ISSN 2047-3338

Performance Evaluation of Bayesian Classifier on Filter-Based Feature Selection Techniques

OLASEHINDE Olayemi O.[1], ALESE B. K.[2], ADETUNMBI A. O.[3]

[1,3]Federal Polytechnic, Ile Oluji, Ondo State, Nigeria
[2]Federal University of Technology, Akure, Ondo State, Nigeria

***Abstract*– Feature selection (FS) is a Machine Learning technique and a preprocessing stage in building intrusion detection system which can be independent of the choice of the learning algorithm or not, it plays important role in eliminating irrelevant and redundant feature in intrusion detection system (IDS); thereby increases the classification accuracy and reduces computational overhead cost of the IDS. it is an efficient way to reduce the dimensionality of an intrusion detection problem. This research examined the features of UNSW-NB15 dataset; a recently published intrusion detection dataset and applied three (3) filtered based feature selection techniques; information gain based, consistency based and correlation based on it to obtained a reduce dataset of attributes to build an intrusion detection system models that reduce the overhead computational cost and increases classification performance accuracy models. The result of the performance evaluations of the IDS model built on the reduced and whole datasets with Naive bayes machine learning algorithm shows that the reduced dataset accuracy and overhead processing cost outperformed the original whole dataset, model built with the consistency based reduced features has highest classification accuracy improvement of 14.16% over the classification accuracy of the whole test dataset, followed by information gain and correlation reduced test dataset with classification accuracy improvement of 13.55% and 10.7% respectively**

***Index Terms*– Attack Categories, Dimensionality, Computational Overhead, Filtered Based Features Selection**

I. INTRODUCTION

NETWORK packets consist several features also known as attributes, some of these attributes are redundant or irrelevant, in the sense that their values do not affect, determine or influence the target class label, the presence of redundant attributes are major reasons for high false alarm rate (FAR), high computational overhead cost and low detection accuracy rate. Feature selection is a method of reducing the number of attributes of dataset to be analyzed; it is an efficient way to reduce the dimensionality of a problem. it is one of the preprocessing stages in building intrusion detection system

The removal of the redundant and irrelevant attributes results in a reduce dataset on which the machine learning algorithm will be applied to learn from it to build an intrusion detection model. FS improves the computational speed of IDS models. [1] This research examined the attributes of UNSW-NB15 dataset and applied three (3) filtered based feature selection methods on it to obtained three reduced datasets from each method used, to build three (3) different intrusion detection system that is efficient and effective computationally

II. LITERATURE REVIEW

Feature selection research has been widely applied in several computing fields such as machine learning, statistics, pattern recognition to mention a few [2], the objectives of FS is to generate a relevant subset of the dataset that will improve classification accuracy of the target class. Feature selection methods can be broadly divided into filter and wrapper approaches. in filter approach, the feature selection is independent of the machine learning algorithm used for building the intrusion system, while in wrapper approach feature selection is tie with the machine learning algorithm. The filter-based approaches are independent of the supervised learning algorithm therefore offer more generality and they are computationally cheaper than the wrapper and embedded approaches. For processing the high-dimensional data, the filter methods are suitable rather than the wrapper and embedded methods [3].

Feature selection has proven in both theory and practice to be effective in enhancing learning efficiency, increasing predictive accuracy and reducing complexity of learned results [4], [5]. Traditional feature selection process consists of four basic steps, namely, subset generation, subset evaluation, stopping criterion, and validation [6]. Subset generation is a search process that produces candidate feature subsets for evaluation based on a certain search strategy. Each candidate subset is evaluated and compared with the previous best one according to a certain evaluation. If the new subset turns to be better, it replaces best one. This process is repeated until a given stopping condition is satisfied.

[7] studied the use of filter feature selection methods in the general problem of image classification. The choice of

features typically depends on the target application.

[8] analyzed the performance of eight different filter-based feature selection methods and three classification methods on two datasets of microarray gene expression data. The best individually performing feature selection methods varied depending on the dataset and the classifier used.

