iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VI **Month of publication:** June 2023

DOI: https://doi.org/10.22214/ijraset.2023.54319

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

Comparative Analysis of Machine Learning and Deep Learning Techniques for Intrusion Detection

K. Teja Reddy[1], K. Suhaas[2], K. Srikar[3], K. Srivathsa[4], Dr. K. Rajeshwar Rao[5]
[1, 2, 3, 4]Student(s), AIML Department, Malla Reddy University
[5]Assistant Professor, AIML Department, Malla Reddy University

***Abstract:** This comparative analysis examines the application of both machine learning and deep learning methods in network traffic classification. Network traffic classification holds significant importance in network security, traffic management, and Quality of Service provisioning. The analysis covers a range of popular machine learning techniques, such as Decision Tree, K-Nearest Neighbours, Naive Bayes, Logistic Regression, Multi-Layer Perceptron, and Feed Forward Neural Network with a sigmoid activation function. Each technique's strengths and weaknesses are discussed, along with the factors that influence the selection of a particular technique. Ultimately, the choice of machine learning approach depends on data characteristics, performance requirements, and available resources. The demand for prompt and accurate classification of Internet traffic has been steadily increasing, driven by the emergence of new applications in the field. Traditional approaches based on port numbers and packet payloads have become insufficient, prompting the adoption of pattern recognition techniques that leverage statistical flow-based features in training samples to classify unknown flows. To ensure real-time identification of traffic types, the chosen method must be capable of swift classification before the entire flow is completed. In this study, a supervised machine learning approach and deep learning techniques are proposed for the identification of various Internet applications. The proposed system exhibits the ability to detect application types based on just a few initial packets within each flow, enabling real-time operation. Promising results were achieved, with the Logistic Regression algorithm attaining the highest accuracy of 80.7%.*

Keywords: **Network, classification, Internet, Packets, Deep Learning, Perceptron**

I. INTRODUCTION

Machine learning and deep learning techniques have gained popularity as an alternative for classifying flows based on application protocol payload-independent statistical features. These features include packet length, inter-arrival times, flow lengths, and others. A consistent set of payload-independent statistical features characterizes each traffic flow. To build a machine learning classifier, a representative set of flow instances with known network applications is used for training. The trained classifier can then be applied to classify unknown flows. The statistical analysis-based approach treats the application classification problem as a statistical challenge. The ML and DL-based approach offers the advantage of being independent of packet payload inspection, making it robust against encryption. In recent years, various supervised and unsupervised, deterministic and probabilistic ML and DL methods have been utilized to classify network traffic flows based on different applications and features. While many existing studies focus on analysing the entire flow lengths, real-time traffic classification has become crucial for addressing complex network management challenges faced by ISPs and equipment vendors. Network operators require prompt knowledge of the traffic flowing through their networks in order to react swiftly and align with their business goals. Achieving timely classification before the completion of the entire flow is essential for identifying specific traffic classes and facilitating rapid network responses. This implies making classification decisions based on a finite subset of packets from each flow.

II. IMPLEMENTATION

With the growing number of applications and internet usage, the increasing traffic within data flows poses a significant challenge, often leading to server crashes or blockages. To address this issue, it becomes essential to implement an efficient system capable of detecting and tracking data packets within applications and websites. Such a system would enable early detection of traffic, even before the packets reach their intended destination. In the proposed approach, a comprehensive collection of packets is curated using a combination of online and offline modes. To optimize the performance of the system in terms of computational complexity and accuracy, meticulous pre-filtering is applied to the captured payload packets. This strategic elimination process aims to streamline the analysis procedure. Subsequently, the system intelligently identifies the precise count of packets traversing the flow, enabling the selection of a judiciously sampled subset for subsequent classification.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

In the proposed methodology, the first step involves calculating the features of each packet within the flow and performing sub-sampling to create a representative subset. Next, an attribute selection method is employed to identify the most significant attributes for machine learning. Subsequently, the packets are classified into different classes using various classifiers such as k-nearest neighbour, logistic regression, decision tree, naive Bayes, Multi-Layer Perceptron, and Feed Forward Neural Network with Sigmoid Neuron. Finally, the results are evaluated based on the classes using support vector machine, allowing for comprehensive assessment of the classification performance.

