Heterogeneous Ensemble with Combined Dimensionality Reduction for Social Spam Detection

https://doi.org/10.3991/ijim.v15i17.19915

Abdulfatai Ganiyu Oladepo(✉), Amos Orenyi Bajeh, Abdullateef Oluwagbemiga Balogun, Hammed Adeleye Mojeed, Abdulsalam Abiodun Salman, Abdullateef Iyanda Bako
University of Ilorin, Ilorin, Nigeria
abdulfataig@gmail.com

Abstract—Spamming is one of the challenging problems within social networks which involves spreading malicious or scam content on a network; this often leads to a huge loss in the value of real-time social network services, compromise the user and system reputation and jeopardize users trust in the system. Existing methods in spam detection still suffer from misclassification caused by redundant and irrelevant features in the dataset as a result of high dimensionality. This study presents a novel framework based on a heterogeneous ensemble method and a hybrid dimensionality reduction technique for spam detection in micro-blogging social networks. A hybrid of Information Gain (IG) and Principal Component Analysis (PCA) (dimensionality reduction) was implemented for the selection of important features and a heterogeneous ensemble consisting of Naïve Bayes (NB), K Nearest Neighbor (KNN), Logistic Regression (LR) and Repeated Incremental Pruning to Produce Error Reduction (RIPPER) classifiers based on Average of Probabilities (AOP) was used for spam detection. To empirically investigate its performance, the proposed framework was applied on MPI_SWS and SAC'13 Tip spam datasets and the developed models were evaluated based on accuracy, precision, recall, f-measure, and area under the curve (AUC). From the experimental results, the proposed framework (Ensemble + IG + PCA) outperformed other experimented methods on studied spam datasets. Specifically, the proposed framework had an average accuracy value of 87.5%, an average precision score of 0.877, an average recall value of 0.845, an average F-measure value of 0.872 and an average AUC value of 0.943. Also, the proposed framework had better performance than some existing approaches. Consequently, this study has shown that addressing high dimensionality in spam datasets, in this case, a hybrid of IG and PCA with a heterogeneous ensemble method can produce a more effective model for detecting spam contents.

Keywords—high dimensionality, ensemble, spam detection

1 Introduction

An increase in penetration and access to the Internet along with developments in mobile technology in recent years has enhanced the popularity of Online Social Networks (OSNs) among Internet users. OSNs such as Twitter, Facebook, Sina Weibo, Instagram and so on, now has about 2.62 billion users across the globe and is expected to reach an estimated 3.02 billion by 2021 [1, 2]. Users on these networks communicate with one another by sharing and discussing both personal and public issues and events. This helps to build an intrinsic trust relationship among cyber friends (*followers/followees*) even though they may not know each other in person. Users usually feel more confident to read messages or even visit links from their cyber friends [3–5]. Micro-blogging Social Networks (MSNs) are also OSNs with specific characteristics such as (i) use of short messages composed of a limited number of characters; (ii) use of domain-specific words; (iii) high content of noisy data. MSN users can share short messages called *micro-post(s)* along with images and multimedia contents with other users [6]. They connect through a process of a follower-followee relationship. For instance, as illustrated in Figure 1, user A initiates a friendship connection with user B without user B acknowledging in return, hence user A is user B's follower and user B is followee to user A, while user B and user C are both follower and followee to each other.

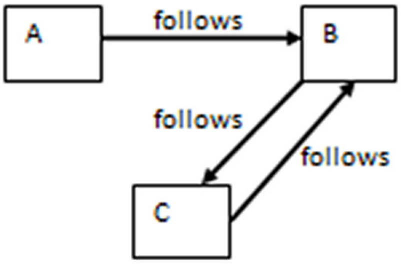

Fig. 1. User follower/followee relationship in MSNs

Location-Based Social Networks (LBSNs) is a type of micro-blogging social network where users share their geographic location, search for interesting places and post *tips* about existing locations. Examples of LBSNs include Apontador, Gowalla and Foursquare. Apontador, a popular Brazilian LBSN system has features that allow users to search for places, register new locations, check in locations and post tips or comment about these locations using smartphones. These tips help users, in addition to finding nearby and interesting places, to also read suggestions about what to order, what to buy or even what to avoid in specific places. Thus, allowing users to post tips and comment on places exposes the platform and other genuine users to spammers who then post unsolicited messages on tips and comments about locations [7–9]. Due to the popularity of micro-blogging social networks and the trust relationship built amongst cyber friends, MSN such as Twitter become a veritable platform for spammers to abuse and post malicious or spam content.

Spam involves the spreading of phishing, malicious, or scam content on a network. Spamming attacks do not only lead to a loss in the value of real-time search services, but they also interfere with statistics presented by tweet mining tools and consume additional resources from users and systems (such as network bandwidth- leading to significant revenue loss for organizations); compromise the user and system reputation; they may also jeopardize users trust on the existing tips in the system [7, 10]. The alarming rate at which spamming activities take place on social networks and the inherent consequences make it worrisome and challenging to both users and providers of online social networks. According to Nexgate's 2013 report on the state of social media spam: during the first half of 2013, the growth of social spam was 355%, much higher than the growth rate of accounts and messages on branded social networks [11].

As it is evident from the foregoing, the need arises for research into methods of identifying spammers and spam content on micro-blogging social networks. Adewole, et al. [12] asserted that a majority of studies on spam detection have been on detecting spammers' accounts and only little has focused on spam message detection. Although many spam/spammer detection methods have been proposed in several studies, most of which are based on content analysis of users' data interaction; learning classifications that use topological features, sociological/behavioural characteristics of nodes within and across the social structure. Few kinds of research on social spam detection and classification used content-based and social structure analysis. Benevenuto, et al. [10] and Zheng, et al. [13] in their respective studies used a support vector machine (SVM) based algorithm for spammer classification. Barushka and Hajek [14], Abulaish and Bhat [15] and Bhat, et al. [16] evaluated the performance of some ensemble learning methods using topology-based learning for social spam detection. However, redundant and irrelevant features as a result of high dimensionality are still a long-term problem for social spam detection. The overhead effect of misclassification in spam detection as a result of low spam detection accuracy caused by this problem can be very risky. Removal of such features with spectral information enhances the classification process as well as accurate classification decisions [17–19]. Dimensionality reduction is the transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation still retains the meaningful properties of the original data. It applies specific techniques for reducing the number of input variables in training data for predictive models. Fewer input dimensions often mean correspondingly fewer parameters or a simpler structure in the machine learning model.

