Grenze International Journal of Engineering and Technology, June Issue

Network Traffic Cyber Attacks Classification using Supervised Machine Learning Techniques

J. Geetha Priya[1], S.Srinivasan[2], Chintha Priyanka[3], Guna Sree.P [4] and Dodla Harini[5]
[1]Assistant Professor, Department of CSE, R.M.D. Engineering College, Chennai, India
Email: jgp.cse@rmd.ac.in
[2]Professor, Department of CSE, R.M.D. Engineering College, Chennai, India
Email: ssn.cse@rmd.ac.in
[3-5]Student, Department of CSE, R.M.D. Engineering College, Chennai, India
Email: {ucs20124, ucs20141, ucs20135}@rmd.ac.in

***Abstract*— Cyber attack classification through the utilization of supervised machine learning methods is investigated in this study. The system is designed to categorize diverse cyber-attacks by employing a meticulously curated dataset encompassing a wide array of attack types, including but not restr to malware, phishing, and (DDoS) attacks. Feature extraction techniques are applied to both network traffic data and behavioral attributes, facilitating the training of a robust classification model. Various supervised learning algorithms, such as the Adaboost classifier, Catboost classifier, and Gaussian Naïve Bayes, are evaluated for their efficacy in accurately predicting attack categories. The training process involves labelling historical attack instances, enabling the model to discern intricate patterns and subtle differentiators among attack types. Regular model updates and retraining with new attack data ensure its relevance in dynamically evolving threat landscapes. The system's predictive accuracy empowers cyber security teams to swiftly identify and respond to cyber threats, thereby bolstering overall defence strategies. Through this research, I contribute to the proactive identification and mitigation of cyber-attacks, ultimately fortifying digital security frameworks.**

I. INTRODUCTION

The classification of cyber-attacks through supervised machine learning techniques stands as a critical component within contemporary cyber security frameworks. As the digital landscape continues to evolve in complexity, the proliferation of cyber threats escalates in both sophistication and frequency. Consequently, the imperative to promptly and accurately categorize these threats becomes paramount. Supervised machine learning emerges as a potent solution, harnessing labeled datasets to instruct algorithms in recognizing and classifying various cyber-attack types. This classification capability equips organizations with the means to mount effective responses, thereby mitigating potential damages and reinforcing defensive postures against future incursions.

However, the realm of cyber attack classification presents multifaceted challenges. These include the diverse array of attack methodologies employed, the adaptive nature of adversaries, and the prevalence of imbalanced data distributions. The dynamic and multifarious tactics employed by cybercriminals render the development of comprehensive detection systems arduous, hindering the creation of universally effective defenses. Furthermore, the perpetual evolution of attackers' strategies necessitates ongoing refinement and enhancement of detection algorithms. Moreover, imbalanced datasets, wherein certain attack types predominate over others, pose a significant hurdle to traditional machine learning methodologies. Biases introduced by skewed data distributions

Grenze ID: 01.GIJET.10.2.84
© Grenze Scientific Society, 2024

can compromise the accuracy and efficacy of classification models, impacting their overall performance.
Nevertheless, the potential applications of supervised machine learning in cyber security are extensive, spanning from intrusion detection and email filtering to malware identification and anomaly detection. These applications underscore the indispensable role played by machine learning algorithms in safeguarding digital infrastructures against malevolent activities. Looking forward, ongoing research endeavors are directed towards refining machine learning models to effectively counter evolving threats. This includes integration with broader security strategies, such as threat intelligence sharing and incident response protocols. Additionally, addressing ethical considerations surrounding the deployment of machine learning technologies in cyber security is imperative to ensure their responsible and ethical utilization in safeguarding digital assets and preserving privacy.

