Journal of Theoretical and Applied Information Technology
15th April 2020. Vol.98. No 07

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN INVESTIGATOR DIGITAL FORENSICS FREQUENCIES PARTICLE SWARM OPTIMIZATION FOR DETECTION AND CLASSIFICATION OF APT ATTACK IN FOG COMPUTING ENVIRONMENT (IDF-FPSO)

AHMAD K. Al HWAITAT[1], SAHER MANASEER[2], RIZIK M. H. Al-SAYYED[3] ,MOHAMMED AMIN ALMAIAH[4], OMAR ALMOMANI[5]

[12]King Abdullah The II IT School, Department Of Computer Science, The University Of Jordan, Jordan.

[3]King Abdullah The II IT School, The University Of Jordan ,Department Of Information Technology, Jordan.

[4] King Faisal University, Department Computer Science, Saudi Arabia.

[5]The World Islamic Siences And Education University, Department Computer Networks, Jordan

E-mail: [1]Ahmad.Hwaitat1@gmail.com,[2]Sahr@ju.edu.jo , [3]r.alsayyed@ju.edu.jo, malmaiah@kfu.edu.sa[4] , Omar.almomani@wise.edu.jo[5].

ABSTRACT

The hough there are several approaches to detect the malware attacks in cloud, the detection techniques could not be applied in FOG based environment. This is because of its possession of distinct features. As FOG computing has been evolving, it is mandatory to develop detection and mitigation schemes of malware attacks. Thus, in this research, an approach for investigation of digital forensics has been developed, where it classifies and detects the APT attack named Shamoon attack from different attack types in FOG environment. Digital Forensics has been recently gaining focus in solving or investing the cybercrimes. Several researches have been developed in this field where they have analyzed several security challenges. Previous technologies, to measure these attacks are completely based on methodology of pattern matching. If an attack is newly occurred, then the detection rate is very low and false negative will be very high. Thus the challenges are highly increased as the data volume increases, and the technology used by attacker is continually developed. As there is a lack in detection technology and the deployment boards, and the low efficient models in FOG computing makes the challenge a difficult one. Thus a proposed scheme has been introduced where Frequency Particle Swarm Optimization (FPSO) has been utilized in investigating digital forensics Particle Swarm Optimization in order to detect and to classify the APT attack (Shamoon attack) in FOG environment. This approach uses four phases. In feature extraction, best set of features are extracted. Using FPSO (Frequencies PSO), best weighed features are predicted. These weighed features are clustered using K-means clustering and classified using k-nearest neighbors (KNN) classifier. The performance of this approach is then evaluated using confusion matrix and results are provided. Finally, the proposed KNN-FPSO classifier is compared with other existing classifiers and the results are recorded.

Keywords: *Digital Forensics, Shamoon Attack, FOG Computing,APT Attacks, Cyber Security, Machine Learning, Information Security.*

1. INTRODUCTION

FOG computing is an extension model of cloud computing where it has the network, storage and computational facilities towards the edge of the networks, during cloud data offloading and thus reducing the latency to end users. But such characteristics are raised with new security and secrecy related challenges [26]. There are various safety and privacy measures for cloud computing. But these measures couldn't be applied in FOG[23]. This is because of the structures such as heterogeneity, mobility, and wider scale of geographic distribution. [1] provided an explanation and overview of conventional privacy

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

and security concerns for FOG. It also highlighted the survey about new research, open challenges, and trends followed to mitigate the security issues in FOG [28].

Digital forensic is a sort of a category in forensic science which is about to investigate the material recognized in digital devices. It is often related as a computer crime. Its investigations has wider variety of applications. It mainly supports a hypothesis provided before the courts. The major goal is to safeguard the evidence in most primary form without compromising the investigation structure of gathering, recognizing and authenticating the digital data, to re-construct the past events. [2] paper introduced major security approaches and challenges in digital forensics.

Shamoon [3]is a considered to be the most destructive wiper malware which harms/corrupts records on a computer that are compromised thereby overwrites the Master Boot Record (MBR) with an effort for rendering an unusable one. Wiper is the malwares class name which in turn wipes out hard drives. Generally, the data that are wiped out are not a recoverable one. So far, it is one of the most popular wiper. After Shamoon, in history, another version of this was the most mysterious wipers and this has been hidden for almost four years. This new version have come up with different features thereby attacking company Aramco in Kingdom of Saudi Arabia (KSA) [23]. This version targets for the huge demolition of the systems in a targeted groups of KSA. This latest version consists of several resemblances with Shamoon, however it is more advanced in performing different techniques and tools. All through the attack of Shamoon, the invaders acquire proprietor authorizations on behalf of victim's network.

In this paper, a proposed scheme (IDF- FPSO) has been introduced where Frequency related PSO has been utilized in investigating digital forensics Particle Swarm Optimization. This scheme will detect and classify the APT attack (Shamoon attack) in FOG environment. This approach uses four phases. In feature extraction, best set of features are extracted. Using FPSO (Frequencies PSO), best weighed features are predicted (features with weight > 0.6). These weighed features are clustered using K-means clustering and classified using KNN classifier. The primary objectives of this algorithm are provided below,

- To effectively detect and to classify the APT attack especially Shamoon attack using investigator digital forensics frequenciesparticle swarm optimization for dectection and classification of apt attack in fog computing enviroment (IDF-FPSO) algorithm
- To obtain best predicted features by analyzing the weights (>0.6) using FPSO optimizer.
- Performance and comparative analysis on the basis of confusion matrix

The organization of the remaining paper is as follows: Section 2 explains the related work and literature survey. Section 3 describes the proposed system design along with the details. Section 4 discusses the performance and experimental results. Section 5 presents the comparative analysis with various classifiers. Finally section 6 presents the conclusion of the study.

2. RELATED WORKS

The term fog computing has been gaining incredible popularity due to mobile computing demand along with small delay. Although, it is a virtual based environment, it is also susceptible to cyber-attacks namely APT – Advanced Persistent Threat Attacks, Shamoon attacks, etc. Recently various approaches have been developed in securing computing networks against APT and Shamoon related attacks. Because of the progressive development of these two attacks, securing fog computing networks completely may not be possible. Because of this, extensive application of fog computing in services like business, etc. has been delayed. To safeguard against these attacks, the security methods were not only emphasized on improving the security methods from the providers of fog computing, but also consider another means of cyber risk management. This section discusses about the researches done in Fog computing, Digital forensics, APT attacks, Shamoon attacks.

2.1. Fog Computing

A group of computers and servers associated in a network are generally clarified as a cloud computing. Recently, several organizations were begin to utilize IOT (Internet of Things)[29], as they needed huge quantity of information that are to be accessed. This is where the term fog computing / fog networking arises [4]. It has been created by Cisco network. It is a sort of distributed infrastructure where some of the application services are managed till end of the network by using edge devices and others in cloud. Usually, the fog network present in the mid layer among the hardware edge devices and cloud. It provides

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

enhanced analysis, storage and processing of data. The fog computing is mainly used to enhance efficient network and minimizes the data amount needed to get transported in analyzing cloud, data processing and storage[23]. [5] defined the extension of cloud computing providing services like data services, end user application, computational storage services. The quality of service was improved and it also eliminates the latency. Cisco stated that due to vast geographical distribution, fog computing could be suitable for big data and real time analytics.

