ANALYSIS OF VARIOUS SUPERVISED MACHINE LEARNING ALGORITHMS FOR INTRUSION DETECTION

KABIR NAGPAL, NIYATI JAIN, AYUSH PATRA, ARNAV GUPTA, ANJANA SYAMALA, SUNITA SINGHAL

DEPARTMENT OF CSE, MANIPAL UNIVERSITY JAIPUR, JAIPUR, RAJASTHAN, INDIA

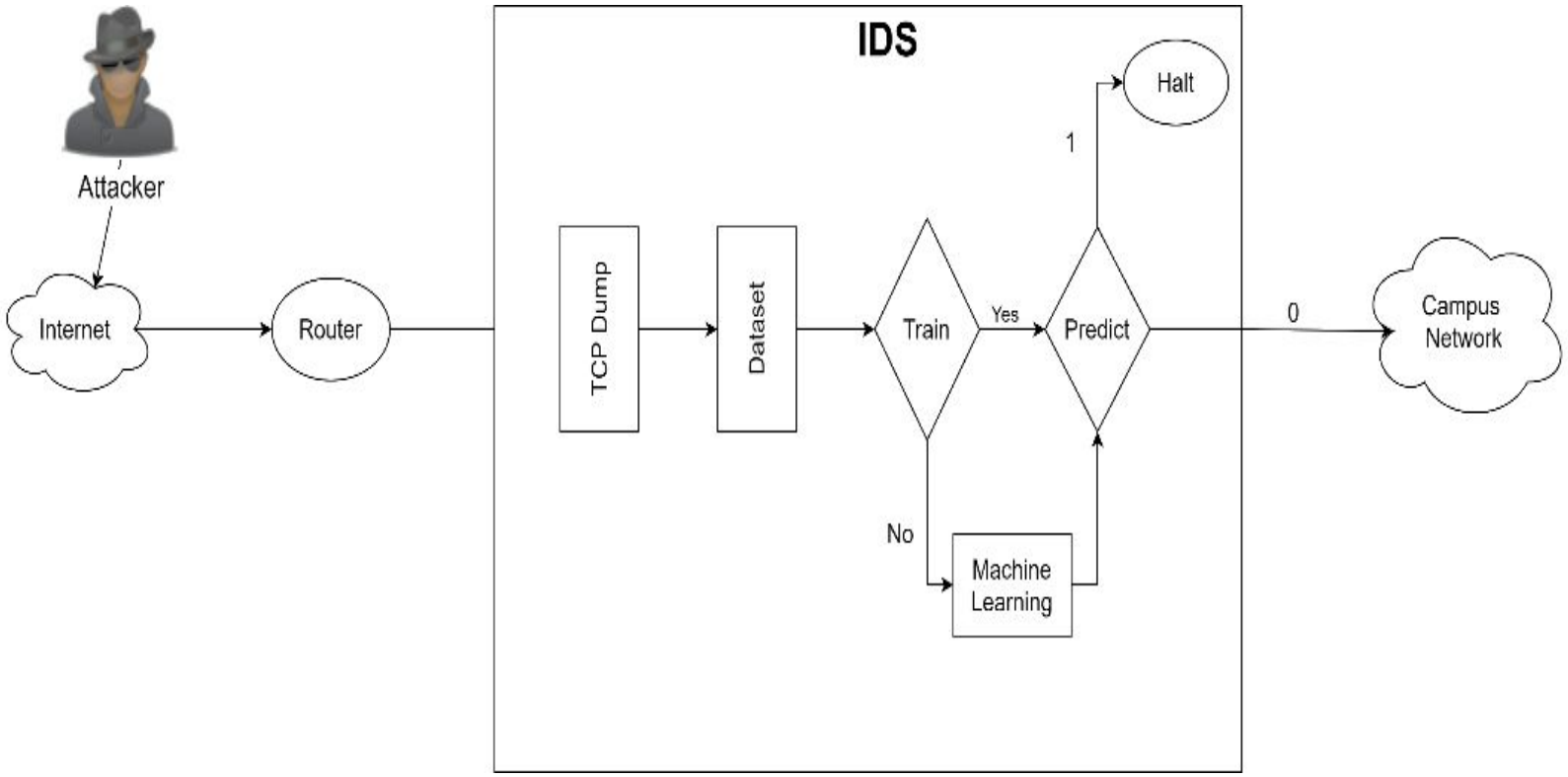
INTRODUCTION

- Intrusion Detection is the ability to block threats and furthermore detect them in modern-day technologies.
- They are capable of Threat classification, Anomaly Detection, Network Traffic Processing, etc.
- There are further classified into two different methods that are:

1. Anomaly-based detection
2. Misuse detection.

- It is a classification task that looks into the behavior of network traffic into normal and anomalous.
- Moreover this paper will use the Feature Engineering strategies to improve the optimal features like decreasing the FAR(False Alarm Rate) and to increase detection accuracy.

