JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Plagiarism Checker X Originality Report

AUTHOR NAME: B VALARMATHI

Statistics: 350 words Plagiarized / 3483 Total words

Remarks: Low Plagiarism Detected - Your Document needs Optional Improvement.

INTRUSION DETECTION SYSTEM USING VOTING BASED NEURAL NETWORK

ABSTRACT With the rapid evolution of the Internet, the landscape of cyber-attacks is constantly changing, leading to a rather pessimistic outlook on cyber security. This article delves into the realm of network analysis for intrusion detection, specifically focusing on the implementation of Machine Learning (ML) and Deep Learning (DL) techniques. A comprehensive tutorial description is provided for each ML/DL method, accompanied by an examination of relevant research papers.

These papers were meticulously indexed, read, and summarized based on their temporal or thermal correlations. Given the paramount importance of data in ML/DL methods, the article also sheds light on commonly utilized network datasets within this domain. Furthermore, it addresses the challenges associated with employing ML/DL for cyber security and offers valuable suggestions for future research directions.

Notably, the KDD data set emerges as a well-established benchmark in the field of Intrusion Detection techniques. Extensive efforts are being made to enhance intrusion detection strategies, with equal emphasis placed on the quality of data used for training and testing the detection model.

This project undertakes a comprehensive analysis of the KDD data set, specifically focusing on four distinct attribute classes: Basic, Content, Traffic, and Host. To categorize these attributes, the Modified Random Forest (MRF) approach is employed. Keywords: Intrusion Detection, Feature Selection, Machine Leaning 1. INTRODUCTION In today's digital era, ensuring the security of computer networks and data has become of utmost importance.

Given the increasing complexity of cyber threats and the interconnected nature of our systems, the necessity for robust network intrusion detection systems (NIDS) has never been more critical. Intrusion detection plays a pivotal role in safeguarding organizations by identifying unauthorized access and mitigating potential threats to information systems.

However, traditional intrusion detection methods often encounter difficulties in adapting to the ever-evolving threat landscape. To overcome these challenges and enhance the effectiveness of intrusion detection, we propose an innovative approach called "Network Intrusion Detection with Two-Phased Hybrid Ensemble Learning and Automatic Feature Selection."

This research endeavors to combine cutting-edge techniques from the fields of machine learning, data science, and cybersecurity. By integrating the power of ensemble learning and automatic feature selection into a two-phased detection system, our aim is to revolutionize the field of network intrusion detection. 1.1 FEATURE SELECTION In today's ever-expanding digital landscape, the security of networks and information systems has

become a top priority.

With the rise of cyber threats, ranging from sophisticated malware to advanced persistent threats, it is crucial to continuously evolve network intrusion detection systems (NIDS) to prevent unauthorized access and malicious activities. The key to effective NIDS lies in selecting the most relevant data attributes, also known as "features."

Feature selection is a critical process in machine learning and data analysis, with the primary objective of identifying and retaining informative attributes while discarding irrelevant or redundant ones. In the context of network intrusion detection, the careful selection of features is essential to improve both the efficiency and accuracy of the detection process.

/ Figure 1. Feature Selection 2. LITERATURE REVIEW Felix Obite [1]et.al. has proposed in this paper that the significant growth in Internet traffic confirms the shift of the telecommunications back bone from a time division multiplexing (TDM) orientation to a focus on Ethernet solutions.

Ethernet PON, which combines low-cost Ethernet and fiber infrastructures, has emerged as the dominant technology in a market previously dominated by DSL and cable modems. This new technology is characterized by its simplicity, affordability, and scalability, enabling the delivery of large amounts of data services to end-users over a single network.

The paper provides a review of the evolution of EPON, with a particular focus on the current development process of future high-data-rate access networks such as NG-PON2, WDM PON, and OFDM PON. Additionally, the recently concluded 100G-EPON is examined to highlight the latest advancements in the field. By offering a comprehensive and up-to-date review, the paper aims to equip network operators and interested practitioners with the necessary knowledge to prioritize and plan their activities. Furthermore, the study aims to identify technical solutions for future investigation.

