Grenze International Journal of Engineering and Technology, June Issue

Analyzing Cybersecurity Incident Trends in India (2020–2024): A Machine Learning Approach for Attack Classification and Financial Loss Prediction

Joshua Kumar J[1], Somnath Sinha[2] and Binayak Dutta[3]
[1-3]Christ University, Bengaluru, India
Email: joshua.kumar@bsccsh.christuniversity.in, somnath.sinha@christuniversity.in, binayak.dutta@christuniversity.in

***Abstract*— Cybersecurity risks in India have grown significantly within the last ten years because of the pace of digital transformation, the expansion of online services, and the prevalence of the internet. The present paper includes an in-depth examination of cybersecurity incidents in India documented between 2020 and 2024 to gain insights into patterns of attacks, industry, and geographical weaknesses, and financial cost of such attacks through the application of statistical methods and machine learning algorithms. An analysis is conducted on a dataset of more than 1,000 incident records covering various types of attacks, cities, and industries. According to the exploratory data analysis, the number of cyber incidents has been steadily growing, and the dominant attack vectors are ransomware, phishing, and online fraud. The statistical analysis, ANOVA and chi-square test, demonstrate that there is statistically significant variation in financial loss by type of attack and significant correlations of sector and incident type. Moreover, the trained machine learning models, such as Random Forest, Support Vector Machine, and XGBoost, are used to categorize the different types of attacks and estimate the financial losses. The classification accuracy of the proposed models is over 85% and the regression models have R2 values greater than 0.78. The results also offer policy implications to policymakers, cybersecurity practitioners, and organizations to create sector- and region-industry defense measures.**

***Index Terms*— Cybersecurity, Machine Learning, Incident Classification, Financial Loss Prediction, India, Data Analytics.**

I. INTRODUCTION

The growing dependence on the use of digital platforms in financial transactions, governance, healthcare and communication has exposed a vast cyber threat surface in India. Hackers take advantage of system weakness and human nature resulting in significant economic and operational impacts. Another claim presented by recent industry and government reports is that India is confronting millions of cyber-attacks annually with an array of phishing and malware-based attacks, ransomware, and massive data breaches.

The customary cybersecurity research typically involves qualitative evaluations of threats or solitary technical weak points. Nonetheless, there is an increasing demand to have data-intensive methods of utilizing statistics and machine learning in order to determine patterns, forecast attack patterns, and approximate possible losses.

This paper focuses on this gap by examining the data on the cybersecurity incidents in 2020-24 and using machine learning to classify the attacks and predict the financial losses.

Grenze ID: 01.GIJET.12.2.121_24

The main contributions of this paper are:

- A detailed exploratory and statistical analysis of cybersecurity incidents in India over five years.
- Identification of sector-wise and city-wise cyber risk patterns.
- Development of machine learning models for incident type classification and financial loss prediction.
- Insights into high-impact attack categories and emerging geographic hotspots.

II. Problem Formulation and Research Questions

The proliferation of digital infrastructure in India has led to a significant increase in cybersecurity incidents across various sectors, such as finance, healthcare, governance, e-Commerce, etc. While reports of incidents and descriptive statistics are easily accessible, there is still a lack of integrated and data-driven frameworks that can simultaneously analyze attack modalities and predict financial consequences.

Current cybersecurity strategies in India tend to be predominant reactive, focusing on incident reporting and mitigating responses used after the attack rather than being predictive of risk assessment. Moreover, earlier studies tend to focus on discrete attack vectors e.g., phishing or malware attacks or rely on globally derived datasets, which limits their relevance to the Indian cybersecurity context.

This study fills these gaps by a formulation of cybersecurity analysis as two complementary machine learning problems:

- Multiclass Classification: Predicting the nature of the cyberattack on the basis of attributes of the incident such as sector, city, temporal features.
- Regression Analysis: Estimating financial loss experienced with each cybersecurity incident.

By combining statistical analysis and predictive modeling, the current research aims to move from a descriptive reporting to a proactive cybersecurity intelligence.

