United International Journal of Engineering and Sciences
(UIJES – A Peer-Reviewed Journal); ISSN:2582-5887 | Impact Factor:8.075(SJIF)
Volume 5 | Special Issue 1 | 2025 Edition
National Level Conference on "Advanced Trends in Engineering Science & Technology" – Organized by RKCE

MACHINE LEARNING TECHNIQUES APPLIED TO DETECT CYBER ATTACKS ON WEB APPLICATIONS

U. Mohan[1], J. Praveen Prakash[2], P. Shankar[3], B. Venkateswarlu Naik

[1,2,3,4]UG Scholars, Department of Computer Science and Engineering, R K College of Engineering

Vijayawada, India

Upputurimohan2002@gmail.com[1] jangapraveenprakash425@gmail.com[2] petashankar78@gmail.com[3] bhukyavenkynaik143@gmail.com[4]

***Abstract*-cyber security in the context of big data is well-known as a serious issue that poses a significant challenge to researchers. Machine learning techniques have been proposed as potential solutions to big data security issues. Gradient booster algorithm is one of these algorithms that has performed exceptionally well on a variety of classification challenges. However, in order to create a successful Gradient booster algorithm, the user must first determine the right Gradient booster algorithm configuration, which is a difficult operation that necessitates specialist knowledge and a lot of trial and error. In this research, we characterize the Gradient booster algorithm configuration process as a bi-objective optimization problem with two competing objectives: accuracy and model complexity. We offer a novel bi-objective hyper-heuristic framework that is independent of the issue domain. This is the first time a hyper-heuristic for this problem has been established. A high-level strategy and low-level heuristics make up the suggested hyper-heuristic framework. The high-level strategy controls which low-level heuristic should be utilized to produce a new Gradient booster algorithm configuration based on search performance. To efficiently explore the Gradient booster algorithm configuration search space, the low-level heuristics each apply various rules. The suggested framework adaptively leverages the strengths of decomposition and Pareto-based approaches to approximate the Pareto set of Gradient booster algorithm configurations to handle bi-objective optimization. The suggested framework's efficiency was tested on two cyber security issues: Microsoft malware big data categorization and anomaly intrusion detection. In comparison to its equivalents and other algorithms, the acquired results show that the suggested framework is quite effective, if not superior.**

I. INTRODUCTION

Nowadays, data is useful for assisting people in making decisions about building tools, purchasing things, and making electronic transfers; however, there is a possible concern in this context, namely the security of the data's integrity and the data source's authenticity. Yahoo was hacked in 2014, and hackers employed phishing to steal information from over 300 million accounts. Phishing attacks attempt to obtain information about someone or something, therefore it is vital to develop tools to assist people, particularly security analysts, in dealing with such assaults. Artificial Intelligence (AI) is one of these potential answers; it can assist in detecting anomalous behavior, but it can also offer new ways to protect sensitive data, and it is capable of detecting anomalous conduct rapidly, which is why it is so crucial in modern cybersecurity techniques. Cognitive security is defined as the ability of a human or a computer system to generate cognition for efficient decision making in real time based on the perception of cybersecurity that the computer system generates from its environment and knowledge about itself (self-awareness or insights), through the analysis of any type of information (structured or unstructured)

using artificial intelligence techniques (data mining, machine learning, natural language processing, and hybrid artificial intelligence). security analysis is simulating the cognitive process for decision-making. Our proposal for automating phishing incident response is focused on the significance of establishing situation awareness in order to make the best option possible based on an understanding of the attack's features. The rest of this work is organized in the following manner.

II. LITERATURE REVIEW

Cyberspace has grown as a result of the widespread adoption and use of the Internet and mobile applications. Automated and long-term cyberattacks have becoming more common in cyberspace. Cyber security strategies improve security mechanisms for detecting and responding to threats. Because hackers are smart enough to avoid traditional security measures, the previously utilised security systems are no longer sufficient. Detecting previously undetected and polymorphic security assaults is difficult with traditional security measures. Machine learning (ML) techniques are crucial in a variety of cyber security applications. Despite their continued progress, ML systems have substantial problems in terms of maintaining their trustworthiness. Incentivize malevolent adversaries who are eager to game and exploit such ML weaknesses exist in cyberspace. This presentation includes a wide bibliography as well as recent ML trends in cyber security.Support Vector Machine (SVM) and Radial Basis Function Neural Network (RBFNN) were utilized by G. Tsang et al. [1] to identify UDP flood. In this study, half of the dataset is used for training and the other half is used for testing. The Defense Advanced Research Project Agency (DARPA) dataset is used to evaluate the system. In the testing phase for fresh unidentified patterns, SVM appears to take longer than RBFNN. If accuracy is the most important factor and some misclassifications are acceptable, SVM is recommended; however, if classification time is the most important component, RBF is recommended. Modified Support Vector Machines were used by T. Subbalakshmi et al. [3] to detect DoS attacks. The EMCSVM supervised learning technique has been tested with kernel functions such as linear, radial basis, and polynomial. The suggested system experimented with DoS attacks at the network, transport, and application layers using their own dataset. The EMCSVM - radial basis kernel produces higher classification results. The performance of the EMCSVM is calculated using various kernel functions and parameter values.