The integration of high dimensionality reduction methods will further enhance the performance of classifiers and ensemble methods in spam detection. This study investigates spam detection in micro-blogging social networks using content and behavioural features from a hybrid dimensionality reduction technique, with a heterogeneous ensemble learning method on Apontador (a location-based social network) datasets. The specific objectives are highlighted as follow:

i. To design a hybrid dimensionality reduction method for spam detection in micro-blogging social networks
ii. To develop a spam detection framework that integrates the designed hybrid dimensionality reduction in (i) with heterogeneous ensemble models
iii. To investigate the performance of the developed framework empirically on publicly available spam detection dataset
iv. To validate the usefulness of the framework by comparing its performance with existing methods proposed in the literature

Summarily, the main contributions of this study are:

i. This study proposed a novel spam detection framework based on heterogeneous ensemble and a combination of dimensionality reduction techniques.
ii. An empirical study to show the impact of dimensionality reduction techniques on ensemble methods in spam detection.

The rest of this paper is organized as follows: Section 2 outlines the review and analysis of existing related studies. Section 3 presents the research methods which include the classifiers, datasets, the experimental framework and performance evaluation metrics used in this study. Section 4 presents the experimental results and a discussion of our findings. Section 5 presents the conclusion and highlights the future works of this study.

2 Related works

A lot of research has been conducted on spam detection in domains such as email, short message service (SMS), webpage, and social networks. More studies are still needed to be done in these areas especially social network domain viz-a-viz micro-blogging social network and location-based social networks. Generally, spam detection methods have focused on various characteristics or features of the messages and/or users via two main approaches—content-based and user/behavioural-based learning. Furthermore, spam detection in online social networks (OSNs) has explored the following technique in spam detection: blacklist, graph-based, and Machine Learning (ML); all of which could adopt either or both content and user/behavioural-based learning.

Grier, et al. [20] as cited in Adewole, et al. [12], applied a blacklist-based approach to detect malicious tweets on the Twitter network. They investigated users' click-through data generated from the phishing URL's clicks to study the effectiveness of using malicious URLs to launch large-scale phishing attacks. They further analyzed the capability of blacklist-based approach in spam detection, but their findings suggested that the approach is very slow in protecting users from being compromised.

For the graph-based approach; Ahmed and Abulaish [21] proposed a Markov clustering algorithm (MCL) to classify a set of profiles on the social network as spam and non-spam. They applied the majority vote technique to examine the overlapping clusters generated using the MCL algorithm; while Ghosh, et al. [22] analyzed link farming activities on Twitter and proposed a CollusionRank algorithm to penalize users that connect with spammers on the network, thereby discouraging the activities of link farming by lowering users' score for connecting with spammers [20].

For the Machine Learning approach; Adewole, et al. [12] in their study, proposed an ensemble streaming framework that is based on classification and clustering for spam detection and risk assessment. They used a combination of Multinomial Naïve Bayes (MNB) and modified K-Nearest Neighbour (KNN) classifiers and the majority vote technique as the ensemble method for classifying messages. The risk assessment function was then computed from the risk score obtained from the outputs of MNB and KNN algorithms. Streaming K-means algorithm was used for the clustering to detect campaign of spam messages. They were, however, constrained to use the SMS spam dataset for training their classifiers via transfer learning due to the non-availability of real-life micro-blogging datasets.

A considerable number of studies have been conducted using the machine learning (ML) approach for spam detection in OSNs and other domains such as SMS, email and so on. By combining graph-based and ML approaches, Abulaish and Bhat [15] proposed an ensemble of classifier algorithms: J48; A variant of C4.5 Decision Tree; and Naïve Bayes (NB), using bagging and boosting methods to identify spam in OSN (Facebook dataset) based on topological and community features from users' interaction network. They observed that the performance of NB and J48 using bagging or boosting ensemble learning methods is better than their respective individual performances. However, the ensemble method using the J48 classifier showed a better performance than that of NB.

Two different works by Benevenuto, et al. [10] and Zheng, et al. [13] considered content and user/behavioural attributes of their datasets, and both applied non-linear SVM classifier with Radial Basis Function (RBF) kernel - for the control of overfitting of the model and degree of nonlinearity. Benevenuto, et al. [10] worked on crawled Twitter dataset, their model identified spammers with 70.1% accuracy and non-spammers with 96.4%. Out of the 96 features trained by the SVM model, only 10 were found to be discriminatory. On the other hand, Zheng, et al. [13] performed their experiment on crawled, manually labeled Sina Weibo dataset and obtained 99.1% spammer detection accuracy and 99.9% non-spammer. The SVM model was found to perform better than the NB and Bayesian Networks (BN) upon the comparison.

While most of the previous studies have approached social/microblogging spam detection as a classification problem; Miller, et al. [23] viewed it as an anomaly detection problem. They proposed a modified StreamKM++ and DenStream clustering algorithm for spam detection on Twitter. Their model achieved 99% recall and 6.4% false-positive rate (FPR) using StreamKM++; and 99% recall and a 2.8% FPR with DenStream. When used together, they achieved 100% recall (meaning it identified all spammers in the test data) and 2.2% FPR (meaning it incorrectly detected just 2.2% of normal users as spammers).

The motivation for this study was derived from the afore-stated researches as they further identified the need for studies involving ensemble methodology and selection of important features for the task of spam detection.

3 Methodology

This section presents the baseline classifiers, dimensionality reduction methods, spam datasets, performance evaluation metrics and experimental framework used in this study.

3.1 Classification algorithm

This sub-section presents the baseline classification algorithms used in this study. These classifiers were selected based on their respective computational complexity which is aimed at introducing diversification to the classification process, hence, the heterogeneity in the ensemble method.

Naïve Bayes (NB). This machine-learning algorithm was derived from the Bayes rule and it assumes that independent attributes of observation(s) are completely independent of each other, given a dependent variable [24]. According to Mitchell [25], when X contains n attributes that are conditionally independent of themselves given Y, the Naïve Bayes algorithm is expressed as

$$P(X_1 \ldots X_n|Y) = \prod_{I=1}^{n} P(X_i|y) \tag{1}$$

Considering the training of a classifier whose output is the probability distribution over possible values of Y, based on new instance X that is to be classified. Also, assuming they (X_i) are conditionally independent given Y, then Equation 1 becomes:

$$P(Y = y_k|X_1 \ldots X_n) = \frac{P(Y = y_k)\prod_i P(X_i|Y = y_k)}{\sum_j P(Y = y_j)\prod_i P(X_i|Y = y_j)} \tag{2}$$

The fundamental of the Naïve Bayes classifier is expressed in Equation 2. However, the most probable value of Y is the actual interest thus, the Naïve Bayes classification rule is expressed in Equation 3 below. First, from Equation 2, we derived

$$Y \leftarrow \arg_{y_k}^{\max} \frac{P(Y = y_k)\prod_i P(X_i|Y = y_k)}{\sum_j P(Y = y_j)\prod_i P(X_i|Y = y_j)}$$

This is then simplified to Equation 3 as the denominator is not dependent on y_k.

$$Y \leftarrow \arg_{y_k}^{\max} P(Y = y_k)\prod_i P(X_i|Y = y_k) \tag{3}$$

Equation 3 above is the simplified Naïve Bayes classification rule that outputs the most probable value of Y having considered all $X_1 \ldots X_n$ values to be independent of each other.

