Machine Learning-Based Intrusion Detection System

[1]Shaik Masood, [2]Dr. A.R. Deepa
[1]2301050161@kluniversity.in,
[2]deepaamuth@kluniversity.in,
[1,2]Department of Computer Science and Engineering,
[1,2]Koneru Lakshmaiah Education Foundation, Vaddeswaram,
Andhra Pradesh, India.

Abstract:

An explanation of IDSs is given in this work. Recent developments in technology have sparked worries about privacy and security. Network security is becoming more and more crucial as cyber networks and their uses grow. Intrusion detection systems (IDSs) that use machine learning are successful; in particular, the Supervised Model raises detection rates. People may find it challenging to comprehend their choices when faced with complex models. The majority of current model interpretation research is concentrated in domains including biology, computer vision, and natural language processing. Experts in cybersecurity find it difficult to maximize choices based on model evaluations in real life. A framework is proposed to handle these issues.

Keywords: datasets, machine learning, and intrusion detection systems Decision tree, Random Fores, and Support Vector Machine.

I.INTRODUCTION

Rapid development has been fueled by the Internet of Things (IoT), cloud computing, and 5G communication. By 2022, it is anticipated that there will be one trillion physical devices online. But new developments in technology have sparked worries about privacy and security. Because Internet equipment is so widely distributed and open, it is more likely to be the target of cyberattacks. Numerous Internet devices gather and handle personal data, which makes them vulnerable to malevolent attacks. Network security is ensured in large part by intrusion detection systems (IDSs). Within the host's network, an IDS finds abuse, misuse, and unauthorized use.

In recent years, IDSs have been developed and assessed using a variety of methodologies. Certain systems use superficial techniques to identify intrusions. For intrusion detection tasks, a variety of classification algorithms have been employed, such as Decision Tree, SVM, K-Nearest Neighbors, and Bayes Classifier. Other methods build intrusion detection models using ensemble classifiers or feature selection. IDSs are increasingly using deep learning techniques, such as CNNs, RNNs, VAEs, and DNNs. When detecting attacks, accuracy. This is a result of DNNs' continued use as "black boxes" that are unable to justify choices.

IDS will create a training model after training with every potential attack signature that could come from malicious user requests in order to identify such attacks. In order to determine whether a new request falls into the regular or attack category, IDS applies it to the train model. Different data mining classification or prediction techniques will be utilized to train and forecast such models. The author of this work assesses the effectiveness of ANNs and SVMs.

II LITERATURE SURVEY

"A review of themes, frameworks, methods, and future directions in cloud computing research" is the title. A meta-analysis of information research on cloud computing is presented in this publication. A review of the literature was carried out over a period of seven years in order to evaluate research frameworks, technique, geographic distribution, level of analysis, and trends. The literature currently available on cloud computing ignores commercial, conceptualization, and application domains in favor of concentrating on technical factors. Studies on cloud computing have steadily increased during the last seven years. Nevertheless, a lot of these investigations are devoid of models and theoretical frameworks. Instead of using qualitative, quantitative, or mixed approaches, the majority of cloud computing studies relied on simulation and experimentation. By offering thorough insights into research issues, methodology, framework, geography, and future directions, this study advances the field of cloud computing research.

The necessity of multidisciplinary approaches that integrate technology and business is emphasized in the review. It recommends using socio-technical viewpoints to have a deeper understanding of cloud consequences. There is a deficiency in research on actual adoption difficulties. Future research is guided by this publication to make more comprehensive contributions.

The survey "A survey of machine learning and data mining techniques for cyber security, including intrusion detection" offers a focused overview of the literature on machine learning (ML) and how data mining (DM) techniques are used to apply it to cyber analytics. It collects important works according to citation counts and provides succinct explanations of key ML/DM techniques. In addition, the paper lists popular cyber data sets, acknowledging the data need of

these techniques have a low false positive rate and are quite accurate.

Still, there are problems with intrusion detection, especially when it comes to system transparency. Experts in cybersecurity usually base their conclusions on IDS. Recommendations necessitate intelligible model predictions. When people are involved, the aforementioned models' growing complexity is a "Anomaly-based intrusion detection systems are increasingly utilizing hybrid machine learning models to enhance their efficiency and performance." It introduced the multi-class SVM model with chi-square feature selection. Chi-square was used to determine each feature's statistical significance, and features with low rankings were eliminated. During the feature selection process, there were 31 features instead of 41. To get the best possible combination of C and gamma values, the RBF-kernel SVM hyperparameters were adjusted. Despite the model's outstanding output, the authors did not use the KDD Test+ to assess it. A hybrid multi-level data mining system with hybrid feature selection was presented by Yao et al. To choose the best machine learning algorithms, the authors ran a number of tests.