III. ARCHITECTURE

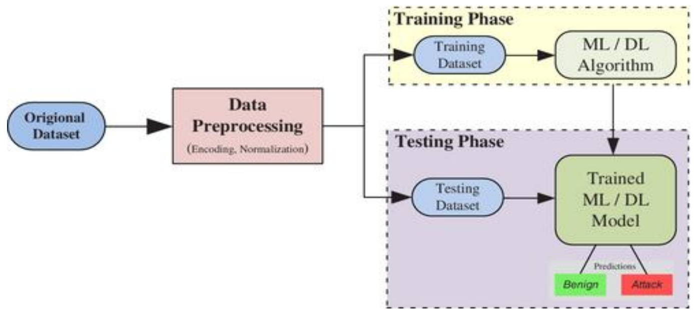


Fig. 1 Architecture and process flow of the network classification

Majorly 4 types of supervised machine learning algorithms and 2 deep learning algorithms are being using in implementing the project. They are

1) K-nearest neighbor (KNN) is a popular machine learning algorithm used for classification and regression tasks. It operates by classifying new data points based on the majority class of their k nearest neighbors in the feature space.
2) Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems.
3) Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
4) Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
5) The Multi-Layer Perceptron (MLPs) breaks this restriction and classifies datasets which are not linearly separable. They do this by using a more robust and complex architecture to learn regression and classification models for difficult datasets.
6) A Feed Forward Neural Network is an artificial neural network in which the connections between nodes does not form a cycle. The opposite of a feed forward neural network is a Recurrent Neural Network, in which certain pathways are cycled. The feed forward model is the simplest form of neural network as information is only processed in one direction. While the data may pass through multiple hidden nodes, it always moves in one direction and never backwards.

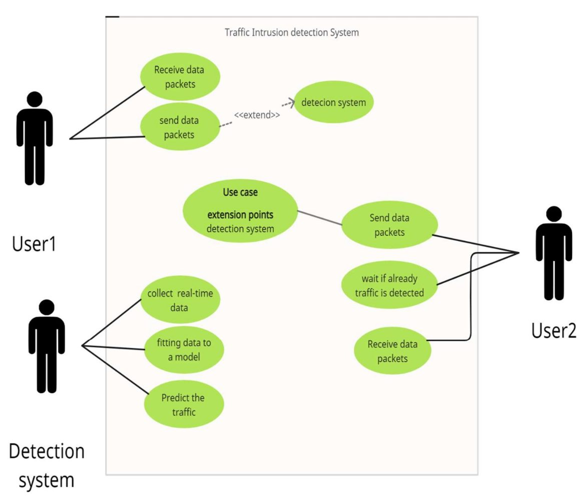


Fig. 2 Use-case diagram of the Network Traffic

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

IV.RESULTS

pandas : 1.3.5
numpy : 1.21.6
matplotlib : 3.2.2
seaborn : 0.11.2
sklearn : 1.0.2
imblearn : 0.8.1
Train set dimension: 125973 rows, 42 columns
Test set dimension: 22544 rows, 42 columns
Original dataset shape Counter({1: 67343, 0: 45927, 2: 11656, 3: 995, 4: 52})
Resampled dataset shape Counter({1: 67343, 0: 67343, 3: 67343, 2: 67343, 4: 67343})