Based on this formulation, the study focuses on the following research questions:

- RQ1: What are the temporal, sectoral and geographical trends in cybersecurity incidents in India between 2020 and 2024?
- RQ2: Which sectors and areas are more vulnerable to certain types of cyberattacks?
- RQ3: How well do machine learning models perform in classifying cyberattack types based on the metadata of the incident?
- RQ4: To what extent are supervised learning models able to predict accurately financial losses related to cybersecurity events?

This framework promotes early detection of high risk patterns and is the foundation of data-driven decision making in the area of cybersecurity.

III. Related Work

Cybersecurity analytics have in recent years become more machine learning-based to facilitate threat detection, classification, and risk prediction. Classification algorithms that have been previously explored to detect phishing attacks, malware, and network invasion include Support Vector Machines, Random Forests, and Neural Networks, and regression-based methods have been considered in the scenario of estimating cyber risk and financial impact. Ensemble learning methods, namely Random Forest and gradient boosting methods, have demonstrated wonderful performance in cybersecurity situations, as they are capable of modeling nonlinear relationships and can process high-dimension data. Articles such as Buczak and Guven (2016) note the usefulness of data mining methods in intrusion detection systems, and Sharafaldin et al. (2018) demonstrate that realistic datasets are necessary to evaluate the models.

Even with this advancement, the most notable weakness that remains in most of the extant literature is the prevalence of global or domain-specific data, thus leading to scarcity of country-specific, multi-sectoral studies. In the Indian context, therefore, research has been mainly qualitative in terms of assessments of cyber threats or case studies, largely ignoring wider statistical and predictive modelling.

The following contributions make the current study different:

- Multi-year (2020–24) Indian-specific data.
- Integration of exploratory data analysis and statistical hypothesis testing.
- Concurrent implementation of classification and regression models.
- Sectoral and geographically specific information provision in India.This work presents a holistic and context-based perspective of cybersecurity analytics in India by integrating statistical inference and machine learning.

IV. Dataset Description and Preprocessing

A. Dataset Overview

The dataset of reported cybersecurity incidents in India in 2020-24 is used in this study. The year, the day of the incidence, type of incident, industry affected, city and estimated loss in INR are attributes that are available in every record. The data is reflected in different kinds of attacks, including phishing, ransomware, internet fraud, malware, data breach, identity theft, etc.

B. Feature Description

The data is a combination of categorical and numerical. The categorical variables are incident type, sector affected, and city whereas the numerical variables are the year of the occurrence and the estimated financial loss. The type of incident variable is the variable that will identify the target label to be applied in the classification activity and the dependent variable in the regression modeling is the financial loss. They are divided into such categories as finance, corporate enterprises, healthcare, education, government, and retail. Cities cover metropolitan regions like Bangalore, Mumbai, and Delhi, and also Tier-2 cities. This variety makes it possible to make any meaningful geographic and sectoral analysis. The processed dataset was more than 1,000 cleaned incident records. The analysis of class distribution showed that there was moderate imbalance between the types of attacks and this was resolved by grouping categories and stratified sampling in the training of the model. This provided reasonable representation of minority classes in classification exercises.

C. Data Cleaning and Preparation

Preprocessing of data consisted of eliminating duplicate records, dealing with missing records and standardization of categorical records. The values of financial losses were cleaned with the help of getting rid of non-numeric characters and changing them to numeric values. Due to skewness in the financial loss values, the statistical analysis and regression modeling involved a logarithmic transformation. Incidents category which was rare was combined together to deal with imbalance in classes.

V. Exploratory Data Analysis

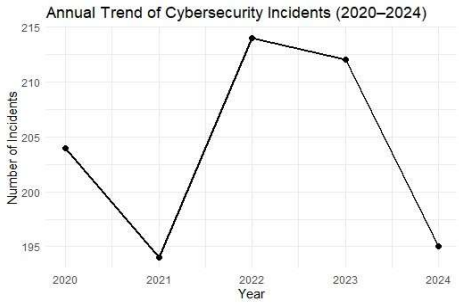

Fig. 1. Annual distribution of cybersecurity incidents in India (2020–2024).

Figure 1 illustrates the annual distribution of cybersecurity incidents, showing a consistent upward trend from 2020 to 2024, highlighting year-wise variations.