III. UNSW-NB15 DATASET

The UNSW_NB 15 (University of New South Wales –NB 2015) was created using IXIA Perfect Storm tool in the Cyber Range Laboratory of the Australian Centre for Cyber Security (ACCS) to generate a hybrid of the realistic modern normal activities and the synthetic contemporary attack behaviors from network traffics. The training and testing sets are made up of 82,332 and 175,341 records respectively as shown in Table I. Attack types were classified into nine groups, namely Analysis, Dos, Exploits, Fuzzers, Generic, Reconnaissance, Shellcode, Worms and Backdrop, The Training and Testing dataset contains 44 features attributes (4- Categorical, 28 Integer, 10 Float and 3 binary).

The UNSW-NB15 dataset is the latest published dataset, which was created in 2015 for research purposes in intrusion detection. The advantages UNSW-NB15 dataset over the NSLKDD data set, include, similarity between training and testing dataset and its suitability to evaluate existing and new attacks in an effective and reliable manner [9]. Fig. 1 shows the attacks and normal connections distribution of both the training and testing dataset, while Fig. 2 shows the attack categories and normal connections in the training and testing and dataset.

Table I: Percentage distribution of attacks and normal record in the training and testing dataset of UNSW-NB15 Dataset

	Training		Testing	
Names of Attack	No of Connection	Percentage Distribution	No of Connection	Percentage Distribution
Reconnaissance	3496	4.25	10491	5.98
Dos	4089	4.9	12264	6.99
Exploit	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	672	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

IV. FILTER BASED FEATURES SELECTION METHODS

Feature selection is a method of identifying most relevant features from a set of given features. The importance of feature selection is taken into account mainly for improving detection rate and detection accuracy in addition to reducing computation time and data size [10], it was necessary to determine the set of attributes of the UNSW-NB15 dataset the are deemed more predictive for the network packets classification. Three Filter-based feature selection (FS) methods namely; information gain based, consistency based and correlation-based feature selections were employed to identify the relevant features attributes among the features of UNSW-NB15 dataset.

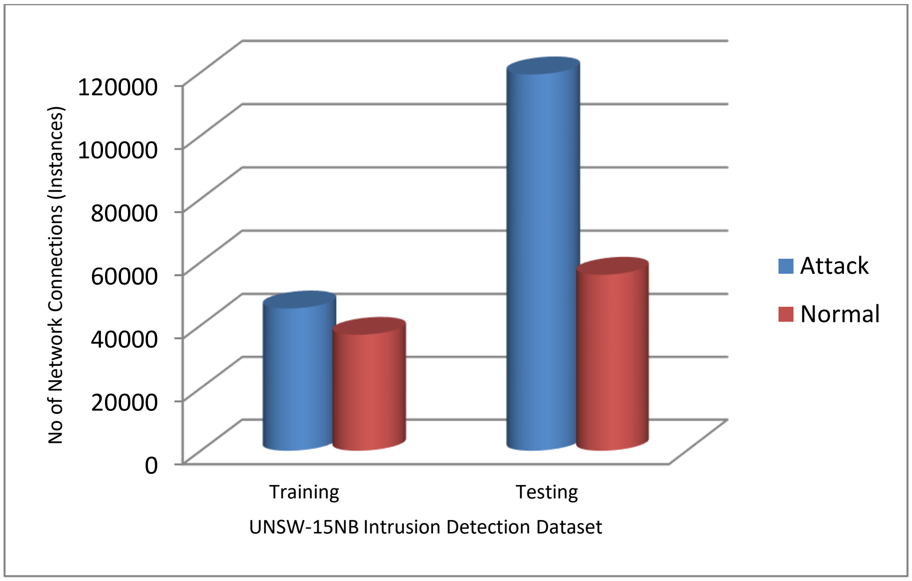


Fig. 1: Distribution of attack and normal connections in both training and testing UNSW-NB15 dataset