K-nearest neighbor (KNN). KNN is an example of instance-based learners Reduction (it is used interchangeably as IBK in this study). Nearest-neighbour classifiers are based on learning by analogy, that is, by comparing a given test tuple with training tuples that are similar to it. The training tuples are described by n attributes. Each tuple represents a point in an n-dimensional space. In this way, all the training tuples are stored in n-dimensional pattern space. When given an unknown tuple, a k-nearest-neighbour classifier searches the pattern space for the k training tuples that are closest to the unknown tuple [26]. These k training tuples are the k "nearest neighbours" of the unknown tuple. KNN is a type of lazy learning where the function is only approximated locally and all computations are deferred until classification. An object is classified by a majority of its neighbours. K is always a positive integer and the neighbours are selected from a set of objects for which the correct classification is known [27, 28].

RIPPER. RIPPER is the short form of Repeated Incremental Pruning to Produce Error Reduction (it is used interchangeably as JRip in this study). It is an optimized version of Incremental Reduced Erro Prunning (IREP) based on association rules with reduced error pruning [29]. This algorithm is a rule induction method that implements a propositional rule learner. It greedily learns rules from a given dataset by employing a divide and conquered strategy [30]. Concerning the class frequencies inherent in a given dataset, sorting of training data is being executed in an ascending manner by class labels. Thus, beginning from the smallest, rules are being generated and learned for n–1 classes. As a result, instances covered by the rules are removed from the original data repeatedly until all instances are completely removed.

Logistic regression (LR). Logistic Regression (LR) is a discriminative ML method that is based on logistic function. LR focuses only on the posterior probability of each class. it is a generalized linear model, mapping the output of linear multiple regression to the posterior probability of each class [31].

3.2 Dimensionality reduction technique

This sub-section presents the dimensionality reduction techniques deployed in this study. Specifically, Information Gain (IG) and Principal Component Analysis (PCA) were selected for the dimensionality reduction techniques.

Information gain (IG). Information Gain (IG) is a feature selection (FS) method for selecting relevant features from available features for any given data. According to Jain and Bhupendra [32], IG generates the best subset of features among the original features based on ‘Entropy’. Usually, the entropy of each feature of data is computed and arranged in descending order. Hence, features with lower entropy scores are discarded while those with high entropy scores are selected for creating a subset of the original data to be used for model development.

To compute IG, the expected information required to categorize a record in a data table is first computed after which the expected information required for each attribute is also computed. To obtain the IG of each feature, the information score for each attribute is subtracted from the expected information of the given data table.

Principal component analysis (PCA). Principal Component Analysis (PCA) is a multivariate statistical method for analyzing several variables to reduce large dimensional data to a relatively small number of dimensions or components [33]. As a tool, PCA is quite applicable in several use-cases such as for the visualization of genetic distance or relatedness between populations [34]. However, in this research, PCA is used for dimensionality reduction (as a method for feature extraction). It is done using the eigenvalue decomposition of a data correlation (or covariance) matrix after executing the normalization phase of the original data.

Algorithmically, PCA orthogonally transforms collections of observations of possibly correlated features into another set of linearly uncorrelated values (i.e. principal components), using the following steps as shown in [35]:

1. Collect the original data having d-dimensional observations ignoring the class label
2. Execute the standardization of the d-dimensional observations
3. Compute the mean vector of the d-dimensional data
4. Compute the covariance matrix of the whole data set
5. Compute the eigenvector as well as the corresponding eigenvalues

6. Sorting of eigenvectors and selection of k eigenvectors with the largest eigenvalues from a d x k dimensional matrix W (where every column represents an eigenvector).
7. Use the obtained d x k eigenvector matrix to transform the observation onto a new subspace.

3.3 Spam dataset

In this study, two spam datasets from Costa, et al. [7] and Dutta, et al. [36] are used for training and testing the proposed models. These datasets (herein referred to as Dataset 1 and Dataset 2) are about "Tip Spam" in location-based social networks. Specifically, Dataset 1 is based on Apontador and consists of 60 attributes, 7076 instances and 2 class labels (spam or non-spam)[7]. Dataset 2 was created by Costa, et al. [37]. The dataset consists of 41 attributes and 2762 instances with 2 class labels (spam or non-spam). Both datasets have an equal distribution of class labels. That is, both datasets are balanced with an equal number of spam and non-spam instances.

3.4 Performance evaluation metrics

For comprehensive performance evaluation, accuracy, precision, recall, f-measure and area under curve (AUC) values are used to measure the efficacy of the spam detection models developed in this study. Our preference for these evaluation metrics is based on their wide usage in existing studies on social spam detection [6, 7, 17, 36–38] and their suitability for achieving the objectives of this study.

I. Accuracy measures the percentage of correctly classified spam instances to the total number of instances [26] and its value is calculated as thus:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{4}$$

II. Precision measure the rate of the actual number of detected spam instances that are spam instances. It is represented as;

$$\text{Precision} = \left(\frac{\text{TP}}{\text{TP} + \text{FP}} \right) \tag{5}$$

III. Recall measures the rate of spam instances that are correctly classified. Its formula is given as:

$$\text{Recall} = \left(\frac{\text{TP}}{\text{TP} + \text{FN}} \right) \tag{6}$$

IV. F-measure measures the harmonic mean of precision and recall.

$$\text{F} - \text{Measure} = 2 \times \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \right) \tag{7}$$

wherein Equations (4), (5), (6) and (7), TP = True Positive which implies the accurate classification; FP = False Positive which implies inaccurate classification; TN = True Negative which implies accurate misclassification; and FN = False Negative which implies inaccurate misclassification.

V. The area under the curve (AUC), which is also known as Area under the ROC (Receiver operating characteristics) curve shows the trade-off between TP rate and FP rate [26, 39]. It provides an aggregate measure of performance across all possible classification thresholds.