A PICTORIAL REPRESENTATION OF INTERNET SETUP WITH IDS



THE UNSW-NB15 DATASET

- The UNSW-NB15 comprises 42 features out of which 3 instances are categorical and 39 are numeric.
- The data categories include normal data and nine types of attacks. The UNSW-NB15 includes the following types of network attacks: Backdoor, Shellcode, Reconnaissance, Worms, Fuzzers, DoS, Generic, Analysis, Shellcode, and Exploits.

ISSUES WITH THE DATASET

1. Unbalanced: F-score and FAR are used to represent the efficiency of the algorithm since the dataset consists of 56,000 samples of label 0 (no attack) and 119,341 of label 1.

a. F-score is the balanced mean of recall and precision.

b. FAR is the probability of falsely rejecting the null hypothesis for a particular test.

2. Skewness: Most columns in the dataset are right-skewed-'trans_depth' and 'response_body_len' have the highest skewness of 176.3 and 76.3, respectively and 30 more columns have a right skewness of greater than 1. The right skewness is corrected using log-transformation while the left is corrected using cubic function.

ISSUES WITH THE DATASET

3.Quantile Transformation: This helps spread out the most frequent values and remove outliers.

4. Standard Scaling: Standardization is the necessary equipment for many algorithms like support vectors, as they assume the dataset is centered, on 0 with a unit standard deviation.

5. New columns were added:

New Feature	Formula	Description
Mean	Smean - dmean	Difference of average transmitted packet size
Pkts	Spkts - Dpkts	Difference in number of packers sent
Bytes	Sbytes - Dbytes	Difference in number of bytes sent
Loss	Sloss + Dloss	Total Packets retransmitted or dropped
Inkpt	Sinkpt - Dinkpt	Different in inter packet arrival time
Load	Sload- Dload	difference in bits transmitted per second
Jit	Sjit + djit	Total jitter
Tcpb	Stcpb - Dtcpb	Difference in TCP window advertisement
Ttl	Sttl + Dttl	Total time to live

PREPROCESSING

- Categorical columns ‘proto’, ‘service’, and ‘state’ were one-hot encoded.
- Only the following categories are kept in the dataset for each column while others are renamed to “Others” and then one hot encoded

a. Service : 'dns', 'http', 'smtp', 'ftp-data', 'ftp’

b. Proto : 'tcp', 'udp', 'unas’

c. State : 'FIN', 'INT', 'CON'

FEATURE SELECTION

- Random Forest classification is used for feature selection which is a sub category of embedded methods.
- Embedded methods are like filter and wrapper methods wherein every algorithm has their own built-in feature selection method.
- The list of twenty-one selected features selected after applying Random forest feature selection are:

1. Sbytes	2. Dbytes	3. Rate	4. Sttl
5. Dttl	6. Sload	7. Dload	8. Dinpkt
9. Tcprtt	10. Synack	11. Ackdat	12. Dmean
13. Ct_sry_src	14. Ct__state_ttl	15.Ct_dst_src_ltm	16. Ct_sry_dst
17. Bytes	18. Ttl	19. Load	20. Mean
21. State_INT			

MACHINE LEARNING ALGORITHMS

The following algorithms were used for the experiment:

