# Comparative Analysis of Classification Algorithms for Email Spam Detection

**Article** *in* International Journal of Computer Network and Information Security · January 2018

DOI: 10.5815/ijcnis.2018.01.07

Maryam Shuaib Bobi
Federal University of Technology Minna
**3** PUBLICATIONS **12** CITATIONS
SEE PROFILE

Oluwafemi Osho
Federal University of Technology Minna
**39** PUBLICATIONS **164** CITATIONS
SEE PROFILE

Ismaila Idris
Federal University of Technology Minna
**42** PUBLICATIONS **269** CITATIONS
SEE PROFILE

John Alhassan
Federal University of Technology Minna
**59** PUBLICATIONS **96** CITATIONS
SEE PROFILE

*I. J. Computer Network and Information Security,* **2018, 1, 60-67**
Published Online January 2018 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijcnis.2018.01.07

# Comparative Analysis of Classification Algorithms for Email Spam Detection

**Shafi'i Muhammad Abdulhamid, Maryam Shuaib, Oluwafemi Osho**
Department of Cyber Security, Federal University of Technology, Minna, Nigeria.
E-mail: shafii.abdulhamid@futminna.edu.ng, maryambobi@gmail.com, femi.osho@futminna.edu.ng

**Idris Ismaila and John K. Alhassan**
Department of Cyber Security, Federal University of Technology, Minna, Nigeria
E-mail: ismi.idris@futminna.edu.ng and jkalhassan@futminna.edu.ng

Received: 24 June 2017; Accepted: 10 August 2017; Published: 08 January 2018

***Abstract*—**The increase in the use of email in every day transactions for a lot of businesses or general communication due to its cost effectiveness and efficiency has made emails vulnerable to attacks including spamming. Spam emails also called junk emails are unsolicited messages that are almost identical and sent to multiple recipients randomly. In this study, a performance analysis is done on some classification algorithms including: Bayesian Logistic Regression, Hidden Naïve Bayes, Radial Basis Function (RBF) Network, Voted Perceptron, Lazy Bayesian Rule, Logit Boost, Rotation Forest, NNge, Logistic Model Tree, REP Tree, Naïve Bayes, Multilayer Perceptron, Random Tree and J48. The performance of the algorithms were measured in terms of Accuracy, Precision, Recall, F-Measure, Root Mean Squared Error, Receiver Operator Characteristics Area and Root Relative Squared Error using WEKA data mining tool. To have a balanced view on the classification algorithms' performance, no feature selection or performance boosting method was employed. The research showed that a number of classification algorithms exist that if properly explored through feature selection means will yield more accurate results for email classification. Rotation Forest is found to be the classifier that gives the best accuracy of 94.2%. Though none of the algorithms did not achieve 100% accuracy in sorting spam emails, Rotation Forest has shown a near degree to achieving most accurate result.

***Index Terms*—**Email spam, classification algorithms, Bayesian Logistic Regression, Hidden Naïve Bayes, Rotation Forest.

## I. INTRODUCTION

Email is a means of information transfer from any part of the world that is extremely fast and cost effective and can be used from personal computers, smartphones, and other last-generation electronic gadgets. [1], [2].

Despite the increase in usage of other forms of online communication such as instant messaging and social networking, emails have continued to take the lead in business communications and still serves as a requirement for other forms of communications and e-transactions. Emails are used by almost all humans. It is estimated that by the end of 2016, there will be over 2.6 billion email account holders worldwide and it is estimated that nearly half of the world population will be using emails by the end of 2020 [3].

The increase in the popularity and use of emails for transactions has led to a rise in the amount of spam emails globally. Spam emails also called junk emails are unsolicited messages that is non-requested and are almost identical sent to multiple recipients via emails. The sender of spam mails has no prior relationship with the receivers but gathers the addresses from different sources such as phone books and filled forms. Spam messages are fast growing to be one of the most serious threats to users of E-mail messages because it is a major means of sending threats, including viruses, worms and phishing attacks [4], [5],[6], [7].

Recently, data mining has drawn attention in the knowledge and information industry because of the immense accessibility of big data and the forthcoming need for converting such data into useful information and knowledge. According to [8], Data mining as an emergent field that requires extracting implicit, previously not known, and potentially helpful information from data is being explored and used as a means of building software that automatically sieves through databases in search of regularities or patterns. Strong patterns identified, are likely to be used to generalize and give accurate predictions.

According to [9], classification or prediction tasks which are supervised methods that seek to discover the hidden associations between the target class and the independent variables are popularly used in data mining. For supervised learning, classifiers allow tags to be attributed to the observations, so that data not observed can be categorized based on the training data. Spam detection systems are built with the use of classification algorithms to group the emails as spam or non-spam[10],[11] The aim of the paper is to evaluate the performance of classification algorithms that are used for grouping emails as spam or not spam including Bayesian Logistic Regression, Hidden Naïve Bayes, RBF Network, Voted Perceptron, Lazy Bayesian Rule, Logit Boost, Rotation Forest, NNge, Logistic Model Tree, REP Tree, Naïve Bayes, J48, Multilayer Perceptron and Random Tree.