3.5 Experimental framework

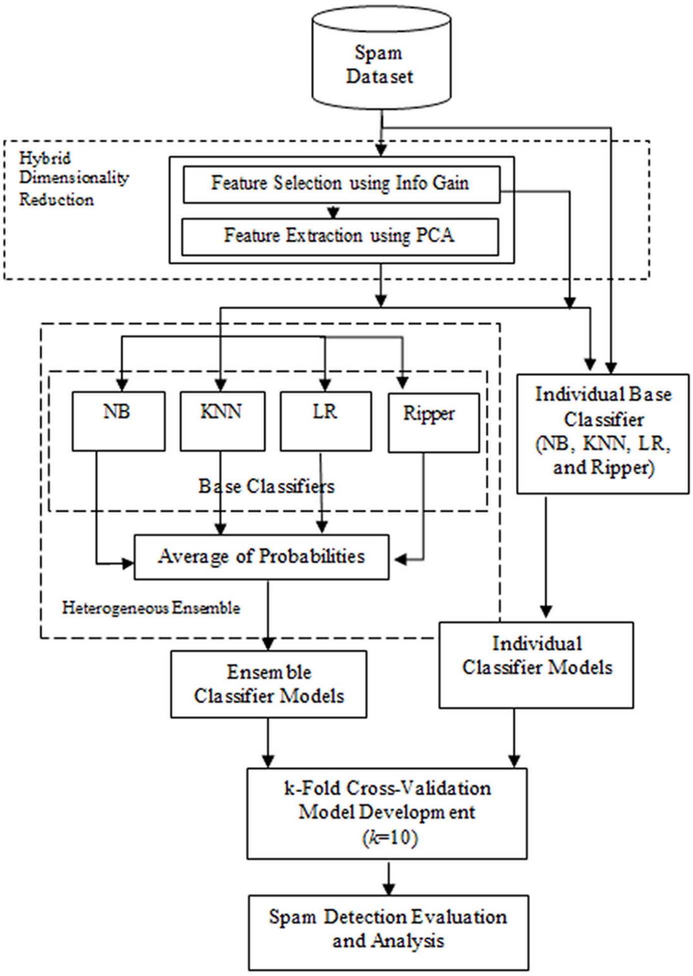
Spam Dataset
Hybrid Dimensionality Reduction
Feature Selection using Info Gain
Feature Extraction using PCA
NB
KNN
LR
Ripper
Base Classifiers
Average of Probabilities
Heterogeneous Ensemble
Individual Base Classifier (NB, KNN, LR, and Ripper)
Ensemble Classifier Models
Individual Classifier Models
k-Fold Cross-Validation Model Development (k=10)
Spam Detection Evaluation and Analysis

Fig. 2. Experimental framework

To validate the efficacy of the proposed framework for spam detection, an experimental framework as shown in Figure 2 is developed. The experimental process is divided into two phases:

1. **Pre-process Phase**: In this phase, spam datasets are pre-processed by a hybrid IG+PCA dimensionality reduction method to reduce the dimensionality of each spam dataset. Specifically, IG based on the Ranker search method is used to select top-ranked log2N relevant features (where N is the total number of features in the dataset). Thereafter, the selected features from IG are passed through PCA to assess and generate the optimum subset of features. The essence of passing the IG selected features through PCA for further processing as proposed is to address the bias of IG towards features with a large range of values [40, 41]. The output from this phase is the pre-processed features from each dataset which are passed into the next phase (model construction phase) for the development of spam detection models.
2. **Model Construction Phase:** Optimal feature subsets from the pre-processing phase are used for spam detection model construction. In this study, a heterogeneous ensemble method based on the average of probabilities rule is developed for spam detection. The goal of the heterogeneous ensemble method is to harness and aggregate the performance of individual baseline classifiers for classification processes [18, 29, 42]. Specifically, NB, IBK, LR and JRip baseline classifiers are used to develop a heterogeneous ensemble framework. Each of the baseline classifiers is based on different computational characteristics, hence, the heterogeneity. The detection models are developed based on the 10-fold cross-validation (CV) technique. The preference for a 10-fold CV is based on its ability to produce models with low bias and variance [43–46]. Also, spam detection models with or without dimensionality reduction were developed to have an unprejudiced comparison and to evaluate the effect of dimensionality reduction and ensemble methods in spam detection.

In the end, the performance of ensuing spam detection models is evaluated and analyzed based on accuracy, precision, recall, f-measure and AUC. All experiments were carried out using the WEKA machine learning tool [47].

3.6 Research method

The research method adopted in this study is the quantitative empirical method. In the empirical research method, the investigation is based on observation and measurement of phenomena as based on direct real-life experience. In this study, a real-life dataset based on “Tip Spam” in location-based social networks were used to perform several experimental investigations to evaluate and validate the suitability, effectiveness and significance of the proposed approach.

4 Results and discussion

The section presents the experimental results of the baseline classifiers and ensembles on the two datasets used in this study. Tables 1–6 present the experimental results of the classifiers and ensembles on Dataset 1 and Dataset 2 based on accuracy, precision, recall, f-measure and area under the curve (AUC).

Table 1. Experimental results of spam models on Dataset 1 data without dimensionality reduction

Classifiers	Accuracy	Precision	Recall	F-Measure	AUC
NB	78.20	0.806	0.782	0.778	0.837
IBK	79.22	0.792	0.792	0.792	0.799
LR	82.33	0.827	0.823	0.823	0.892
JRip	83.92	0.845	0.839	0.839	0.878
Ensemble	84.71	0.847	0.847	0.846	0.913

Table 2. Experimental results of spam models on Dataset 2 without dimensionality reduction

Classifiers	Accuracy	Precision	Recall	F-Measure	AUC
NB	62.54	0.708	0.625	0.584	0.858
IBK	79.89	0.799	0.799	0.799	0.789
LR	85.40	0.856	0.854	0.854	0.926
JRip	87.08	0.872	0.871	0.871	0.907
Ensemble	87.92	0.873	0.879	0.879	0.933

From Table 1 and Table 2, it can be observed that the heterogeneous ensemble of the baseline classifiers (NB, IBK, LR, and Jrip) based on average of probabilities (AOP) outperforms all the considered baseline classifiers on all performance in both datasets. Specifically, the heterogeneous ensemble method had the highest accuracy value (84.71%), precision (0.847), recall (0.847), f-measure (0.847), and AUC (0.913) on Dataset 1 and accuracy value (87.92%), precision (0.873), recall (0.879), f-measure (0.879), and AUC (0.933) on Dataset 2 when compared with other baseline classifiers as presented in Table 1 and Table 2. Amongst the baseline classifiers, JRip performed best on all performance metrics with accuracy value (83.92%), precision (0.845), recall (0.839), f-measure (0.839), and AUC (0.878) on Dataset 1 and accuracy value (87.08%), precision (0.872), recall (0.871), f-measure (0.871), and AUC (0.907) on Dataset 2. Although, the margin (in terms of performance metric values) between the heterogeneous ensemble method models may not be statistically significant, the adverse effect of allowing such predictive margin could be dangerous if single classifiers are used instead of ensemble methods. Besides, these results give further credence to the application and adoption of ensemble methods for prediction processes as ensemble methods have been proven to be better than single classifiers [6, 29].

