INTRUSION DETECTION USING ML
A SURVEY

GROUP-14

1814061 TIRTH THAKER

1814062 GOPALKRISHNA WAJA

1814063 DHAIRYA UMRANIA

SOMAIYA
VIDYAVIHAR UNIVERSITY

Somaiya
TRUST

CONTENTS

- ABSTRACT
- INTRODUCTION
- DATASET DESCRIPTION
- SUPERVISED LEARNING
- UNSUPERVISED LEARNING
- ENSEMBLE LEARNING
- DISCUSSION
- CONCLUSION
- REFERENCES

ABSTRACT

- In our literature survey paper, we have studied various machine learning techniques for intrusion detection.
- We have broadly classified papers based on the use of machine learning technique into supervised, unsupervised and ensemble learning algorithms and compared their performance in detecting intrusions.
- For supervised learning, we have surveyed algorithms such as: KNN, SVM, Logistic regression, naive bayes classifier. For unsupervised learning we looked at algorithms such as: k-means, isolation forest, EM, SOM. For ensemble learning we have looked at a diverse range of boosting, bagging and stacking algorithms.
- We have also given a dataset description of each dataset that has been covered for any of the algorithms included in the paper.

INTRODUCTION

- An intruder is an attacker or unauthorized user who tries to steal or damaging confidential data from the victim’s system.
- Detection techniques can either be signature-based or anomaly-based.
- Anomaly based systems use ML to train a model which uses statistical quantities to define the thresholds of different parameters for normal activity and any deviations from these normal thresholds can indicate the occurrence of a malicious activity
- Over the past decade researchers have experimented with plenty of ML techniques with the aim of augmenting the detection-rate, minimizing false positives and maximizing the accuracy of IDS.

DATASET DESCRIPTION

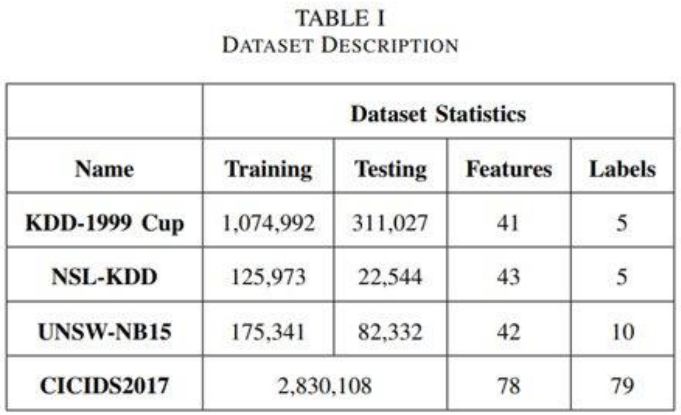
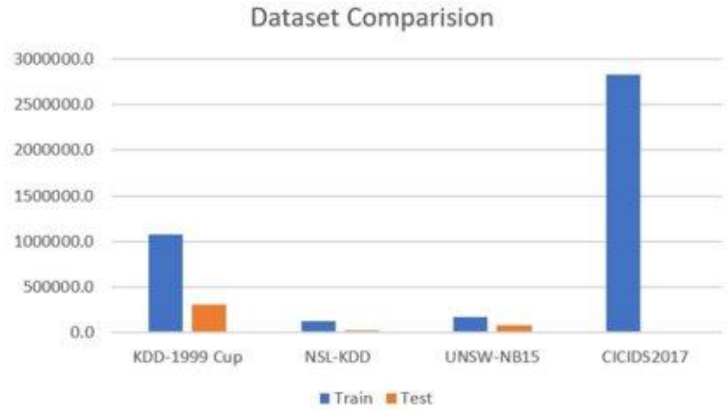


TABLE I
DATASET DESCRIPTION

	Dataset Statistics			
Name	Training	Testing	Features	Labels
KDD-1999 Cup	1,074,992	311,027	41	5
NSL-KDD	125,973	22,544	43	5
UNSW-NB15	175,341	82,332	42	10
CICIDS2017	2,830,108		78	79



- KDD-1999 Cup dataset

It is constructed by collecting data on a wide range of intrusions simulated in a military network environment. The dataset comprises of 41 features with 5 labels: DOS,R2L, U2R and probing.