The increase in data traffic and the growing number of online users, who spend more time online and use bandwidth-intensive applications, necessitate broadband services that can support high-speed internet transmission. This is expected to contribute to economic improvement. Therefore, future access networks must possess large bandwidth and mobility capabilities to accommodate these new and real-time broadband applications.

In recent years, there has [2] been a significant focus on the fifth generation of wireless broadband connectivity, commonly known as '5G', which is currently being deployed by Mobile Network Operators. However, surprisingly, there has been less attention given to 'Wi-Fi 6', the new IEEE 802.1ax standard in the Wireless Local Area Network technology family, specifically designed for private, edge-networks. In this paper, Edward J. Oughtonet.al.

revisits the suitability of both cellular and Wi-Fi technologies in delivering high-speed wireless Internet connectivity. Both cellular and Wi-Fi technologies aim to provide enhanced performance, enabling faster wireless broadband connectivity and supporting the Internet of Things and Machine-to-Machine communications. As a result, these technologies can be seen as technical substitutes in various usage scenarios.

The authors conclude that both technologies will play important roles in the future, serving as competitors and complements simultaneously. It is anticipated that 5G will remain the preferred technology for wide-area coverage, while Wi-Fi 6 will continue to be the preferred choice for indoor use due to its lower deployment costs. Recently, Somayye Hajiheidari [3] et.al.

proposed a system that introduces a new dimension of intelligent objects by reducing the power consumption of electrical appliances. This system upgrades daily physical objects by incorporating electronic devices and connecting them to the Internet, enabling local intelligence and communication with cyberspace. This concept is known as the Internet of Things (IoT), which refers to the network of interconnected objects.

However, due to the direct connection of IoT objects to the Internet, they are vulnerable to attacks from malicious individuals. These attacks, known as internal attacks, exploit the resource constraints of IoT devices to infect internal nodes and carry out attacks on the network. Therefore, the importance of Intrusion Detection Systems (IDSs) in the IoT cannot be overstated.

Despite this significance, there is a lack of comprehensive and systematic reviews discussing and analyzing the mechanisms of IDSs in the IoT environment. To address this gap, this paper presents a Systematic Literature Review (SLR) of IDSs in the IoT. The paper provides detailed categorizations of IDSs based on their approach (anomaly-based, signature-based, specification-based, and hybrid), architecture (centralized, distributed, hybrid), evaluation method (simulation, theoretical), and types of attacks they detect (denial of service attack, Sybil attack, replay attack, selective forwarding attack, wormhole attack, black hole attack, sinkhole attack, jamming attack, false data attack).

In the realm of cybersecurity research, the ensemble of classifiers [4], also known as an ensemble learner, has garnered significant attention, particularly in the domain of intrusion detection systems (IDSs). IDSs are crucial in preventing cyberattacks, and to enhance their detection capabilities, there is a need to design an improved detection framework, especially when utilizing ensemble learners. The design of an ensemble often poses two main challenges, namely the selection of available base classifiers and combiner methods.

This paper presents an overview of how ensemble learners are utilized in IDSs through a systematic mapping study. The study involved the collection and analysis of 124 prominent publications from existing literature, which were then categorized based on years of publication, publication venues, datasets used, ensemble methods, and IDS techniques.

Additionally, the study reports and analyzes an empirical investigation of a new classifier ensemble approach, called stack of ensemble (SoE), for anomaly-based IDS. The SoE is an ensemble classifier that adopts a parallel architecture to combine three individual ensemble learners, namely random forest, gradient boosting machine, and extreme gradient boosting machine, in a homogeneous manner.

The performance of classification algorithms is statistically examined in terms of their Matthews correlation coefficients, accuracies, false positive rates, and area under ROC curve metrics. This study fills the gap in current literature by providing an up-to-date systematic mapping study and an extensive empirical evaluation of recent advances in ensemble learning techniques applied to IDSs. Muhamad Erza Amina [5] et al.

have presented a system that addresses the security challenges posed by the widespread use of IoT-enabled devices in our daily lives, which is a result of recent advancements in mobile technologies. The main concern lies in the open nature of wireless networks, such as Wi-Fi, which makes them vulnerable to impersonation attacks. These attacks involve an adversary disguising themselves as a legitimate party within a system or communications protocol.