The remainder of the paper are organized as follows: section II presents related literatures in Comparative analysis of classification algorithms in the field of email spam detection and filtering. Section III shows the materials and methods employed in the research. Section IV chronicles the results obtained in the analysis of the classification algorithms and section V describes the conclusion and future recommendations.

## II. Related Works

The rise in the number of email users has made the task of handling large volumes of email challenging for data mining and machine learning due to the rise in spam emails during the previous years. This has led a number of researchers to carryout comparative studies on the performance of classification algorithms in correctly classifying emails using a combination of performance metrics. It is therefore, necessary to determine which algorithm performs best for any chosen metric to assist in proper classification of emails as spam or non-spam is vital.

Many works have been carried out to compare the performances of some classification algorithms in grouping emails. Classification algorithms whose performances have been so far compared include Naïve Bayes[1], [12]–[17], other algorithms compared include C-PLS, ANN, C-RT, CS-CRT, CS-MC4, CS-SVC, Continouns PLS-DA, PLS-LDA, LDA[1], Bayesnet[4], [12], [13], Multilayer perceptron [1], [15], SVM [1], [4], [12]–[14], [16], [17]. Table 1 shows the summary of algorithms used in previous comparative research.

Particle Swarm Optimization and Artificial Neural Network were combined for feature selection and Support Vector Machine was used to classify and separate spam by[18]. Their method was compared with other methods such as data classification Self Organizing Map and K-Means based on criteria Area under Curve. The results indicate that the Area under Curve (AUC used as benchmark for performance evaluation) in the proposed method is better than other methods.

[19]in their paper titled Spam Mail Detection using Classification carried out an experiment on many data mining techniques to the dataset of spam in an attempt to search the most suitable classifier to email classification as spam and non-spam. they checked the performance of many classifiers with the use of feature selection algorithm and found out that in the result analysis part the Naïve Bayes classifier provides finer accuracy of 76% with respect to other two classifiers such as support vector machine and J48 and also that time taken for Naïve Bayes classifier is lesser than other two classifiers which means that Naïve Bayes classifier is the best classifier among the other two classifier which are used for classifying the spam mails.

A lot of conventional anti-spam techniques for evading spam such as Bayesian based sort, rule based system, IP blacklist, Heuristic based filter, White list and DNS black holes were identified by [20]. They used RBF, a neural network technique in which neurons were trained. The proposed mechanism improves the accuracy, precision, recall Frr and Far. The proposed mechanism is compared with SVM and the results were comparatively better.

[12] in their paper Spam Mail Detection through Data Mining – A Comparative Performance Analysis, analyzed various data mining approach to spam dataset in order to find out the best classifier for email classification. In this paper they analyzed the performance of various classifiers with feature selection algorithm and without feature selection algorithm. The Best-First feature selection algorithm was applied in order to select the desired features and then apply various classifiers for classification. They found that results are improved in terms of accuracy when feature selection process is embedded in the experiment and also found Random Tree to be the best classifier for spam mail classification with accuracy = 99.72%. Still none of the algorithm achieves 100% accuracy in classifying spam emails but Random Tree is very nearby to that.

[21] paper on Content-Based Spam Filtering and Detection Algorithms- an Efficient Analysis & Comparison focused on Spam as one of the major problems faced by the Internet community. The content of each item is represented as a set of descriptors or terms. The terms are typically, the words that occur in a document. User profiles are represented with the same terms and built up by analyzing the content of items seen by the user. Their research paper mainly contributes to the comprehensive study of spam detection algorithms under the category of content based filtering. Then, the implemented results were benchmarked to examine how accurately they have been classified into their original categories of spam. The efficient technique among the discussed techniques is chosen as Bayesian method to create a spam filter.

[1] paper on Comparative Study on Email Spam Classifier using Data used spam data set analyzing with the use of TANAGRA data mining tool to explore the efficient classifier for email spam classification. Initially, feature construction and feature selection is done to extract the relevant features. Then various classification algorithms are applied over this dataset and cross validation is done for each of these classifiers. Finally, The Rnd tree classifier for email spam is identified as the best based on the error rate, precision and recall.