- Canadian network traffic NSL-KDD dataset

An enhancement of the KDD-CUP 1999 dataset and is a standard dataset which is used for experimental analysis and training models for an Intrusion detection system

UNSW-NB15

Created by Cyber Range Lab of the Australian Site of Cyber curity. The instances mentioned in the dataset have 42 atures and are categorized in one of the 10 attack classes

CICIDS2017

Simulation of a network system which is used in the real world, and it consists of 78 features with 79 labels

SUPERVISED LEARNING

- Anomaly base Intrusion detection systems over the past few years have generally made use of labelled dataset for training a model. This approach in which the machine learns for a set of labelled instance is called Supervised machine learning. Here, in case of IDS the dataset instances can be visualized as a pair containing host or network features and the category or class to which the instance belongs to
- Paper 1 (UNSW-NB 15 dataset for binary classification using 18 out of 42 features selected using feature selection technique):
  - KNNeighbors (KNN)
  - Support Vector Machine
- Paper 2(UNSW-NB 15 dataset for multi-class classification (10 attack Classes)):
  - Random Forest Classifier
  - Decision Tree
  - Multinomial Logistic
  - Multinomial Naive Bayes

RESULTS

- Paper 1:
  - **KNN**: The model finds the K nearest neighbors of the example using some distance metric like the Euclidean or Manhattan distance and then uses the majority label of these nearest neighbors to classify the instance. **Accuracy of 89.4 %, AUC score of 0.97.**
  - **SVM**: algorithms tries to identify a set of datapoints from the training examples known as support vectors which can be used for the creation of an hyperplane such that the margin of the hyperplane is maximum which helps to obtain a decision boundary which most optimally separates the two classes **Accuracy of 92%, AUC score of 0.97.**
- Paper 2:
  - Random Forest: It uses an ensemble of decision trees which are trained by subsets created by bootstraping the original dataset**. Accuracy of 0.85 received**
  - Decision Tree: In this learning algorithm a tree is constructed where the leaf nodes represent the class labels and the internal nodes represent the features. Here the internal nodes are selected based on the entropy of the attributes. Once a node is selected the dataset is split based on the values of the selected attribute and this process is repeated recursively until a pure class is received. **Accuracy of 0.86**
  - Multinomial Logistic Regression: is an extension of Logistic Regression which can be used for multiclass classification where cross entropy can be used as the loss function for optimization of the model. For this model an **accuracy of 0.76** was received.
  - Multinomial Naive Bayes: It is used for multi-class classification by making use of the posterior probability of the class labels **Accuracy of 0.71**

UNSUPERVISED LEARNING

Unsupervised learning is a subset of machine learning algorithms where the inferences are made from the input dataset and not from the tagged labels. Unsupervised learning algorithms use techniques such as clustering to group the data and draw inferences from them. It uses the features of each tuple to check how similar they are and groups the most similar ones together. A few examples of unsupervised learning that we will have a look at in this paper are: K-means clustering, isolation forest, Expectation Maximization and Self-organizing map.

- Paper 1 (UNSW-NB 15 dataset) :
  - Here the authors have used K-means clustering and Isolation forest to find out the different types of attacks we can identify from the dataset. For K-means three attributes are used for clustering, the optimal number of clusters was found out to be 5, based on the elbow method. For Isolation forest the number of trees they decided to use in the forest were 100, they chose 256 samples to select the estimate value of each tree,
- Paper 2 (CICIDS2017 testing dataset) :
  - Authors have used the CICIDS2017 testing dataset for their unsupervised machine learning models: k-means, EM and SOM. Their k-means clustering model found out the best number of clusters to be 4, and maximum iterations to be 300

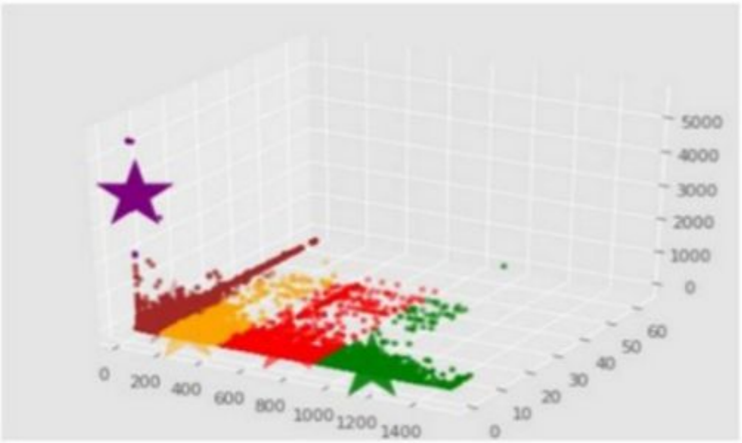


Fig. 2. K-means Clusters formed in paper [1]

RESULTS

- Paper 1:
- K-means: Optimal Clusters: 5
- Isolation Forest: Accuracy of 0.45, AUC score of 0.49.

