

PAPER • OPEN ACCESS

## Effect of Feature Selection on Performance of Internet Traffic Classification on NIMS Multi-Class dataset

To cite this article: Jonathan Oluranti *et al* 2019 *J. Phys.: Conf. Ser.* **1299** 012035

View the [article online](#) for updates and enhancements.

You may also like

- [Network intrusion detection method based on deep learning](#)  
Shuai Zou, Fangwei Zhong, Bing Han et al.
- [Network Traffic Classification Based on Deep Learning](#)  
Jun Hua Shu, Jiang Jiang and Jing Xuan Sun
- [A Semi-Stack Approach for Accurate Network Traffic Classification Using Multi-View Stacking](#)  
Adil Fahada




**ECS** The Electrochemical Society  
Advancing solid state & electrochemical science & technology



### 242nd ECS Meeting

Oct 9 – 13, 2022 • Atlanta, GA, US

Presenting more than 2,400 technical abstracts in 50 symposia

  
**ECS Plenary Lecture**  
featuring  
**M. Stanley Whittingham**,  
Binghamton University  
Nobel Laureate –  
2019 Nobel Prize in Chemistry

 Register now!



# Effect of Feature Selection on Performance of Internet Traffic Classification on NIMS Multi-Class dataset

Jonathan Oluranti<sup>1</sup>, Nicholas Omoregbe<sup>1</sup>, Sanjay Misra<sup>2</sup>

Email: [jonathan.oluranti@covenantuniversity.edu.ng](mailto:jonathan.oluranti@covenantuniversity.edu.ng)

**Abstract.** The challenges faced by networks nowadays can be solved to a great extent by the application of accurate network traffic classification. Internet network traffic classification is responsible for associating network traffic with the application generating them and helps in the area of network monitoring, Quality of Service management, among other. Traditional methods of traffic classification including port-based, payload-load based, host-based, behavior-based exhibit a number of limitations that range from high computational cost to inability to access encrypted packets for the purpose of classification. Machine learning techniques based on statistical properties are now being employed to overcome the limitations of existing techniques. However, the high number of features of flows that serve as input to the learning machine poses a great challenge that requires the application of a pre-processing stage known as feature selection. Too many irrelevant and redundant features affect predictive accuracy and performance of the learning machine. This work analyses experimentally, the effect of a collection of ranking-based filter feature selection methods on a multi-class dataset for traffic classification. In the first stage, the proposed Top-N criterion is applied to the feature sets obtained, while in the second stage we generate for each Top-N set of features a new dataset which is applied as input to a set of four machine learning algorithms (classifiers). Experimental results show the viability of our model as a tool for selecting the optimal subset of features which when applied, lead to improvement of accuracy and performance of the traffic classification process.

**Keywords:**-Traffic Classification, Network Management, Feature Selection, Multi-class dataset

## 1. Introduction

The challenges faced by networks can, to a great extent, be solved by the use of accurate traffic classification [1, 2, 3]. In terms of network management, traffic classification can assist service providers to manage, control and understand the bandwidth requirements and behaviors of applications such as Voice over IP (VoIP) and video conferencing [4, 5]. On the other hand, traffic classification can assist in security by blocking attackers and unwanted traffic [6]. However, the continuous development of modern technologies have resulted in large number of computer and internet applications leading to the generation of large amounts of data at an unprecedented rate [2]. Such data include video, text, voice, data from social media, Internet of things as well as cloud computing to mention a few [1, 2]. Also, the massive data exhibit the characteristic of high dimensions which makes it difficult to conduct data analysis and decision-making. Feature selection has proved to be effective in pre-processing high-dimensional data [1, 5]. Using feature selection, it is possible to eliminate those features in a dataset, which are not relevant and redundant based on certain criterion [7]. The benefits of doing this include: reduction in computation time, improvement of accuracy of the learning machine and better understanding of resulting model is achieved [7, 8]. Feature selection is only one of the two ways to reduce the dimensionality of a dataset. The second method is feature extraction. Feature extraction transforms the original data to features that can be used to recognize patterns easily and quickly [9]. However, the emerging features from feature extraction appear new and difficult to correlate with original set of features [1]. Therefore most researchers prefer to use feature selection, as it maintains the structure of the original set of features. Some areas of application of feature selection include image recognition, image retrieval, text mining, intrusion detection [, bioinformatics data analysis, fault diagnosis, to mention a few. There are several divisions of feature selection. They include; the use of training data which can be labeled, unlabeled or semi-labeled giving to rise to the supervised,



unsupervised, and semi-supervised models [1,2,9]; division based on the relationship with the learning methods giving rise to filter, wrapper, and embedded models; division based on various evaluation criteria like correlation, Euclidean distance, consistency, dependence, and information measure; division based on the search strategies giving rise to methods like forward increase, backward deletion, random, and hybrid models and finally division based on the type of output giving rise methods like feature ranking (involving weights) and subset selection models. In the case of filters, the relationship between the feature and target class label is the focus while for wrapper model, there is need to verify the selected subset of features with a classifier. Since there could be a large number of subsets to verify, the wrapper method incurs a large amount of computational cost. Also for the filter, the evaluation criterion is critical. The embedded model on the other hand, selects feature while carrying out the training process of learning model such that the feature selection output is available once the training process is completed [1]. The performance of the feature selection method is usually evaluated by the machine learning model [7, 8, 9]. The commonly used machine learning models includes Naïve Bayes, K-Nearest Neighbor, C4.5, Support Vector Machine, Back Propagation- Neural Network, Radial Basis Function-Neural Network, K-means, Hierarchical clustering[3], to mention a few. A good feature selection method should have high learning accuracy but less computational overhead (time complexity and space complexity [7]. In this work, we analyse experimentally, the effect of a collection of ranking-based filter feature selection methods on a multi-class dataset for traffic classification which is not so common as most analysis have been done datasets with only a pair of classes. In the first stage, the proposed Top-N criterion is applied to the feature sets obtained, while in the second stage we generate for each Top-N set of features a new dataset which is applied as input to a set of four machine learning algorithms (classifiers).

## 2. Related Works

In this section, we highlight some related proposals with respect to feature selection upon which this work is based. In the research domain of Internet traffic classification [10], feature selection has been attracting special attention for over a decade now as a result of the explosion in the scale of Internet traffic feature sets [8]. The benefits of feature selection include among others, improvement in computational performance while ensuring that the accuracy of the classifier is not affected negatively when conducting traffic flow identification. [10] proposed Fast Correlation-Based Filter, which was combined with a novel wrapper-based method to determine threshold. The model was used to select useful features. [11] proposed a new feature selection method called BFS. The same dataset used in [10] was reused and BFS was found to be more competitive in maintaining the balance of multi-class classification results in comparison with FCBF based on two metrics namely g-mean and classification accuracy. In [12] an application-based feature subset selection was proposed using parameter estimation for each logistic regression model established for the corresponding application class. The focus of [13] was to effectively resolve the imbalance problem caused by elephant flows. Real-time Internet traffic identification was the focus of the work in [14] with emphasis on special requirements of simplicity and effectiveness on feature subsets. Correlation based, Consistency based, and PCA feature selection were performed in [15] on a real-time Internet traffic dataset obtained by a packet capture tool. In [16], a mutual-information-based feature selection and automatic determiner of the number of relevant features was proposed. In [17] a number of new evaluation metrics namely goodness, stability and similarity were used to assess the advantages and defects of existing feature selection methods. Six useful feature selection methods were integrated with the aim of combining their strengths. The six feature selection methods include Information Gain, Gain Ratio, PCA, Correlation-Based Feature Selection, Chi-square, and Consistency-Based Feature Selection. In [18] a wrapper method was proposed which used Bees Algorithm as search strategy and Support Vector Machine as the classifier. It was found that Bees algorithm yielded better results than other feature selection methods such as Rough-DPSO, Rough Set, Linear Genetic Programming, MARS and Support Vector Decision Function Ranking. In [19], a hybrid feature selection method called LGP BA was proposed. It combined Linear Genetic Programming and Bee Algorithm that achieved better accuracy and efficiency. In this work, a set of six ranking-based filter feature selection methods is applied to the NIMS multi-class dataset. A scheme for arriving at the optimal feature subset based on the Top-N criterion is proposed and the resulting feature subset for each

of the set thresholds is used to evaluate the accuracy of a set of four classifiers. The next section provides details of the new scheme.

### 3. Methodology

#### 3.1 Dataset Used

NIMS data set is one of the classical datasets used for the study. It includes packets collected from a tested network and made available by the original authors for download. The dataset consists of 22 features and one class Among its classes are SSH servers outside connection and application behaviors traffic such as DNS, HTTP, SFTP and P2P traffic. The detailed characteristics of the datasets as well as the list and description of the features are presented in Tables 1 and 2.

**Table 1.** Description of the Instances of NIMS dataset

s/n	Name/Class	Number of Instances
1	DNS	38,016
2	FTP	1,728
3	HTTP	11,904
4	LFD	2,557
5	RFD	2,422
6	SCP	2,444
7	SFTP	2,412
8	SHELL	2,491
9	TELNET	1,251
10	X11	2,355

**Table 2.** Description of Features of NIMS Dataset

SN	Attribute	Description	Type
1	Minfpktl	minimum flow packet length	Numeric
2	meanfpktl	mean flow packet length	Numeric
3	maxfpktl	maximum flow packet length	Numeric
4	Stdfpctl	standard deviation flow packet length	Numeric
5	minbpctl	minimum byte packet length	Numeric
6	meanbkctl	mean byte packet length	Numeric
7	maxbkctl	maximum byte packet length	Numeric
8	Stdbkctl	standard deviation byte packet length	Numeric
9	Minfiat	minimum flow interarrival time	Numeric
10	Meanfiat	mean flow interarrival time	Numeric
11	Maxfiat	maximum flow interarrival time	Numeric
12	Stdfiat	standard deviation flow interarrival time	Numeric
13	Minbiat	minimum byte interarrival time	Numeric
14	meanbiat	mean byte interarrival time	Numeric
15	Maxbiat	maximum byte interarrival time	Numeric
16	Stdbiat	standard deviation byte interarrival time	Numeric
17	durartion	Duration	Numeric
18	Proto	Protocol	Numeric

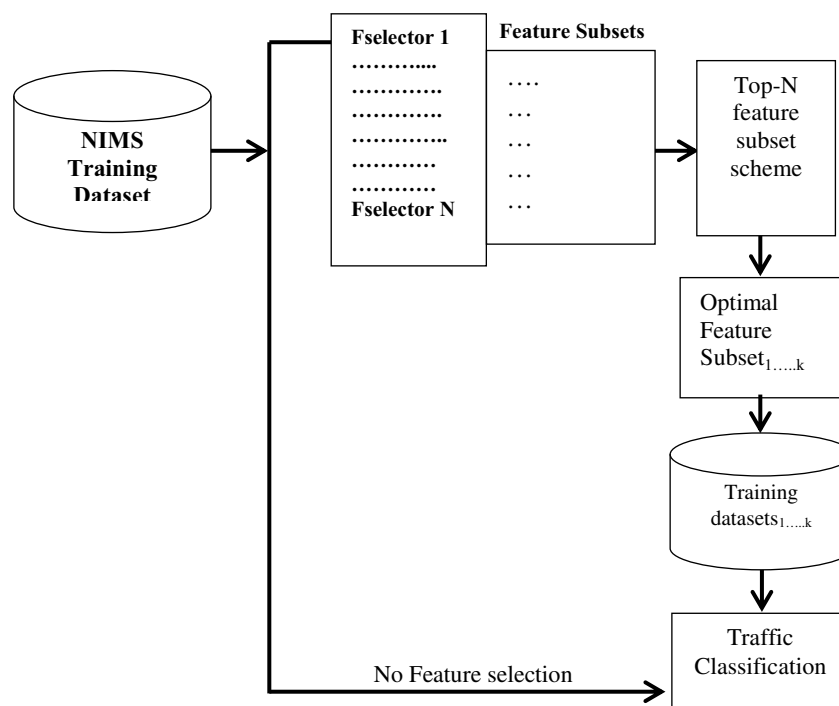
19	Totalfpkt	total flow packet	Numeric
20	Totalfvol	total flow volume	Numeric
21	totalbpkt	total byte packet	Numeric
22	totalbvol	total byte volume	Numeric
23	Class		Character

