iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** V **Month of publication:** May 2024

DOI: https://doi.org/10.22214/ijraset.2024.61461

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

Designing IDS to Analyze Malicious Attacks using Machine Learning

Prof. S. L. Dawkhar[1], Nikita Borate[2], Nandini Bangad[3], Prashant Patil[4], Sudeep Joshi[5]
Department of Information Technology, Sinhgad College Of Engineering, Pune,India

Abstract: ***This study presents a novel approach to enhance network security through the design and implementation of an Intrusion Detection System (IDS) employing machine learning (ML) techniques. The primary aim is to address current challenges in real-time threat detection by seamlessly integrating ML models, including Logistic Regression, Random Forest, and XGBoost. The research tackles drawbacks in existing systems, emphasizing the struggle with real-time threat detection, the complexity of ML integration. The working model classifies incoming requests as normal or intrusions, categorizing the latter into DDoS, R2L, U2R, or Probing attacks. Leveraging insights from four referenced papers, this study builds on efficient activation functions, optimized feature selection, and empirical analyses of ML models for adaptive network intrusion detection. Results showcase promising comparisons between ML algorithms which include Logistic Regression, Random Forest, XG Boost.***
Keywords: Intrusion Detection System, Machine Learning

I. INTRODUCTION

In the dynamic internet landscape, the surge in network security challenges is evident alongside technological advancements. Ensuring the security of internet applications is crucial as our reliance on them grows. Among defense mechanisms, the Intrusion Detection System (IDS) plays a pivotal role, akin to a fortress guarding against threats.

Operating as a server-side protection system, an IDS continuously analyzes network activities, much like a sentinel, to detect and thwart attacks. Categorized as Host-Based Intrusion Detection System (HIDS) and Network-Based Intrusion Detection System (NIDS), these systems offer retrospective and real-time analysis, respectively.

While traditional methods like feature matching and machine learning are prevalent, the emergence of machine learning algorithms has revolutionized intrusion detection, offering superior accuracy and self-learning capabilities. This paper focuses on a network-based IDS design based traditional IDS dataset, integrating machine learning for enhanced accuracy and real-time defense. This paper further more compares various machine learning models to provide accurate and precise values using various parameters for training and testing. This helps to achieve a model having high accuracy to detect intrusion in the system.

The proposed system features a modular development approach, enabling easy expansion and development of functionalities. This ensures high-precision intelligent defense, balancing overall operating efficiency. This report details the system architecture, implementation methods, and CNN integration, heralding a new era of intrusion detection powered by advanced deep learning techniques.

II. LITERATURE REVIEW

Sr. No	Paper Title and Year	Author	Key Findings and Results
1.	Design And implementation of IDS Using Deep Learning (2023 IEEE 3rd International Conference on Electronic Technology, Communication and Information)	Shihao Wang*, Zhefan Chen, Jin Chen, Peiwen Zhu	This paper implements networkbased intrusion detection systems with convolutional neural networks, enabling efficient coordination and highprecision defense. The system collects and processes traffic data for real-time intrusion detection using efficient activation functions and optimizers

©IJRASET: All Rights are Reserved

2.	An Optimized Network Intrusion Detection Mode (2022 IEEE 5th Advanced Information Management, Communicates, Electronic and Automation Control Conference)	LIU Dongdong, DOU Hongtao,HAN Bo,NIU Lei	This paper introduces a featurebased network intrusion detection model, highlighting its effectiveness with the KDD CUP99 dataset, where it outperforms other models in terms of detection rates and false alarmsThis paper presents an optimized network intrusion detection model using feature selection, showing improved performance compared to other models. However, further research is required to assess its scalability and real-world applicability
3.	Empirical Analysis of Machine Learning Models towards Adaptive Network Intrusion Detection Systems (2022 4th International Conference on Smart Systems and Inventive Technology)	Dr. Aranga Arivarasan, Mr. S. P. Senthilkumar	This paper empirically analyzes machine learning models for adaptive network intrusion detection, investigating their effectiveness in detecting network intrusions. The research shows that machine learning models effectively detect network intrusions using tools like TensorFlow and scikit-learn for analysis and evaluation.
4.	Intrusion detection system analysis using ML (2022 IEEE 2nd Mysore Sub Section International Conference)	Rupak Dutta, Nirupama B.K, Niranjanamurthy	This research employs machine learning for efficient invader detection, evaluating models within the context of Intrusion Detection Systems (IDS), encompassing both Host-based (HIDS) and Network-based (NIDS) classifications. Among the four algorithms, KNeighborsClassifier performed best with an accuracy score of 0.867, making it a reliable choice for future research or implementations

III. SYSTEM ARCHITECTURE

The project's modeling and analysis encompass several key components, beginning with data collection and preprocessing. A diverse dataset of network traffic data is collected and then subjected to rigorous preprocessing, including cleaning, normalization, and feature engineering, to enhance its quality and relevance.

Next, three machine learning (ML) models—Logistic Regression, Random Forest, and XGBoost—are developed for intrusion detection. These models are trained on the preprocessed dataset to distinguish between normal network activity and various types of attacks, such as DDoS, R2L, U2R, and Probing to provide accuracy and precision.

In real-time intrusion detection, incoming network requests are analyzed by the ML models for classification. The models determine whether a request constitutes normal activity or falls into one of the predefined attack categories which will further detect intrusion. This real-time analysis allows for immediate identification and response to potential security threats.

This will help to notify the administrator of the intrusion and direct to take protective measures towards the attack.

This comprehensive approach to modeling and analysis integrates machine learning with IDS, enabling effective and efficient intrusion detection in complex network environments.

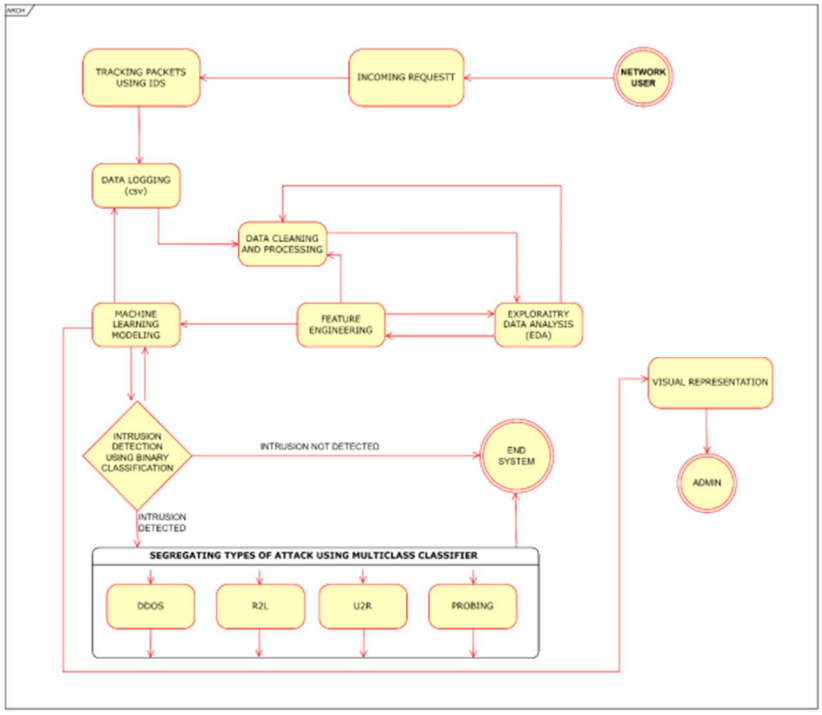


Diagram 1: Architecture of the model

IV. PROPOSED METHODOLOGY

To achieve the objectives outlined in our research, we employed a systematic methodology, integrating machine learning (ML) with the Intrusion Detection System (IDS). The following steps delineate the key components of our approach:

1) *Dataset Selection and Preprocessing:* We utilized a diverse dataset containing network traffic data to train and evaluate our ML models. Data preprocessing involved cleaning, normalization, and feature engineering to enhance the quality and relevance of the input data.
2) *Training and Testing:* Using NSL-KDD datasets for training and testing the various considered machine learning models for identifying the attacks. This step will help in predicting the accuracy of the models and considering which model to be used in the IDS which will provide better and optimized output.
3) *Machine Learning Model Development:* Three ML models—Logistic Regression, Random Forest, and XGBoost—were implemented for intrusion detection. The models were trained on labeled data, distinguishing between normal network activity and four types of attacks: DDoS, R2L, U2R, and Probing.
4) *Integration with IDS:* We seamlessly integrated the developed ML models with the IDS to enable real-time intrusion detection. The integration involved the creation of interfaces to facilitate communication and data exchange between the ML models and IDS.
5) *Evaluation Framework:* We established a comprehensive evaluation framework to assess the performance of our IDS. Metrics such as accuracy, precision, recall, and F1 score were used to quantify the effectiveness of the ML models in detecting and classifying intrusions.
6) *Validation and Comparison:* The performance of our system was validated through rigorous testing using both simulated and real-orld scenarios. Comparative analyses were conducted against existing IDS solutions and traditional methods to gauge the efficacy of our approach. By meticulously executing these steps, our methodology aimed to create a robust IDS that not only leverages ML for enhanced threat detection but also seamlessly integrates with existing infrastructure while providing an intuitive visualization.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

V. RESULT ANALYSIS

Two techniques data scientists can use to balance datasets are Oversampling and Undersampling Oversampling is appropriate when data scientists do not have enough information. One class is abundant, or the majority, and the other is rare, or the minority. In oversampling, the scientist increases the number of rare events. The scientist uses some type of technique to create artificial events. One technique to create artificial events is synthetic minority oversampling technique (SMOTE).

Undersampling is appropriate when there is plenty of data for an accurate analysis. The data scientist uses all of the rare events but reduces the number of abundant events to create two equally sized classes. Typically, scientists randomly delete events in the majority class new to end up with the same number of events as the minority class. Undersampling is a technique to balance uneven datasets by keeping all of the data in the minority class and decreasing the size of the majority class. It is one of several techniques data scientists can use to extract more accurate information from originally imbalanced datasets.

Random oversampling involves randomly duplicating examples from the minority class and adding them to the training dataset. Examples from the training dataset are selected randomly with replacement. This means that examples from the minority class can be chosen and added to the new "more balanced" training dataset multiple times; they are selected from the original training dataset, added to the new training dataset, and then returned or "replaced" in the original dataset, allowing them to be selected again.

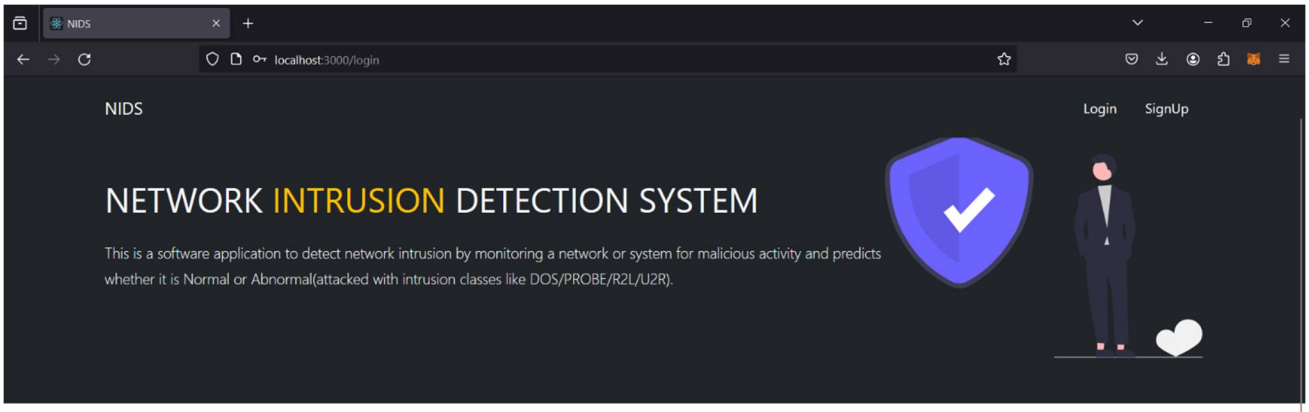
SMOTET omek is somewhere upsampling and downsampling. SMOTETomek is a hybrid method which is a mixture of the above two methods, it uses an under-sampling method (Tomek) with an oversampling method (SMOTE). This is present within imblearn.combine module.

