International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 12– No. 39, April 2022 – www.ijais.org

A Hybrid Model for Classification of E-mail Fraud

T.O. Oyegoke
Department of Computer Science and Engineering,
Obafemi Awolowo University, Ile-Ife. Nigeria

ABSTRACT
The study pre-processed e-mail data, formulated and validated a Particle Swarm Optimization (PSO)-based Back Propagation model for email fraud detection. This was done by the hybridization of two algorithms namely; Nature Inspired Algorithm and Artificial Neural Network. The dataset collected for the purpose of developing the model contained fraudulent mails (46.3%), Spam (32.6%) and Ham (21.1%) e-mails. 12,831 features were extracted after data preparation and cleaning, in which only 6,382 (49.7%) relevant features were selected using PSO. The model was simulated using 70% and 80% for training while 30% and 20% of datasets were used for testing respectively. The results of using the 30% and 20% testing dataset for the gradient-based BP algorithm showed that using the relevant features selected by PSO improved the accuracy by a value of 0.27% and 0.35% respectively while for the PSO-based BP algorithm, using the relevant features selected by PSO improved the accuracy by a value of 1.51% and 1.46% respectively. The results showed that using PSO-based BP had a better performance than gradient-based BP by a value of 1.48% and 2.72% for 30% training dataset and a value of 1.46% and 2.57% using the original features and the features selected using PSO respectively. The study concluded that the PSO-based BP algorithm was able to improve the performance of the Multi-Layer Perceptron compared to the Gradient-Based Back Propagation algorithm which has implications on improving advance fee fraud detection.

Keywords
Email, Machine Learning, Advance Fee Fraud; Fraud detection, Artificial Neural Network (ANN), Particle Swarm Optimization (PSO)

1. INTRODUCTION
Today, criminals have sought the use of computing devices in perpetuating fraudulent activities which has turned out very lucrative for them. The increasing rate of adoption of these devices is much higher than the rate of development of innovative strategies for providing defense mechanisms by cyber forensic analysts and researchers. Fraudulent activities have led to large financial losses incurred by companies and individuals (the victims) has motivated various research towards the development of various defense mechanisms required for tackling fraud thus preventing and detecting the onset of fraudulent activities. The rapid growth of the Internet which has led to a significant increase in the number of email users at the same time has also led to an increase in spam emails rate and in general, electronic frauds by fraudsters [1]. [2] indicated in 2009 that e-mail spam had increased by 30% over 2007 amounting to loses of $130 billion worldwide. Additional problems caused by spam mails include waste of network traffic and storage space alongside the computational process involving the spam receivers which is however very irritative and a violation of human rights according to many victims [3]. Electronic fraud is any type of activity which is conducted using computing devices such as e-mails, websites, instant messaging or social networks with the purpose of luring unsuspecting victims into financial loss [4]. Fraud detection uses background server-based processes that examine users' and other defined entities' access and behavior patterns, and typically compares this information to a profile of what is expected. The failure is as a result of the inability to deal with some unique various challenging properties of fraudulent activities which include: experience imbalance, online learning, adaptive adversaries, concept drift, noise, unequal misclassification costs and fast processing and large volume of data [5]. Various researches have adopted artificial intelligence and statistics to the development of fraud detection however many of which have a centralized control leading to difficulty in understanding the changing dynamics of fraudulent environments [6]. Fraud detection is not intrusive to a user unless the user's activity is suspected, however it tries to detect and recognize fraudulent activities entering the systems in order to report them to the system manager [7]. According to [8], the popularity of statistical methods in fraud detection was provided in a comprehensive survey of statistical approaches for fraud detection which described the tools available for statistical fraud detection. Among the mentioned techniques, most of them fail to deal with changing environments [9]. On the other hand, there are also other artificial intelligence algorithms called nature inspired techniques which were inspired by the way natural systems work in order to solve complex problems and are part of the class of computational intelligence algorithms [10, 11]. This is because natural systems, such as an animal's organ or groups of animals which seem to behave randomly and imprecise, are usually most times robust and are resilient enough to deal with redundancy and incomplete information, for instance, the flocking movements of birds in search for food motivated the development of Particle Swarm Optimization (PSO) algorithms [12] or the adoption of evolutionary algorithms which were motivated by the Darwinian Law of Natural Evolution [13]. The versatility of nature-inspired algorithms makes them effective at solving challenging classification and optimization problems [14]. Nature-inspired algorithms require no assumptions about the problem area with a capacity to provide a wider search space of feasible solutions which is another important reason for adoption in solving complex problem thus enabling them to be able to solve problems without knowledge of the problem domain or underlying factors [15]. These algorithms can also handle optimizations problems effectively in the presence of noise and other uncertainties thus making them adaptive unlike many artificial intelligence methods which have difficulty in dealing with changing and dynamic environments using the population of diverse individuals [15,16].

The existence of loose and flexible connections which exist between parts made system dynamism possible thereby leading to multiple response in the presence of diverse tasks [10]. This adaptability which is possessed by natural systems has lured scientist to develop computational methods motivated by nature

and natural systems. Also, the advantage of nature inspired algorithms over other artificial intelligence methods resides in their ability to adopt the use of diverse populations with specific specialized tasks with the purpose of handling changes within their environment. These algorithms require no assumptions about the problem area with a capacity to provide a wider search space of feasible solutions which is another important reason for adoption in solving complex problem thus enabling them to be able to solve problems without knowledge of the problem domain or underlying factors [14]. Each individual on the other hand is good at something, flexible and capable of responding to changes in environment. The application of nature-inspired algorithms to the detection of fraudulent e-mails can assist in the identification of the most relevant e-mail features which can maximize the detection of fraudulent mails from illegitimate (or spam) mails thereby influencing real-time fraud incident reporting to possible victims of electronic fraud.

A number of related works which have applied the Artificial Neural Network (ANN) using the gradient-based Back Propagation (BP) algorithm to perform the training task required for the development of fraud detection models have been identified to show a number of drawbacks. Some of these drawbacks include slow training convergence speed and the likelihood of the solution getting stuck in a local minimum easily instead of searching for global solutions [11]. The use of meta-heuristic algorithms such as nature-inspired algorithms has been identified in recent research to overcome a number of these drawbacks, some of which include inability to adapt to changing features in e-mail writing styles, limited number of identified e-mail features, incapacity to respond to false alarms, no feedback mechanisms in place, high fraud misclassification rates.

The detection of fraudulent e-mails provides a means of safeguarding the interests of unsuspecting online users by providing a real-time and early detection of fraudulent e-mails on arrival. The availability of meta-heuristic algorithms such as nature-inspired algorithms have provided a means by which optimization techniques can be used to identify the best solution of the identified problem also called global solutions. The identification of global solution in turn provide a means of identifying the variables (or e-mails features) that fall under such solution space and also required by the model for providing the best solution in the presence of a candidate set of many local solutions.

