International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 08 | Aug 2026 www.irjet.net p-ISSN: 2395-0072

An Integrated Geo-Temporal System for Forecasting Cyber Attacks Using Machine Learning

S.V.S. Santhi[1], N. Hima Chakradhar[2], M. Satwik[3], P. Bipin Satyankar[4]

[1,2,3,4]*Dept. of Computer Science and Engineering (Data Science), Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh, India*

Abstract - *In today's digital era, the increasing volume and sophistication of cyber threats demands intelligent and automated defences. Machine learning techniques are used to detect, analyse, and predict cyber-attacks, enabling proactive cybersecurity measures and timely threat mitigation. Existing models analyse location and time of cyber-attacks in isolation, without combining them effectively or providing accurate forecasting before the threat occurs. The proposed model overcomes these limitations by integrating location and time-based features in a unified framework. The system uses encoded regional information along with time patterns to understand attack behaviour. A multi-output regression model is trained to predict the different attack types and network protocols for two weeks and identifies the most dominant type of attack likely to occur; model performance is evaluated using R^2, MAE and RMSE. Hence, the proposed work for forecasting cyber-attacks proves to be novel by leveraging geo-temporal data in a unified and predictive framework.*

Key Words: Cyber Attack Prediction, Random Forest, Time Series Forecasting, Multi-Output Regression, Predictive Modeling, Threat Intelligence, Geo-Temporal Analysis, Machine Learning.

1. INTRODUCTION

The growing use of digital technology in government services, financial systems, healthcare systems, and business networks has created a larger cyberspace threat environment. The increasing dependence of countries on their digital networks has led to more complex cyber-attacks which include Distributed Denial of Service (DDoS) attacks, malware, ransomware, and network hacking. The Indian government has made protecting essential cyber systems a national priority because the country is expanding its digital transformation projects through cloud technology and e-government services. Traditional cybersecurity mechanisms depend on reactive defence models which only detect and deal with threats after attackers have already entered the system.

The latest developments in data science and machine learning technology create fresh possibilities for predictive cybersecurity analytics through their recent advancements. Machine learning models use historical cyber incident datasets together with network telemetry data to reveal hidden time-based patterns that link to attack incidents and forecast future threats [3]. Organizations can use predictive analytics to shift from defensive security methods to proactive protection systems because it helps them detect threats early, distribute resources effectively, and prepare for incidents more efficiently.

Ensemble learning methods provide superior performance for modelling complex non-linear data patterns which are typical in cybersecurity datasets according to machine learning research. Random Forest algorithms achieve strong predictive accuracy through their method of combining multiple decision trees while using bootstrap aggregation to decrease prediction uncertainty. The Random Forest model exhibits effective prediction performance across various types of cybersecurity data which comprises attack frequency patterns, network abnormality detection, and state-based incident distribution.

Predictive analytics is increasingly becoming significant within the context of cybersecurity, as shown by various scholarly findings. According to recent research on intrusion detection systems based on machine learning, the use of ensemble algorithms proves to be efficient in detecting complicated attack patterns [5]. Recent research on the analysis of cyber threat intelligence indicates that predictive modelling should be recognized as the primary method to increase the capacity of national cybersecurity [6]. Time-series models, along with predictive analytics, have developed systems capable of predicting future attacks and measuring threats in various regions [7].

The main contributions of this study are briefly listed below:

1) A predictive cybersecurity framework that is aimed at predicting the distribution of cyber-attacks on Indian states through machine learning algorithms.

2) A Random Forest-based multi-target regression framework that can be used to predict both time and spatial cyber-attack patterns.

3) A data processing, feature extraction, forecasting, and evaluation pipeline for analysing cyber-attacks using machine learning forecasting approaches.

4) Data visualization dashboards for monitoring attacks and supporting decision making in cybersecurity.

5) The integration of machine learning forecasting and cybersecurity analysis to develop a predictive cyber threat defence mechanism for large-scale infrastructures.

2. RELATED WORK

Cybersecurity research has mainly concentrated on three areas which include intrusion detection systems, machine learning-based threat classification, and time-series cyber-attack forecasting. The systems detect security threats by studying both network traffic patterns and user behaviour patterns. Cyber threat detection has traditionally been done through machine learning methods which include Support Vector Machines, Decision Trees, and Deep Belief Networks to solve spam, malware, and intrusion detection problems. The techniques achieve high accuracy for detecting existing attack patterns, but they remain reactive methods which cannot accurately predict upcoming cyber-attack patterns [1], [2].