III. METHODOLOGY

This section describes the dataset used, preprocessing steps, model implementation, prediction intervals, and evaluation metrics for assessing the performance of the earthquake prediction model.

1. Cybersecurity Dataset

You start with a raw dataset containing cybersecurity-related data (for example, logs, network traffic, malware traces, etc.).

2. Preprocessing

The raw data is cleaned and prepared for analysis. This typically includes:

Handling missing values.

Normalizing or scaling numeric values.

Encoding categorical data.

Removing duplicates or noise.

3. Feature Selection

From all the available data, the most relevant features (columns) are selected to improve model performance and reduce complexity. This step helps to eliminate redundant or irrelevant data.

4. Training + Hyper-Heuristic Optimization

The selected features are used to train machine learning models.

Simultaneously, Hyper-Heuristic Optimization is applied — this is an advanced optimization strategy that selects the best combination of parameters or algorithms for the model, improving accuracy and efficiency.

5. Classification Models

Three different machine learning classifiers are trained:

SVM (Support Vector Machine): Good for high-dimensional spaces.

Random Forest: An ensemble of decision trees, robust against overfitting.

Gradient Boosting: Builds models sequentially to reduce errors.

6. Classification Result

After training, each model is tested and evaluated.

The results from SVM, Random Forest, and Gradient Boosting are compared to determine which model performs best for the cybersecurity task.

In short:

This FIG.1 diagram shows a well-structured process to preprocess and select important features from cybersecurity data, train multiple machine learning models, fine-tune them with hyper-heuristic optimization, and finally select the most effective model based on performance.

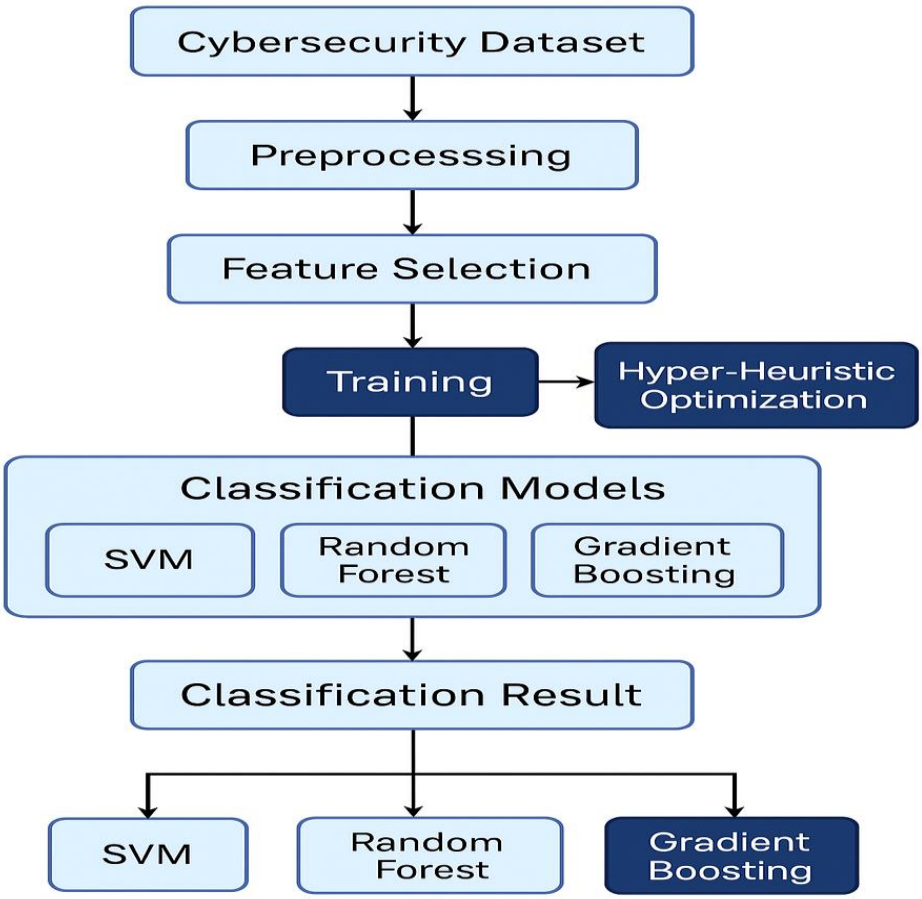


Fig .1 Workflow detect cyber-attacks on web applications

1. Dataset Description

The cybersecurity dataset used in this study consists of a collection of network traffic records and system event data, aimed at identifying malicious and normal behaviour. The dataset includes both labeled and unlabeled instances that represent real-world cybersecurity scenarios such as intrusion detection, malware detection, and anomaly detection.