Table 3. Experimental results of spam models on Dataset 1 with dimensionality reduction (IG)

Classifiers	Accuracy	Precision	Recall	F-Measure	AUC
NB+IG	79.69	0.813	0.797	0.794	0.859
IBK+IG	82.33	0.824	0.823	0.823	0.880
LR+IG	81.24	0.816	0.812	0.812	0.874
JRip+IG	82.69	0.835	0.827	0.826	0.848
Ensemble+IG	85.04	0.850	0.847	0.850	0.918

Table 4. Experimental results of spam models on Dataset 2 with dimensionality reduction (IG)

Classifiers	Accuracy	Precision	Recall	F-Measure	AUC
NB+IG	78.50	0.799	0.785	0.783	0.870
IBK+IG	79.48	0.795	0.795	0.795	0.824
LR+IG	81.84	0.818	0.815	0.815	0.892
JRip+IG	81.77	0.830	0.818	0.816	0.862
Ensemble+IG	88.79	0.875	0.878	0.881	0.942

Table 3 and Table 4 present the experimental results of spam models with one of the feature selection methods, Information Gain, which is a form of dimensionality reduction technique. This is to further improve the performance of the spam models (ensemble and base classifiers) as feature selection has been known to improve prediction models [48–51]. The heterogeneous ensemble method still outperforms the baseline classifiers on all performance metrics on both datasets. On Dataset 1, the heterogeneous ensemble had the highest accuracy value (85.04%), precision (0.85), recall (0.847), f-measure (0.85), and AUC (0.918). While on Dataset 2, the heterogeneous ensemble had the highest accuracy value (88.79%), precision (0.875), recall (0.878), f-measure (0.881), and AUC (0.942). There was a slight improvement in the accuracy values (+0.33; +0.87), AUC values (+0.05; +0.09) of the heterogeneous ensemble method with IG when compared with accuracy value without IG on Dataset 1 and Dataset 2 respectively. IG improved the predictive performance of the heterogeneous ensemble method.

Table 5. Experimental results of prediction models on Dataset 1with dimensionality reduction (IG+PCA)

Classifiers	Accuracy	Precision	Recall	F-Measure	AUC
NB+IG+PCA	80.77	0.819	0.808	0.806	0.868
IBK+IG+PCA	82.44	0.826	0.824	0.824	0.883
LR+IG+PCA	81.72	0.821	0.817	0.817	0.880
JRip+IG+PCA	82.44	0.834	0.824	0.823	0.852
Ensemble+IG+PCA	85.82	0.869	0.862	0.862	0.928

Table 6. Experimental results of prediction models on Dataset 2 with dimensionality reduction (IG+PCA)

Classifiers	Accuracy	Precision	Recall	F-Measure	AUC
NB+IG+PCA	79.72	0.819	0.797	0.794	0.879
IBK+IG+PCA	79.46	0.795	0.795	0.795	0.824
LR+IG+PCA	80.52	0.808	0.805	0.805	0.885
JRip+IG+PCA	82.28	0.829	0.823	0.822	0.871
Ensemble+IG+PCA	89.18	0.885	0.828	0.881	0.957

Table 5 and Table 6 present the experimental results of the proposed framework **(Ensemble+IG+PCA)** in comparison with base classifiers. In this experiment, the Ensemble+IG+PCA models outperformed other and base classifier models on all performance metrics. On Dataset 1, the Ensemble+IG+PCA model had the highest accuracy value (85.82%), a precision score of 0.869, an F-Measure value of 0.862 and an AUC value of 0.928. A similar case was observed in the performance of the proposed framework on Dataset 2. Ensemble+IG+PCA outperform all other methods as presented in Table 6. Figures 3 and 4 show the graphical illustration of the accuracy values of the Ensemble+IG+PCA model and other baseline classifiers on Dataset 1 and Dataset 2 respectively. Also, Figures 5 and 6 present graphically the performance metric values (AUC, f-measure, precision and recall) of the heterogeneous ensemble model and other experimented methods on both datasets respectively.

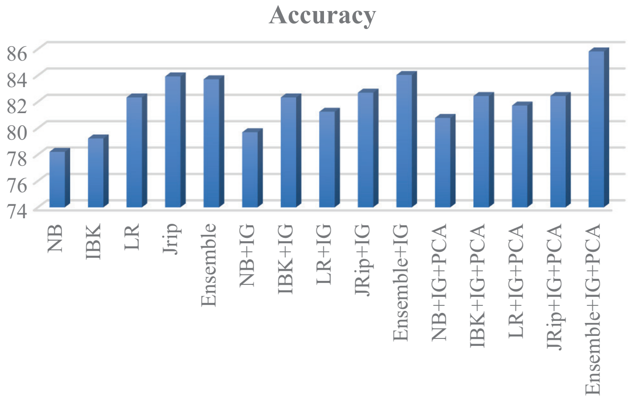

Fig. 3. Performance accuracies of all models on Dataset 1

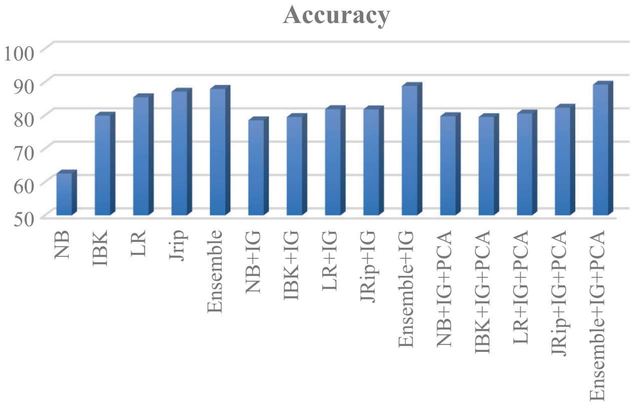

Fig. 4. Performance accuracies of all models on Dataset 2

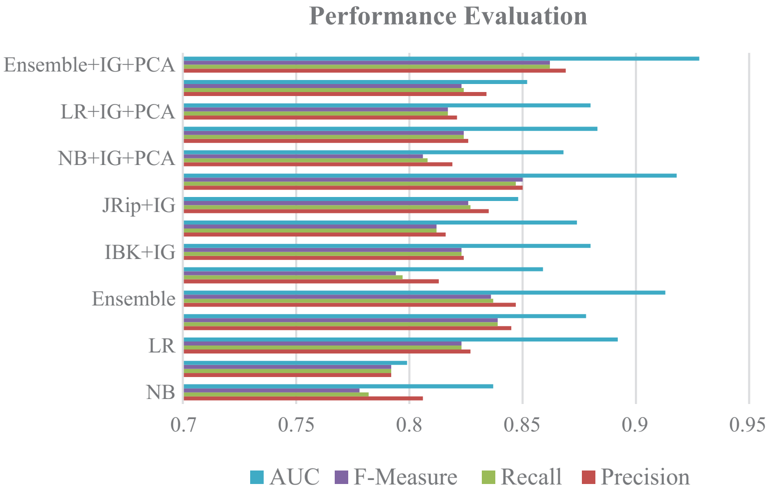

Fig. 5. Performances of the models on Dataset 1

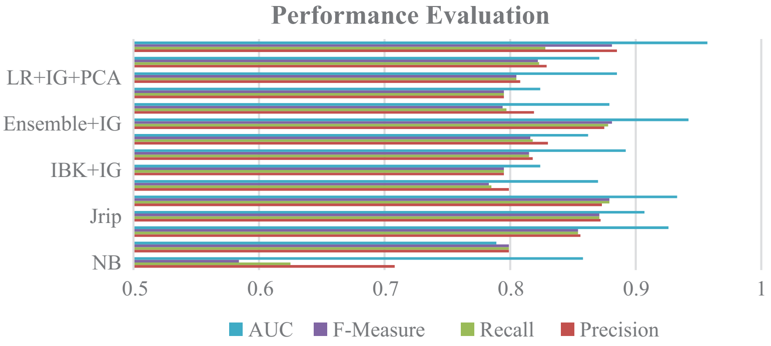

Fig. 6. Performances of the models on Dataset 2

Table 7. Average experimental results values for the spam models on Dataset 1 and Dataset 2