- Paper 2:
- K-means clustering: Optimal Clusters: 4, Accuracy of 0.23.
- EM classifier: Optimal Clusters: 4, Accuracy = 0.6.
- SOM: Accuracy of 0.59.

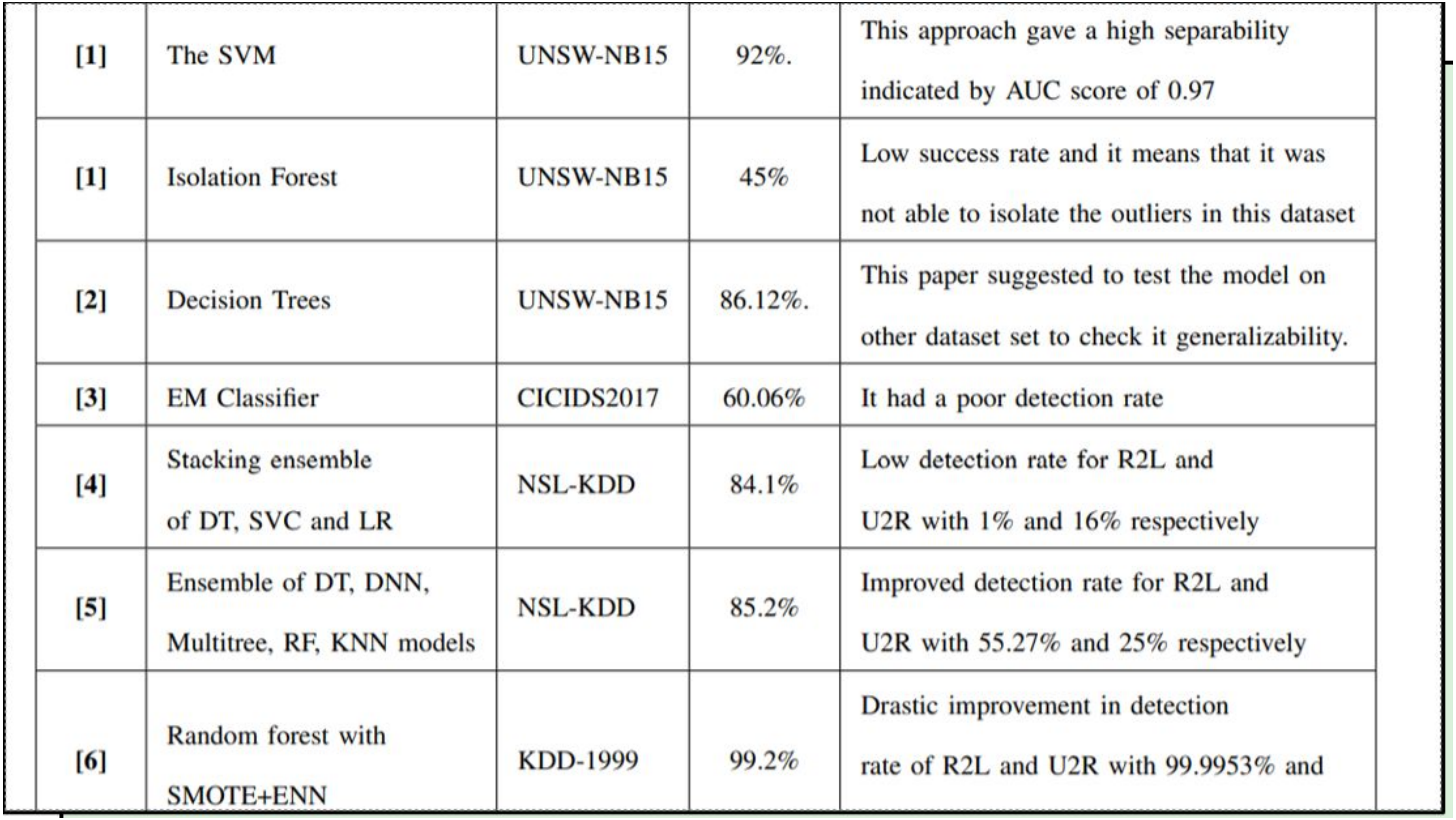
ENSEMBLE LEARNING

Ensemble Learning is a type of supervised learning technique in which a combination of ML models known as weak learners is used to create a strong learner, where individual decisions of the weak learners are combined to get to the consensus.

