Cybercrime Prediction Using Statistical Models and Deep Learning

Harshit Narang
B.Tech in Computer Science and Engineering
Vellore Institute of Technology, Vellore
India
harshit.narang2023@vitstudent.ac.in

Rishit Shivam
B.Tech in Computer Science and Engineering
Vellore Institute of Technology, Vellore
India
rishit.shivam2023@vitstudent.ac.in

Pratyush Singh
B.Tech in Computer Science and Engineering
Vellore Institute of Technology, Vellore
India
pratyush.singh2023@vitstudent.ac.in

Garv Bansal
B.Tech in Computer Science and Engineering
Vellore Institute of Technology, Vellore
India
garv.bansal2023@vitstudent.ac.in

Abstract

In an era where cybercrime is evolving at a staggering pace, traditional security measures increasingly fall short in detecting and mitigating emerging threats. This research presents an innovative AI-driven framework that combines Long Short-Term Memory (LSTM) networks and Graph Neural Networks (GNNs) to predict cybercrime incidents before they occur. By leveraging a comprehensive dataset covering a decade (2013–2023) of cybercrime data, our approach not only captures long-term temporal trends but also models the complex interdependencies between diverse cyber threat vectors. The LSTM component is specifically designed to learn sequential patterns and forecast future incident rates, while the GNN component uncovers hidden relationships among cybercrime events, linking geographical hotspots with specific attack types and evolving digital environments. Furthermore, our model integrates anomaly detection techniques to identify outliers in cyber activity, enabling preemptive countermeasures. This holistic methodology provides actionable insights for cybersecurity professionals and law enforcement, facilitating proactive defense strategies that could significantly mitigate the economic and social impacts of cyber attacks. The research also discusses potential improvements such as real-time threat intelligence integration, enhanced interpretability through explainable AI (XAI), and the adoption of federated learning for privacy-preserving data analysis, ensuring the framework remains robust and adaptable to the ever-changing landscape of cyber threats.

1 Introduction

Cybercrime has become a growing threat in today's digital world, ranging from phishing scams and ransomware attacks to identity theft and financial fraud. Traditional cybersecurity defenses focus on reaction rather than prevention, leading to significant financial and data losses. While advancements in security protocols have helped mitigate risks, they remain largely reactive. The ability to anticipate and prevent cyber threats through predictive

modeling can be a game-changer in modern cybersecurity. This paper explores AI-driven cybercrime prediction by analyzing historical cybercrime trends and building a machine-learning model capable of forecasting future threats.

With the increasing digitalization of services, cybercriminals are employing more sophisticated attack techniques. The lack of predictive mechanisms in traditional security models means that organizations are often left reacting to threats rather than preventing them. AI-driven models have the potential to change this by identifying attack patterns before they escalate, allowing for timely intervention.

2 Literature Review

2.1 Summary of Existing Research

The application of Machine Learning (ML) in cybersecurity has grown significantly, with researchers exploring various methods for threat detection, intrusion prevention, and cybercrime analysis. Several studies have focused on using ML models for detecting cyber threats in real time, but relatively few have attempted to predict cybercrime trends before they occur.

- **Intrusion Detection Systems (IDS):** Traditional IDS rely on rule-based mechanisms or supervised learning models to identify network anomalies. While methods like Random Forest, Support Vector Machines (SVM), and XGBoost have shown promise, they primarily operate in a reactive manner rather than proactively predicting threats.
- **Phishing and Fraud Detection:** Researchers have utilized Natural Language Processing (NLP) and deep learning models like CNNs to detect phishing websites, email scams, and fraudulent transactions. However, these approaches often focus on classification rather than forecasting future cybercrime activities.
- **Malware Analysis and Threat Intelligence:** ML models such as Decision Trees and Deep Neural Networks (DNNs) have been applied to malware detection by analyzing executable files and network traffic patterns. Despite advancements, these studies primarily focus on identifying existing threats rather than anticipating new attack trends.
- **Cybercrime Trend Prediction:** Limited research has been conducted on forecasting cybercrime patterns using time-series models. Approaches leveraging AutoRegressive Integrated Moving Average (ARIMA) and basic neural networks have been explored but lack robustness in capturing complex dependencies over time.
- **Deep Learning for Cybersecurity:** While LSTM and CNNs have been widely applied in domains like finance and healthcare, their use in cybercrime forecasting is still emerging. Most existing studies focus on anomaly detection rather than predicting crime rates over time.

A major limitation in existing research is the lack of a comprehensive framework that integrates probability, statistics, and deep learning to predict cybercrime trends. Our study aims to bridge this gap by leveraging LSTM and CNN models alongside probabilistic approaches to improve forecasting accuracy and law enforcement preparedness.

2.2 Identified Research Gaps

While previous studies have successfully applied ML techniques such as Random Forest and XGBoost for cyber threat detection, they exhibit several limitations:

- **Lack of Temporal Awareness:** Many models fail to incorporate the sequential nature of cybercrime trends over time.
- **Limited Relationship Modeling:** Most approaches treat cybercrime incidents as independent events rather than interconnected occurrences across geographies and categories.
- **Bias in Training Data:** A lack of diverse datasets makes it challenging to generalize results across different regions and attack types.
- **High False Positive Rates:** Traditional ML techniques often misclassify legitimate activities as cyber threats.

2.3 How Our Model Addresses These Gaps

To bridge these gaps, our research introduces an **LSTM-GNN hybrid model** that:

- Leverages LSTMs to capture long-term temporal dependencies in cybercrime patterns.
- Uses GNNs to analyze relationships between different cybercrime types, geographical locations, and evolving trends.
- Incorporates an 80-20 dataset split for effective generalization and robust performance evaluation.
- Applies anomaly detection techniques to minimize false positives and improve prediction reliability.