Models	Average Accuracy (%)	Average Precision	Average Recall	Average F-Measure	Average AUC
NB	70.37	0.757	0.703	0.681	0.8475
IBK	79.56	0.795	0.7955	0.7955	0.794
LR	83.87	0.841	0.8385	0.8385	0.909
Jrip	85.5	0.858	0.855	0.855	0.8925
Ensemble	86.31	0.860	0.863	0.8625	0.923
NB+IG	79.09	0.806	0.791	0.7885	0.8645
IBK+IG	80.91	0.809	0.809	0.809	0.852
LR+IG	81.54	0.817	0.8135	0.8135	0.883

(Continued)

Table 7. Average experimental results values for the spam models on Dataset 1 and Dataset 2 *(continued)*

Models	Average Accuracy (%)	Average Precision	Average Recall	Average F-Measure	Average AUC
JRip+IG	82.23	0.832	0.8225	0.821	0.855
Ensemble+IG	86.91	0.862	0.8625	0.8655	0.9300
NB+IG+PCA	80.24	0.819	0.8025	0.800	0.8735
IBK+IG+PCA	80.95	0.810	0.8095	0.8095	0.8535
LR+IG+PCA	81.12	0.814	0.811	0.811	0.8825
JRip+IG+PCA	82.36	0.831	0.8235	0.8225	0.8615
*Ensemble+IG+PCA	87.50	0.877	0.845	0.8715	0.9425

Note: *Indicates proposed method.

Table 7 presents the performance metric values for the spam models on Dataset 1 and Dataset 2. As depicted in Table 7, using the NB as a single classifier produced an average accuracy of 70.37% which improved to 79.09% when implemented with the IG FS method (NB+IG) and lastly, its accuracy slightly increased to 80.25% when combined with IG and PCA (NB+IG+PCA). Also, the IBK classifier produced an average accuracy of 79.56% which increased to 80.91% when combined with the IG (IBK+IG) and slightly increased to 80.95% with IBK+IG+PCA. As for the LR algorithm, the model's initial average accuracy was 83.87% but a reduction of the accuracy to 81.54% was recorded when combined with IG (LR+IG) and is further reduced to 80.95% when PCA feature extraction was implemented. The JRip algorithm had an initial average accuracy of 85.5% but dropped to 82.23% average accuracy when combined with the IG feature selection technique and increased to an average accuracy of 82.36% when the PCA feature extraction technique was combined. Finally, the ensemble method had an initial average accuracy of 86.315% and increased to 86.915% when combined with IG feature selection. However, the proposed framework (Ensemble+IG+PCA) had the highest average accuracy value (87.5%), average precision value (0.877), average recall (0.845), average f-measure (0.8715) and average AUC value (0.9425).

Table 8. Performance comparison of proposed methods with existing methods on Dataset 1

	Accuracy (%)	Precision	Recall	F-Measure	AUC
*Ensemble+IG+PCA	85.82	0.869	0.862	0.862	0.928
Dutta, et al. [36]	81.04	–	–	0.809	–

Note: *Indicates proposed method.

Table 9. Performance comparison of proposed methods with existing methods on Dataset 2

	Accuracy (%)	Precision	Recall	F-Measure	AUC
*Ensemble+IG+PCA	89.18	0.885	0.828	0.881	0.957
Costa, et al. [7]	87.8	–	–	0.873	–
Agrawal and Velusamy [52]	82.5	–	–	–	–

Note: *Indicates proposed method.

Furthermore, Table 8 and Table 9 present the performance comparison of the proposed method (Ensemble+IG+PCA) and some recent approaches from existing studies on Dataset 1 and Dataset 2. It shows that the proposed methods outperform some of the existing recent approaches based on the considered performance metrics. Conclusively, it is evident that the proposed method can detect spam messages more effectively than some existing methods.

5 Conclusions and future works

This study focused on proposing an effective machine-learning-based spam message detection framework by implementing machine learning techniques (KNN, LR, RIPPER, and NB), dimensionality reduction method (feature selection: IG and feature extraction: PCA), and ensemble methods (AOP technique). Specifically, a spam message detection framework based on a heterogeneous ensemble framework and a combination of dimensionality reduction techniques was proposed and implemented. Evidently, from the results of the experiments, it was observed that removing redundant and irrelevant features from spam datasets using hybridized feature selection and feature extraction method in conjunction with the heterogeneous ensemble method provides an effective method for detecting social spam contents. This proves that better methods for spam detection can be developed by addressing underlining issues such as the high dimensionality of datasets. Consequently, it is recommended that more studies can be conducted by combining other dimensional reduction techniques as well as other forms of ensemble method to provide a generalizable social spam message detection model(s) with effective detection rates.

6 References

[1] D. V. Dimitrova and J. Matthes, “Social media in political campaigning around the world: Theoretical and methodological challenges,” ed: SAGE Publications Sage CA: Los Angeles, CA, 2018. https://doi.org/10.1177/1077699018770437

[2] H. Shen and X. Liu, “Detecting spammers on Twitter based on content and social interaction,” presented at the 2015 International Conference on Network and Information Systems for Computers, 2015. https://doi.org/10.1109/ICNISC.2015.82

[3] M. Singh, A. Singh, D. Bansal, and S. Sofat, “An analytical model for identifying suspected users on Twitter,” *Cybernetics and Systems,* vol. 50, no. 4, pp. 383–404, 2019. https://doi.org/10.1080/01969722.2019.1588968

[4] M. Almseidin, A. A. Zuraiq, M. Al-Kasassbeh, and N. Alnidami, “Phishing detection based on machine learning and feature selection methods,” *International Journal of Interactive Mobile Technologies,* vol. 13, no. 12, 2019.

[5] L. F. Hussein, A. B. Aissa, I. A. Mohamed, S. Alruwaili, and A. Alanzi, “Development of a secured vehicle spot detection system using GSM,” *International Journal of Interactive Mobile Technologies,* vol. 15, no. 4, 2021. https://doi.org/10.3991/ijim.v13i12.11411

[6] K. S. Adewole, N. B. Anuar, A. Kamsin, and A. K. Sangaiah, “SMSAD: a framework for spam message and spam account detection,” *Multimedia Tools and Applications,* vol. 78, no. 4, pp. 3925–3960, 2019. https://doi.org/10.1007/s11042-017-5018-x

[7] H. Costa, F. Benevenuto, and L. H. Merschmann, "Detecting tip spam in location-based social networks," presented at the Proceedings of the 28th Annual ACM Symposium on Applied Computing, 2013. https://doi.org/10.1145/2480362.2480501

[8] D. Ibrahim and N. Alruhaily, "Anomaly detection in wireless sensor networks: a proposed framework," *International Journal of Interactive Mobile Technologies,* vol. 14, no. 10, 2020. https://doi.org/10.3991/ijim.v14i10.14261

[9] A. Odeh, I. Keshta, and E. Abdelfattah, "Efficient detection of phishing websites using multilayer perceptron," *International Journal of Interactive Mobile Technologies,* vol. 14, no. 11, 2020. https://doi.org/10.3991/ijim.v14i11.13903

[10] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on Twitter," presented at the Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (CEAS), 2010.