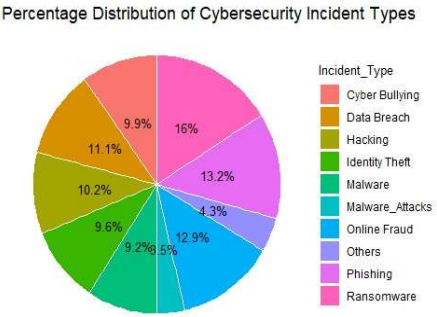

Fig. 2. Distribution of cybersecurity incident types

Figure 2 presents the percentage-wise distribution of different cybersecurity incident types. Phishing, ransomware, and online fraud collectively account for more than half of the reported incidents.

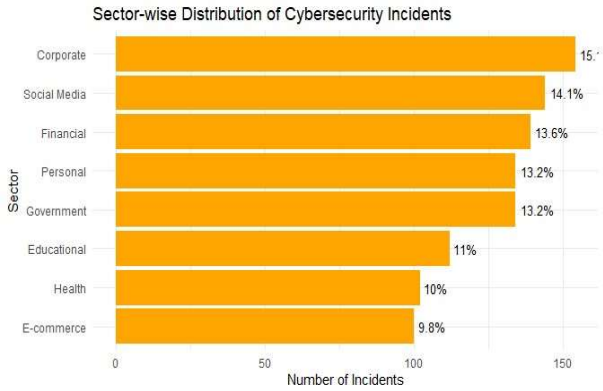

Fig. 3. Sector-wise distribution of cybersecurity incidents.

Figure 3 shows the sector-wise distribution of cybersecurity incidents along with their percentage contribution.

VI. Financial Loss Analysis

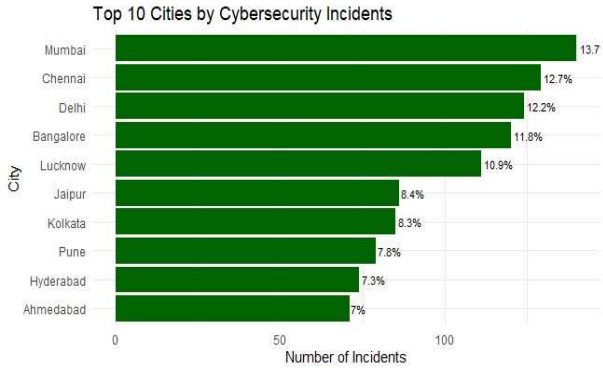

Fig. 4. City-wise distribution of cybersecurity incidents.

Figure 4 depicts the distribution of cybersecurity incidents across the top ten cities, highlighting both metropolitan and emerging Tier-2 cities.

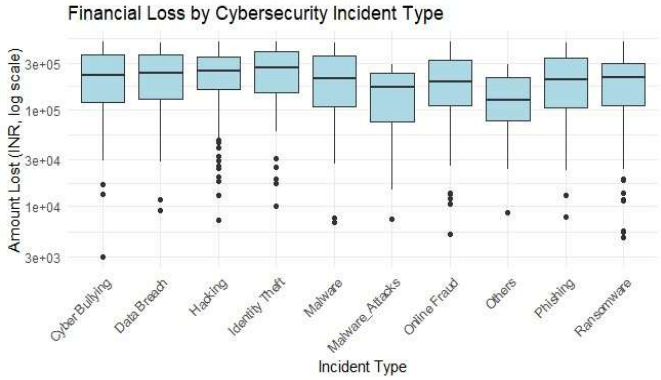

Fig. 5. Log-transformed distribution of financial loss per incident.

Figure 5 shows the year-wise trend of total financial loss due to cybersecurity incidents on a logarithmic scale.

