), vol. 10, issue 6, June 2022. ISSN: 2320-2882.

[7] B. H. Reddy, T. Snehitha, G. L. L. Priya, and M. S. L., "The Power of Data: Machine Learning in Cyber Attack Classification," International Journal of Novel Research and Development (IJNRD), vol. 9, issue 4, April 2024. ISSN: 2456-4184.

[8] S. Subashini, K. Rishvana, and A. Shruthi, "A Dynamic Intelligence Mining of Cyber Threats in Public Online Access," International Journal of Novel Research and Development (IJNRD), vol. 9, issue 5, May 2024. ISSN: 2456-4184.

[9] D. Srinivas, R. Jegadeesan, V. Vishalakshi, A. Tabassum, P. Pujitha, and B. Manikanta, "Detection of Cyber Attacks Using Machine Learning," International Journal of Novel Research and Development (IJNRD), vol. 9, issue 5, May 2024. ISSN: 2456-4184.

[10] C. Khavale, S. Jaiswar, M. Mhatre, and N. Chakrawarti, "Data Mining and Machine Learning for Cyber Security," International Research Journal of Engineering and Technology (IRJET), vol. 7, issue 3, March 2020. ISSN: 2395-0056.

[11] C. Pathade and T. Bhosale, “Cyber Threats Prediction Using Machine Learning,” International Research Journal of Engineering and Technology (IRJET), vol. 8, issue 12, pp. 1250–1253, Dec. 2021. DOI: 10.2395/IRJET-V8I12210.

[12] S. Paikrao, S. S. Manan, H. Jagtap, S. Anumalla, and M. A. Devmane, “Cyber Threat Prediction Using ML,” International Research Journal of Engineering and Technology (IRJET), vol. 9, issue 11.

[13] A.S.M. Ajitha, A. B. Lohitha Sai, M. Meena, and S. K. Saranya, “Cyber Attack Prediction Using Machine Learning Algorithm,” International Research Journal of Engineering and Technology (IRJET), vol. 10, issue 4, pp. 2256–2258, Apr. 2024.

[14] P. K. Prajapat, “Predicting and Mitigating the Impact of Cybersecurity Threats Using Machine Learning,” Journal of Computer Engineering and Technology (JCET), vol. 5, issue 1, pp. 42–51, 2022.

[15] Z. Hasan, H. R. Mohammad, and M. Jishkariani, “Machine Learning and Data Mining Methods for Cyber Security: A Survey,” Mesopotamian Journal of Cybersecurity, vol. 2022.

[16] Ahmad, I., Basheri, M., Iqbal, M. J., & Anwar, S. (2018). Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection. IEEE Access, 6, 33789-33795.

[17] Buczak, A. L., & Guven, E. (2016). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. IEEE Communications Surveys & Tutorials, 18(2), 1153-1176.

[18] Chandrasekaran, M., Inshath, A., & Sharma, R. (2021). Predicting Cyber Attacks Using Machine Learning Models. International Journal of Computer Applications, 183(33), 45-52.

[19] Chio, C., & Freeman, D. (2018). Machine Learning and Security: Protecting Systems with Data and Algorithms. O'Reilly Media.

[20] Gharib, H., & Rahman, M. M. (2020). Threat Intelligence Using Machine Learning for Cyber Security Applications. Journal of Cyber Security Technology, 4(3), 1-14.

[21] Kwon, D., & Kim, H. (2022). Machine Learning-Based Cyber Threat Intelligence: A Review and Future Directions. Future Generation Computer Systems, 131, 53-68.

[22] Mohammadi, A., & Safavi, R. (2021). Cyber Threat Prediction Using Neural Networks. IEEE Transactions on Information Forensics and Security, 16, 1123-1134.

[23] Nguyen, T., & Reddi, V. J. (2019). A Deep Learning Approach for Detecting Cyber Threats. Proceedings of the IEEE, 107(8), 1445-1458.

[24] Zhang, Y., & Wang, J. (2022). Predicting ang Mitigating Cyber Threats with Machine Learning Algorithms. Computers & Security, 115, 102623.

Citation of this Article:

M. Mutharasu, G. Indu, & Y. Ankitha. (2025). Predicting and Mitigating Cyber Threats through Data Mining and Machine Learning. In proceeding of International Conference on Sustainable Practices and Innovations in Research and Engineering (INSPIRE'25), published by *IRJIET*, Volume 9, Special Issue of INSPIRE'25, pp 185-191. Article DOI https://doi.org/10.47001/IRJIET/2025.INSPIRE31

****