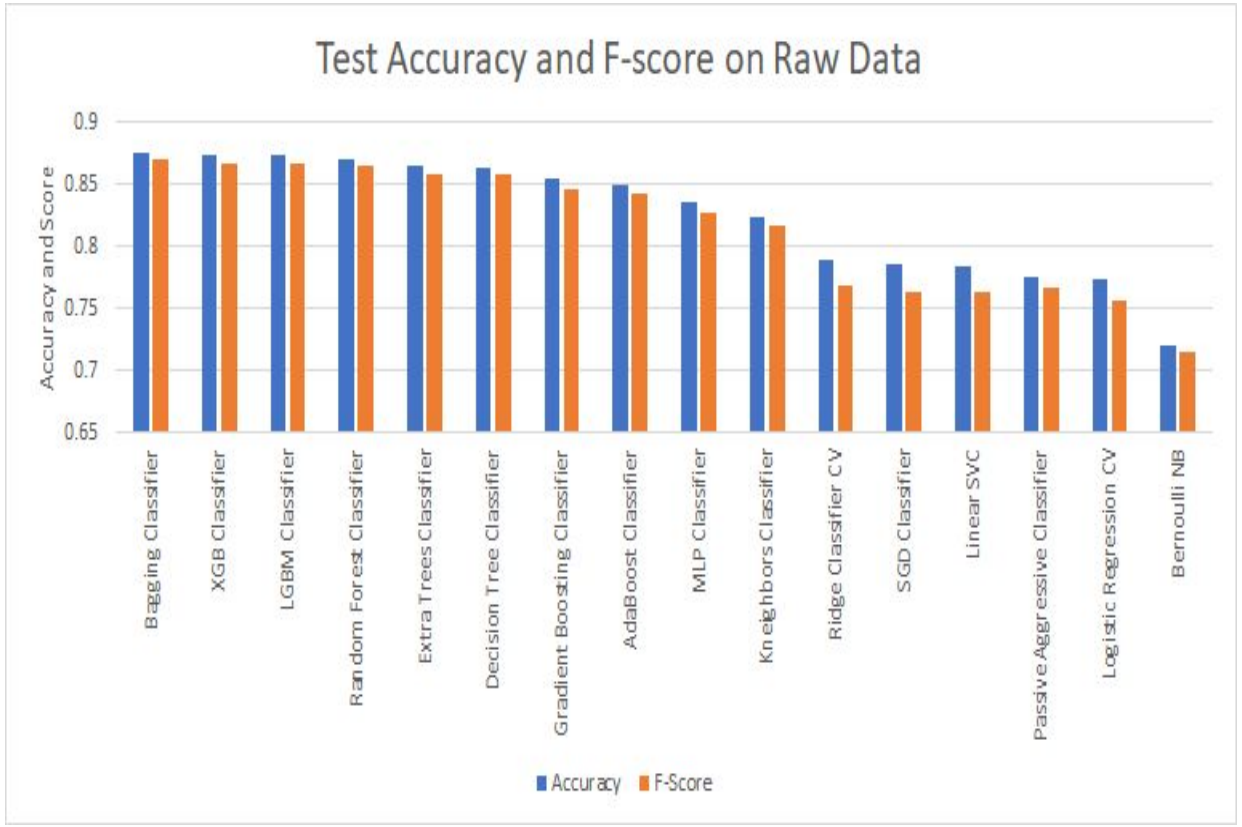
- Logistic Regression
- SVM
- Naïve Bayes
- K-Nearest Neighbor (KNN)
- Passive Aggressive
- Stochastic Gradient Descent classifier
- Multilayer Perceptron
- Decision Tree
- Random Forest
- Extra Trees Classifier
- Bootstrap Aggregating
- Gradient Boosting
- Adaptive Boosting
- Ridge regression
- XGBoost
- Light Gradient Boosting Machine (LGBM)