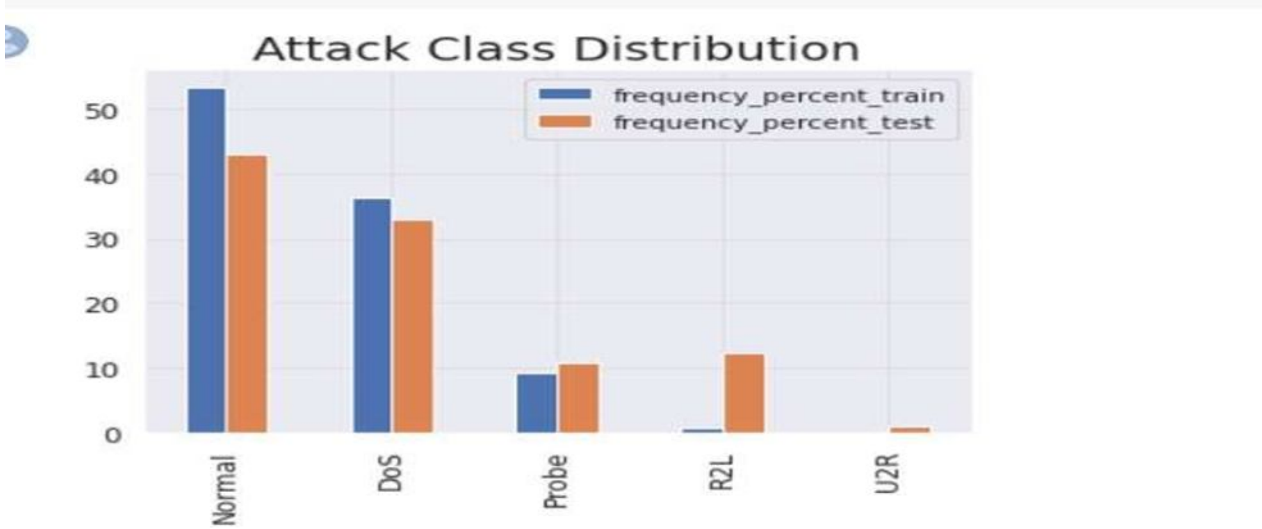


Fig. 3 Attack class distribution of different attack types in train and test data

	attack_class	frequency_percent_train	attack_class	frequency_percent_test
Normal	67343	53.46	9711	43.08
DoS	45927	36.46	7458	33.08
Probe	11656	9.25	2421	10.74
R2L	995	0.79	2754	12.22
U2R	52	0.04	200	0.89

Fig. 4 Frequency of each type of attack in train and test data

['src_bytes', 'dst_bytes', 'logged_in', 'root_shell', 'serror_rate', 'srv_serror_rate', 'dst_host_srv_count', 'dst_host_serror_rate', 'dst_host_srv_serror_rate', 'service']

