Journal of Information Security, 2019, 10, 155-176
http://www.scirp.org/journal/jis
ISSN Online: 2153-1242
ISSN Print: 2153-1234

An Intelligent Model for Online Recruitment Fraud Detection

Bandar Alghamdi, Fahad Alharby

Naïf Arab University (NAUSS), Riyadh, KSA
Email: binnawar2006@hotmail.com, fmalharby@nauss.edu.sa

How to cite this paper: Alghamdi, B. and Alharby, F. (2019) An Intelligent Model for Online Recruitment Fraud Detection. *Journal of Information Security*, **10**, 155-176.
https://doi.org/10.4236/jis.2019.103009

Received: January 6, 2019
Accepted: July 8, 2019
Published: July 11, 2019

Abstract

This study research attempts to prohibit privacy and loss of money for individuals and organization by creating a reliable model which can detect the fraud exposure in the online recruitment environments. This research presents a major contribution represented in a reliable detection model using ensemble approach based on Random forest classifier to detect Online Recruitment Fraud (ORF). The detection of Online Recruitment Fraud is characterized by other types of electronic fraud detection by its modern and the scarcity of studies on this concept. The researcher proposed the detection model to achieve the objectives of this study. For feature selection, support vector machine method is used and for classification and detection, ensemble classifier using Random Forest is employed. A freely available dataset called Employment Scam Aegean Dataset (EMSCAD) is used to apply the model. Pre-processing step had been applied before the selection and classification adoptions. The results showed an obtained accuracy of 97.41%. Further, the findings presented the main features and important factors in detection purpose include having a company profile feature, having a company logo feature and an industry feature.

Keywords

Online Recruitment Fraud, Intelligent Model, Privacy

1. Introduction

In the last decade which is called the Internet and social Era, the integral parts of the modern landscape considered are the Internet and social media. In modern organizations, there is a wide use of the Internet and social media deployed in employee recruitment [1]. Recently, the cloud was integrated to the procedure of recruiting new members, where the managed cloud services or solutions are used

by human resource managers. Nevertheless, there are many violating risk threats increased by scams and frauds along with the wide interest and adopting such embedded software [2].

Cybercrime is one of the present risky crimes that face the world and threaten the individuals and organizations security causing substantial losses [3]. Based on cyber security ventures report 2021, the cost of cybercrime damages in the world is around $6 trillion annually [4]. Thus, we have an urgent need for Information Security to ensure the Confidentiality, Integrity and Availability (CIA) to combat these crimes. It can be done through the implementation of known information security strategies such as prevention, detection, and response.

In Saudi Arabia, the 2030 vision predicts that there will be a growth in job generation [5]. This growth requires government to assure the inclusion of CIA in job recruitment process to protect the individuals and organizations from cybercrime occurrence. ORF is considered as one type of cybercrimes that has appeared recently. ORF violates the privacy and financial funds of individuals and organizations by exploiting Internet technology and web service. It allows non-legitimate users to damage the reputations of the organizations [6].

Data mining methods have added to data analysis, knowledge mining, prediction and detection of cybercrime. They can be used to create an intelligent model which is very effective in detecting fraud and scams of the network. ORF disrupts the privacy of job seeker and bothers the reputation of organizations. Moreover, it causes loss of money for individuals. It happens when criminals post fake ads exploiting the automation recruitment to trap job seekers. To the best of our knowledge, there has only been one empirical research conducted up to now on this kind of cybercrime. This is the only research that has examined the online fraud issues and developed a new solution for detection.

Vidros, Kolias, Kambourakis, & Akoglu (2017) added many features of ORF to the public dataset (EMSCAD). The researchers recommended the research community to find a reliable detection model of ORF. Thus, we need to obtain a new reliable model to enhance the performance of classification based on pre-processing and feature selection phase [2].

2. Background

2.1. Online Recruitment System

The online recruitment system utilizes web-based tools like public internet or intranet to recruit staff [7]. Recruitment brings many benefits for firm's success by getting the best applicants in short time, highlighting the professional requirements, assessing applicants via interviews and welcoming the newly selected employees. Furthermore, this process makes the hiring process more affordable and productive without spending a lot of money. The critical components of online recruitment include tracking the status of candidates, employer's website, job portals, online testing, and social networking [8]. The advantages of e-recruitment include effectiveness, high value, easiness, and efficiency [9]. The

literature highlighted various vital benefits of e-recruitment such as Time reduction, Cost Reduction, Reach huge masses (employers and candidates), Filter Functions, and Development of Brand image.

Two different methods for online recruitment as defined by (Prasad & Kapoor, 2016) include posting the company's profile as well as the job requirements on job portals and creating an online recruitment page on the company's website [10].

2.2. Knowledge Discovery from Data (KDD)

Knowledge Discovery from Databases (KDD) can be defined as a recognition and extraction of valuable, genuine, useful, unique, and comprehensible correlations or patterns in the data [11]. **Figure 1** summarizes the KDD concept that depends on multiple sequence of performed processes. Each process depends on the success and output of prior phase or process. It affirms the iterative nature of the process and various feedback loops to indicate revision activities [12]. The conventional model for KDD process is represented in **Figure 1**.

The KDD model includes five main phases: Selection of relevant prior knowledge; Acquiring or creating targeted dataset; Preprocessing to handle missing values, noise and errors in the data; Transformation to create dataset form suitable for easily implementing data mining algorithms; Data mining, the decision-making activity to define models such as regression, classification or clustering to obtain patterns of interest, representational form, or rule sets and trees; and Interpretation and Evaluation with respect their validity, and visualization of the patterns and models.

Web mining investigates the data quantities available in the World Wide Web by extracting information from all available web documents and services. Web content is very dynamic due to the rapid growth nature and update [13]. Text mining or text classification refers to text analytics by extracting high qualified information from digital text [14]. Classification, clustering and regression, Classification techniques use labelled datasets, a supervised learning method, and involves learning and training phase that classifies data into various and multiple classes based on assigned attributes derived from dataset [15]. Clustering defines similar classes of items based on the similarity among objects or

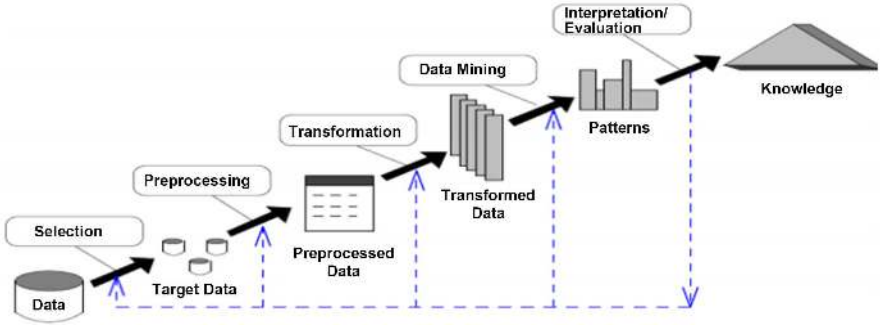

Figure 1. The conventional KDD model [2].

items. It can be used to perform a preprocessing phase for selection tasks [16]. Regression is one of predictive and statistical techniques used in numerical and continuous prediction through training process to determine the correlation among various attributes [17].