II. Literature Survey

1. Preetish Ranjan etal [5] proposed a Apriori Viterbi Model for Prior Detection of Socio Technical Attacks in a Social Network .presents an innovative strategy to counteract harmful activities within social platforms. By amalgamating the powerful Apriori algorithm for detecting patterns with the adaptable Viterbi algorithm for analyzing sequences, this model offers proactive defense against socio technical attacks. Utilizing these algorithms, the model identifies suspicious behavior patterns and sequences of events that may indicate potential threats, allowing for timely intervention. Through extensive testing, the efficacy of the proposed model is validated, highlighting its potential to enhance security measures within social network ecosystems.
2. Seraj Fayyad etal [3] proposed a "New Attack Scenario Prediction Methodology" project aims to develop a cutting-edge approach to anticipate emerging cyber threats. By analyzing diverse datasets using advanced machine learning and anomaly detection techniques, the project seeks to forecast potential attack scenarios before they occur. Integrated with threat intelligence feeds, the methodology enhances predictive capabilities and enables proactive defense strategies. Through rigorous validation, the project aims to empower organizations with actionable insights to stay ahead of evolving cyber threats.
3. Shamsun Nahar Edib etal [3] proposed a The "Cyber Restoration of Power Systems" project focuses on developing strategies to quickly recover power systems from cyber attacks. It involves designing resilient grid architectures, implementing effective incident response protocols, deploying advanced monitoring technologies, ensuring secure communication networks, providing training, and fostering collaboration for enhanced cyber resilience in the power sector.
4. Wentao Zhao etal [4] The "Prediction Model for DoS Attack's " focuses on developing a model to predict the distribution of Denial of Service (DoS) attacks using discrete probability methods. This research aims to provide insights into the likelihood and patterns of DoS attacks occurring within a network or system. By analyzing historical data and employing discrete probability techniques, the model seeks to forecast the probability distribution of future DoS attacks, enabling proactive measures to mitigate their impact and enhance network security.
5. Pan He etal [2] Proposed a Attacks and Defenses for Deep Learning" explores methods of manipulating input data to deceive deep learning models, along with strategies to defend against such attacks. This research aims to enhance the security and robustness of deep learning systems against adversarial threats.

III. Existing Work

The application of invariants in the development of security mechanisms has garnered significant attention within the realm of exploration, primarily due to their eventuality to help and descry attacks CPS. In substance, an steady represents a property expressed using design parameters alongside Boolean drivers, which constantly works under normal operation of system, particularly within CPS surroundings. These invariants can be deduced through the analysis of functional data about colorful parameters in an functional CPS, or by verifying the conditions and documents. Both approaches parade substantial implicit in detecting and precluding cyber-attacks targeting CPS surroundings. The generation of design- driven invariants generally requires significant homemade intervention. In this paper, the focus is on expounding the limitations of data- driven unvariying by showcasing a series of inimical attacks targeting similar invariants.
To address these failings, I propose a result strategy aimed at detecting similar attacks by supplementing data-driven invariants with design- driven counterparts. The efficacity of this approach is validated through a series of trials carried out on a real- world hydro theraphy test bed. The results demonstrate that this mongrel approach can effectively alleviate false cons while achieving high delicacy in detecting attacks within CPSs.

IV. Proposed Work

The proposed system presents a comprehensive approach to cyber attack classification through the application of supervised machine learning techniques. An extensive and diverse dataset covering various cyber attack types is compiled, capturing their distinct characteristics. Relevant features are extracted from network traffic, system logs, and attack patterns to construct a robust feature set. Utilizing supervised learning algorithms such as the Adaboost classifier, the Catboost classifier, and Gaussian Naive Bayes, I train a classification model. This model undergoes refinement using labelled historical data, enabling accurate categorization of incoming cyber threats.
Real-time network monitoring facilitates the deployment of the model, enabling swift analysis of ongoing activities and flagging potential attacks. Regular updates and continuous retraining maintain the model's efficacy against evolving attack methodologies. Ultimately, the system enhances cyber defence by providing rapid, accurate, and proactive identification of cyber attacks, empowering timely response and mitigation strategies.

V. Working Process

- Download and install Anaconda and acquire the most useful packages for machine learning in Python.
- Load a dataset and comprehend its structure using statistical summaries and data visualization.
- Implement machine learning models, select the best, and ensure confidence in the reliability of accuracy.

Work Flow Diagram

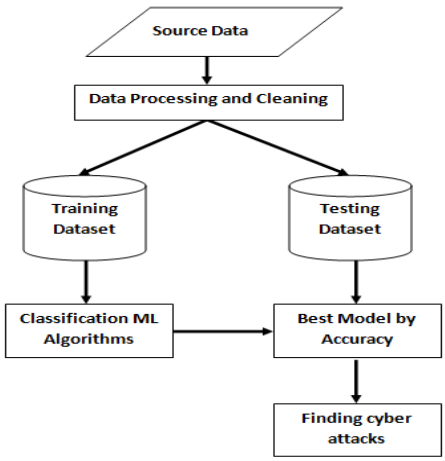


Fig : Work Flow Diagram

Data Set

Our attack prediction process is based on real-time data, which is followed by data preprocessing and data visualization. Data will be sent to a computer using an algorithm to obtain the highest accuracy. The most accurate algorithm will be turned into an AI model once all three have been compared.