It may also be defined as platform that is cloud-like using same storage, data, and computation and application type services, however it is fundamentally different. Additionally, the fog networks have the potentiality to process large amount of local data, and are completely convenient which could be fitted on varied hardware. Due to this nature, it could be employed in applications based on location or time sensitive. For example, devices of IoT should do a quick process on huge amount of data. The extensive sort of functional applications deepens some security concerns like malware threats, etc. [6] surveyed several fog related applications to recognize security gaps. This survey also included similar methodologies like Cloudlets, Edge computing, etc. The major fog applications were inspired to desire end-user requirements. This paper also evaluated the security issues and other possible solutions, keeping in mind that to provide safety related information to developers who are accountable for designing, developing and maintenance of fog systems.

As safety and security are of utmost importance to provide guarantee in service quality, an IDS-Intrusion Detection system for fog computing using smart data approach had been proposed [24] [7]. This approach is based on Artificial Immune System (AIS). [25] [27] This IDS had been comprised of three layers which include cloud, edge layers, and fog. In first Cloud layer, major network traffic gets clustered and trained its indicators. In middle layer (Fog), smart data concept is utilized for analyzing intrusion alerts. In last (edge) layer, detectors were deployed in edge devices. Smart data is considered to be an auspicious method which enabled efficient and lightweight IDS in order to provide a detection path for silent attacks namely botnet attacks in IoT systems[29]. Usually edge devices of IoT are attached to fog devices. Such devices of Fog are resided in close contiguity to the users and were in charge for intermediate operations and storage.

Some of the major challenge in consecutive applications of IoT are resource allocations and scheduling of tasks. [8] surveyed the recent trends, requirements, its architectures, and research directions. Thus the work helped the industry and research community for synthesizing and identifying fog computing's requirements. [9] focused on overcoming various security related issues that are occurred while outsourcing the data from fog client to node. This research work implemented a safety and access control cross area procedure among fog client and node for enhanced secured communication between respective client and node. [10] developed a dual encryption of data using emoji technique along with the combination of steganography and cryptography. The major purpose of this work is to process the closeness of data to the edged devices. Here, the data is encrypted first and then the encrypted data is covered using text like emoticons. [11] stated that Fog computing could be referred as an addition of cloud computing and provided edge of the network services. Cloud computing can able to retain up with current data processing but it is not sure whether it has the ability to retain the field IoT. Thus Fog networking provides an architectural solution to provide solution to the problems.

2.1 Digital Forensics

Digital forensics can be considered as a portion of computer forensics. In cloud, several challenges prevents the method of cloud forensics such that none of the typical background could be designed. [12] paper surveyed some of the challenges and respective solutions. It also presented numerous contests in each cloud forensics step by all possible solutions in order to eliminate those challenges. Digital triage is considered to be the first step in forensic examination. This digital triage has a power in quickly identifying things that are most probably comprise data that were evidential. It seemed to be the solution to the case backlog problems. Several existing methods of digital triage have some of the drawbacks in the forensic context. [13] paper reviewed several available study mechanisms and some suggested solutions for digital triage and explained it in 4 stages namely triage tools, mobile device triage, post-mortem triage and live triage.

As digital forensics is considered to be an approach in dealing cyber-crimes, it has a progressive important. [14] reviewed some of the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

existing forensic models and focused on few challenges in this domain. There were several tools that had been undertaken several questions regarding the future of this domain. In recent days, cybercrimes are happening in increased rate, and involves serious threats towards individual security, even in developed countries. [15] paper reviewed in detail about various cyber-crimes followed by respective digital forensics that are involved in those investigations. It also studied and compared various tools used for digital forensics along with its merits, demerits, challenges, and issues. This paper also recommended the purpose of training programs for the respond and judgment of authentication.

2.2 Advanced Persistent Threats Attacks (APT)

It is considered that APT attacks are one of the major threats in IT, recently. It is a complex phenomenon, also a danger to several organizations. [16] showcased the problem of APT, associated threats, and selected methods and tools that mitigate the APT attacks. This paper also outlined efficient as well as multi-layered model for defense. APTs are considered to be cyber-attacks implemented by well-trained and sophisticated trainers who target particular type of information in high profile governments similar to long term campaign with different steps. [17] [23] presented the consequences of complete revision on APT, describing the unique features and model attack. This paper also enumerated few measures that help in eliminating APTs.

[18] stated that due to many recent discoveries of Advanced Persistent Threats (APTs), it is mandatory to know the purpose of its operations hence to effectively eliminate the attacks. It also analyzed certain characteristics of APTs and compared various life-cycle models with each other and examined the real world APTs such as Energetic Bear, Regin, APT1, Duqu 2.0 and fit of the model. This research had been done in order to examine the validity of the selected model and to utilize it based on practical attack example that explained specific techniques and tools used by APTs. By associating attack vectors using best practices and eliminate specific strategies with no single technology or technique which assures safety from APTS. [19] defined that APTs can be referred as 'security marketing buzzword', as it can be represented as nightmare of attacks. This research is developed based on set of contributions on individual works inscribed by six master students in Information Security at GUC.

2.3 Shamoon Attacks

The most famous cyber-attacks among all is Shamoon, which was against at least 2 organizations in Middle East energy sector. The malware is made in a way which overwrite and wiped the records along with (MBR) Master Boot Record of target hosts for the purpose of making them unusable. [20] [31] stated that some of the vicious targeted attacks namely Shamoon and Samas caused remarkable damages o target systems, leads to cause some disturbances in critical business operations. As these attacks accessed many hosts continuously, a fast response is needed to avoid certain severe damages. To get that response rapidly, the responders needed to locate all the hosts of victim after the 1st victim host alarmed for detection. Hence, it proposed a novel method known as SAST – Suspicious Activity Spike Train which is used for locating potential victims by examining the similarity amongst the activity patterns of the 1st victim and other hosts. Thus it is considered to be the most robust method.

The Middle East region is currently the targeted place of cyber-attacks that are carried out by unknown parties, especially energy industries. These attacks involved deploying sophisticated type malwares. [21] stated that a campaign were opened by Stuxnet malware 2010 and then progressed through Flame, Gauss, Duqu, and Shamoon malware. This paper provided a technical survey on malwares namely Flame, Stuxnet, Shamoon. It also described various main modules and their spreading capabilities. It also pointed out the recent trends that were pervaded by new type of malware into cyber-attacks.

2.4 Optimizer Algorithms

Various optimization algorithms have been employed in providing solutions to cyber-attacks such as PSO, GA, FPSO, and so on [30]. As said earlier, several IT based systems are fragile to cyber-attacks. Several research works had been conducted to deploy a defensive strategy. A research work [22] proposed a defensive strategy to attackers and defenders who are attempting to attack the system. It represented the problem as competitive optimization problem which can be solved by using PSO – Particle Swarm Optimization. As PSO [23], got inspired by the nature of birds flocking, every particle term is denoted as individual swarms. The major aim is to discover the gbest solution by improvising the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

movement of every particle on accordance with gbest and pbest positions. Here every particle got converged to best solution which can be known as 'fitness'. These optimization algorithm will provide solutions to real world problems.

Though PSO is a remarkable optimizer in solving the optimization problems like providing potential search space points in dynamic environment, it has the tendency to impact the convergence early on trying to solve complex problems. Thus an enhanced or modified PSO has been introduced named as FPSO (Frequency Particle Swarm Optimization) [24] . This FPSO does optimization via impersonating three characteristics of waves named as frequency, amplitude and wavelength. For each and every iterations, the optimizer will extract the fitness function with best weighed feature has been selected and stored.

