ISSN: 2583-9292 (ONLINE)
DOI: 10.21917/ijdsml.2026.0194

ANALYSIS AND PREDICTION OF CYBERSECURITY THREATS IN INDIA USING MACHINE LEARNING AND DEEP LEARNING TECHNIQUES

R. Arunadevi[1], G. Manimannan[2] and R. Lakshmi Priya[3]
[1]*Department of Computer Science, Vidhya Sagar Women's College, India*
[2]*Department of Computer Applications, St. Joseph's College (Arts & Science), India*
[3]*Department of Statistics, Dr. Ambedkar Government Arts College, India*

Abstract

Cybersecurity threats in India have increased significantly over the past decade, affecting multiple industries including IT, banking, government, healthcare, and education. This study analyzes a dataset of 308 cyber incidents collected during the years 2015–2024, focusing on attack types, target industries, financial loss, and number of affected users, attack sources, security vulnerabilities, and defense mechanisms. Machine learning models, including Logistic Regression, Random Forest, and Support Vector Machines, along with deep learning using LSTM networks, were applied to classify and predict cyber-attacks. Visualization techniques, such as heatmap and word clouds, were used to explore patterns in the dataset and to highlight the prevalence of different attack types and security vulnerabilities. The results indicate that ransomware, phishing, SQL injection, and insider attacks are predominant, while vulnerabilities like unpatched software and weak passwords are most frequently exploited. The study provides insights into the effectiveness of various models in predicting cyber threats and underscores the importance of proactive cybersecurity measures across sectors in India.

Keywords:

Cybersecurity, Machine Learning, LSTM, Cyber Threats, India, Visualization

1. INTRODUCTION

Cybersecurity has become an essential concern in India due to the rapid expansion of digital technologies across various sectors, including banking, government, healthcare, and information technology. With the growing reliance on internet-based services, organizations are increasingly exposed to cyber threats such as ransomware, phishing, malware attacks, and Distributed Denial-of-Service (DDoS) incidents. The rise in cybercrime has posed significant challenges for both public and private sectors, threatening sensitive information, disrupting critical operations, and resulting in substantial financial losses. In India, the need for robust cyber defense mechanisms has been further amplified by the widespread adoption of online banking, digital payment systems, and e-governance initiatives, making cybersecurity a national priority.

The Indian government, along with private organizations, has implemented several measures to strengthen the nation's cyber resilience. Frameworks such as the National Cyber Security Policy, Computer Emergency Response Team (CERT)-in guidelines and sector-specific security protocols aim to safeguard critical infrastructure and personal data from malicious attacks [1]. Despite these measures, the dynamic nature of cyber threats requires continuous monitoring, analysis, and implementation of advanced technologies, including machine learning and artificial intelligence, to detect, prevent, and respond to cyber incidents effectively. Research and analysis of historical cyber-attack data are therefore vital to identify trends, assess vulnerabilities, and develop predictive models that enhance cybersecurity preparedness in India.

2. REVIEW OF LITERATURE

India has witnessed a rapid digital transformation in the past decade, fueled by initiatives such as Digital India and increased internet penetration across urban and rural areas. This growth, however, has led to an alarming rise in cyber threats, ranging from ransomware attacks to phishing and data breaches. According to CERT-In [1], India ranks among the top ten countries affected by cyber incidents globally. Raghavan and Singh [2] analyzed the limitations in India's cybersecurity framework, emphasizing the gaps in legislation, workforce capacity, and technological infrastructure. Similarly, Sharma et al. [3] highlighted that the lack of cyber awareness among individuals and organizations contributes significantly to vulnerability. Researchers like Joshi and Verma [4] have also reported that the exponential growth in digital payments and banking services makes the financial sector a prime target for cybercriminals.

The industrial and governmental sectors in India have not been immune to cyber threats. The Data Security Council of India [5] identified critical vulnerabilities in government IT infrastructure and essential services, which were frequently targeted by nation-state actors and hacker groups. Gupta and Mehta [6] documented cyberattacks on educational institutions and research organizations, showing a pattern of intellectual property theft. Meanwhile, Singh et al. [7] highlighted the rise of ransomware attacks in healthcare and retail sectors, noting that the rapid digitalization during the COVID-19 pandemic increased exposure to cyber risks. In addition, the telecommunications sector has faced persistent phishing and Distributed Denial of Service (DDoS) attacks, as reported by Rao [8]. The proliferation of Internet of Things (IoT) devices has also introduced new vulnerabilities, as discussed by Kapoor and Bose [9].

