Unmasking unbalanced network traffic by intruders based on Machine-learning Algorithms.

C. Siva Kumar[1]
Associate Professor, School of Computing,
Dept of Data Science at Mohan Babu Univesity
(ERSTWhile Sree Vidyanikethan Engineering College), Tirupati, Andhra Pradesh, India
sivakumar.c@vidyanikethan.edu

A. Rajalingam[2]
Senior Lecturer, College of Engineering and Technology
University of Technology and Applied Sciences, Shinas, Oman
rajalingam.a@gmail.com

G. Charulatha[3]
Associate Professor, Department of ECE, PERI Institutue of Technology Chennai
drcharulatha79@gmail.com

M. Ramkumar Prabhu[4]
Professor, Department of ECE, PERI Institutue of Technology Chennai
ramkumarprabhu@gmail.com

Abstract

Obtrusion unmasking involves identifying and detecting malicious or anomalous network activity. Imbalanced network traffic refers to a scenario where the ratio of normal to abnormal traffic is significantly skewed, with one type of traffic significantly outweighing the other. Machine learning and deep learningtechniques can be used to analyze network traffic and identify patterns and characteristics that may indicate the presence of obtrusions or anomalies. These techniques can be applied to various types of network data, including packets, flows, and logs, and can be trained on large datasets to improve their accuracy and effectiveness. For the purpose of predicting the intrusion in imbalance network traffic, this research employs a variety of categorization techniques. The classification algorithms are Random Forest, Support Vector Machine, convolution neural network, and principal component analysis. The dataset, which included several meteorological parameters, was obtained via the UCI repo. With a testing data ratio of 70:30. Based on precision, accuracy, recall, f1-score, and execution time, the efficacy of the classification algorithms was assessed.

Keywords—Machine learning, Deep learning, Principal component analysis, Convolutional neural network, Random Forest.

I. INTRODUCTION

Obtrusion unmasking in imbalanced network traffic refers to the process of detecting and identifying malicious or anomalous activity within a network environment where the distribution of normal and abnormal traffic is unbalanced. This can be a challenging task, as traditional machine learning algorithms are often designed to perform well on balanced datasets and may struggle to accurately classify anomalous traffic in an imbalanced setting. To overcome this issue, advanced ML and deep learning techniques can be utilized to analyze network traffic and identify patterns and characteristics that may indicate the presence of obtrusions or anomalies. These techniques can be applied to various types of network data, including packets, flows, and logs, and can be trained on large datasets to improve their accuracy and effectiveness. The use of ML and DL techniques for obtrusion unmasking in imbalanced network traffic can play a critical role in improving the security and reliability of networks by detecting and mitigating the impact of malicious activity and other types of anomalous behavior. This can be especially important in today's interconnected world, where networks are constantly under attack and the risk of cyber threats is constantly evolving. When employing machine learning and deep learning techniques to perform obtrusion unmasking in unbalanced network traffic, a number of difficulties may occur. These difficulties include:

Unbalanced network traffic, as previously established, is characterized by a situation in which the ratio of normal to abnormal traffic is noticeably skewed, with one form of traffic greatly outweighing the other. Due to this, it may be challenging for conventional ML algorithms to precisely recognize and categorize abnormal traffic.

Restricted data: In some circumstances, the amount of data that can be used to train and evaluate ML and DL models for obtrusion unmasking may be limited. Because of this, it could be challenging to recognize trends and traits that might point to the existence of anomalies.

Threats: Cyber dangers are continuously changing, and it may be challenging to stay on top of the newest tricks and strategies employed by attackers. This may make it difficult to appropriately identify and categorize novel obtrusion kinds.

False positives: Sometimes ML and DL models may create false positives, misclassifying regular traffic as abnormal. This may result in unneeded alarms and interruptions and necessitate further human analysis to ascertain the real nature of the traffic.

Complex networks: Due to the volume and diversity of the data they generate, large and complex networks can make it more difficult to detect abnormalities.