- Paper 1 (NSL-KDD dataset):
  - Here they have compared the performance of individual base learners with 5 different Stacking ensembles created by using different permutations of Decision trees, SVC, Logistic regression and gradient boost algorithms.
- Paper 2 (NSL-KDD dataset)
  - Authors have trained several ML models like KNN, Decision trees, random forest, KNN, logistic regression, SVM, DNN and multi-tree on the NSL-KDD train dataset and then they have created ensemble of 5 model selected 5 models giving the best results to be on the cross-validation test and here instead of using hard voting, they have used weighted voting where the weights to the base learners were given based on the results of the cross-validation.
- Paper 3 (KDD-1999 dataset with 25 features present after feature selection):
  - The authors here have used SMOTE+ENN imbalance reduction technique in combination with Random forest which is a bagging based ensemble on the KDD-1999 dataset with 25 features present after feature selection based on variance to enhance the performance of models to classify R2L and U2R attacks

RESULTS

- Paper 1:
  - Using recall as the evaluation metric they found that the ensemble learners outperformed the individual base learners and the best result of 84.1% accuracy was obtained for the ensemble using Decision Tree, KNN and Logistic regression as the base learners. The performance of this ensemble methods was high for the normal ,DoS and probe classes they were not able to identify the examples instances belong to R2L and U2R classes and had a recall of 1% and 16% respectively.
- Paper 2:
  - The result of including sophisticated models like DNN, Multitree and Random forest and using a weighted voting was that there was a significant increase in the performance measures than those obtained in previous paper with R2L and U2R classes giving a recall of 55.27% and 25% respectively.
- Paper 3:
  - Following the methodology in this paper the authors received a dramatic increase in the recall for classification of R2L and U2R for 55.27\% to 99.9953\% and 25\% to 99.997\% respectively



[1]	The SVM	UNSW-NB15	92%.	This approach gave a high separability indicated by AUC score of 0.97
[1]	Isolation Forest	UNSW-NB15	45%	Low success rate and it means that it was not able to isolate the outliers in this dataset
[2]	Decision Trees	UNSW-NB15	86.12%.	This paper suggested to test the model on other dataset set to check it generalizability.
[3]	EM Classifier	CICIDS2017	60.06%	It had a poor detection rate
[4]	Stacking ensemble of DT, SVC and LR	NSL-KDD	84.1%	Low detection rate for R2L and U2R with 1% and 16% respectively
[5]	Ensemble of DT, DNN, Multitree, RF, KNN models	NSL-KDD	85.2%	Improved detection rate for R2L and U2R with 55.27% and 25% respectively
[6]	Random forest with SMOTE+ENN	KDD-1999	99.2%	Drastic improvement in detection rate of R2L and U2R with 99.9953% and

DISCUSSION

- From the surveyed papers the performance of algorithms based on supervised learning is far better than the performance of unsupervised algorithms.
- The supervised Learning give consistently an accuracy better than 80% and therefore when these supervised algorithms are amalgamated to form ensembles a remarkable improvement in the performance is observed with accuracy going up to 99% and it should be used for intrusion detections.
- Some classifiers perform better on 1 type of dataset when compared to others and hence before using any model for a practical implementation of an IDS the performance of the model on several datasets must be studied.
- Finally, To avoid data reduction and leverage huge datasets like the CICIDS2017 the effect of deep learning models for Intrusion detection systems can be studied.

OUR IMPLEMENTATION

DATASET DESCRIPTION

KDD-1999 Cup dataset:

- This dataset was provided during the 3rd International Knowledge Discovery and Data Mining Tools Competition and it has standard data on a gamut of intrusions simulated in a military network environment. The dataset has 4,898,431 network-traffic training instances out of which 1,074,992 are unique and 311,027 network-traffic test instances out of wich 77,289 are unique with each instance having 41 features and one of the 5 labels: DOS, R2L, U2R and probing. A DoS attack leads to denial of services to legitimate users. Probe attacker is the type of attack in which the attacker collects sensitive information and fingerprints of the target machine which can be used to plan future attacks. In R2L, the attacker's aim is to obtain unauthorized access to the target system while in U2R, the attacker gains access to an unauthorized machine which is then used to obtain root access.