3. PROPOSED WORK (INVESTIGATOR DIGITAL FORENSICS FREQUENCIES PARTICLE SWARM OPTIMIZATION FOR DECTECTION AND CLASSIFICATION OF APT ATTACK IN FOG COMPUTING ENVIROMENT (IDF-FPSO)

3.1 FPSO Optimizer

The enhanced version of PSO also referred as Frequency Particle Swarm Optimization (FPSO) algorithm [24], impersonates the wave characteristics by utilizing 3 parameters such as Amplitude, Frequency and wavelength. The movement of each and every particle is same as movement of the waves. These nature of movement are entirely depend upon the given frequency. Moreover, every particle has distinctive frequency, which means that, not every particle moves along the same direction with respective to every iteration. First of all, the random position of every particle has to be initiated followed by initiating the velocity based frequency. Based upon the frequency, the movement of the particle is decided whether to move up or down. For these every movement, the optimizer will derive the fitness function and deliver the best converged solution to the real world problem. This optimizer involves four types of processes namely 1) Initialization, 2) Evaluation, 3) Updation, 4) Selection. In first step, the particle positions along with frequency got initiated. In second step, fitness function is evaluated. In third step, for the provided number of iterations, pbest and gbest solution are updated. In final step, the best weighed particle is selected by eliminating the least or worst particle. The major aim of FPSO optimizer is to improvise the initial random solution which got converged to best search point. In proposed system of investigator digital forensics frequencies particle swarm optimization for dectection and classification of apt attack in fog computing enviroment (IDF-FPSO), it plays major role in predicting and selecting the best weighed features.

3.2 Shamoon Attack – An Industrial Espionage

Shamoon, can also be referred as W32.Disttrack, is a malware used in attacks against minimum two energy sector organizations in the Middle East country [22]. It is not a usual malware, as its major goal is to deliver maximum possible destruction. It was designed in a way such that it will wipe or overwrite the files and making Master Boot Record of the system unusable. Figure 1 depicts the Shamoon malware components of its main file known as TrkSvr.exe. Its main components are dropper, wiper, and reporter.

Dropper – It plays major role in installing the malware and initiating the process. It follows the technique of network sharing to attain maximum spread. Once the target is determined, the malware tries to remotely access the file, copying it and executing it by using psexec.exe.

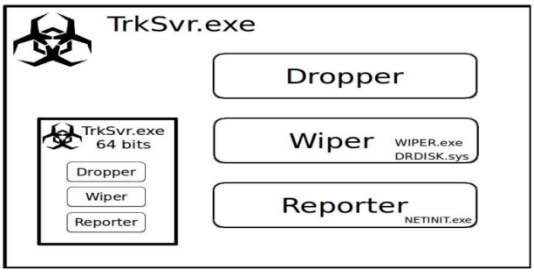


Figure 1: Shamoon's Malware Components

Wiper – It is responsible for destroying the files. Execution will be done only after obtaining hardcoded data. (E.g., Aug 15, 2012, Saudi Aramco attack).

Reporter – It is responsible for transferring infected information back to attacker. Information is sent in GET request. Though it is less sophisticated than other malwares like Stuxnet, Duqu, etc. But this malware has the potentiality of carrying out vast attack on a larger organization.

3.3 System Architecture

The proposed system's architecture – IDF-FPSO is depicted as a block diagram in figure 2. Two datasets have been used for training sets for Shamoon attack set and for different types of

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

attacks. Each dataset are represented with hash signatures. The primary goal of the proposed IDF-FPSO scheme is to correctly classify the Shamoon attack (a malware type under APT) from different attacks. In order to do so, certain modules has been employed. 1) Feature extraction, 2) Predicting best weighed feature set using FPSO Optimizer, 3) Classification using KNN classifier. These trained data are sent for feature extraction process, where each row possess 272 features. It will find frequency for single character and then find frequency for two characters. The extracted features including frequencies of single and two characters are sent to FPSO optimizer. In FPSO, the data is classified based on weight value. Thus the weight value is determined for each and every feature. If the determined weight > 0.6, then the weight is predicted. Meanwhile, the features are extracted for test data as well by applying PSO algorithm and weight is predicted accordingly. Then K- means clustering has been performed for both testing and training data**.** For the APT attack type called Shamoon type dataset, the centroid (C_1) possess two cluster classes ($c1_a$, $c1_b$). For the different attack type dataset, the centroid (C_2) possess two cluster classes ($C2_a$, $C2_b$). Finally, fitness function is evaluated to get the optimal weight. Using the separated classes, KNN classifier classifies the training and testing data and predicts the sample data point. The result is compared with other machine learning approaches and analyzed using various performance measures.

3.4 Proposed System Modules (investigator digital forensics frequencies particle swarm optimization for dectection and classification of apt attack in fog computing enviroment (IDF-FPSO)

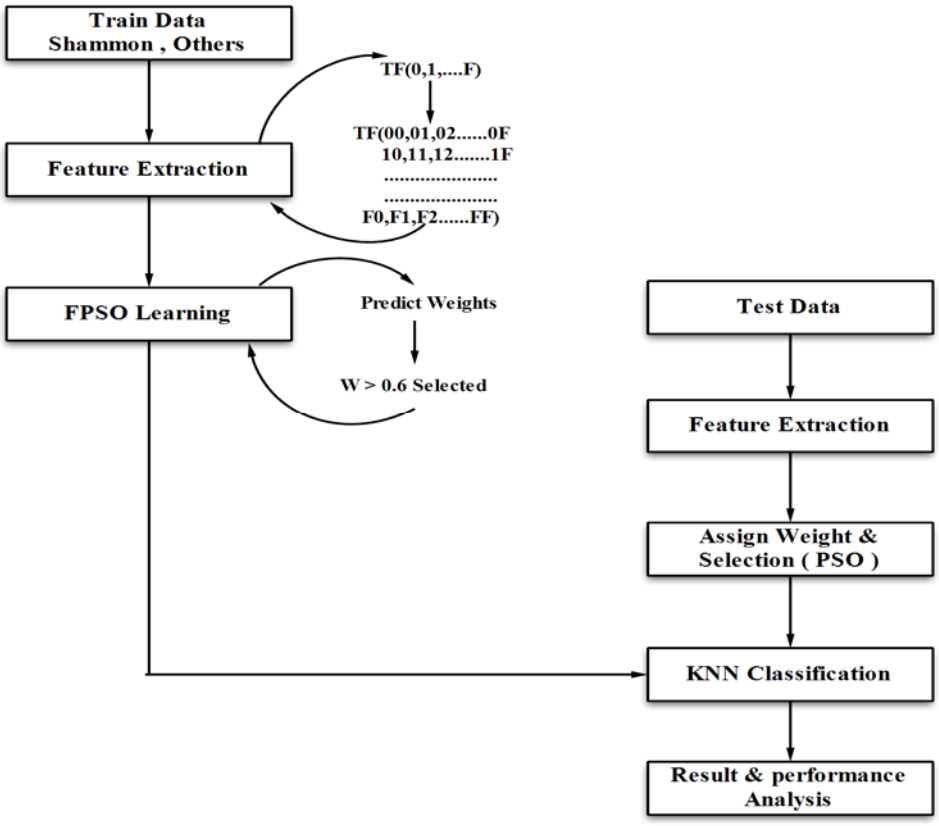


Figure 2: IDF-FPSO – Block Diagram

Journal of Theoretical and Applied Information Technology
15th April 2020. Vol.98. No 07

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.4.1 Feature Extraction:

The training set includes two datasets such as,

1) Dataset used for classifying APT attack type called Shamoon attack and

2) Dataset used for classifying different attack types.