In order to properly address these issues, sophisticated ML and DL methods must be used, together with thorough planning and analysis to guarantee that models are correctly trained and optimised for the particular network context. The manner of finding and resolving troubles that expand whilst the quantity of incoming and outgoing community site visitors is surprisingly choppy is referred to as obstruction unmasking in imbalanced community site visitors. This can show up whilst one aspect of a community connection is sending a disproportionate quantity of information in comparison to the other, that could reason overall performance troubles and possibly pose protection

Threats: Machine gaining knowledge of (ML) and deep gaining knowledge of can be hired in lots of methods to fight impediment unmasking in unbalanced community site visitors (DL). Several such techniques include:

Data balancing: Using techniques like oversampling or under sampling to alter the distribution of information is one manner to stabilize the extent of incoming and outgoing community site visitors. This may also useful resource in minimizing the impact of unbalanced information at the effectiveness of ML and DL models.

ML and DL: An opportunity approach is to hire ML and DL strategies to identify and spotlight peculiar styles in community site visitors that would factor to the feasible unmasking of a blockage. To do this, one may also educate a version to recognize common community site visitors styles after which hire it to identify departures from those styles.

Managing the waft of community site visitors with the aim of improving overall performance and keeping off congestion is referred to as "site visitors shaping." Traffic shaping algorithms can be progressed the usage of ML and DL to higher control impediment unmasking in unbalanced community site visitors.

Network optimization: ML and DL will also be used to decorate the overall performance of the community as an entire and optimize how community assets are configured. This may also entail streamlining community useful resource allocation, streamlining routing algorithms, and finding and putting off bottlenecks that is probably inflicting blockage unmasking.

In conclusion, using ML and DL to remedy impediment unmasking in unbalanced community site visitors can help to decorate the overall performance and protection of the community and ensure that it could correctly meet the desires of the customers and packages it serves.

II. RELATED WORK

Sofia Mathew [1] Established a machine learning (deep learning)-based intrusion detection system for cyber security with the use of ensemble language to combine many classifier techniques for improved prediction performance with the help of Alexnet classifier and mini VGGnet classifier

Punam Bedi [2] Without employing conventional class balancings techniques like oversampling, random under sampling, or SMOTE, the proposed Siam-IDS addresses the issue of class imbalance in IDSs. The NSL-KDD dataset was used for the training and testing of the suggested system.

Shone [3] submitted a non-symmetrical deep automobile encoder(NDAE) established deep education for alone feature knowledge. Their approach has happened tested in Tensor flow accompanying GPU support and proven utilizing the standard Knowledge Discovery in Databases Cup '99 and NSL-Knowledge Discovery in Databases datasets. According to the authors, the veracity of their resolution on the NSL-KDDdataset is 85.4% for the 5-class categorization and 89.2% for the 13-class categorization

Lloret et al. [4] proposed a novel network traffic categorization method that combines the RNN and CNN deep learning models. When tested against the RedIRIS dataset, the model achieves an accuracy of 96.3%.

Dada [5] Support Vector Machine, K Nearest Neighbor, and Primal-Dual Particle Swarm

Optimization were used in a hybrid technique to obtain high accuracy for IDSs.

Quamar Niyaz [6] bestowed deep education-based methods for designing specific fruitful and adaptable NIDS. On NSL-KDD, a gauge dataset for network interruption, we engage in Self-instructed Learning, a deep learning-located approach.

Mrutyunjaya Panda [7] Principal component analysis is secondhand as a filtering design in the submitted study, which connects discriminatory parameter education accompanying a Naive Bayes classifier to improve categorization veracity compared to prior designs.

wan Syarif [8] The research was done into how well different clustering algorithms performed when used for anomaly detection. There are five distinct clustering techniques: k-Means, enhanced Algorithms for network intrusion detection using k-Means, k-Medoids, EM clustering, and distance-based outlier identification

Chuanlongyin [9] Recurrent neural networks have happened bestowed as a deep education action for intrusion detection. The act of the projected model in twofold categorization and multiclass categorization was likewise checked, as well as the belongings of the number of neurons and knowledge rate on the results.