Table 1. Summary of Relevant Algorithms Compared in Related Research Works

| | | Algorithms | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| S/No. | Reference | C-PLS | ANN | C-RT | CS-CRT | RVM | CS-MC4 | CS-SVC | Continous PLS-DA | PLS-LDA | LDA | ID3 | Neural Network | KNN | Bayesnet | compliment Naïve Bayes | DMNBText | Naïve Bayes | Logistic | Multilayer Perceptron | SMO/SVM | IBK | kstar | LWL | PART | ADTree | FT | J48 | RandomForest | RandomTree | SimpleCart |
| | [4] | | | | | | | | | | | | | | √ | | | | | | √ | √ | | | | | | √ | | | |
| | [22] | | | | | | | | | | | | | | | | | | | | | | | | √ | | | | √ | | |
| | [12] | | | | | | | | | | | | | | √ | | | √ | | | √ | | | | | | √ | √ | √ | √ | |
| | [1] | √ | | √ | √ | | √ | √ | √ | √ | √ | √ | | √ | | | | √ | √ | √ | √ | | | | | | | | | √ | |
| | [13] | | √ | √ | | | | | | | | | | | √ | | | √ | | | √ | √ | √ | √ | | | | √ | √ | | |
| | [14] | | | | | | | | | | | | √ | √ | | | | √ | | | √ | | | | | | | | | | |
| | [15] | | | | | | | | | | | | | | | | | √ | | √ | | | | | | | | √ | | | |
| | [23] | | | | | | | | | | | √ | | | | | | | | | | | | | | √ | | √ | | | √ |
| | [16] | | | | | √ | | | | | | | √ | | | | | √ | | | √ | | | | | | | | | | |
| | [17] | | | | | | | | | | | | √ | | | | | √ | | | √ | | | | | | | √ | | | |

Table 2. Summary of relevant Performance Metrics used for Comparison in Related Research Work

| | Performance Metric | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Reference | Correctly Classified Instances | Incorrectly Classified Instances | Kappa Statistic | Mean Absolute Error | Root Mean Squared Error | Relative Absolute Error | Root Relative Squared Error | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Runtime |
| [4] | √ | | | | | | | | √ | √ | √ | √ | | |
| [22] | | | | | | | | √ | | | | | | √ |
| [12] | | | √ | √ | √ | √ | √ | √ | | | | | | |
| [1] | | √ | | | | | | | | √ | √ | | | |
| [13] | √ | √ | | | | | | | √ | √ | √ | √ | | |
| [14] | | | | | | | | √ | | √ | √ | | | |
| [15] | √ | √ | | | | | | √ | | | | | | √ |
| [23] | | | | | | | | √ | | √ | √ | | | |
| [16] | | | | | | | | √ | | √ | √ | | | |
| [17] | | | | | | | | √ | | √ | √ | | | |

[14] Looks at Machine Learning Methods for Spam E-Mail Classification. The authors reviewed some of the most popular machine learning methods (Bayesian classification, k-NN, ANNs, SVMs, Artificial immune system and Rough sets) and of their applicability to the problem of spam Email classification. Descriptions of the algorithms were presented, and the comparison of their performance on the Spam Assassin spam corpus was presented.

The researchers employed the use of a combination of some performance metrics including Correctly Classified Instances, Kappa Statistics, Mean Absolute Error, Root Mean Squared Error, Relative Absolute Error, Root Relative Squared Error [12]. Other performance metrics used are TP Rate, FP Rate, Precision, Recall, F-Measure and ROC [4], [13]. A few researchers also considered the time taken to load models in determining the performance of the algorithms [15], [22]. Table 2 shows performance metrics employed by previous research works.

Spam classifiers are built and tested on publicly available datasets for evaluation. For example Naïve Bayes, Bayesnet, SMO/SVM, ID3, FT, J48, Random Forest, Random Tree, C-PLS, C-RT, CS-CRT, CS-MC4, CS-SVC, Continuous PLS-DA and PLS-LDA is used on the Spambase dataset from UCI Library [1], [12], [23]. In some research works, two or more datasets are used for comparative analysis [16], [22]. The datasets are made publicly available and normally contain proper ham or spam ratio.

There are still a number ofclassificaion algorithms that are yet to be compared in terms of their performance and accuracy in email spam classification including Spegasos, voted perceptron, IB1, MIWrapper, LWL, CitationKNN, AdaBoostM1, HyperPipes, Dagging, Deecorate, END, FilteredClassifier, Grading, LogitBoost, MetaCost,MultiBoostAB, DecisionTable, Multi Scheme, Ordinal Class Classifier, Raced Incremental, Logit Boost, RandomCommittee, RandomSubSpace, MIBoost, MISMO, IBK, kstarSimpleMI, Bagging,VFI, ConjuctiveRule, Multi Class Classifier,DTNB, Jrip, Nnge, OneR, PART, Ridor, ZeroR.

## III. Materials and Methods

In carrying out this research three steps were involved: Dataset Preparation, Pre-Processing and Application of various machine learning classifiers and evaluating the performance of machine learning classifiers.

### *A. Dataset Preparation, Pre-Processing and Algorithm Application*

The Spambase dataset gotten from the UCI Machine Learning Repository was used. The dataset has 57 attributes of different variable types in 4601 instances. The Spambase dataset is converted into .arff format (a format compatible for machine learning) supported by the WEKA tool for input data that was used for the analysis.