Both datasets are in hash MD5 format type. Each row of feature set comprises of 272 features. In feature extraction, the first set features of the training data are selected in order to determine frequency term value. For a single character, the frequency value is calculated in the form of one digit.

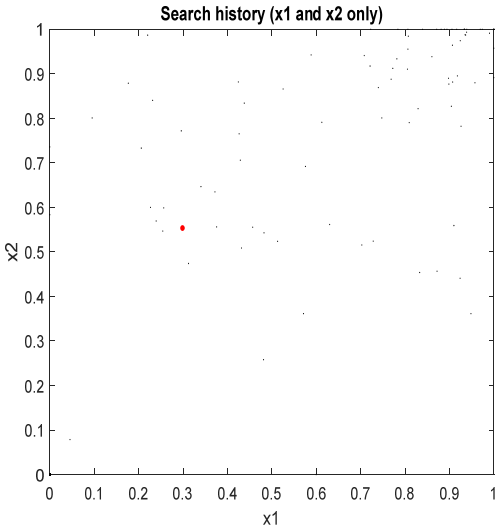


Figure 3: Search History of X1 and X2 Values.

3.4.2 FPSO Learning:

FPSO is a Frequency – Particle Swarm Optimization learning. It is a modified PSO, where the optimization can be done through Frequency, wave and sound. Usually, it mimics the characteristics of waves using 3 types of parameters such as frequency, amplitude, and wavelength. It is an alternative learning algorithm where it is utilized for resolving global optimization problem in a potential way. In selection process of FPSO, the best feature is selected by eliminating the worst feature. In this paper, FPSO is utilized selecting best weighted features by neglecting other weighted features.

As the feature extracted dataset consists of two types of data (i.e., one form Shamoon attack and remaining from other attack types), it is classified based on weight and feature selection properties. Since the dataset is in hash MD5 format, the encoded variable are determined first as below,

$0 \leq W \leq 1$

If the determined variables' weight value is greater than 0.6. (W >0.6), then those features are selected for Shamoon attack type. Remaining weight values can be neglected. Thus the size of selected feature set is reduced automatically. The resultant feature set will be,

$$X1 = W.X; \qquad (1)$$

Where X => features of training data
W => Weight value
X1=> New Features Value
$0 \leq W \leq 1^{min}$ *(A/B)*

Thus the best weighted (>0.6) features are predicted using FPSO. Meanwhile, testing dataset, which also contains two types of dataset undergoes feature extraction process and forwarded to PSO optimizer. The initialization parameters of the PSO optimizer utilizes the number of particle (count as 30). Here the fitness function, also referred as objective function, is used as the performance index of the particular population. It is noted that the higher the fitness value, the better the performance. The major aim of this proposed work is to provide the proposer fault decision. Thus, the fitness function should be evaluated in this case which fixes the level, the fault detection error that are ought to be minimized.

The application of FPSO algorithm in proposed is represented graphically in figure 4. From that graph, it can be observed that best cost value (best solution) can be calculated over iterations.

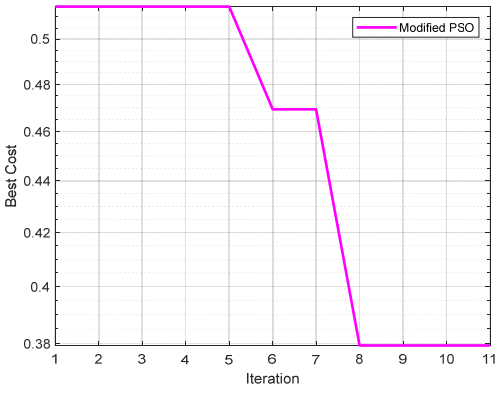


Figure 4: Determining the Best Weighed Variable

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.4.3 KNN Classification:

Before classification process, clustering process has been implemented. K-Means clustering has been used for clustering process, as it will divide the observations into k-clusters. As the amount of clusters are known, it can be used for data classification into clusters which possess equal or more than number of classes. Then the centroids created are classified into classes by using a classifier. Here KNN classifier has been utilized for classification of classes. In proposed, clusters are created for both Shamoon type features and other attack features. For each feature types, a centroid has been created along with two classes. On considering Shamoon feature set, the Shamoon Centroid (C_1) has two classes as ($c1_a$, $c1_b$). Similarly, the centroid for other attack feature set (C_2) has two classes as ($C2_a$, $C2_b$). The classifier results could be evaluated as below,

Minimum:

$$A = sum\ [|C1_a - C1_b|] + sum[|(C2_a - C2_b)| \quad (2)$$

Maximum:

$$B = sum\ [|C1_a - C2_a|] + sum[|(C1_b - C2_b)|] \quad (3)$$

3.4.4 Fitness Function:

The major goal of fitness function is to assess the quality of the solution. The input of this function is denotes the class with selected features. Then the KNN classifier is built by using this selected features. Finally, the output will be classifier's classification accuracy. The fitness function of the proposed scheme has been represented as below,

$$Fitness = \left(\frac{A}{B}\right) \quad (4)$$

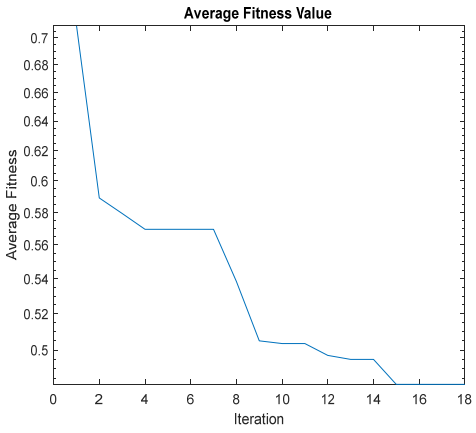


Figure 5: Average Fitness Value

The average fitness value determination has been represented graphically in figure 5. Form that graph, it has been observed that the average fitness value could be calculated over the number of iterations provided. Then find the fitness value, the fitness function is calculated based on controller to switch latency and inter-controller latency. And the controller to switch latency value is calculated based on the worst case latency and average latency, inter-controller latency is calculated based on the worst latency and average latency. The pseudo code for entire proposed scheme modules are provided below.

Pseudo Code:

***Input**: Read Shamoon Data and other Data Signatures*

***Output**: Classify the Signatures*

i) Preprocess

for i = 1 : NData
  freq (single char) = count1;
  freq (two char) = count2;
end
feature = [count1, count2]

ii) FPSO Learning:

Weight = rand (1, Feature size);
Weight (weight <0.6) = [ ];
Evaluate Fitness min (B/A);
Optimal weight (weight);*

iii) KNN Classify:

*Weight**
Train KNN Classifier;
Model = KNN Train (Train Data, Train label);
Predict = Classify (model, Test Data);

iv) Result:

Evaluate Performance: Accuracy, Sensitivity, Specificity, F-measure, G-mean

4. EXPERIMENTAL RESULTS

This section presents the metrics used for evaluating the proposed scheme using confusion matrix. And also evaluated the performance of the proposed IFD-FPSO by using various measures.