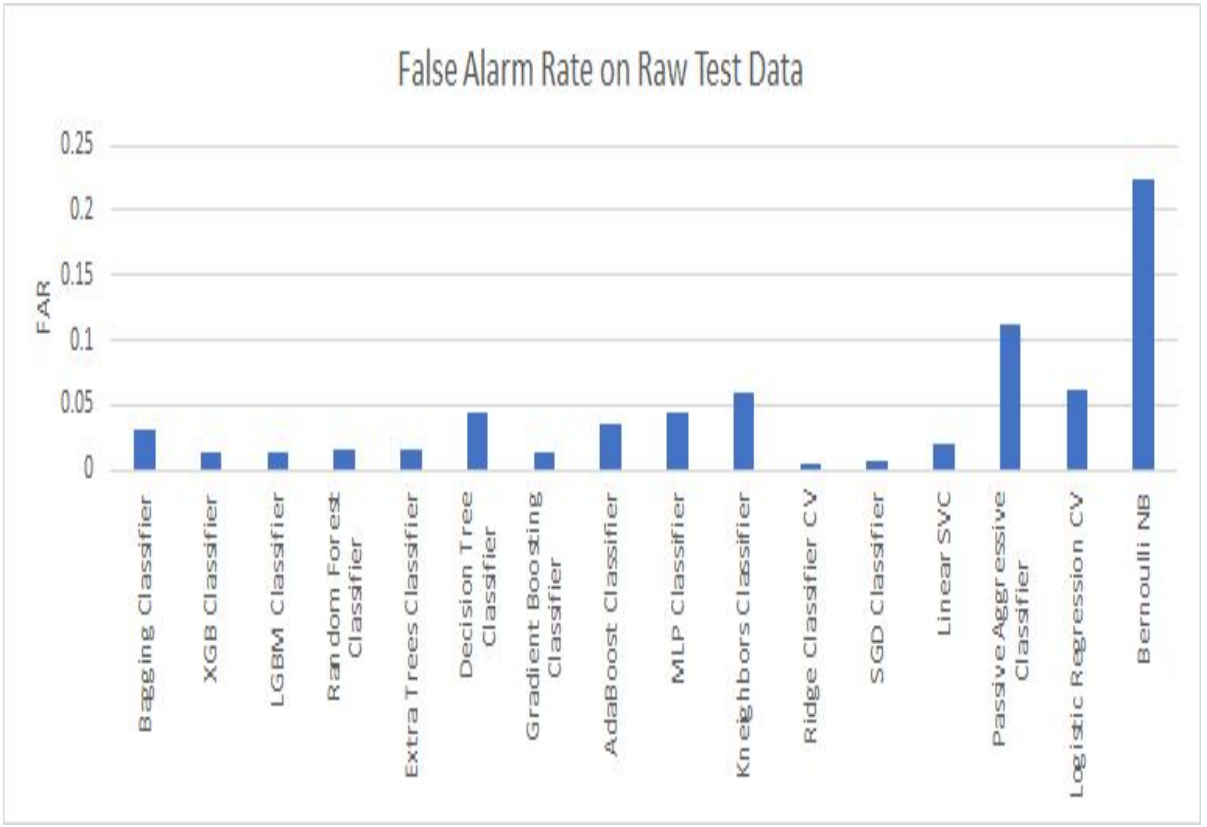
RESULTS

- Machine Learning models applied on raw dataset comes to a conclusion of attaining the best F-Score of 87% with FAR of 0.03.
- Secondly when such classifiers are applied on processed data that results in achieving a F-score and accuracy of 88.35% and 88.78% respectively.
- Different classifiers are applied on the reduced data to check out the accuracy as well.
- XGBoost and LGBM classifiers have topped the methods.

