Reduced Features Intrusion Detection Systems Classification Accuracy Improvement

*OLASEHINDE Olayemi O.
Department of Computer Science
Federal Polytechnic
Ile Oluji, Ondo State, Nigeria
Olaolasehinde@fedpolel.edu.ng

Williams Kehinde
Department of Computer Science
Federal Polytechnic
Ile Oluji, Ondo State, Nigeria
Kehwilliams@fedpolel.edu.ng

Adegoke, B. O.
Department of Computer Engineering
Federal Polytechnic
Ile Oluji, Ondo State, Nigeria
benadegoke@fedpolel.edu.ng

ABSTRACT

Intrusion detection system plays an important role in network security, machine learning algorithms have severally been used to build intrusion detection models, the accuracy of the intrusion detection models (classifiers) does not only depend on the classification algorithm but also on the selected relevant features used to train the machine learning algorithms, the presence of irrelevant and redundant features used to built intrusion detection models are major causes of decreasing detection rate and high false alarm rate. Feature selection (FS) techniques are preprocessing techniques used to select the relevant features of a dataset to build intrusion detection system. This paper investigates the classification accuracy improvement of intrusion detection models built with reduced features obtained from three features selection techniques; Consistency FS, Correlation FS and Information Gain FS over the intrusion detection models built using the whole features of the UNSW-NB15 intrusion dataset. Each of the reduced features and the whole features were used to train three machine learning algorithms; K Nearest Neighbour, Decision Tree and Naïve Bayes. Decision tree models recorded the highest classification accuracy among all models, 75.71% with the whole features model, accuracy of 86.77% with consistency reduced features, accuracy of 87.18% with information gain reduced features and accuracy of 85.30% with correlation reduced features. Naive Bayes models recoded the least classification accuracy 56.04% with the whole features, 70.20% with consistency reduced features, 69.59% with information gain reduced features and 66.74% with correlation reduced features. Naive Bayes Models recorded the highest classification accuracy improvements of 25.27% with consistency reduced features model, 24,18% with information gain reduced features model and 19.09% with correlation reduced features model over the accuracy of the whole feature models. Decision Tree models recorded the least classification accuracy improvement of 14.61% with consistency reduced feature model, 15.15% with information gain reduced features model and 12.67% with correlation reduced features model.

Keywords: *Intrusion, Feature Selection, Relevant Analysis, Redundancy Analysis, Relevant Features, Machine Learning Algorithms, Accuracy, Classification Improvement*

1. INTRODUCTION

Intrusion Detection Systems (IDSs) are network security tools and predictive models used to analyse and classif network traffics as either normal or intrusive. IDSs can further be used to classify intrusive traffics into various attacks categories. IDSs are built by the application of Machine Learning (ML) algorithms to learn from the features of network traffics or intrusion detection dataset. Machine learning methods have difficulty in dealing with large number of input features, which poses an interesting challenge for researchers. Feature Selection (FS) is one of the most frequent and important techniques in data preprocessing, and has become an indispensable component of the ML processes[1]. The accuracy of the intrusion detection models (classifiers) does not only depends on the classification algorithm but also on the selected relevant features used to build the IDS classifier. The irrelevant and redundant features can confuse the classifiers and lead to incorrect results if they are not remove from the feature used to train the classification model, Figure 1 shows the framework of feature selection process. The Relevant analysis carried out on the intrusion dataset is used to determine relevant features subset that are highly correlated to the target class (Attack types), the redundancy analysis is used to extract and remove redundant feature from the selected relevant features subset in order to obtain optimal feature subset. The most correlated feature subset to the target class (attack type) is used to train the ML algorithm. Feature selection methods seeks to identify features or feature subset according to the given evaluation criterion, that are highly correlated to determine the target class of a given unlabelled dataset instance or network packet.

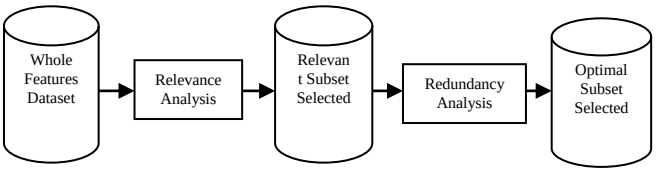

Figure 1: Framework of Feature Selection Process

FS is an efficient way to reduce the dimensionality of a problem. It is part of the ML models preprocessing stage,. many researchers have used FS to improve the classification accuracy and computational speed of intrusion detection models. This research work investigated the classification accuracy improvement of Intrusion Detection models built from the reduced feature subset of consistency, correlation and information gain FS techniques and the whole features of the UNSW-NB15 Intrusion Dataset.

