Temporal Analysis and Forecasting of Cyber Attack Intensity

Shrikant Khawshe[1] Vedanshu Thune[2] Yogvid Wankhede[3] Prof. Madhuri Sahu[4]
[4]Professor
[1,2,3,4]Department of Artificial Intelligence
[1,2,3,4]G H Raisoni College of Engineering, Nagpur, India

Abstract — As the frequency and sophistication of cyber-attacks continue to escalate in today's digital landscape, there is an urgent need for innovative approaches to understand and predict their intensity. This project focuses on leveraging temporal analysis techniques to forecast cyber-attack intensity by identifying patterns within historical data. Through the exploration of temporal correlations in attack volumes, the aim is to develop a robust forecasting system capable of anticipating future cyber incidents.

Keywords: Machine Learning, Multinomial Naïve-Bayes, React Native, Data Extraction, Text Classification, Firebase, NER

I. INTRODUCTION

In the ever-evolving digital landscape of today, the escalating threat posed by cyber-attacks continues to challenge even the most robust security defenses deployed by organizations worldwide. Cybercriminals relentlessly seek ways to exploit vulnerabilities across a myriad of applications within organizational infrastructures, as exemplified by recent incidents like the SolarWinds update compromise. These events underscore the critical need for innovative approaches to cybersecurity.

The exponential growth of data generated from diverse sources, including network traffic logs, system logs, sensor data, and Internet of Things (IoT) devices, presents both opportunities and challenges in cybersecurity. Leveraging big data alongside machine learning and artificial intelligence has emerged as a promising strategy for timely intrusion detection and identification across heterogeneous networks. By harnessing big data analytics, organizations can gain insights into malicious activities, enabling effective mitigation to minimize financial losses and reputational damage.

Despite advancements in cybersecurity technologies, accurately predicting cyber-attacks remains a formidable challenge due to the sophistication of modern attacks and the expansive attack surface available to cybercriminals. Traditional security approaches are proving inadequate in the face of evolving threats, necessitating a shift towards predictive methods capable of detecting and pre-empting sophisticated attacks.

The escalating cost of cyber-attacks, with the average data breach estimated to cost organizations millions of dollars and significant time to detect and contain, underscores the urgency of developing proactive cybersecurity measures. With the proliferation of connected devices and the persistence of Advanced Persistent Threats (APTs) groups employing complex techniques to evade detection, the intensity of cyber-attacks is expected to continue unabated. Predictive approaches offer a promising avenue for redirecting technical resources towards mitigating emerging threats, thereby enhancing cybersecurity resilience.

In our study, we delve into the realm of cyber event forecasting, with the primary goal of reinforcing our ability to detect and prevent cyber-attacks accurately. While forecasting methodologies are common in fields like weather prediction and stock markets, their adaptation to the domain of cyber security is still relatively new. Our research seeks to bridge this gap by harnessing the power of unstructured big data along with robust machine learning techniques to predict cyber intrusions.

Our methodology encompasses a comprehensive suite of machine learning algorithms, including Logistic Regression, Naive Bayes, Support Vector Machines, Decision Tree, Random Forest, XGBoost Regressor, Artificial Neural Network, ARIMA (Autoregressive Integrated Moving Average), and LSTM (Long Short-Term Memory). These algorithms are applied to a rich dataset collected by the University of New South Wales (UNSW), encompassing nine distinct types of cyber-attacks:

1) Fuzzers: Attacks employing randomized data to assess system resilience and pinpoint vulnerabilities.
2) Analysis: Techniques involving system scrutiny to identify weaknesses and potential targets for exploitation.
3) Backdoors: Intrusions establishing covert entry points into systems for subsequent unauthorized access.
4) DoS (Denial of Service): Offensives aiming to disrupt system functionality, rendering it inaccessible to legitimate users.
5) Exploits: Intrusions leveraging system vulnerabilities to gain illicit access or control.
6) Generic: A diverse category encompassing various attack modalities not falling under specific classifications.
7) Reconnaissance: Maneuvers gathering intelligence on target systems, including vulnerabilities and entry points, in readiness for future attacks.
8) Shellcode: Malicious code execution, often in the form of shell scripts, on target systems.
9) Worms: Self-replicating malware proliferating across systems, often inflicting damage in the process.