This study is motivated by the use of two theories, namely: the theory of Group Dynamics by Max Wertheimer and the Dempster-Shafer Theory of evidence by Arthur Dempster and Glenn Shafer. The theory of group dynamics states that there are entities where the behaviour of the whole cannot be derived from its individual elements nor from the way these elements fit together; rather the opposite is true: the properties of any of the parts are determined by the intrinsic structural laws of the whole [17,18]. The theory of group dynamics enforces that, information about the whole group is sufficient to describe the members of the group rather than concluding the whole from the collective individual qualities of its members.

Group theory relates to this study by considering the problem of fraud detection as composing of three groups: fraudulent and non-fraudulent e-mails which are either spam or ham mails. According to this theory, each swarm of birds can be likened to the different classes of e-mails namely: Fraudulent, Spam and Ham mails. The birds that find themselves within each swarm are members of that group and thus can be likened to the features that are found in each classes of e-mails identified. However, the characteristics of the swarm to which a bird belongs cannot be deduced from the collective characteristics of the birds that belong to the swarm. Hence, the ability to classify an e-mail to a class of either Spam, Ham or Fraudulent mail is a function of the understanding of the characteristics of the features extracted from a group of e-mail classified as either Spam, Ham or Fraudulent. Therefore, the identification of characteristics of these features can be used to determine to which a class an e-mail belongs to by assessing the presence and (or absence) of these features.

The theory of Evidence was postulated by Arthur P. Dempster in the context of statistical inference which was extended by Glenn Shafer for modeling evidence [19,20]. The theory allows one to combine evidence from different sources and arrive at a degree that takes into account all the available evidence thus avoiding conflicts. Based on the theory of Evidence, suffice it to say that evidence of the occurrence of a fraudulent activity cannot be derived by inferring facts from a single fraudulent e-mail but from the identification of fraudulent and non-fraudulent features found in as many related e-mails (evidence) as possible. Therefore, each e-mail type identified as either fraudulent, spam or ham serves as the source of evidence while the features that are common to the different types of identified e-mails are believed to be the evidence required for the identification of fraudulent e-mails.

The identification of the most relevant features within e-mails also has implication in reducing the computational space and computational complexity of fraud detection systems which in turn reduces the processing time of such systems.

2. METHODOLOGY

This section presents the materials and methods that were deployed for the development of the classification model required for the detection of fraudulent e-mails. The materials and methods used for data identification and collection, model formulation and simulation alongside performance evaluation were also presented.

2.1 Method of Data Identification and Collection

Data containing fraudulent and non-fraudulent e-mails were extracted from the CLAIR collection of fraudulent e-mails (unstructured data) located at https://www.kaggle.com/rtatman/fraudulent-e-mail-corpus and UCI Spambase structured repository which is a collection of spam and ham e-mails, located at https://archive.ics.uci.edu/ml/datasets/spambase. This study adopted the Spambase datasets and the CLAIR dataset which consists of e-mails that have been already classified according to their respective classes. The Spambase dataset contained 4601 pre-processed datasets consisting of 1813 spam mails (39.4%) and 2788 ham mails (60.0%). Each dataset record is composed of 57 continuous features and a target class named 1 for spam and 0 for ham. Therefore, there were three (3) classes for identifying the e-mails collected from the two (2) data sources, namely: spam mail (suspicious but not necessarily fraudulent e-mails), ham mails (non-fraudulent e-mails) and fraudulent e-mails. The dataset that was used for this study eventually consisted of a total of 8575 e-mail samples such that: 46.3% consists of fraudulent e-mails, 21.1% consists of ham e-mails while the remaining 32.6% consists of spam e-mails as shown in Table 1.

Table 1: Dataset Source and Size

S/N	Data Source	E-mail Type	Frequency	Percentage (%)
1.	Spambase	Ham mail	2788	32.6
		Spam Mail	1813	21.1
2.	CLAIR	Fraudulent mail	3974	46.3
	Total		8575	100.0

2.2 Method of E-Mail Text Preprocessing

For the purpose of the development of the fraud detection model for e-mails, there was the need of converting the unstructured contents within the CLAIR e-mails collected into a structured format similar to the Spambase features (but not necessarily the same features). The features extracted are the words that were found within the body of the CLAIR e-mail contents. The text preprocessing of e-mails required the use of the Python® Natural Language Tool-Kit (NLTK) for the purpose of performing the different text preprocessing stages required for converting the unstructured data into a structured format.

Using the Python® NLTK, the unstructured CLAIR e-mails was tokenized in order to convert every content of the e-mails into sets of words found within each e-mail which was followed by the removal of stop words from the extracted e-mail contents. The stemming process was applied to the extracted contents in order to convert all words into their root word (for example, families, familiar becomes famili and so on). The Porter's Algorithms was adopted for the purpose of performing the stemming of the words extracted from the e-mails. The application of the stemming algorithm reduced the feature space of the words in each document to their root words following which frequency of occurrence of the words found in each document is taken into account.

2.3 Method of Text Classification

In order to convert the preprocessed words in the e-mails into a structured form required for fraud detection, the words identified from both e-mail samples (CLAIR and Spambase) was used to form a term-document matrix, $\mathcal{D}_{ij}$ which represents the occurrence of each term w_i. within each document d_j. In the term-document matrix $\mathcal{D}_{ij}$ the rows represent the occurrence (or absence) of a word w_i in a document. The value is 0 when the word does not occur in a document and the value if greater than 0 based on the technique of representation to be adopted. For this study, the Boolean model was adopted. Using the Boolean model, weight $w_{ij} > 0$ is assigned to each term $w_i \epsilon d_j$ while for any term that does not appear in d_j then, $w_{ij} = 0$.

Following the process of creating the term-document for all 8575 e-mail samples collected, the class of each e-mail was used to map each row of the term-document matrix. Therefore, the term-document matrix consisting of 8575 rows and n columns (features/words/terms from e-mails) was mapped to a fraudulent classification column vector consisting of 8575 values of either Ham, Spam or Fraudulent by a function f. This relationship is presented by the function shown in equation 1.

$$f: \mathcal{D} \rightarrow \mathcal{F} \quad (1)$$

Defined as: $f(w_1, w_2, \dots\dots\dots, w_n) = \begin{cases} Ham \\ Spam \\ Fraudulent \end{cases}$

The final product of the process performed yielded the structured form of the dataset required for the development of the fraud detection model required for this study. The dataset was also subjected to feature selection in order to extract the most relevant features out of the final features in the vector space model developed for this study using the term-document matrix. This process was expected to improve the classification performance of the fraud detection model by providing a classification model for fraud detection with a lower computational complexity, time complexity and model complexity.