2. LITERATURE REVIEW

Network packets are made up of several features, some of these feature(s) are not relevant to determination of the target class while some others are redundant. The Presence of redundant and irrelevant features are the major causes of low classification accuracy and high false alarm rate. FS is the techniques used to select relevant feature set that are highly correlated to the determination of the target class. FS improves classification accuracy and efficiency of the IDS, reduces the computational time for the IDS, and reduces the dimensionality of the dataset. Building a reliable IDS involves the removal of noisy, irrelevant and redundant features from the features of the network packets or intrusion detection dataset [2]. The selected reduced features subset are used to train the ML algorithm to build the IDS model

According to [3], selected relevant features from input dataset simplifies the intrusion detection problem with improved classification accuracy detection rate. [4] proposed an algorithm for consistency-based feature selection, experiments were performed with several large datasets to compare the efficiency of the algorithm against INTERACT and LCC. There are two instances of consistency-based algorithms with potential real world applications. The algorithm performed better in terms of time efficiency and accuracy when compared with the result of INTERACT and LCC. [5] develops a measure that is monotonic and fast to compute and select the best relevant feature subset. This guaranteed the search for relevant features to be complete but not exhaustive. An empirical study was conducted to show that the algorithm indeed lives up to what it claims. [6] presents a FS using Genetic Algorithm with Support Vector Machine (SVM) for mining the medical dataset, results from the work show that models built with reduced features gives an higher diagnoses rate and lower miscalculation rate. Findings from [7] shows an improved performance in terms of efficiency, accuracy, with reduced selected features and understandability of the learning process. Empirical evaluation of the consistency feature selection measures with wrapper method, shows consistency method performing efficiently more than the wrapper method. [8] proposes a feature selection method using ant colony optimization for face recognition system. In this approach, the nearest neighbor classifier was adopted for evaluating the generated subset using ant colony optimization based learning . In this paper, we applied consistency, correlation and information gain FS techniques on UNSW-NB15 dataset to obtain relevant reduced features/attributes used to build K nearest Neigbour , Naive Bayes and Decision Tree IDS to classify incoming Network Packets or Intrusion Detection Dataset into attacks and normal packet.

2.1 FEATURE SELECTION TECHNIQUES

Three (3) feature selection methods; Information gain, consistency based method and correlation based were used to identify reduced features attributes that can be used to determine the class label (attacks and attacks categories) of UNSW-NB15 dataset, the consistency based method being a subset selector, select the best subset of the feature attributes that is best in determining the target class, while the information gain and correlation based were attributes ranking selector, ranking used scoring function which measures the relevant between each features to the target feature to orders the features.

2.1.1 Consistency Features Selection Technique

Consistency measures the attempt to find a minimum number of features that distinguish between the classes as consistently as the full set of features. An inconsistency arises when multiple training samples have the same feature values, but different class labels.

Given a training sample *S* the *inconsistency count(IC)* of an instance subset $A \in S$ is given in equation 1

$$IC_{X'}(A) = X'(A) - \max_k X'_k(A) \quad (1)$$

Where $X'(A)$ is the number of instances in S equal to subset A using only the features in $X'$and $X'_k(A)$is the number of instances in S of class k equal to A using only the features inX'.