2.3. Data Mining Tools

Data mining tools provide different open source tools to allow access and perform data mining such as rapid miner, weka, and orange. Rapid Miner is an open source tool that utilizes the client-server model and deals with various data file format [18]. Waikato Environment for Knowledge Analysis (Weka) is a machine learning tool lunched by Waikato University [19] and offers a comprehensive range of preprocessing data and data modelling algorithms [18]. Orange is a very powerful open source data mining tool launched by the University of Ljubljana to support a wide range of widgets found in toolbox visualization and data analysis [18].

Data mining utilizes different methods for trend pattern and prediction tasks. The main methods deployed for data mining and diverse implementation methods include Classification is based on finding the basic rules to distribute items into specific defined classes, which is considered a predicting task [20].

3. Literature Review

There is rich literature about cybercrime detection models in different fields. However, there are only two studies one descriptive. One of them is an empirical study that has addressed the fraud and scams in the online recruitment. The related works often studied data mining techniques for various other detection purposes. A few research efforts addressed the online recruitment frauds.

Vidros *et al.* (2016) determined the frauds exposed by job seekers through online recruitment services. They found that ORF is a new field of current severe vulnerable. Three current methods of fraud and scams in the online recruitment include Fake Job Advertisement, Economic trickery using fake job advertisements published on online recruitment boards, and Reputed and real business enterprise publishing fake vacancy announcement [21].

Data Mining Techniques Related Works

Many studies have carried out data mining such as Yasin & Abuhasan (2016) that has provided an intelligent classification model to detect phishing emails using knowledge discovery, data mining and text processing techniques. A model based on knowledge discovery (KD) was proposed to build an intelligent email classifier to classify a new email message into legitimate or spam. The knowledge discovery model achieved high accuracy rates in classification of phishing emails that outperformed other schemes. Using the Random Forest algorithm and J48, 99.1% and 98.4% accuracy was achieved respectively. Using MLP classifier, TP rate and FP rate were 0.977 and of 0.026 respectively, while MLP achieved ROC

area of 0.987. The results of this study confirmed that the proposed model achieves high rates of accuracy in the classification of phishing e-mail [22].

Al-garadi, *et al.* (2016) introduced a study that has investigated cybercrime detection in online communications especially cyber bullying in Twitter. The main aim was to develop a number of unique features derived from Twitter. They included network, activity, user, and tweet content. A model to detect cyber bullying in Twitter was proposed using engineering features. The number of friends (followers), the number of users being followed (following), the following and followers ratio, and account verification status were collected through a survey. Users' activity features were also employed to measure the online communication activity of a user. The features implemented and included personality, gender and age. Naïve Bayes (NB), Support vector machine (SVM), Random forest and KNN were applied. Random forest showed f-measure 93%. The results of this study indicate that the proposed model contributes to providing a suitable solution for the detection of cyberbullying in online communication environments [23].

Sharaff, Nagwani, & Swami, 2015 investigated the impact of feature selection technique on email classification through studying the effect of two feature selection methods. A comparison was conducted between Bayes algorithm, tree-based algorithm J48 and support vector machine. Feature selection techniques included a Chi-Square (χ^2) and information gain. The best performance was gained using SVM classification technique which gave the overall best results without employing any feature selection techniques. There is no effect of Naïve Bayes on feature selection techniques. Further, J48 showed slight improvement with feature selection, whereas info-Gain performed better than Chi-square feature selection technique [24].

The research of Sornsuwit & Jaiyen (2015) created an Intrusion Detection Model Based on Ensemble Learning for User to Root (U2R) and Remote to Local (R2L) Attacks. The ensemble learning was concentrated to detect network intrusion data and reduce redundant features using a correlation-based algorithm. It can improve the accuracy of classifier by solving the determined problems applied on U2R and R2L attacks in KDD Cup'99 intrusion detection dataset. They applied an Adaboost algorithm to construct a strong classifier as linear combination of weak classifiers. Naïve Bayes was used to determine the appropriate class of unseen data and to compute the posterior probability for each class. A multilayer Perceptron (MLP) network was also used to perform linear mapping from input space to hidden space and from hidden space to output space. Support Vector Machine (SVM) approach was used to solve the classification problems based on an optimal hyper plane in a high-dimensional space. The result of this study show that reducing features contributes to improved efficiency in detecting attacks in works in many weak scales [25].

Gaikwad & Thool (2015) applied the bagging ensemble method of machine learning in order to provide a novel intrusion detection technique. Two instruments were utilized based on five modules including feature selection, REP Tree

design, and construction of main classifier, packet sniffer and detector. In addition, they proposed an intrusion detection system called bagging ensemble method algorithm. Weak classifier was used to improve the classification accuracy. The ensemble bagging machine-learning technique provided highest classification accuracy of 99.67%. They also revealed that the model building time and false positives exhibited by the method were lower as compared to AdaBoost algorithm with Decision stump base classifiers. The results of the study confirm that the bagging group with REPTree displays the highest accuracy of the classification. One advantage of using bagging method is that it takes less time to build the model. The proposed group method provides low false positives in comparison to other machine learning techniques [26].

Zuhaira, Selmat, & Salleh (2015) investigated the effect of feature selection on phish website detection by examining the effects of the feature selection approach on classification performance. An empirical test was conducted on a specific test-bed set to extract a large number of hybrid features. Four feature selection algorithms (FSAs) included CBF, WFS, χ^2, and IG. A comparison between classification models was performed to qualify the way of the features selective subsets shift detection accuracy, specificity and sensitivity of the classification model to the best rates. Some feature selection methods significantly outperformed their competitors by exhibiting better robustness, prediction, and performance. The results of the experiment showed a signification improvement in detection accuracy with low latency and accuracy of observation in the sensitivity of hardness and predictability. The results of the study contributed to providing the best possible subset of features for strong selection and effective phishing detection [27].

