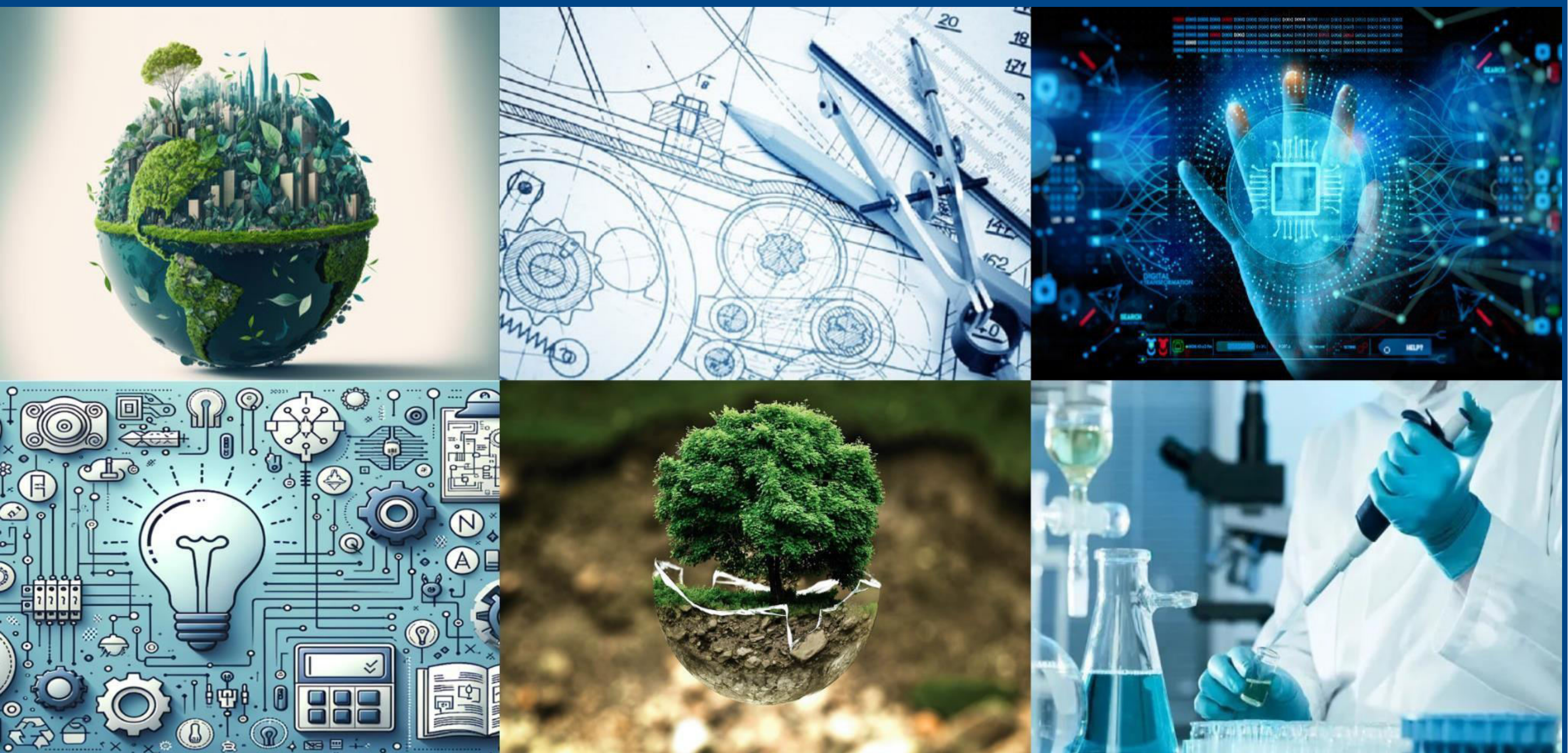
ISSN: 2582-7219

International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206

Volume 8, Issue 10, October 2025

www.ijmrset.com ijmrset@gmail.com +91 6381 907 438 +91 63819 07438

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

A Data Science and Machine Learning Framework for Cyber Attack Prediction and Classification

Sowmya J[1], Pooja S L[2], Anusha L[3]

Assistant Professor, Dept. of Computer Science & Application, The Oxford College of Science, Bangalore, India[1]

PG Students [MCA], Dept. of Computer Applications and Science, The Oxford College of Science, Bangalore, India[2-3]

ABSTRACT: Cybersecurity isn't just a technical concern anymore—it touches all our lives, especially with how connected we are today. The threats we face online have become smarter, faster, and trickier, making it tough for old-school security tools to keep up. For example, systems that rely on set rules or known signatures often miss new and evolving attack methods.