To adequately classify the Spambase dataset, Bayesian Logistic Regression, Hidden Naïve Bayes, RBFNetwork, Voted Perceptron, Lazy Bayesian Rule, Logit Boost, Rotation Forest, NNge, Logistic Model Tree, REPTree, Naïve Bayes, J48, Multilayer Perceptron and Random Tree were used and a 10 folds cross validation was used in this research. The choice of 10 folds was due to results obtained from broad tests on various datasets, with varying learning procedures, that have demonstrated that 10 is about the correct number of folds to get the best gauge of error [8]. For cross-validation, a specified number of folds is chosen, the data is partitioned arbitrarily into 10 parts in which the class is represented in approximately the same proportions as in the full dataset. Each partition is held out in turn and the learning scheme trained on the remaining nine-tenths; then its error rate is processed on the holdout set. Hence, the learning procedure is carried out a total of 10 times on various training sets (each of which have a lot in common). Finally, the averages of the 10 error estimates are taken to give an overall error estimate.

For Comparative reasons, the dataset was also run using percentage split which allows you to take out a certain percentage of the data for testing, 66% split was employed for this research work.

## IV. Results

The entire dataset was used for the experiment with 10 folds cross validation and 66% split. The comparison of performance in terms of Accuracy, Precision, Recall, F-Measure, Root Mean Squared Error, Receiver Operator Characteristics Area and Root Relative Squared Error is summarised here.

### *A. Accuracy*

The Accuracy is used to show the level of correct predictions. It does not consider positives or negatives independently and thus other measures for performance analysis aside from the accuracy are also used. The value 1 is the largest indicating highest accuracy, in this research work, the highest Accuracy is 0.942 gotten when the 10-folds cross validation was applied on Rotation Forest algorithm and the lowest was 0.891 gotten when 66% split was used with the REPTree algorithm. Fig 1 and Table 4 shows the Accuracy

### *B. Precision, Recall and F-Measure*

Precision is the fraction of relevant recollected instances, while recall is the fraction of relevant instances that are recollected. Precision and recall depend on an understanding and measure of relevance. When discussing, precision and recall scores, either values for one measure are likened for a specific level at the other measure or both are combined as a single measure. In this research the F-measure is used. A high F-measure is required since both precision and recall are desired to be high and Rotation forest has the highest F-measure of 0.942 the charts are presented in Table 4 and Fig 2 to Fig 4.

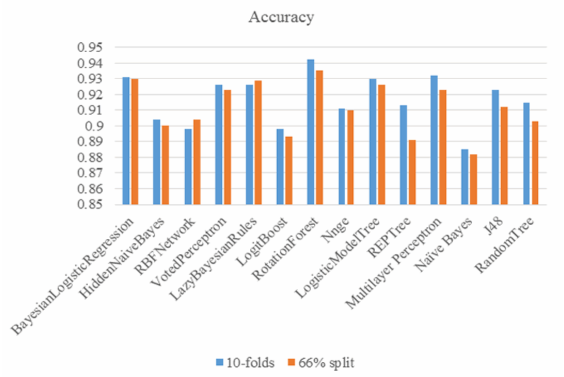

Fig.1. Comparison of Accuracy

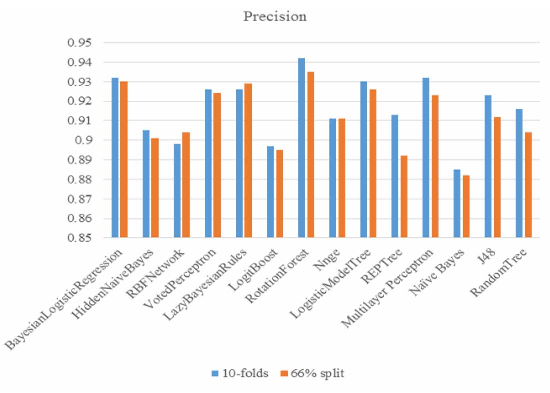

Fig.2. Comparison of Precision

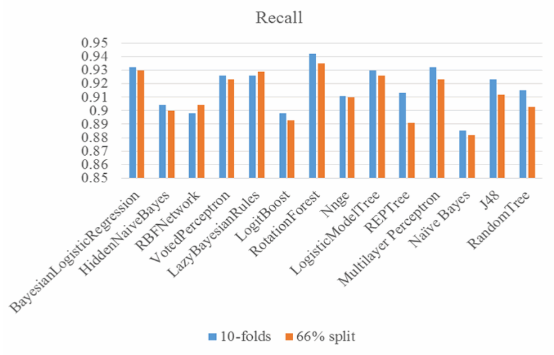

Fig.3. Comparison of Recall

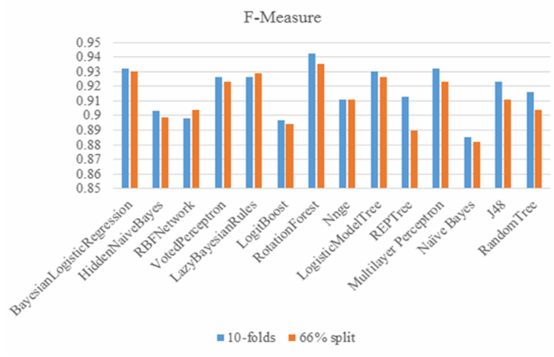

Fig.4. Comparison of the F-Measure

### C. *ROC Area*

The ROC (AUC) Area of a classifier/algorithm is equal to the probability of the classifier ranking a randomly selected positive instance higher than a randomly selected negative instance. Fig 5 shows the areas under ROC curves of classifiers used in this research with Rotation forest having the highest with 0.98 and Random Tree having the lowest with 0.905

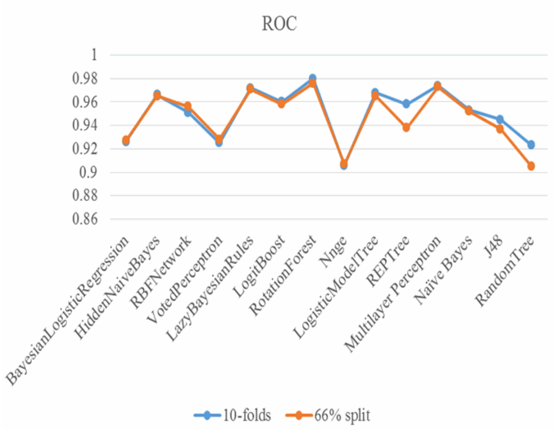

Fig.5. Comparison of ROC Area

### D. *Kappa Statistics*

The Kappa characteristic gives the level of agreements between the true classes and the classifications. The value 1 is the highest showing total agreement, in this study, the highest kappa characteristics is 0.879 which was gotten when the test was carried out on Rotation Forest with 10 folds cross validation. Table 4 and Fig 6 shows the respective kappa characteristics.

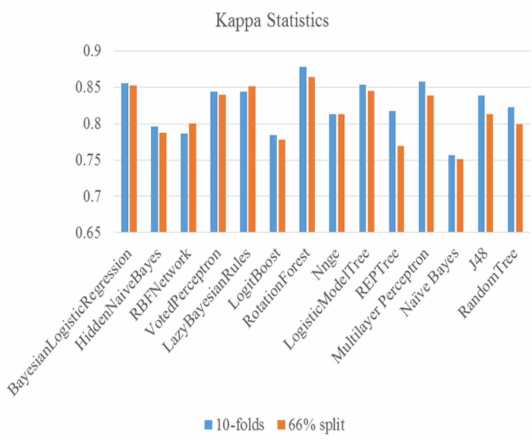

Fig.6. Kappa Statistics

### E. *Root Mean Squared Error*

According to root mean square error a low value is an indication of an excellent classifier. A low value for the root mean square error was recorded for Rotation Forest using 10-folds cross validation with 0.216. Fig 7 and Table 4 shows the Root Mean Squared Error.

### F. *Root Relative Squared Error*

The relative squared error normalizes the total squared error by dividing it by the total squared error of the simple predictor. The error is reduced to the same dimension as the quality being predicted by taking the square root of the relative squared error. Fig 8 and Table 4 gives the respective values of the Root Relative Squared Error.

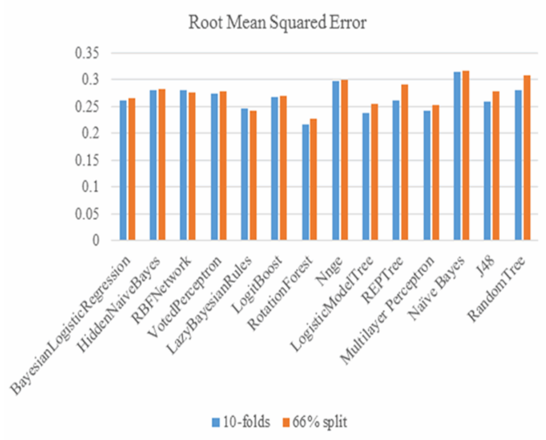

Fig.7. Root Mean Squared Error

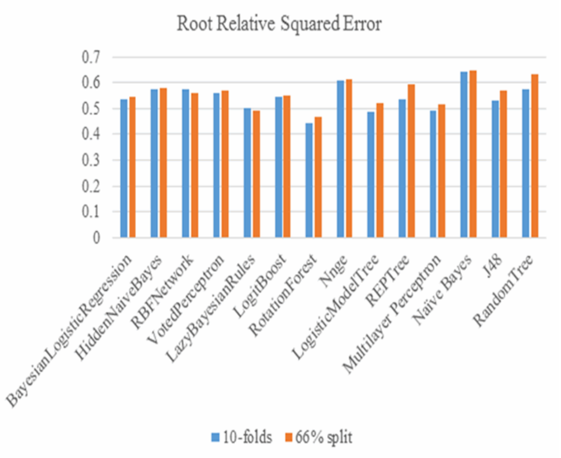

Fig.8. Root Relative Square Error

Table 3. Results of Accuracy, Precision, Recall, F-Measure, ROC Area, Kappa Statistic, RMSE and RRSE

| | | Accuracy | | Precision | | Recall | | F-Measure | | ROC Area | | Kappa Statistic | | RMSE | | RRSE | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| S/N | Algoithm | 10-folds | 66% Split | 10-folds | 66% Split | 10-folds | 66% Split | 10-folds | 66% Split | 10-folds | 66% Split | 10-folds | 66% Split | 10-folds | 66% Split | 10-folds | 66% Split |
| 1 | BayesianLogisticRegression | 0.931 | 0.93 | 0.932 | 0.93 | 0.932 | 0.93 | 0.932 | 0.93 | 0.926 | 0.927 | 0.856 | 0.853 | 0.261 | 0.265 | 0.535 | 0.543 |
| 2 | HiddenNaiveBayes | 0.904 | 0.900 | 0.905 | 0.901 | 0.904 | 0.9 | 0.903 | 0.899 | 0.966 | 0.965 | 0.796 | 0.787 | 0.281 | 0.282 | 0.574 | 0.577 |
| 3 | RBFNetwork | 0.898 | 0.904 | 0.898 | 0.904 | 0.898 | 0.904 | 0.898 | 0.904 | 0.951 | 0.956 | 0.786 | 0.8 | 0.281 | 0.275 | 0.576 | 0.562 |
| 4 | VotedPerceptron | 0.926 | 0.923 | 0.926 | 0.924 | 0.926 | 0.923 | 0.926 | 0.923 | 0.925 | 0.928 | 0.844 | 0.84 | 0.273 | 0.277 | 0.558 | 0.567 |
| 5 | LazyBayesianRules | 0.926 | 0.929 | 0.926 | 0.929 | 0.926 | 0.929 | 0.926 | 0.929 | 0.972 | 0.971 | 0.844 | 0.851 | 0.245 | 0.241 | 0.502 | 0.494 |
| 6 | LogitBoost | 0.898 | 0.893 | 0.897 | 0.895 | 0.898 | 0.893 | 0.897 | 0.894 | 0.96 | 0.958 | 0.784 | 0.778 | 0.267 | 0.27 | 0.546 | 0.552 |
| 7 | RotationForest | 0.942 | 0.935 | 0.942 | 0.935 | 0.942 | 0.935 | 0.942 | 0.935 | 0.98 | 0.976 | 0.878 | 0.864 | 0.216 | 0.228 | 0.442 | 0.467 |
| 8 | Nnge | 0.911 | 0.91 | 0.911 | 0.911 | 0.911 | 0.91 | 0.911 | 0.911 | 0.906 | 0.907 | 0.813 | 0.813 | 0.298 | 0.3 | 0.61 | 0.612 |
| 9 | LogisticModelTree | 0.93 | 0.926 | 0.93 | 0.926 | 0.93 | 0.926 | 0.93 | 0.926 | 0.968 | 0.965 | 0.854 | 0.845 | 0.237 | 0.255 | 0.486 | 0.522 |
| 10 | REPTree | 0.913 | 0.891 | 0.913 | 0.892 | 0.913 | 0.891 | 0.913 | 0.89 | 0.958 | 0.938 | 0.817 | 0.769 | 0.261 | 0.29 | 0.534 | 0.592 |
| 11 | Multilayer Perceptron | 0.932 | 0.923 | 0.932 | 0.923 | 0.932 | 0.923 | 0.932 | 0.923 | 0.974 | 0.973 | 0.858 | 0.839 | 0.241 | 0.252 | 0.494 | 0.516 |
| 12 | Naïve Bayes | 0.885 | 0.882 | 0.885 | 0.882 | 0.885 | 0.882 | 0.885 | 0.882 | 0.953 | 0.952 | 0.757 | 0.751 | 0.315 | 0.317 | 0.644 | 0.649 |
| 13 | J48 | 0.923 | 0.912 | 0.923 | 0.912 | 0.923 | 0.912 | 0.923 | 0.911 | 0.945 | 0.937 | 0.839 | 0.813 | 0.259 | 0.278 | 0.53 | 0.569 |
| 14 | RandomTree | 0.915 | 0.903 | 0.916 | 0.904 | 0.915 | 0.903 | 0.916 | 0.904 | 0.923 | 0.905 | 0.823 | 0.799 | 0.281 | 0.308 | 0.574 | 0.631 |

## V. Conclusion and Recommendations

This research work was driven by the increasing rate of spam emails across the globe and the knowledge from literature review of the availability of classification algorithms that have not been compared in terms of their performance on email datasets. From the experiment and results obtained from running fourteen different classification algorithms (including commonly used algorithms) using two test options it has been established that some uncommon algorithms perform relatively well on the Spambase dataset our training and testing dataset on WEKA, the testing environment with Rotation Forest emerging as the best classifier.