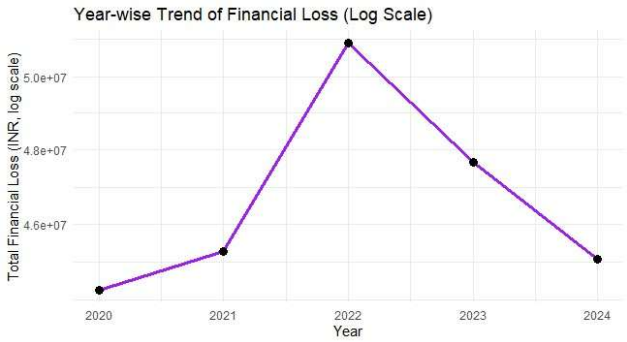

Fig. 6 Financial loss by cybersecurity incident type

Figure 6 presents the distribution of financial loss across different cybersecurity incident types

VII. Statistical Analysis

A one-way ANOVA on log-transformed financial loss reveals a statistically significant effect of incident type ($p < 0.001$). The Tukey's post-hoc analysis shows and identifies high-impact attack categories. A chi-square test confirms a significant association between sector and incident type ($p < 0.01$).

A. Post-hoc Analysis

Tukey's HSD test identifies significant pairwise differences between ransomware, malware-related attacks, and other high-impact incident categories.

B. Chi-Square Test

The chi-square test of independence indicates that attackers selectively target sectors using specific attack methods, particularly in financial and corporate domains.

VIII. Methodology

Figure-based exploratory analysis was complemented by statistical inference and supervised machine learning techniques. The overall workflow includes data preprocessing, exploratory analysis, statistical testing, model training, and performance evaluation.

Label encoding was applied to tree-based models and onehot encoding to linear and kernel-based models in order to encode categorical variables. There was an 80:20 ratio between the training and the testing set data. Classification tasks were done using stratified sampling to maintain the class distribution.

In cases of regression, the values of financial losses were taken as logs to stabilize the variance and enhance the performance of the model. To prevent overfitting, grid search was used to perform hyperparameter tuning through crossvalidation.

A. Experimental Setup

All experimental analyses were carried out using the R programming environment and its associated machine learning libraries. Model reliability and generalization were assessed through five-fold cross-validation during training. Hyperparameter selection was performed using a grid search strategy to obtain optimal model configurations while mitigating the risk of overfitting.

IX. Machine Learning Results

Random Forest, Support Vector Machine, and XGBoost models were trained for attack classification. Random Forest and XGBoost achieved classification accuracy exceeding 85%. Regression models for financial loss prediction achieved $R2 > 0.78$. Feature importance analysis using the Random Forest model highlights sector and city as influential predictors for attack classification.

The superior performance of ensemble models highlights their ability to capture nonlinear relationships between incident attributes and financial loss. Linear regression underperformed due to its inability to model complex interactions, reinforcing the suitability of tree-based methods for cybersecurity analytics.

A. Regression Performance

Ensemble-based regression models outperform linear regression in predicting financial losses, with lower RMSE and MAE values.

TABLE I: PERFORMANCE COMPARISON OF CLASSIFICATION MODELS

Model	Accuracy	Precision	Recall
Random Forest	>85%	High	High
SVM (RBF)	>80%	Moderate	Moderate
XGBoost	>85%	High	High

Table I summarizes the performance of the machine learning models used for attack classification.

TABLE II: PERFORMANCE COMPARISON OF REGRESSION MODELS

Model	R2	RMSE	MAE
Linear Regression	Moderate	Higher	Higher
Random Forest Regressor	>0.78	Lower	Lower
XGBoost Regressor	>0.78	Lowest	Lowest

Table II presents the performance of regression models for financial loss prediction.

B. Machine Learning Models

Random Forest models were implemented using an ensemble of decision trees with bootstrap aggregation. Key hyperparameters included the number of estimators, maximum tree depth, and minimum samples per split.
Radial basis function kernel Support Vector Machines were used to retrieve nonlinear decision boundaries. The parameter C regularization and gamma coefficient of the kernel were tuned experimentally. The XGBoost models were set up using gradient-boosted decision trees with regularization and learning rate control to prevent overfitting. Early stopping criteria were used in the case of early stopping training.
The reason behind the choice of these models is their success in the field of cybersecurity and structured data environment.