The pervasiveness of connected devices leads to the generation of large-scale, high-dimensional data, making simultaneous detections complex. However, the study proposes a novel approach called Deep-Feature Extraction and Selection (D-FES) to overcome this challenge. D-FES combines stacked feature extraction and weighted feature selection techniques.

By utilizing stacked auto encoding, meaningful representations are obtained by reconstructing relevant information from raw inputs. This is then combined with modified weighted feature selection inspired by a shallow-structured machine learner. The effectiveness of the proposed D-FES is demonstrated through experimental results on the Aegean Wi-Fi Intrusion Dataset (AWID), a well-referenced Wi-Fi network benchmark dataset. The results showcase a remarkable detection accuracy of 99.918% and a false alarm rate of 0.012%.

These findings establish the proposed D-FES as the most accurate method for detecting impersonation attacks reported in the literature. Additionally, the condensed set of features derived from D-FES not only reduces the bias of machine learning models but also minimizes computational complexity. 3.

RELATED WORK Traditional firewalls and data encryption techniques are no longer sufficient to meet the demands of modern network security due to the rise in both the quantity and variety of network threats. Consequently, to address network threats, intrusion detection systems have been proposed. Although machine learning helps the existing mainstream intrusion detection algorithms, they still have issues with low detection rates and a high feature engineering overhead. This paper proposes a deep learning model for network intrusion detection (DLNID) to address the problem of low detection accuracy.

It does this by combining an attention mechanism with a bidirectional long short-term memory (Bi-LSTM) network, which extracts sequence features of data traffic first using a convolutional neural network (CNN) network, then uses the attention mechanism to reassign the weights of each channel, and finally uses Bi-LSTM to learn the network of sequence features. In public data sets for intrusion detection, there are typically significant data imbalances.

This paper addresses problems with data imbalance by using adaptive synthetic sampling (ADASYN) to expand minority class sample sizes and create a relatively symmetric dataset. It also reduces data dimensionality using a modified stacked auto encoder, which improves information fusion. .4. METHODOLOGY The MODIFIED RANDOM FOREST (MRF) algorithm is utilized by the proposed network intrusion detection system (IDS) to classify network traffic as either normal or malicious.

The system categorizes data attributes in the KDD dataset into four categories: Basic, Content, Traffic, and Host, and trains a MRF classifier on each category. The system collects network traffic data from the monitored network, preprocesses it to extract relevant features, and feeds these features into the MRF classifiers. The classifiers produce a prediction for each data point, and the system takes the average prediction to make a final prediction.

The proposed methodology draws inspiration from negative selection-based detection generation and is evaluated using the NSL-KDD dataset, a modified version of the widely used KDD CUP 99 dataset. Additionally, the system's adaptability and flexibility are increased by automatically selecting parameter values based on the training dataset used. A.

Probability Model In this module, we preprocess the probability model utilized to capture a user's normal mentioning behavior and the training of the model. We define a post in a social network stream by its mention count, denoted ask, and the set V of mentioned users' names (IDs). Two types of infinity must be considered in this context. The first type pertains to the number of users mentioned in a post, denoted ask.

While it is impractical for a user to mention hundreds of other users in a single post, we aim to avoid imposing an artificial limit on the number of mentioned users. Instead, we adopt a geometric distribution and integrate out the parameter to eliminate any implicit limitation. The second type of infinity relates to the number of users that can potentially be mentioned.

To prevent constraining the number of possible mentions, we employ the Chinese Restaurant Process (CRP) for estimation, which is known for its utilization in handling infinite vocabularies. B. Computing The Link-Anomaly Score In this module, we present a method for calculating the deviation of a user's behavior from the normal mentioning behavior that has been modeled.

To determine the anomaly score of a new post by user u at time t, which includes k mentions to users V, we calculate the probability using the training set T(t)u. The training set T(t)u consists of the posts made by user u within the time period [t-T, t], where T is set to 30 days in this project. The link-anomaly score is then defined based on this calculation.