Evaluation of Classifier using Confusion Matrix:

The performance of the proposed scheme is based essentially on examining the confusion matrix that is obtained from classification of a testing subset, i.e. on a set of test data, the true values are known. The concept of this matrix is relatively simple to understand, but its terminologies could be confusing.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

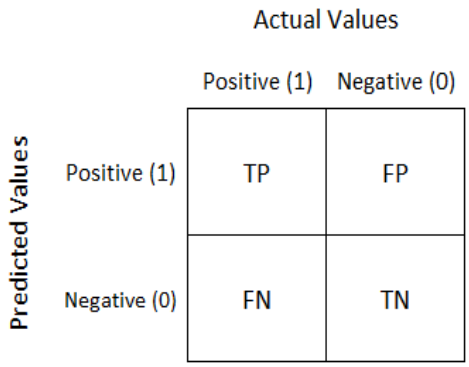


Figure 6: The Confusion Matrix

This matrix is utilized for measuring recall, precision, specificity, accuracy and AUC-ROC curve. Consider an example **confusion matrix for a binary classifier**. The actual and predicted values are listed in table as below

Table1: Actual and Predicted values

Actual	Predicted
0	0
0	0
0	1
0	0
0	0
1	0
1	0
1	1
1	1
1	1

TP = 3 (TP → True Positive)
TN=4 (TN→True Negative)
FP=1 (FP→False Positive)
FN=2 (FN→False Negative)

The above confusion matrix for binary classifier could be represented as,
0 → the Shamoon Attack Data
1 → the Normal Data

Classification Accuracy

It can be represented as:
Accuracy $=(TP+TN)/TP+TN+FP+FN$

Table 2: Notion of Classification accuracy

TP	TN	%
FN	FP	%
%	%	% Accuracy

There are few problems with accuracy. As it assumes equal prices for both errors. It can be said that the 99% of accuracy can be related to good, better, mediocre, terrible depends on the problem.

Recall

It can be calculated by defining as the ratio of perfectly classified positive variables divided by total number of all positive variables. The variable with high recall denotes that it is classified correctly.
It can be represented as below,

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

Precision

Precision can be obtained by dividing the total perfectly classified positive instances by the total number of predicted positive instances. An instance with high precision denotes as positive. (Small number of FP)
Precision can be represented as:

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

There might be two situations, as
1) High recall, low precision: Here, the positive instances are identified correctly along with lot of FP.
2) Low recall, high precision: Here, all the predicted values are positive instances. But at the same time, we missed lot of positive instances, while predicting. (High FN – Low FP).

F-measure:

It uses Harmonic mean, as it ignores the risky values. Its resultant value will always be nearer to small value of recall or precision.

$$F-Measure = \frac{2*Recall*Presicion}{Recall+Precision} \quad (7)$$

The confusion matrix of proposed methodology has been represented graphically in figure 7. In that figure, X-axis denotes targeted class and Y-axis denotes output/predicted class. It has 4

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

choices 00,01,10,11. By using this matrix, accuracy, precision, recall, specificity can be calculated. From this figure, the diagonal 2 (green color) indicates correctly classified sample, whereas red color indicates misclassified samples.

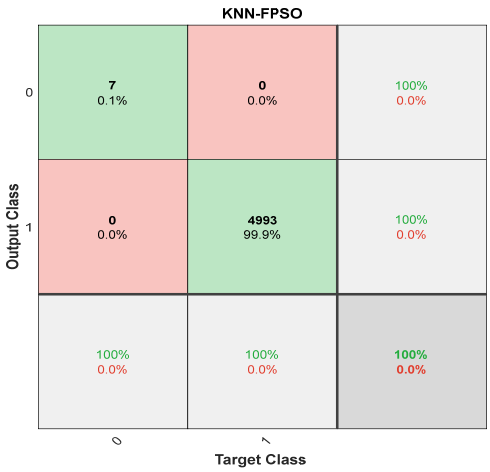


Figure 7: Confusion matrix for proposed work

White color indicates the accuracy of whole sample and grey color represents accuracy of particular class only.

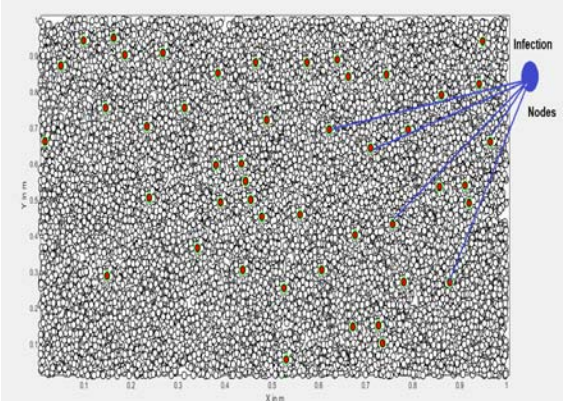


Figure 8: Shamoon Attack Capture Rate

The Shamoon attack identification has been graphically represented in figure 8. This graph contains both the attack data and normal data. In this graph, the Green with Red dot indicates correctly identified Shamoon attack and the Red dot indicates the misidentified Shamoon attack, and the Green dot indicates actual Shamoon attack. The major goal of the proposed scheme is to correctly classify the Shamoon attack from other malware attacks. As we can see that, out of all Shamoon attacks, only one or two got left behind (i.e. misidentified) whereas remaining got captured. From this graph, it is obvious that, most of the Shamoon attacks got captured accurately from all the other attacks.

5. COMPARATIVE ANALYSIS

This section presents the comparative analysis of the proposed scheme with few Machine Learning (ML) algorithm approaches in a detailed manner. The ML approaches used for comparison are SVM classifier, Naïve Bayes Classifier, KNN classifier, Decision tree.

5.1 IDF–FPSO Approach Versus SVM Classifier

SVM – Overview:

SVMs are considered as supervised learning models which are utilized for analyzing and recognizing patterns, during classification and regression analysis. It uses the concept of hyper planes in order to define the decision boundaries that are separated between data points of different classes. It handles both linear and non-linear classification. The major idea is to map original data points to high dimensional or infinite dimensional. A training dataset is considered as $\{x_i, y_i\}_{i-1}^{N}, with\ x_i \in R^d$ with input vectors and class labels$y_i \in \{-1, +1\}$. The SVM classifier can be formulated step by step,

$$w^T \varphi(x_i) + b \geq +1 \text{ For} y_i = +1, \quad (8)$$

$$w^T \varphi(x_i) + b \geq -1 \text{ For} y_i = -1,$$

Thus the classifier can be written as below,

$$f(x) = sign(w^T \varphi(x) + b) \quad (9)$$

SVM is a general algorithm based on assured risk bounds of statistical learning theory. It is an expected risk bound by the summation of Vapnik-Chervonenkis (VC) confidence. It can also be utilized for solving regression estimation, pattern recognition, and density estimation problems. It can also be applied in various applications like bioinformatics, database marketing, etc. The procedure followed in SVM are provided below,

Step 1: Implement all the provided training samples to train initial SVM, resulted in $l_1$support vectors along with respective decision functions,

$$\{SV_i^{l_n}, i = 1,2, \ldots, l_1\} \quad (10)$$

$$d_1(\vec{x}). \quad (11)$$

Step 2: Exclude certain support vectors from the training set, whose projections have largest curvatures, by

- Finding its projections
- Compute the generalized curvature

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Step 3: Retrain the remaining samples.

Result:

A confusion matrix has been drawn for SVM classifier which is used for comparing it with proposed scheme. It is represented in figure 9, where X-axis denotes the targeted class and Y-axis – predicted class. By using this matrix, various performance measures like accuracy, specificity, recall, and precision has been computed.