Ahmed et al. [1] conducted research to study how machine learning techniques could be used to predict cyber-attacks through structured time-series data. The framework utilized feature selection methods which included Information Gain, and it used multiple regression and deep learning models with an LSTM model to enhance prediction results. The study succeeded in developing successful results but faced two main obstacles which included dataset limitations and the necessity to forecast results over extended periods. Shaukat et al. [2] studied different machine learning algorithms to show that different cyber threat types lead to distinct model performance outcomes which require research dataset variety and correct model choice.

Recent studies have investigated how cyber threats develop over time and how security measures work to prevent those threats. Al-Hawamleh [3] studied current cyber-attack patterns to discover common security threats which included malware, phishing, and ransomware attacks together with human mistakes that caused security breaches. The study underscored the essential nature of developing security systems which will defend against emerging cyber threats.

Researchers now use time-series analysis techniques to forecast upcoming cyber-attacks with greater frequency. Loumponias et al. [4] developed a machine learning framework which combines Random Forest and LSTM models to predict attacks through network traffic analysis. Their research demonstrated that LSTM models provide superior performance to traditional techniques when it comes to understanding time-dependent patterns, yet some attack types continue to present challenges for precise forecasting. Hakim and Wulandhari [5] created an LSTM prediction system which accurately predicts cyber-attack sequences while showing high accuracy for particular types of attacks.

Time-series analysis and data-driven methods have emerged as essential elements for cybersecurity analytics according to extensive research studies. Landauer et al. [6] examined several time-series analysis methods which included ARIMA statistical models along with RNN and LSTM deep learning frameworks, and they found multiple challenges which included handling high dimensional data, managing noisy information, and dealing with nonstandard data patterns. The results indicate that machine learning and temporal analysis should be combined to develop effective proactive cybersecurity systems; however, more research is necessary to enhance prediction accuracy for different types of cyber-attacks [6].

3. METHODOLOGY

A. System Architecture

Recent studies have explored the combination of machine learning with cybersecurity prediction systems. In the first stage of the research project, the emphasis is placed on the processing of data into a standard form that is amenable to analysis. In the second stage, the detection of patterns is achieved using feature generation along with ensemble learning to create a model of prediction. The third stage includes forecasting, model evaluation, and result visualization using the dashboard.

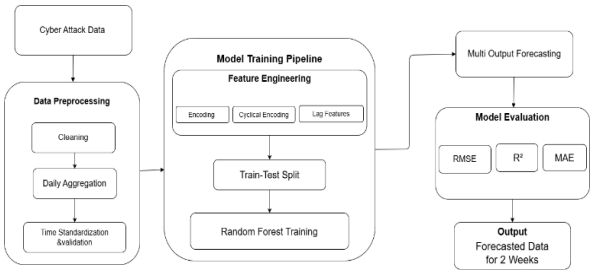

Fig -1: Overall System Architecture of the Proposed Cyber-Attack Forecasting Framework

Phase-1: Data Pre-Processing and Aggregation

Currently, the stage of the project involves transforming unstructured data of cyber-attack logs into structured data ready for analysis. The generated data, which include timestamp data, state data, type of attack data, and protocol data, face numerous accuracy issues due to lack of data, duplicating data, and inaccurate time data. The process of cleaning the data improves the quality of data using three main approaches.

The process of temporal standardization begins after data refinement because researchers extract date components and transform timestamps into a standard format. The dataset groups records by state and date to create daily aggregated data which is used to determine attack patterns and protocol usage. The 7-day rolling mean helps to reduce temporary changes while making it easier to see long-term trends. The dataset has been cleaned, time alignment has been achieved, and the dataset has been prepared for the upcoming modelling stage.