[11] H. Nguyen, "2013 state of social media spam," *Publication of NexGate, USA, from websites,* 2013. http://nexgate.com/wpcontent/uploads/2013/09/Nexgate-2013-State-of-Social-Media-Spam-Research-Report.pdf

[12] K. S. Adewole, N. B. Anuar, A. Kamsin, K. D. Varathan, and S. A. Razak, "Malicious accounts: dark of the social networks," *Journal of Network and Computer Applications,* vol. 79, pp. 41–67, 2017. https://doi.org/10.1016/j.jnca.2016.11.030

[13] X. Zheng, Z. Zeng, Z. Chen, Y. Yu, and C. Rong, "Detecting spammers on social networks," *Neurocomputing,* vol. 159, pp. 27–34, 2015. https://doi.org/10.1016/j.neucom.2015.02.047

[14] A. Barushka and P. Hajek, "Spam filtering in social networks using regularized deep neural networks with ensemble learning," presented at the IFIP International Conference on Artificial Intelligence Applications and Innovations, 2018. https://doi.org/10.1007/978-3-319-92007-8_4

[15] M. Abulaish and S. Y. Bhat, "Classifier ensembles using structural features for spammer detection in online social networks," *Foundations of Computing and Decision Sciences,* vol. 40, no. 2, pp. 89–105, 2015. https://doi.org/10.1515/fcds-2015-0006

[16] S. Y. Bhat, M. Abulaish, and A. A. Mirza, "Spammer classification using ensemble methods over structural social network features," presented at the Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 02, 2014. https://doi.org/10.1109/WI-IAT.2014.133

[17] A. G. Akintola, A. O. Balogun, F. Lafenwa-Balogun, and H. A. Mojeed, "Comparative analysis of selected heterogeneous classifiers for software defects prediction using filter-based feature selection methods," *FUOYE Journal of Engineering and Technology,* vol. 3, no. 1, pp. 133–137, 2018. https://doi.org/10.46792/fuoyejet.v3i1.178

[18] A. O. Ameen, A. O. Balogun, G. Usman, and G. S. Fashoto, "Heterogeneous ensemble methods based on filter feature selection," *Computing, Information Systems, Development Informatics & Allied Research Journal,* vol. 7, no. 4, pp. 63–78, 2016.

[19] P. Petrov, S. Ivanov, P. Dimitrov, G. Dimitrov, and O. Bychkov, "Projects management in technology start-ups for mobile software development," *International Journal of Interactive Mobile Technologies,* vol. 15, no. 7, 2021. https://doi.org/10.3991/ijim.v15i07.19291

[20] C. Grier, K. Thomas, V. Paxson, and M. Zhang, "@ spam: the underground on 140 characters or less," presented at the Proceedings of the 17th ACM conference on Computer and communications security, 2010. https://doi.org/10.1145/1866307.1866311

[21] F. Ahmed and M. Abulaish, "An mcl-based approach for spam profile detection in online social networks," presented at the 2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications, 2012. https://doi.org/10.1109/TrustCom.2012.83

[22] S. Ghosh *et al.*, "Understanding and combating link farming in the Twitter social network," presented at the Proceedings of the 21st international conference on World Wide Web, 2012. https://doi.org/10.1145/2187836.2187846

[23] Z. Miller, B. Dickinson, W. Deitrick, W. Hu, and A. H. Wang, "Twitter spammer detection using data stream clustering," *Information Sciences,* vol. 260, pp. 64–73, 2014. https://doi.org/10.1016/j.ins.2013.11.016

[24] A. O. Balogun, S. Basri, S. J. Abdulkadir, V. E. Adeyemo, A. A. Imam, and A. O. Bajeh, "Software defect prediction: analysis of class imbalance and performance stability," *Journal of Engineering Science and Technology,* vol. 14, no. 6, pp. 3294–3308, 2019.

[25] T. M. Mitchell, "Generative and discriminative classifiers: Naive Bayes and logistic regression," *Machine Learning,* pp. 1–17, 2010.

[26] J. Han and M. Kamber, "Data Mining: C d h Concepts and Techniques," 2012.

[27] P. Cunningham and S. J. Delany, "k-Nearest neighbour classifiers," *Multiple Classifier Systems,* vol. 34, no. 8, pp. 1–17, 2007.

[28] M. A. Mabayoje, A. O. Balogun, H. A. Jibril, J. O. Atoyebi, H. A. Mojeed, and V. E. Adeyemo, "Parameter tuning in KNN for software defect prediction: an empirical analysis," *Jurnal Teknologi dan Sistem Komputer,* vol. 7, no. 4, pp. 121–126, 2019. https://doi.org/10.14710/jtsiskom.7.4.2019.121-126

[29] A. O. Balogun, A. M. Balogun, P. O. Sadiku, and V. E. Adeyemo, "Heterogeneous ensemble models for generic classification," *Scientific Annals of Computer Science,* vol. 15, no. 1, pp. 92–98, 2017.

[30] Y. K. Jain, "Upendra: an efficient intrusion detection based on decision tree classifier using feature reduction," *International Journal of scientific and research Publications,* vol. 2, no. 1, 2012.

[31] P. Ranganathan, C. Pramesh, and R. Aggarwal, "Common pitfalls in statistical analysis: logistic regression," *Perspectives in Clinical Research,* vol. 8, no. 3, p. 148, 2017. https://doi.org/10.4103/picr.PICR_123_17

[32] A. Jain and L. Bhupendra, "Classifier selection models for intrusion detection system (IDS)," *Informatics Engineering, an International Journal (IEIJ),* vol. 4, no. 1, pp. 1–11, 2016.

[33] R. d. O. Santos, B. M. Gorgulho, M. A. d. Castro, R. M. Fisberg, D. M. Marchioni, and V. T. Baltar, "Principal component analysis and factor analysis: differences and similarities in nutritional epidemiology application," *Revista Brasileira de Epidemiologia,* vol. 22, p. e190041, 2019. https://doi.org/10.1590/1980-549720190041

[34] P. E. Jorgensen, S. Kang, M.-S. Song, and F. Tian, "Dimension reduction and kernel principal component analysis," *arXiv preprint arXiv:1906.06451,* 2019.

[35] S. Raschka, "Implementing a Principal Component Analysis (PCA) in Python step by step," ed, 2014.