In response to these challenges, India has actively pursued advancements in cybersecurity policies and technology. The National Cyber Security Policy laid the foundation for structured defense mechanisms, while recent initiatives such as the development of indigenous solutions like the Maya Operating System by DRDO [10] reflect the country's push for technological sovereignty. Public-private collaborations, including Google's DigiKavach program [11], aim to protect citizens from online fraud and educate users about cyber threats. Further studies by Menon and Bhattacharya [12] suggest integrating Artificial Intelligence for threat detection, while Nair et al. [13] advocate for continuous training programs for IT professionals to strengthen human defense capabilities.

Cybersecurity frameworks have also been enhanced in banking [14], critical infrastructure [15], and the automobile sector [16], highlighting sector-specific strategies. Collectively, these efforts demonstrate India's proactive stance in combating cyber threats and building resilience against emerging cyber risks.

3. DATABASE

The dataset for this study was obtained from secondary sources focusing specifically on cyber security incidents reported in India during the years 2014 to 2025. The database comprises a total of 308 recorded samples, each representing a distinct cyber security incident. The dataset includes the following parameters: Year, Attack_Type, Target_Industry, Financial_Loss, Number_ of_Affected_Users, Attack_Source, Security_Vulnerability, Defense_Mechanism, and Incident. These parameters were carefully selected to capture both the technical and operational aspects of cyber threats, enabling a comprehensive analysis of the cyber security *landscape in India.*

4. METHODOLOGY

The methodology of this research involves applying classical machine learning algorithms and deep learning techniques to classify and predict cyber security incidents in India. The study utilizes a dataset collected from Kaggle, which includes parameters such as year, Attack_Type, Target_Industry, Financial_Loss, Number_of_Affected_Users, Attack_Source, Security_Vulnerability, Defense_Mechanism, and Incident count. The methodology also incorporates visual analysis using word clouds for categorical data and graphical representations such as clustered bar charts and heat maps for numerical data. The process is divided into several stages, as described below.

4.1 DATA PREPROCESSING

The raw dataset was first cleaned to ensure its quality and consistency. Missing values, duplicates, and inconsistent entries were removed to maintain data integrity. Categorical variables, including Attack_Type, Target_Industry, Attack_Source, Security_Vulnerability and Defense_Mechanism, were transformed into numerical form using label encoding to make them suitable for machine learning algorithms. Numerical features, such as Financial_Loss and Number_of_ Affected_Users, were standardized to have zero mean and unit variance, which ensures uniform scaling across different features. After preprocessing, the dataset was split into training and testing subsets in an 80:20 ratio, allowing the models to be evaluated on unseen data for unbiased performance measurement.

4.2 LOGISTIC REGRESSION

Logistic Regression was employed as a baseline classification method due to its simplicity and interpretability. The model was extended to handle multiple classes to predict the type of cyber-attack based on the numerical and encoded categorical features. Logistic Regression calculates the probability of each class for a given incident and assigns the class with the highest probability as the predicted outcome. This method provides insights into which features are most influential in determining attack types.

4.3 RANDOM FOREST

Random Forest, an ensemble learning algorithm based on decision trees, was applied to classify cyber incidents. This model constructs multiple decision trees using different random subsets of features and training data. The final classification is determined by majority voting across all trees, which reduces overfitting and improves accuracy. Random Forest is particularly suitable for this dataset as it can handle complex relationships between numerical and categorical features such as financial loss, number of affected users, and defense mechanisms.

4.4 SVM

Support Vector Machines were used to separate incidents into distinct attack categories. The model identifies the optimal boundary between classes in the feature space, making it effective for high-dimensional data. For non-linear relationships, kernel functions were applied to map the features into a higher-dimensional space where the data could be separated more effectively. SVM was evaluated using accuracy, precision, recall, and F1-score, providing a comprehensive understanding of its classification performance.