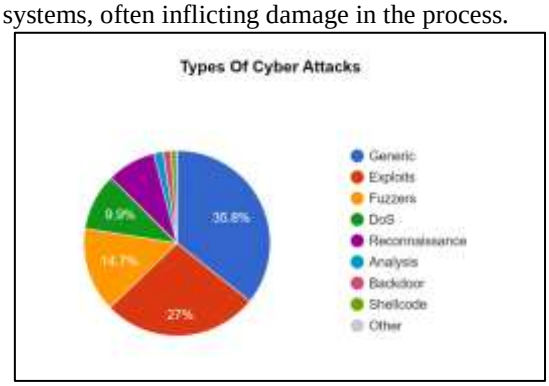

Fig. 1: Types of Cyber Attack

In our study, we employ time series resampling and evaluate various forecasting models, such as Linear Regression and Long Short-Term Memory (LSTM), with the aim of predicting cyber-attack occurrences within predefined time frames. We assess the effectiveness of our forecasts using metrics like Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

II. RELATED WORK

Traditional intrusion detection systems primarily rely on a signature-based approach, matching monitored events against known signatures of previously observed incidents. However, these systems often struggle to detect new events with unknown signatures and may require significant time to deploy signatures across the network. Anomaly detection approaches, on the other hand, focus on identifying deviations from normal behaviour but suffer from a high false alarm rate due to their sensitivity to previously unseen actions.

Recent research has explored frameworks for attack projection, which aim to model how cyber-attacks unfold over time by focusing on the footprints and interdependencies of multi-stage cyber-attacks. These frameworks generate hypothesized attack strategies based on observed footprints to project future actions of ongoing attacks. Effective algorithms such as attack graphs and probabilistic reasoning with dynamic Bayesian networks have been proposed to correlate isolated alerts into coherent attack scenarios and predict forthcoming attacks.

Machine learning methodologies have emerged as key tools for intrusion detection and cyber-attack prediction. Various techniques, including artificial neural networks, clustering, ensemble learning, and support vector machines, have been extensively utilized in this domain. Some studies have specifically focused on predicting multi-stage cyber-attacks, employing methodologies such as Longest Common Subsequence (LCS) algorithms and variable-length Markov models (VLMM) to quantify similarities between attack progressions and capture sequential properties of attacks for future activity projection.

- ARIMA (Autoregressive Integrated Moving Average): ARIMA is a time series forecasting algorithm that models the relationship between a series of observations and their lagged values. In the context of cyber-attack forecasting, ARIMA can be used to analyze historical patterns of cyber-attacks and predict future attack occurrences based on the time series data.
- LSTM (Long Short-Term Memory): LSTM is a type of recurrent neural network (RNN) architecture designed to model sequence data and capture long-term dependencies. In cyber-attack forecasting, LSTM networks can be trained on historical attack data to learn complex temporal patterns and make predictions about future attack events.
- Decision Tree: Decision trees are widely recognized in the machine learning domain for their versatility in both classification and regression tasks. In the field of cyber-attack forecasting, they prove valuable by allowing analysis of features derived from network traffic data. This analysis helps in spotting patterns that could signify potential attack activities.
- Random Forest: Random Forest is an ensemble learning algorithm that constructs multiple decision trees and combines their predictions to improve accuracy and robustness. In cyber-attack forecasting, random forest models can leverage the collective intelligence of multiple decision trees to make more reliable predictions about future attack events.

III. PROPOSED METHODOLOGY

In our study, we introduced a cyber event forecasting model based on time series analysis. Given the complexity of cyber-attacks, our model aims to provide valuable insights to forecast such events. By adopting forward-thinking strategies, it enables security teams and senior management to implement robust measures safeguarding their network infrastructure. We utilized an openly accessible dataset containing various attack indicators gathered from a realistic and secure network environment.