RESULTS OF ML CLASSIFIER MODELS ON RAW DATA

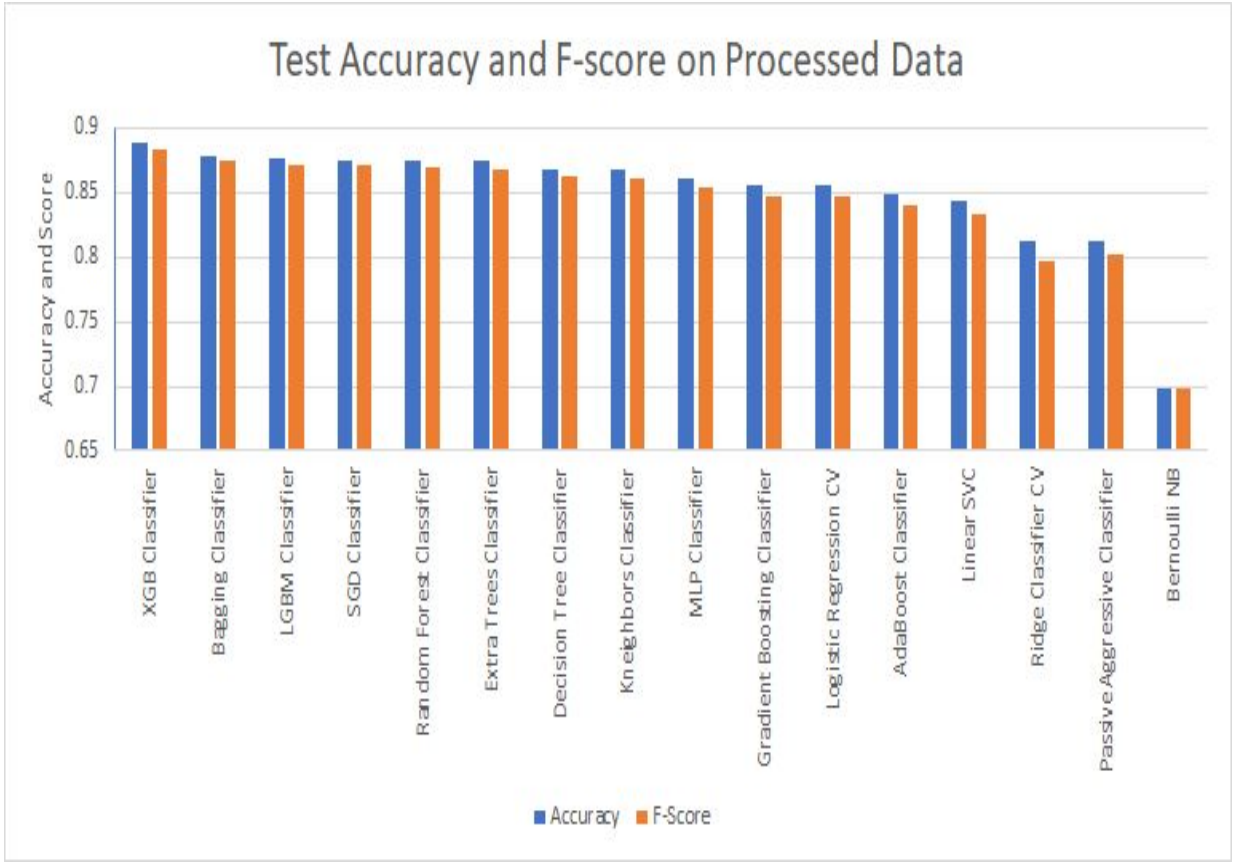
Model	F-Score	FAR	Accuracy
Bagging Classifier	**0.87079**	0.0308	**0.87554**
XGB Classifier	0.86745	0.01436	0.8734
LGBM Classifier	0.86721	0.01379	0.87322
Random Forest Classifier	0.86513	0.01511	0.87127
Extra Trees Classifier	0.85933	0.0165	0.86596
Decision Tree Classifier	0.85776	0.04454	0.86286
Gradient Boosting Classifier	0.84713	0.01361	0.85528
Ada Boost Classifier	0.84216	0.03609	0.84927
MLP Classifier	0.82786	0.04535	0.83585
KNN Classifier	0.81665	0.06066	0.82471
Ridge Classifier CV	0.76847	**0.00609**	0.78995
Passive Aggressive Classifier	0.76688	0.11219	0.77633
Linear SVC	0.7644	0.02142	0.78463
SGD Classifier	0.76325	0.00818	0.78551
Logistic Regression CV	0.75732	0.06168	0.77367
Bernoulli NB	0.71544	0.2245	0.72051