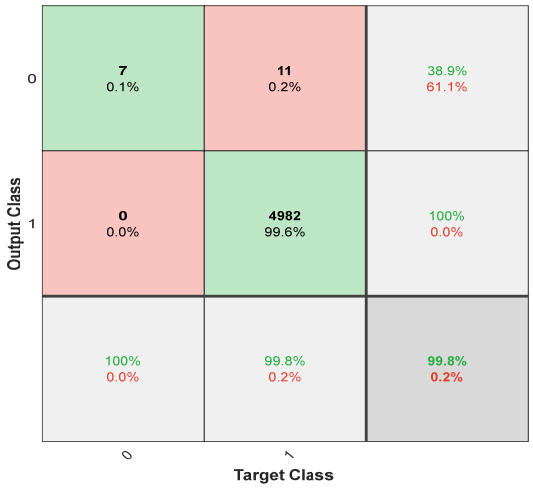


Figure 9: Confusion matrix – SVM classifier

Here the accuracy of SVM is 0.996 .However, the KNN-FPSO shows better values for specificity, precision, and G-Mean than existing SVM classifier.

5.2 IDF – FPSO Approach Versus Naïve-Bayes classifier:

Naïve-Bayes – Overview:

This technique is based on Bayesian theorem and is suited when the input dimensionality is higher. These classifiers handle a random numbers of independent variables. Let us assume, for a given set of variables, $X = \{x_1, x_2, \dots, x_d\}$ a posterior probability for the event C_j has been constructed for the set of outcomes $C = \{c_1, c_2, \dots, c_d\}$. So by using Bayes' rule,

$$p(C_j|x_1, x_2, \dots, x_d) \propto p(x_1, x2 \ , \dots, x_d|C_j)p(C_j) \qquad (12)$$

Where $p(C_j|x_1, x_2, \dots, x_d)$ is the posterior probability of class membership. As this classifier considers the conditional probabilities of the independent variables. Thus the likelihood of the product of terms would be,

$$p(X|C_j) \propto \prod_{l-1}^{d} p(x_k|C_j) \qquad (13)$$

Naïve Bayes will reduce the high-dimensionality density estimation task into 1-D density estimation. Moreover, the theory does not affect the posterior probabilities specifically in the regions nearer to decision boundaries leaving the classification task unaffected.

Result:

A confusion matrix has been drawn for Naïve Bayes classifier which is used for comparing it with proposed scheme. It is represented in figure 10, where X-axis denotes the targeted class and Y-axis – predicted class. By using this matrix, various performance measures like accuracy, specificity, recall, and precision has been computed.

Here the accuracy of Naïve Bayes is 0.248, It proves the KNN-FPSO classifier have better values for specificity, precision, and G-Mean than existing classifier.

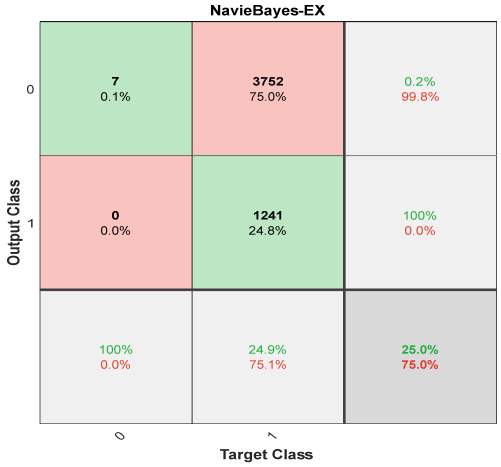


Figure 10: Confusion matrix – Naïve Bayes classifier.

5.3 IDF – FPSO Approach versus Decision Tree:

Decision Tree – Overview:

A decision tree can be defined as a decision support tool which utilizes a model or tree-like graph and their possible consequences, including costs and utility. It is more like a flowchart where each and every internal node

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

represents a test attribute, each branch denotes the outcome of the test, and each and every leaf node denotes a class label and the path from root to leaf represents the classification rules. These trees are closely associated with influence diagram and are utilized in analytical and visual decision, where the predicted values are computed. It has 3 types of nodes, namely 1) Decision nodes (Squares) 2) Chance nodes (Circles), 3) End nodes (triangles).

Decision tree model:

Decision Tree Model:

In decision tree setting, the algorithm can be viewed as a Boolean function of,

$f: \{0,1\}^n \rightarrow \{0,1\}$, where input is the series of questions and output will be the decisions.

Result:

A confusion matrix has been drawn for Decision tree which is used for comparing it with proposed scheme. It is represented in figure 11, where X-axis denotes the targeted class and Y-axis – predicted class. By using this matrix, various performance measures like accuracy, specificity, recall, and precision has been computed.

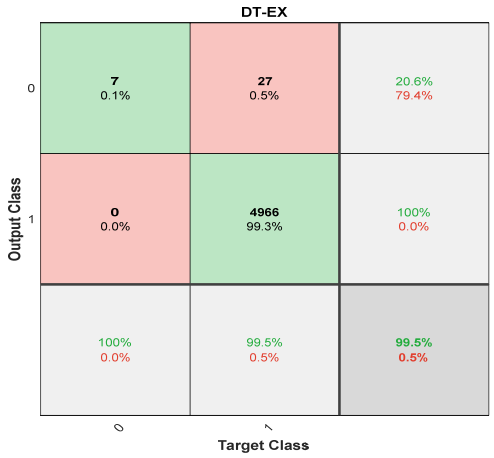


Figure 11: Decision Tree Based Confusion Matrix

Here the accuracy of Decision tree is 0.993, the proposed KNN-FPSO classifier shows better values for sensitivity, and specificity, and precision, Recall, F-Measure, and G-Mean than existing classifier.

5.4 IDF – FPSO Approach versus KNN classifier:

KNN is considered to be the classification algorithm in ML. As it is parametric, it is widely reusable in real-life applications. The procedure of the KNN classifier is provided below, where m is the number of training samples and p is an unknown point.

1. The training samples are stored in array arr[].for i=0 to m:

Compute the Euclidean distance d(arr[i],P).

2. Make sets S of K from the smallest distances. Each of these distances assiciated to an already classified data point.
3. Return the majority label among S.

Results

A confusion matrix has been drawn for KNN which is used for comparing it with proposed scheme. It is represented in figure 12, where X-axis denotes the targeted class and Y-axis – predicted class. By using this matrix, various performance measures like accuracy, specificity, recall, and precision has been computed.

Here the accuracy of KNN is 0.968, which is lesser than KNN-FPSO classifier. Also, the proposed classifier shows better values for sensitivity, specificity, and precision, Recall, F-Measure, and G-Mean than existing classifier.

Table 3 shows the comparison between KNN-FPSO classifier and other existing classifiers using various performance measures like, precision, sensitivity, specificity, recall, accuracy, F-measure, G-Mean.

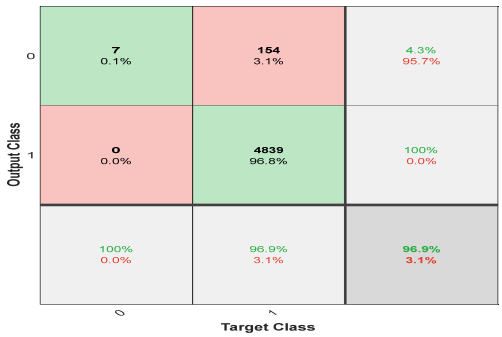


Figure 12. KNN Classifier Based Confusion Matrix

From that table, shows the comparison between KNN-FPSO classifier and other existing classifiers such as Existing KNN, Decision tree,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Naive bayes and SVM, by using various performance measures like, precision, sensitivity, specificity, recall, accuracy, F-measure, G-Mean. Moreover, the sensitivity, and specificity, recall, precision, and finally G-mean, F-measure values of proposed are remarkable higher when compared to other existing classifiers.

are remarkable higher when compared to other existing classifiers. It is represented graphically in figure 13.