Bostani and Sheikhan used a modified Optimum-path Forest model (OPF) to present a graph-based machine learning framework. The original NSL-KDD dataset was divided into K distinct training sets by the authors using K-Means in the framework.

The OPFs are trained using subsets. In order to expedite the OPF stage, a pruning module selected the best predictive samples from subsets derived from K-Means implementation using the social network analysis concepts of centrality and prestige.

III METHODOLOGY

1.Data analysis:

After the compilation of datasets from several sources. Prior to training the model, the dataset has to be pre-processed. Data pre-processing can be done in a number of ways.

Reading the gathered dataset and then cleaning it are the steps in the procedure. Some redundant features are eliminated from career prediction models during the data cleaning process. Unwanted characteristics and datasets with missing values must be eliminated or replaced with nan values in order to increase accuracy.

2. Data Preprocessing:

After the compilation of datasets from several sources. Prior to training the model, the dataset has to be pre-processed. Data pre-processing can be done in a number of ways.

Reading the gathered dataset and then cleaning it are the steps in the procedure. Some redundant features are eliminated from career prediction models during the data cleaning process. Unwanted characteristics and datasets ML/DM. It also discusses the intricacy of ML/DM algorithms, the difficulties in applying them to cyber security, and provides suggestions for their efficient implementation.

IV SYSTEM ARCHITECTUREAND DESIGN

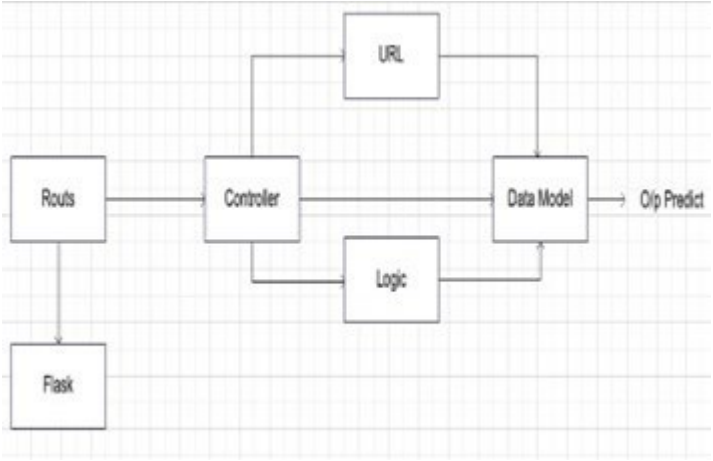


Fig 4.1Architecture Diagram

Incoming and outgoing network packets, flows, and logs are all considered network traffic. Numerous servers, network devices, and applications produce system logs. Information Gathering: Gather unprocessed data from many sources. Data cleaning: Get rid of extraneous features, noise, and missing values.

Extracting pertinent characteristics from unprocessed data is known as feature extraction. Frequency distributions and statistical data may be examples of this.

Normalization/Scaling: To guarantee consistency, scale characteristics to a predetermined range. Select the features that are most pertinent to the machine learning model's training.

Machine Learning Models:

Training Phase: The machine learning model is trained using labeled data. Neural networks, decision trees, and support vector machines are examples of common algorithms.

Validation and testing: Use distinct test data to assess the model's performance in order to guarantee generalizability. Model Selection: Considering criteria and performance indicators, choose the best machine learning method.

Model Selection: Based on this, choose the best machine learning algorithm.

Intrusion Detection Engine: Monitoring in real-time entails keeping an eye on system records and incoming network traffic. The process of extracting pertinent aspects from these incoming data streams, like traffic volume, source and destination IP addresses, and protocol types, is known as feature extraction. This data is essential for spotting patterns and irregularities that can point to performance problems or security risks. Through the analysis of the traits that were extracted, businesses can optimize resource allocation and adopt efficient security measures.

with missing values must be eliminated or replaced with nan values in order to increase accuracy.