MODELS AND ALGORITHMS

- Naive Bayes:

  Naive Bayes is a classification algorithm that is suitable for binary and multiclass classification. It is a supervised classification technique used to classify future objects by assigning class labels to instances/records using conditional probability

- Decision Tree

  Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features

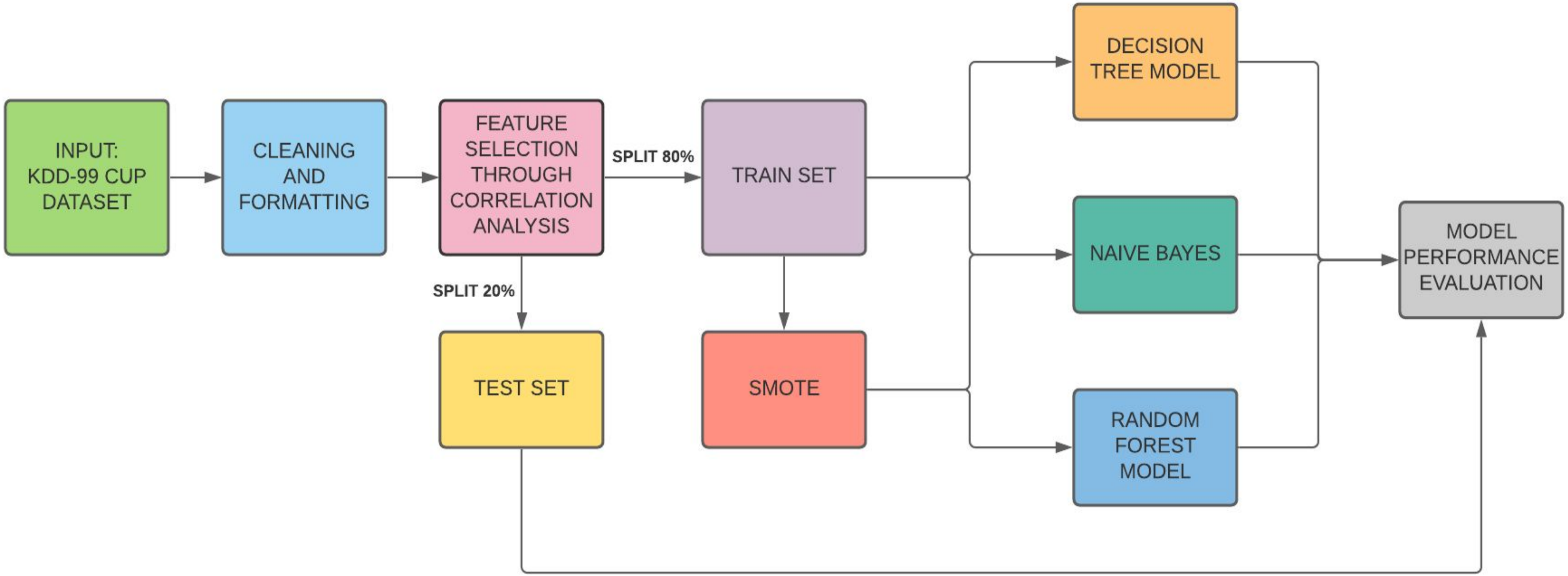
- Random Forest

  It is ensemble model made of many decision trees using bootstrapping, random subsets of features, and average voting to make predictions

- SMOTE

  It is one of the most commonly used oversampling methods to solve the imbalance problem. It aims to balance class distribution by randomly increasing minority class examples by replicating them