The Table 3 has been represented graphically in figure 13. In that graph, the test cases like accuracy, sensitivity, specificity, Precision, Recall, and F-Measure has been placed in X-axis and values in Y-axis. This comparative graph compares proposed KNN classifier (red color) with conventional classifiers like exisiting KNN, decision trees, Naïve Bayes, and SVM. The proposed KNN classifier shows remarkable accuracy (0.999998) attainment when compared

with existing KNN classifier (0. 9692). i.e. the proposed KNN classifier has 0.8% higher than exisiting KNN. Similarly for other metric like specificity, precision, recall and F-Measure, the proposed system shows remarkably significant value than other existing classifiers.

From figure 14, it has been noted that, the prediction time of proposed KNN is higher when compared to classifiers like DT, NB, and SVM. However on comparing proposed KNN-FPSO with existing KNN, the prediction time of proposed seems to be remarkable lower. This proves that KNN-FPSO is better than existing KNN classifier.

Table 3: Comparison Average 5 Iterations Of Various Classifiers With KNN-FPSO Classifier Using Performance Measures

Measures Classifiers	Accuracy	Sensitivity	Specificity	Precision	Recall	F-Measure	G-Mean
FPSO- KNN	0.99998	0.99998	1.0000	1.0000	0.99998	0.99996	0.9997
Existing KNN	0.9692	0.9692	1.0000	1.0000	0.9692	0.9843	0.9845
Decision Tree	0.9946	0.9946	1.0000	1.0000	0.9946	0.9973	0.9973
Naïve Bayes	0.2496	0.2485	1.0000	1.0000	0.2485	0.3981	0.4985
SVM	0.9978	0.9978	1.0000	1.0000	0.9978	0.9989	0.9989

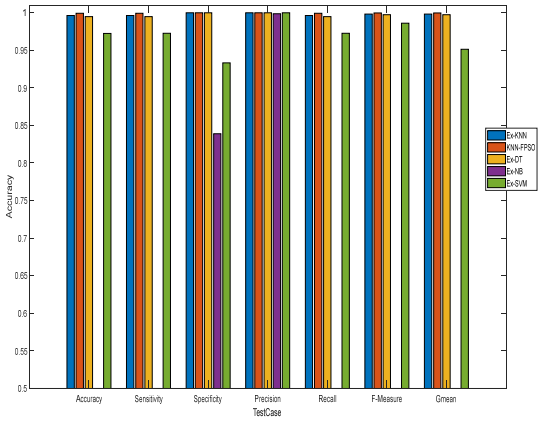
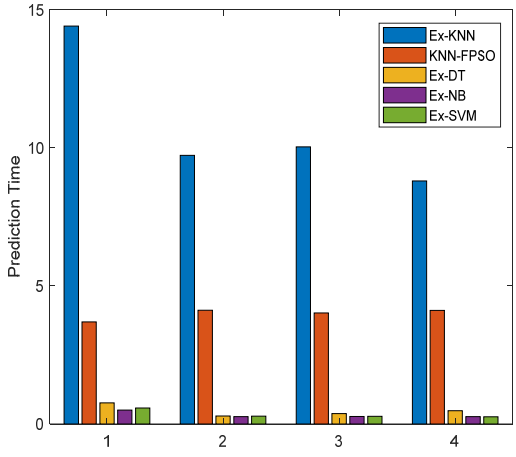


Figure 13: Comparison Of Various Classifiers With KNN-FPSO Classifier Using Performance Measures.



Figur 14: Comparison Of Existing Classifiers With KNN-FPSO Classifier Based On Prediction Time

The main aim of these classifiers and optimizers is to have reduction in prediction and training time. The KNN-FPSO classifier is supposed to train with reduced feature subset and compared with

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

other classifiers in order to evaluate the reduction in time for training. The training time of KNN-FPSO compared with other existing classifiers are represented graphically in Figure 15.

From figure 15, it has been noted that, the KNN-FPSO classifier has very low training time (< 0.1). It is even lesser than existing KNN. Thus, it is obvious that KNN-FPSO classifier resulted in low time consumption, when compared to other classifiers, which means proposed KNN-FPSO can be easily trained than existing KNN.

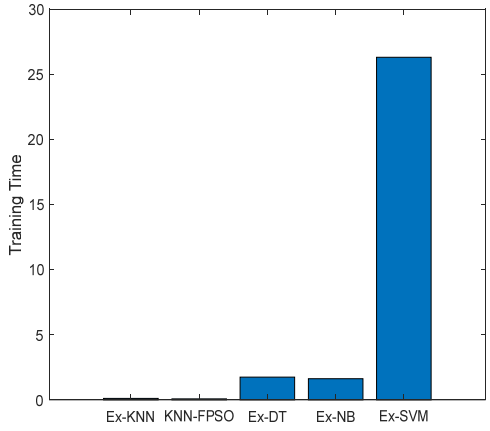


Figure 15: Comparison Of Existing Classifiers With KNN-FPSO Classifier Based On Training Time

6. CONCLUSIONS

TThe actual performance of proposed IDF-FPSO has been explained by using two various types of datasets namely 1) Dataset used for classifying APT attack type called Shamoon attack, 2) Dataset classification of different attack types. The Frequency based FPSO algorithm[24] has been utilized in this proposed scheme. As FPSO is renowned for solving the search space optimizations and providing finest solution, the proposed scheme utilized its feature for the betterment. In first module, the features from two data set has been extracted, in order to reduce the size of feature set. The best weighed features has been predicted as best solutions using FPSO (W > 0.6). The KNN classifier has been utilized in order to cluster the best classes and the best fitness function has been evaluated.

The performance of proposed IDF-FPSO has been evaluated using confusion matrix and through comparisons with various existing classifiers (SVM, KNN, Naïve Bayes, Decision trees). On evaluating its performance through confusion matrix, the values of accuracy, specificity, sensitivity, recall, precision, F-Measure, G-Mean has been obtained. KNN-FPSO classifier managed to come closer to that accuracy (0.99998).

Moreover, the performance of the KNN-FPSO classifier is further evaluated using specificity, sensitivity and precision, where it shown remarkable recognition than other conventional classifiers. Due to this remarkable performance by KNN-FPSO, the system has perfectly identified the Shamoon attacks from different attacks. Furthermore, the training time and prediction time of proposed with existing classifiers are compared and evaluated. From the result, it has been obvious that the prediction time is remarkable lower than existing KNN classifier and it can be easily trained, when compared to other classifiers.

The main objective was to preserve evidence without compromising investigation structure of collecting, identifying and validating digital data in order to reconstruct historical events.

To overcome problems, IDF-FPSO an investigator digital forensic algorithm was utilized to detect and categorize advanced persistent and Shamoon attacks in fog environment.

As the fog computing had been progressed, then it is necessary to improve detection and mitigation of malware attack methods. An original performance of proposed investigator digital forensic had been elucidated by utilizing two types of datasets such as advanced persistent and Shamoon attack that could be extracted to optimize the set of feature size. An optimal weighted feature solution could be predicted by utilizing frequency particle swarm optimization. A k-nearest neighbor was used to group best classes and at last fitness function had been estimated.