VI. List Of Modules

1. Data Pre-processing
2. Data Analysis of Visualization
3. Adaboost Classifier
4. Catboost Classifier
5. Gaussian Naive Bayes
6. Deployment

VII. Module Description

A. Data Pre-processing

Confirmation ways in machine literacy are essential for estimating the error rate of a Machine literacy(ML) model, which approximates the true error rate of the dataset. While large datasets may represent the population adequately,

real- world scripts frequently involve working with samples that may not be completely representative. Addressing missing and indistinguishable values and describing data types(e.g., pier or integer) are pivotal way in datapre-processing. Incorporating confirmation data into model configuration can introduce bias, hence frequent evaluation is necessary. Understanding data parcels aids in opting applicable algorithms for model structure. Python's panda library facilitates colorful data drawing tasks, with a focus on handling missing values efficiently. Data can be missing due to arbitrary miscalculations, programming crimes, or deliberate elision by druggies. Understanding the reasons for missing data influences the approach to handling them, including introductory insinuation and statistical styles. Variable Recognizable evidence with Uni-variate,Bi-variate, andMulti-variate Examination Consequence vital libraries and pored the dataset. dissect the general parcels of the dataset. Display the dataset as a data frame. Show columns, shape, and describe the data frame. Check data types and information. Identify and handle indistinguishable data. Address missing values. Examine unique and count values. Brand or drop data frame columns. Specify value types and produce fresh columns. These way form the foundation of datapre-processing and variable identification, pivotal stages in preparing data for machine literacy analysis.

B. Data visualization

It provides tools for qualitative understanding. Exploring datasets visually can aid in identifying patterns, outliers, and other characteristics that may not be apparent through quantitative analysis alone.

Visualizations offer a visceral representation of data, making it easier for stakeholders to grasp key relationships and insights. With domain knowledge, visualizations can effectively convey complex information in a more intuitive manner than traditional statistical measures.

Exploratory data analysis (EDA) and data visualization are entire fields in themselves, with numerous techniques and tools available. Mastery of these skills is essential for anyone working with data, whether in statistics or machine learning.

In applied statistics and machine learning, the ability to quickly and effectively visualize data samples is invaluable. Python offers a wide range of plotting libraries and tools, allowing practitioners to create various types of plots and charts to better understand their data.

Understanding different plot types and how to use them is crucial for effective data visualization. By mastering these techniques, practitioners can gain deeper insights into their data and communicate findings more effectively to stakeholders.

For further exploration of data visualization and exploratory data analysis, it is recommended to delve into additional resources such as books dedicated to these topics.

C. Algorithm implementation

It is crucial to consistently compare the performance of diverse machine learning algorithms, and creating a test harness in Python with sci kit-learn serves as a valuable template for this purpose. This template can be adapted for various machine learning problems, allowing the inclusion of algorithms for comparison. Every model exhibits different performance characteristics, and by employing re sampling techniques one can estimate their accuracy on unseen data. These estimates are instrumental in selecting the one or two best-performing models from the array created. When faced with a new data set, exploring data through different visualization techniques becomes imperative to gain varied perspectives. This principle extends to model selection, where employing various methods to depict accuracy, variance, and other properties aids in discerning the distribution of model accuracies. In the following section, you will learn precisely how to execute this process using Python and Sci kit-learn. The crucial element for impartially comparing machine learning algorithms lies in ensuring that each algorithm undergoes evaluation consistently on identical data sets. This uniformity is established by mandating that each algorithm be assessed within a standardized test harness. This approach guarantees that every algorithm is appraised using the same data set under identical conditions, thereby enabling a fair and objective comparison of their respective performances.

The below 3 different algorithms are compared:

- Adaboost Classifier
- Catboost Classifier
- Gaussian Naive Bayes

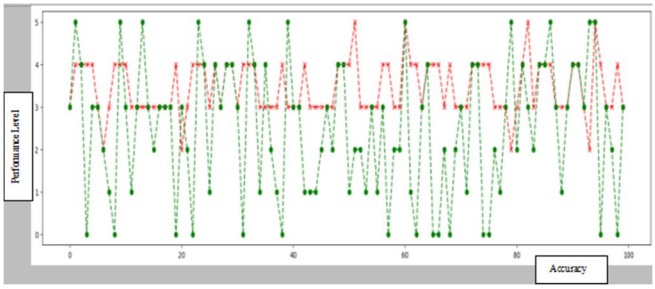
Adaboost Classifier:

The AdaBoost classifier is a meta-estimator, initiating by a classifier on the original data and subsequently fitting additional copies on the same data set. However, the weights of incorrectly classified instances are adjusted, emphasizing challenging cases for subsequent classifiers. This adaptive approach allows AdaBoost to boost the

performance of various machine learning algorithms, particularly when employed with weak learners—models that achieve accuracy only slightly above random chance in classification problems. The algorithm is notably effective when paired with decision trees of minimal depth.