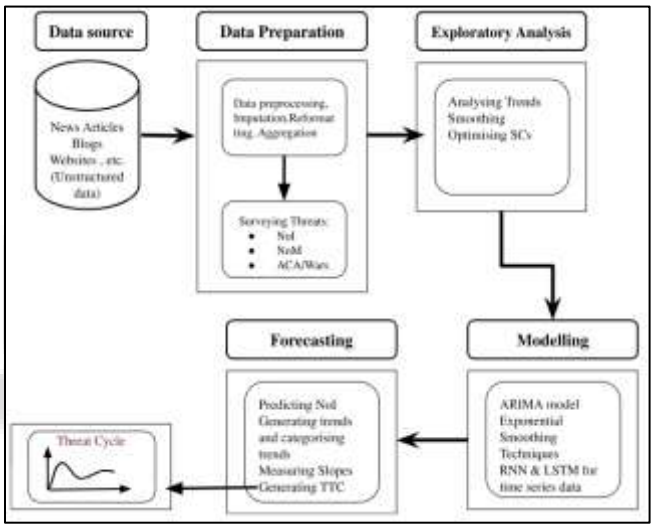

Fig. 2: Architecture

A. *Problem Definition and Scope*

Our aim with this research is to create a time series-based model specifically tailored for forecasting the intensity of cyber-attacks. Through our project titled "Temporal Analysis and Forecasting of Cyber Attack Intensity," we aspire to equip security teams and senior management with innovative strategies to bolster the resilience of their network systems. The scope of our study encompasses various types of cyber-attacks, including but not limited to Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic attacks, Reconnaissance, Shellcode, and Worms. The timeframe for forecasting spans a significant duration, enabling meaningful analysis and prediction of cyber-attack events.

B. *Data Collection*

We made use of a readily accessible dataset comprising records of cyber-attacks gathered from a realistic and secure network environment. This dataset encompasses various attack labels and covers a significant timeframe, enabling us to conduct thorough and comprehensive analysis.

C. Data Preprocessing

Prior to analysis, the dataset underwent thorough preprocessing to ensure data quality and suitability for modeling. Irrelevant or sparse features were removed, and data sampling was performed at regular intervals to produce a time-series dataset. Stationarity of the time-series data was verified to ensure the stability of our forecasting model.

D. Feature Selection

To identify the most relevant features for our forecasting model, we employed feature selection techniques such as Information Gain (IG). By selecting a subset of features that contribute significantly to predicting cyber-attack intensity, we aimed to enhance the accuracy and efficiency of our model.

E. Model Development

Our model development process involved the implementation of a time series forecasting model using the selected features. We explored various methodologies, including autoregressive integrated moving average (ARIMA), exponential smoothing techniques, and machine learning algorithms suitable for time series data, such as recurrent neural networks (RNNs) and long short-term memory (LSTM) networks.

F. Model Training and Testing

The dataset was split into training and testing sets to facilitate model training and evaluation. We utilized the training set to train the forecasting model and the testing set to estimate its performance. Model parameters were tuned as needed to optimize forecasting accuracy.

G. Forecasting

Employing the trained model, we forecasted the intensity of cyber-attacks within a specified time frame. This intensity was measured by Predicting NoI, generating trends and categorising trends, generating TTC (The Threat Cycle). Based on these and predictions based on historical data and selected features, we aimed to provide insights into future attack events.

H. Performance Evaluation

The performance of our forecasting model was assessed using relevant criteria such as mean absolute error (MAE), root mean squared error (RMSE), and accuracy. Comparing predicted attack intensities with factual observations validated the effectiveness of our model.

I. Attack Classification

Additionally, we applied supervised learning algorithms to classify and categorize observed cyber-attacks based on their characteristics. Evaluation of classification algorithms such as Bayes Net, SVM, and random forest provided insights into the effectiveness of attack classification techniques.

J. Documentation and Reporting

The entire methodology, including data preprocessing steps, feature selection, model development, and evaluation procedures, was thoroughly documented. Findings, insights, and recommendations were presented in a comprehensive report suitable for publication or presentation, highlighting the significance of our project in enhancing cybersecurity resilience through temporal analysis and forecasting of cyber-attack intensity.