3. Machine Learning Prediction uses statistical and machine learning methods to provide precise forecasts based on past data. Our system groups data using supervised techniques like classification and predicts continuous outcomes using regression. By successfully spotting patterns and trends that influence choices, these methods enhance decision-making and advance strategic initiatives. Organizations may take advantage of opportunities and proactively handle problems in a changing environment by utilizing these insights.

V DATA SET

In order to train machine learning to identify unusual dangers, the dataset is necessary. Many researchers still employ antiquated techniques, according to this study. A variation of KDD00, the NSL-KDD datasets, have drawn criticism for being out of date and unrelated to contemporary network architecture. This dataset is almost 20 years old, having been created in 1999. Network infrastructure is changing as a result of new technologies like social media, cloud computing, and the Internet of Things. Threat attacks are evolving as a result of these changes. Because of this, certain research findings—despite their great accuracy—are viewed as overstated.

1.Upload NSL-KDD Data Set

2.Preprocess the Data Set

VI EVALUATION

Support Vector Machine (SVM): SVM is a well-liked supervised learning technique for regression and classification. But in machine learning, it is mostly applied to classification issues. In order to make it simpler to allocate fresh data points to the proper category, the SVM method seeks to establish a decision boundary that separates an n-dimensional space into classes. A hyperplane is the optimal choice boundary.

Linear regression:

Prompt incident response is made possible by real-time monitoring, which allows for the instant identification of suspicious activities. By adjusting to shifting network patterns, machine learning approaches improve the accuracy of threat detection. Sustaining a secure network environment requires a strong monitoring and feature extraction strategy. Organizations can find vulnerabilities before they are exploited by routinely examining data, and automated alerts speed up decision-making. This proactive strategy raises security awareness and lowers risks.

Decision Tree:

A decision tree is a tree-like structure composed of nodes that indicate choices or decisions depending on properties or aspects of the data. The data is iteratively divided into subsets according to the feature values, producing a tree structure where each internal node denotes a "decision" based on a feature, each branch denotes the decision's conclusion, and each leaf node denotes the final prediction.

Random Forest:

Machine learning problems involving regression and classification frequently employ Random Forest, a potent ensemble learning technique. It creates a lot of decision trees during training and outputs the class that represents the mean prediction (regression) or the mode of the classes (classification) of the individual trees. An outline of the Random Forest classifier algorithm may be found here.

Upload Test Data and Detect Attack:

One popular supervised machine learning approach for predicting is linear regression. In supervised machine learning models, the model is constructed using training data, and its accuracy is then evaluated using the loss function.

One popular time series forecasting method for predictive modeling is linear regression. The dependent variable (of interest) and independent variables are implied to have a linear connection by the terminology. The linear regression model's equation, which is displayed here, is one that we are all familiar with.

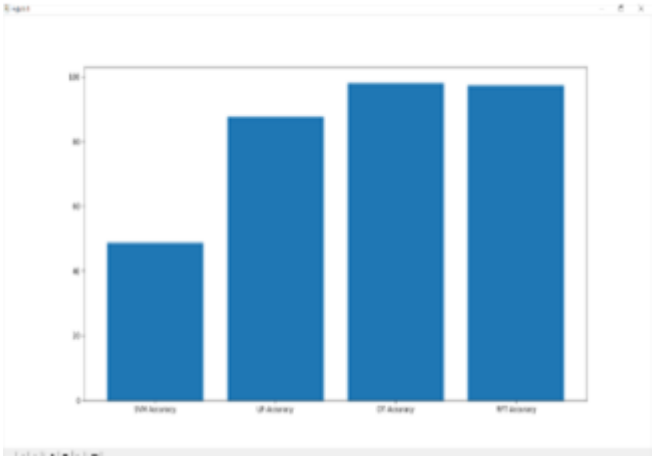
$y = mx + c$

where the dependent variable is denoted by y.

C is a constant, X is the independent variable, and M is the slope of the line.

VI RESULTS AND DISCUSSION

The project's goal was to develop an intrusion detection system strategy based on machine learning. The highest validation accuracy of 97.188 percent was attained by the implemented model. A number of indicators were used to assess the model's performance.



The Random Forest classifier method has achieved the highest accuracy out of all the algorithms.

VII FUTURE SCOPE

Enhanced DetectionAccuracy:

Higher detection accuracy for both known and new threats may result from further study into machine learning techniques, such as deep learning and ensemble approaches. One method that can help models become more resilient to evasion attacks is adversarial training.