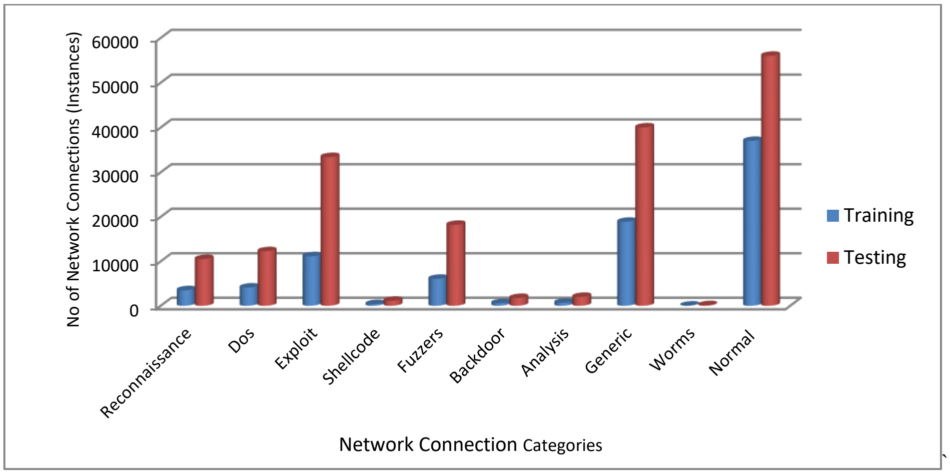


Fig. 2: Distribution of attack categories and normal connections in both training and testing UNSW-NB15 dataset

The basic algorithm for a filter-based feature selection algorithm is shown in the following Fig. 3 for the given UNSW-NB15 dataset X and feature set F. Usually, the algorithm may start with one of the following subsets of X' such as $X' = \{\varphi\} or X' \subset X$. The independent measure *I_m*(using the consistency, correlation or the information-based criterion) evaluates each generated subset I_g and compares it to the previous optimal subset. The search iterates until the stopping criteria θ is not met. Finally, the algorithm outputs the current optimal feature subset X_{opt}. The algorithm is presented as follows [14]:

```
INPUT:
D = {X, F}        //A training data set with n Intrusion Dataset
                  //$X = \{X_1, X_2, ., X_n\}$ – Attributes of intrusion
Data
                  // and F labels – Attack Class Label
        of Records
X'                //Predefined initial feature
        //subset/single attribute/

/($X' = \{\varphi\} or X' \subset X$)
$\theta$                        //Stopping criterion
OUTPUT: $X'_{opt}$        //An optimal subset of Initial Attributes
Begin:
Initialize:
        $X_{opt} = X'$;          //applying a search algorithm of
choice
        $\theta_{opt} = E(X', I_m)$;   //evaluate X' by using an
                              //independent measure $I_m$
do begin
        $X_g = generate(X)$;     //select next subset/attribute
                              //for evaluation
        $\theta = E(X_g, I_m)$;       //$X_g$current subset/attribute evaluation
by $I_m$
        If($\theta > \theta_{opt}$)
                $\theta_{opt} = \theta$;
                $X'_{opt} = X_g$;
repeat(until $\theta$ is not reached);
end
returnX'_{opt};     //optimal subset of
                    //attribute/ranked list of attributes
end;
```

Fig. 3: Filter based feature selection algorithm

A) Information Gain Method

Information Gain Based Feature Selection Method:

1. Compute the Information Gain (IG) for each feature of the UNSW-NB15 dataset
2. Rank each of the feature based on their IG value in descending order
3. Validate set of the ranked attributes in terms of classification accuracy on the dataset
4. Set of attributes with the highest accuracy is returned and selected

Information gain methods are used to determine the nominal valued feature Y (target class, C) estimating the individual probabilities of the values $y \in Y$ (network intrusion attack type, c) from the training data containing the initial attribute set. If this model is used to estimate the value of *target class* (attack type) for a sample drawn from the training data, then the *entropy* of the model (and hence of the attribute) is the number of bits it would take, on average, to correct the output of the model. Entropy is a measure of the *uncertainty* or unpredictability in a system. The entropy of the target class Y is given by equation (2).

