**Network Intrusion Detection System Using Machine Learning**

**Project Guide:** Dr. Bheemappa Halavar

**Team members:** S. Sai Tarun (S0190010157)

L. Kranthi Kusal (S20190010107)

P. Sree Charan (S20190010146)

**Abstract**

The network size and associated data have greatly increased as a result of the quick developments in the internet and communication areas. The resulting proliferation of innovative threats has made it difficult for network security to reliably identify breaches. Furthermore, it is impossible to overlook the existence of intruders who intend to conduct a variety of attacks against the network. One such technology is an intrusion detection system (IDS), which guards against potential intrusions by examining network traffic to verify its availability, confidentiality, and integrity. We first trained models using the NSL-KDD dataset. incorporating feature selection and extraction with machine learning techniques including decision trees, naive bayes, logistic regression, and random forests. In addition, we bagged naïve bayes and decision trees, which improved accuracy correspondingly from 72.2% to 98.3% and 99.6% to 99.1%. Unlike prior studies, we also examined this model on a different dataset (CIC-IDS).

**Introduction**

The present interest in and developments in internet and communication technologies over the past 10 years have made network security a crucial research issue. It uses tools like firewalls, antivirus software, and intrusion detection systems to ensure the security of the network and all of its associated assets in cyberspace (IDS). One of these is the network-based intrusion detection system (NIDS), which is the attack detection mechanism that provides the needed protection by continuously examining network traffic for malicious and suspicious behavior.

In 1980, Jim Anderson introduced the idea of IDS for the first time. Since then, a wide range of IDS systems have been developed and enhanced to satisfy network security requirements. But in the last ten years, technology has advanced so quickly that both the size of networks and the number of applications they can accommodate have significantly increased. As a result, a sizable amount of important data is generated and distributed among many network nodes. Considering the increase of new attacks,either as mutations of more established attacks or entirely new attacks, protecting these data and network nodes has become a difficult task.

**Related Work**

In order to effectively detect intrusions across the network, machine learning (ML) and deep learning (DL)-based IDS systems have recently been introduced as promising solutions. IDS is defined in [1], which also offers a taxonomy based on prominent ML and DL approaches used in the creation of network-based IDS (NIDS) systems. The discussion of the advantages and disadvantages of the suggested solutions provides a thorough evaluation of the recent NIDS-based studies. The most recent developments in ML and DL-based NIDS are then presented in terms of the suggested approach, assessment metrics, and dataset choice.To clarify the importance of feature selection in the classification and training phase of ML IDS, feature selection that affects the effectiveness of ML IDS is explored in [2].

[3] suggests an adaptive ensemble learning model that may integrate the benefits of each method for various forms of data detection and use ensemble learning to produce the best outcomes. On the NSL-KDD dataset, the advantage of ensemble learning is that it combines the predictions of numerous base estimators to increase generalizability and robustness in comparison to a single estimator.

**Problem Statement**

Building and comparing Intrusion Detection System implementations using various Machine Learning Models that inspects network traffic and predicts various harmful cyber attacks.

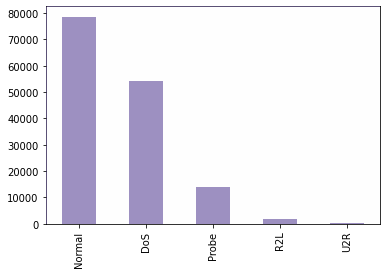
**Proposed Methodology**

The proposed model applies various machine learning algorithms on 2 different datasets, namely NSL-KDD and CSE-CIC-IDS2018, which contain various attack classes and features that help to determine the nature of the request as to whether it is an anomaly or a normal request.

* **NSL-KDD Dataset**

This dataset is an iteration of the KDD-CUP '99 dataset.