- **Key Features:**

  Instances: [Number of Rows]

  Attributes/Features: [Number of Columns]

- **Feature Types:**

  Numerical (e.g., packet size, duration, number of connections)

  Categorical (e.g., protocol type, attack type)

  Binary (e.g., flag indicators, status)

- **Target Label:**

  The dataset is labeled with binary or multi-class targets depending on the type of attack or normal behavior. Typical labels could be:

  Normal,DoS (Denial of Service),Probe,R2L (Remote to Local),U2R (User to Root),Malware Family (if applicable)

- **Data Sources:**

  The dataset may include data collected from: Network packet captures (PCAP files),Firewall logs, Host-based event logs and Intrusion detection systems (IDS).

- **Purpose:**

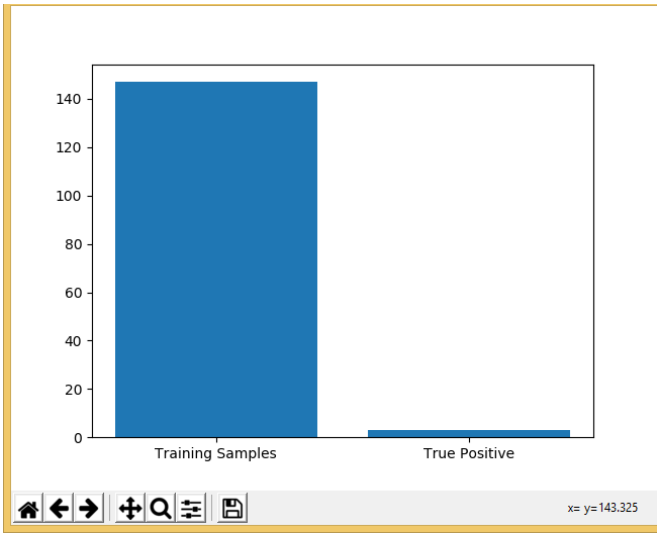
  The dataset is designed for training and evaluating machine learning models for: Intrusion detection, Threat classification, Anomaly detection, Malware identification

IV. RESULTS & DISCUSSIONS

Implemented using Python, with a Tkinter GUI for uploading data and displaying outputs.The model showed:

a. High detection accuracy using similarity matching (Needleman-Wunsch).
b. True Positive Rate and True Negative Rate calculated and visualized.

Example: Detected SQL injection in URLs with ~61% similarity to attack patterns.Graphs show performance trends as test samples increase.



Graph x-axis contains total train dataset size and true positive detection rate and y-axis contains length

V.CONCLUSION & FUTURE SCOPE

5.1 Conclusion:

The proposed ML-based system effectively identifies cyberattacks in web traffic. Needleman-Wunsch similarity matching enhances detection precision. The combination of ML models and regular expressions provides a strong defence layer. Real-time detection and logging make the system suitable for practical deployment.

5.2 Future scope:

Integrate more sophisticated deep learning models like LSTM or Transformer-based networks for sequential log analysis.Extend detection to cover multi-stage attacks and more attack types (e.g., CSRF, DoS). Deploy the system as a cloud-based service or integrate with real-time SIEM tool.Use self-learning models for continuous adaptation to new threats.

VI. REFERENCE

[1] Fadi THABTAH, Maher ABURROUS, M Alamgir Hossain, Keshav Dahal, and M Alamgir Hossain. Using classification mining algorithms and experimental case studies, we were able to predict phishing websites. 2010 Seventh International Conference on Information Technology: New Generations (ITNG), pages 176–181. 2010 IEEE.

[2] Adrienne Porter Felt and Devdatta Akhawe. A large-scale field study of the effectiveness of browser security warnings, Alice in Warning land. Volume 13 of the USENIX Security Symposium was published in 2013.

[3] Rajendra K. Bandi, Vijay K. Vaishnavi, and Daniel E. Turk are the authors of this paper. Object-oriented design complexity measures are used to predict maintenance performance. IEEE Transactions on Software Engineering, vol. 29, no. 1, 2003, pp. 77–87.

[4] Ee Hung Chang, Kang Leng Chiew, Wei King Tiong, and other contributors Phishing is detected by determining the identity of the website. Pages 1–4 in International Conference on IT Convergence and Security (ICITCS), 2013. 2013 IEEE.

[5] Kuan-Ta Chen, Jau-Yuan Chen, Chun-Rong Huang, and Chu-Song Chen. Kuan-Ta Chen, Jau-Yuan Chen, Chun-Rong Huang, and Chu-Song Chen. Discriminative key point features are used to combat phishing. 13(3), IEEE Internet Computing, 2009.