By summing all the inconsistency counts and averaging over the size of the training sample size, a measure called the inconsistency rate (IR) for a given subset is defined. The *inconsistency rate* of a feature subset A in a sample S is given by equation 2

$$IR(X') = \frac{\sum_{A \in S} IC_{X'}(A)}{|S|} \quad (2)$$

2.1.2 Correlation Features Selection Technique

Correlation features extraction generates all the possible subset S of the dataset and then calculate Merit M for each of the subset S, using equation 2, the subset with the highest Merit will be selected and returned as follows;

$$M_s = \frac{k\,\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}} \quad (3)$$

where $\overline{r_{cf}}$ is the average attack categories to features, $\overline{r_{ff}}$ is the average features to features correlations and k is the number of features in the subset S

2.1.3 Information Gain Features Selection Technique

Information Gain (IG) measures the amount of information in bits about the class prediction, it examines each feature independently and evaluate its information gain and how important and relevant it is to the class label. Information Gain (IG) for attribute(s) x is given by;

$$IG(X) = H(Y) - H(Y \mid X) \quad (4)$$

Where H(y) is Entropy of Y and H(Y | X) is Entropy of Y given X

$$H(Y) = -\sum_{i=1}^{n} p(y_i) log_2 p(y_i) \quad (5)$$

$$H(Y \mid X) = -\sum_{i=1}^{n} p(x_i) \sum_{j=i}^{k} p(y_i \mid x_i)\, log_2\, p(y_i \mid x_i) \quad (6)$$

Where n: is number of instance in the UNSW-NB15 dataset
k: is the number of attack categories in the UNSW-NB15 dataset

$P(y_i)$: is the probability of occurrence of attack categories value of instance i

$P(y_i|x_i)$: is the probability of attack categories value of instance i will occur given the occurrence of attribute value x of instance i

2.2 ML CLASSIFICATION ALGORITHMS

Three ML classification algorithms used in this work are; Naive Bayes, K Nearest Neighbor and C4.5 Decision Tree, each of these algorithms were used to build classification models with the whole features training dataset and each of the three (3) reduced features dataset generated from the three (3) features selection methods;

2.2.1 K-Nearest Neighbour

K-Nearest Neighbour is based on Euclidean distance between the training set and the testing set. Given that p_i is the instance to be classified having features ranging from 1 to *n*, q_t is the other instances in a data set ranging from 1 to *k* with having the same number of features as P. The Euclidean distance between p_i and q_t can be defined as:

$$d(p_i, q_t) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \quad (7)$$

From equation (3), a given instance will be classified as the attack categories having the majority attacks among top n closest instance to the given instance.

2.2.2 Naive Bayes

Naive Bayes classification is expressed as follows; given the UNSW-NB15 dataset that have X number *of* attributes called the predictors ($X = x_1, x_2,...,x_n$) and another attribute y called the class label, having ten members that ranges from $y_1,....y_{10}$, the probability that a class y_j with be assigned to a given unlabeled instance X is given as follows;

$$p(y_j \mid x_1,.,x_{43}) = \frac{p(y_j)p(x_i|y_j)}{p(x_i)} \quad (\forall_j = 0,1) \quad (8)$$

Maximum posterior probability for classifying attack categories to a new instance is given as:

$$y = \arg\max_{y} py_j \prod_{j=0}^{1} p(y_j) p(x_1, x_2, .x_{43} \mid y_j) \quad (9)$$

Where : y_i is the attack class,

$x_1, x_2 \ldots, x_{43}$: are the predictor attributes of UNSW-NB15 dataset

2.2.3 Decision Tree

Decision Tree (DT), (C4.5) is a classification model consisting of nodes that are attribute names of UNSW-NB15 and arcs which are attribute values connection to other nodes all the way to the leaves which are the attack categories (class label). Decision Tree (DT) builds a classification tree, which will be used to predict the attack categories of a new instance in the test dataset; DT calculates the Gain Ratio of all the attributes of the training dataset, by dividing the information gain of an attribute with splitting value of that attribute. The formula for Gain Ratio is;

$$Gain\ Ration\ A_i = \frac{Information\ \ Gain\ A_i}{Split\ Information\ A_i} \quad (10)$$

Split value of an attribute is chosen by taking the average of all the values in the domain of current attribute. It is given by (11)

$$Split\ info(A_i) = -\sum_{i=1}^{n} \frac{|t_j|}{|T|} \cdot \log_2 \frac{|t_j|}{|T|} \quad (11)$$

Where |T| is the number of values of the current attribute
t is the values of attributes A_i
n is the number of values in attribute A_i