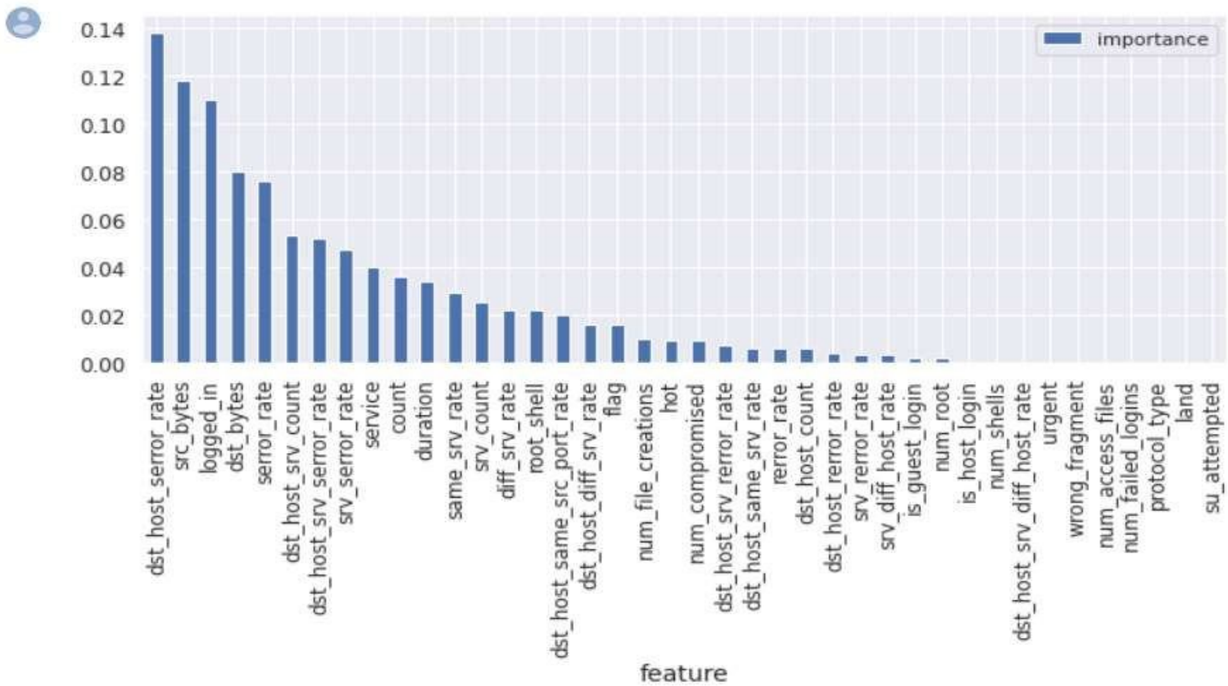


Fig. 5 Extracting the required features to train the model

Naive Baye Classifier Model Test Results
Model Accuracy:0.7335313646688799
Confusion matrix: [[3227 4231]
[344 9367]]

Decision Tree Classifier Model Test Results
Model Accuracy:0.2588386044615295
Confusion matrix: [[3344 4114]
[8611 1100]]

KNeighborsClassifier Model Test Results
Model Accuracy: 0.695497699341837
Confusion matrix: [[2730 4728]
[500 9211]]

LogisticRegression Model Test Results
Model Accuracy:0.8076183819674996
Confusion matrix: [[5627 1831]
[1427 8239]]

V. CONCLUSION

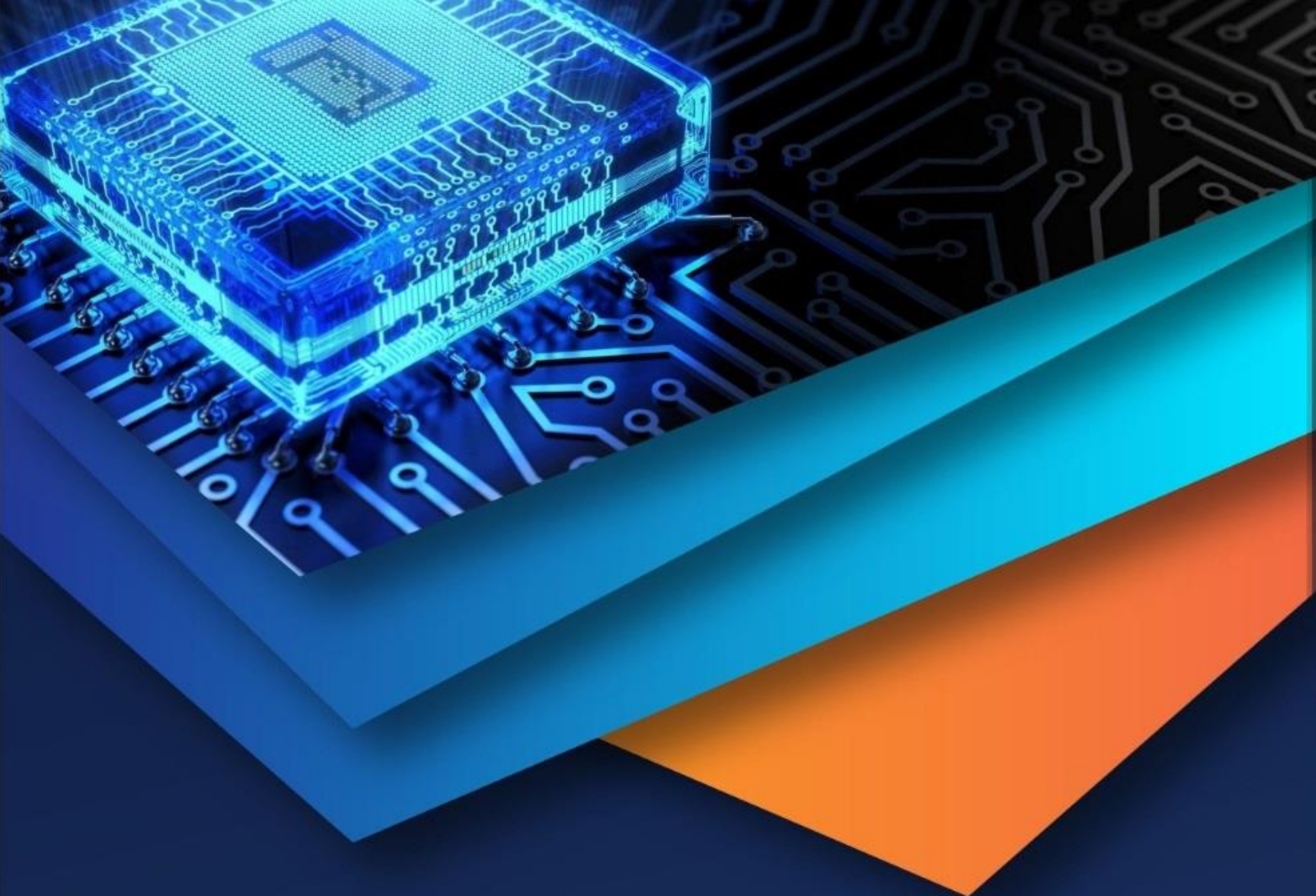
In this project, a system has been proposed to detect Internet traffic intrusion using supervised machine learning and deep learning methods. The system relies solely on flow-based and packet-based statistical features, eliminating the need for payload inspection and addressing associated limitations. Furthermore, the study explores the discriminatory power of different feature types for distinguishing traffic types, suggesting specific attribute subsets for each category. Notably, the system demonstrates real-time traffic type detection capabilities by analysing just a few packets within each flow, making it highly valuable for online traffic monitoring in support of Quality of Service (QoS) and security objectives.