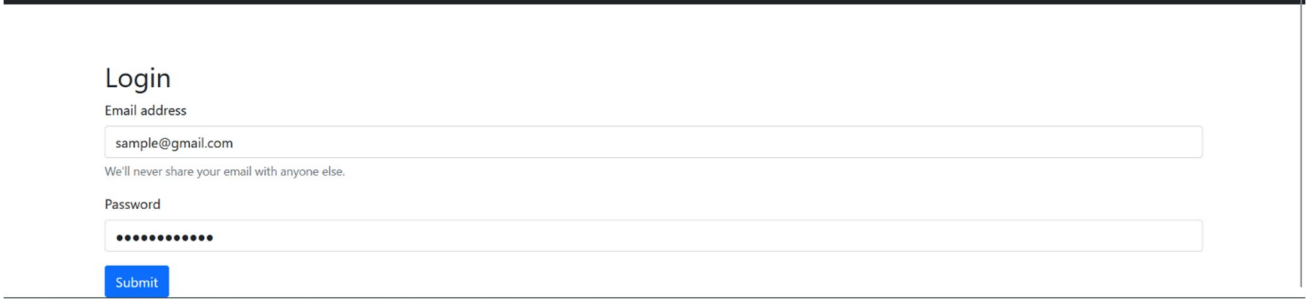
Lastly, by going through the analysis of the types of attacks considered i.e. Probe, U2R, R2L and DDoS. DDoS attack has the most dominance over the other attacks in the training and the testing phase.