3.0 UNSW-NB15 DATASET

UNSW-NB15 dataset was used to evaluate the KNN IDS, This dataset was published in 2015 and it contains nine different moderns attack types with the normal connection, its training datasets contain 82, 332 records while the testing dataset contain 175, 341 records, [9]. Table 1 shows the Percentages Distribution of Attacks and Normal Connections in both the Training and Testing Datasets. According to [10], The UNSW-NB15 data set is reliable and effective than NSLDD and KDD datasets in detecting existing and new attacks, other advantages of this dataset include; the similarity in the probability distribution of the training and testing sets, it contains real modern normal behaviors and contemporary synthesized attack activities, it features from the payload and packets header efficiently reflect real network packet, similarity between training and testing dataset and its suitability to evaluate existing and new attacks in an effective and reliable manner [11] are some of the advantages UNSW-NB15 dataset over the NSLKDD data set.

4.0 METHODOLOGY

The architecture of the proposed Intrusion detection system is shown in Figure 2, the three FS; Consistency, Information Gain and Correlation techniques are used to obtain three different reduced feature datasets. The three reduced features of the training dataset with the whole features are used to trained the three ML algorithms; K-Nearest Neighbour, Naive Bayes and Decision Tree to build the IDS models, the twelve (12) models built were evaluated with the testing Dataset

4.1 Models Performance Measurement

Confusion Matrix was used to measure the performance of the IDS models, it classification outcome has four possible outcomes, which are; True Positive (correct positive classification), True Negative (correct negative classification), False Positive (incorrect positive classification), and false negative (incorrect negative classification). Classification accuracy, false alarm rate and performance improvement are the three metrics used to measure the IDS perfromances

4.1.1 Accuracy

Accuracy (ACC) is the ratio of all correct classification to the total number of instances in the test dataset, it is given by Equation 3.24. An accuracy of 1 implies error rate of 0 and an accuracy of 0 indicate error rate of 12

$$ACC = \frac{TP + TN}{FN + FP + FN + TP} \quad (12)$$

4.1.2 False Alarm Rate (FAR)

False Positive Rate (FPR) or False Alarm Rate (FAR) is the proportion of the wrongly model negative as positive by the model, FPR should be as low as possible to avoid unwanted false alarms. it is given by Equation 13

$$\boldsymbol{FPR = FAR = \frac{FP}{TN + FP}} \quad (13)$$

4.1.3 Performance Improvement

Performance Improvement (PI) is the ratio of changes in performance to the initial performance, it is given in Equation 14

$$\boldsymbol{PI = \frac{Final\ Performance - Initiatial\ Performance}{Initial\ Performance}} \quad (14)$$

Table 1: Percentage distribution of Attacks and Normal Network Connection Categories in the Training and Test UNSW-NB15 Dataset

	Training		Testing	
Names of Attack	No. of connections	**% Distribution**	No. of Connections	**% Distribution**
Reconnaissance	3496	4.25	10491	5.98
Dos	4089	4.97	12264	6.99
Exploits	11132	13.52	33393	19.04
Shellcode	378	0.46	1133	0.65
Fuzzers	6062	7.36	18184	10.37
Backdoor	583	0.71	1746	1.00
Analysis	677	0.82	2000	1.14
Generic	18871	22.92	40000	22.81
Worms	44	0.05	130	0.07
Total No of Attacks	**45332**	**55.06**	**119341**	**68.06**
Normal	37000	44.94	56000	31.94
Total No of Connections	**82332**	**100.00**	**175341**	**100.00**