The two terms in the equation mentioned above can be computed using the predictive distribution of the number of mentions and the predictive distribution of the mentioned users. C. Change Point Analysis and Dto This method is an expansion of the proposed Change Finder technique, which identifies changes in the statistical dependence structure of a time series by monitoring the compressibility of new data.

Instead of using the plug-in predictive distribution, this module utilizes a Modified Random Forest (NML) coding known as MRF coding as a coding criterion. The detection of a change point involves two layers of scoring processes. The first layer identifies outliers, while the second layer detects change points. In each layer, the criterion for scoring is based on the predictive loss using the MRF coding distribution for an autoregressive (AR) model.

Although the optimal NML code length is difficult to compute, the proposed SNML provides an approximation that can be computed sequentially. Additionally, the MRF employs discounting in the learning of the AR models. Finally, in our approach, the change-point scores are converted into binary alarms by applying a threshold. D.

Modified Random Forest Detection Method In the previous sections, we discussed the change-point detection based on MRF followed by DTO. In this module, we have tested our method in combination with Kleinberg's Modified Random Forest-detection method. To be specific, we have implemented a two-state version of Kleinberg's Modified Random Forest-detection model.

The reason behind choosing the two-state version is that we expect a nonhierarchical structure in this experiment. The Modified Random Forest-detection method is based on a probabilistic automaton model with two states, Modified Random Forest state and non-Modified Random Forest state.

The occurrence of certain events, such as the arrival of posts, is assumed to happen according to a time-varying Poisson process whose rate parameter depends on the current state. 5. ALGORITHM DETAILS Machine Learning (ML) and Deep Learning (DL) approaches are used, with a particular emphasis on the Modified Random Forest (MRF) approach, to analyze the KDD dataset.

Intrusion Detection with Modified Random Forest Step 1: Data Pre-processing Load the KDD dataset Pre-process the data, handle missing values, encode categorical features, etc. Step 2: Feature Engineering Extract relevant features from the dataset Optionally, perform dimensionality reduction techniques Step 3: Split the Dataset Split the dataset into training and testing sets Step 4: Modified Random Forest (MRF) Training Initialize the MRF model with hyper parameters Train the MRF model using the training set Step 5: Model Evaluation Use the trained MRF model to make predictions on the testing set Evaluate the model's performance using Detection Rate (DR) and False Alarm Rate (FAR) Step 6: Attribute Analysis Analyze the contributions of each attribute class (Basic, Content, Traffic, Host) to DR and FAR Optimize the dataset by adjusting features to achieve maximum DR while minimizing FAR 6.

RESULT ANALYSIS The Modified Random Forest (MRF) technique to empirical analysis of the KDD dataset produces informative results for the Intrusion Detection Systems (IDS) field. The study reveals the unique contributions of each attribute class to the Detection Rate (DR) and False Alarm Rate (FAR) by classifying the dataset into four categories: Basic, Content, Traffic, and Host.

Through this detailed analysis, the dataset may be optimized to maximize detection ratio (DR), which is a measure of successful intrusion detection, while decreasing false alarm rate (FAR) prevents needless false alerts. The results highlight how crucial attribute class considerations are when creating reliable intrusion detection models and offer insightful information for improving the effectiveness of cyber security measures. algorithm _accuracy _ _NB, and DT _75 _ _MRF _88 _ _ Table 1. Comparison table / Figure 3.

Comparison graph In the context of a particular investigation, the table displays the accuracy results of several

algorithms, including Naive Bayes (NB), Decision Trees (DT), and the Modified Random Forest (MRF). Interestingly, the combined accuracy of Naive Bayes and Decision Trees is 75%, whereas the Modified Random Forest (MRF) performs substantially better, scoring 88%.

These accuracy metrics demonstrate the MRF algorithm's improved performance above its NB and DT counterparts, highlighting its efficacy in the context under analysis. The MRF algorithm is shown in the table as a possible option for the current assignment, highlighting the significance of algorithm selection in reaching improved accuracy rates. 7.