Real-time Detection and Response:

The period between intrusion and mitigation can be

VIII CONCLUSION

A novel framework for understanding both local and global occurrences is put forth in this research. This approach can be used with any model and has a strong theoretical basis. IDS predictions can be interpreted using this paradigm. Key characteristics, feature correlations, and attack kinds are described in the global description. The NSL-KDD dataset is used to confirm the viability of the suggested framework.

The interpretation of our approach provides intuitive findings consistent with some attack features. While multiclass classifiers manage many classes at once, frequently increasing detection precision, one-vs-all classifiers simplify classification by treating each class as a binary problem. Security professionals can improve their capacity to handle a variety of attacks and fortify network security by utilizing these tactics. Furthermore, performance is greatly impacted by the classifier selection, necessitating a thorough assessment of the particular environment in order to maximize detection capabilities.

IX REFERENCES

[1 Aljabri, M.; Altamimi, H.S.; Albelali, S.A.; Maimunah,A.H.; Alhuraib, H.T.; Alotaibi, N.K.; Alahmadi, A.A.; Alhaidari, F.; Mohammad, R.M.A.; Salah, K.Detecting malicious URLsusing machine learning techniques: Review and research directions. IEEE Access 2022.

[2] Liu, J.; Dong, Y.; Zha, L.; Tian, E.; Xie, X. Event-based security tracking control for networked control systems against stochastic cyber-attacks. Inf. Sci. 2022

.[3] Zha, L.; Liao, R.; Liu, J.; Xie, X.; Tian, E.; Cao, J. Dynamic event-triggered output feedback control for networked systems subject to multiple cyber attacks. IEEE Trans. Cybern. 2021

.[4] Zheng, Q.; Zhao, P.; Wang, H.; Elhanashi, A.; Saponara, S. Fine-grained modulation classification using multi-scale radio transformer with dual-channel representation. IEEE Commun. Lett. 2022

[5] Elhanashi, A.; Gasmi, K.; Begni, A.; Dini, P.; Zheng, Q.; Saponara, S. Machine Learning Techniques forAnomaly-Based Detection System on CSE-CIC-IDS2018 Dataset. In

shortened by concentrating efforts on creating IDS that can identify and react to threats instantly. Rapid threat containment and mitigation can be achieved by integration with automated response mechanisms.

Scalability and Efficiency:

To handle the increasing amount and complexity of network traffic, scalable machine learning methods and distributed computing frameworks might be researched. Investigating lightweight models suitable for resource-constrained settings such as edge devices and Internet of Things networks is part of this.

Integration with Security Ecosystem:

To offer complete threat detection and response capabilities, intrusion detection systems (IDS) can be connected with other security tools and technologies, including firewalls, intrusion prevention systems (IPS), and Security Information and Event Management (SIEM) systems.

[10] Lopez-Martin, M.; Sanchez-Esguevillas, A.; Arribas, J.I.; Carro, B. Contrastive Learning Over Random Fourier Features for IoT Network Intrusion Detection. IEEE Internet Things J. 2023.

[11] Lopez-Martin, M.; Carro, B.; Arribas, J.I.; Sanchez-Esguevillas, A. Network intrusion detection with a novel hierarchyofdistances betweenembeddingsofhashIPaddresses. Knowl.-Based Syst. 2021

Applications in Electronics Pervading Industry, Environment and Society: APPLEPIES 2022; Springer: Berlin/Heidelberg, Germany, 2023

[6] Liu, L.; Wang, P.; Lin, J.; Liu, L. Intrusion detection of imbalanced network traffic based on machine learning and deep learning. IEEEAccess 2020

[7] Solani, S.; Jadav, N.K. A Novel Approach to Reduce False-Negative Alarm Rate in Network-Based Intrusion Detection System Using Linear Discriminant Analysis. In Inventive Communication and Computational Technologies; Springer: Berlin/Heidelberg, Germany, 2021

[8] Dini, P.; Begni, A.; Ciavarella, S.; De Paoli, E.; Fiorelli, G.; Silvestro, C.; Saponara, S. Design and Testing Novel One-Class Classifier Based on Polynomial Interpolation With Application to Networking Security. IEEE Access 2022

[9] Moualla, S.; Khorzom, K.; Jafar, A. Improving the performance of machine learning-based network intrusion detection systems on the UNSW-NB15 dataset. Comput. Intell. Neurosci.2021