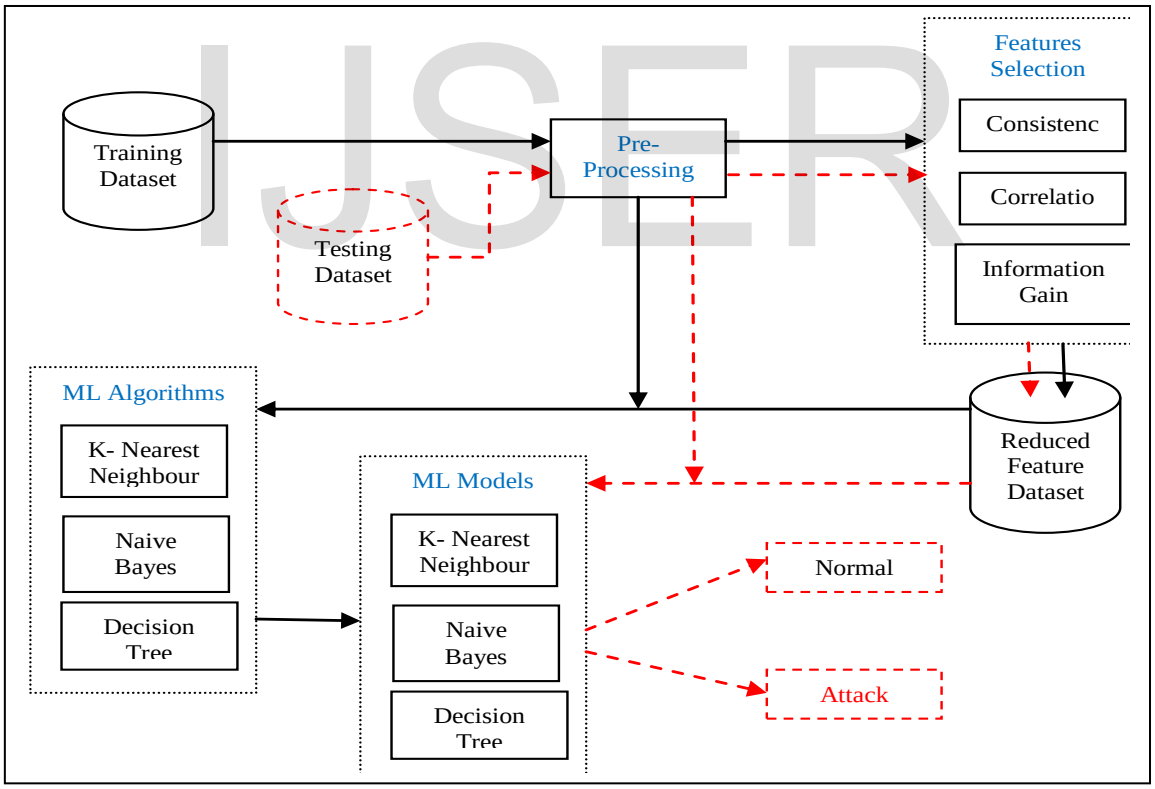

Figure 2: System Architecture of the Reduced Features Intrusion Detection System

5.0 Results and Discussion

Three feature selections techniques are applied to the UNSW-NB15 dataset, Table 2 shows the number of features selected by each of the FS techniques, Consistency selects 28 features, correlation selects 24 features, while Information Gain selects 23 least number of features. Figure 3 shows the classification accuracy of the evaluation of each intrusion detection models with the test dataset

Table 2: Features selected by the Feature Selection Techniques

All Feature (43)	**Consistency Reduced Features (28)**	**Information gain Reduced Features (23)**	**Correlation Reduced Features (24)**
Proto, state, dur, sbytes, dbytes, sttl, dttl, sloss, dloss, rate, service, sload, dload, spkts, dpkts, swin, dwin, stcpb, dtcpb, smean, dmean, trans_depth, res_bdy_len, sjit, djit, sintpkt, dintpkt, tcprtt, synack, ackdat, is_sm_ips_ports, ct_state_ttl, ct_flw_http_mthd, is_ftp_login, ct_ftp_cmd, ct_srv_src, ct_srv_dst, ct_dst_ltm, ct_src_ ltm, ct_src_dport_ltm, ct_dst_sport_ltm, ct_dst_src_ltm, attack_cat	dur, proto, service, spkts, sbytes, dbytes, rate, sttl, sload, dload, sinpkt, sjit, djit, tcprtt, synack, ackdat, smean, dmean, trans_depth, ct_srv_src, response_body_len, ct_dst_ltm , ct_src_dport_ltm, ct_srv_dst, ct_dst_sport_ltm, ct_dst_src_ltm, ct_src_ltm, attack_cat	attack_cat, sbytes, smean, sload, dbytes, service, dmean, sinpkt, synack, ct_dst_sport_ltm, proto, rate, ct_state_ttl, dur, spkts, dttl, ct_src_dport_ltm, ct_ srv_dst, dinpkt, dpkts, dload, ct_srv_src, tcprtt	attack_cat, dtcpb, stcpb, ct_dst_sport_ltm, sttl, swin, state, rate, ct_src_dport_ltm, ct_srv_dst, ct_srv_src, dwin, ct_dst_src_ltm, service, ct_dst_ltm, ct_src_ltm, ackdat, synack ct_state_ttl, proto, dttl, dload, dmean, tcprtt,