[36] S. Dutta, S. Ghatak, R. Dey, A. K. Das, and S. Ghosh, "Attribute selection for improving spam classification in online social networks: a rough set theory-based approach," *Social Network Analysis and Mining,* vol. 8, no. 1, p. 7, 2018. https://doi.org/10.1007/s13278-017-0484-8

[37] H. Costa, L. H. Merschmann, F. Barth, and F. Benevenuto, "Pollution, bad-mouthing, and local marketing: the underground of location-based social networks," *Information Sciences,* vol. 279, pp. 123–137, 2014. https://doi.org/10.1016/j.ins.2014.03.108

[38] K. S. Adewole, T. Han, W. Wu, H. Song, and A. K. Sangaiah, "Twitter spam account detection based on clustering and classification methods," *The Journal of Supercomputing,* vol. 76, no. 7, pp. 4802–4837, 2020. https://doi.org/10.1007/s11227-018-2641-x

[39] S. Whalen and G. Pandey, "A comparative analysis of ensemble classifiers: case studies in genomics," presented at the 2013 IEEE 13th International Conference on Data Mining, 2013. https://doi.org/10.1109/ICDM.2013.21

[40] F. Salo, A. B. Nassif, and A. Essex, "Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection," *Computer Networks,* vol. 148, pp. 164–175, 2019. https://doi.org/10.1016/j.comnet.2018.11.010

[41] P. Nskh, M. N. Varma, and R. R. Naik, "Principle component analysis based intrusion detection system using support vector machine," presented at the 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2016. https://doi.org/10.1109/RTEICT.2016.7808050

[42] A. O. Balogun *et al.*, "Rank Aggregation Based Multi-filter Feature Selection Method for Software Defect Prediction," in *International Conference on Advances in Cyber Security,* 2020, pp. 371–383: Springer. https://doi.org/10.1007/978-981-33-6835-4_25

[43] Y. A. Alsariera, V. E. Adeyemo, A. O. Balogun, and A. K. Alazzawi, "AI meta-learners and extra-trees algorithm for the detection of phishing websites," *IEEE Access,* vol. 8, pp. 142532–142542, 2020. https://doi.org/10.1109/ACCESS.2020.3013699

[44] Y. A. Alsariera, A. V. Elijah, and A. O. Balogun, "Phishing website detection: forest by penalizing attributes algorithm and its enhanced variations," *Arabian Journal for Science and Engineering,* pp. 1–12, 2020. https://doi.org/10.1007/s13369-020-04802-1

[45] A. O. Balogun *et al.*, "Empirical analysis of rank aggregation-based multi-filter feature selection methods in software defect prediction," *Electronics,* vol. 10, no. 2, p. 179, 2021. https://doi.org/10.3390/electronics10020179

[46] V. E. Adeyemo, A. O. Balogun, H. A. Mojeed, N. O. Akande, and K. S. Adewole, "Ensemble-Based Logistic Model Trees for Website Phishing Detection," in *International Conference on Advances in Cyber Security,* 2020, pp. 627–641: Springer. https://doi.org/10.1007/978-981-33-6835-4_41

[47] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter,* vol. 11, no. 1, pp. 10–18, 2009. https://doi.org/10.1145/1656274.1656278

[48] M. A. Mabayoje, A. O. Balogun, A. O. Bajeh, and B. A. Musa, "Software defect prediction: effect of feature selection and ensemble methods," *FUW Trends in Science & Technology Journal,* vol. 3, no. 2, pp. 518–522, 2018.

[49] A.-B. Verónica Bolón, M. Amparo, and C. N. Sánchez, *Artificial Intelligence: Foundations, Theory, and Algorithms Feature Selection for High-Dimensional Data*. Springer, 2017.

[50] M. A. Mabayoje, A. O. Balogun, S. M. Bello, J. O. Atoyebi, H. A. Mojeed, and A. H. Ekundayo, "Wrapper feature selection based heterogeneous classifiers for software defect prediction," *Adeleke University Journal of Engineering and Technology,* vol. 2, no. 1, pp. 1–11, 2019.

[51] A. O. Balogun *et al.*, "Impact of feature selection methods on the predictive performance of software defect prediction models: an extensive empirical study," *Symmetry,* vol. 12, no. 7, p. 1147, 2020. https://doi.org/10.3390/sym12071147

[52] M. Agrawal and R. L. Velusamy, "PRISMO: priority based spam detection using multi optimization," presented at the International Conference on Big Data Analytics, 2018. https://doi.org/10.1007/978-3-030-04780-1_27

7 Authors

Abdulfatai Ganiyu Oladepo is an IT Service Management practitioner with a keen interest in Data Science, Machine Learning, and IT Project Management. He can be reached via his email address (abdulfataig@gmail.com).

Amos Orenyi Bajeh has a BSc and an MSc degree in Computer Science from the University of Ilorin where he is currently a Senior Lecturer in the Department of Computer Science at the same University. He has a PhD in Information Technology from Universiti Teknologi PETRONAS. Software measurement, software maintenance, machine learning and fuzzy inference system are his areas of research interest. He can be reached via his email address (bajehamos@unilorin.edu.ng).

Abdullateef Oluwagbemiga Balogun received his B.Sc. and M.Sc degrees in Computer Science from the University of Ilorin, Nigeria. Currently on his PhD in Information Technology at the Universiti Teknologi PETRONAS, Perak, Malaysia. He is an academic staff in the Department of Computer Science, Faculty of Communication and Information Sciences, University of Ilorin, Nigeria. His research interests include Search-Based Software Engineering, Software Quality Assurance, Machine Learning, Data Science. He can be reached via his email address (balogun.ao1@unilorin.edu.ng).

Hammed Adeleye Mojeed is a Lecturer in the Department of Computer Science, University of Ilorin, Ilorin Nigeria. He received a Master of Science in Computer Science with distinction from the University of Ilorin, Ilorin, Nigeria in 2019, a Diploma in Computer Networking from SIIT Global, New Delhi, India in 2014 and a Bachelor of Science in Computer Science with First Class Honors from University of Ilorin, Ilorin Nigeria in 2013. His research interests fall in the field of Empirical Search-Based Software Engineering, Software Project Planning and Management, Machine Learning, Optimization and Text Mining. He has authored/co-authored over 20 publications in reputable outlets. He is a member of the IEEE Nigeria Computer Chapter and a Graduate Member of Computer Professionals of Nigeria (GMCPN). He can be reached via his email address (mojeed.ha@unilorin.edu.ng).

Abdulsalam Abiodun Salman is an Associate Professor and Head of the Department of Library and Information Science, Faculty of Communication and Information Sciences, University of Ilorin, Ilorin, Nigeria. He can be reached via his email address (salman.aa@unilorin.edu.ng)

Abdullateef Iyanda Bako is an Associate Professor and Dean of the Faculty of Environmental Sciences, University of Ilorin, Ilorin, Nigeria. He can be reached via his email address (bako.ai@unilorin.edu.ng)

Article submitted 2020-11-18. Resubmitted 2021-07-02. Final acceptance 2021-07-03. Final version published as submitted by the authors.