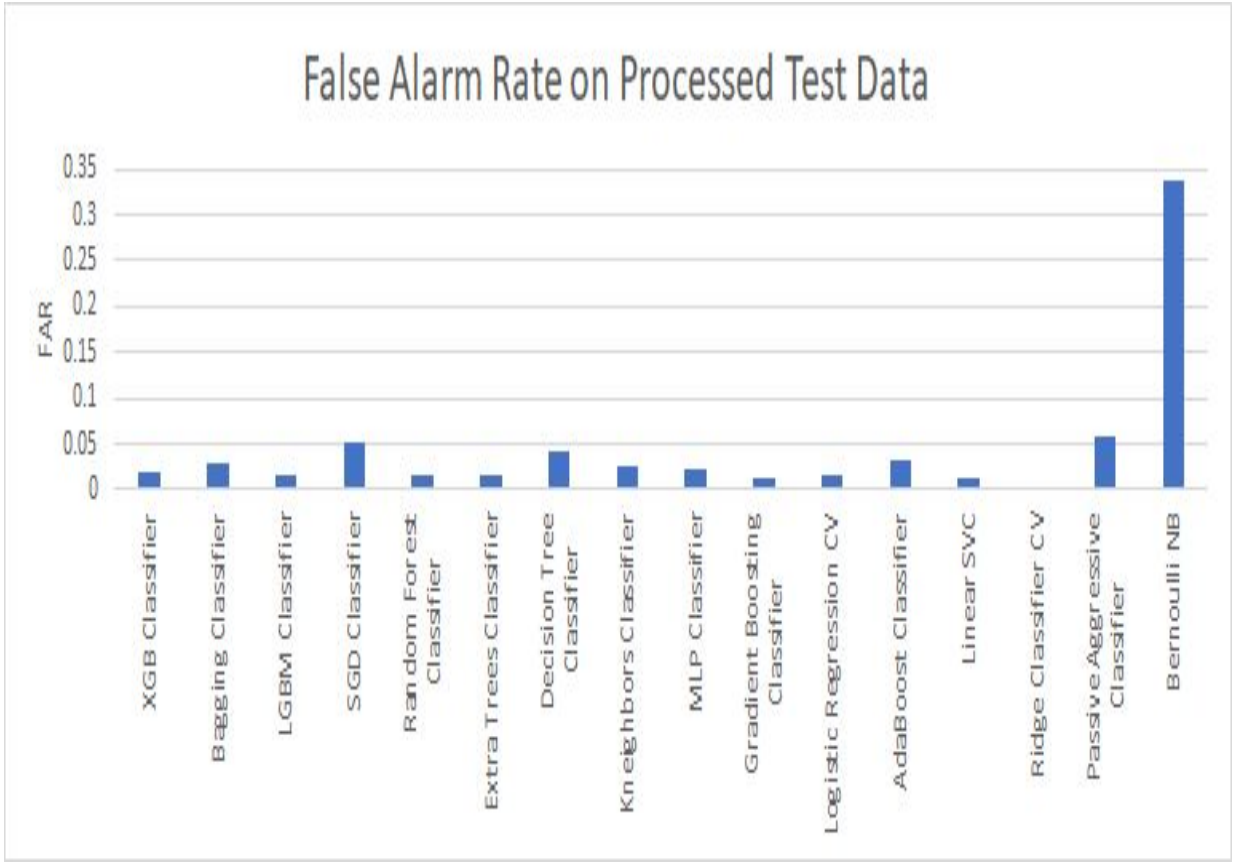

Test Accuracy and F-score on Raw Data
Accuracy and Score
0.9
0.85
0.8
0.75
0.7
0.65
Bagging Classifier
XGB Classifier
LGBM Classifier
Random Forest Classifier
Extra Trees Classifier
Decision Tree Classifier
Gradient Boosting Classifier
AdaBoost Classifier
MLP Classifier
Kneighbors Classifier
Ridge Classifier CV
SGD Classifier
Linear SVC
Passive Aggressive Classifier
Logistic Regression CV
Bernoulli NB
Accuracy
F-Score


False Alarm Rate on Raw Test Data
FAR
0.25
0.2
0.15
0.1
0.05
0
Bagging Classifier
XGB Classifier
LGBM Classifier
Random Forest Classifier
Extra Trees Classifier
Decision Tree Classifier
Gradient Boosting Classifier
AdaBoost Classifier
MLP Classifier
Kneighbors Classifier
Ridge Classifier CV
SGD Classifier
Linear SVC
Passive Aggressive Classifier
Logistic Regression CV
Bernoulli NB