$$H(Y) = -\sum_{y \in Y} p(y) \log_2\big(p(y)\big) \qquad (2)$$

If the observed values of target class in the training data are partitioned according to the values of input features X, and the entropy of Y with respect to the partitions induced by X is less than the entropy of the target class prior to partitioning, then there is a relationship between the target class Y and the indicator variables X. Equation (3) gives the entropy of Y after observing X.

$$H(Y|X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log_2\big(p(y|x)\big) \quad (3)$$

Information gain is the amount by which the entropy of Y decreases in relation to the target class Y and the indicator variables X. Thus, information is given by:

$$IG = H(Y) - H(Y|X) = H(X) - H(X|Y) = H(Y) + H(X) - H(X|Y). \qquad (4)$$

B) Consistency-based Method

The consistency-based feature selection method:

1. Generate all possible features subset of the dataset,
2. Compute the inconsistency count INC_i of all pattern p_i of the subset S
3. then compute the inconsistency rate INCR of subset S
4. Repeat steps 2 and 3 for all other subset of the dataset
5. Select subset with the lowest inconsistency rate INCR

Two instances are considered inconsistent if they have the same feature subset values but different attack categories value, inconsistency count INC_i of all pattern p_i of the subset S is given by:

$$INC_I = N - N_I \qquad (5)$$

where N : is the total number pattern p_i instance in subset s
N_i : is the number of instances of pattern p_i of subset s with the highest no of attacks

The inconsistency rate INCR of subset S of the UNSW-NB15 dataset given by equation 6:

$$INCR = \frac{\sum_i^h INC_i}{M} \qquad (6)$$

where
h : is the number of all possible patterns from subset s of the UNSW-NB15 dataset
M: is number attributes (features) contained in the subset *S* of the UNSW-NB15 dataset

Feature Subset with lowest of inconsistency rate will be selected as the feature selection.

C) Correlation-based (CFS) Method

Correlation based feature selection method:

1. Generates all possible attributes (features) subset S of the UNSW-NB15 dataset
2. then calculate $Merit_s$ for each of the subset S, using the merits function (equation 8).
3. Rank each of the feature based on their calculated $Merit_s$ in descending order
4. Validate set of the ranked attributes in terms of classification accuracy on the dataset
5. Set of attributes with the highest accuracy is returned and selected

Correlation based features selection measures closeness or dependency among features of a dataset. [11] stated that features are relevant if their values vary systematically with the attack category (class label) thus, a feature is useful if it is correlated with or predictive of the class label; otherwise it is irrelevant. Thus, a feature subset of the UNSW-NB15 dataset V_i is said to be relevant (predictive of the attack categories) if and only if there exist some v_i (values of feature subset – nominal or numeric) and *c* (target class – attack categories) for which $p(V_i = v_i) > 0$ such that [12]:

$$p(C = c|V_i = v_i) \neq p(C = c) \qquad (7)$$

The implication of this is that the relevant feature subset is one that contains highly correlated with (predictive of) the attack categories, yet uncorrelated with each other. It is but important to state that a feature subset of our intrusion dataset that are highly correlated with the target variable will at the same time bear low correlations with each other [13]. Equation (8) is the heuristic measure for the merit of feature subset *S* containing *k* features in supervised classification:

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \qquad (8)$$

where; $\overline{r_{cf}}$ = average feature-class correlation
$\overline{r_{ff}}$ = average feature-feature correlation

The values of cf and ff were computed from equation 9:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}} \qquad (9)$$

where *c and f* represents x *and y* in the equation 9.

Feature subset with the highest $Merit_s$value is selected and returned as the determinant of the attack categories (class label).