PIPELINE



INPUT: KDD-99 CUP DATASET

CLEANING AND FORMATTING

FEATURE SELECTION THROUGH CORRELATION ANALYSIS

SPLIT 80%

SPLIT 20%

TRAIN SET

TEST SET

SMOTE

DECISION TREE MODEL

NAIVE BAYES

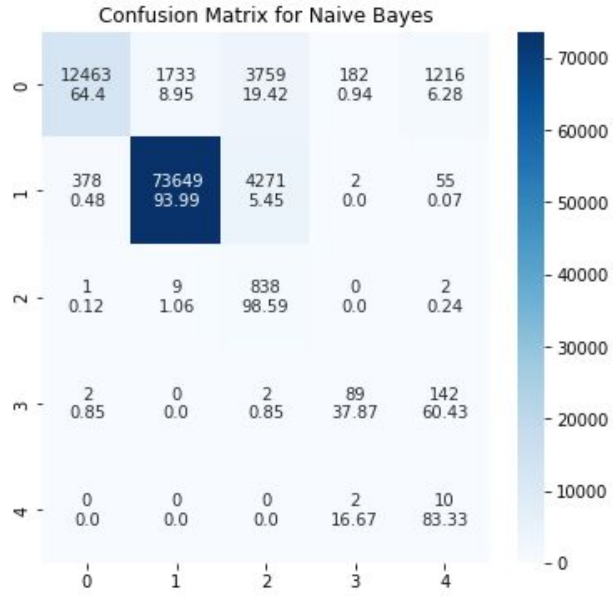
RANDOM FOREST MODEL

MODEL PERFORMANCE EVALUATION

RESULTS

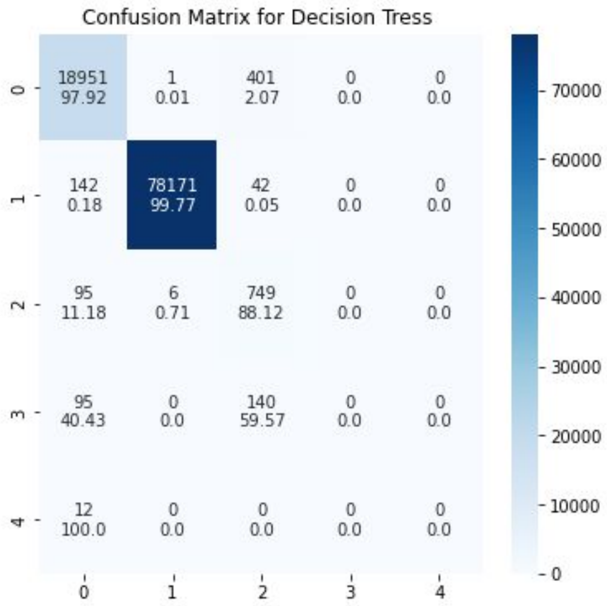
MODEL	ACCURACY	F1	TIME(s)
Naive Bayes	0.88	0.45	0.184
Decision Tree	0.99	0.53	1.233
Random Forest	1.00	0.95	12.122
Naive Bayes + SMOTE	0.79	0.78	0.969
Decision Tree + SMOTE	0.94	0.94	9.665
Random Forest + SMOTE	1.00	1.00	104.879

Confusion Matrices:



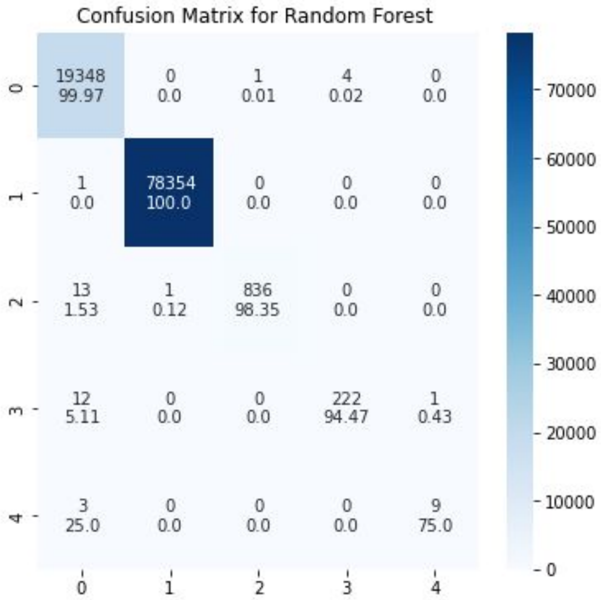
Confusion Matrix for Naive Bayes

	0	1	2	3	4
0	12463 64.4	1733 8.95	3759 19.42	182 0.94	1216 6.28
1	378 0.48	73649 93.99	4271 5.45	2 0.0	55 0.07
2	1 0.12	9 1.06	838 98.59	0 0.0	2 0.24
3	2 0.85	0 0.0	2 0.85	89 37.87	142 60.43
4	0 0.0	0 0.0	0 0.0	2 16.67	10 83.33