Nizamani, *et al.* (2014) proposed a fraudulent email detection model based on advanced feature choice. J48 classification algorithms technique was used due to its simplification and inductive nature. Support Vector Machine (SVM) was also used to transform non-linearly separable data to a new linearly separable data by using kernel trick. Moreover, Naive Bayes' (NB) classification algorithm was used to calculate the probabilities of the feature values for each of the classification categories. Further, cluster-based classification model (CCM) was applied to perform the classification task by grouping the data points based on obvious features. The dataset contains 8000 emails in total. The frequency-based features attain high accuracy for the task of fraudulent email detection regardless of choice of classification method. The model employed features extracted from the content of the emails achieving accuracy as high as 96%. The results of the study showed that the level of correctness was pretentious by the kind of determined features rather than the classifiers' type [28].

Shrivas & Dewangan (2014) presented an approach based on ANN-Bayesian Net-GR approach that combines an Artificial Neural Network (ANN) and a Bayesian Net. By using Gain Ratio (GR) advantage selection approach, the authors used Classification and Regression Tree (CART) to build a binary decision tree by splitting the record at each node. According to a function of a single

attribute, the Artificial Neural Network (ANN) was utilized to mine data for classification. The ensemble approach was used to build a hybrid model to improve classification accuracy. Further, a feature selection approach was used to overcome bias. To reduce the irrelevant features and improve classification accuracy, various classification techniques were applied on NSL-KDD and KDD99 dataset. The proposed model provided accuracy of 99.42% with KDD99 dataset and 98.07% with NSL KDD dataset [29].

Balakrishnan, Venkatalakshmi, & Kannan (2014) studied the intrusion detection system using feature selection and classification technique. The study aimed to provide and employ an intrusion detecting system to deal with possible attacks. The authors adopted various techniques, where the needed data acquired from the KDD'99 cup dataset. A rule based classifier was used to perform effective decision making on intrusions, in addition to a support vector machine method to make binary classification and regression estimation tasks. A proposed algorithm for optimal feature selection was applied through calculating information gain ratio on attribute selection. The proposed feature selection algorithm selected only the important features that helped in reducing the time taken for detecting and classifying the record. Rule based classifier and support vector machine helped achieve a greater accuracy. The provided intrusion detection system reduced the false positive rates and reduces the computation time. This study contributed to the selection of the optimal advantage by calculating the percentage of information gain in characterization. It also helped reduce the time taken to discover and classify the record. Help the base-based workbook and the support transfer machine to achieve greater accuracy. The intrusion detection system reduced false positive rates and reduces the calculation time [30].

Riyad & Ahmed (2013) designed an ensemble classification approach for intrusion detection. The Support Vector Machine (SVM) algorithm was used to maximize the classification by sub dividing feature space into sub spaces and to classify the new data, the Random Forest tree predictor was used to construct the tree with different bootstrap samples, and Artificial Neural Network was used process information. These combined models can increase the accuracy of prediction over a single model. Each classification from the base algorithms is given a weight 0 to 1 depending on their accuracy. The result of the study indicated that the ensemble method is one of the main developments in the field of machine learning [31].

Ozarkar & Patwardhan (2013) implemented an efficient spam classification using Random Forest and Partial Decision Trees algorithms to classify spam vs. non-spam emails. A Chi-square test was used in order to decide whether effects were present or not. Information Gain measure was applied to reduce in entropy caused by partitioning the examples according to a given attribute. In addition, Symmetrical Uncertainty measure was employed to determine desirable property for a measure of feature-feature inter-correlation to have. The authors also used One R algorithm to infer a rule that predicts the class given the values of

the attributes. The study acquired the best percentage accuracy of 99.918% with Random Forest which is 9% better than previous spam base approaches and 96.416%. The results of the study showed that the use of Random Forest and Partial Decision Trees algorithms to classify spam are more effective than other algorithms that have been implemented in terms of accuracy and time complexity [32].

Rathi & Pareek (2013) used data mining to investigate spam mail detection analyzing various data mining approaches on spam dataset to find the best classifier for email classification. Support vector machine was used to analyze data and was mainly used for classification purpose. A Naïve Bayes classifier was used to determine the presence or absence of a particular feature of a class was unrelated to the presence/absence of any other feature, given the class variable depending on the nature of probability model. Moreover, a feature selection method was used to analyze the data, by removing irrelevant and redundant features from the data. The results showed that promising accuracy of the classifier Random Tree is 99.715% with best-first feature selection algorithm and accuracy is 90.93% [33].

4. Research Methodology

4.1. Based Up on, the Main Questions of This Study Are

Q1—How to determine the relevant features used in Online Recruitment Fraud?

Q2—What is the best classification algorithm to be used for Online Recruitment Fraud?

Q3—Is the Ensemble approach suitable for the detecting Online Recruitment Fraud?

The key purpose of this research is to protect individuals and organization protection from compromised privacy and loss of money through constructing a suitable reliable model to detect the fraud exposure in the online recruitment environments. To achieve this objective, the research will apply algorithms to detect this behavior. This research aims to achieve a set of various objectives:

- To utilize the unique existing dataset for online recruitment through preprocessing data EMSCAD to enhance the accuracy of the model.
- To determine the relevant features by applying feature selection techniques which assist to reduce dimensionality.
- To build a reliable model which helps to effectively detect fraud ads with highest accuracy.

This research is an empirical study based on observation, testing and evaluation. Weka tool is used to implement and evaluate the performance the proposed model. The proposed model through following steps to solve the problem of research. The proposed model involves three main stages of scrutiny:

- First stage: Pre-processing stage which conducted by EMSCAD.
- Second stage: Features Selections, where the support vector machine (SVM)

supervised learning was used for feature extraction and exposure importance feature ranking.

- Third Stage: Classification, where the ensemble approaches Random Forest classifier was used. The Random forest classifier is considered by the main and most approach technique used in recent research models and algorithms.

A lot of research work is performed on automatic detection of spam emails using classification techniques such as SVM, NB, MLP, KNN, ID3, J48, Random Tree, etc. but Among all classification technique SVM is the best performer and gives overall best results without using any feature selection techniques, and use of Random Forest, These algorithms outperformed the previously implemented algorithms in terms of accuracy and time complexity.

4.2. The Proposed Online Recruitment Fraud Detection Model

A pre-processing step was required with the data set due to its nature, before being applied to the classifier. However, the dataset's characteristics were determined the suitable feature or variable selection method to use through the model. The core idea of the model is about using ensemble classifier and Support Vector Machine algorithm (SVM) for feature selection in order to distinguish the scams or fraud items in dataset. The proposed model can be conducted in three steps: preparing and processing Dataset, Feature extraction, and classification as shown in **Figure 2**.

The data is represented in an Excel sheet, where each record contains various data genre, structured and unstructured data, and each field contains one of the following four types: String, HTML fragment, Binary, and nominal, the following is detailed descriptive **Table 1** is a detailed description of the dataset's contents.

4.3. Pre-Processing

Pre-processing data is a phase involves the text cleaning function and converting the text in the form suitable to the classification method. It includes extracting noise and uninformative characters and words in the text, such as HTML tags, where such words do not influence the general orientation of text [34].