Table 3: Classification Accuracy of the Whole Feature and Reduced Features Dataset

Classification Models	Dataset with All Features (%)	Consistency Feature Reduced Set (%)	Information Gain Feature Reduced Set (%)	Correlation Feature Reduced Set (%)
Naïve Bayes	56.04	70.20	69.59	66.74
K Nearest Neighbor	70.70	82.05	82.35	82.62
Decision Tree	75.71	86.77	87.18	85.30

The three models built from the whole features dataset recorded the least classification accuracy of 56.04% with Naive Bayes, 70.70% for KNN and 75.71% for Decision Tree. Decision Tree models perform better that Naive Bayes and KNN models, it records the highest classification accuracy for each of the reduced dataset and the whole features dataset, 75.71% for the whole features dataset, 86.77% for consistency reduced dataset, 85.30% for correlation reduced dataset and the overall highest classification accuracy of 87.18% for the information Gain reduced dataset. Figure 3 shows the graphical representation performances of each of the classification models. Table 4 shows the classification improvement of the models with the reduced selected features against the models with the whole features, Naive Bayes models recorded the three best classification accuracy improvement against the whole feature models; 25.27% improvement with consistency model, 24.18% with information Gain model and 19.09% with correlation model. Decision Tree models recorded the least classification accuracy improvement of 12.67% with correlation model, 14.61% with consistency model and 15.15% with information Gain Model. Figure 4 shows the performance improvement of the reduced features models over the whole feature models

Table 4: Classification Accuracy Improvement of the Reduced Feature Dataset over Whole Feature Dataset

Classification Models	Consistency Feature Reduced Set	Information Gain Feature Reduced Set	Correlation Feature Reduced Set
Naïve Bayes	25.27%	24.18%	19.09%
K Nearest Neigbour	16.05%	16.48%	16.86%
Decision Tree	14.61%	15.15%	12.67%