D) Bayesian Classifier

Bayesian classifier is the most straightforward and widely tested method for probabilistic induction [15]. In a Bayesian classifier, a probabilistic model of the features is built and the learning algorithm, uses that model to predict the classification of a new example [16]. Naive Bayes is the Bayesian classifier used in this work. The Naive Bayes algorithm is a classification algorithm based on Bayes rule, that assumes the attributes X1 ...Xn are all conditionally independent of one another, given Y. The value of this assumption is that it dramatically simplifies the representation of P(X|Y), and the problem of estimating it from the training data, it treats all features independently, with no feature depends on others features values [17]. Naïve Bayes algorithm is a significant classifier; it is easy to construct, does not requires parameter estimation, it is easy to interpret. Therefore, expert and inexpert data mining developers can perform Naïve Bayes. It generally performs well in comparison with other data mining methods (Arzucan, 2004).

Naïve Bayes' classifier in this research work is expressed as follows:

Let k_{ij} be the UNSW-NB15 dataset containing records of i number of attributes, for j number of instance in the dataset such that, k_i is the set of attributes. $k = k_1,....,k_j$ are the predictors in the dataset. C is the class label for each predictors, C comprises of ten classes;: (0 for normal network with no attack, 1...9 for various network attacks) is given as follows:

$$p(c_i \mid k_1, \ldots, k_j) = \frac{p(C_i)p(k|C_i)}{P(k_1, \ldots, k_n)} \qquad (10)$$

$$where\ i = 0\ or\ 1$$

Maximum posterior probability for classifying the class of a network instance is given as:
Naive bayes predicts attack category with the highest probability

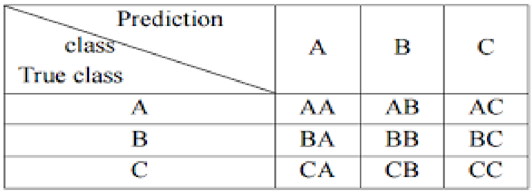
V. MODELS PERFORMANCE MEASURE

The performance of intrusion detection models was carried out by evaluating the measures from the values in the coincidence matrix also known as the confusion matrix Fig. 4, Confusion Matrix is an N X N matrix, where N is any integer greater than 1, The diagonal elements represent the number of points for which the predicted label is equal to the true label, while off-diagonal elements are those that are mislabeled by the classifier. The higher the diagonal values of the confusion matrix the better, indicating many correct predictions. **it** produced four outcomes, which are; true positive, true negative, false positive and false negative.

a) True positive (TP): correct positive classification
b) False positive (FP): incorrect positive classification
c) True negative (TN): correct negative classification
d) False negative (FN): incorrect negative classification

From Fig. 4, the overall performance of the classification model, and its possible four outcomes are defined as follows:

a) TP = TN: Sum of all correctly classified instances, it is sum of all instances along the diagonal from left to right.
b) FP: Sum of instances which are incorrectly classified as belonging to the Class. it is the sum of all entries in the matrix apart from the positively actual classified entries (diagonal entries)
c) FN: All instances that were not classified as belonging to the positive class but should have been not. it is the sum of all entries in the matrix apart from the positively actual classified entries (diagonal entries),
d) FN = FP



Prediction class True class	A	B	C
A	AA	AB	AC
B	BA	BB	BC
C	CA	CB	CC

Fig. 4: Confusion Matrix

Classification 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 11. An accuracy of 1 implies error rate of 0 and an accuracy of 0 indicate error rate of 1.

$$ACC = \frac{TP + TN}{FN + FP + TN + TP} \qquad (11)$$

A) Results and Discussion

The reduced datasets from the three (3) filter based feature selection methods were used to build an intrusion detection system and evaluated with naive bayes machine learning algorithm in terms of classification accuracy, the time taken to build each models and the time taken by the built models to classify a given set of new instance, Table II shows the number of attributes selected by each of the FS methods and the performance of each of the models built from the reduced set of each of the FS methods.