4.4. Selection Feature

Based on the dataset's nature, as discussed in a previous section, the dataset was labelled dataset. The researcher estimated that the suitable feature selection and extraction method was the supervised learning algorithm; further the target feature was a label in the fraudulent indication in dataset. SVM designed for binary classification was utilized by which items were distributed on specific groups based relevancy. Each group and its prior group were separated by specific margin, maximizing the margin means to minimize the error. The SVM is the hyper plane separating the closest data points groups. The following **Figure 3** is the selection technique of the SVM [35].

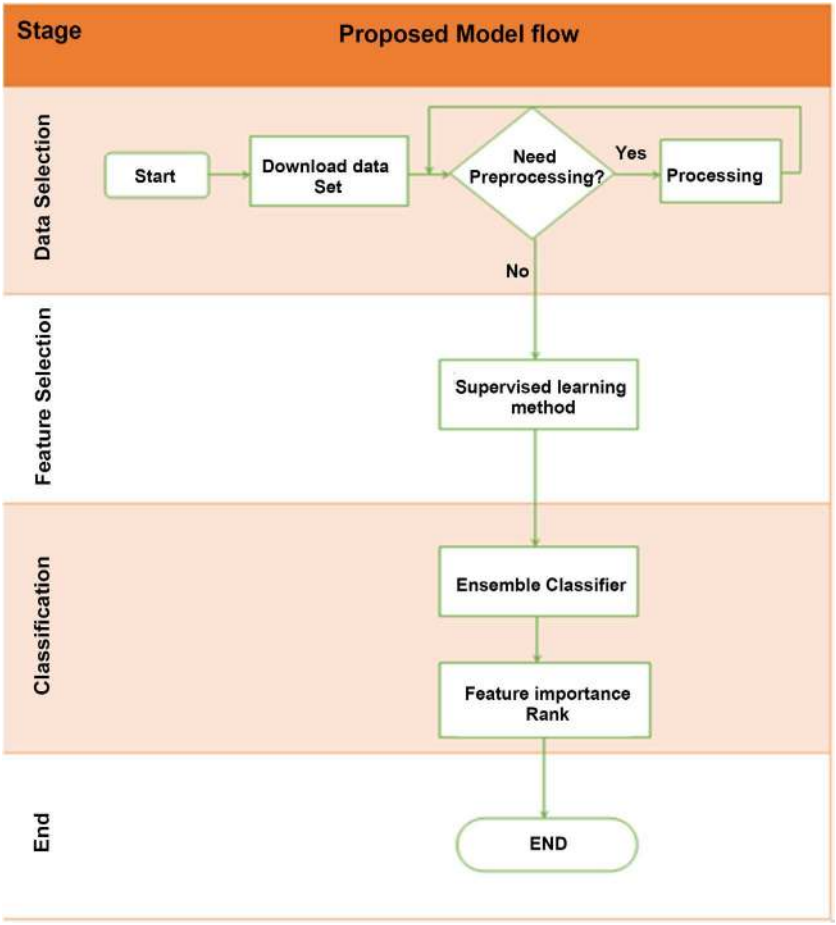

Figure 2. Preparing and processing dataset, feature extraction, and classification.

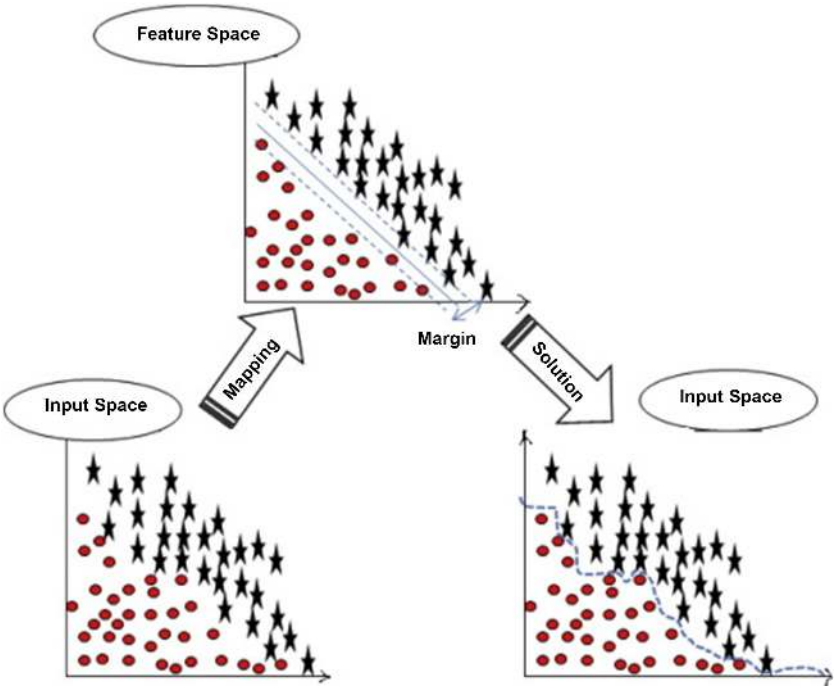

Figure 3. Support vector machine SVM [3].

Table 1. Dataset content description.

Name	Description
String Data	
Tile	The job advertisement header
Location	The location of the job adviser
Department	Job relevant department like sales
Salary range	Suggested Salary Range such as 50,000 - 60,000$
HTML Fragment	
Company Profile	A brief description of the company
Description	Advertised Job details
Requirement	Required list for job
Benefits	Benefits list offered by employer
Binary	
Telecommuting	True for Telecommuting positions
Company Logo	True if company logo exists
Questions	True if screening question exists
Fraudulent	Classification attribute
Nominal	
Employment Type	Full-type, Part-time, Contract, etc.
Required Experience	Executive, Entry level, Intern, etc.
Required Education	Doctorate, Master's Degree, Bachelor's, etc.
Industry	Automotive, IT, Health care, Real estate, etc.
Function	Consulting, Engineering, Research, Sales etc.

4.5. Classification Algorithm

SVM has several important features because of this it is acquire publicity and have hopeful experimental performance. SVM creates a hyper level in authentic input space to divide the data points. Occasionally, it is challenging to implement divide of data points in authentic input space, therefore, to make divide easier the authentic limited dimensional space charted into novel upper dimensional space. Kernel functions are utilized to non-direct charting of teaching examples to great dimensional space (see Figure 3).

The Classification algorithm used was the Random forest ensemble classifier, was constructed based a combination of tree-structured Classifiers. Figure 2 summarizes the concept of ensemble classifier (see Figure 4).

In the procedure of construction Forest, around one-third of the sample was removed because of sampling and replacement executed in accumulation the group of trees. In the following the random forest flow chart represented in Figure 5.