2.4 Method of Extraction of Relevant Features

Particle Swarm Optimization (PSO) mimics the flocking behavior of birds (which represents the features extracted from the e-mails) whenever they are in search for food. The birds fly in a solution space and their flocking behavior determines the optimum solution (class of e-mail be it spam, ham or fraud) of the e-mails classification problem. Particles tend to move towards its local best position (solution) (pBest) found by them thus keeping track of the global best (gBest) solution, the best (shortest) path found at any instance. Birds communicate with each other to find the most optimum (best) path to reach its food sources.

In the PSO algorithm, features extracted from the e-mails are the swarm particles. The Stopping criteria is the amount of information about a swarm (fraudulent e-mail or not) that is possessed by a particle. The fitness value of each particle is the heuristic merit (initial velocity) possessed by the features extracted from e-mail contents and is used to evaluate the best position (local solution) of the particle and that of the swarm (global solution). If stopping criteria is not met, then the amount of information missing in a feature (or particle) is determined by the amount of change in velocity needed to move a particle

from its best position to that of the swarm's.

Using the Particle Swarm Optimization (PSO) algorithm, an initial set of feature space containing a number of e-mail features was selected which formed the basis of the initial position of each particle $x_i(t)$ in the swarm with an initial velocity, $v_i(t)$ assigned to each particle. Each particle containing a set of features/attributes takes note of its best position $P_i(t)$ and the global position $G_i(t)$ of the swarm based on the best positions of other particles containing another set of attributes/features and uses them to determine the new position $x_i(t+1)$ and velocity $v_i(t+1)$of the particle.

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (2)$$

$$v_i(t+1) = w\,v_i(t) + c_1(P_i(t) - x_1(t)) + c_2\big(G_i(t) - x_1(t)\big) \quad (3)$$

The expression for determining the new position from the initial position and final velocity is as shown in equation (2) while the expression for determining the new velocity from the particle's initial position, its personal best position and the global solution of the swarm is presented in equation (3). Using the 2 equations for updating the position and velocity of particles, the PSO was used to identify the most relevant features among the initially identified features from the e-mails.

3. 3. RESEARCH PROPOSITION AND FORMULATION

LetE = $\{E_1, E_2, E_2 \ldots\ldots E_i\}$ be the set of E-mails gathered for fraud detection. Let Y= $\{Ham, Spam, Fraudulent\}$ represent the set of outcome for each e-mail collected. Let F_{ri} represent the set of features r extracted from e-mails i following tokenization. By applying PSO, the feature set of e-mail is reduced from r to m where m<r to F_{mi} as shown in equation (4). It is the attempt of this study to show that using PSO to update the weights of back-propagation algorithm, the performance will improve.

$$PSO\big(E_{F_{ri}}\big) = F_{mi} \quad (4)$$

Therefore, it is expected that the fraud detection model which is based on PSO-based BP algorithm $BP_{PSO}(F_{ri})$, using the features F_{ri}extracted with PSO should perform better than the model $BP_{PSO}(F_{mi})$developed using the original features F_{mi}. This in turn is expected to perform better than using the fraud detection model based on the gradient-based BP algorithm $BP(F_{ri})$ using the features F_{ri}extracted with PSO which should also perform better than the model $BP(F_{mi})$developed using the original features F_{mi} extracted from the e-mails according to equation 5.

$$(F_{mi}) > BP_{PSO}(F_{ri}) > BP(F_{mi}) > BP(F_{ri}) \quad (5)$$

The procedure provided above was used for the development of the classification model development process as shown in figure 1. The PSO algorithm was used to extract relevant features after which the features identified were used as a basis of developing the classification model using the PSO-based BP algorithm.

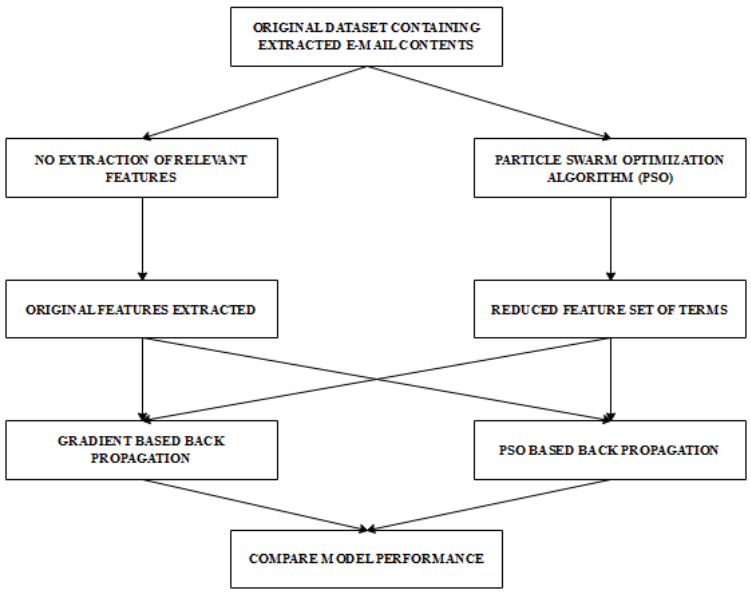

Figure 1: Conceptual Diagram of the Model Development Process

The performance of the fraud detection model developed using the PSO-based BP algorithm was compared with that of the gradient-based BP algorithm using the original features extracted from e-mails and relevant features selected by PSO in order to identify the most effective classification model for advance fee fraud detection.

3.1 Identification of the Optimization Problem

During the feature selection process of this study, the swarm intelligence algorithms selected for this study PSO handled the problem as an optimization problem with its own objective function and constraints. The objective function of the study requires the selection of the optimal number of features x_r from the initial identified number of features x_i which were extracted from the collected e-mails and will maximize the classification accuracy of the fraud detection problem according to equation 6. As a result of this, the irrelevant e-mail features are likely to

have the coefficient α_j tending towards 0 while more relevant e-mail features are likely to tend to 1.

$$\max_{x_r \subset x_i} \left(f\left(\sum_{j=1}^{r} \alpha_j x_j \right) \right) \quad subject\ to: 0 < \alpha_j < 1 \tag{6}$$

$\alpha_j \epsilon \mathbb{R};\ x_j\ is\ the\ set\ of\ selected\ relevant\ features$

However, during the back-propagation process, the swarm intelligence algorithms selected for this study using PSO handles the problem of selecting the optimal value of weights that were used by the back-propagation algorithm of the multi-layer perceptron as an optimization problem with its objective function and constraints. The objective function requires the selection of value for weights w_{ik} attaching an input i to a node k that minimizes the mean square error of the back-propagation algorithm according to equation 7. By comparing the models developed according to figure 1, then one can formulate the fraud detection model which can be used to classify e-mails as spam, ham or fraudulent according to equation 8.

$$\min_{w_{rs} \subset w_{ij}} \left(f\left(\sum_{k=1}^{r} \sum_{i=1}^{s} w_{ik} O_{kj} \right) - A_{ij} \right)^2 \quad subject\ to:\ -1 < w_{ik} < 1 \tag{7}$$

$w_{ik} \epsilon \mathbb{R}\ such\ that\ k \epsilon (1, r)\ and\ i \epsilon (1, s);$

$O_{kl}\ is\ set\ of\ node\ output$

$$BP_{SI}\left(SI\left(E_{F_{ri}}\right)\right) = \begin{cases} Ham \\ Spam \\ Fraudulent \end{cases} \tag{8}$$

3.2 Formulation of PSO-based Fraud Detection Model

The PSO based Back propagation (PSO-BP) algorithm requires the combination of the PSO algorithm with the back propagation algorithm of the neural network. The PSO has been identified as a global algorithm because of its strong ability to find global optimistic result. This PSO algorithm, however, has a disadvantage that the search around global optimum is very slow. The BP algorithm, on the contrary, has a strong ability to find local optimistic result, but its ability to find the global optimistic result is weak. By combining the PSO with the BP, a new algorithm referred to as SI–BP hybrid algorithm is formulated in this study.