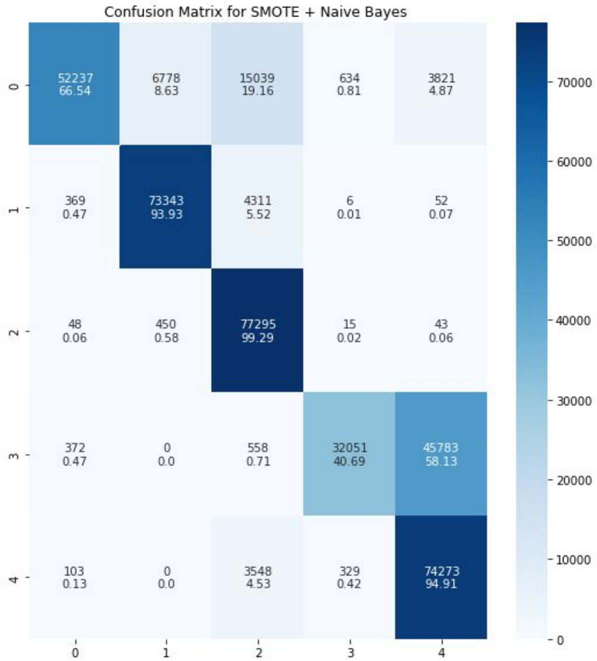
Confusion Matrix for Decision Tress

	0	1	2	3	4
0	18951 97.92	1 0.01	401 2.07	0 0.0	0 0.0
1	142 0.18	78171 99.77	42 0.05	0 0.0	0 0.0
2	95 11.18	6 0.71	749 88.12	0 0.0	0 0.0
3	95 40.43	0 0.0	140 59.57	0 0.0	0 0.0
4	12 100.0	0 0.0	0 0.0	0 0.0	0 0.0



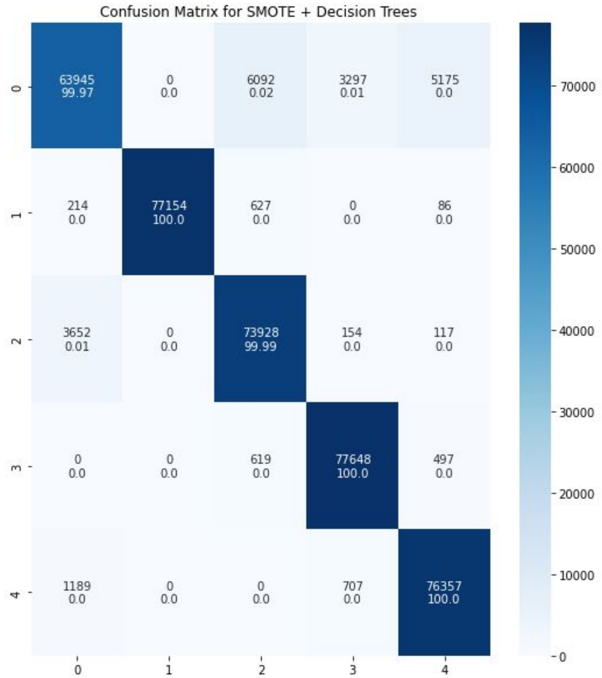
Confusion Matrix for Random Forest

	0	1	2	3	4
0	19348 99.97	0 0.0	1 0.01	4 0.02	0 0.0
1	1 0.0	78354 100.0	0 0.0	0 0.0	0 0.0
2	13 1.53	1 0.12	836 98.35	0 0.0	0 0.0
3	12 5.11	0 0.0	0 0.0	222 94.47	1 0.43
4	3 25.0	0 0.0	0 0.0	0 0.0	9 75.0



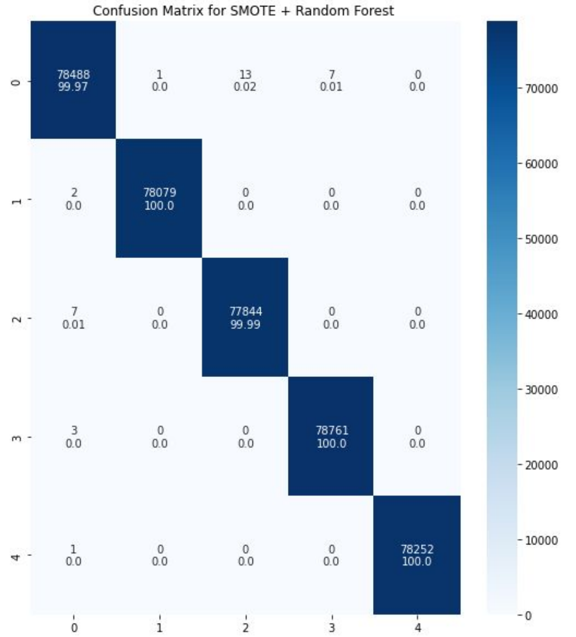
Confusion Matrix for SMOTE + Naive Bayes