Partha Sarathi Bhattacharjee [10] Various network assaults may be detected using a Genetic Algorithm based with the aim of finding an appropriate Vectorised Fitness function for chromosomal assessments, an intrusion detection system is used. For effective intrusion detection, GA-based IDS with weighted Vectorizeda Fitness function is suggested and tested using the NSL-KDD data set.

A.A. Ojugo [11] attempts to offer evolutionary unsupervised model(s) as a computational option that modifies the data locality feat and compares convergence results produced by the unsupervised hybrid classifiers in order to give intelligent adaptive mail assistance that learns user preferences. which accomplishes these accomplishments by including local judgment heuristics into its categorization procedures such that the spam filter or filters are built with email genres in mind.

Yali Yuan [12] The C5.0 approach, the Naive Bayes algorithm, and twain Layers Multi-class Detection method are bestowed for adjusting network interruption discovery, which increases two together the discovery rate and the needless alarm or worry rate. The submitted TLMD method likewise takes care of various challenges that stand while data excavating, to a degree talking unstable datasets, dealing with unending traits, and threatening clamor in training datasets

III. EXISTING METHOD - XGBOOST

A well-liked machine learning technique called XGBoost Extreme Gradient Boosting is frequently used for supervised learning tasks like regression and classification. It is a member of the family of techniques known as gradient boosting, which incrementally adds additional models to increase prediction accuracy.

Gradient boosting is a method used by XGBoost to merge a number of poor predictive models into a single strong model. Gradient boosting is a method for improving prediction accuracy by iteratively adding new models after fitting a basic model to the data in the beginning. The method seeks to fit a model that can account for the mistakes produced by the prior models at each iteration.

The capacity of XGBoost to handle big datasets with plenty of features is one of its main advantages. It is ideal for use in distributed computing environments because it is also very efficient and simple to parallelize.

Many applications, such as fraud detection, web page ranking, picture classification, and many more, have successfully exploited XGBoost. Also, its effectiveness has been acknowledged in a number of machine learning competitions, including the well-known Kaggle events.

Overall, XGBoost is a robust and adaptable algorithm that is well-liked by the machine-learning community and has gained widespread adoption among data scientists and practitioners.

Using XGBoost for intrusion unmasking in unbalanced networks has various benefits, including:

Dealing with Unbalanced Data: Unbalanced datasets are typical in intrusion detection when the majority of the data comes from one class (for example, regular traffic) and the minority class (for example, malicious traffic) is underrepresented. Unbalanced data may be handled by XGBoost using built-in techniques like weighing the samples or modifying the decision threshold.

Each variable in the dataset receives a feature significance score from XGBoost. This makes it possible for the user to comprehend which characteristics are most crucial for forecasting the target variable, which may aid in feature engineering and feature selection.

Scalability: XGBoost can handle huge datasets with a lot of features and is very scalable. It is suitable for usage in situations that employ distributed computing since it can also be parallelized.

Regularization: L1 and L2 regularisation techniques, which are included in XGBoost, can assist reduce over fitting and enhance the generalization capabilities of the model.

Performance: XGBoost has won several machine learning contests because of its excellent performance and accuracy. It has been demonstrated that it performs better in many intrusion detection settings than other well-known algorithms like Random Forest and Support Vector Machines.

In conclusion, XGBoost is a strong and adaptable algorithm that excels at handling unbalanced data, providing feature importance, scaling well, regularising, and delivering good performance, making it a great option for obstruction unmasking in imbalanced networks using machine learning.

IV. PROPOSED METHOD – PCA IN RF

PCA could be a broadly utilized dimensionality decrease strategy in machine learning that can be utilized in conjunction with SVM to progress the execution of the classifier.