The fundamental idea for this hybrid algorithm is that at the beginning stage of searching for the optimum, the PSO was employed to accelerate the training speed. When the fitness function value has not changed for some generations, or value changed is smaller than a predefined number, the searching process is switched to gradient descending searching according to this heuristic knowledge. The PSO–BP algorithm's searching process is also started from initializing a group of random features (or particles) which are updated until a new generation set of particles are generated, and then those new particles are used to search the global best position in the solution space.

Finally, the PSO-BP algorithm was used to search around the global optimum. In this way, this hybrid algorithm may find an optimum solution more quickly. As stated before, the task at hand is an optimization problem which requires the need to identify the values of the weights that will minimize the squared-error in detecting fraudulent e-mails using the identified features $X_i$as defined in equation 9. In this equation, the values w_{ij} represent the set of weights connected to features, O_{kl} is the set of output values from each node and A_{kl} is the actual value of the class to which an e-mail belongs to while X_i is the set of features applied. This was done by limiting the constraint of the weights attached to the interval$-1 < w_{kj} < 1$.

$$\min_{w_{rs} \subset w_{ij}} f(X_i) = \left| \sum_{k=1}^{r} \sum_{l=1}^{s} w_{kl} O_{kl} - A_{kl} \right|^2 \tag{9}$$

Therefore, the hybrid model which combines the PSO with the BP algorithm $PSO - BP(X_r)$was required to collect the values of the r features identified by PSO to perform the following classification of e-mails as either spam, ham or fraudulent e-mails as shown in equation 10 using a mapping function g(.). The function g(.) was used to map the values of the set of relevant features X_r to the output interval {0, 1, 2} such that 0, 1 and 2 corresponded to Ham, Spam and Fraudulent Mails respectively.

$$PSO - BP(X_r) = g(X_r) = \begin{cases} Ham\ Mail & if\ g(X_r) = 0 \\ Spam\ Mail & if\ g(X_r) = 1 \\ Fraudulent\ Mail & if\ g(X_r) = 2 \end{cases} \tag{10}$$

Figure 3.6 shows the flowchart of the hybrid back-propagation algorithm which requires the use of the PSO algorithm for the purpose of training the back-propagation algorithm required by the artificial neural network to update the weights attached to its

node during model training. The hybrid algorithm requires that the weight are initialized to a random value within the interval of [-1 1] following which the feed-forward algorithm was used to propagate the sum of product of the weights attached to its respective inputs (or hidden layer output) attached to their successive nodes to which sigmoid functions was used as the activation function.

The simulation of the back-propagation algorithm involved the use of the Google drive for the movement of e-mail data in and out of the process and the TensorFlow back-end for modeling, 2 multi-layer perceptron models with 2 layers were constructed. The perceptron models generated were modeled using input neurons equal to the number of attributes n in the dataset used (either 12831 or 6382 input neurons). The input neurons (i1 to iN) were attached to the first layer using n weights attached to each neuron present, in all there were 20 neurons (r1 to r20) in the first layer of the network as shown in Figure 3.7.

Therefore, 20 sets of n weights were attached to n inputs to the neurons in the first layer of the multi-layer perceptron network. The second hidden layer had 10 neurons (s1 to s10) to which were attached 10 sets of 20 weights coming from the output of the first hidden layer. Finally, the output of the second hidden layer were attached to three output layer (o1 to o3) namely: Fraud, Ham and Spam. The outputs produced by each neuron in the hidden layer and outer later were created using activation functions, namely: the rectifier linear unit (ReLU) and the Softmax function also called the normalized exponential function. The values provided at the output nodes were used to determine the error in prediction which is required by the back propagation for weight update.

The mean square error of the prediction made in comparison to the actual values recorded in the validation dataset was used to adjust the values of the weights based on the results of the Particle Swarm Optimization (PSO) algorithm. The weights attached to the nodes from the output layer through the hidden layers to the input layers were adjusted following which another iteration is performed in order to perform further adjustments to the model. This process continued until there was no more error detected between the predicted and actual values of the validation dataset such that there was no need to adjust the weights. The final model developed following this procedure was then validated in order to determine the performance of the fraud detection model needed for the development of the fraud detection and incident reporting system.

3.3 System Development Tools

In order to develop the fraud detection and incident reporting system, the choice of the programming language was the Python programming language. This is because the python program supports the analysis of unstructured text by making use of the python natural language toolkit library available from Python. It also supports the collection, pre-processing and the analysis of unstructured text with the use of supported libraries such as the Python Natural Language Toolkit (NLTK) library. The dataset containing the e-mails that were required to be processed were imported to a bag of words and parsed to the Python NLTK following which the process of tokenization which was required for extracting all the words within the body of the e-mail. All words were extracted following which the process of stop word removal was done in order to remove the most frequently occurring words in English sentences such as pronouns, articles and prepositions to mention a few. The process of stop word removal was necessary so as to facilitate the thorough reduction of the vector space of the feature set which were constructed. The process of stemming was performed using the Porter's stemming algorithm, the process involved the reduction of related words to their base forms and was done with the aim of further reducing the feature space of the dataset. The set of words that remained in the bag of words were the basis of generating the final feature set which was used in converting all unstructured e-mails into a term-frequency matrix which contained the words along the column with their respective binary value provided. The term-frequency matrix was generated by parsing all the words in the bag through each e-mail following which a value of 1 was entered if the word was present otherwise a value of 0 was provided. This process was repeated for all e-mails following which a dataset containing the extracted words as the feature set while each row represented the presence (or absence) of each word in each e-mail. This dataset was used as the basis of the development of the fraud detection model using the PSO-based BP algorithm.

The process of model development was done using the Python Machine Library via which the PySwarms library was used to implement the Particle Swarm Optimization (PSO) algorithm that was used to extract the most relevant features from the initially identified features in the dataset. The features extracted by PSO were used to generate another dataset which contained e-mail records which had binary values for the extracted features excluding the columns of non-relevant features. Following the process of the feature selection of relevant features using the PSO algorithm, the PSO-based back propagation algorithm was implemented using the Python ML library by hybridizing the neural network with the PSO algorithm as presented in this study.