That's why this study shifts gears, exploring how machine learning can help spot and predict different kinds of cyber attacks before they happen. We focused on four major types that cause serious problems: Denial of Service (DoS), Remote to Local (R2L), User to Root (U2R), and various kinds of malware.

By working with a solid dataset—a collection of 40,000 samples of network activity, each packed with 25 features—we tested out four popular machine learning methods: Logistic Regression, Decision Tree, Random Forest, and Support Vector Classifier (SVC). Throughout the process, we paid close attention to getting the data cleaned up, choosing the best features, and using a range of metrics (like accuracy, precision, recall, and more) to judge how well each model performed.Random Forest stood out as the most dependable choice overall, but each algorithm had its own strengths depending on the attack being detected. The findings from this research can help shape smarter, more adaptive cybersecurity tools for the future.

KEYWORDS: Cybersecurity , Machine Learning ,Attack Prediction, Random Forest, Network Security, Predictive Models, Data-driven Security.

I. INTRODUCTION

Technology is advancing fast, and today, businesses, governments, and individuals depend heavily on connected systems to get important work done. But this reliance also opens the door for cyber attackers to cause serious trouble.Because these threats keep evolving, we need smarter and more flexible defenses that can catch not just the usual attacks but also new and unexpected ones.

Traditional security tools like firewalls and signature-based intrusion detection systems often react after the damage is done and focus only on known threats. They work based on fixed rules, which means they struggle to spot sophisticated or brand-new attacks. This is why there's growing excitement around using machine learning and data science in cybersecurity. Machine learning models don't just follow preset rules—they learn from past data, notice subtle behavior patterns, and can even tackle threats they haven't seen before.We focus on four popular models—Logistic Regression, Decision Tree, Random Forest, and Support Vector Classifier—and test how well they detect four major attack types: Denial of Service (DoS), Remote to Local (R2L), User to Root (U2R), and various types of malware. Our aim is not only to check their accuracy but also to understand where each model shines or falls short in real-world situations. By analyzing their performance and the importance of different features, this research hopes to add valuable knowledge to the push for smarter, automated systems that can keep up with today's fast-changing cyber threats.

1.1 Objective

- The goal is to create and test machine learning models that can predict cyber attacks.
- We want to see how different algorithms stack up against each other in terms of performance.
- It's important to identify which key features have the biggest impact on how accurate the predictions are.
- Finally, we aim to offer useful insights that can help improve automated systems for detecting intrusions.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

II. LITERATURE REVIEW

2.1 Cyber Security Attacks Prediction Results: Traditional cybersecurity tools have mostly relied on signature-based methods, but these often can't keep up with new and changing threats. To tackle this, researchers have looked into other options like Bayesian networks and time-series models. Bayesian methods use probabilities to make decisions, while models based on Long Short-Term Memory (LSTM), a type of neural network, are especially good at spotting patterns over time. This makes them effective for detecting attacks like spyware or flooding.

2.2 Machine Learning used in The Cyber Security Prediction:- Supervised learning methods like decision trees, random forests, and support vector machines have shown great promise in identifying and predicting cyber threats. These models learn from labeled data, so they're good at recognizing attack patterns they've seen before. Meanwhile, unsupervised techniques such as clustering are helpful for spotting new or zero-day attacks that haven't been labeled yet.

2.3 Research Agenda:- Many existing studies don't thoroughly evaluate multiple types of attacks or compare a range of algorithms. On top of that, there's often a shortage of consistent ways to measure performance and limited access to diverse real-world data. This research addresses these gaps by using a solid experimental design and performing a detailed comparison to provide clearer insights.

Process Of Machine Learning

III. METHODOLOGY

3.1 Datasets Used:- The dataset we used in this study includes 40,000 samples, each with 25 different features that capture details like network activity, protocol types, packet sizes, timing, and signs of threats. It focuses on four main types of attacks:

- DoS (Denial of Service), which tries to overwhelm system resources.
- R2L (Remote to Local), where attackers attempt to access a system from afar.
- U2R (User to Root), involving attacks that escalate privileges from a regular user to the root level.
- Malware, which is malicious software created to damage or disrupt systems.