* Since redundant data are not included in the train set, the classifiers won't be biased in favor of more frequent records.
* The suggested test sets do not contain any duplicate data; as a result, the performance of the learners is unaffected by approaches that have higher detection rates for frequent records.
* The proportion of records in the original KDD data set that are selected from each difficulty level group is inversely related to the number of records chosen from those groups. Because of this, there is a broader range of variation in the classification rates of various machine learning algorithms, which makes it easier to evaluate various learning methods accurately.
* The train and test sets contain a manageable number of records, making it feasible to conduct the experiments on the entire set without the need to randomly pick a subset. As a result, evaluation findings from various research projects will be comparable and consistent.



*Fig1. Attack classes of NSL-KDD Dataset before preprocessing*

We have observed that the attack classes are highly imbalanced i.e, The labels are not in correct proportions(as in Fig1). As it is not good for the classifiers, we have oversampled the data using random oversampler. The size of the dataset after the preprocessing is (148,517x43).

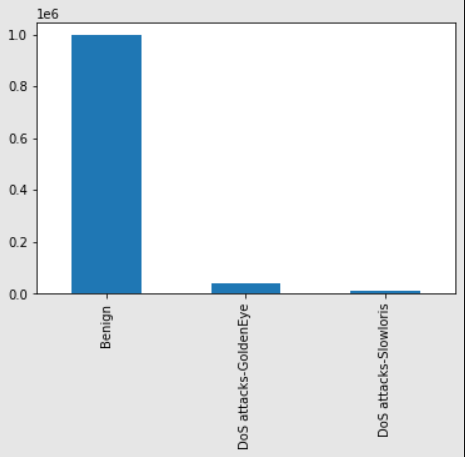
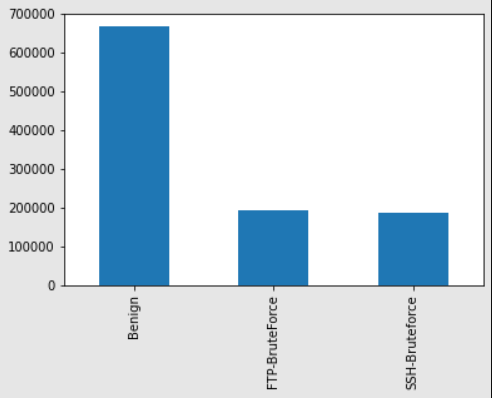
After balancing the data, we had to extract important features from the dataset which contribute the most for the model in identifying the data point and classifying it as an anomaly or a normal request. As we have applied the correlation matrix on all the 42 features, we found that there are 11 features which are highly correlated with the other features, hence we removed them and after we have obtained the independent features and balanced attack classes, we have applied various machine learning algorithms to train and test the data.

After observing the results, we thought that ensembling the data would provide better results as it would improve the data and help the models analyze better. Hence we have applied a bagging algorithm to ensemble the features and obtained the results.

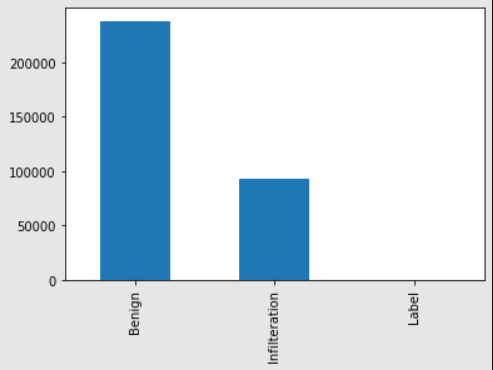
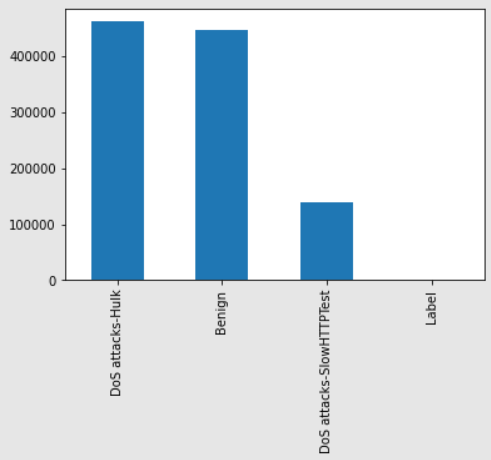
* **CSE-CIC-IDS2018 Dataset**