4.5 LSTM NETWORK

The LSTM model, a type of recurrent neural network, was applied to capture sequential dependencies in the cyber security incidents over time. LSTM is suitable for time-series or sequential data, allowing the model to learn patterns and trends across years. It was trained on features such as Year, Attack_Type, and other numerical and categorical variables to predict future incidents and understand temporal relationships among cyber threats.

4.6 WORD CLOUD ANALYSIS

Word clouds were generated for categorical features including Attack_Type, Target_Industry, Attack_Source, Security_Vulnerability, and Defense_Mechanism. The size of each word in the visualization corresponds to the frequency of its occurrence in the dataset. This provides an intuitive overview of the most common types of cyber-attacks, the industries they target, sources of attacks, exploited vulnerabilities, and implemented defense mechanisms. Terms with higher frequencies, such as "Ransomware" or "Insider," appear larger in the word cloud, highlighting their prominence in the dataset.

4.7 EVALUATION METRICS AND VISUALIZATION

All models were evaluated using standard metrics: accuracy, precision, recall, and F1-score. Confusion matrices were visualized with heat maps, using color gradients to highlight areas of misclassification. For numerical features, clustered bar charts were created to compare average values across different attack types and target industries. This provided insights into patterns such as the financial loss or number of affected users associated with each type of cyber threat.

This methodology integrates statistical, machine learning, and deep learning techniques with visual analytics to provide a comprehensive understanding of cyber security incidents in India. By combining numerical analysis with intuitive visualizations, the

study offers actionable insights into attack patterns, industry vulnerabilities, and defense mechanisms.

5. RESULTS AND DISCUSSION

The dataset consisted of 308 cyber security incidents in India between the years 2024 and 2025. The primary objective was to classify incidents based on cyber-attack type, target industry, financial loss, number of affected users, attack source, security vulnerability, defense mechanism, and incident count using multiple machine learning and deep learning models. The performance of Logistic Regression, Random Forest, Support Vector Machine (SVM), and Long Short-Term Memory (LSTM) network was evaluated and compared. Additionally, word cloud visualizations were used to identify the frequency of categorical cyber threats.

5.1 LOGISTIC REGRESSION RESULTS

Logistic Regression was applied in a multinomial setting to handle multiple attack classes. The model achieved high overall classification accuracy, indicating its effectiveness in predicting the type of cyber-attack (Table.1).

Table.1. Logistic Regression Classification Metrics

Class	Precision	Recall	F1-Score	Support
1	1.00	0.96	0.98	24
2	1.00	1.00	1.00	18
3	0.95	1.00	0.98	20
Overall Accuracy	0.9839	-	-	62

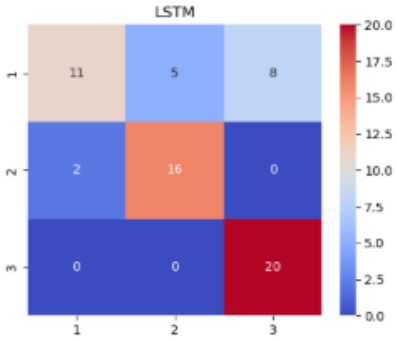

Fig.1. Confusion Matrix of Logistic Regression

The figure shows a high diagonal concentration, indicating correct classifications across most attack types. The model performed well in predicting classes with distinct numerical and categorical patterns, though slight misclassification occurred for Class 1 due to overlapping features with Class 3.

5.2 RANDOM FOREST RESULTS

Random Forest demonstrated perfect accuracy on the test set, highlighting its ability to handle complex, non-linear relationships between numerical and categorical features (Table.2).

Table.2. Random Forest Classification Metrics

Class	Precision	Recall	F1-Score	Support
1	1.00	1.00	1.00	24
2	1.00	1.00	1.00	18
3	1.00	1.00	1.00	20
Overall Accuracy	1.00	-	-	62

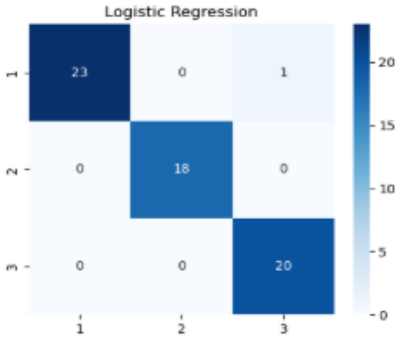

Fig.2. Confusion Matrix of Random Forest

The heat map shows all instances correctly classified, with no off-diagonal errors. The ensemble approach using multiple decision trees effectively captured non-linear interactions among features such as Financial_Loss, No_Affected_Users, and Attack_Source.