3.2 Machine Learning Algorithms Used

- **Logistic Regression** is a simple model that predicts the chance of something happening. It's a good starting point for figuring out classifications.
- **Decision Tree** works like a flowchart, splitting data step-by-step to make decisions. It's easy to follow and understand how it reaches conclusions.
- **Random Forest** combines lots of decision trees to make better and more reliable predictions. It's stronger and less likely to make mistakes than just one tree.
- **Support Vector Classifier** finds the best line or boundary that separates different groups in the data. It's great for handling complex and detailed datasets.

3.3 Performance Measures:- We checked how well each model did by looking at several key measures:

- **Accuracy,** which tells us how often the model got things right overall.
- **Precision,** showing how many of the positive predictions were actually correct.
- **Recall (also called Sensitivity),** which measures how well the model caught actual positives.
- **F1-Score,** a balance between precision and recall to give an overall effectiveness score.
- **Specificity,** indicating how well the model identified the negatives.
- **Confusion Matrix,** which gives a full picture of true and false predictions.

© 2025 IJMRSET | Volume 8, Issue 10, October 2025| DOI:10.15680/IJMRSET.2025.0810008

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

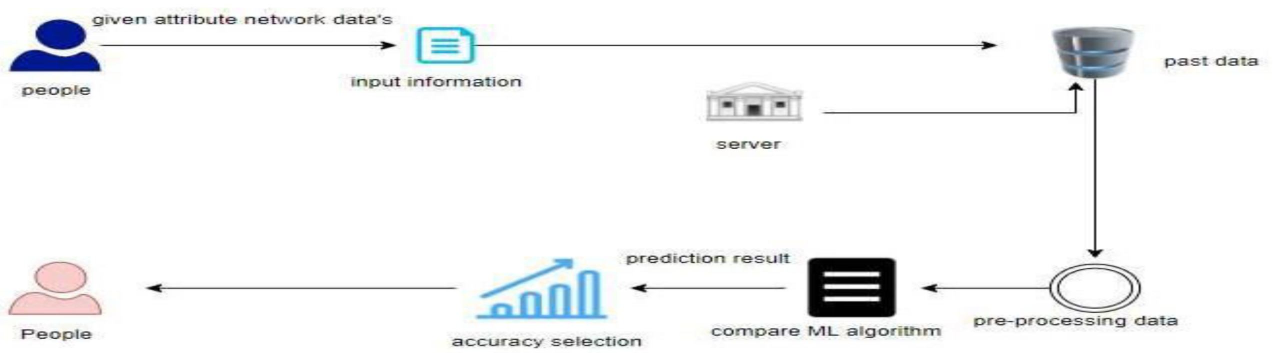
3.4 Implementation Tools Used:- The models were built using Python with popular libraries like scikit-learn, pandas, and numpy, all run in Jupyter Notebook on a Windows system. We gathered our datasets from Kaggle and used both frontend and backend tools—Flask for the backend and React.js for the frontend—to create an interactive interface. The results were displayed using clear, easy-to-understand graphs. To make sure our models performed well and consistently, we applied Hyper parameter tuning and cross-validation techniques.