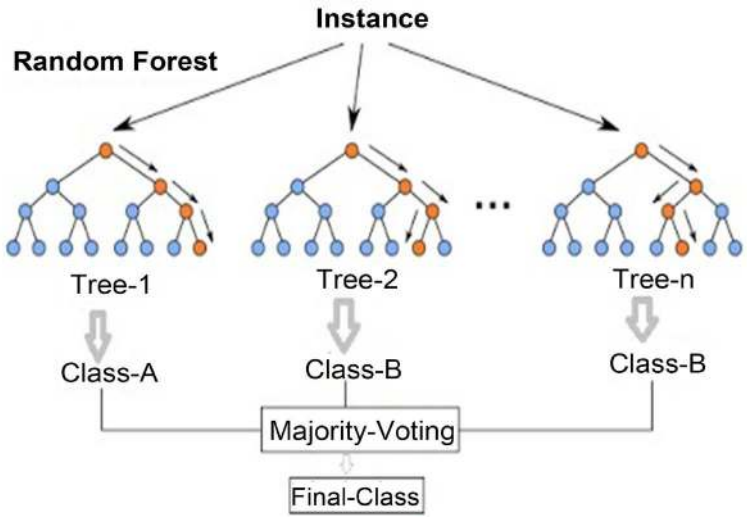

Figure 4. Ensemble classifier-random forest classifier general concept [6].

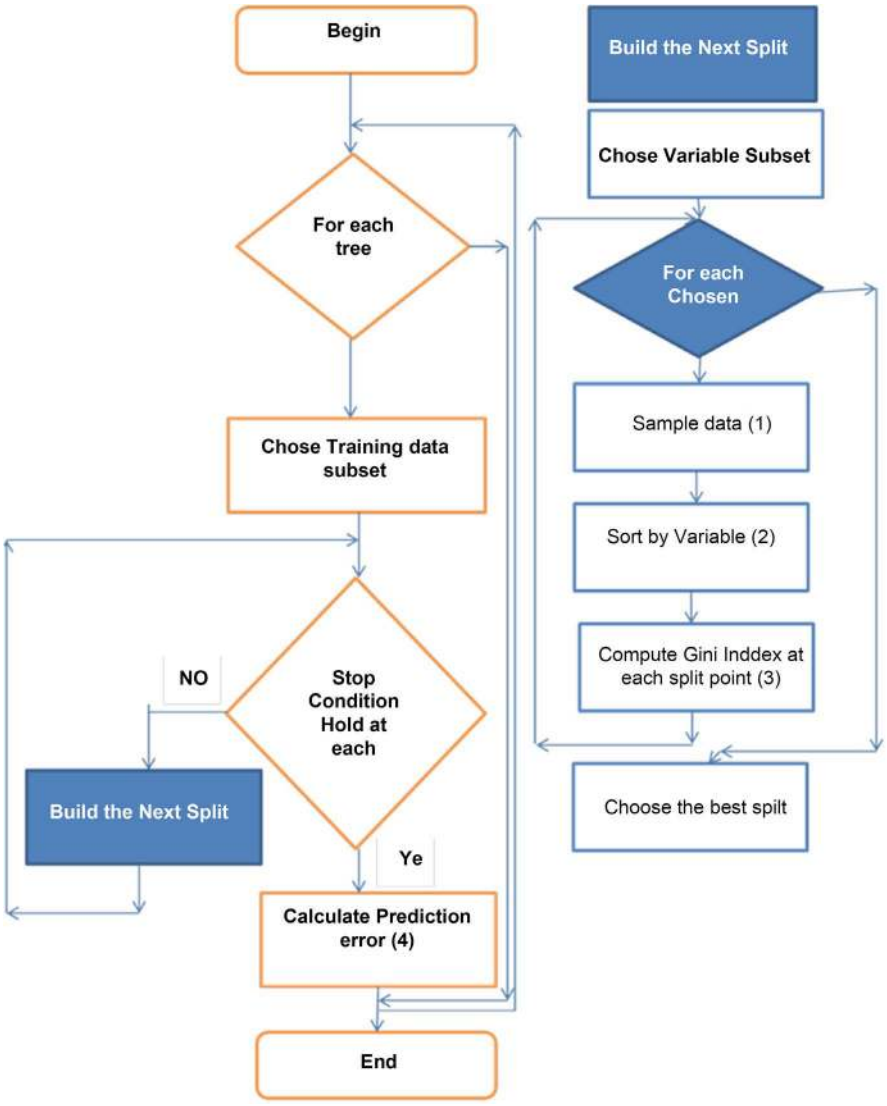

Figure 5. Random forest flow chart.

The pseudo code for the Random Forest Classifier was written as follows (see Figure 6).

```
Procedure Random Forest
For I to T do
Draw n points D1 with replacement from D
Build Full decision/regression tree on D1
But: each split only considers K feature, Pick uniformly at
random
New feature for every split
Prune tree to minimize out-of-bag error
End for
Average all trees
End procedure
```

4.6. Evaluation Parameter

The evaluation of classier considers the fault notion; the fault can have induced due to wrong prediction when compared to the actual selected case. It total number is important to observe the total number of faults to assess the classifiers works [36]. The evaluation parameters represented on the following [37]:

1) Accuracy is the quality measurement for classifier, which can be calculated through mathematically finding the ratio of correct classified attempts number according to the total number of all classification attempts.

$$\text{accuricy} = (TP + TN)/(TP + TN + FP)$$

2) Precision refers to classifier exactness indications, where the high precision means a large number of true positive, in some trade-off between accuracy and precision, low accuracy accepted if the precision is high.

$$\text{Precision} = TP/(TP + FP)$$

3) Recall is another evaluation parameter for classifier, which reflects the classifier competence, where high recall indicates high true positives; recall is mathematical represented as the amount of true positives over the summation of both true positives and false negatives.

$$\text{Recall} = TP/(TP + FN)$$

4) ROC is an abbreviation of the receiving operating characteristics, ROC is another evaluation parameter considered for classier performance. ROC graph is a graph plotting true positives rates which indicates the sensitivity of detection, corresponding to the false positive rate.

```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 96
        Merit of best subset found:    0.098

Attribute Subset Evaluator (supervised, Class (nominal): 17 fraudulent):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 5,10,15 : 3
                     company_profile
                     has_company_logo
                     industry
```

Figure 6. _best first filter attribute extraction.

5. Results and Discussion

The device used for whole procedure is Acer laptop, Core i3 processor, with 3GB RAM, 32-bit operating system, and windows 7 operating system. Weka version 3.8.2 was also used.