IV. IMPLEMENTATION

In this project titled "Temporal Analysis and Forecasting of Cyber Attack Intensity," we developed a cyber-attack detection system (IDS) to identify and classify different types of network attacks using the KDDCup 1999 dataset, a widely recognized benchmark for network cyber-attack detection tasks. The dataset consists of a series of TCP connection records, each containing 41 features and a label indicating whether the connection is normal or an attack. The dataset includes four main categories of attacks: Denial of Service (DoS), Probing (Probe), Remote to Local (R2L), and User to Root (U2R). For our project, we utilized a 10% subset of the original dataset.

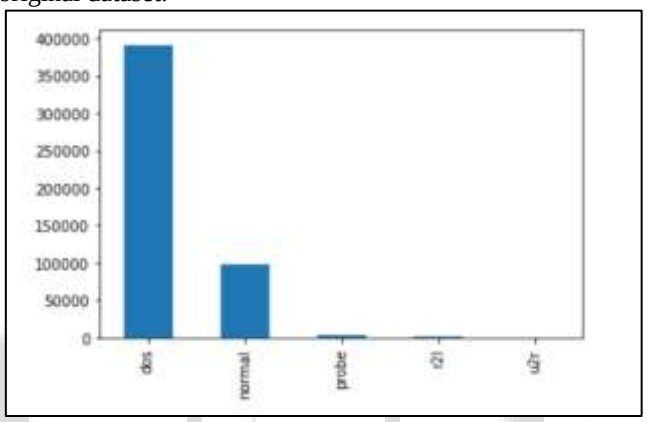

Fig. 3: Categories Of Attacks

The dataset was loaded into a Pandas Data Frame, and an 'Attack Type' column was added to label each connection as either normal or one of the four attack categories. After checking for missing values, we prepared the dataset for modeling by identifying and separating the categorical and numerical features. To reduce the dimensionality of the feature space and potentially improve model performance, we performed feature selection by dropping columns with missing values and columns with only a single unique value, as they do not provide any discriminatory information. Additionally, we computed and visualized a correlation matrix using a heatmap to identify and potentially remove highly correlated features.

We implemented and evaluated seven different machine learning models for the cyber-attack detection task, including Gaussian Naive Bayes, Decision Tree, Random Forest, Support Vector Machine (SVM), Logistic Regression, Gradient Boosting Classifier, and a custom Artificial Neural Network (ANN) architecture. The Random Forest Model emerged as the best model for cyber-attack detection, outperforming the other models in terms of accuracy with both the dependent feature label and attack category. To gain insights into the rules governing cyber-attack detection, we employed the Decision Tree Model. This analysis enabled us to identify the most important features that play a crucial role in detecting cyber-attacks accurately.

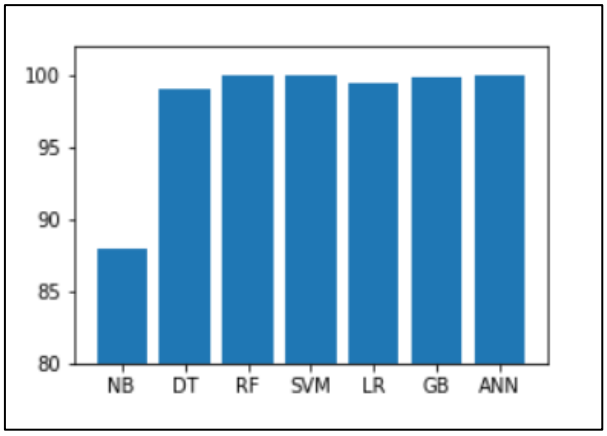

Fig. 4: Accuracy of different models

The Random Forest Model's superior performance can be attributed to its ensemble learning approach, which combines multiple decision trees to provide more robust and accurate predictions. By leveraging the collective wisdom of multiple trees, the Random Forest Model can effectively handle complex patterns and relationships in the data, making it well-suited for the cyber-attack detection task. Furthermore, its ability to handle high-dimensional data and its resilience to overfitting contributed to its strong performance.

For each model, we split the dataset into training and testing sets, with 67% of the data used for training and 33% for testing. The models were trained on the training set, and their performance was evaluated on the testing set using metrics such as accuracy score, training time, and testing time. We compared the performance of the different models, analyzing their strengths and weaknesses in the context of the intrusion detection task, and discussed potential reasons for the superior performance of certain models.