PCA works by changing the first highlights into a set of modern highlights, called vital components, which are direct combinations of the first highlights. These vital components capture the most extreme sum of change within the information with the least number of measurements. Utilizing PCA in conjunction with SVM can have a few preferences:

Made strides Execution: By diminishing the dimensionality of the information, PCA can offer assistance progress the execution of the SVM classifier, particularly when managing with high-dimensional information.

Decreased Over fitting: PCA can moreover offer assistance to decrease over fitting by expelling commotion and unessential data from the information, which can offer assistance to the SVM to generalize way better-unused information.

Quicker Preparing: With fewer measurements, the SVM can be prepared speedier, which can be critical when managing huge datasets.

Way better Visualization: PCA can offer assistance visualize the information in a lower-dimensional space, which can offer assistance in understanding the structure of the data and recognizing designs.

Dimensionality Lessening: PCA can decrease the dimensionality of the information by changing the initial highlights into a set of new features, called central components. This may offer assistance to diminish the revile of dimensionality, move forward the SVM's generalization execution, and avoid overfitting.

Feature Selection: PCA can offer assistance distinguish the foremost instructive highlights within the dataset by positioning the vital components based on their change or commitment to the information. This may offer assistance to diminish the number of highlights utilized by the SVM, which can speed up preparation and decrease the hazard of over fitting.

Moved forward Separation: PCA can offer assistance progress the partition of the classes by changing the information into a lower-dimensional space where the classes are more divisible. This could lead to way better classification execution and decrease the chance of misclassification.

Diminished Computational Complexity: With fewer measurements, the SVM can be prepared quickly and requires less memory. This may be vital when managing expansive datasets.

Visualization: PCA can offer assistance visualize the information in a lower-dimensional space, which can offer assistance in understanding the structure of the information and distinguishing designs.

In any case, it's vital to note that utilizing PCA with SVM may not continuously lead to superior execution. It depends on the particular dataset and the issue being illuminated. In some cases, utilizing PCA may lead to a diminish in execution, particularly on the off chance that the first highlights are as of now exceedingly enlightening and pertinent to the classification errand.

In outline, utilizing PCA with SVM can be a capable procedure for dimensionality diminishment and making strides in the execution of the classifier. Be that as it may, it's critical to carefully assess the effect of PCA on the particular dataset and issue being illuminated.

V. COMPARISON OF RESULTS

After applying Principal Component Analysis in Random Forest on Knowledge discovery in databases the Random Forest algorithm predicts the Obtrusion in imbalance network traffic efficiently with an accuracy of 99.33% on the Knowledge Discovery in Databases dataset. Table1 shows the different metrics evaluated on imbalance network traffic in the network and figure2 shows the graphical comparison between the algorithms used.

	Accuracy	Precision	Recall	F1-Score
XGBoost	0.7715	0.8107	0.7715	0.7365
PCA in SVM	0.7966	0.8384	0.8066	0.7615

Table 1: prediction of imbalance network traffic using machine learning models.

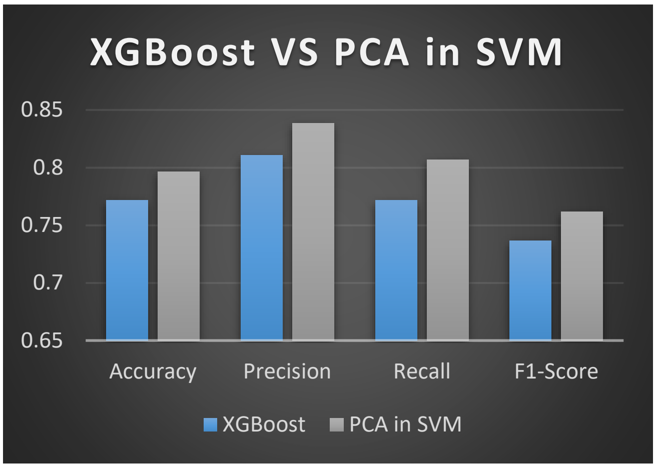

Figure 2: Comparison between XGBoost and Principal component analysis in Support vector machine

VI. CONCLUSION

Prediction of obtrusion unmasking in imbalance network traffic can be done with many algorithms but still, there is a problem of accuracy. In this paper, we have concentrated on different ML algorithms, and with the output accuracy from the algorithm, we have done some analysis from the imbalance network traffic data. We see that including principal component analysis in support vector machine will work efficiently when compared with the other algorithms.