5.1. Pre-Processing Data Set

The dataset utilized in this research work was EMSCAD, which is available online. The downloaded dataset file was in CSV file format with all the data inserted in one column. Utilizing MS-Excel potentials, the data were divided into an individual column. The dataset comprises 17,860 rows, each row indicates one online recruitment advertisement and each column including specific relevant information from aforementioned Table 2. The dataset combines various data types such as String, HTML fragment, Binary, and nominal. The following is a small portion of one item in the dataset before reprocessing.

There are a lot of missing values in the dataset. The missing values were filled automatically using the mathematical and logical potentials in Excel to comply with the Weka environment's conditions. The researcher used IF statements to fill all missing value in the dataset file. The value "None" was used to fill the blank cells in the department category, the value "Not specified" was used to fill the blank cells in the employment type category and the value "No Information" was used to fill the blank cells in the required experience, required education, Industry, and function categories. After completing the missing values, specific text mining tools were required to handle the string data type. To avoid going through complex procedures and simplify the works, this data type was converted to binary values based on its contents. This step also included using IF statements in MS-Excel, and spaces and special characters were deleted from the data. Another issue rose in pre processing, the dataset was that some categories,

Table 2. Specific relevant information from aforementioned.

location	department	salary_range	company_ profile	telecommuting	has_company_ logo	has_ questions	Employment _type
US, NY, New York	Marketing	We're Food 52, and we've created a groundbreaking and award-winning cooking site. We support, connect, and celebrate home cooks, and give them everything they need in one place. We have a top editorial, business, and engineering team. We're focused on using technology to find new and better ways to connect people around their specific food interests, and to offer them superb, highly curated information about food and cooking. We attract the most talented home cooks and contributors in the country; we also publish well-known professionals like Mario Batali, Gwyneth Paltrow, and Danny Meyer. And we have partnerships with Whole Foods Market and Random House.	f	t	f	Other	

such as location, department, employment type, required experience, required education, industry and function, were non-numeric data types, and SVM does not handle non-numeric values. Therefore, they had to be converted into numeric values. The researcher used MATLAB potentials to convert the categories' values to numeric values for this purpose.

The previous version of the dataset was saved in CSV file. Table 3 below is a screen of previous version of dataset.

5.2. Selection Features

The algorithm used for extracting the main features was the SVM attributes selection, which is efficient with labelled dataset. To conduct this phase, the researcher used Weka tools. First, a Weka tool was used to convert the last version dataset file format from.CSV to ARFF file, to import it successfully in Weka. After importing the processed dataset on Weka and choosing SVM as attribute selection, specifically sixteen features were extracted: Title, Location, Department, Salary range, Company profile, Description, Requirements, Benefits, Telecommuting, Has company logo, Has questions, Employment type, Required experience, Required education, Industry, and Function.

To minimize the attribute list, the researcher used attribute _best first filter, the filter which extracted three main features: company profile, has company logo and in Industry.

5.3. Ensemble Classification

After extracting the features and minimizing them, the ensemble classification was applied in order to apply the ensemble classification, the researcher utilized Weka tool library and choose the Random Forest classifier, the training set and testing set can be defined by specifying the ratio of data, which set as 66% training, and the rest is testing set. The training data set contained 11,788 items, while the test dataset contained 6072 items. The cross-validation value, which specifies the statistical performance of learning and the accuracy of predicting

Table 3. Processed dataset screen.

location	salary_range	company_profile	description	has_company_logo	has questions	employment_type
1	0	1	1	1	0	1
1	0	1	1	1	0	2
1	0	1	1	1	0	3
1	0	1	1	1	0	2
1	0	1	1	1	1	2
1	0	0	1	0	0	3
1	1	1	1	1	1	2
1	0	1	1	1	1	3
1	0	1	1	1	1	2

model, was set to be 10 folds. The model achieved 97.41% accuracy rate as shown in Figure 7 below.

The corresponding confusion matrix is shown in Figure 8 below.

The precision and recall and ROC values were also calculated by Weka. The classification performance was evaluated based on the aforementioned parameters. Precision refers to classifier exactness indications, where the high precision means a large number of true positive. There are some trade-offs between accuracy and precision; for example, low accuracy can be accepted if the precision is high. In the current case, the precision achieved is 97.2% which consider high and adequate. Recall is another evaluation parameter for classifier that reflects the classifiers' competence; high recall indicates high true positives. Recall is mathematically represented as the number of true positives over the summation of both true positives and false negatives, in the current case, the recall was around 97.4%. Figure 9 shows the corresponding value of evaluation parameters.

Receiving Operating Characteristics (ROC) is another evaluation parameter considered for classier performance. ROC graphs plot a true positives rate, which indicates the sensitivity of detection, corresponding to the false positive rate. The ROC graph, created with visualisation potentials in Weka, is shown in Figure 10 below.

It is very important to mention that the cross-validation number has impacts on the precision, recall, ROC and accuracy of classifier, in the current case the cross-validation obviously affected the ROC value with around 0.4 differences. The accuracy difference between two folds' attempts around 0.2, as the comparison is shown in Table 4.

Feature importance ranking, is a ranking technique available through Weka that ranks attributes by their individual evaluations. The ranks were determined via two cross validations; the first attempt used 5 folds, and the second used 10 folds. In this case, the threshold set as −1.7976931348623157E308 the feature can

```
Time taken to build model: 15.74 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       17399               97.4188 %
Incorrectly Classified Instances       461                2.5812 %
Kappa statistic                          0.6578
Mean absolute error                      0.043
Root mean squared error                  0.1444
Relative absolute error                 46.6265 %
Root relative squared error             67.2497 %
Total Number of Instances            17860
```

Figure 7. Classification accuracy summary.

		Predicted	
		Legitmate (F)	Fraud (P)
Actual	Legitmate(F)	16930 True Positve	64 False Negaive
	Fruad(T)	397 Fasle Positive	469 True Negative

Figure 8. Confusion matrix for random forest classification.

```
Total Number of Instances            17860

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.996    0.458    0.977      0.996   0.987      0.679    0.958     0.997     F
               0.542    0.004    0.880      0.542   0.670      0.679    0.958     0.763     T
Weighted Avg.  0.974    0.436    0.972      0.974   0.971      0.679    0.958     0.985
```

Figure 9. Precision, recall, F-measure and ROC values.

be discarded based on the ranks determined by the two cross validations. The 5 cross validation folds ranking is shown in Table 5.

The 10 folds across validation optimized the first ranking results and established new ranking, as shown in Table 6 below. This ranking was determined the suitable associating with attribute _best first filter result.