The AdaBoost algorithm functions on the sequential growth of learners, where each subsequent learner is built upon the knowledge of previously trained ones. In essence, weak learners evolve into stronger ones over the course of sequential learning. This boosting technique is characterized by a unique approach, enhancing the focus on instances that previous learner found challenging.

In terms of its fundamental operation, given input data, the AdaBoost algorithm aims to produce an output that signifies the achieved accuracy. The essence of AdaBoost lies in its ability to iteratively refine the learning process, emphasizing instances that prove difficult for the model, ultimately enhancing the overall predictive accuracy.



The accuracy score of adaboost classifier is: 37.540661881334515

Catboost Algorithm

CatBoost stands out as a gradient-boosting algorithm designed for supervised machine learning tasks, focusing on classification and regression challenges. Renowned for its impressive accuracy, efficient treatment of categorical features, and resistance to over fitting, CatBoost employs a unique approach.

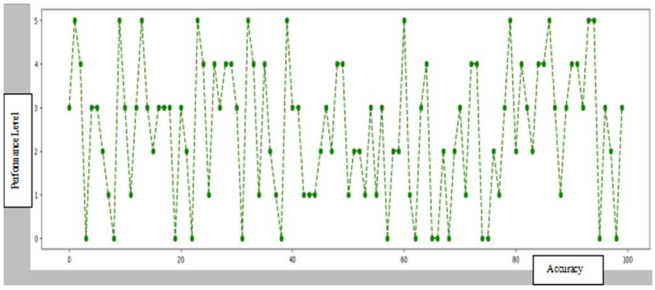
A distinctive feature of CatBoost lies in its adept handling of categorical features, eliminating the need for extensive preprocessing. Leveraging "ordered boosting," it assigns numerical values to categories based on target variable statistics during training, seamlessly incorporating categorical information.

The construction of decision trees in CatBoost follows a depth-first approach, optimizing a loss function to determine the best split points for each tree node and employing techniques like leaf-wise splits and oblivious trees for enhanced efficiency. To combat overfitting, the algorithm incorporates built-in regularization techniques, utilizing L2 regularization on leaf values and feature combinations.

A learning rate shrinkage strategy contributes to model stability and generalization by controlling the impact of each new tree on the ensemble, mitigating overfitting risks. CatBoost also offers solutions for handling imbalanced datasets through class weights adjustment or custom loss functions.

For preventing overfitting, CatBoost incorporates an early stopping mechanism based on holdout validation data, halting training when further iterations may lead to overfitting. Post-training, the model is ready for predictions, providing class probabilities or labels for classification tasks based on majority class determination from the trees' leaf nodes.

With an array of tunable hyper parameters like tree depth, learning rate, and regularization strength, CatBoost allows fine-tuning for optimal performance. Notably, its scalability ensures efficiency with large datasets and complex models, enabling parallelization for faster training. In essence, CatBoost stands as a robust gradient-boosting algorithm, particularly valuable for handling real-world classification and regression tasks with precision and efficiency.



The accuracy score of cat boost classifier is: 99.85213861332903

Gaussian Naive Bayes

Gaussian Naive Bayes, a probabilistic classification algorithm rooted in Bayes' theorem, proves invaluable for addressing classification and prediction challenges. Despite its apparent simplicity and the inherent "naive" assumptions it makes, the algorithm consistently demonstrates robust performance across various applications.