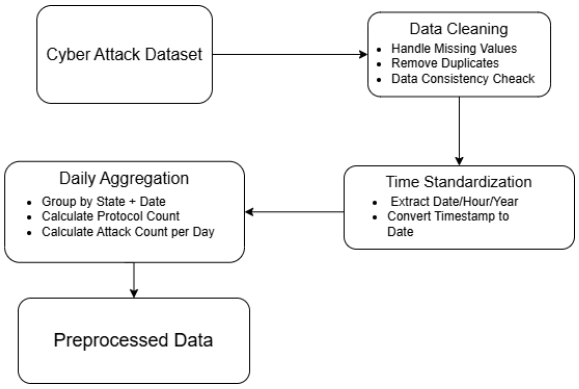

Fig -2: Cyber Attack Data Preprocessing and Aggregation Pipeline

Phase-2: Model Processing and Predictive Learning

The pre-processed dataset transforms into its feature-rich state which serves as the foundation for training a predictive model. First, the process enters its initial stage through feature engineering which enables the creation of new variables that improve the model's learning capability. The algorithm processes categorical attributes through encoding techniques which transform state identifiers into numerical representations. The system generates lag features (t-1, t-2, t-3) which enable the model to learn from prior attack patterns in order to establish time dependencies. Cyclical encoding applies to time-related attributes which enables the system to track periodic patterns that occur during cyber-attack events. The system implements a smoothing mechanism which decreases noise while stabilizing data fluctuations.

The dataset undergoes time-based partitioning after researchers complete their feature development work. The modelling training process utilizes historical data from 2020 to 2022 while the testing phase uses data from 2023 to assess how the model handles previously unencountered future data. To capture complex data relationships in cybersecurity, the training process implements a Random Forest multi-output regression model. The model achieves better performance through its decision tree combination which enhances system reliability while decreasing the risk of overfitting. The system acquires knowledge about how time elements, different regions, and previous attack patterns interact with each other during its training phase. The trained model generates predictions from new data in the testing phase which serve as the basis for the next forecasting step.

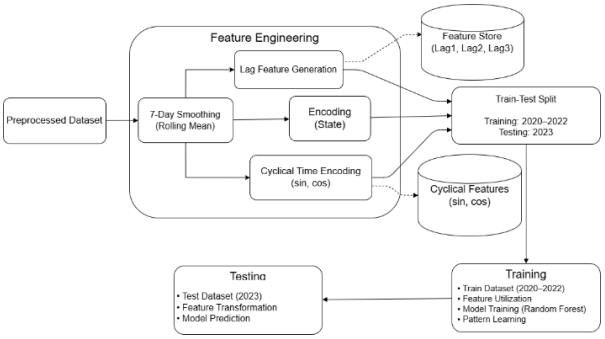

Fig -3: Feature Engineering and Model Training Workflow

Phase-3: Forecasting, Evaluation, and Visualization

The last stage of the project converts model results into practical online results through the development of an interactive display system. The process starts with a forecasting method that generates predictions for future time periods through its recursive prediction system which uses temporal intervals of 14 days. The system uses its initial predicted value to produce multi-step predictions through all following forecast periods. The system uses percentage normalization to convert its generated predictions into attack type and protocol usage distributions that show raw values as proportional distributions. The transformation process makes results easier to understand while allowing users to compare different areas of study.

Based on these normalized outputs, the system identifies the dominant attack category by selecting the class with the highest predicted probability at each time step. The performance of the regression model is measured using standard regression metrics. RMSE measures the size of the prediction error while MAE represents the average absolute difference between the prediction and the actual value. R-Squared measures how well the model captures the variation in the dataset. The formulas used for computing the metrics are as follows:

$$RMSE = \sqrt{[\ (1/n)\ \Sigma\ (y_i - \hat{y}_i)^2\ ]}$$

$$MAE = (1/n)\ \Sigma\ |y_i - \hat{y}_i|$$

$$R^2 = 1 - [\ \Sigma(y_i - \hat{y}_i)^2\ /\ \Sigma(y_i - \grave{y})^2\ ]$$

where y_i represents actual values, $\hat{y}_i$ denotes predicted values, $\grave{y}$ is the mean of actual values, and n is the number of observations.

In addition to numerical evaluation, visual analysis techniques such as scatter plots and line charts are used to examine prediction trends and deviations. The final results are presented through a web-based dashboard, where users can select a specific state and access forecast tables, pie chart distributions, temporal trend graphs, and weekly summaries for effective monitoring and analysis of cyber-attack patterns.