After analyzing the standard NSL-KDD dataset, we have chosen the most recent big data, openly accessible intrusion detection dataset that covers a variety of attacking methods, that is the CIC-IDS 2018 dataset. CIC-IDS2018 Dataset includes seven different attack scenarios of the network from inside. The attacking infrastructure includes 50 machines and the victim organization has 5 departments and includes 420 machines and 30 servers.

The dataset consists of more than 10 files, which contains the recorded raw data including the network traffic (Pcaps), and event logs (windows and Ubuntu event Logs) per machine along with 80 features extracted from the captured traffic using CICFlowMeter-V3. The dataset is therefore severely unbalanced and useless for analysis. In order to select the labels that contained the various types of attacks from the data collected across all days, we first looked at the data that was collected each day.



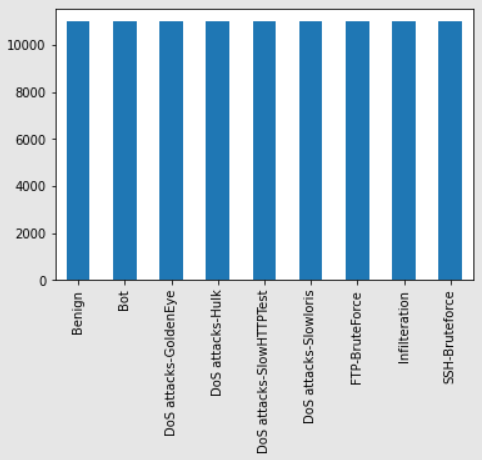
*Fig2.1 Fig2.2*



*Fig2.3 Fig2.4*

*Fig2. Attack labels of a day in CIC-IDS2018 Dataset*

The dataset has a very large number of data points(as in Fig2), hence undersampling must be done. We discovered that 10,990 data points represent the smallest non-negligible number of reported attacks. As a result, we equalized the number of non-negligible attack labels by undersampling them altogether. As a result, our dataset's size following preprocessing is (98910x78) with 9 different attack classes, i.e, Benign, Botnet, Infilteration, FTP-BruteForce, SSH-BruteForce, DoS attacks-Hulk, DoS attacks-GoldenEye, DoS attacks-SlowHTTPTest(as in Fig3).



*Fig3. CIC-IDS2018 Dataset after preprocessing*

This is the dataset after preprocessing with all the attack labels have the same number of data points, i.e. The dataset is balanced. Then, In order to help the model identify a data point and categorize it as an anomaly or a regular request, we had to extract crucial properties from the dataset. We removed the 24 features that were highly correlated with the other features after applying the correlation matrix to all 80 features to obtain independent features and balanced attack classes. Later, we applied a variety of machine learning algorithms to train and test the data using the independent features and balanced attack classes that we had obtained.

**Experimental Results:**

* **NSL-KDD Dataset**

| **Classifier** | **Accuracy** |
| --- | --- |
| Naive-bayes | 0.89 |
| Random Forest | 0.93 |
| Logistic Regression | 0.90 |

*Table1. Performance of the models used*

| **Classifier** | **Validation Score** | **Accuracy** | **Bagging** |
| --- | --- | --- | --- |
| **Decision Tree Classifier** | 99.6 | 98.8 | 99.1 |
| **KNN Classifier** | 99.3 | 98.9 | 98.9 |
| **Naive bayes classifier** | 72.2 | 72.2 | 98.3 |