CONCLUSION To summarize, the utilization of a modified random forest (MRF) algorithm in a network intrusion detection system (IDS) shows promise in accurately detecting network intrusions while addressing issues such as overfitting, adaptability, flexibility, and resilience against emerging threats. Implementing and training MRF-based IDS systems is relatively straightforward, and they can effectively monitor large networks.

These systems are capable of identifying various network attacks, including denial-of-service attacks, port scanning attacks, and malware attacks. However, it is crucial to acknowledge that no IDS system is flawless. MRF-based IDS systems are susceptible to evasion techniques, similar to other IDS systems. Moreover, the training and operation of MRF-based IDS systems can be computationally demanding. 8.

FUTURE WORK MRF classifiers are renowned for their exceptional precision in classification tasks; however, there exists potential for enhancement. Subsequent efforts may concentrate on the creation of novel MRF algorithms that exhibit heightened accuracy and efficiency. Intruders persistently devise fresh tactics to elude IDS systems.

Future endeavors may prioritize the development of MRF classifiers that possess greater resilience against these evasion techniques. The process of training and operation can incur substantial computational expenses. Future research could emphasize the formulation of innovative training algorithms and optimization techniques capable of diminishing the computational burden associated with MRF classifiers. 9. REFERENCES [1]. R. Kumar, A. Malik, and V.

Ranga conducted a study titled "An intellectual intrusion detection system using hybrid hunger games search and remora optimization algorithm for IoT wireless networks" which was published in the journal Knowledge-Based Systems in November 2022. [2]. W. Wang, S. Jian, Y. Tan, Q. Wu, and C. Huang developed a network intrusion detection system based on representation learning and capturing explicit and implicit feature interactions.

Their research was published in the journal Computer Security in January 2022. [3]. J. Oughton, W. Lehr, K. Katsaros, I. Selinis, D. Bubley, and J. Kusuma explored the comparison between wireless internet connectivity options 5G and Wi-Fi 6. Their findings were published in the journal Telecommunication Policy in June 2021. [4]. B. A. Tama and S.

Lim conducted a systematic mapping study and cross-benchmark evaluation on ensemble learning for intrusion detection systems. Their research was published in the journal Computer Science Review in February 2021. [5]. S. Lei, C. Xia, Z. Li, X. Li, and T. Wang proposed a novel model called HNN for studying intrusion detection based on multi-feature correlation and temporal-spatial analysis.

Their work was published in the IEEE Transactions on Network Science and Engineering in October 2021. [6] "Leveraging semisupervised hierarchical stacking temporal convolutional network for anomaly detection in IoT communication," by Y. Cheng, Y. Xu, H. Zhong, and Y. Liu IEEE Internet Things Journal, Jan. 2021, vol. 8, no. 1, pp. 144–155.

[7] "Sustainable ensemble learning driving intrusion detection model," IEEE Trans. Dependable Secure Comput., vol. 18, no. 4, pp. 1591–1604, Jul./Aug. 2021, X. Li, M. Zhu, L. T. Yang, M. Xu, Z. Ma, C. Zhong, H. Li, and Y. Xiang [8] Building an effective intrusion detection system based on feature selection and ensemble

classifier, Y. Zhou, G. Cheng, S. Jiang, and M.

Dai Journal of Computer Networks, vol. 174, June 2020, Article no. 107247. [9] "MLEsIDSs: Machine learning-based ensembles for intrusion detection systems—A review," by G. Kumar, K. Thakur, and M. R. Ayyagari Nov. 2020; J. Supercomput., vol. 76, no. 11, pp. 8938–8971 [10] "An enhanced anomaly detection in web traffic using a stack of classifier ensemble," B. A. Tama, L. Nkenyereye, S. M. R.

Islam, and K. Kwak, IEEE Access, vol. 8, pp. 24120–24134, 2020.