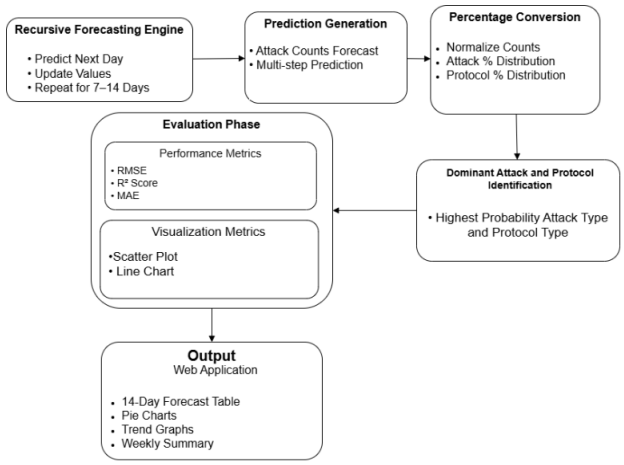

Fig -4: Forecasting, Evaluation, and Visualization Process

B. Algorithmic Workflow

Input: Structured cyber-attack dataset containing attributes such as Timestamp, State, Attack Type, and Protocol.

Output: State-wise cyber-attack forecasts, evaluation metrics, and interactive dashboard visualizations.

Step 1: Dataset Initialization and Organization

• Step 1.1: Load the processed dataset generated from the pre-processing phase.
• Step 1.2: Arrange records in chronological order based on state and time for sequential analysis.

Step 2: Feature Engineering and Data Transformation

• Step 2.1: Convert categorical attributes such as state into numerical representations using encoding techniques.
• Step 2.2: Generate lag-based features (t-1, t-2, t-3) to capture historical attack behaviour.
• Step 2.3: Apply cyclical encoding to temporal attributes to preserve periodic patterns.
• Step 2.4: Perform smoothing using a rolling mean to reduce noise and stabilize trends.

Step 3: Dataset Preparation for Modelling

• Step 3.1: Separate input features and target variables from the dataset.
• Step 3.2: Format the dataset to ensure compatibility with the machine learning model.

Step 4: Time-Based Data Partitioning

• Step 4.1: Allocate historical records (2020-2022) for training purposes.
• Step 4.2: Reserve recent observations (2023) for testing and validation.

Step 5: Model Training and Learning

• Step 5.1: Initialize the Random Forest algorithm with suitable parameters.
• Step 5.2: Train the model using engineered features and historical cyber-attack data.
• Step 5.3: Enable multi-output prediction to estimate multiple attack indicators simultaneously.

Step 6: Prediction and Recursive Forecasting

• Step 6.1: Generate predictions on the test dataset using the trained model.
• Step 6.2: Perform recursive forecasting to estimate future attack trends for the next 14 days.

Step 7: Post-Processing and Interpretation

• Step 7.1: Convert predicted values into percentage distributions for attack types and protocols.
• Step 7.2: Identify the dominant attack category based on the highest predicted probability.

Step 8: Model Evaluation and Analysis

• Step 8.1: Compute performance metrics such as RMSE, MAE, and R^2 to assess prediction accuracy.
• Step 8.2: Generate visual analyses including scatter plots and line charts to examine trends and deviations.

Step 9: Visualization and Deployment

• Step 9.1: Integrate results into a web-based dashboard interface.
• Step 9.2: Enable users to explore state-wise forecasts, graphical trends, and summarized insights.

4. EXPERIMENTAL EVALUATION

A. Setup

Assessment of the model is done using unseen data, which determines how well the system performs, as required for a thorough analysis across different datasets. The system determines whether it is efficient in predicting cyber-

attack trends in the future by conducting assessments in actual operational settings. The data used includes cyber-attack records at the state level, compiled through data cleansing and aggregation before being transformed into percentages. A time-based splitting technique is adopted because of the chronological order of the dataset. Training of the model uses data collected between 2020 and 2022, whereas model testing uses the most current data, that of 2023. The design of the system necessitates the use of future data to make predictions, hence creating an accurate evaluation process. The Random Forest model is used in a multi-output format, predicting different attack indicators simultaneously. Three evaluation metrics - RMSE, MAE, and R^2 - are used to assess the model.