REFERENCES

[1] Sofiya Mathew "Intrusion Detection of Imbalanced Network Traffic Based On Ensemble Learning ",vol. 1 , No. 6 , pp. 1630-1635, 2022

[2] Punam Bedi-"Handling class imbalance problem in intrusion detection systems using siamese neural network,"Procedia Comput. Sci., vol. 2, pp. 780–789, 2020

[3] Shone - "A deep learning approach to network intrusion detection,'' IEEE Trans. Emerg.Topics Comput. Intell, vol. 3, no. 1, pp. 41–50, Feb. 2018

[4] Lloret et al[4] - ("Deep Learning Nature, vol. 4, pp.436-444, 2015.)

[5] Dada - "A hybridized SVM-kNN-pdAPSO approach to intrusion detection system" Proc. Fac. Seminar Series: 14-21, 2017.

[6] J. R. Arunkumar, S. Velmurugan, B. Chinnaiah, G. Charulatha, M. Ramkumar Prabhu et al., "Logistic regression with elliptical curve cryptography to establish secure iot," Computer Systems Science and Engineering, vol. 45, no.3, pp. 2635–2645, 2023.

[7] Mrutyunjaya Panda[7] -" Ensemble of Classifiers for Detecting Network Intrusion (ICAC3'09)", pp. 510-515, 2009

[8] Wan Syarif-""Faster R-CNN Implementation Method for Multi-Fruit Detection Using Tensorflow Platform"", vol. 8.

[9] R. Anusuya, M. Ramkumar Prabhu, Ch. Prathima, J. R. Arun Kumar" Detection of TCP, UDP and ICMP DDOS attacks in SDN Using Machine Learning approach" Journal of Survey in Fisheries Sciences, Vol. 10 No. 4S (2023): Special Issue 4. 964-971.

[10] Chuanlong Yin - ""A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks"", vol.9.

[11] Partha Sarathi Bhattacharjee-""intrusion detection system for NSL-KDD data set using vectorised fitness function in geneticalgorithm,"" Adv. Comput. Sci. Technol., vol. 10, no. 2, pp. 235–246, 2017.

[12] Yali Yuan[11] - "Two Layers Multi-class Detection method for network Intrusion Detection System"IEEE Symposium on Computers and Communications (ISCC). Heraklion, Greece: IEEE: 767-772.

[13] P. Nirmala, T. Manimegalai, J. R. Arunkumar, S. Vimala, G. Vinoth Rajkumar, Raja Raju, "A Mechanism for Detecting the Intruder in the Network through a Stacking Dilated CNN Model", Wireless Communications and Mobile Computing, vol. 2022, Article ID 1955009, 13 pages, 2022. https://doi.org/10.1155/2022/1955009.

[14] A.A. Ojugo -"Genetic Algorithm Rule-Based Intrusion Detection System", CIS, Vol. 12, No. 8.ISSN 2079-8407, 2012

[15] Prathima Chilukuri , J.R. Arun Kumar , R. Anusuya , M. Ramkumar Prabhu. (2022). Auto Encoders and Decoders Techniques of Convolutional Neural Network Approach for Image Denoising In Deep Learning. Journal of Pharmaceutical Negative Results, 13(4), 1036–1040. https://doi.org/10.47750/pnr.2022.13.04.142 (Original work published November 4, 2022).

[16] Quamar Niyaz-"A deep learning approach for network intrusion detection system" vol.6, 2016.