Timestamp	Source IP Addr	Destination	Source Po	Destinatic	Protocol	Packet Le	Packet Ty	Traffic Typ	Payload D	Malware I	Anomaly S	Alerts/Wa	Attack Typ
30-05-2023 06:33	103.216.15.12	84.9.164.252	31225	17616	ICMP	503	Data	HTTP	Qui natus	IoC Detect	28.67		Malware
26-08-2020 07:08	78.199.217.198	66.191.137.1	17245	48166	ICMP	1174	Data	HTTP	Aperiam	IoC Detect	51.5		Malware
13-11-2022 08:23	63.79.210.48	198.219.82.1	16811	53600	UDP	306	Control	HTTP	Perferend	IoC Detect	87.42	Alert Trigg	DDoS
02-07-2023 10:38	163.42.196.10	101.228.192.	20018	32534	UDP	385	Data	HTTP	Totam maxime beat		15.79	Alert Trigg	Malware
16-07-2023 13:11	71.166.185.76	189.243.174.	6131	26646	TCP	1462	Data	DNS	Odit		0.52	Alert Trigg	DDoS
28-10-2022 13:14	198.102.5.160	147.190.155.	17430	52805	UDP	1423	Data	HTTP	Repellat quas illum		5.76		Malware
16-05-2022 17:55	97.253.103.59	77.16.101.53	26562	17416	TCP	379	Data	DNS	Qui		31.55		DDoS
12-02-2023 07:13	11.48.99.245	178.157.14.1	34489	20396	ICMP	1022	Data	DNS	Amet	IoC Detect	54.05	Alert Trigg	Intrusion
27-06-2023 11:02	49.32.208.167	72.202.237.9	56296	20857	TCP	1281	Control	FTP	Veritatis r	IoC Detect	56.34	Alert Trigg	Intrusion
15-08-2021 22:29	114.109.149.11	160.88.194.1	37918	50039	UDP	224	Data	HTTP	Consequatur ipsum		16.51	Alert Trigg	Malware
20-07-2022 13:28	177.21.83.200	196.218.124.	35538	35006	ICMP	661	Data	HTTP	Sequi		24.91	Alert Trigg	Malware
26-06-2022 15:15	92.4.25.171	112.43.185.2	10903	36817	TCP	281	Control	HTTP	Nihil praesentium as		86.07		Malware
30-09-2020 21:35	57.91.207.84	98.96.110.38	53471	38048	ICMP	64	Control	DNS	Earum sit est et eaqu		74.2	Alert Trigg	Intrusion

Fig 1. Datasets which have used for Comparison

3.5 System Architecture

Fig.2 System Architecture Diagram



The diagram shows how the system works step by step. People start by providing network data, which gets sent in as input information. This input goes to a server, which also has access to previous, stored data. The system prepares and cleans the data (pre-processing) before running different machine learning algorithms to compare how well they perform. Once the models make their predictions, the system picks out the most accurate results and shares them back with people in a way that's easy to understand. The whole process is designed to help people see which model works best for predicting new cyber threats, using helpful graphs and clear comparisons.

3.6 Workflow Diagram:- The diagram lays out a simple step-by-step process for using machine learning to catch network attacks. It starts with the raw source data, which first goes through a cleaning and processing stage to make sure it's ready for analysis. The prepped data is then split in two: one part is used to train machine learning models, and the other part is set aside for testing them. After the models are trained, their accuracy is evaluated using the testing data, and the model that performs best is chosen. This best model is then used to spot and identify suspicious network activity, helping to detect potential attacks more reliably.

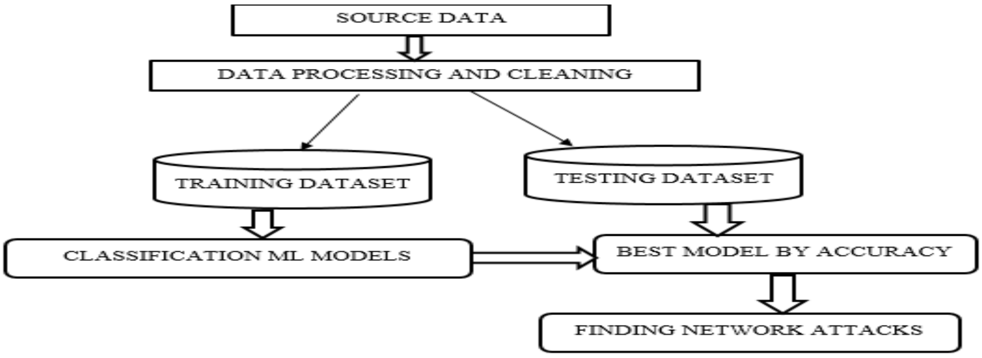


Fig.3 WorkFlow Diagram

IV. RESULTS

4.1 Overall Results:- Among all the models we tested, Random Forest consistently gave the best results across most of the evaluation measures. Although Logistic Regression is easier to understand and interpret, its accuracy wasn't as high as some of the more complex models like Random Forest.