By using a graph-based approach, we can model cybercrime as a network of interconnected events rather than isolated incidents, making our prediction model more robust and adaptable to evolving threats.

3 Methodology

3.1 Data Collection

We utilize an extensive dataset covering monthly cybercrime statistics from 2013 to 2023, containing records of reported cybercrime incidents, including:

- Phishing attacks
- Tampering of source code
- Online fraud
- Malware-based attacks
- Obscene and threatening emails/SMS

Each record includes:

- Number of cases registered
- Number of cases solved
- Pending cases
- Time-based trends (monthly data)
- Location-based statistics to capture geographical crime hotspots

The dataset is structured across multiple sheets corresponding to different years (2013–2023), allowing analysis of seasonal trends, geographical hotspots, and crime category distributions over time. By understanding these patterns, our model can identify high-risk periods and locations for cybercrime activity, enabling law enforcement to take proactive measures.

3.2 Machine Learning Techniques Used

To predict cybercrime trends, we employ:

1. **Long Short-Term Memory (LSTM) networks** – Captures temporal patterns in cybercrime trends, predicting future occurrences based on past patterns.
2. **Graph Neural Networks (GNNs)** – Models relationships between cybercrimes across regions and types, identifying interdependencies between cyber threats.
3. **Anomaly Detection (Autoencoders, Isolation Forest)** – Detects unusual spikes in cybercrime activity.
4. **Hybrid Approaches** – Combining LSTM with GNN to leverage both sequential dependencies and relational structures in cybercrime trends.

3.3 Data Preprocessing & Feature Engineering

- Handling missing values where case data is incomplete.
- Encoding categorical data (e.g., crime types, location) using label encoding.
- Standardization & normalization to ensure model accuracy.
- Constructing an adjacency matrix for GNNs to establish cybercrime relationships.
- Extracting temporal features such as seasonal trends, year-wise variations, and cyclic patterns.
- Splitting the dataset into an 80-20 ratio, with 80% for training and 20% for testing to ensure model generalization.
- **Feature Extraction:** Identifying key indicators such as economic factors, digital penetration rates, and cybersecurity infrastructure in different regions.
- **Data Augmentation:** Using synthetic data generation methods to enhance dataset diversity and robustness.

4 Challenges in AI-Based Cybercrime Prediction

1. **False Positives:** Some legitimate activities may be wrongly flagged as cyber threats.
2. **Bias in Training Data:** The dataset mostly covers one region, limiting global applicability.
3. **Computational Complexity:** GNNs require high processing power due to relational data modeling.
4. **Ethical Concerns:** Predictive policing in cybercrime raises privacy concerns.
5. **Data Imbalance:** Certain types of cybercrime may be underrepresented, affecting model accuracy.
6. **Model Interpretability:** Deep learning models, especially GNNs, can act as black boxes, making their predictions difficult to explain.

5 Conclusion & Future Work

This research highlights how LSTM and GNN models can be leveraged to predict cybercrime trends, offering law enforcement and cybersecurity professionals a powerful preemptive tool. Future work will focus on:

- Expanding the dataset to include global cybercrime trends.
- Enhancing model accuracy through reinforcement learning.

- Integrating real-time cyber threat intelligence feeds.
- Addressing ethical concerns in AI-driven law enforcement.
- Implementing explainable AI (XAI) techniques to improve model transparency and decision-making.
- Exploring federated learning to enable privacy-preserving cybercrime prediction models.
- Conducting real-world testing by partnering with cybersecurity agencies to validate the model's effectiveness in active threat detection scenarios.

References

1. P. N. V. Kumar, "Growing cyber crimes in India: A survey," 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE), Ernakulam, India, 2016, pp. 246–251, doi: 10.1109/SAPIENCE.2016.7684146.
2. M. Xu, K. M. Schweitzer, R. M. Bateman and S. Xu, "Modeling and Predicting Cyber Hacking Breaches," in *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2856–2871, Nov. 2018, doi: 10.1109/TIFS.2018.2834227.
3. T. Arora, M. Sharma and S. K. Khatri, "Detection of Cyber Crime on Social Media using Random Forest Algorithm," 2019 2nd International Conference on Power Energy, Environment and Intelligent Control (PEEIC), Greater Noida, India, 2019, pp. 47–51, doi: 10.1109/PEEIC47157.2019.8976474.
4. Y. Goyal and A. Sharma, "A Semantic Approach for Cyber Threat Prediction Using Machine Learning," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 435–438, doi: 10.1109/ICCMC.2019.8819694.
5. M. Arshey and K. S. Angel Viji, "Thwarting Cyber Crime and Phishing Attacks with Machine Learning: A Study," 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2021, pp. 353–357, doi: 10.1109/ICACCS51430.2021.9441925.
6. A. Swaminathan, B. Ramakrishnan, K. M and S. R, "Prediction of Cyber-attacks and Criminality Using Machine Learning Algorithms," 2022 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), Sakheer, Bahrain, 2022, pp. 547–552, doi: 10.1109/3ICT56508.2022.9990652.
7. E. F. Al Jarboua, M. Bte Md. Din and A. A. Bakar, "Cyber-Crime Detection: Experimental Techniques Comparison Analysis," 2022 International Visualization, Informatics and Technology Conference (IVIT), Kuala Lumpur, Malaysia, 2022, pp. 124–129, doi: 10.1109/IVIT55443.2022.10033332.

8. N. Aggarwal, M. Sehgal and A. Arya, “An empirical analysis of Cyber Crimes, their prevention measures, and laws in India,” 2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC), Solan, Himachal Pradesh, India, 2022, pp. 570–575, doi: 10.1109/PDGC56933.2022.10053354.