The extracted attributes matched the attributes extracted in Vidros, Kolias, Kambourakis, & Akoglu (2017) experiment. Further, the current research showed that the company profile, and has company logo had the higher attributes ranking, which aligns with the correlation attributes measured by Vidros, *et al.* (2017) experiments, their study showed that company profile , and company logo is the highest correlation attributes. The current experiment

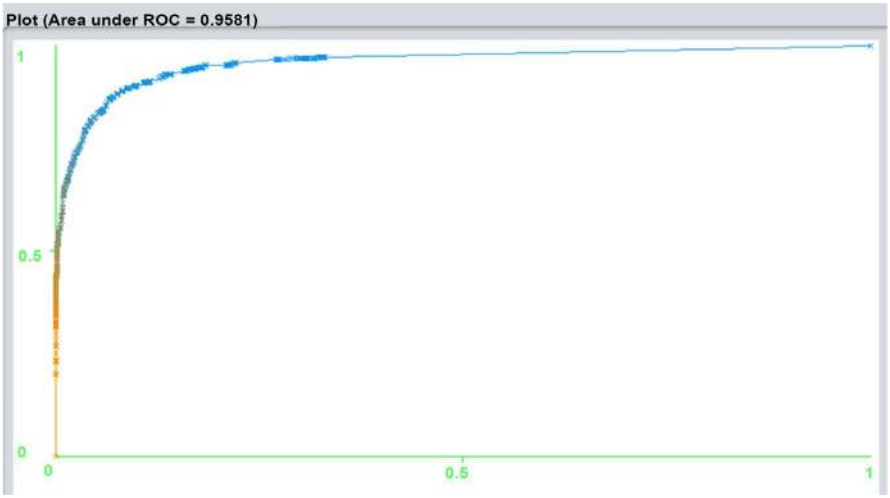

Figure 10. ROC Curve for ensemble classifier.

Table 4. Cross-validation impacts on the classifier performance.

Parameters	Accuracy	Recall	Precision	ROC
Cross-validation 5	97.2	97.2	97.0	95.4
Cross-validation 10	97.4	97.4	97.2	95.8

Table 5. Attributes rank (cross validation 5).

Average rank	Attribute
1 ± 0	5 company profile
2± 0	10 has company loge
3 ± 0	15 industry
4 ± 0	16 function
5 ± 0	14 required education
6 ± 0	11 has questions
7 ± 0	13 required experience
8 ± 0	12 employment type
9 ± 0	4 salary range
10 ± 0	9 telecommuting
11.8 ± 0.98	2 location
12 ± 0	3 department
13.4 ± 0.96	7 requirements
13.4 ± 0.49	8 benefits
14.4 + 0.49	6 description
16 + 0	1 title

achieved accuracy of 97.4%, which outperformed the highest accuracy value obtained by Vidros *et al.* (2017), around 91.22%. This is summarized in Table 7 below, which indicates the research approach and model enhance the detection performance of the online recruitment fraud [2].

Table 6. Attributes rank (cross validation 10).

Average rank	Attribute
1 ± 0	5 company profile
2 ± 0	10 has company loge
3 ± 0	15 industry
4 ± 0	16 function
5 ± 0	14 required education
6 ± 0	11 has questions
7 ± 0	13 required experience
8 ± 0	12 employment type
9 ± 0	4 salary range
10 ± 0	9 telecommuting
11.8 ± 0.98	2 location
12 ± 0	3 department
13.4 ± 0.96	7 requirements
13.4 ± 0.49	8 benefits
14.4 + 0.49	6 description
16 + 0	1 title

Table 7. Result comparison of current research and Vidros *et al.* research.

Experiments	Algorithms	Accuracy
Vidros *et al.* 2017 experiment	ZeroR	50
	Logistic regression	77.22
	OneR	77.33
	J48	84.77
	Naive Bayes	86.33
	Random forest	91.22
Current Experiment	Random forest	97.41

6. Conclusion and Future Work

Online recruitment systems are a promising platform that many companies and enterprises depend on in their recruitment and hiring process. According to Saudi 2030 vision, the percentage of adoption and utilization of online recruitment system will increase gradually. This prediction sheds lights on these systems. Nevertheless, online recruitment systems are abused by criminals conducting scams. Despite this abuse, only two research studies have been done on this topic. This first was a descriptive study that reviewed the fraud and scam types detected in the online recruitment systems, and the second proposed a detection model. The current research proposes detection model for online recruitment fraud.

The proposed model used two data mining algorithms, Support Vector Machine (SVM) for feature selections and Random Forest as ensemble classifier. The Model achieved 97.41% accuracy, which is the best accuracy result that has been obtained until now.

This research emphasized that the online recruitment contains similarities of previously well-studied scopes such as email spam phishing, cyber bullying and so forth. Furthermore, this research utilized the EMSCAD dataset which is the only free available dataset for this scope. The main contributions of this research are that it is only the second experimental research in this scope and that it enhances fraud detection classifiers for online recruitment systems. The work's limitation can be summarized in scarcity of various datasets, the complexity and variety of the data types in the datasets, and that the data set contained only English language online advertisements.

Future Work could examine various different data mining algorithms. Also, it could utilize the available EMSCAD dataset to analyze the company profile, company logo, and required experience as main string attributes for detection purposes. Further, future research could conduct several data mining techniques to create a new dataset of online recruitment advertisement in English language, or to create an Arabic dataset of online recruitment advertisements for detection purposes.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Scanlon, J.R. and Gerber, M.S. (2014) Automatic Detection of Cyber-Recruitment by Violent Extremists. *Security Informatics*, **3**, 5. https://doi.org/10.1186/s13388-014-0005-5

[2] Vidros, S., Kolias, C., Kambourakis, G. and Akoglu, L. (2017) Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset. *Future Internet*, **9**, 6. https://doi.org/10.3390/fi9010006

[3] Novaković, J.D., Veljovic, A., Ilić, S.S., Papic, Z. and Tomovic, M. (2017) Evaluation of Classification Models in Machine Learning. *Theory and Applications of Mathematics & Computer Science*, **7**, 39-46.

[4] Morgan, S. and Menlo Park, C. (2017) Cybercrime Report from the Editors at Cybersecurity Ventures. Herjavec Group, Toronto.

[5] Gov, K. (2018) Vision of 2030. http://vision2030.gov.sa/en

[6] ACRON (2018) Australian Cybercrime Online Reporting Network (ACORN). http://www.acorn.gov.au/learn-about-cybercrime

[7] Armstrong, A. (2006) Handbook of Human Resource Management Practice. 10th Edition, Kogan Page Limited, London.

[8] Hada, B. and Gairola, S. (2015) Opportunities and Challenges of E-Recruitment. *Journal of Management Engineering and Information Technology*, **2**, 1-4.

[9] Kaur, P. (2015) E-Recruitment: A Conceptual Study. *International Journal of Ap-*

plied Research, **1**, 78-82.