4.2 Attack wise Results

- Denial of Service (DoS) attacks are designed to make systems or resources unavailable to legitimate users by overwhelming them with excessive requests or exploiting vulnerabilities. These attacks can target features, bugs, or misconfigurations and disrupt normal operations on the Internet. Distributed Denial of Service (DDoS) attacks are more severe, using thousands of sources to flood a victim, making mitigation challenging. In this study, DoS and DDoS attacks were simulated, with Random Forest models achieving an accuracy of 94.2% in detection. The results confirm that machine learning can be highly effective in identifying and mitigating denial of service threats in a technical environment.
- Securing digital communications is a significant challenge, as sophisticated attacks like Remote to User (R2L) pose real threats to organizations worldwide. R2L attacks occur when an external attacker exploits system vulnerabilities to gain illegal, local user access, often using social engineering or crafted network packets. These attacks are subtle and hard to detect, making them a persistent risk for internet-connected systems. In this study, R2L attack detection proved difficult, with machine learning accuracy dropping to 85.7%, highlighting the attack's deceptive nature. This result underscores the ongoing need to improve detection strategies, as effective recognition of R2L intrusions is vital for robust cybersecurity defense.
- User to Root (U2R) attacks involve escalating privileges from a normal user to root access by exploiting system vulnerabilities, such as buffer overflow. Attackers start with legitimate access but probe for weaknesses to gain full control, often targeting passwords or bugs. Detecting U2R attacks is challenging due to their sophisticated and stealthy nature. In this study, the Random Forest algorithm achieved a detection accuracy of 78.9%, the lowest among attack types tested, emphasizing the difficulty in spotting these threats. This highlights the need for more advanced techniques to effectively detect and prevent privilege escalation attacks
- All models did a good job detecting Malware, with Random Forest scoring a strong 92.1%.

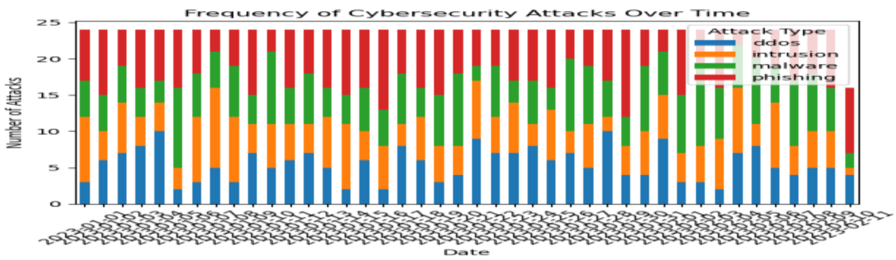


Fig.3 Overall Results of the Attacks

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

4.3 Final Result Comparison:-While Random Forest gave us strong and accurate results, it required more computational power and resources to run. On the other hand, Decision Tree offered a good balance between speed and accuracy, making it a better choice when working in environments with limited computing capabilities.

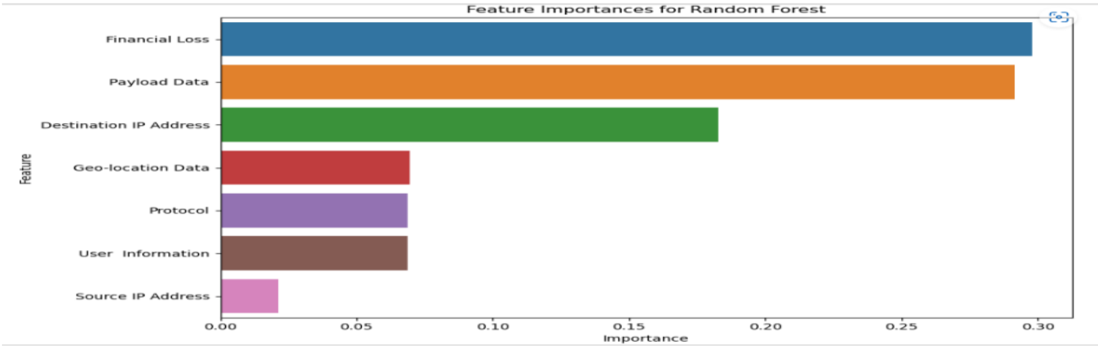


Fig.4 Results of features comparison in dataset using random forest algorithm

V. ANALYSIS

5.1 Practical Outcomes

- **Algorithm Suitability:** Random Forest is ideal for general use, while Decision Trees are better for real-time applications.
- **Feature Design:** Focusing on traffic and protocol details helps improve prediction accuracy.
- **Attack Specialization:** Tailoring models to specific types of attacks can boost results.
- **Resource Considerations:** Decision Trees require less computing power, making them a good fit for systems with limited resources.