B. Model Performance Analysis

Figure 5 shows the comparative assessment of three different regression algorithms, including Random Forest, Linear Regression (Ridge), and Support Vector Regression (SVR). The Random Forest model performs better than other models because it attains the minimum value of RMSE (0.1365), MAE (0.1072), and maximum R^2 (0.6921). On the contrary, Linear Regression and SVR models perform poorly due to high error values and low R^2 values. Generally, Random Forest is the best regression algorithm for the given dataset.

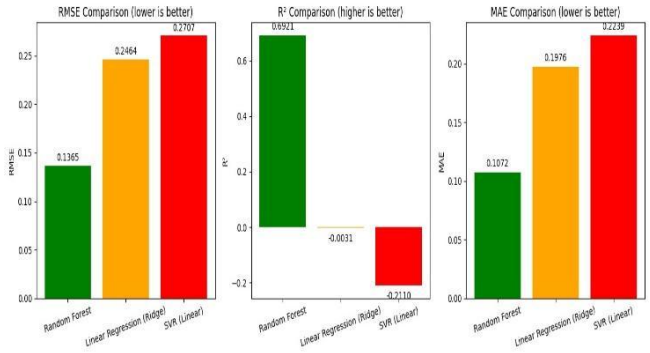

Fig -5: Performance Evaluation of the Proposed Prediction Model Using MSE, RMSE, and R^2 Metrics

Figure 6 illustrates the performance of the proposed predictive model using three key evaluation metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2). The model achieves a high R^2 value of approximately 0.69, indicating a strong ability to explain the variance in the data and good predictive accuracy. The RMSE value is relatively low, suggesting that the average prediction error is within an acceptable range. Additionally, the MSE value is minimal, reflecting a low squared difference between actual and predicted values. Overall, the results demonstrate that the model performs effectively with reliable and consistent predictions.

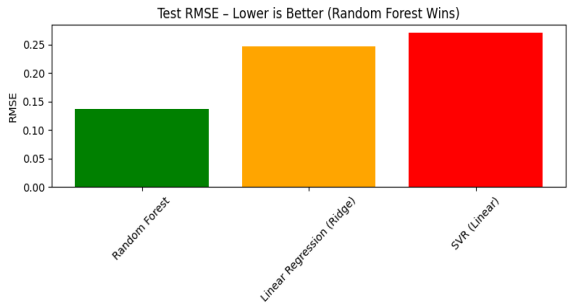

Fig -6: Performance Comparison of Machine Learning Models Using RMSE

C. Model Comparison

Table -1: Performance Evaluation of Proposed Model

Model	RMSE	R^2 Score	MAE
Random Forest	0.1365	0.6921	0.1072
Linear Regression (Ridge)	0.2464	-0.0031	0.1976
SVR (Linear)	0.2707	-0.2110	0.2239

This table presents the performance comparison of three regression models - Random Forest, Linear Regression (Ridge), and Support Vector Regression (SVR) - based on the evaluation metrics RMSE, R^2 Score, and MAE. The Random Forest algorithm outperformed all other models by achieving the lowest RMSE value of 0.1365 and MAE value of 0.1072, and the highest R^2 score of 0.6921, which indicates strong predictive accuracy and good model fit. The Linear Regression and SVR models demonstrate extremely poor performance because they produce negative R^2 values which indicate that these models fail to capture the underlying patterns in the data effectively. The Random Forest model proves to be the best option for this prediction task because it delivers higher accuracy and better reliability than other models.

5. SYSTEM OUTPUTS

As can be seen, the approach is capable of predicting the patterns of cyber-attacks. It maintains very small error values and produces results that match the observed data pattern movements. Its efficiency is confirmed since both qualitative and graphical analysis confirm its effectiveness. The system achieves consistent predictions through the combination of appropriate data preprocessing methods, feature engineering techniques, and Random Forest learning methods. The results demonstrate that the model

functions as an effective forecasting tool which organizations can use to predict cyber-attack trends and develop security measures.