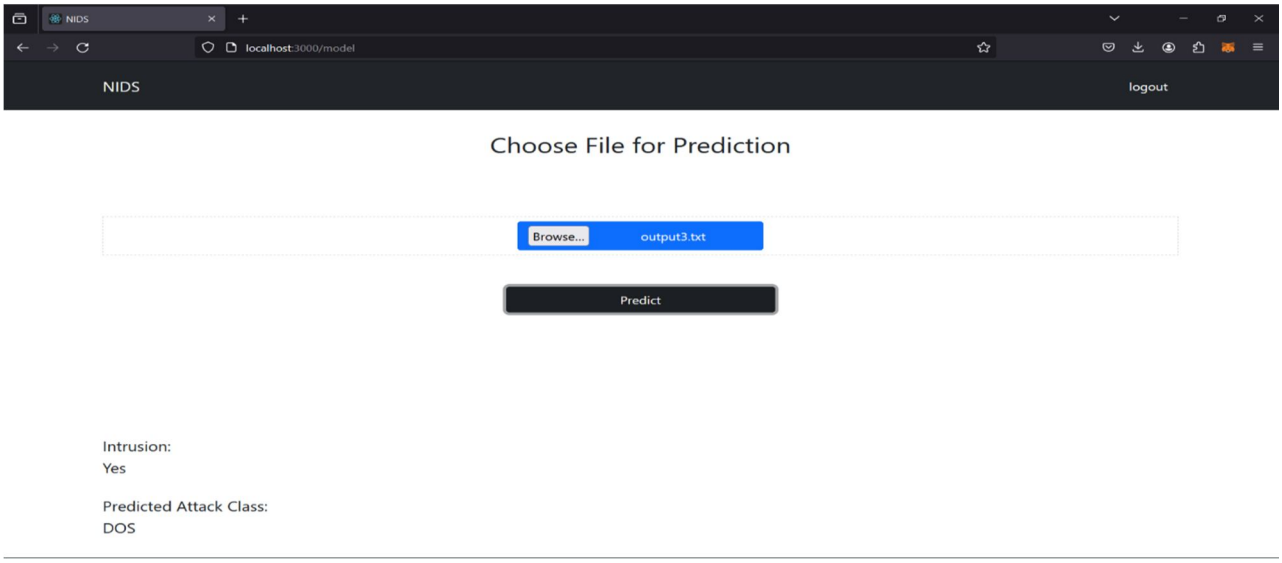
Model	Train Log Loss	Test Log Loss	Misclassified Points
Logistic Regression without over sampling or under sampling	0.2623	0.9640	25.0793
Random Forest without over sampling or under sampling	0.0162	0.4468	13.1851
XGBOOST without over sampling or under sampling	0.0206	0.3990	8.6721
Logistic Regression with under sampling	1.0356	2.3678	35.3235
Random Forest with under sampling	0.1186	1.5287	61.4075
XGBOOST with under sampling	0.1616	1.6396	62.5891
Logistic Regression with over sampling using RandomOverSampler	1.3815	1.3819	25.0741
Random Forest with over sampling using RandomOverSampler	0.0106	0.6264	14.7363
XGBOOST with over sampling using RandomOverSampler	0.0377	0.6415	12.3470
Logistic Regression with over sampling using SMOTETomek	1.3823	1.3802	25.0689
Random Forest with over sampling using SMOTETomek	0.0207	0.4894	11.7692
XGBOOST with over sampling using SMOTETomek	0.0361	0.7281	10.5980