[10] Prasad, L. and Kapoor, P. (2016) Topic: E-Recruitment Strategies. *International Journal of Business Quantitative Economic and Applied Management Research*, **2**, 80-95.

[11] Panov, P., Soldatova, L. and Džeroski, S. (2013) OntoDM-KDD: Ontology for Representing the Knowledge Discovery Process. 16*th International Conference on Discovery Science*, Singapore, 6-9 October 2013, 126-140. https://doi.org/10.1007/978-3-642-40897-7_9

[12] Cios, K.J., Pedrycz, W., Swiniarski, R.W. and Kurgan, L.A. (2007) Data Mining: A Knowledge Discovery Approach. Springer, New York.

[13] Hussain, S. (2017) Survey on Current Trends and Techniques of Data Mining Research. *London Journal of Research in Computer Science and Technology*, **17**, 7-15.

[14] Sinoara, R., Antunes, J. and Rezende, S. (2017) Text Mining and Semantics: A Systematic Mapping Study. *Journal of the Brazilian Computer Society*, **23**, 9. https://doi.org/10.1186/s13173-017-0058-7

[15] Diwathe, D. and Dongare, S. (2017) Classification Model Using Optimization Technique: A Review. *International Journal of Computer Science and Network*, **6**, 42-48.

[16] Singh, G. and Singh, A. (2017) A Review Paper: Using Data Mining Clustering Technique to Predict Criminal Behavior. *International Journal of Computer Science and Mobile Computing*, **6**, 160-167.

[17] Witten, I. and Frank, E. (2005) Data Mining Practical Machine Tools and Techniques. Morgan Kaufmann Elsevier, San Francisco.

[18] Kukasvadiya, M. and Divecha, N. (2017) Analysis of Data Using Data Mining Tool Orange. *International Journal of Engineering Development and Research*, **5**, 836-1840.

[19] Rehman, N. (2017) Data Mining Techniques Methods Algorithms and Tools. *International Journal of Computer Science and Mobile Computing*, **6**, 227-231.

[20] Jyoth, P., Siva Ranjani, R., Mishra, T. and Mishra, S.R. (2017) A Study of Classification Techniques of Data Mining Techniques in Health Related Research. *International Journal of Innovative Research in Computer and Communication Engineering*, **5**, 13779-137876.

[21] Vidros, S., Kolias, C. and Kambourakis, G. (2016) Feature: Online Recruitment Services: Another Playground for Fraudsters. *Computer Fraud & Security*, **2016**, 8-13. https://doi.org/10.1016/S1361-3723(16)30025-2

[22] Yasin, A. and Abuhasan, A. (2016) An Intelligent Classification Model for Phishing Email Detection. *International Journal of Network Security & Its Applications*, **8**, 55-72. https://doi.org/10.5121/ijnsa.2016.8405

[23] Al-garadi, M.A., Varathan, K.D. and Ravana, S.D. (2016) Cybercrime Detection in Online Communications: The Experimental Case of Cyberbullying Detection in the Twitter Network. *Computers in Human Behavior*, **63**, 433-443. https://doi.org/10.1016/j.chb.2016.05.051

[24] Sharaff, A., Nagwani, N.K. and Swami, K. (2015) Impact of Feature Selection Technique on Email Classification. *International Journal of Knowledge Engineering,* **1**, 59-63. https://doi.org/10.7763/IJKE.2015.V1.10

[25] Sornsuwit, P. and Jaiyen, S. (2015) Intrusion Detection Model Based on Ensemble Learning for U2r and R2l Attacks. In: 7*th International Conference Information Technology and Electrical Engineering*, IEEE, Chiang Mai, 354-359. https://doi.org/10.1109/ICITEED.2015.7408971

[26] Gaikwad, D.P. and Thool, R.C. (2015) Intrusion Detection System Using Bagging Ensemble Method of Machine Learning. *Computing Communication Control and Automation*, Pune, 26-27 February 2015, 291-295. https://doi.org/10.1109/ICCUBEA.2015.61

[27] Zuhaira, H., Selmat, A. and Salleh, M. (2015) The Effect of Feature Selection on Phish Website Detection: An Empirical Study on Robust Feature Subset Selection for Effective Classification. *International Journal of Advanced Computer Science & Applications*, **1**, 221-232. https://doi.org/10.14569/IJACSA.2015.061031

[28] Nizamani, S., Memon, N., Glasdam, M. and Nguyen, D.D. (2014) Detection of Fraudulent Emails by Employing Advanced Feature Abundance. *Egyptian Informatics Journal*, **15**, 169-174. https://doi.org/10.1016/j.eij.2014.07.002

[29] Shrivas, A.K. and Dewangan, A.K. (2014) An Ensemble Model for Classification of Attacks with Feature Selection Based on KDD99 and NSL-KDD Data Set. *International Journal of Computer Applications*, **99**, 8-13. https://doi.org/10.5120/17447-5392

[30] Balakrishnan, S., Venkatalakshmi, K. and Kannan, A. (2014) Intrusion Detection System Using Feature Selection and Classification Technique. *Journal of Computer Science and Application*, **3**, 145-151. https://doi.org/10.14355/ijcsa.2014.0304.02

[31] Riyad, A.M. and Ahmed, M.I. (2013) An Ensemble Classification Approach for Intrusion Detection. *International Journal of Computer Applications*, **80**, 37-42. https://doi.org/10.5120/13836-1402

[32] Ozarkar, P. and Patwardhan, M. (2013) Efficient Spam Classification by Appropriate Feature Selection. *Global Journal of Computer Science and Technology*, **13**, 49-57.

[33] Rathi, M. and Pareek, V. (2013) Spam Mail Detection through Data Mining—A Comparative Performance Analysis. *International Journal of Modern Education and Computer Science*, **5**, 31-39. https://doi.org/10.5815/ijmecs.2013.12.05

[34] Haddi, E., Liu, X. and Shi, Y. (2013) The Role of Text Pre-Processing in Sentiment Analysis. *Procedia Computer Science*, **17**, 26-32. https://doi.org/10.1016/j.procs.2013.05.005

[35] Tomar, D. and Agarwal, S. (2015) Twin Support Vector Machine: A Review from 2007 to 2014. *Egyptian Informatics Journal*, **16**, 55-69. https://doi.org/10.1016/j.eij.2014.12.003

[36] Novakovic, J.D., Veljovic, A., Ilic, S.S., Papic, Z. and Tomovic, M. (2017) Evaluation of Classification Models in Machine Learning. *Theory and Applications of Mathematics & Computer Science*, **7**, 39-46.

[37] Dharamkar, B. and Singh, R.R. (2014) Cyber-Attack Classification Using Improved Ensemble Technique Based on Support Vector Machine and Neural Network. *International Journal of Computer Applications*, **103**, 7.