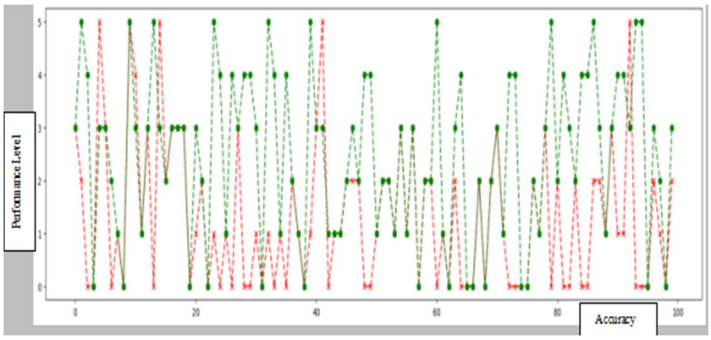
Central to the Naive Bayes algorithm is Bayes' theorem, a fundamental probability concept relating the conditional probability of an event A given an event B. In the context of classification, this translates to A representing the class label to predict and B signifying the features or attributes of the data.

The algorithm's "naive" characteristic arises from the assumption of conditional independence among features, given the class label. This implies that each feature contributes independently to the probability of an instance belonging to a specific class, simplifying calculations and ensuring computational efficiency.

Training a Naive Bayes classifier necessitates labeled data where both class labels and features are known. The algorithm computes two sets of probabilities: Class Probabilities (Prior Probabilities) derived from the frequency of each class in the training data, and Conditional Feature Probabilities (Likelihoods) representing the probability of each feature given a class, estimated from the training data.

During prediction, when presented with a new data point containing features but lacking a class label.The predicted class label is assigned based on the highest conditional probability.

In essence, Naive Bayes provides an effective and efficient solution to classification tasks, leveraging probabilistic principles to deliver accurate predictions. The algorithm's applicability extends across diverse domains, making it a valuable tool in data science and machine learning endeavors.



The accuracy score of gaussian naive bayes classifier is: 56.81909831438019

D. Deployment

Django (Web FrameWork) :

Django is a micro web framework designed in Python, recognized for its classification within the micro-framework category owing to its minimalistic approach that avoids the necessity for specific tools or libraries. A distinctive feature is its lack of a built-in database abstraction layer, form validation, and other standard components, as it opts not to depend on preexisting third-party libraries for common functionalities. Despite these simplified attributes, Django exhibits a versatile architecture that accommodates extensions. This distinctive feature enables the integration of additional application features with ease, creating a seamless experience as if these features were originally inherent components of Django itself.

VIII. Conclusion

The analytical way commenced with comprehensive data cleaning and processing, addressing issues such as missing values, followed by an exploratory analysis. Subsequently, the focus shifted towards model building and evaluation. The goal was to identify the most accurate algorithms by testing them on a public test set, ultimately selecting the one with the highest accuracy score. This chosen algorithm is then integrated into the application, serving as a valuable tool for identifying and categorizing different types of cyber attacks.

References

[1] E. Al Neyadi, S. Al Shehhi, A. Al Shehhi, N. Al Hashimi, Q. Mohammad, S. Alrabaee, "Discovering public Wi-Fi vulnerabilities using Raspberry Pi and Kali Linux," 2020 12th Annual Undergraduate Research Conference on Applied Computing (URC), IEEE, pp. 1-4 (2023).

[2] International Telecommunication Union, "Measuring digital development: Facts and figures," Accessed on June 25, 2023.

[3] Azab, "Packing resistant solution to group malware binaries," Int. J. Secur. Network., vol. 15, no. 3, pp. 123-132 (2022).
[4] H. Mohajeri Moghaddam, "Skypemorph: Protocol Obfuscation for Censorship Resistance," Master's Thesis, University of Waterloo (2022).
[5] Chabaa1, A. Zeroual1, J. Antari1, "Identification and Prediction of Internet Traffic Using Artificial Neural Networks" (Scientific Research, 2021) Vol. 2, pp. 147-155.
[6] C.F. Hai-liang, L. Q. Chen, "Multi-scale Internet Traffic Prediction Using Wavelet Neural Network Combined Model" (IEEE, 2009) ISBN: 1-4244-0462-2.
[7] Qin, X. Quan, W. Li, P. Wang, "Monitoring Malicious Traffic Flow Based on Independent Component Analysis" (IEEE, 2009) ISBN: 978-1-4244-3434-3.
[8] N. C. Anand, C. ScoglioS, B. Natarajan, "GARCH Non-Linear Time Series Model for Traffic Modeling and Prediction" (IEEE, 2008) ISBN: 978-1-4244-2065-0.
[9] C. Quang, G. Jian, D. Wei, "A Time series Decomposed Model of Network Traffic" (Springer, 2005) ISBN: 338-345.
[10] E. S. Yu, C.Y.R. Chen, "Traffic Prediction Using Neural Networks" (IEEE, 1993) ISBN: 0-7803-0917-0.