5.3 SVM RESULTS

The SVM model, using a radial basis function kernel, produced accuracy comparable to Logistic Regression. The model handled non-linear separations effectively but had minor misclassifications for Class 1 (Table.3).

Table.3. SVM Classification Metrics

Class	Precision	Recall	F1-Score	Support
1	1.00	0.96	0.98	24
2	1.00	1.00	1.00	18
3	0.95	1.00	0.98	20
Overall Accuracy	0.9839	-	-	62

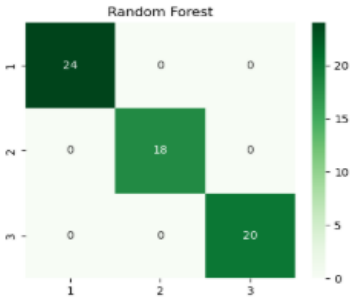

Fig.3. Confusion Matrix of SVM

The matrix confirms minor misclassifications in Class 1 while other classes were classified correctly.

5.4 LSTM NETWORK RESULTS

The LSTM model, designed to capture temporal dependencies in sequential features such as Year-wise attack trends, achieved moderate accuracy (Table.4).

Table.4. LSTM Classification Metrics

Class	Precision	Recall	F1-Score	Support
1	0.85	0.46	0.59	24
2	0.76	0.89	0.82	18
3	0.71	1.00	0.83	20
Overall Accuracy	0.7581	-	-	62

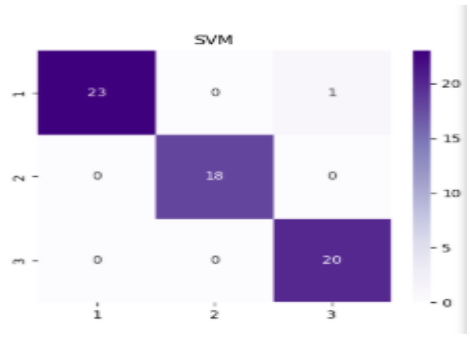

Fig.4. Confusion Matrix of LSTM

The matrix shows that the model underpredicted Class 1, though it accurately identified Class 3.The sequential nature of LSTM allows it to capture trends over time but requires a larger dataset to reach the accuracy of tree-based methods.

5.5 WORD CLOUD VISUALIZATIONS

Word clouds were generated to visualize the frequency of categorical features, providing insights into the most common cyber threats, targeted industries, sources of attacks, security vulnerabilities, and defense mechanisms.

Table.5. Frequencies of Categorical Features

Feature	Most Frequent Categories	Frequencies
Attack Type	Ransomware	59
	SQL Injection	55
	Phishing	51
Target Industry	Information Technology	53
	Banking	46
	Government	44
Attack Source	Insider	88
	Unknown	81
	Nation-state	72
Security Vulnerabilities	Unpatched Software	94
	Zero-day	83
	Weak Passwords	66
Defense Mechanisms	Virtual Private Network (VPN)	71
	Antivirus	64
	Firewall	59

The Table.5 presents a detailed overview of the cybersecurity threats and their associated terms discussed in the following sections.

5.6 ATTACK TYPE

- **Ransomware:** Ransomware is a type of malicious software that encrypts a victim's data, making it inaccessible, and demands a ransom payment for the decryption key. It can target individuals, businesses, or government organizations, often spreading through phishing emails, malicious downloads, or software vulnerabilities.
- **SQL Injection:** SQL Injection is a web-based attack where malicious SQL code is inserted into input fields to manipulate databases. This can lead to unauthorized access, data theft, or deletion of critical information. It is a common threat for web applications lacking proper input validation.
- **Phishing:** Phishing involves fraudulent attempts to acquire sensitive information such as usernames, passwords, or financial details, usually through deceptive emails or websites. Attackers impersonate trusted entities to trick victims.

