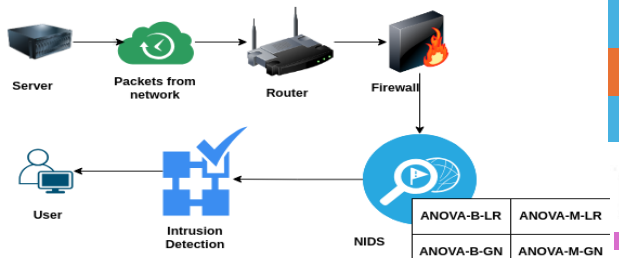
Optimizing Intrusion Detection Systems (IDS) through Significant Feature Selection By Machine Learning Techniques

Group Leader: M. Usman Akram **Team**: M. Ahtsham Akram, Ahsan Tariq **Supervised By**: Dr. Reehan Ali Shah

Introduction:

In this study, we address these challenges by employing a range of ML techniques: **ANOVA-B-LG, ANOVA-B-GN, ANOVA-B-GB, ANOVA-M-LG, ANOVA-M-GN, and ANOVA-M-GB, ANOVA-M-LG L-1, ANOVA-B-LG L-1, ANOVA-B-LG L-2, ANOVA-M-LG L-2, ANOVA-B-GN L-1, ANOVA-M-LG L-2, ANOVA-M-GB-GSCV, ANOVA-M-GB-XGB, ANOVA-B-GB-XGB, ANOVA-M-GB-RF.** These methods are tailored for binary and multi-class classification on the **UNSW-NB-15** benchmark dataset.



ANOVA-B-LR	ANOVA-M-LR
ANOVA-B-GN	ANOVA-M-GN
ANOVA-B-GB	ANOVA-M-GB

Fig: Architecture Diagram Of IDS

Literature Review

Methodology	Dataset Used	Key Findings
Data analytics in cyber security	Not specified	Big Data analytics offers new avenues for enhancing cyber security through analysis of large volumes of network data.
Machine learning algorithms in IoT for network security	Not specified	Machine learning algorithms in IoT bolster network security by ensuring reliability, security, availability, and survivability of security assets.
Intrusion detection system with integrated classification model	UNSW-NB15 dataset	Novel intrusion detection system capable of identifying five distinct types of threats with a significantly higher accuracy rate of 83.8% on real-time data.
Feature reduction with XGBoost for intrusion detection	UNSW-NB15 dataset	XGBoost-based feature reduction method improves test accuracy of binary classification scheme from 88.13% to 90.85%, addressing challenges of high-dimensional and imbalanced datasets.
Network Intrusion Detection Systems (NIDS) in IoT security	UNSW-NB15 dataset	SVM outperforms other methods with 85.99% accuracy for binary classification and 75.77% for multi-classification.
Comparative analysis of classification algorithms	UNSW-NB15 dataset	Random Forest classification model achieves highest accuracy of 97.49% compared to Decision Tree and Naïve Bayes models.
Development of network intrusion detection systems	UNSW-NB15 dataset	Stacking machine learning models with feature selection techniques achieves accuracy of 96.24%, outperforming recent competing models.

Problem Statement

The main challenge of IDS lies in optimizing feature representation to enhance accuracy and increase processing performance. Extensive features set demand a strategic approach to address redundancy and ensure relevance in cyber threat detection. Strike a balance between accuracy and processing performance.

Aim:

- The primary aim of This project is to develop a Network intrusion detection system (NIDS) using Machine Learning Techniques.

Objectives:

- Optimize IDS performance
- Reduction of imbalance redundant features and selection of the significant features.
- Analyze IDS on various ML techniques.

Methadology:

Given Figure described The methadology :

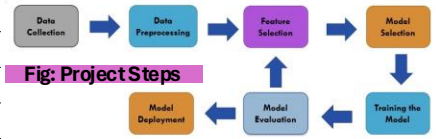


Fig: Project Steps

The Given Figures Described The Work Flow Of Our Project:

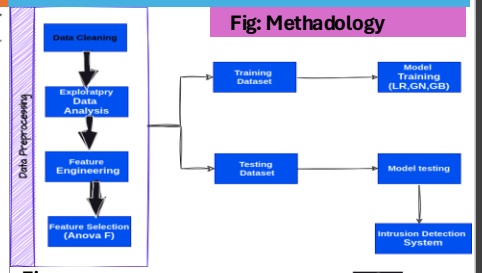


Fig: Methadology

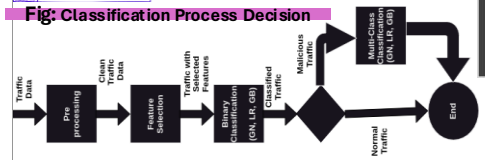


Fig: Classification Process Decision

Categorical Features

Attack	Numbers	States	Numbers	Services	Numbers	Protocols	Numbers
Normal	37000	FIN	39339	None	47153	TCP	43095
Generic	18871	INT	34163	DNS	21367	UDP	29418
Exploits	11132	CON	6982	HTTP	8287	UNAS	3515
Fuzzers	6062	REQ	1842	SMTP	1851	ARP	987
Dos	4089	ACC	4	FTP	1552	OSPT	676
Reconnaissance	3496	RST	1	FTP-DATA	1396		
Analysis	677	CLO	1	POP-3	423		
Backdoor	583			SSH	204		
Shellcode	378			SSL	30		
Worms	44			SNMP	29		
				DHCP	26		
				RADIUS	9		
				IRC	5		

Machine Learning Models:

- **Naive Bayes Model**

$$P(y|x_1, x_2, \ldots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1, x_2, \ldots, x_n)}$$

- **Gradient Boosting Model**

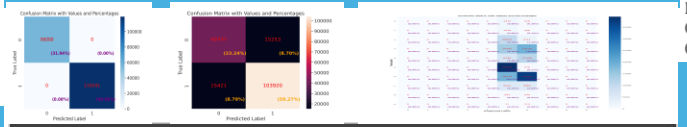
$$F_m(x) = F_{m-1}(x) + \lambda h_m(x)$$

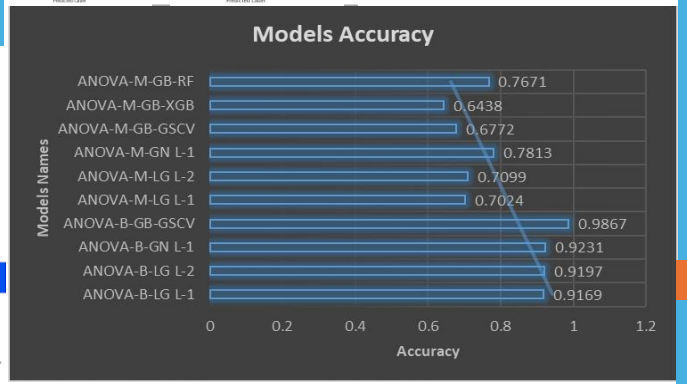
- **Logistic Regression Model**

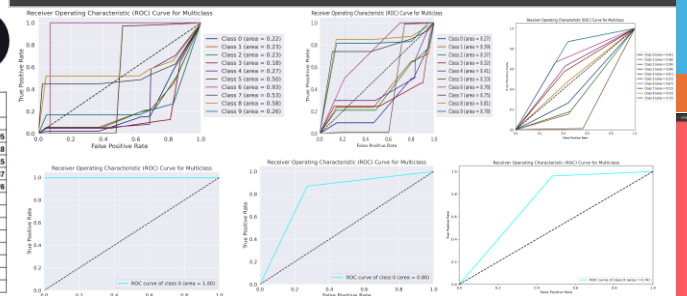
$$P(y = 1|x) = \frac{1}{1 + e^{-\beta^T x}}$$

Results And Discussions:

The Given Figures Described The Accuracy Score Achieved By ML Models For the IDS:







TIME COMPLEXITY:

Algorithm	Training Time	Prediction Time
Logistic Regression	$O(nm^2)$	$O(n)$
Gradient Boosting	$O(nm \log m)$	$O(n \log m)$
Gaussian Naive Bayes	$O(nm)$	$O(n)$

Future Work:
In the future, we are set to elevate our intrusion detection systems by deploying our machine learning models on live networks, transitioning from static web-based applications to dynamic, real-time defenses.

CONCLUSION:
In case of **binary classification**, the results highlighted that **Gradient Boosting, coupled with Grid-Search Cross Validation (ANOVA-B-GB-GSCV)**, exhibited the highest accuracy, achieving an impressive score of **0.9867**.
On the other hand, in the domain of multi-classification, **Gaussian Naive Bayes in conjunction with L1-Regularization (ANOVA-M-GN L-1)** emerged as the most proficient technique, attaining a commendable accuracy of **0.7813**.

This evolution includes a strategic shift towards federated learning models, enabling enhanced security through collaborative data sharing across multiple decentralized sources. By leveraging the power of federated learning, we aim to create a robust, adaptive intrusion detection system that continuously learns and improves, providing unprecedented protection against emerging threats.

Web Deployment

Backend : Django Framework(Python)
Frontend : HTML, CSS , JAVASCRIPT, JQUERY
DATABASE: SQL ORM

WEB PAGE LAYOUTS