Table II: Number and List of Attributes Selected by Features Selection Methods

Consistency (27 Attributes)	Information Gain (22 Attributes)	Correlation (23 Attributes)
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 ,	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,	ct_dst_sport_ltm , sttl, swin, state, ct_src_dport_ltm, ct_srv_dst, ct_srv_src, dwin, ct_dst_src_ltm, service, ct_dst_ltm, ct_src_ltm, rate, dtcpb, stcpb, ct_state_ttl, proto, dttl, dload, dmean, tcprtt, ackdat, synack

Table II, shows the classification accuracy of Naive Bayes Classification model on each of the reduced test dataset and the whole test dataset, Consistency selected feature subset has the highest classification accuracy of 70.20%, closely followed by the ranked attributes selected by the information Gain attributes selector, features subset selected by the correlation methods has the least accuracy of 66.74% among the reduced test dataset. the accuracy of 56.04% obtained by the whole test dataset without feature selection justify the need for feature selection. Fig. 5 shows the graphical representation of the performance of the Naive Bayes classification model of each of the test dataset evaluated.

Table III: Classification Accuracy of Naive Bayes Model on Different Test Dataset Evaluated

		Correct Classification		Incorrect Classification	
		No. of Instance	%	No. of Instance	%
Reduced Feature Set	Information Gain (22 Attributes)	122,020	69.59	53,321	30.41
	Consistency Based (27 Attributes)	123,081	70.20	52,260	29.80
	Correlation Based (23 Attributes)	117,026	66.74	58,315	33.26
Whole Dataset (43 attributes)		98,262	56.04	77,079	43.96

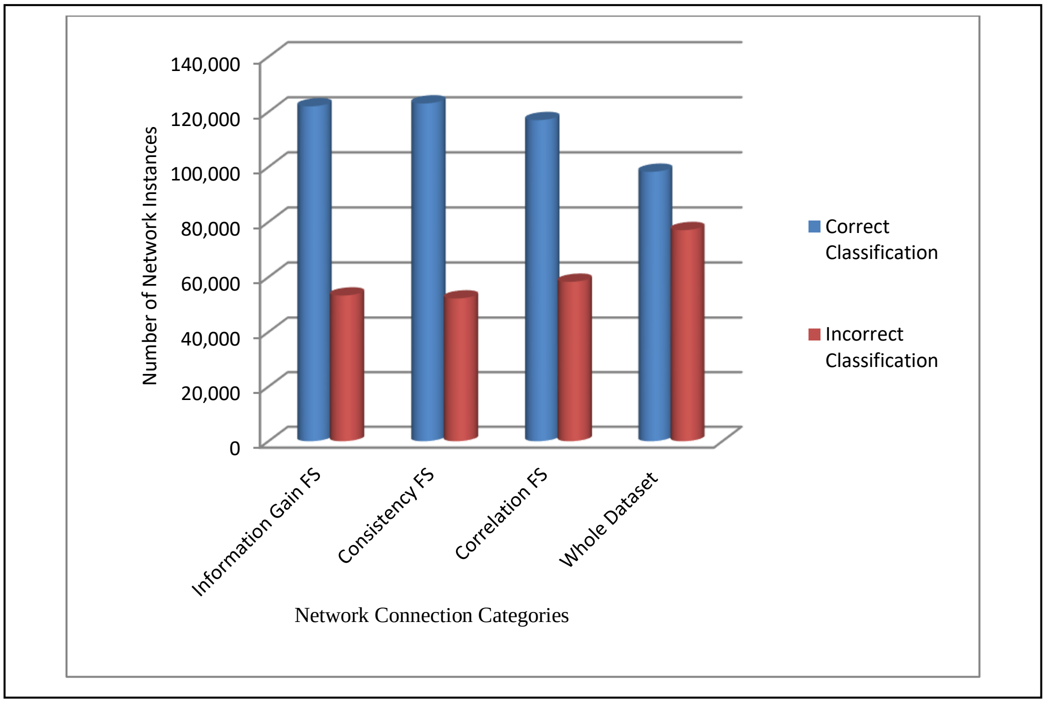


Fig. 5: Performance of the Naive Bayes classification model on the Evaluation of each of the test dataset