V. Result

In this study, our experiment aimed to evaluate the effectiveness of various classification and forecasting techniques in predicting cyber events using a comprehensive dataset of cyber-attack records. Here are the key findings:

Initially, in the baseline classification phase, random forest emerged as the most accurate classifier with an impressive accuracy of 95.4%, accompanied by a False Alarm Rate (FAR) of 0.6%. Following closely behind was ANN with an accuracy of 92.6%, while BayesNet and Naive Bayes yielded lower accuracy scores.

Transitioning to the classification of forecasted events, we employed linear regression and SMOreg techniques. The results indicated a minor drop in accuracy in comparison to the baseline results. SVM and random forest achieved the highest classification accuracy of 88.3% for events forecasted using linear regression. Similarly, for events forecasted using SVM and random forest, the highest classification accuracy reached 89.4%.

Moreover, when evaluating time series-forecasted events through Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) metrics, we uncovered interesting insights. SMOreg's forecasted events showed better performance in terms of MAE, outperforming both LSTM and Linear Regression techniques. However, it's important to mention that when we looked at the Root Mean Square Error (RMSE) evaluation, linear regression showed the best prediction accuracy in three of the top five features.

In addition to the experimental findings, the dataset itself provided comprehensive insights into cyber threats across multiple dimensions, including spam, ransomware, local infection, exploit, malicious mail, network attack, on-demand scan, and web threat. Statistical analysis shed light on the varying intensity and frequency of different types of cyber-attacks across different countries.

In summary, the results underscore the efficacy of machine learning techniques in classifying and forecasting cyber events. The comprehensive dataset and the developed cyber-intelligence solution provide valuable resources for researchers, policymakers, and cybersecurity professionals to enhance their understanding of cyber threats and mitigate potential risks effectively.

VI. Conclusion

In this research, we constructed an extensive model designed to forecast and predict cyber events, drawing from a publicly available IDS dataset and employing advanced machine learning techniques. Through meticulous data preparation and resampling, we ensured the reliability and relevance of our analysis. Initial classification results on the baseline dataset were great, with random forest leading with an accuracy of 95.4%. Subsequently, we turned our focus to cyber event forecasting, utilizing SMOreg, Linear Regression and LSTM methods to anticipate future events. When machine learning classification algorithms were applied to SMOreg-forecasted events, they exhibited the highest performance, achieving a score of 88.3%. Assessment using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) metrics underscored the superior performance of SMOreg-forecasted events, particularly in three of the top five features.

This study makes a significant contribution to the realm of cyber-attack prediction, providing more precise forecasts informed by detected cyber occurrences within the network. Our discoveries offer valuable insights for security professionals and decision-makers, empowering them to effectively plan proactive security measures to safeguard systems and critical infrastructure.

While our forecasting was limited to particular hourly intervals due to dataset constraints, this timeframe still allows ample opportunity for security professionals to react and mitigate potential threats proactively. Future research endeavours aim to expand upon this work by encompassing the entire dataset spanning multiple days, with the goal of extending the forecasting interval to between 4 to 8 days. We recognize the rapid evolution of threats, and we believe that extending the forecasting horizon within this timeframe will provide actionable insights without rendering leading indicators obsolete. Furthermore, we addressed a significant gap in the field of cyber-attack detection for IoT by presenting a novel DL-based intrusion detection method using focal loss. Our approach, CNN-Focal, outperformed baseline models and state-of-the-art approaches by substantial margins across various evaluation metrics and

datasets. This underscores the efficacy and potential of our methodology in enhancing IoT security.

In conclusion, our research lays a solid foundation for advancing cyber event forecasting and cyber-attack detection, paving the way for more proactive and effective security measures in an increasingly interconnected and vulnerable digital landscape.