The dataset containing the initially identified features and that containing the finally extracted features were used to build the fraud detection model using the gradient-based and PSO-based BP algorithm. The simulation process provided four (4) fraud detection models which were compared based on a number of performance evaluation metrics following which the most effective fraud detection algorithm was selected. The most effective model that was identified was integrated into the implementation of the fraud detection and incident reporting system using the Python programming language.

4. RESULTS AND DISCUSSION

The section presents the results of the particle swarm algorithm (PSO) algorithm that was used to identify the most relevant features alongside the features that were extracted. It also presents the results of the hybrid back-propagation algorithm which uses the PSO algorithm to optimize the selection of optimal weights using the back-propagation algorithm.

4.1 Results of the Extraction of Relevant Features using PSO Algorithm

Using the PySwarms.py library, the Particle Swarm Optimization (PSO) algorithm that was used for the selection of the relevant features from the initially identified features was performed. The simulation required for the process was performed by applying the PSO algorithm to the selection of features and testing the effectiveness of the features using logistic regression. The set of attributes with the lowest error rate for the logistic regression plot was selected as the optimal feature set. The PSO algorithm was simulated using 30 particles with a dimension equal to the attributes extracted from the e-mail records (=12831) which was used to define the shape of the dataset.

Using the PySwarms.py library to run the PSO algorithm for

feature selection, a total of 6382 features were selected from the initially extracted 12831 features. Using the 6382 features reduced the dimensionality of the data by about 50% thus offering possibilities for improved performance in terms of reduced error rates and reduced space-time complexities. Following the selection of relevant features using the PSO algorithm on the original dataset, the results of the simulation and the validation of the hybrid Particle Swarm Optimization-Back Propagation (PSO-BP) algorithm using the e-mail dataset was presented.

4.2 Results of the Formulation and Simulation of the Hybrid PSO-BP Algorithm

The results of the process of model formulation and simulation for fraud detection involved the use of the dataset containing the originally extracted 12831 features and that containing the 6382 features selected by PSO according to the conceptual framework presented in the previous chapter. As a result of this, a multi-layer perceptron network which depended only the back-propagation algorithm and gradient descent was developed using the 2 dataset followed by the development of the hybrid PSO-BP algorithm. The four (4) models developed were then compared based on the values of the performance evaluation metrics estimated from their respective confusion matrix. The experiment was performed by using the dataset in such a manner that the dataset was split into training/testing dataset proportion of (70/30) % and (80/20) % such that there were 6002/2573 and 6860/1715 for training and testing respectively. Table 2 shows the distribution of the target class labels among the datasets used for the simulation process within the testing dataset used in this study.

Table 2: Distribution of Target Class labels among Testing Dataset

Dataset Data	30% Testing Data		20% Testing	
	Frequency	Percentage (%)	Frequency	Percentage (%)
Spam Mail	837	32.53	557	32.49
Ham Mail	544	21.14	363	21.17
Fraud Mail	1192	46.33	795	46.36
Total	2573	100.00	1715	100.00

The results of the correct and incorrect classifications made by the model which were projected into the confusion matrix were later used to estimate the performance of the model developed in order to select the most accurate model for fraud detection. The results of the error rate for the training and testing datasets presented in Figure 2 shows the behaviour of the model using the training and testing dataset for model validation for model performance evaluation.

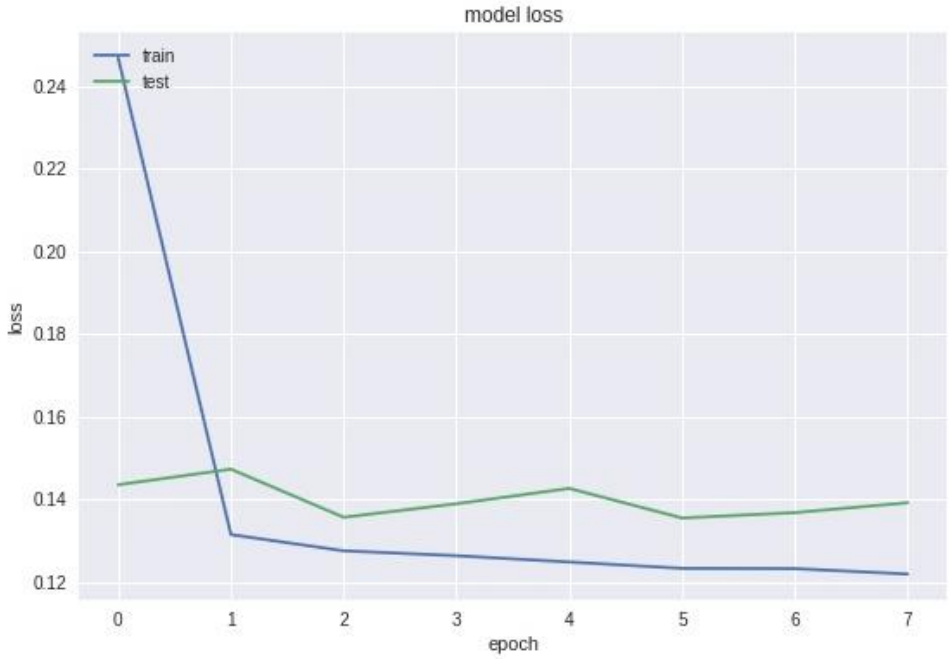

Figure 2: Graphical Plot of Training and Testing Errors using PSO based Back Propagation at 7 epochs

The plot showed that the model attained lower error rate within the training dataset compared to that of the testing dataset due

to the fact that the training dataset was adopted for model development. Since the dataset collected had been split into 2 sets of training/testing dataset 70/30 and 80/20, both datasets were adopted for the simulation and validation of the fraud detection algorithm for this study. In the dataset consisting of 30% testing dataset, there was 837 Spam (32.3% of Spam), 544 ham (21.14% of Ham) and 1192 Fraud (46.33%) e-mails which totals to 2573 (30% of datasets) while in the dataset containing the 20% testing dataset there were 557 Spam (32.49%), 363 Ham (21.17%) and 795 Fraud (46.36%) e-mail datasets.

These datasets were used for the simulation and validation of the model using original and relevant features on the gradient-based and PSO-based BP algorithms. The results of the simulation and validation of the gradients-based PSO model using the 30% and 20% testing dataset are presented based on the use of its original features and using the relevant features selected by the PSO algorithm. The results of the use of the 20% testing dataset was presented followed by that of the 30% testing dataset.