	0	1	2	3	4
0	52237 66.54	6778 8.63	15039 19.16	634 0.81	3821 4.87
1	369 0.47	73343 93.93	4311 5.52	6 0.01	52 0.07
2	48 0.06	450 0.58	77295 99.29	15 0.02	43 0.06
3	372 0.47	0 0.0	558 0.71	32051 40.69	45783 58.13
4	103 0.13	0 0.0	3548 4.53	329 0.42	74273 94.91



Confusion Matrix for SMOTE + Decision Trees

	0	1	2	3	4
0	63945 99.97	0 0.0	6092 0.02	3297 0.01	5175 0.0
1	214 0.0	77154 100.0	627 0.0	0 0.0	86 0.0
2	3652 0.01	0 0.0	73928 99.99	154 0.0	117 0.0
3	0 0.0	0 0.0	619 0.0	77648 100.0	497 0.0
4	1189 0.0	0 0.0	0 0.0	707 0.0	76357 100.0



Confusion Matrix for SMOTE + Random Forest

	0	1	2	3	4
0	78488 99.97	1 0.0	13 0.02	7 0.01	0 0.0
1	2 0.0	78079 100.0	0 0.0	0 0.0	0 0.0
2	7 0.01	0 0.0	77844 99.99	0 0.0	0 0.0
3	3 0.0	0 0.0	0 0.0	78761 100.0	0 0.0
4	1 0.0	0 0.0	0 0.0	0 0.0	78252 100.0

Conclusion

- Without any under/over-sampling techniques, Naive Bayes and Decision Tree have poor results, especially in R2L and U2R classes which are grossly undersampled.

- Random Forest still performs relatively well.

- But with the SMOTE oversampling technique, the results of Naive Bayes, Decision Tree and Random Forest are drastically improved.

CONCLUSION

- The implementation of various machine learning models were surveyed to combat intrusion in network systems.ML Models like supervised learning models such as SVM, unsupervised learning models like k-means, isolation forest, and ensemble models such as random forest were used.
- The ensemble models in general and the Random forest with SMOTE+ENN in particular, was the best model in terms of accuracy for detecting intrusions in network systems.
- There is still a lot of room for further studies and improvement, intruders are also continuously evolving their strategies of intrusion, hence the data must be regularly updated so that the machine learning models can keep up with them.
- Different ML models and dataset can be tried and researched to improve constantly.

REFERENCES

- [1] F. G. Portela, F. Almenares Mendoza and L. C. Benavides, "Evaluation of the performance of supervised and unsupervised Machine learning techniques for intrusion detection," 2019 IEEE International Conference on Applied Science and Advanced Technology (iCASAT), 2019, pp. 1-8, 2019.
- [2] A. Srivastava, A. Agarwal and G. Kaur, "Novel Machine Learning Technique for Intrusion Detection in Recent Network-based Attacks," 2019 4th International Conference on Information Systems and Computer Networks (ISCON), 2019, pp. 524-528, 2019.
- [3] Z. K. Maseer, R. Yusof, N. Bahaman, S. A. Mostafa and C. F. M. Foozy, "Benchmarking of Machine Learning for Anomaly Based Intrusion Detection Systems in the CICIDS2017 Dataset," in IEEE Access, vol. 9, pp. 22351-22370, 2021.
- [4] M. Raihan-Al-Masud and H. A. Mustafa, "Network Intrusion Detection System Using Voting Ensemble Machine Learning," 2019 IEEE International Conference on Telecommunications and Photonics (ICTP), pp. 1-4, 2019.
- [5] X. Gao, C. Shan, C. Hu, Z. Niu and Z. Liu, "An Adaptive Ensemble Machine Learning Model for Intrusion Detection," in IEEE Access, vol. 7, pp. 82512-82521, 2019.
- [6] T. Lu, Y. Huang, W. Zhao and J. Zhang, "The Metering Automation System based Intrusion Detection Using Random Forest Classifier with SMOTE+ENN," 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT) pp. 370-374, 2019.
- [7] U. S. Musa, M. Chhabra, A. Ali and M. Kaur, "Intrusion Detection System using Machine Learning Techniques: A Review," 2020 International Conference on Smart Electronics and Communication (ICOSEC), pp. 149-155, 2020.
- [8] U. S. Musa, S. Chakraborty, M. M. Abdullahi and T. Maini, "A Review on Intrusion Detection System using Machine Learning Techniques," 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), pp. 541-549, 2021.