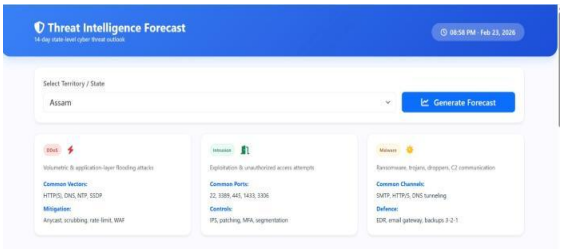

Fig -10: Cyber-attack Forecasting Dashboard Interface

Fig -11: Attack Type Trend Visualization Over the Forecast Period

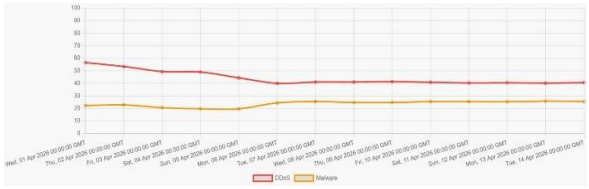

Fig -12: Declining Trend in DoS Attacks Over Time, While Malware Incidents Remain Relatively Stable

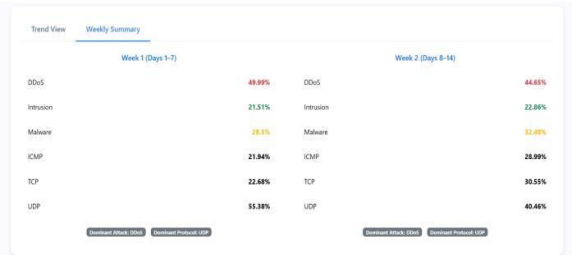

Fig -13: Weekly Summary of Predicted Attack and Protocol Distributions

6. CONCLUSION AND FUTURE WORK

The research developed a framework which uses machine learning to predict future cyber-attack patterns based on historical data. The system implemented a predictive model which enables forecasting of future attack patterns across multiple states to replace traditional security methods that only respond to existing threats. The framework successfully captured crucial time-based and location-based changes in cyber-attack patterns through its combination of data preprocessing, feature engineering, and the Random Forest model.

The experimental results show that the model performs reliably through its ability to maintain low error rates while achieving accurate results. The system demonstrates its ability to track actual attack patterns through both numerical metrics and visual analysis while producing valid predictive outcomes. The findings suggest that machine learning algorithms can be successfully employed in analyzing cybersecurity data in order to create effective early warning systems and improve planning efforts. However, there is room for improvement.

The next phase of research will concentrate on achieving two main goals: integrating real-time data and exploring the capabilities of deep learning algorithms and hybrid models to increase future forecasting accuracy. Moreover, the usability of the model will be improved by two strategies: adding an alert system and extending the dashboard. The research establishes a solid base for creating advanced cybersecurity systems which use real-time data to detect threats.

REFERENCES

[1] Ahmed, Y., Azad, M. A., & Asyhari, T. (2024). Rapid forecasting of cyber events using machine learning-enabled features. Information, 15(1), 36. https://doi.org/10.3390/info15010036

[2] Shaukat, N., et al. (2020). Machine learning techniques for cyber security: A survey. IEEE Access, 8, 128820-128835. https://doi.org/10.1109/ACCESS.2020.3007909

[3] Al-Hawamleh, A. M. (2023). Cyber-attack trends and security measures. International Journal of Information Security Science, 12(2), 45-56.

[4] Loumponias, K., et al. (2023). Machine learning-based cyber-attack forecasting. Future Internet, 15(6), 210. https://doi.org/10.3390/fi15060210

[5] Hakim, L., & Wulandhari, A. (2024). LSTM-based cyber threat prediction. Journal of Big Data, 11(1), 25.

[6] Landauer, M., et al. (2024). Time-series analysis techniques for cybersecurity analytics. ACM Computing Surveys, 56(4), 1-36.

[7] Loumponias, K., Raptis, S., Darra, E., Tsikrika, T., Vrochidis, S., & Kompatsiaris, I. (2023). Forecasting cyber-attacks to destination ports using machine learning. In Proceedings of the International Conference on Information Systems Security and Privacy (ICISSP) (pp. 757-764).

[8] Zhang, J., et al. (2019). Random forest-based network intrusion detection. IEEE Access, 7, 123456-123467.

[9] Wang, W., et al. (2019). Malware traffic classification using convolutional neural networks. IEEE Access, 7, 191033-191041. https://doi.org/10.1109/ACCESS.2019.2963706

[10] Hindy, H., et al. (2020). A taxonomy of network threats and intrusion detection datasets. IEEE Communications Surveys & Tutorials, 22(3), 2485-2524. https://doi.org/10.1109/COMST.2020.2992808