5.2 Limitations

- Using just one dataset can limit how well the model works on different kinds of data or in new situations.
- Models may struggle to keep up with changes in cyber threats unless they are regularly retrained with fresh data.
- It's important to reduce false alarms (false positives) so the system can be trusted and effectively used in real-life settings.
- The complexity of cybersecurity systems can require specialized skills and ongoing maintenance, which can be challenging for some organizations.
- Rapidly evolving cyber threats demand continuous updates and adaptations to security measures, making constant vigilance necessary.

5.3 Future Work

- The network sector aims to automate the detection of packet transfer attacks in real-time based on connection details.
- The prediction results should be displayed through a web or desktop application for easier user interaction.
- The system needs to be optimized for implementation within an Artificial Intelligence environment.
- Investigate ensemble and hybrid machine learning models to improve detection accuracy.
- Explore deep learning techniques to extract richer features from network data.
- Implement real-time detection systems that are robust against adversarial inputs to ensure reliability and security.

5.4 Understand ability and Justification of AI Predictions in Cyber security

Machine learning models like Random Forest and Support Vector Classifier are great at predicting cyber threats, but they can be complex and hard to understand when it comes to how they make decisions. Explaining these predictions—known as understand ability—is especially important in cybersecurity, where trust and clarity are vital for using these tools effectively. Simpler models like Logistic Regression and Decision Trees are easier to interpret because they clearly show which features matter and how decisions are made. However, these simpler models sometimes aren't as accurate as the more advanced ones.

To address this, new explainable AI methods such as SHAP and LIME have been developed. These tools help explain the predictions of complex models by showing which factors influenced each decision and exposing any biases or unusual patterns. By making cybersecurity models more transparent, analysts can trust the results better, troubleshoot issues more easily, and meet regulatory standards more effectively.

Going forward, research should focus on creating machine learning models that are not only powerful but also easy to interpret, bridging the gap between high accuracy and operational transparency. This will help cybersecurity teams better understand, trust, and improve their AI-driven defenses.

VI. CONCLUSION

This research highlights the strong potential of machine learning, especially Random Forest, for predicting cyber attacks with high accuracy across various attack types, making it a valuable tool in cybersecurity systems. However, different algorithms offer unique benefits, so each may be better suited for certain applications. The study underscores the importance of careful data preprocessing, building models tailored to specific attacks, and continuous evaluation to keep up with rapidly evolving threats. Future research should focus on using more diverse datasets and advanced learning techniques to improve how well models adapt and withstand new challenges.

The analysis started with thorough data cleaning and processing, handling missing values, and exploratory data analysis before moving on to model building and evaluation. The best model was chosen by comparing accuracy scores on public test sets, focusing on how well each algorithm performed for different types of network attacks in predicting future threats. This approach provides valuable insights into diagnosing attacks in new network connections. The goal is to create an AI-powered prediction model that surpasses human accuracy and offers early detection capabilities. Overall, this work shows that machine learning techniques can help network sectors speed up attack diagnosis and reduce human errors, making cybersecurity more efficient and reliable.

REFERENCES

[1]"A Comprehensive Review on Detection of Cyber-Attacks" (2023) — Reviews modern machine learning techniques for cyberattack detection and classification. [ScienceDirect]
[2] "Advanced Machine Learning Algorithm for Cyber Attack Prediction and Prevention" (2024) — Proposes a hybrid ML model focusing on real-time detection and minimizing false positives. [IJISAE Journal]
[3] "Evaluating Deep Learning Variants for Cyber Attacks in IoT Environments" (2024) — Analyzes deep learning approaches for cyber threat detection specifically in IoT networks. [PMC]
[4] "Machine Learning Algorithms for Cyber Attack Detection" (2023) — Examines multiple classification algorithms applied to cyberattack datasets for prediction accuracy. [ACM Digital Library]
[5] "Application of Classification Algorithms of Machine Learning for Cybercrime Detection" (2022–2024) — Discusses the role of ML classifiers in detecting various types of cybercrime. [ScienceDirect]

INNO SPACE

SJIF Scientific Journal Impact Factor

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com