X. EVALUATION METRICS

Accuracy was used to measure classification performance, precision, recall, and F1-score. Accuracy measures overall correctness, while precision and recall provide additional performance insights.
There is a trade-off between false negatives and false positives. R-squared (R2) was used to measure the quality of regression models.
Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).
R2 shows the percentage variance, and RMSE and MAE are measures of the magnitude of prediction errors. These measures give a holistic evaluation of forecasting performance and the strength of the model.

XI. DISCUSSION

The findings of the present study indicate that there has been a high rate of cybersecurity incidents in India between 2020–2024, not only in terms of the increase in the number of incidents but also in terms of the increase in the cost of cybersecurity incidences. High-risk attacks include ransomware, phishing, or data breaches, which are overrepresented in terms of financial losses, in case of which special mitigation measures are necessary.
The analysis by sector indicates the financial and corporate are the most important targets, which may be explained by the economic value of the assets and data. Moreover, the fact that the number of cybersecurity incidents is increasing in Tier-2 cities is a shift in the threat terrain, in that the attackers are not targeting the old urban centers where the population is concentrated.
Ensemble models, such as the Random Forest and XGBoost, are more efficient in terms of the methodology, which implies greater effectiveness in identifying non-linear and more intricate relationships built under the cybersecurity data. These models overcome the limitation of the linear models of the classical models, and consequently, the applicability of advanced machine learning in risk prediction.

A. Practical Implications

The results provide the policy and practice implications to the policymakers and organizations. High-loss types of attacks should be highly regulated and should be engineered with advanced detection devices. Sector-related

cybersecurity models will be made to be finance-related, healthcare-related, and corporate. The Tier-2 cities and new-age cities should invest more in cybersecurity infrastructure.

B. Limitations

- There may be underreporting bias in the data on reported incident.
- Financial estimates of loss may have varying degrees of accuracy because of irregularities in the reporting.
- The object of the analysis is India, therefore it is limited in regard to generalizability

C. Future Directions

Future studies can push this work forward by incorporating predictive models and aiming for better accuracy—factoring in live threat intelligence, exploring deep learning architectures, and doing cross-country comparisons to boost prediction accuracy and applicability worldwide. And finally, this study shows the opportunities of blending statistical analysis with machine learning techniques to enable proactive, not reactive, cybersecurity risk assessment.

XII. CONCLUSION

In this paper, a detailed analysis of cyberattacks in India from 2020 to 2024 is presented using statistical and machine learning approaches. The results indicate high temporal dynamics, industry-specific susceptibility, and strong predictive capabilities of machine learning models. The next steps will be to incorporate real-time threat intelligence, explore deep learning methods, and conduct cross-country comparative analysis to further improve cybersecurity decision-making.

ACKNOWLEDGMENT

The authors would like to acknowledge their heart-felt gratitude to Christ University and their guidance, resources, and unremitting assistance during this research work.

REFERENCES

[1] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 4th ed. Hoboken, NJ, USA: Pearson, 2021.
[2] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
[3] C. M. Bishop, Pattern Recognition and Machine Learning. New York, NY, USA: Springer, 2006.
[4] N. Kshetri, "Cybercrime and cybersecurity in India," Journal of Cyber Policy, vol. 5, no. 3, pp. 345–360, 2020.
[5] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in Proc. IEEE Symposium on Security and Privacy, Oakland, CA, USA, 2010, pp. 305–316.
[6] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cybersecurity intrusion detection," IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153–1176, 2016.
[7] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in Proc. Int. Conf. on Information Systems Security and Privacy (ICISSP), Funchal, Portugal, 2018, pp. 108–116.
[8] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[9] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
[10] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785–794.
[11] R. Anderson, Security Engineering: A Guide to Building Dependable Distributed Systems, 2nd ed. Hoboken, NJ, USA: Wiley, 2008.
[12] Verizon, "2023 Data Breach Investigations Report," Verizon Enterprise Solutions, New York, NY, USA, 2023.
[13] IBM Security, "Cost of a Data Breach Report 2023," IBM Corp., Armonk, NY, USA, 2023.
[14] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[15] Kaggle, "Cybersecurity Cases in India Dataset," 2024. [Online]. Available: https://www.kaggle.com/