REFERENCES

[1] Rildo Antonio de Souza, Vitor de Castro Silva, Sylvio Barbon Junior, Bruno Bogaz Zarpelão, "Forecasting Malware Incident Rates in\\xa0Higher Education Institutions", Advanced Information Networking and Applications, vol.202, pp.226, 2024.
[2] Yussuf Ahmed, Muhammad Ajmal Azad, Taufiq Asyhari, "Rapid Forecasting of Cyber Events Using Machine Learning-Enabled Features", Information, vol.15, no.1, pp.36, 2024.
[3] Pavol Sokol, Richard Staňa, Andrej Gajdoš, Patrik Pekarčík, "Network security situation awareness forecasting based on statistical approach and neural networks", Logic Journal of the IGPL, vol.31, no.2, pp.352, 2023.
[4] Richard Staňa, Patrik Pekarčík, Andrej Gajdoš, Pavol Sokol, "Network Security Situation Awareness Forecasting Based on Neural Networks", Theory and Applications of Time Series Analysis and Forecasting, pp.255, 2023.
[5] Supreeth Nigam, Abhishek Srivastava, Akansh Singh, Aniket Behera, Ashish Kumar, "Recognizing phishing site using Machine Learning-A Comparative Approach using MultinomialNB & Logistic Regression", Journal of Advanced Zoology Vol 44, p955, 2023.
[6] Dina, A.S.; Siddique, A.; Manivannan, D. A deep learning approach for intrusion detection in Internet of Things using focal loss function. Internet Things 2023, *22*, 100699.
[7] Almahmoud, Z.; Yoo, P.D.; Alhussein, O.; Farhat, I.; Damiani, E. A holistic and proactive approach to forecasting cyber threats. Sci. *Rep.* 2023, *13*, 8049.
[8] Schmidl, S.; Wenig, P.; Papenbrock, T. Anomaly detection in time series: A comprehensive evaluation. Proc. VLDB Endow. 2022
[9] Gurung, S.; Ghose, M.K.; Subedi, A. Deep learning approach on network intrusion detection system using NSL-KDD dataset. Int. J. Comput. Netw. Inf. Secur. 2019
[10] Yaffee, Robert A., and Monnie McGee. An introduction to time series analysis and forecasting: with applications of SAS® and SPSS®. Elsevier, 2000.
[11] Choudhary, Rishabh, and Hemant Kumar Gianey. "Comprehensive review on supervised machine learning algorithms." 2017 International Conference on Machine Learning and Data Science (MLDS). IEEE, 2017.
[12] Samek, Wojciech, et al. "Explaining deep neural networks and beyond: A review of methods and applications." Proceedings of the IEEE 109.3 (2021): 247-278.
[13] Yu, Yong, et al. "A review of recurrent neural networks: LSTM cells and network architectures." Neural computation 31.7 (2019): 1235-1270.
[14] Gers, Felix A., Nicol N. Schraudolph, and Jürgen Schmidhuber. "Learning precise timing with LSTM recurrent networks." Journal of machine learning research 3.Aug (2002): 115-143.
[15] Pascanu, Razvan, Tomas Mikolov, and Yoshua Bengio. "On the difficulty of training recurrent neural networks." International conference on machine learning. PMLR, 2013.
[16] Elsworth, Steven, and Stefan Güttel. "Time series forecasting using LSTM networks: A symbolic approach." arXiv preprint arXiv:2003.05672 (2020).
[17] Yu, Yong, et al. "A review of recurrent neural networks: LSTM cells and network architectures." Neural computation 31.7 (2019): 1235-1270.
[18] Jakaša, Tina, Ivan Andročec, and Petar Sprčić. "Electricity price forecasting—ARIMA model approach." 2011 8th International Conference on the European Energy Market (EEM). IEEE, 2011.
[19] Contreras, Javier, et al. "ARIMA models to predict next-day electricity prices." IEEE transactions on power systems 18.3 (2003): 1014-1020.
[20] Meyler, Aidan, Geoff Kenny, and Terry Quinn. "Forecasting Irish inflation using ARIMA models." (1998): 1-48.
[21] Box, GEORGE EP, Gwilym M. Jenkins, and G. Reinsel. "Time series analysis: forecasting and control Holden-day San Francisco." BoxTime Series Analysis: Forecasting and Control Holden Day1970 (1970).
[22] Contreras, Javier, et al. "ARIMA models to predict next-day electricity prices." IEEE transactions on power systems 18.3 (2003): 1014-1020.
[23] Sepp Hochreiter, Jürgen Schmidhuber; Long Short-Term Memory. Neural Comput 1997; 9 (8):1735–1780.doi: https://doi.org/10.1162/neco.1997.9.8.1735.