5.7 TARGET INDUSTRIES

- **IT (Information Technology):** The IT sector is highly susceptible to cyber threats due to the digital nature of its services, data handling, and online infrastructure. Attacks on IT companies can lead to intellectual property theft and operational disruptions.
- **Banking:** The banking sector faces constant cyber threats targeting online transactions, customer data, and financial records. Threats like ransomware and phishing are prevalent in this industry.
- **Government:** Government agencies are prime targets for cyber espionage and data breaches, often aimed at confidential information or disrupting public services. Nation-state actors frequently attack government networks.

5.8 ATTACK SOURCES

- **Insider:** Insider threats come from employees or contractors who intentionally or unintentionally compromise organizational security. These attacks are difficult to detect due to the trusted access insiders have.
- **Unknown:** Unknown sources refer to cyber-attacks whose origin cannot be identified. This category includes attacks masked by proxies, anonymizing services, or advanced evasion techniques.
- **Nation-state:** Nation-state attacks are sophisticated cyber operations conducted by governments to gather intelligence, disrupt infrastructure, or gain economic advantage. These often involve Advanced Persistent Threats (APTs).

5.9 SECURITY VULNERABILITIES

- **Unpatched Software:** Software vulnerabilities that remain unpatched allow attackers to exploit known weaknesses.

Regular updates and patch management are critical to preventing exploitation.

- **Zero-day:** Zero-day vulnerabilities are flaws unknown to the software vendor and exploited by attackers before a patch is available. These are highly dangerous due to the lack of defenses.
- **Weak Passwords:** Weak or reused passwords are a major source of security breaches. Attackers use brute-force or dictionary attacks to gain unauthorized access.

5.10 DEFENSE MECHANISMS

- **VPN (Virtual Private Network):** VPNs provide a secure, encrypted connection over the internet, protecting data in transit and masking the user's IP address to prevent eavesdropping and man-in-the-middle attacks.
- **Antivirus:** Antivirus software detects and removes malicious programs from devices, providing real-time protection against malware, ransomware, and other threats.
- **Firewall:** Firewalls act as a barrier between a trusted network and external threats, filtering incoming and outgoing traffic based on predefined security rules. They are fundamental for network defense.

Fig.5. Word Cloud: Attack Type

Fig.6. Word Cloud – Target Industry

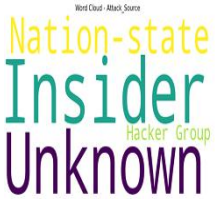

Fig.7. Word Cloud – Attack Source

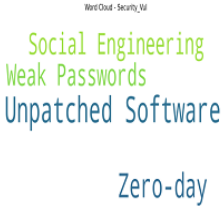

Fig.8. Word Cloud – Security Vulnerabilities

Fig.9. Word Cloud – Defense Mechanism

The word clouds clearly depict that Ransomware and Insider attacks are the most prevalent threats, while Unpatched Software and Virtual Private Networks (VPNs) are the most frequently exploited vulnerability and defense mechanism, respectively (Fig.5 to Fig.9).

5.11 DISCUSSION

From the results, Random Forest achieved the highest accuracy, followed closely by Logistic Regression and SVM. This indicates that ensemble tree-based models are most effective for classifying cyber security incidents with mixed numerical and categorical features. LSTM, although lower in accuracy, provides insights into temporal trends, which are valuable for predicting emerging threats. Word clouds complement these findings by highlighting frequent attack types and industries, providing actionable insights for security planning. Overall, the combination of statistical, machine learning and deep learning approaches provides a comprehensive understanding of cyber threats and their characteristics in India.

6. CONCLUSION AND SUGGESTIONS

The present study highlights the critical state of cybersecurity in India, demonstrating that both public and private sectors face significant threats from a variety of cyber-attacks, including ransomware, phishing, SQL injection, and DDoS. Analysis of the dataset indicates that certain industries such as IT, banking, and government are disproportionately affected, while insiders and unknown actors are among the most common sources of attacks. Additionally, prevalent vulnerabilities such as unpatched software, zero-day exploits, and weak passwords continue to be exploited by cybercriminals. Defense mechanisms such as Virtual Private Networks (VPNs), antivirus software, firewalls, and AI-based detection have been widely deployed; however, gaps remain in effective implementation and awareness. Overall, the findings underscore the urgent need for comprehensive cybersecurity strategies that combine technological solutions, policy frameworks, and human awareness initiatives.