The results obtained shows that even with less feature selection employed, the Rotation Forest classification algorithm with 0.942 performs relatively well in email classification, even better than some commonly used classification algorithms including J48 which records 0.923 accuracy, Na ïve Bayes with 0.885 and Multilayer Perceptron with 0.932.

We recommend that the results obtained be compared with more spam datasets from different sources and using different Machine Learning tools. Also, more classification algorithms should be analysed with email spam datasets.

## Acknowledgements

The authors would like to acknowledge and appreciate the Department of Cyber Security, Federal University of Technology, Minna, Nigeria for their support.

## References

[1] R.. Kumar, G. Pookuzhali, and P. Sudhakar, "Comparative Study on Email Spam Classifier using Data Mining Techniques," 2012, vol. I.

[2] J. M. Carmona-cejudo, G. Castillo, M. Baena-garc ı́a, and R. Morales-bueno, "Knowledge-Based Systems A comparative study on feature selection and adaptive strategies for email foldering using the ABC-DynF framework," vol. 46, pp. 81–94, 2013.

[3] R. Group, "Email Statistics Report , 2016-2020," vol. 44, no. 0, pp. 0–3, 2016.

[4] A. Sharaff, N. . Nagwani, and A. Dhadse, "Comparative Study of Classification Algorithms for Spam Email Detection," *Springer*, no. January, 2016.

[5] R. M. Alguliev, R. M. Aliguliyev, and S. A. Nazirova, "Classification of Textual E-Mail Spam Using Data Mining Techniques," *Appl. Comput. Intell. Soft Comput.*, vol. 2011, pp. 1–8, 2011.

[6] A. F. Yasin, "Spam Reduction by using E-mail History and Authentication (SREHA)," *Int. J. Inf. Technol. Comput. Sci.*, vol. Vol.8, no. No.7, p. pp.17-22, 2016.

[7] M. Iqbal, M. A. Malik, A. Mushtaq, and K. Faisal, "Study on the Effectiveness of Spam Detection Technologies," *Int. J. Inf. Technol. Comput. Sci.*, vol. Vol.8, no. 1, pp. 11–21, 2016.

[8] I. H. Witten and F. Eibe, *Data mining : practical machine learning tools and techniques*, 2nd ed. Morgan Kaufmann Publishers, 2005.

[9] O. Maimon and L. Rokach, *The data mining and knowledge discovery handbook*, 2nd ed. Springer, 2010.

[10] S. M. Abdulhamid *et al.*, "A Review on Mobile SMS Spam Filtering Techniques," *IEEE Access*, 2017.

[11] Adebayo, O. S., D. O. Ugiomoh, and M. D. AbdulMalik, "The Design and Development of Real-Time E-Voting System in Nigeria with Emphasis on Security and Result Veracity.," *Int. J. Comput. Netw. Inf. Secur.*, vol. 5, no. 5, p. 9, 2013.

[12] M. Rathi and V. Pareek, "Spam Mail Detection through Data Mining – A Comparative Performance Analysis," *Int. J. Mod. Educ. Comput. Sci.*, vol. 5, no. December, pp. 31–39, 2013.

[13] P. Panigrahi, "A comparative study of supervised machine learning techniques for spam E-mail filtering," in *Proceedings - 4th International Conference on Computational Intelligence and Communication Networks, CICN 2012*, 2012, pp. 506–512.

[14] W. . Awad and S. . Elseuofi, "Machine Learning Methods for Spam E- mail Classification," vol. 3, no. 1, pp. 173–184, 2011.

[15] D. . Renuka, T. Hamsapriya, M. . Chakkaravarthi, and P. . Surya, "Spam Classification based on Supervised Learning using Machine Learning Techniques," in *Process Automation, Control and Computing (PACC)*, 2011, pp. 1–7.

[16] B. Yu and Z. Xu, "A comparative study for content-based dynamic spam classification using four machine learning algorithms," *Knowledge-Based Syst.*, vol. 21, no. 14, pp. 355–362, 2008.

[17] S. Youn and D. Mcleod, "A Comparative Study for Email Classification," *Adv. Innov. Syst. Comput. Sci. Softw. Eng.*, pp. 387–391, 2007.

[18] M. Zavvar, M. Rezaei, and S. Garavand, "Email Spam Detection Using Combination of Particle Swarm Optimization and Artificial Neural Network and Support Vector Machine," *Int. J. Mod. Educ. Comput. Sci.*, vol. 7, no. July, pp. 68–74, 2016.

[19] P. Parveen and P. G. Halse, "Spam Mail Detection using Classification," vol. 5, no. 6, pp. 347–349, 2016.