4.3 Results of the Simulation of the PSO-based BP Algorithm

By using the 30% of the dataset containing the originally extracted features from e-mails to build the PSO-based BP algorithm the following were observed: out of the actual 837 spam mails, 781 were correctly classified while 56 were incorrectly classified as ham mails; out of the actual 544 Ham mails, 474 were correctly classified while 70 was incorrectly classified as Spam mails; while out of the actual 1192 Fraud e-mail, 1190 were correctly classified while 2 were incorrectly classified as Spam mail. In total there were 2445 correct classification out of 2575 owing for an accuracy of 95.03% as shown in the confusion matrix in Figure 3 (a).

By using the 30% of the dataset containing the features selected using PSO from the initially extracted features from the e-mails to build the PSO-based BP algorithm the following were observed: out of the actual 837 spam mails, 792 were correctly classified while 45 were incorrectly classified as ham mails respectively; out of the actual 544 Ham mails, 501 were correctly classified while 43 were incorrectly classified as Spam; while out of the actual 1192 Fraud e-mail, 1191 were correctly classified while 1 was incorrectly classified as Spam. In total there were 2414 correct classification out of 2575 owing for an accuracy of 96.54% as shown in the confusion matrix in Figure 3 (b).

	SPAM	HAM	FRAUD
SPAM	781	70	2
HAM	56	474	0
FRAUD	0	0	1190

(a)

	SPAM	HAM	FRAUD
SPAM	792	43	1
HAM	45	501	0
FRAUD	0	0	1191

(b)

Figure 3: Results of Simulation and Validation of 30% Testing Dataset using PSO based BP Algorithm

By using the 20% of the dataset containing the originally extracted features from e-mails to build the PSO-based BP algorithm the following were observed: out of the actual 557 spam mails, 517 were correctly classified while 40 were incorrectly classified as ham mails respectively; out of the actual 363 Ham mails, 313 were correctly classified while 50 was incorrectly classified as Spam e-mails; while out of the actual 795 Fraud e-mail, 789 were correctly classified while 6 were incorrectly classified as Spam e-mail. In total there were 1619 correct classification out of 1715 owing for an accuracy of 94.40% as shown in the confusion matrix in Figure 4 (a).

By using the 20% of the dataset containing the features selected from the initially features extracted from the e-mails to build the PSO-based BP algorithm the following were observed: out of the actual 557 spam mails, 518 were correctly classified while 39 were incorrectly classified as ham mails respectively; out of the actual 363 Ham mails, 335 were correctly classified while 28 was incorrectly classified as Spam e-mails; while out of the actual 795 Fraud e-mail, 791 were correctly classified while 3 and 1 were incorrectly classified as Spam and ham mail respectively. In total there were 1644 correct classification out of 1715 owing for an accuracy of 95.96% as shown in the confusion matrix in Figure 4 (b).

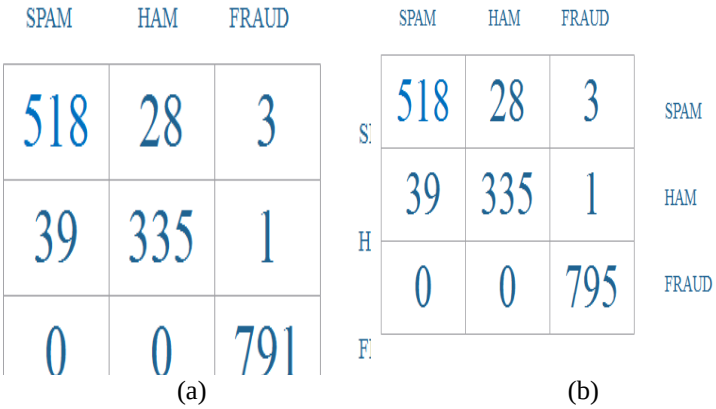

Figure 4: Results of Simulation and Validation of 20% Testing Dataset using PSO Based BP Algorithm

4.4 Discussion of the Results of Model Simulation and Validation

Following the process of the development of the hybrid model for the classification of fraudulent e-mails, the results of the process of simulation and validation have been presented. The results of using the 30% testing dataset for the gradient-based BP algorithm showed that using the relevant features selected by PSO improved the accuracy of the gradient-based BP by a value of 0.27%. Also the results of using the 20% testing dataset for the gradient-based BP algorithm showed that using the relevant features selected by PSO improved the accuracy of the gradient-based BP by a value of 0.35%.

The results of using the 30% testing dataset for the PSO-based BP algorithm showed that using the relevant features selected by PSO improved the accuracy of the PSO-based BP by a value of 1.51%. The results of using the 20% testing dataset for the PSO-based BP algorithm showed that using the relevant features selected by PSO improved the accuracy of the PSO-based BP by a value of 1.46%. The results also showed that using the PSO-based BP improved the performance of the BP compared to gradient-based BP based on the 30% testing dataset by a value of 1.48% and 2.72% using the original features and the features selected using PSO respectively. Also, using the PSO-based BP improved the performance of the BP compared to gradient-based BP based on the 20% testing dataset by a value of 1.46% and 2.57% respectively using the original features and the features selected using PSO as shown in Figure 5. Overall, the results of the formulation of the gradient-based BP algorithm showed that the best performance was achieved when using the 30% dataset for testing based on features that were selected by PSO from the initially extracted features from the e-mails.

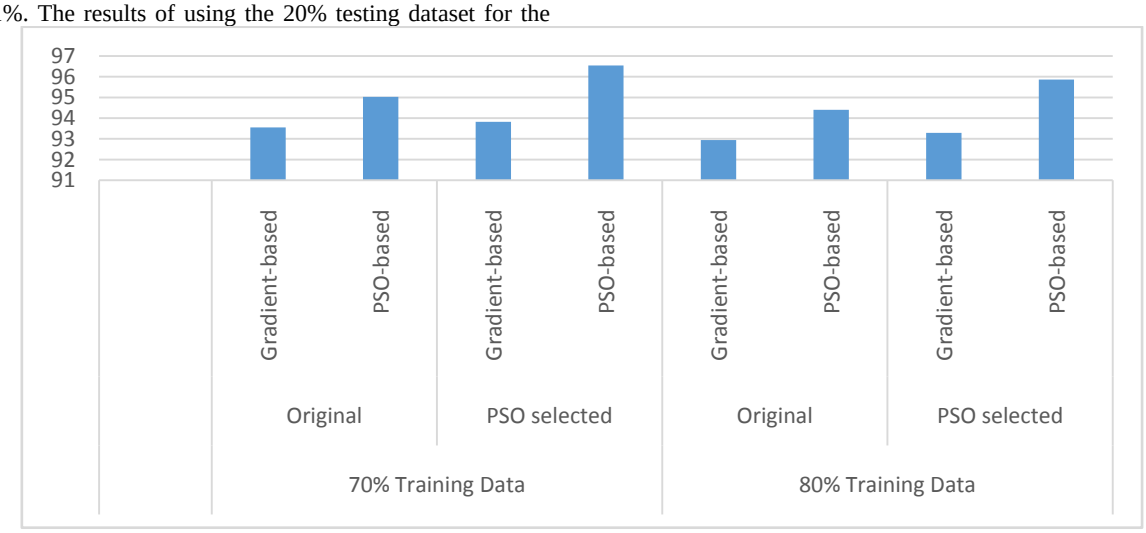

Figure 5: Results of the Accuracy of the simulated Models

Table 3 shows a summary of the results of the evaluation of the performance of each model validated in this study. The results showed that the values of the TP rate, FP rate and the Precision of the models that were developed in this study was proportional to the performance of the variation of the BP algorithms developed in this study. The results clearly showed the model with the best performance had the highest value for the TP rate, accuracy and precision but lower values for the FP rate. The BP algorithm with the best performance overall was the gradient-based BP algorithm that was formulated using the features that were selected using the PSO algorithm from the initial features extracted from the fraudulent e-mails using 30% testing dataset.