*Table2. Performance of the Ensembled Models*

* **CSE-CIC-IDS2018 Dataset**

| **Classifier** | **Validation Score** | **Accuracy** |
| --- | --- | --- |
| **Decision Tree Classifier** | 0.8506 | 0.849 |
| **SVM Classifier** | 0.852 | 0.854 |
| **Random Forest Classifier** | 0.842 | 0.843 |

*Table1. Performance of the models used on CIC-IDS2018 Dataset*

**Conclusion and Future Work**

The significant rise in computer networks and network applications has been accompanied by a noticeable rise in cyberattacks. Predictive models have been developed using a number of intrusion detection datasets, notably CIC-IDS2018. CIC-IDS2018 is class-imbalanced, multi-class, and has around 16,000,000 instances. Based on this dataset, we actively looked for relevant papers as late as September 22, 2020.

When performance ratings were available for each trial, we generally noticed that they were unusually high. This can be the result of overfitting. Additionally, we observe that only a small number of the papers evaluated examined solutions for the CIC-IDS2018 class imbalance. Class imbalance can affect the outcomes of an experiment, especially with massive data.

Finally, we want to stress how poorly the CICIDS2018 data cleaning was done compared to our expectations. This issue has an impact on how reproducible experiments are.

The current research has a few holes that have been found. The literature is lacking in discussions of massive data processing frameworks, concept drift, and transfer learning.

These gaps should be filled by future research.

**References**

[1] Ahmad, Zeeshan, et al. "*Network intrusion detection system: A systematic study of machine learning and deep learning approaches*." Transactions on Emerging Telecommunications Technologies 32.1 (2021): e4150.[Link](https://onlinelibrary.wiley.com/doi/full/10.1002/ett.4150)

[2] Kovarasan, Rajesh Kambattan, and Manimegalai Rajkumar. "*An effective intrusion detection system using flawless feature selection, outlier detection and classification*." Progress in Advanced Computing and Intelligent Engineering. Springer, Singapore, 2019. 203-213.[Link](https://link.springer.com/chapter/10.1007/978-981-13-1708-8_19)

[3] Gao, X., Shan, C., Hu, C., Niu, Z., & Liu, Z. (2019). ”*An adaptive ensemble machine learning model for intrusion detection*”. IEEE Access, 7, 82512-82521.[Link](https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8740962)

[4] Kumar, Neeraj, and Upendra Kumar. "*Knowledge computational intelligence in network intrusion detection systems*." Knowledge Computing and Its Applications. Springer, Singapore, 2018. 161-176.[Link](https://link.springer.com/chapter/10.1007/978-981-10-6680-1_8)

[5] Farhan, R. I., Maolood, A. T., & Hassan, N. (2020). “*Performance analysis of flow-based attacks detection on CSE-CIC-IDS 2018 dataset using deep learning*”. Indones. J. Electr. Eng. Comput. Sci, 20(3), 1413-1418.[Link](https://www.researchgate.net/profile/Abeer-Tariq-2/publication/344905847_Performance_analysis_of_flow-based_attacks_detection_on_CSE-CIC-IDS2018_dataset_using_deep_learning_Deep_learning_Flow-based_intrusion_detection_Internet_of_thing_IOT/links/5f98744ca6fdccfd7b84a7e4/Performance-analysis-of-flow-based-attacks-detection-on-CSE-CIC-IDS2018-dataset-using-deep-learning-Deep-learning-Flow-based-intrusion-detection-Internet-of-thing-IOT.pdf)

[6] Seth, S., Chahal, K. K., & Singh, G. (2021). ”*A novel ensemble framework for an intelligent intrusion detection system*”. IEEE Access, 9, 138451-138467.[Link](https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9551879)

[7] Leevy, J. L., & Khoshgoftaar, T. M. (2020). “*A survey and analysis of intrusion detection models based on cse-cic-ids 2018 big data*”. Journal of Big Data, 7(1), 1-19.[Link](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00382-x)