[20] R. Sharma and G. Kaur, "E-Mail Spam Detection Using SVM and RBF," no. April, pp. 57–63, 2016.

[21] R. Malarvizhi and K. Saraswathi, "Content-Based Spam Filtering and Detection Algorithms - An Efficient Analysis & Comparison," *Int. J. Eng. Trends Technol.*, vol. 4, no. 9, pp. 4237–4242, 2013.

[22] P. Ozarkar and M. Patwardhan, "Efficient Spam Classification by Appropriate Feature Selection," vol. 13, no. 5, 2013.

[23] A. K. Sharma and S. Sahni, "A Comparative Study of Classification Algorithms for Spam Email Data Analysis," *Int. J. Comput. Sci. Eng.*, vol. 3, no. 5, pp. 1890–1895, 2011.

## Authors' Profiles

**Shafi'i Muhammad Abdulhamid** received his PhD in Computer Science from Universiti Teknologi Malaysia (UTM), MSc in Computer Science from Bayero University Kano (BUK), Nigeria and a Bachelor of Technology in Mathematics/Computer Science from the Federal University of Technology Minna, Nigeria. His current research interests are in Cyber Security, Cloud computing, Soft Computing and BigData. He has published many academic papers in reputable International journals, conference proceedings and book chapters. He has been appointed as an Editorial board member for UPI JCSIT and IJTRD. He has also been appointed as a reviewer of several ISI and Scopus indexed International journals such as JNCA Elsevier, ASOC Elsevier, EIJ Elsevier, JKSU-CIS Elsevier, NCAA Springer, BJST Springer, IJNS, IJST, IJCT, JITE:Research, JITE:IIP, JAIT, IJAER and JCEIT SciTechnol. He is a member of IEEE, International Association of Computer Science and Information Technology (IACSIT), Computer Professionals Registration Council of Nigeria (CPN), International Association of Engineers (IAENG), The Internet Society (ISOC), Cyber Security Experts Association of Nigeria (CSEAN) and Nigerian Computer Society (NCS). Presently he is a lecturer at the Department of Cyber Security Science, Federal University of Technology Minna, Nigeria.

**Maryam Shuaib** is a Postgraduate student in the Department of Cyber Security Science, Federal University of Technology, Minna, Nigeria. She has a B.Tech. degree in Mathematics with Computer Science. She was the Special Assistant to the Governor of Niger State, Nigeria on ICT Development between 2012-2015. Her research interests include cybersecurity, IoT and Database Security. Maryam is an Oracle Database 11g

Certified Associate and a member of the Association of Nigerian Authors.

**Oluwafemi Osho** is currently a lecturer in the Department of Cyber Security Science, Federal University of Technology, Minna, Nigeria. He holds an M.Tech. degree in Mathematics, and a B.Tech. degree in Mathematics/Computer Science. Before joining the institution, he served as Head of the IT Department of one of the leading mortgage banks in Nigeria. His current research interests include cybersecurity, mobile security, and security analysis. Oluwafemi is a Certified Ethical Hacker (CEH), and a member of the Cyber Security Experts Association of Nigeria (CSEAN), and a host of other professional associations.

**Dr. Ismaila Idris** is with the Deparment of Cyber Security Science. He obtain his Bachelor degree with Federal University of Technology, Minna. M.Sc. with university of Ilorin and PhD degree with University of Teknologi Malaysia. His research interest are Information Security, Data Mining, Machine Learning, Evolutionary Algorithm.

**Dr. J. K. Alhassan** was born at Ganmu-Alhaeri, in Kwara State, Nigeria on 9th January, 1974 and obtained Bachelor of Technology in Mathematics/Computer Science, at Federal University of Technology, Minna, Niger State, Nigeria in 2000. Then Master of Science in Computer Science, at University of Ibadan, Nigeria in 2006, and Doctor of Philosophy in Computer Science, at Federal University of Technology, Minna, Niger State, Nigeria in 2014. The major field of study is computer science. He carried out part of his PhD research at United Institute of Informatics Problems, National Academy of Sciences of Belarus (UIIP NASB) Minsk, Republic of Belarus. He is currently the Ag. Head, at the Department of Cyber Security Science, Federal University of Technology, Minna, Niger State, Nigeria. He has published twelve journal articles and four conference proceedings. His research interest includes Artificial Intelligence, Data Mining, Internet Technology, Database Management System, Software Architecture, Machine Learning, Human Computer Interaction and Computer Security. Dr. Alhassan is a member of Computer Professionals Registration Council of Nigeria (CPN).

**How to cite this paper:** Shafi'i Muhammad Abdulhamid, Maryam Shuaib, Oluwafemi Osho, Idris Ismaila, John K. Alhassan,"Comparative Analysis of Classification Algorithms for Email Spam Detection", International Journal of Computer Network and Information Security(IJCNIS), Vol.10, No.1, pp.60-67, 2018.DOI: 10.5815/ijcnis.2018.01.07