INTERNET SOURCES:

<1% - https://www.researchgate.net/publication/353623479_Intrusion_detection_system_using_voting-based_neural_network
<1% - https://typeset.io/papers/a-review-of-benchmark-datasets-and-its-impact-on-network-1dei2etg
<1% - https://www.mdpi.com/1424-8220/22/10/3744
1% - https://www.forbes.com/sites/benjaminlaker/2023/01/12/how-technology-will-shape-leadership-in-2023/
<1% - https://journalofbigdata.springeropen.com/articles/10.1186/s40537-024-00886-w
<1% - https://ieeexplore.ieee.org/document/10121454
<1% - https://www.analyticsvidhya.com/blog/2021/06/feature-selection-techniques-in-machine-learning-2/
<1% - https://www.hindawi.com/journals/cin/2022/6420799/
<1% - https://standards.ieee.org/standard/802_11-2020.html
<1% - https://www.sciencedirect.com/science/article/pii/S030859612100032X
<1% - https://arxiv.org/pdf/2010.11601
<1% - https://en.wikipedia.org/wiki/Internet_of_things
<1% - https://cybersecurity.springeropen.com/articles/10.1186/s42400-021-00077-7
<1% - https://link.springer.com/chapter/10.1007/978-981-16-3915-9_1
<1% - https://www.researchgate.net/publication/348049939_Ensemble_learning_for_intrusion_detection_systems_A_systematic_mapping_study_and_cross-benchmark_evaluation
<1% - https://www.researchgate.net/figure/Analysis-of-different-types-of-attacks-in-NSL-KDD-data-set_tbl2_330361204
<1% - https://www.researchgate.net/figure/Architecture-of-intrusion-detection-systems_fig1_301798790
<1% - https://www.researchgate.net/figure/Performance-evaluation-for-the-classifiers-A-SVM-B-Logistic-regression-C-Naive-Bayes_fig3_332837590
<1% - https://www.mdpi.com/2079-9292/12/8/1901
<1% - https://arxiv.org/pdf/2204.01682.pdf
<1% - https://repository.unsri.ac.id/29045/1/Automatic_Features_Extraction_Using_Autoencoder_in.pdf
<1% - https://link.springer.com/chapter/10.1007/978-981-13-1444-5_6
<1% - https://research.monash.edu/en/publications/deep-abstraction-and-weighted-feature-selection-for-wi-fi-imperso
<1% - https://link.springer.com/article/10.1007/s10994-023-06327-8
<1% - https://www.mdpi.com/2079-9292/11/6/898
<1% - https://www.sciencedirect.com/science/article/pii/S1877050915020190
<1% - https://www.sciencedirect.com/science/article/pii/S0140366422004601
<1% - https://ieeexplore.ieee.org/document/6113948
<1% - https://home.ttic.edu/~ryotat/papers/ICDM2011.pdf
<1% - http://rylanschaeffer.github.io/content/learning/bayesian_nonparametrics/chinese_restaurant_process.html
<1% - https://www.tandfonline.com/doi/full/10.1080/09537287.2024.2320790
<1% - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9414798/
<1% - https://www.hindawi.com/journals/wcmc/2022/6155925/

<1% - https://www.sciencedirect.com/topics/engineering/false-alarm-rate
<1% - https://dl.acm.org/doi/10.5555/3001460.3001502
<1% - https://www.sciencedirect.com/science/article/pii/S1568494623011432
<1% - https://www.semanticscholar.org/paper/Representation-learning-based-network-intrusion-by-Wang-Jian/e005b3fba841c7fbf39232d6b68c6629b661f9f2
<1% - https://discovery.researcher.life/article/ensemble-learning-for-intrusion-detection-systems-a-systematic-mapping-study-and-crossbenchmark-evaluation/85c9e6ff130134c7a5c7a2683ee79ffd
<1% - https://typeset.io/journals/computer-science-review-1rjmsi4d/2021
<1% - https://link.springer.com/article/10.1007/s12652-022-04461-0
<1% - https://dl.acm.org/doi/10.1109/TDSC.2022.3186918
<1% - https://arxiv.org/abs/1904.01352
<1% - https://www.growkudos.com/publications/10.1007%252Fs11227-020-03196-z/reader