Table 3: Summary of Result of Model Simulation

Training Data	Attribute Set	BP Algorithm	Correct	Accuracy (%)	TP rate			FP rate			Precision		
					Spam	Ham	Fraud	Spam	Ham	Fraud	Spam	Ham	Fraud
70% Training Data	Original	Gradient-based	2407	93.55	0.914	0.833	0.997	0.052	0.04	0.001	0.894	0.86	0.999
		PSO-based	2445	95.03	0.933	0.871	0.998	0.041	0.03	0	0.916	0.894	1
	PSO selected	Gradient-based	2414	93.82	0.898	0.866	0.999	0.043	0.04	0	0.91	0.847	1
		PSO-based	2484	96.54	0.946	0.921	0.999	0.025	0.02	0	0.947	0.918	1
80% Training Data	Original	Gradient-based	1594	92.94	0.912	0.813	0.995	0.06	0.04	0.001	0.88	0.853	0.999
		PSO-based	1619	94.4	0.928	0.862	0.992	0.048	0.03	0	0.902	0.887	1
	PSO selected	Gradient-based	1600	93.29	0.898	0.848	0.996	0.049	0.04	0	0.898	0.842	1
		PSO-based	1644	95.86	0.93	0.923	0.995	0.027	0.03	0	0.944	0.893	1

The results showed that using the gradient-based BP algorithm classifier that adopted the relevant attributes extracted by PSO showed a better performance compared to that developed using the original attributes selected from the e-mail dataset. This can be seen in figure 6.

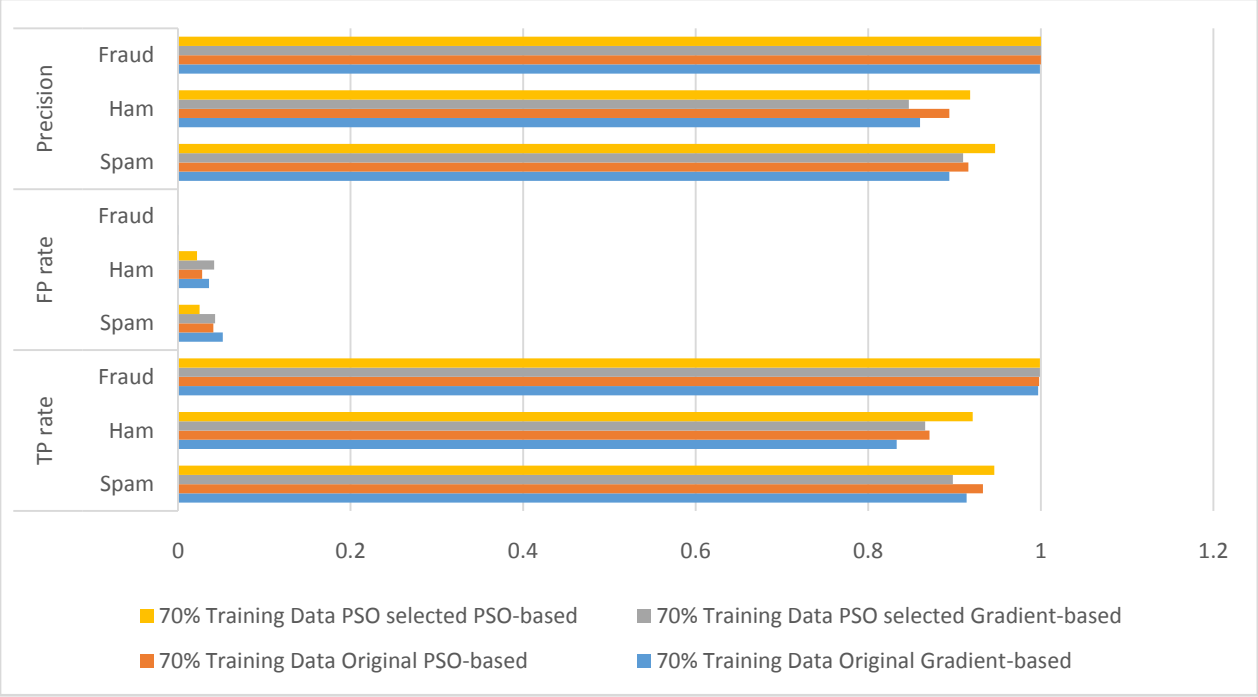

Figure 6: Results of the TP rate, FP rate and Precision for 30% testing data

Also, the PSO-based BP algorithm classifier adopted using the attributes selected by the PSO algorithm outperformed the classifier formulated using the original extracted attributes. However, on a general note, the overall best performance was achieved using the PSO-based BP classifier which adopted the relevant attributes extracted from the e-mails using PSO basedon the 20% testing dataset. This is as shown in Figure 7.