©IJRASET: All Rights are Reserved

VI. IMPLEMENTATIONS







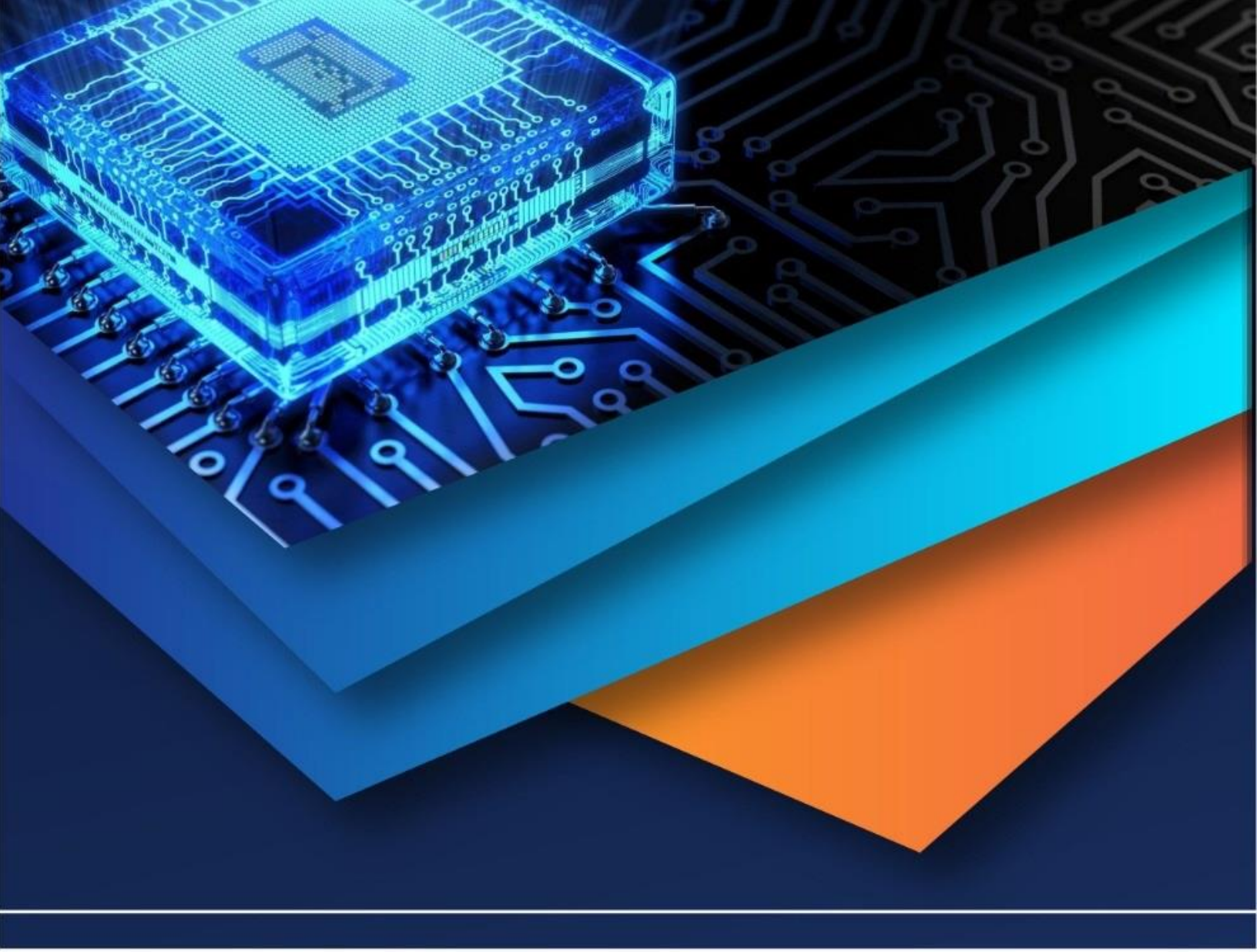
VII. CONCLUSION

By seamlessly integrating machine learning (ML) techniques, including Logistic Regression, Random Forest, and XGBoost, with the Intrusion Detection System (IDS), we have successfully addressed longstanding challenges in real-time threat detection. Predicted the Intrusion-based signal using a combination of different classification algorithms. Also used some over-sampling and under-sampling techniques to make changes in the data.

The data were imbalanced, so the Recall score/Sensitivity score will be the most effective performance metric to evaluate the best model for binary classification and log loss score for multi-class classification. So based on the best recall score, finally chosen the Random forest algorithm on the over-sampled data for binary classification part and XGBOOST algorithm on the raw data for multi-class classification part. Comparative analyses against IDS solutions underline the competitive advantage of our ML-driven approach, showcasing its robust performance and adaptability across diverse network environments However, the system is currently weak against malicious attacks against artificial intelligence systems.

REFERENCES

[1] Shihao Wang*, Zhefan Chen, Jin Chen, Peiwen Zhu, "Design And implementation of IDS Using Deep Learning" IEEE 3rd International Conference on Electronic Technology, Communication and Information, 2023
[2] LIU Dongdong, DOU Hongtao,HAN Bo,NIU Lei , "An Optimized Network Intrusion Detection Model" IEEE 5th Advanced Information Management, Communicates, Electronic and Automation Control Conference, 2022
[3] Dr.Aranga.Arivarasan, Mr. S. P. Senthilkumar, "Empirical Analysis of Machine Learning Models towards Adaptive Network Intrusion Detection Systems" 4th International Conference on Smart Systems and Inventive Technology, 2022
[4] C Rupak Dutta, Nirupama B.K, Niranjanamurthy, "Intrusion detection system analysis using ML", IEEE 2nd Mysore Sub Section International Conference, 2022



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)