6.1 SUGGESTIONS

- Strengthen Organizational Cybersecurity Policies and Training: Organizations should implement regular cybersecurity audits, update software patches promptly, and conduct continuous training for employees on recognizing phishing, social engineering, and other cyber threats. Emphasis should be placed on sectors like banking, IT, and government, which are frequent targets of cyberattacks.
- Enhance National-Level Cybersecurity Infrastructure: The government should invest in developing advanced threat

detection systems, support AI-based security solutions, and promote public-private partnerships to monitor and respond to cyber incidents in real time. Creating awareness campaigns and promoting secure digital practices among citizens will further reduce the vulnerability of individuals and small businesses.

REFERENCES

[1] CERT-In, "India Cyber Threat Report 2023", Data Security Council of India, Available at file:///C:/Users/user/Downloads/ANUAL-2024-0001.pdf, Accessed in 2023.

[2] S. Raghavan and A. Singh, "Challenges and Gaps in India's Cybersecurity Framework", *Criminal Law Journal*, Vol. 5, No. 1, pp. 412-423, 2021.

[3] R. Sharma, K. Patel and S. Agarwal, "Cyber Awareness and Threat Mitigation in India: A Review", *Journal of Information Security*, Vol. 10, No. 2, pp. 88-101, 2022.

[4] M. Joshi and P. Verma, "Digital Banking Vulnerabilities in India: A Threat Assessment", *International Journal of Financial Technology*, Vol. 3, No. 1, pp. 34-45, 2020.

[5] DSCI, "India Cybersecurity Domestic Market 2023", Data Security Council of India, Available at https://www.dsci.in/files/content/knowledge-centre/2023/India%20Cybersecurity%20Domestic%20Market%202023%20Report.pdf, Accessed in 2023.

[6] N. Gupta and R. Mehta, "Cyberattacks on Indian Academic Institutions", *Education and Cybersecurity Journal*, Vol. 7, No. 2, pp. 57-68, 2019.

[7] A. Singh, V. Kumar and S. Reddy, "Ransomware Attacks in Indian Healthcare and Retail Sectors", *International Journal of Cybersecurity Research*, Vol. 12, No. 3, pp. 120-134, 2021.

[8] P. Rao, "Phishing and DDoS Attacks in Indian Telecommunications", *Journal of Network Security*, Vol. 8, No. 4, pp. 45-57, 2021.

[9] R. Kapoor and S. Bose, "IoT Security Challenges in India: An Assessment", *Journal of Emerging Technologies*, Vol. 6, No. 1, pp. 15-28, 2022.

[10] DRDO, "Maya Operating System: Enhancing Cybersecurity in Defence Systems", Defence Research and Development Organisation, Available at https://www.scribd.com/document/928536570/Maya-OS, Accessed in 2023.

[11] Google, "DigiKavach: Protecting Indian Users from Online Fraud", Google India, Available at https://safety.google/intl/en_in/safety/engineering-center/engineering-center-india/, Accessed in 2023.

[12] S. Menon and A. Bhattacharya, "AI-based Threat Detection in Indian Cyberspace", *Journal of Intelligent Systems*, Vol. 14, No. 2, pp. 99-112, 2021.

[13] R. Nair, J. Thomas and K. Pillai, "Cybersecurity Workforce Training and Development in India", *Indian Journal of Information Security*, Vol. 5, No. 1, pp. 23-36, 2020.

[14] C.P. Krishna, "Cybersecurity Threats in Digital Banking in India: An Analytical Perspective", *ShodhKosh Journal*, Vol. 5, No. 1, pp. 2206-2218, 2024.

[15] V. Chopra and L. Varma, "Cybersecurity Framework for Critical Infrastructure in India", *International Journal of Critical Infrastructure Protection*, Vol. 11, No. 1, pp. 11-25, 2019.

[16] A. Desai and P. Kulkarni, "Cybersecurity Challenges in the Indian Automobile Sector", *Indian Journal of Cybersecurity*, Vol. 8, No. 2, pp. 77-90, 2022.