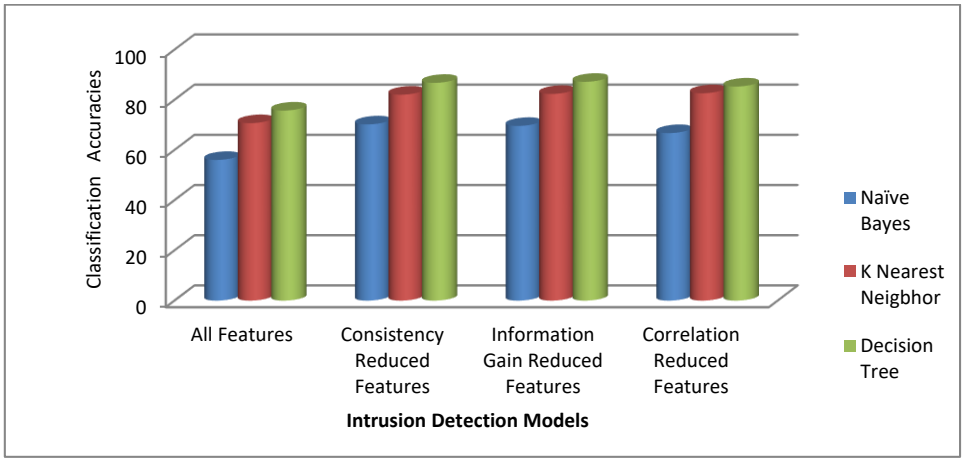

Figure 3: Classification Performances Accuracy of each of the models

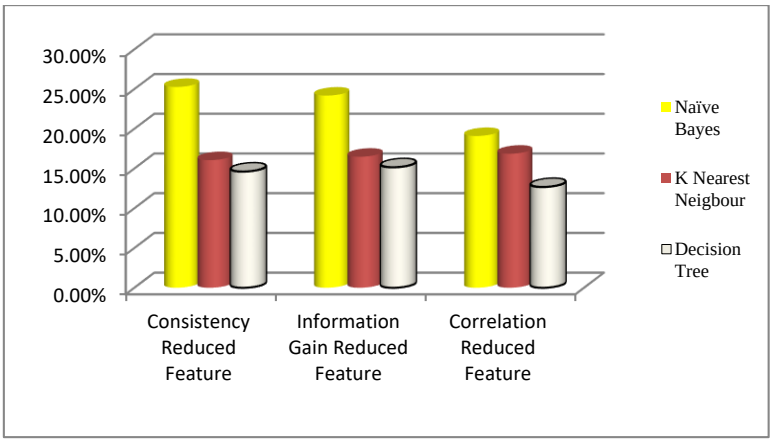

Figure 4: Performance Improvement of the Reduced Features Models over the Whole Feature Models

CONCLUSION

This work investigates the classification accuracy improvement of reduced features Intrusion Detection Systems, three features selection techniques are applied on the UNSW-NB15 dataset to select relevant features/feature subset that are capable to determine the attack label of the dataset. The selected features alongside the whole features of the dataset were used to train three ML algorithms. The models built from each of the ML algorithm training with each of the reduced and whole feature dataset are evaluated with the test dataset. The results obtained shows that all IDS built with the reduced relevant features set records higher detection accuracy than the IDS built with the whole feature set. Models built with the Information Gain reduced FS recorded the highest accuracy, closely followed by correlation FS models, while consistency has the least accuracy. Decision Tree models performs more than the other two Models, KNN models has the least classification accuracy, Decision Tree model with information Gain FS is therefore recommended for development of Intrusion Detection System.

Ethical Standard

Funding: This research work is a self funded research undertaken by the corresponding author with the other two co-authors in the Department of Computer Science, Federal Polytechnic, Ile Oluji, Ondo State, Nigeria

Conflict of Interest: The authors declare that they have no conflict of interest.

REFERENCES

[1] Kalousis A., Prados J. and Hilario M., (2007) “Stability of Feature Selection Algorithms: a study on high dimensional spaces,” Knowledge and information System, vol. 12, no. 1, pp. 95-116.

[2] Hall, M. A. (1999). Correlation-based Feature Selection for Machine Learning. The University of Waikato

[3] Adetunmbi A.O., Adeola S.O., and Daramola O. A. (2010) "Analysis of KDD ’99 Intrusion Detection Dataset for Selection of Relevance Features", Proceedings of the World Congress on Engineering and Computer Science WCECS San Francisco, USA 1: 20-22.

[4] Zhao Z. and Liu H.(2007) Searching for interacting features. In Proceedings of International Joint Conference on Artificial Intelligence, pages 1156 – 1161,

[5] Liu H., Motoda H., and Dash M. (1998) A Monotonic measure for optimal Feature Selection. In Proceedings of European Conference on Machine Learning, 1998

[6] Welikala, R.A, Fraz, MM, Dehmeshki, J, Hoppe, A, Tah, V, Mann, S, Williamson, TH & Barman, SA (2015) Genetic Algorithm Based Feature Selection combined with dual classification for the Automated Detection of proliferative Diabetic Retinopathy", Computerized Medical Imaging and Graphics, vol. 43, pp.64-77

[7] Arauzo-Azofra A., Benitez J. M., and Castro J. L. (2008) Consistency measures for Feature Selection. Journal of Intelligent Information Systems, 30(3):273–292.

[8] Kanan, H. R. and Faez, K. 2008, An Improved Feature Selection method based on Ant Colony Optimization (ACO) evaluated on face Recognition System", Applied Mathematics and Computation, vol. 205, no.2, pp.716- 725

[9] Moustafa N. and Slay J., 2015, “Unsw-nb15: A comprehensive Dataset for Network Intrusion Detection,” in MilCIS-IEEE Stream, Military Communications and Information Systems Conference. Canberra, Australia, IEEE publication, 2015.

[10] Moustafa N. and Slay J., (2015b) “The significant features of the UNSW-NB15 and the KDD99 sets for Network Intrusion Detection Systems”, the 4th International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS 2015), collocated with RAID 2015, 2016, [Online] available:

[11] Lee, W., Stolfo, S.J. and Mok, K.W. (2000) Adaptive Intrusion Detection: A Data Mining Approach. Artificial Intelligence Review, 14, 533-567. http://dx.doi.org/10.1023/A:1006624031083

.