VI. CONCLUSION

Feature selection (FS) is a very important step in intrusion detection model building, the goal of feature selection is to improve computational efficiency and classification accuracy, In this work, we employed three (3) filter based features methods; Consistency Subset Selection, Correlation Subset Selection and Information Gain Ranked attribute selection to build a Naive Bayes Classification model for each of the three (3) reduced training dataset from the FS methods and whole training dataset, the models were evaluated on the test dataset, the result of the evaluation shows classification improvement with the reduced dataset containing relevant feature to the target class.

REFERENCES

[1]. Yu Lei and Liu Huan. (2004) Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research, 5:1205–1224.

[2]. Mitra P., Murthy C. A. and Pal S. K.. (2002) “Unsupervised feature selection using feature similarity,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 3, pp. 301–312.

[3]. Gnana D. Antony, Appavu S. and and LeavlineJebamalar (2016) Literature Review on Feature Selection Methods for High-Dimensional Data, International Journal of Computer Applications, Vol. 136, No. 1.

[4]. Almuallim H. and Dietterich T. G. (1994). “Learning Boolean concepts in the presence of many irrelevant features,” Artificial Intelligence, Vol. 69, No. 1-2, pp. 279–305.

[5]. Koller D. and Sahami M., (1996) “Toward optimal feature selection,” In Proceedings of the Thirteenth International Conference on Machine Learning, pp. 284–292, 1996.

[6]. Dash M.and Liu H.,(1997) “Feature Selection for Classification,” An International Journal of Intelligent Data Analysis, vol. 1, no. 3, pp.131-156,

[7]. Bins J. and Draper B. A., (2001), "Feature selection from huge feature sets," in: Proc. 8th International Conference on Computer Vision (ICCV-01), Vancouver, British Columbia, Canada, IEEE Computer Society, pp. 159–165,

[8]. Abusamra H. (2013) "A comparative study of feature selection and classification methods for gene expression data of glioma," Procedia Computer Science, vol. 23, pp. 5–14,

[9]. Moustafa N. and Slay J. (2015) A hybrid feature selection for network intrusion detection systems: Central points, Australian Information Warfare and Security Conference, 2015

[10]. Jothi, L. U. (2013)., An Anomaly Intrusion Detection using Feature Relevance and Negative Selection Algorithm, International Journal of Technological Exploration and Learning (IJTEL), Vol. 2, Issue 5, pp. 223-229.

[11]. Gennari, J. H., Langley, P., & Fisher, D. (1989). Models of incremental concept formation. Artificial Intelligence, 40, 11–61.

[12]. Kohavi R, John M. (1996) Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. Portland, Oregon: AAAI Press; 1996. 1996. Error-based and entropy-based discretization of continuous features; pp. 114–119.

[13]. Hall Mark A. and Smith Lloyd A. (1999) Feature Selection for Machine Learning: Comparing a Correlation-based Filter Approach to the Wrapper. Proceedings of the Twelfth International FLAIRS Conference. Copyright © 1999, AAAI (www.aaai.org). All rights reserved.

[14]. Kumar V. and Minz S. (2014) Feature Selection: A literature Review. Journal of Smart Computing Review, vol. 4, no. 3, June 2014.

[15]. Langley, P., & Sage, S. (1994). Induction of selective Bayesian classifiers. In Proceedings of the Tenth international conference on Uncertainty in artificial intelligence (pp. 399- 406). San Francisco, CA: Morgan Kaufmann http://dx.doi.org/10.1016/b978-1-55860-332-5.50055-9

[16]. Poole D., & Mackworth, A. (2010). Bayesian Classifiers. Retrieved November 4, 2015, from http://artint.info/html/ArtInt_181.html#id1

[17]. Mitchell, T., (1997) Machine Learning, McGraw-Hill.