RESULTS OF ML CLASSIFIER MODELS ON PROCESSED DATA

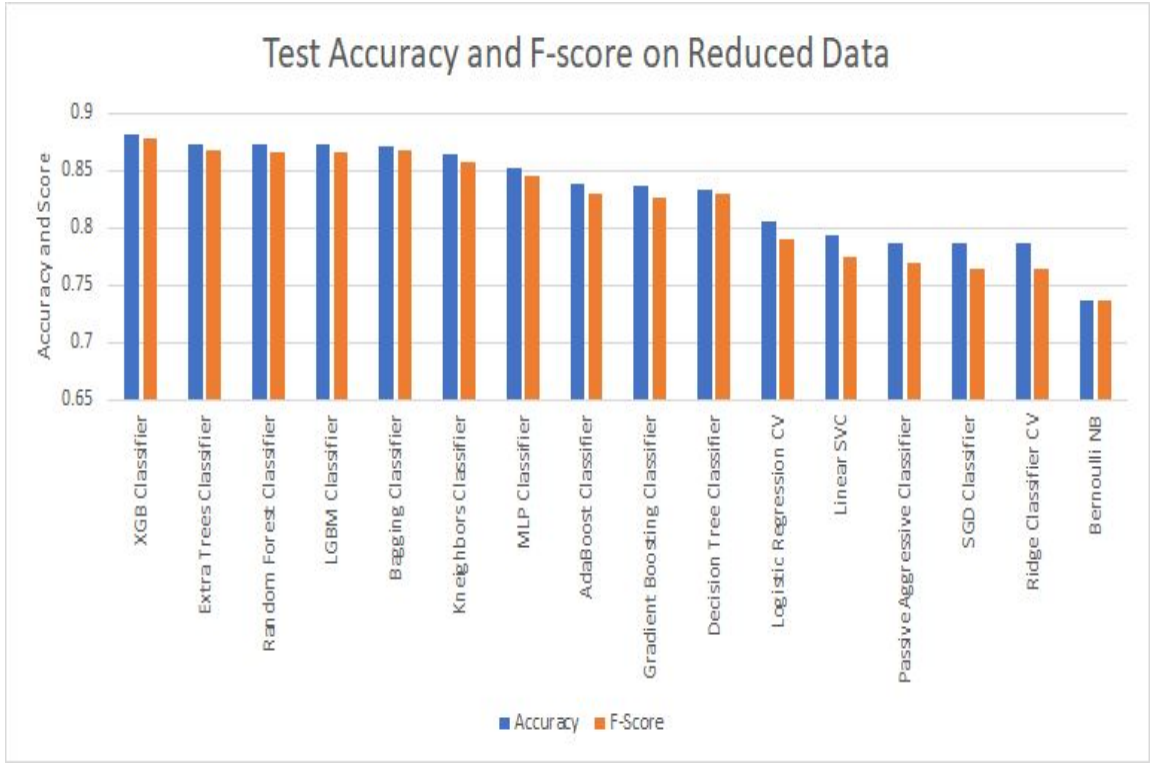
Model	F-Score	FAR	Accuracy
XGB Classifier	0.8835	0.0178	0.88787
Bagging Classifier	0.87375	0.02841	0.87838
SGD Classifier	0.87147	0.05052	0.8752
LGBM Classifier	0.87091	0.01445	0.87654
Random Forest Classifier	0.86874	0.01637	0.87446
Extra Trees Classifier	0.8682	0.01672	0.87395
Decision Tree Classifier	0.86207	0.04185	0.86695
KNN Classifier	0.86108	0.02722	0.86688
MLP Classifier	0.85316	0.02371	0.85996
Gradient Boosting Classifier	0.84692	0.01204	0.85521
Logistic Regression CV	0.84659	0.01507	0.8547
Ada Boost Classifier	0.84012	0.03124	0.84779
Linear SVC	0.8324	0.01114	0.84258
Passive Aggressive Classifier	0.80171	0.05943	0.81166
Ridge Classifier CV	0.79595	0.00185	0.8127
Bernoulli NB	0.69769	0.33912	0.69795