After that, the performance of investigator digital forensic-frequency particle swarm optimization (IDF-FPSO) could be calculated by utilizing confusion matrix over several current classifiers such as support vector machine, k-nearest neighbor, naïve bayes and decision trees. The values for accuracy, specificity, sensitivity, recall, precision, F-measure and G-mean could be acquired and evaluated by performing confusion matrix method. This maximum recorded accuracy result could be attained by utilizing support vector machine. A k-nearest neighbor method had been utilized to achieve maximum accuracy.Due to the performance of proposed classifier, the malware system had been detected Shamoon attacks perfectly from different attacks. Additionally, the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

training and prediction time of proposed with existing classifiers are compared and evaluated. From the result, it had been observed that prediction time would be less remarkable than existing k-nearest neighbor that could be trained effortlessly when equated with other classifiers.

REFERENCES:

[1] M. Mukherjee, R. Matam, L. Shu, L. Maglaras, M. A. Ferrag, N. Choudhury, *et al.*, "Security and privacy in fog computing: Challenges," *IEEE Access,* vol. 5, pp. 19293-19304, 2017.

[2] M. Conti, A. Dehghantanha, K. Franke, and S. Watson, "Internet of Things security and forensics: Challenges and opportunities," ed: Elsevier, 2018.

[3] S. Alelyani and H. Kumar, "Overview of Cyberattack on Saudi Organizations," 2018.

[4] P. Sangle, R. Deshmukh, R. Ghodake, A. Yadav, and J. Musale, "Data Security System in Cloud by Using Fog Computing and Data Mining," *International Journal,* vol. 7, 2017.

[5] M. T. Dong and X. Zhou, "Fog computing: Comprehensive approach for security data theft attack using elliptic curve cryptography and decoy technology," *Open Access Library J,* vol. 3, p. 1, 2016.

[6] S. Khan, S. Parkinson, and Y. Qin, "Fog computing security: a review of current applications and security solutions," *Journal of Cloud Computing,* vol. 6, p. 19, 2017.

[7] F. Hosseinpour, P. Vahdani Amoli, J. Plosila, T. Hämäläinen, and H. Tenhunen, "An intrusion detection system for fog computing and IoT based logistic systems using a smart data approach," *International Journal of Digital Content Technology and its Applications,* vol. 10, 2016.

[8] R. K. Naha, S. Garg, D. Georgakopoulos, P. P. Jayaraman, L. Gao, Y. Xiang, *et al.*, "Fog Computing: Survey of trends, architectures, requirements, and research directions," *IEEE access,* vol. 6, pp. 47980-48009, 2018.

[9] S. Zahra, M. Alam, Q. Javaid, A. Wahid, N. Javaid, S. U. R. Malik, *et al.*, "Fog computing over iot: A secure deployment and formal verification," *IEEE Access,* vol. 5, pp. 27132-27144, 2017.

[10] S. S. Harish Kumar, Pratvina Talele, "Secure Fog Computing System using Emoticon Technique," *ijritcc,* p. pp. 808, 2017.

[11] S. Chen, T. Zhang, and W. Shi, "Fog computing," *IEEE Internet Computing,* vol. 21, pp. 4-6, 2017.

[12] R. Rani and P. Lourdu Sravani, "Challenges of Digital Forensics in Cloud Computing Environment," *Indian Journal of Science and Technology,* vol. 9, 05/25 2016.

[13] V. Jusas, D. Birvinskas, and E. Gahramanov, "Methods and tools of digital triage in forensic context: survey and future directions," *Symmetry,* vol. 9, p. 49, 2017.

[14] A. Joseph and K. Singh, "A SURVEY ON LATEST TRENDS AND CHALLENGES IN CYBER FORENSICS International Journal of Advances in Electronics and Computer Science, ISSN: 2393 - 2835," *International Journal of Advances in Electronics and Computer Science, ISSN: 2393 - 2835,* vol. http://ijaecs.iraj.in/volume.php?volume_id=295#, 09/01 2016.

[15] N. Rana, G. Sansanwal, K. Khatter, and S. Singh, "Taxonomy of Digital Forensics: Investigation Tools and Challenges," *arXiv preprint arXiv:1709.06529,* 2017.

[16] A. Rot and B. Olszewski, "Advanced Persistent Threats Attacks in Cyberspace. Threats, Vulnerabilities, Methods of Protection," in *FedCSIS Position Papers*, 2017, pp. 113-117.

[17] P. Chen, L. Desmet, and C. Huygens, "A study on advanced persistent threats," in *IFIP International Conference on Communications and Multimedia Security*, 2014, pp. 63-72.

[18] L. Herløw and S. J. Hansen, "Detection and Prevention of Advanced Persistent Threats."

[19] M. Ask, P. Bondarenko, J. E. Rekdal, A. Nordbø, P. Bloemerus, and D. Piatkivskyi, "Advanced persistent threat (APT) beyond the hype," *Project Report in IMT4582 Network Security at GjoviN University College,* 2013.

[20] N. Kawaguchi, H. Tomimura, T. Komiyama, K. Kubota, and M. Tsuichihara, "Locating victims of destructive targeted attacks based on Suspicious Activity Spike Train," in *2017 IEEE Symposium on Computers and Communications (ISCC)*, 2017, pp. 871-878.

[21] S. Zhioua, "The middle east under malware attack dissecting cyber weapons," in *2013 IEEE 33rd International Conference on Distributed Computing Systems Workshops*, 2013, pp. 11-16.

[22] D. Tarakanov, "Shamoon the Wiper in details," *Kaspersky Labs,* 2012.

[23] A. Al Hwaitat, S. Manaseer, &R. Al-Sayyed,. (2019). A Survey of Digital Forensic

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Methods under Advanced Persistent Threat in Fog Environment, Journal of Theoretical and Applied Information Technology , Vol.97. No 18,PP. 4934-4954. 26.

[24] A. Al Hwaitat, S. Manaseer ,(2018), Centralized Web Application Firewall Security System, Modern Applied Science; Vol. 12, No. 10; PP.164-170. https://doi.org/10.5539/mas.v12n10p164 27.

[25] A. Al Hwaitat, S. Manaseer and R. Jabri ,(2018) ,Distributed Detection and prevention of Web Threats in Heterogeneous Environment , Modern Applied Science; Vol. 12 ,No10 ,PP.13-22. https://doi.org/10.5539/mas.v12n10p13 28.

[26] A. Al Hwaitat, S. Manaseer ,(2017), Validation and Integrity Mechanism for Web Application Security, International Journal of Engineering Research & Science , Vol. 3, No.11,PP.34-38. 29.

[27] O. rababha , A. Al Hwaitat , S. Manasser ,(2016) Web Threats Detection and Prevention Framework, communications and Network, Vol. 8, No.8, PP. 170- 178. https://doi.org/10.4236/cn.2016.83017.

[28] A. Al Hwaitat , M. Qasem ,R. Fabozzi ,(2020)," Security of Data Access in Fog Computing using Location-based Authentication", International Journal of Advanced Trends in Computer Science and Engineering ,Vol. 9,No. 1 ,PP.247-253.

[29] A. Al Hwaitat , M. Qasem ,(2020)," A Survey on Li Fi Technology and Internet of Things (IOT)", International Journal of Advanced Trends in Computer Science and Engineering ,Vol. 9,No. 1 ,PP.225-253

[30] A. Hudaib , A. Al Hwaitat ,(2018), " Movement Particle Swarm Optimization Algorithm" Modern Applied Science; Vol. 12, No. 1,pp.148-164.

[31] A. Al Hwaitat, S. Manaseer, R. Al-sayyed3 ,m almaiah, o almomani ,(2020) , A Threat Intelligence Scheme For Incident Reporting And Investigation Of Shamoon Attack Behaviour In Fog Computing, Journal of Theoretical and Applied Information Technology, Vol.98 ,Issues 07