### 3.2 Proposed Method: Top-N Criteria-Based Feature Selection

There are 23 features in the NIMS dataset as shown in TABLE 2. Some features are more important and play a vital role in classification and some are less important. Removing some features may increase the classification accuracy as well as reduce the computational time. Our proposed scheme of feature selection is termed Top-N Criterion-Based Feature Selection (TNCFS). The scheme consists of the steps listed below:

- i) A set of six filter-based ranking feature selection algorithms is applied to the training dataset (NIMS)
- ii) For each feature selection result, the TOP-N features are retrieved.
- iii) A number K is set as threshold for features which are common to all from the results obtained in (i) above
- iv) Steps (ii) and (iii) are repeated for different values N to obtain the different sets of K features.
- v) The resulting subset of features obtained (ii) and (iii) above are each used to generate a new training dataset
- vi) A set of four different classifiers is applied first to training dataset containing the complete features and then to the various training datasets obtained from (v) above.
- vii) The set of K features which yields best results in terms of accuracy, precision and recall upon evaluation.

Figure 1 represents the various steps of the proposed TNCFS model.



**Figure1.** Model of the Proposed TNCFS Method

### 3.3 Experimental Setup

All the experiments were carried out using a Dell Laptop Core i7, 12GB RAM, 1TB HDD and 64-bit Windows 10 Operating System. WEKA, a popular machine learning workbench is used for feature selection and classification. The features selection methods used in the study include: CfsSubsetEvaluation, Correlation AttributeEvaluation, Gain Ratio AttributeEvaluation, Info Gain AttributeEvaluation, OneR AttributeEvaluation, ReliefF AttributeEvaluation and Symmetrical Uncertainty AttributeEvaluation. We proposed a simple feature selection method after using seven ranking-based feature selections to generate features. Three different thresholds N of the number of features to use for our experiment such that if N is 15, then the specific features that appear in all the rankings returned by the feature selection methods used was determined. Therefore experiment was conducted using set N for 15, 10 and 5. The result of the rankings returned by each feature selection method and the result obtained following the threshold that was set, are presented in Table 3 and Table 4. When N is equal to 15 features, we obtained a total of 10 features. When N is 10 we obtained a total of 6 features and 2 features for N = 6.

**Table 3.** Features returned by various Ranking Methods

SN	Feature Selection Technique	Top 15 Features Returned
1	Classifier Attribute selector	22,8,9,6,7,5,21,2,3,4,10,11,12,19,20
2	Correlation Attribute selector	18,7,1,8,6,2,15,11,3,20,17,21,19,4,9
3	Gain Ratio	18,7,5,3,9,7,4,2,6,13,15,20,11,8,21
4	Information Gain	15,11,9,3,13,2,4,6,7,10,14,20,8,16,12
5	OneR	11,15,9,2,3,13,6,4,10,14,7,12,16,8,20
6	Relief	18,7,1,8,6,5,2,3,12,11,15,10,4,16,9
7	Symmetric Uncertainty	1,18,3,5,9,7,4,2,6,15,13,11,20,8,10

**Table 4.** Features obtained after setting Threshold Points N

N	Number of Features selected	Top 15 Features Returned
15	10	9,8,6,7,2,3,4,10,11,20
10	6	9,6,7,2,3,4,
6	2	9,7

Based on Table 4 above, we generated three new datasets and re-run our experiments to check if there is any improvement and the feature set that produces the best result will be termed the optimal for the dataset used. Note that this is the first time we introduce feature selection for improving learning and classification in this work after ensuring that we have reduced the number of features that can be used. The three new datasets are called NDT1 (10 features), NDT2 (6 features) and NDT3 (2 features).

### 3.4 Evaluation Metrics

True Positive (TP) is a state when instance is correctly classified as intrusion by an IDS. True Negative (TN) is a state when normal traffic is correctly classified as normal. False Positive (FP) is a state when normal traffic is misclassified as an intrusion. False Negative (FN) is a state when an intrusion is misclassified as normal. Detection Rate (DR), False Positive Rate (FPR), Precision (P), Recall (R), F-Measure (F-M), Area Under ROC (AUC) were considered to analyse the performance of the classifiers.

A high precision value simply means that the algorithm returned to a large extent, more relevant results than irrelevant. On the other hand, high recall is an indication that an algorithm returned most of the relevant results. F-Measure is a combination of Precision and Recall. It is the weighted harmonic mean of precision and recall. Another performance evaluation metric is Receiver Operator Characteristic (ROC) curve which is a graphical plot that illustrates the performance of the classifier as the threshold is varied. It is drawn by plotting False Positive Rate (FPR) against True Positive Rate (TPR)). The area under the ROC curve (ROC Area or AUC) is equal to the probability that a classifier will rank a randomly chosen instance higher than a randomly chosen normal instance.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (2)$$

$$Specificity = \frac{TN}{FP+TN} \quad FPR \quad (3)$$

$$AUC = \frac{1 + TPR - FPR}{2} \quad (4)$$

Since Specificity = FPR and Sensitivity = TPR

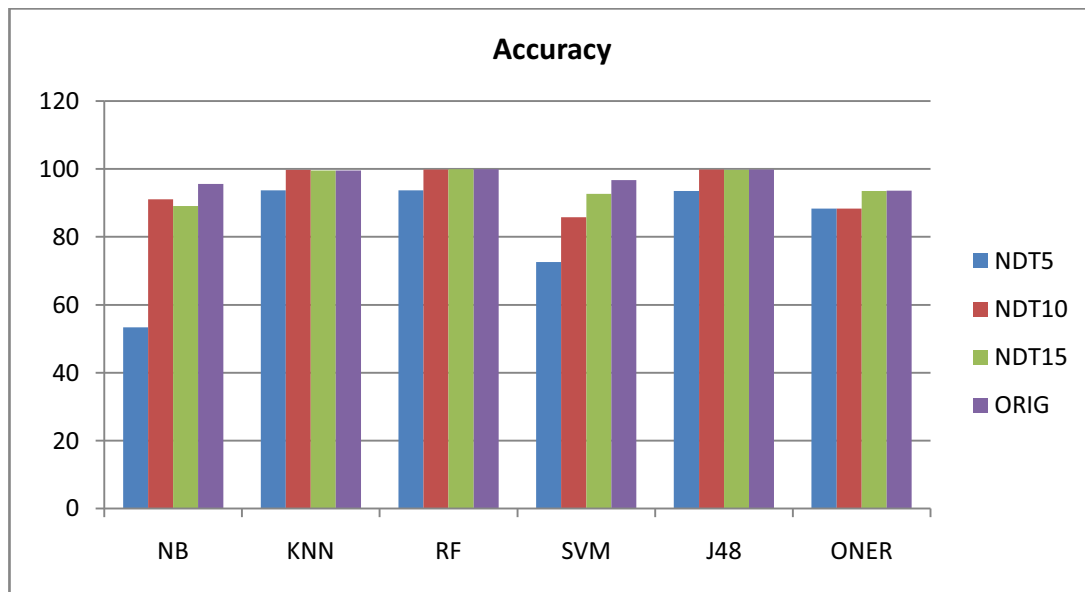
#### 4. Results and Discussion

Table 5 below presents a detail comparison of the performance of the classifiers between the dataset without feature selection (ORIGDT) but 5-fold cross validation, against the datasets with feature selection and 5-fold cross validation. This therefore amounts to experiments on 4 different datasets three of which were derived based on the sets of feature subsets derived from our model. The presentation is supported with appropriate charts and explanation afterwards.

**Table 5.** Comparison of the Performance of five classifiers on four datasets with different feature subsets

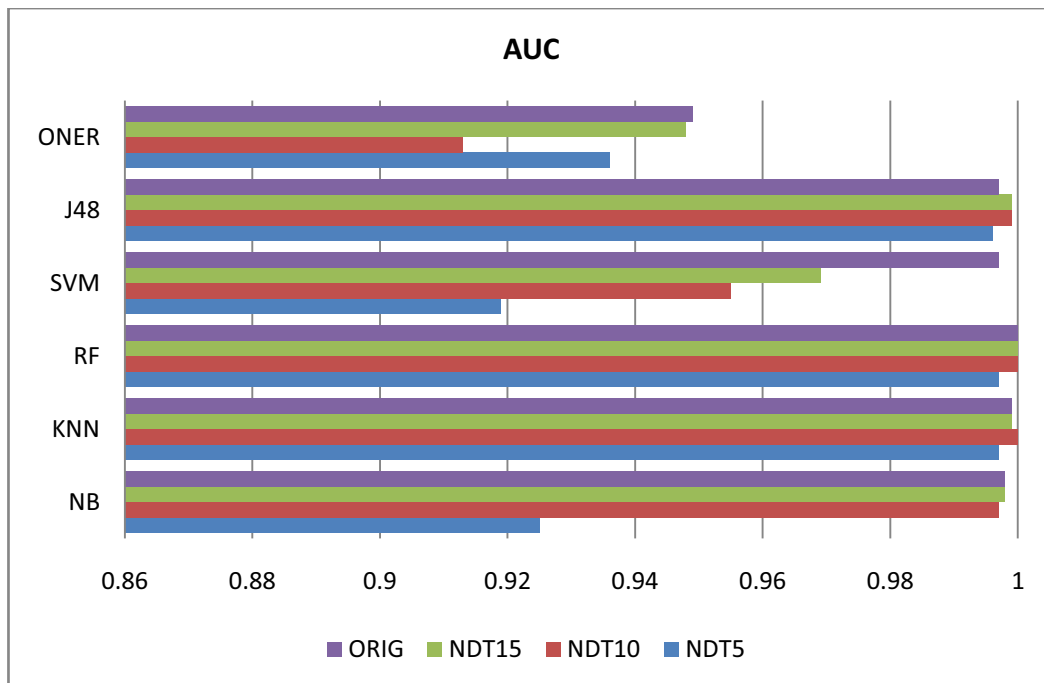
Classification Algorithm	Datasets	Accuracy	Precision	Recall	F-Measure	AUC
NB	NDT5	53.34	??	0.533	??	0.925
	NDT10	91.09	??	0.911	??	0.997
	NDT15	89.07	0.958	0.891	0.896	0.998
	<b>ORIGDT</b>	<b>95.61</b>	<b>0.964</b>	<b>0.956</b>	<b>0.953</b>	<b>0.998</b>
KNN	NDT5	93.67	0.949	0.937	0.936	0.997
	<b>NDT10</b>	<b>99.72</b>	<b>0.997</b>	<b>0.997</b>	<b>0.997</b>	<b>1.000</b>
	NDT15	99.59	0.996	0.996	0.996	0.999
	ORIGDT	99.51	0.995	0.995	0.995	0.999
RF	NDT5	93.66	0.949	0.937	0.936	0.997
	NDT10	99.83	0.998	0.998	0.998	1.000
	<b>NDT15</b>	<b>99.89</b>	<b>0.999</b>	<b>0.999</b>	<b>0.999</b>	<b>1.000</b>
SVM	ORIGDT	99.88	0.999	0.999	0.999	1.000
	NDT5	72.6	??	0.725	??	0.919
	NDT10	85.80	??	0.858	??	0.955
	NDT15	92.7	??	0.927	??	0.969

J48	<b>ORIGDT</b>	<b>96.73</b>	<b>0.970</b>	<b>0.967</b>	<b>0.967</b>	<b>0.997</b>
	NDT5	93.48	0.948	0.935	0.934	0.996
	<b>NDT10</b>	<b>99.81</b>	<b>0.998</b>	<b>0.998</b>	<b>0.998</b>	<b>0.999</b>
	NDT15	99.78	0.998	0.998	0.998	0.999
ONER	ORIGDT	99.79	0.998	0.998	0.998	0.997
	NDT5	88.32	0.887	0.883	0.876	0.936
	NDT10	88.32	0.874	0.883	0.871	0.913
	NDT15	93.51	0.934	0.935	0.933	0.948
	<b>ORIGDT</b>	<b>93.60</b>	<b>0.935</b>	<b>0.936</b>	<b>0.935</b>	<b>0.949</b>



**Figure 2.** Comparison of the Accuracy of six classifiers on four datasets with different feature subsets (Threshold – 5, 10 and 15)





**Figure 3.** Comparison of Area under ROC (AUC) of six classifiers on four datasets with different feature subsets (Threshold – 5, 10 and 15)

Table 5 contains the details we require to draw some insight with respect to the impact of feature selection on classification of network traffic of NIMS dataset. In terms of precision, it is clear from the results obtained and showed on table 5 that the precision value is high for all classifiers when only 10 features are considered (NDT15). Also, the recall result obtained for dataset with 10 features has the highest value for all classifiers except for Naïve Bayes. When compared with dataset NDT5, the set with all Features (ORIGDT), NDT15 is a little lower for the 5 classifiers. In terms of F-measure, NDT5 exhibits lower F-Measure values for the five classifiers while NDT15 which has 10 features presents a higher F-Measure for J48, RF, and Naïve Bayes classifiers. With respect to ROC, dataset NDT15 has the highest ROC Area for the 6 classifiers Naïve Bayes, KNN, RF, SVM, J48 and OneR. Among the datasets with reduced Features, that is, among the datasets NDT15, NDT10 and NDT5, NDT15 with 10 Features has the highest Precision, Recall, F-Measure and ROC Area for all the classifiers. Considering the overall performance and the time taken to build the models, it can be concluded that NDT15 dataset with 10 features exhibits the best performance. The dataset NDT15 also presents a good balance between Accuracy, Precision and Recall and so the 10 features selected by the proposed method can be considered as the superior and most important features, notwithstanding the classifier used.

### Conclusion

In this work, a simple hybrid feature scheme called Top-N-based Feature Selection (TNCFS) scheme was proposed. The viability of the model was tested on the NIMS multi-class dataset and the result obtained indicate that with feature selection, the number of features required to classify the NIMS dataset could be reduced from the initial 23 features to 10 features (**9,8,6,7,2,3,4,10,11,20**) and still achieve the same or even better results in terms of accuracy and performance. Thus if the learning machine is trained on the NIMS dataset with 10 features, the result will be better than when the 23 features are all applied. Based on the results obtained we conclude that feature selection has a positive effect on the NIMS multi-class dataset.

## References

- [1] JieCai, JiaweiLuo, Shulin Wang, Sheng Yang (2018) "Feature selection in machine learning: A new perspective", *Neurocomputing*
- [2] Boutaba, R., Salahuddin, M. A., Limam, N., Ayoubi, S., Shahriar, N., Estrada-Solano, F., &Caicedo, O. M. (2018). A comprehensive survey on machine learning for networking: evolution, applications and research opportunities. *Journal of Internet Services and Applications*, 9(1).doi:10.1186/s13174-018-0087-2
- [3] Adda M, Qader K, Al-Kasassbeh M. Comparative analysis of clustering techniques in network traffic faults classification. *Int J Innov Res ComputCommun Eng*. 2017;5(4):6551–63.
- [4] Villmann, T., Kaden, M., Hermann, W., &Biehl, M. (2016). Learning vector quantization classifiers for ROC-optimization. *Computational Statistics*, 33(3), 1173–1194. doi:10.1007/s00180-016-0678-y
- [5] Kulin, Merima, Carolina Fortuna, Eli De Poorter, Dirk Deschrijver, and Ingrid Moerman."Data-DrivenDesign of Intelligent Wireless Networks: An Overview and Tutorial", *Sensors*,2016.
- [6] NourMoustafa, Jiankun Hu, Jill Slay. "A holistic review of Network Anomaly Detection Systems: A comprehensive survey", *Journal of Network and Computer Applications*, 2019
- [7] Ferri, C.."An experimental comparison of performance measures for classification",*Pattern Recognition Letters*, 2009.
- [8] Muhammad Shafiq, Xiangzhan Yu, Dawei Wang. "Robust Feature Selection for IMApplications at Early Stage Traffic Classification Using Machine Learning Algorithms", 2017 IEEE 19th International Conference on High Performance Computing and Communications; IEEE 15th International Conference on Smart City; IEEE 3<sup>rd</sup>International Conference on Data Science and Systems (HPCC/SmartCity/DSS), 2017.
- [9] Muhammad Shafiq, Xiangzhan Yu, Ali Kashif Bashir, Hassan NazeerChaudhry, DaweiWang. "A machine learning approach for feature selection traffic classification usingsecurity analysis", *The Journal of Supercomputing*, 2018
- [10] Yu, L., Liu, H., "Feature selection for high-dimensional data: a fast correlation-basedfilter solution", *Conf. Machine Learning (ICML 03)*, 2003.
- [11] Zhen, L., Qiong, L., "A New Feature Selection Method for Internet Traffic Classification Using ML". *Physics Procedia Internet Conference on Medical Physics andBiomedical Engineering (ICMPBE 2012)*, 2012.
- [12] En-Najjary, T., Urvoy-Keller, G., Pietrzyk, M., "Application-based Feature Selectionfor Internet Traffic Classification", *22nd International Teletraffic Congress*, 2010.
- [13] Chen, Z., Peng, L., Zhao, S., Zhang, L., Jing, S., "Feature Selection Toward Optimizing Internet Traffic Behavior Identification, Algorithms and Architectures for ParallelProcessing", *Lecture Notes in Computer Science*, vol. 8631, pp. 631–644, 2014.
- [14] Ding, C., Zhou, C., He, X., Zha, H., "R1-PCA: Rotational Invariant L1-norm Principal - Component Analysis for Robust Subspace Factorization", *23rd International Conference on Machine Learning (ICM L2006)*, 2006.
- [15] Kuldeep, S., Agrawal, S., "Performance Evaluation of Five Machine Learning Algorithms and Three Feature Selection Algorithms for IP Traffic Classification", *IJCASpecial Issue on Evolution in Networks and Computer Communications*, vol. 1, pp.25–32, 2011.

- [16] Amiri, F., Yousefi, M., Lucas, C., Shakery, A., Yazdani, N., “Mutual information-based feature selection for intrusion detection systems”, *Journal of Network and Computer Applications*, vol. 34, issue 4, pp. 1184–1199, 2011.
- [17] Fahad, A., Tari, Z., Khalil, I., Habib, I., Alnuweiri, H., “Toward an efficient and scalable feature selection approach for Internet traffic classification”, *Computer Networks*, vol. 57, issue 9, pp. 2040–2057, 2013.
- [18] Alomari, O., Othman, Z., “Bees Algorithm for feature selection in Network Anomaly detection”, *Journal of Applied Sciences Research*, vol. 8, issue 3, pp. 1748–1756, 2012.
- [19] Hassani, S.R., Z.A. Othman and S.M.M. Kahaki, “Hybrid feature selection algorithm for intrusion detection system”, *Journal of Computer Science*, vol. 10, issue 6, pp. 1015–1025, 2014.