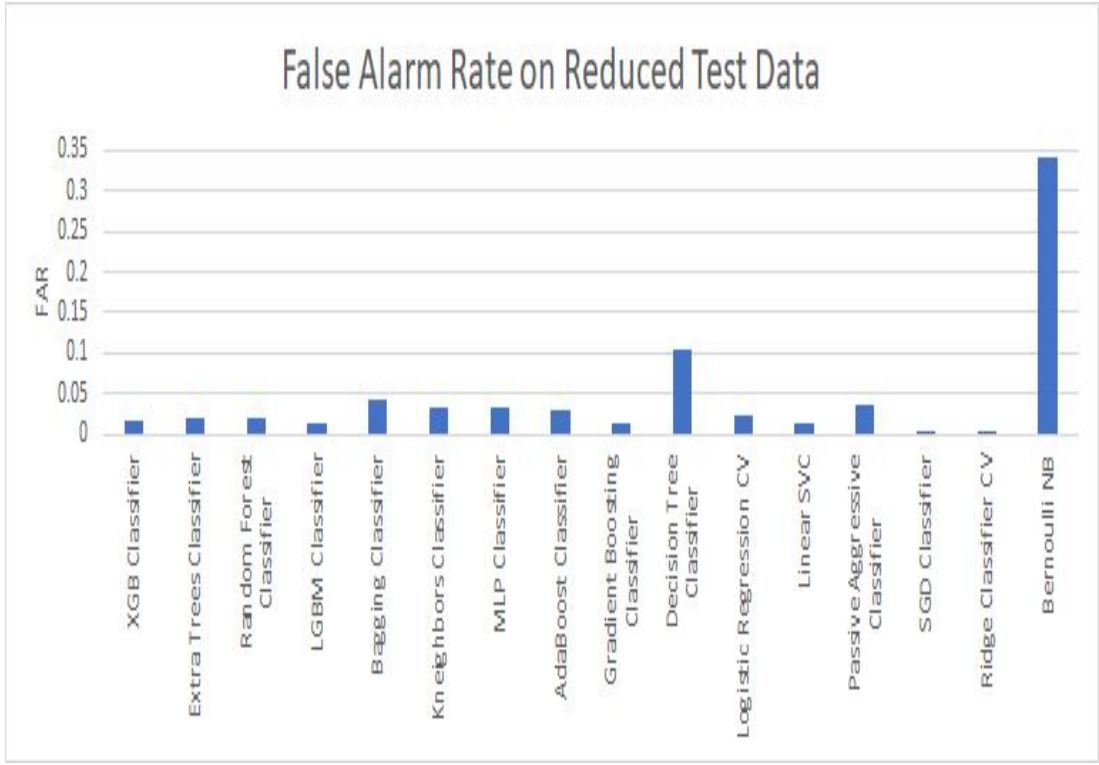

Test Accuracy and F-score on Processed Data
Accuracy and Score
0.9
0.85
0.8
0.75
0.7
0.65
XGB Classifier
Bagging Classifier
LGBM Classifier
SGD Classifier
Random Forest Classifier
Extra Trees Classifier
Decision Tree Classifier
Kneighbors Classifier
MLP Classifier
Gradient Boosting Classifier
Logistic Regression CV
AdaBoost Classifier
Linear SVC
Ridge Classifier CV
Passive Aggressive Classifier
Bernoulli NB
Accuracy
F-Score


False Alarm Rate on Processed Test Data
FAR
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
XGB Classifier
Bagging Classifier
LGBM Classifier
SGD Classifier
Random Forest Classifier
Extra Trees Classifier
Decision Tree Classifier
Kneighbors Classifier
MLP Classifier
Gradient Boosting Classifier
Logistic Regression CV
AdaBoost Classifier
Linear SVC
Ridge Classifier CV
Passive Aggressive Classifier
Bernoulli NB

RESULTS OF ML CLASSIFIER MODELS ON REDUCED DATA

Model	F-Score	FAR	Accuracy
XGB Classifier	**0.8774**	0.01729	**0.88229**
Extra Trees Classifier	0.86771	0.02038	0.87329
Bagging Classifier	0.86748	0.04262	0.87188
Random Forest Classifier	0.86684	0.01853	0.87261
LGBM Classifier	0.86646	0.01454	0.8725
KNN Classifier	0.85821	0.03373	0.86389
MLP Classifier	0.84501	0.03309	0.85203
Ada Boost Classifier	0.83021	0.03002	0.83911
Decision Tree Classifier	0.82919	0.10472	0.83318
Gradient Boosting Classifier	0.82619	0.01271	0.83706
Logistic Regression CV	0.78962	0.02402	0.80494
Linear SVC	0.7742	0.01211	0.79374
Passive Aggressive Classifier	0.76962	0.03545	0.78706
SGD Classifier	0.7641	**0.00243**	0.787
Ridge Classifier CV	0.7634	0.00377	0.78626
Bernoulli NB	0.7362	0.34234	0.73624


Test Accuracy and F-score on Reduced Data
Accuracy and Score
0.9
0.85
0.8
0.75
0.7
0.65
XGB Classifier
Extra Trees Classifier
Random Forest Classifier
LGBM Classifier
Bagging Classifier
Kneighbors Classifier
MLP Classifier
AdaBoost Classifier
Gradient Boosting Classifier
Decision Tree Classifier
Logistic Regression CV
Linear SVC
PassiveAggressive Classifier
SGD Classifier
Ridge Classifier CV
Bernoulli NB
Accuracy
F-Score


False Alarm Rate on Reduced Test Data
FAR
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
XGB Classifier
Extra Trees Classifier
Random Forest Classifier
LGBM Classifier
Bagging Classifier
Kneighbors Classifier
MLP Classifier
AdaBoost Classifier
Gradient Boosting Classifier
Decision Tree Classifier
Logistic Regression CV
Linear SVC
PassiveAggressive Classifier
SGD Classifier
Ridge Classifier CV
Bernoulli NB

CONCLUSION

- Presents an analysis of the UNSW-NB15 dataset with several models.
- Intuitive pre-processing techniques along with State of the art Machine Learning algorithms outperform traditional methods and scores.
- Despite XGBoost's better accuracy, LGBM is preferred due to lower FAR.
- This paper focuses more on critical processing, rather than on feature selection methods.
- Accuracy can be increased by using complex feature selection with mentioned pre-processing techniques to increase the accuracy of detection.
- Decision tree-based approaches may be invented and used as observations prove better performance.