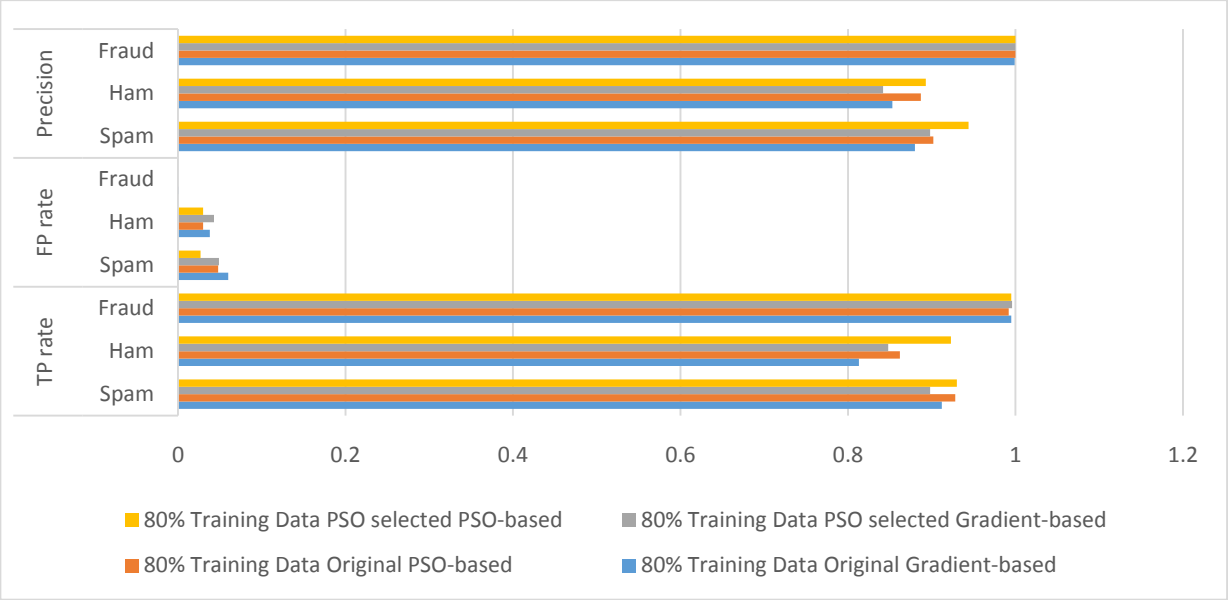

Figure 7: Results of the TP rate, FP rate and Precision for 20% testing data

Therefore, the PSO-based BP algorithm formulated using features extracted by PSO based on a 20% testing dataset size was integrated into the implementation of the fraud detection and incident reporting system required for the classification of fraudulent e-mails.

5. CONCLUSION

The results of this study showed that following the adoption of the PSO algorithm for the selection of relevant words, it was observed that 6382 (49.7%) words were considered most important for speeding up and improving the performance of the fraud detection model. The study concluded that using the PSO algorithm for the selection of the most relevant features was reduced by about 50% of the originally extracted features from the collected e-mail dataset.

Furthermore, the results of the formulation and validation of the fraud detection model showed that using the relevant features selected by PSO increased the performance of the gradient-based BP by 0.27% and 0.35% using 30% and 20% testing data respectively. Also, the selection of relevant features using PSO increased the performance of the PSO-based BP by 1.51% and 1.46% using 30% and 20% testing data respectively. Therefore, the PSO-based BP outperformed the gradient-based BP with a value of 1.48% and 2.72% using 30% testing data based on original features and features selected by PSO respectively. Also, the results showed that the PSO-based BP outperformed the gradient-based BP with a value of 1.46% and 2.57% using 20% testing data based on original features and features selected by PSO respectively.

Based on the results, the study concluded that adopting the use of PSO for the selection of the most relevant features improved the performance of the fraud detection model while adopting the PSO for the optimization of the back-propagation algorithm also improved the performance of the BP algorithm compared to using the gradient-descent method. Furthermore, the study concluded that the fraud detection model developed using the PSO-based BP which used the most relevant features was integrated into a prototype implementation of the fraud detection and incidence reporting system. The incidence reporting was handled by using a blacklist to keep track of known fraudulent e-mails while unknown e-mails were analyzed in order to classify the incoming e-mail as either Spam, Ham or Fraud mail.

6. REFERENCES

[1] Oyegoke T. O., Amoo A. O., Aderounmu G. A. and Adagunodo E. R. (2020). An Email Classification Model for Detecting Advance Fee Fraud: A Conceptual Approach. Computing, Information Systems & Development Informatics Journal 11 (2), 91 -104

[2] Jennings, R. (2009). Cost of Spam is Flattening — Our 2009 Predictions. Retrieved from http,//e-mail-museum.com/2009/01/28/cost-of-spam-is-flattening-our-2009-predictions/ on July 23, 2017.

[3] Blanzieri, E. and Bryl A. (2008). A Survey of Learning-Based Techniques of E-Mail Spam Filtering. Artificial Intelligence Review 29(1), 63 – 92.

[4] Australian Federal Police. (2010). Internet Fraud and Scams. Retrieved from http, //www.afp.gov.au/policing/e-crime/internet-fraud-and-scams.aspx on July 23, 2017.

[5] Behdad, M., Barone, L., Bennamoun, M. and French, T. (2012). Nature-Inspired Techniques in the Context of Fraud Detection. IEEE Transactions on Systems, Man and Cybernetics – Part C, Applications and Reviews 42(6), 1273 – 1290.

[6] Carcillo, F., Dal Pozzolo, A., Le Borgne, Y.-A., Caelen, O., Mazzer, Y. and Bontempi, G. (2017). Retrieved from http,//creativecommons.org/licenses/by-nc-nd/4.0/ on September 23, 2017.

[7] Oyegoke T. O., Akomolede K. K., Aderounmu G. A. and Adagunodo E. R. (2021). A Multilayer Perceptron Model for Email Classification.

[8] Bolton, R.J. and Hand, D.J. (2002). Statistical fraud detection, A review. Statistical Science Journal 17(3), 235 – 249.

[9] Marrow, P. (2000). Nature-inspired computing technology and applications. BT Technology Journal 18, 13 – 23.

[10] Dorigo, M. and Stutzl, T. (2004). Ant Colony

Optimization. Bradford Company Publishers, Minnesota, MI.

[11] de Castro, L.N. (2007). Fundamentals of Natural Computing, An Overview. Journal of Physics of Life Reviews 4, 1 – 36.

[12] Nasser, M. and Seyed, J.M. (2014). Comparison of Particle Swarm Optimization and Back Propagation Algorithms for Training Feed-forward Neural Network. Journal of Mathematics and Computer Science 12, 113 – 123

[13] Eiben, A.E. and Smith, J.E. (2003) Introduction to Evolutionary Computing. Springer-Verlag, Berlin, Heidelberg.

[14] Fleming, P.J. and Purshouse, R.C. (2002). Evolutionary algorithms in control systems engineering, A survey. Control Engineering Practice 10(11), 1223 – 1241.

[15] Nikolos, I., Valavanis, K., Tsourveloudis, N. and Kostaras, A. (2003). Evolutionary algorithm based offline/online path planner for UAV navigation. IEEE Transaction Systems, Man and Cybernetics 33(6), 898 – 912.

[16] Jin, Y. and Branke, J. (2005). Evolutionary optimization in uncertain environments - A Survey. IEEE Transactions in Evolutionary Computing 9(3), 303 – 317.

[17] Wertheimer, G. (1999). Gestalt theory reconfigured, Max Wertheimer's anticipation of recent developments in visual neuroscience. Journal of Perception. 28 (1), 5 – 15. PMID 10627849. doi,10.1068/p2883

[18] Backstrom, L., Huttenlocher, D., Kleinberg, J. and Lan, X. (2006). Group formation in large social networks. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06, 44 - 53. ISBN 1595933395. doi,10.1145/1150402.1150412.

[19] Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. The Annals of Mathematical Statistics 38(2), 325–339.

[20] Shafer, G. (1976). A Mathematical Theory of Evidence. Princeton University Press, New Jersey, USA.