VI. ACKNOWLEDGEMENT

We would like to extend our sincere appreciation to Dr. K. Rajeshwar Rao[5], our project guide, and our Head of the Department (CSE - AI&ML) Dr. Thayyaba Khatoon, for their invaluable guidance, insightful feedback, and continuous support throughout this research. Lastly, we would like to thank our families, friends, and loved ones for their unwavering support and encouragement during the completion of this research.

REFERENCES

[1] Moor, and K. Papaiannaki, “Toward the accurate identification of network application”, PAM'05, pp. 41-54, USA, 2005.
[2] H. Patrick et al. “ACAS: Automated Construction of Application Signatures”, Proceedings of the 2005 ACM SIGCOMM workshop on Mining network data, pp: 197 – 202, Philadelphia, Pennsylvania, 2005.
[3] T. S. Tabatabaei and S. Krishnan, "Towards robust speech signal processing", Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, Istanbul, Oct 2010.
[4] T. S. Tabatabaei and S. Krishnan, "SVM-based classification of digital modulation signals", Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, Istanbul, Oct 2010.
[5] S. Sen, O. Spatscheck, and D. Wang, “Accurate, scalable innetwork identification of P2P traffic using application signatures,” in Proceedings of the 13th International Conference on World Wide Web, pp. 512-521, New York, USA, 2004.
[6] M. Roughan, S. Sen, O. Spatscheck et al., “Class-of-service mapping for QoS: a statistical signature-based approach to IP traffic classification,” In Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement, pp. 135-148 Sicily, Italy, 2004.
[7] K. Singh and S. Agrawal, “Comparative analysis of five machine learning algorithms for IP traffic classification”, Proceedings of IEEE International Conference on Emerging Trends in Networks and Computer Communications (ETNCC), pp 33-38, 2011.
[8] T. Nguyen and G. Armitage, “A survey of techniques for Internet traffic classification using machine learning.” In Proceedings of IEEE Comm. Surv. & Tutor, pp 56–76, 2008.
[9] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, Berlin, 1995.



10.22214/IJRASET

45.98

IMPACT FACTOR:
7.129

IMPACT FACTOR:
7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY