Received: September 20, 2020. Revised: October 23, 2020. 192

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Feature Selection Models Based on Hybrid Firefly Algorithm with Mutation Operator for Network Intrusion Detection

Karrar Mohsin Alwan[1]* **Ahmed H. AbuEl-Atta[1]** **Hala Helmy Zayed[1]**

[1]*Department of Computer Science, Faculty of Computers and Artificial Intelligence, Benha University, Benha, Egypt*
* Corresponding author's Email: kr.moh93@yahoocom

Abstract: Accurate intrusion detection is necessary to preserve network security. However, developing efficient intrusion detection system is a complex problem due to the nonlinear nature of the intrusion attempts, the unpredictable behaviour of network traffic, and the large number features in the problem space. Hence, selecting the most effective and discriminating feature is highly important. Additionally, eliminating irrelevant features can improve the detection accuracy as well as reduce the learning time of machine learning algorithms. However, feature reduction is an NP-hard problem. Therefore, several metaheuristics have been employed to determine the most effective feature subset within reasonable time. In this paper, two intrusion detection models are built based on a modified version of the firefly algorithm to achieve the feature selection task. The first and, the second models have been used for binary and multi-class classification, respectively. The modified firefly algorithm employed a mutation operation to avoid trapping into local optima through enhancing the exploration capabilities of the original firefly. The significance of the selected features is evaluated using a Naïve Bayes classifier over a benchmark standard dataset, which contains different types of attacks. The obtained results revealed the superiority of the modified firefly algorithm against the original firefly algorithm in terms of the classification accuracy and the number of selected features under different scenarios. Additionally, the results assured the superiority of the proposed intrusion detection system against other recently proposed systems in both binary classification and multi-classification scenarios. The proposed system has 96.51% and 96.942% detection accuracy in binary classification and multi-classification, respectively. Moreover, the proposed system reduced the number of attributes from 41 to 9 for binary classification and to 10 for multi-classification.

Keywords: Intrusion detection system, Feature selection, Metaheuristics, Firefly algorithm, Mutation operator.

1. Introduction

An intrusion detection system (IDS) is usually employed to provide efficient security for information and communication systems. The IDS mainly focus on the detection of traffics that are dangerous to a network. An IDS has a similar functionality to other security approaches such as firewalls and antivirus software and can also have access to the control schemes[1-3]. The classification of IDS as presented in [4] is depicted in Fig. 1. Generally, IDS are classified either signature-based or anomaly-based detection systems. The signature-based IDS can identify the pattern of traffic or application data as dangerous and will require its database to be updated to store the signature of the identified attack. On the other side, anomaly-based IDS detect anomalies by comparing all activities against an established defined behavior. The IDS are mainly designed to detect network attacks and for subsequent notifications when a system is under attack.

The current IDS systems are not completely accurate in their detection ability; hence, this study focuses on improving and increasing the performance accuracy of the IDS systems. IDS can detect abnormal network patterns via analyzing network packets. Therefore, the machine learning (ML) methods are normally used for abnormal traffic pattern identification. The ML techniques for attack

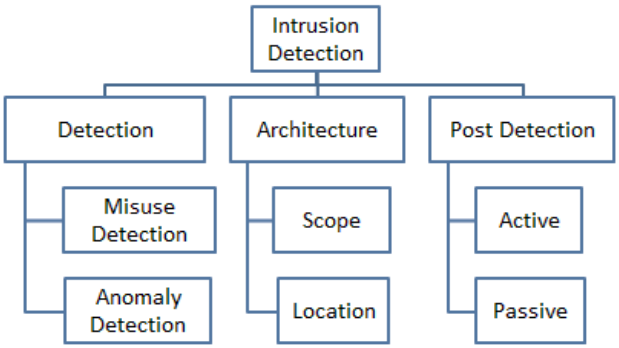

Figure. 1 The classification of the IDS

detection are mainly of two types; these are classification-based and clustering-based ML methods. The effectiveness of some of these ML methods is lost when faced with large data volume.

An IDS system is assumed effective if it is adaptable and accurate in attack detection. However, the performance of IDS is highly dependent on the performance of the used ML method which in turn highly affected by the quality of the used feature vectors. There may be some irrelevant features of the training data which do not contribute to the process of detection. In addition, in most cases these irrelevant features may be redundant or introduce noise into the design of the classifier. Furthermore, the presence of redundant and irrelevant features in such large data sets requires the use of feature selection algorithms to eliminate such irrelevant features. It is necessary to determine the most significant features which enhance the performance of the classifiers[5]. Another problem is the high data dimensionality that needs being reduced to make them tractable. The methods of data dimensionality reduction offer a way of better data understanding, enhancing the performance of prediction and computational time reduction in applications for pattern recognition.

Feature selection or reduction is a data pruning process that selects the important attributes and removes the irrelevant ones from a given dataset to formulate a better model for data learning. The pruned data set retains a good and accurate representation of the entire data features which is enough for data description [6-8]. However feature selection process is an NP-hard problem. Therefore, metaheuristics can be effectively employed to provide near optimal solution within reasonable time. In this paper, a metaheuristic algorithm called the Firefly Algorithm (FA) is employed to perform the feature reduction process to enhance the performance of IDSs. A mutation operator is adopted to enhance the original firefly algorithm. Also, the proposed method is validated using Naïve Bayes classifier on the "NSL-KDD" dataset in terms of the suitable performance metrics. The contributions of this paper can be summarized as follows:

- Proposing a modified version of the firefly algorithm with enhanced exploration ability by adopting the mutation operator.
- Proposing an intrusion detection system that combines the modified firefly algorithm and the Naïve Bayes classifier.
- Achieve better classification accuracy compared to the standard firefly algorithm and other intrusion detection approaches using a benchmark dataset.

The rest of this paper is arranged as follows: Sec. 2 outlines the related works. Sec. 3 explains the suggested feature selection method. Sec. 4 presents the conducted experiments and the obtained results. Finally, Sec. 5 contains the conclusion and future work suggestions.

2. Literature review

Due to the large number of attributes contained in network traffic records, many optimization algorithms have been proposed to reduce the number of these attributes. Hadeel et al. [9] have presented a pigeon inspired optimizer-based wrapper feature selection method for IDS. The suggested algorithm uses the cosine method instead of the sigmoid method to binarize a continuous pigeon inspired optimizer. It has been assessed by utilizing three well-known datasets: KDDCUP99, NLS-KDD, and UNSW-NB15. The obtained results revealed that their algorithm is better than many state-of-the-art feature selection algorithms in the literature. However, there is a large room for improvement regarding the detection accuracy and the number of selected features. Soodeh et al. [10] have suggested a hybrid intrusion detection model that consists of two stages: feature selection and attack detection. Feature selection process was performed using Support Vector Machine (SVM) and Genetic Algorithm (GA). On the other side, attack detection process was performed using an Artificial Neural Network (ANN). The used classifier was trained using a hybrid algorithm that combined both the Hybrid Gravitational Search (HGS) and the Particle Swarm Optimization (PSO). The efficacy of the suggested method has been assessed using NSL-KDD dataset. However, our proposed system is highly simpler than their system with a very competitive performance. Muhammad et al. [11] have introduced a hybrid feature selection that consists of both a filter and a wrapper. The suggested model adopted the Correlation Feature Selection (CFS) technique with

the help of three approaches namely the best-first, the greedy stepwise, and the GA for the search process. The features were first selected using the filter method. Then, the wrapper was used to assess the selected features with the help of a Random Forest (RF) classifier. The proposed model was tested on both KDD99 and DARPA1999 datasets. The obtained results revealed that the proposed feature selection model has a satisfactory performance. However, more efforts are needed to improve the obtained results, particularly, in terms of the number of selected features. Another filter and wrapper based feature selection method has been proposed by Selvakumar et al. [12]. The suggested method employed the firefly algorithm (FA) to determine the selected features. Two classifiers were employed namely, the C4.5 decision tree and the Bayesian Networks (BN). The suggest method was tested using the KDD CUP 99 dataset and the obtained results revealed that only ten attributes were enough to obtain high accuracy. However, the original FA executes the local search more than global search, making it highly prone to local optima entrapment. A new hybrid model was suggested by Mehrnaz et al. [13] for anomaly-based IDS. The suggested model aimed at achieving high detection rate and decreasing the false-positive rate. Feature selection process was performed using the Artificial Bees Colony (ABC) algorithm while the AdaBoost classifier was employed to assess the proposed method on the NSL-KDD and ISCXIDS2012 datasets. The achieved performance was comparable to legendary methods. However, more efforts are needed to improve the classification accuracy. Kumar et al. [14] have introduced a feature selection method called Multi-Linear Dimensionality Reduction (ML-DR). The suggested method aimed at decreasing both the features dimensionality as well as the training time. It was adopted a Multi-class SVM (M-SVM) which is used as a multi-attack classification to detect whether the action is a normal or abnormal. The NSL-KDD dataset was utilized to assess the proposed approach. However, Unauthorized Access from a Remote Machine (R2L) and Unauthorized Access to Local Super User (U2R) attacks have modest classification accuracy. Thaseen et al. [15] have presented a feature selection method based on combining chi-square and M-SVM. The suggested scheme aimed at improving the network attacks' classification accuracy. The NSL-KDD dataset was utilized to assess the proposed approach. However, the false alarm rate needs to be reduced. Further work was presented by Desale et al.[16]. They have introduced a feature selection method based on the GA. A Naïve Bayes (NB) classifier has been utilized to assess the quality of selected features. The obtained results using the NSL-KDD dataset revealed that the proposed scheme has achieved good accuracy and minimized the number of selected features. However, the classification accuracy needs to be improved. Malik et al. [17] aimed to detecting the network PROBE attacks through developing a binary variant of the multi-objective PSO which attempted to balance between the intrusion detection rate and false positive rate. A RF classifier was utilized to assess the significance of the selected features from the KDD99Cup dataset. However, the proposed system is mainly designed to only detect the PROBE attack. Schiezaro et al. [18] investigated, implemented, and analyzed a feature selection method using the ABC algorithm to classify network attacks using different datasets. However, the performance of the proposed feature selection method has not been evaluated using large datasets in intrusion detection scenario. Finally, Srinoy et al. [19] have combined the PSO and the SVM to present an efficient feature selection method. The proposed scheme adopted the one-versus-rest method as a utility function for the classification problem. The model was evaluated against KDD'99 dataset and the result shows that the model has obtained high classification accuracy. However, R2L and U2R attacks have modest classification accuracy.

3. The proposed methodology

3.1 Standard firefly algorithm (FA)

The firefly algorithm (FA) was first invented by Yang [20-21] as a nature-based global stochastic optimization approach that was inspired by the firefly population. Each firefly in the search space is considered a potential solution. The concept of the FA relies on the mating and light flashing pattern of the fireflies; the light flashing pattern of the fireflies serves as their information exchange mechanism. This subsection presents the main characteristics of natural fireflies, the artificial FA, and the variations that have been proposed earlier to the basic algorithm [22-24]. The behavior of the artificial fireflies can be described based on three idealized rules as proposed by [20]. These rules are as follows:

1. Fireflies have no gender differences (unisex); hence they are attractive to each other without any form of sex consideration.
2. The level of attractiveness of an individual firefly is determined by the intensity of the light it is emitting; hence, fireflies that emit brighter lights will naturally attract those that emit lesser light intensity. However, the level of

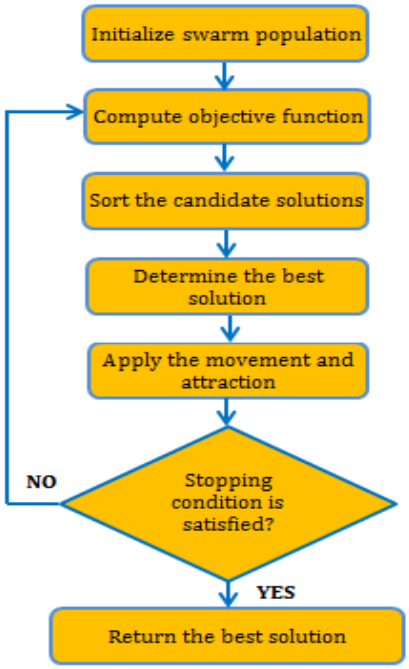

Figure. 2 The flowchart of the original FA

attractiveness of each firefly depends on their closeness to each other since light intensity relates inversely with distance. Fireflies will move randomly when there is no nearby firefly with a brighter light intensity.

3. The factor that determines the level of brightness of the light emitted by the fireflies is the landscape of the fitness function. Hence, the brightness of the emitted could be proportional to the value of the fitness function upon problem maximization.

Based on these idealized rules, the main steps of the FA are shown in Fig. 2.

3.2 The proposed models

Since the development of the FA, it has been utilized to solve many forms of optimization-related problems as evidenced in the literature. Although the original FA performs well in many applications, it is still prone to certain problems, such as the inability to achieve a good balance between global and local searches. The original FA executes the local search more than global search, making it highly prone to local optima entrapment. Hence, this study tries to improve the global search ability of the original FA by introducing the mutation operator of the Genetic algorithm into the original FA.

Two models have been built, one for binary-classification and the other for multi-classification. Both models used a modified version of the firefly algorithm with a mutation operator. The Naïve

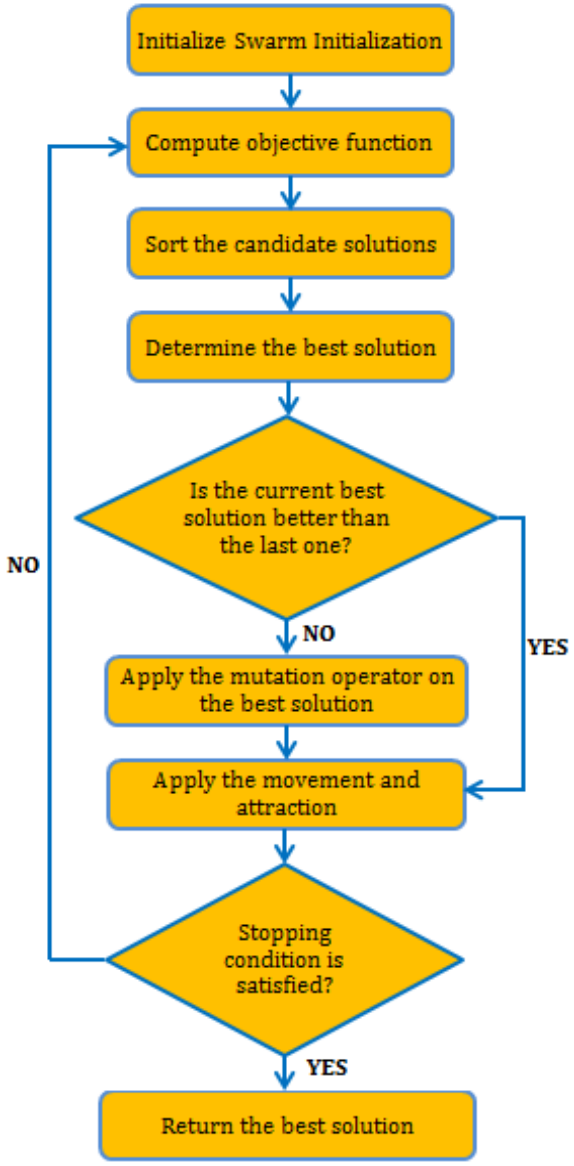

Figure. 3 The flowchart of the modified FA

Bayesian classifier (NBC) has been used to evaluate the selected features.

The best firefly in the original FA does not move as all the other fireflies are meant to move closer to it. Any failure of the algorithm in finding an improved position upon several iterations will affect the performance of the FA; hence, the GA mutation operator will help the FA by facilitating the random movement of the best firefly in the population towards finding a new position. This will imply random updating of the best-found solution using the GA mutation operator, thereby improving its ability to search for improved positions and consequently enhance the algorithmic performance. The flowchart of the modified firefly is shown in Fig. 3.

The following are the main steps of the suggested feature selection algorithm.

i. Swarm initialization

This step involves random initialization of each firefly in the population. This initialization is done in a continuous domain using a uniform distribution as shown by Eq. (1).

$$X_i = (UB - LB) \times Rand + LB \quad (1)$$

where *Rand* is a random variable bounded by the interval [0, 1], and *UB* and *LB* represent the upper bound (1, 0) and lower bound (0, 0), respectively.

To address the problem of features selection, each firefly is first encoded by a binary representation. Hence, the sigmoid function, which is defined by Eq. (2), is first used for the conversion of the continuous values into binary (0s and 1s), where the selected features are coded as 1, while the non-selected features are coded as 0.

$$\mathrm{B_i} = \begin{cases} 1 & \frac{1}{1+\mathrm{e}^{-\mathrm{X_i}}} > Rand \\ 0 & \text{Otherwise} \end{cases} \quad (2)$$

where X_i and B_i are the continuous and binary value for the positions of the firefly, respectively.

ii. Fitness function calculation

The fitness function aids in guiding the search via assigning a quality value to any potential solution. The objective function relies on both the accuracy and number of features for the evaluation of the solutions using Eq. (3).

$$\min \mathrm{f(x)} = (100 - \mathrm{Accuracy}) \quad (3)$$

The accuracy is obtained using a Naïve Bayesian classifier (NBC), which can be calculated based on the probability of the features. The predicted class is calculated based on the output of the maximum probabilities for all possible values. The next step after calculating the error rate, the light intensity for every Firefly is calculated by using Eq. (4).

$$\mathrm{I(F}_i) = \frac{1}{1+error^2} \quad (4)$$

iii. Distance calculation

The range from f_i to f_j is expressed as f_{ij} and described using Eq. (5).

$$f_{ij} = \| X_i - X_j \| = \sqrt{\sum_{d=1}^{D} (X_{id} - X_{jd})^2} \quad (5)$$

where X_{id} is the position of firefly i. The Euclidean distance method has been used to calculate the distance between any two fireflies. In the suggested work, D represents the total number of features related to network intrusion detection which is 41 features.

iv. Attractiveness

For every Firefly, the attractiveness β can be calculating by using Eq. (6).

$$\beta_f = \beta_0 e^{-\gamma f^2} \quad (6)$$

where f symbolizes the range between two Fireflies and ($\beta_0 = 1$) symbolize the attractiveness at $r = 0$ (first attractiveness).

v. Updating the position of fireflies

The fireflies in the population move to other fireflies with greater light intensity based on Eq. (7); this implies that for each firefly, the position is continuously updated. Hence, the values of these fireflies must be converted back to binary form by following the approach in Step 1.

$$X_{new} = X_{old} + \beta \times (X_j - X_i) + \alpha\,(Rand - 0.5) \quad (7)$$

where the first section of the equation symbolizes the current position, the second section contains the attractiveness between the position of F_i and F_j, the third section symbolizes the random movement, α is the randomization parameter, while *Rand* is a uniformly distributed random number ranging from 0 to 1. Hence, the expression $(Rand - 0.5)$ ranges from -0.5 to 0.5 to accommodate both positive and negative changes.

vi. Mutation operator for best firefly

As earlier described, the best firefly maintains its position in the original FA, and this slows down the search process and makes the algorithm more prone to local optima entrapment. However, in the proposed GA-FA variant, the GA mutation operator is employed to update the position of the best firefly via random exchange of the random features and variables. Fig. 4 illustrates this operator. In order to execute this operator, a few steps should be performed; first, an even random number represents how many features swapped inside the best firefly is set as *RF*. Then, half of this *RF* is swapped with the other half. Fig. 4 illustrates an example when *RF* is equal to 2, which means two random positions (r_1, r_2) are swapped. The resulted solution is evaluated using the evaluation function (Eq. (2)). If the updated solution is better than the original, then, keep it, otherwise, return to the previous solution.

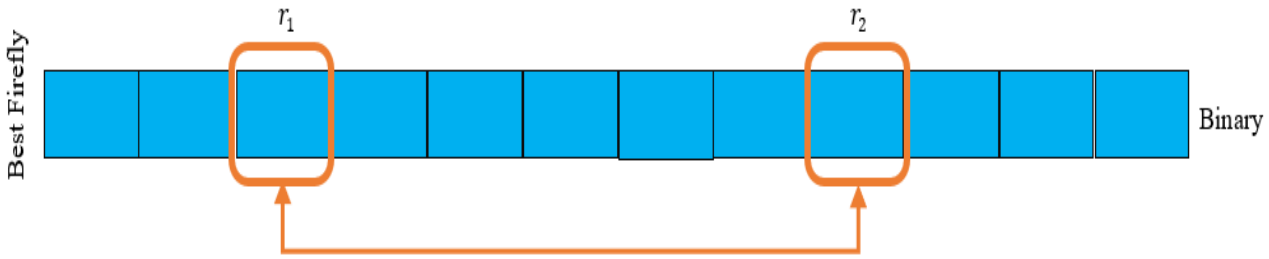

Figure. 4 A schematic view for mutation operator

The mutation operator enhances the searching process of fireflies and decreases the chances for the algorithm when getting stuck in the search space. A mutation operator is performed on both (binary and continuous) values, meaning that the chosen random positions affect both the continuous and discrete values.

4. Results and discussions

4.1 NSL-KDD dataset

Some machine learning and pattern classification algorithms have been used to fix the intrusion detection problems based on the KDD Cup dataset. The KDD Cup '99 dataset has many problems; for example, it contains several redundant and duplicate records, and the difficulty level of the different records and the percentage of records in the original KDD dataset are not inversely proportional. These deficits result in a poor evaluation of different proposed ID techniques. To solve these problems, Tavallaee et al. [24] provided a modified version of the KDD Cup 1999 dataset which is the NSL-KDD dataset. The NSL-KDD dataset was proposed to overcome some of these inherent problems of the KDD Cup 1999 data set. The proposed new dataset consists of selected records of the complete KDD dataset [24]. The high points of the new NSL-KDD over the complete KDD dataset are as follows [24]:

1. Exclusion of redundant features in the training set; hence, the issue of bias towards more frequent records is eliminated.
2. There is an inverse relationship between the selected number of records from each group level and the percentage of records in the original KDD data set.
3. When the number of records in the testing and training dataset is enough, experiments can be economically tested on the whole set without the need for sampling at a reduced scale.

In the NSL-KDD dataset, each record consists of 41 features (e.g., protocol type, service, and flag) and is labeled as normal or one of the specific attack names (DoS, Probe, U2R, and R2L).

There are 4 basic attack categories in the KDD data set; these are:

- **Denial of Service (DoS):** Here, the attacker can gain access to the system and intentionally make it busy just to ensure normal requests are not considered.
- **Surveillance and Other Probing:** Here, an attacker performs network scanning to identify the areas of system vulnerabilities that can be exploited based on the acquired system information.
- **Unauthorized Access from a Remote Machine (R2L):** Here, the attacker sends a packet to a network machine to exploit the machine vulnerabilities to illegally access the network as a disguised genuine user.
- **Unauthorized Access to Local Super User (U2R):** This situation involves an attacker having access to a system as a genuine user in order to exploit the weaknesses of the system and again unhindered access to the system.

4.2 Performance metrics and experimental results

The proposed model was evaluated based on the following metrics – Classification accuracy and the number of selected features. These metrics have been used in the evaluation of IDS performances in the previous studies. The metrics are commonly defined as follows:

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

where:

- *TP*: is the True Positives which means positive cases are correctly identified.
- *TN*: is the True Negatives which mean positive cases are incorrectly identified.
- *FP*: is the False Positives which mean negative cases are incorrectly identified as positive.
- *FN*: is the False Negatives which mean positive cases are incorrectly identified as negative.

All the experiments are conducted on a personal computer with a 3.00 GHZ i5 CPU, 8.00GB RAM, and 64-bit Windows 10 Pro operating system.

The classification accuracy and the number of features is essential to evaluate the performance of the intrusion detection model. The proposed models were used for the best subset features selection for the classification problem of IDSs. The dataset was subdivided into two parts, 70% of the dataset is used for training and 30% of the dataset is used for testing. Due to the possible statistical fluctuations, both the standard firefly algorithm (SFA) and the modified firefly algorithm (MFA) are executed 15 times. The best and worst results are reported. Additionally, four swarm sizes of 10, 20, 30, and 40 agents were used, and the number of iterations was fixed to 500 iterations.

The best and worst accuracy of the SFA and MFA are shown in Fig. 5, and the average accuracy for the best and worst accuracy of the SFA and MFA algorithms have been presented in Fig. 6.

From Fig. 5 and Fig. 6, it is observed that increasing the population size improves the performance of both algorithms. Additionally, it is noticed that MFA outperforms SFA in all cases. Moreover, the reported results confirm the performance stability of MFA compared to SFA.

The number of selected features for each algorithm using different population sizes is reported in Fig. 7.

Based on Fig. 7, it noticed that increasing the population sizes decreases the number of selected features in almost all cases for both algorithms. In addition, it is observed that MFA presents the minimum number of selected features in almost all cases. Hence, MFA is better than SFA in terms of the number of selected features and classification accuracy.

The first binary-classification model has been compared to other binary classification models as well as the number of selected features including Taher et al. [25], Najeeb et al. [26]. The results of the comparison are shown in Table 1.

Taher et al. [25] have conducted a comparison study between Artificial Neural Network (ANN) based machine learning with wrapper feature selection and Support Vector Machine(SVM). From the experiment, the Artificial Neural Network (ANN) outperforms the Support Vector Machine (SVM) technique while classifying network traffic. NSL-KDD dataset is used to assess the performance using SVM and ANN supervised machine learning techniques. Najeeb et al. [26] have proposed a new binary firefly attribute reduction algorithm to selects the best number of attributes from the NSL dataset.

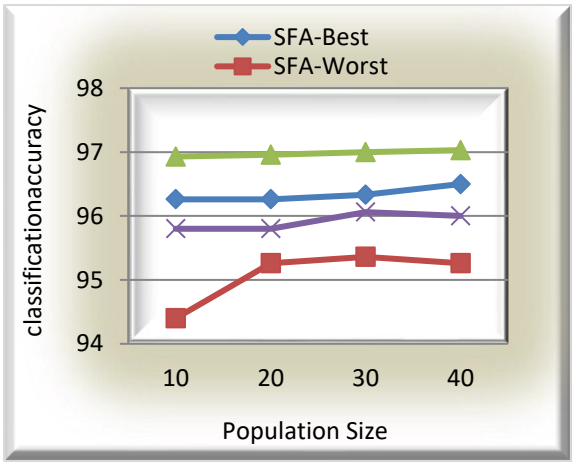

Figure. 5 The obtained classification accuracy of the SFA and MFA using different population sizes

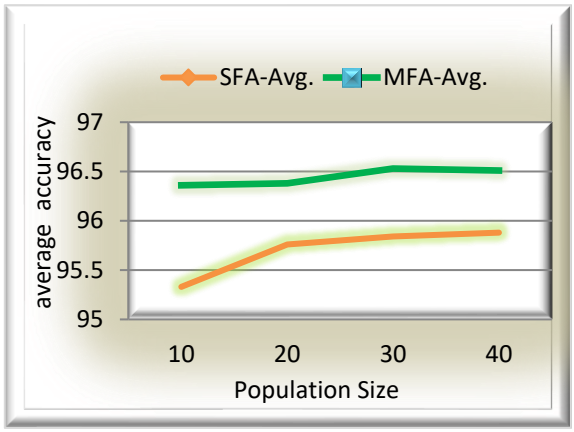

Figure. 6 The obtained average accuracy of the SFA and MFA using different population sizes

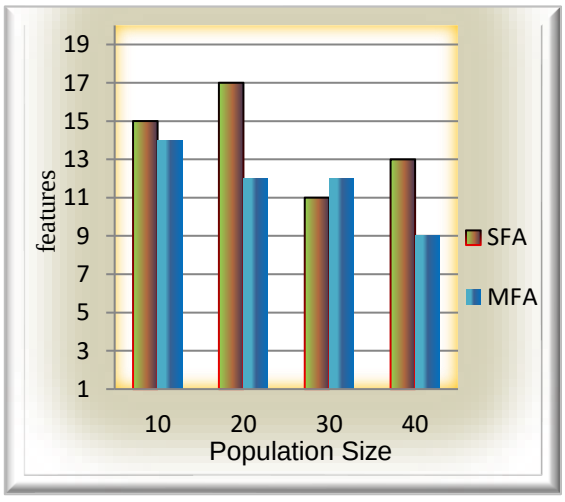

Figure. 7 The number of selected features of SFA and MFA using different population sizes

Besides, the FA was employed with multi-objectives depending on the classification accuracy and the number of attributes at the same time. The proposed classification and feature selection

algorithms enhance the performance of the IDS in the detection of attacks.

The second multi-classification model has been compared to other multi-classification models as well as the number of selected features including Kumar et al. [14], Thaseen et al. [15], and Kuang et al. [27]. The results of the comparison are shown in Table 2. A new Multi class SVM (Support Vector Machine) is proposed by Kumar et al. [14] as dimensionality reduction method for intrusion detection. For feature extraction method, Multi-Linear Dimensionality Reduction (ML-DR) is suggested to minimize the data dimension in order to minimize the training time. A Multi class SVM (M-SVM) is utilized to perform multi-attack classification. For performance evaluation, the NSL-KDD data set is used for the proposed approach. Chi-square feature selection and multi class support vector machine (SVM) is suggested by Thaseen et al. [15] as an intrusion detection model.

A multi class SVM has been constructed to minimize the training and testing time and maximize the classification accuracy of the network attacks. The investigational results on NSL-KDD dataset shows that the suggested model results improved detection rate and reduced false alarm rate. Kuang et al. [27] have proposed a hybrid kernel principal component analysis (KPCA) with a genetic algorithm(GA). KPCA is adopted to extract the principal features of intrusion detection data, and a multi-layer SVM classifier is employed to estimate whether the action is an attack. GA is used to select suitable parameters for the SVM classifier, which avoids over-fitting or under-fitting of the SVM model**.** As we can see from Table 1, it noticed the proposed approach outperform the other approaches in terms of the number of selected features and accuracy. The results is obtained with 40 firefly swarm and the number of iterations was fixed to 500 iterations.

Based on Table 2, it noticed the proposed approach provides the best result in terms of the number of selected features, while the works presented in [14] and [27] have the worst result. On the other side, the works presented in [14] and [15] have the best multi-classification accuracy. However, the proposed work provides a very comparative performance in terms in classification accuracy.

The results presented in Table 2 are obtained with 30 firefly swarms and the number of iterations was fixed to 500 iterations. Additionally, the classification accuracy of each approach regarding each attack is given in Table 3. The obtained results are reported for four types of attacks including DoS, Probe, U2R, and R2L in addition to the normal case.

Based on Table 3, it is noticed that the best performance for the different approaches is obtained with the normal case and the DOD attack. On the other side, the worst performance for the different approaches is obtained with U2R and R2L attacks. In addition, it is observed that the proposed approaches have a close performance in all cases. Hence, the proposed approach can be considered the best compared to the other approach

Table 1. Comparison between binary-classification MFA and other algorithms

Method		Dataset	Selected features	Accuracy%
Artificial Neural Networks - Support Vector Machine (ANN – SVM) 2019	[25]	NSL-KDD	17	94.02
Binary Firefly Algorithm (BFA – NBC) 2018	[26]	NSL-KDD	11	95.84
Proposed Algorithm (MFA)		NSL-KDD	9	**96.51**

Table 2. Comparison between multi-classification MFA and other algorithms

Method		Dataset	SF	Accuracy%
Multi-Linear Dimensionality Reduction (ML–DR) **2018**	[14]	NSL-KDD	41	98.44
chi-square – support vector machine (Chi – SVM) **2017**	[15]	NSL-KDD	31	98.02
Support Vector Machine - Genetic Algorithm (SVM-GA) **2014**	[27]	NSL-KDD	41	95.18
Proposed Algorithm (MFA)		**NSL-KDD**	**10**	96.942

Table 3. Classification accuracy for each attacks class

Method	Normal	DoS	Probe	U2R	R2L
ML-DR [14] **2018**	95.738	95.996	94.971	79.775	78.669
Chi-SVM [15] **2017**	96.1	98.87	95.8	76.92	96.37
SVM-GA [27] **2014**	95.446	94.853	94.472	76.213	76.928
Proposed multi-classification MFA	97.337	98.583	98.462	96.664	96.835

The superiority of the proposed system is achieved through empowering the original firefly algorithm with the mutation operator borrowed from the genetic algorithm. This modification has improved the exploration capability of the original firefly, which allowed the modified algorithm to search effectively uncovered areas from the search space. In addition, this modification allowed the modified firefly to avoid trapping into local optima.

5. Conclusion

Intrusion detection has become a major task in computer networks to preserve its privacy, availability, and security. However, the large number of features may complicate the intrusion detection process. Therefore, there is a necessity to reduce the number of features that can be used for intrusion detection. In this paper, we have presented a modified version of the firefly algorithm to select the optimal subset of features. The modified firefly algorithm adopts the mutation operator to avoid trapping into local optima. The modified firefly algorithm has been utilized to construct two models, one for binary-class classification and the other for multi-class classification. The performance of the proposed models is evaluated using a Naïve Bayes classifier over the NSL-KDD dataset in terms of the classification accuracy and the number of selected features against several related works. The binary and multi-class classification accuracy and the number of selected features of the proposed approaches have been reported using different population sizes. Additionally, the obtained results showed that the two proposed models have improved the classification accuracy and reduced the number of selected features by selecting the most relevant subset of features. In binary classification, the proposed system has 96.51% classification accuracy and reduces the number features from 41 to 9. On the other side, in multi-classification, the proposed system has 96.942% overall classification accuracy and reduces the number of features from 41 to 10. The results enhancement is achieved due to improved exploration capability of the modified firefly algorithm, which has been obtained by employing the mutation operator borrowed from the genetic algorithm.

For future studies, our algorithm can be applied for solving different case studies of feature selection, such as email spam filtering problem, or medical datasets such as hear the diagnosis, disease diagnosis. Moreover, it can be used for handling different optimization problems other than feature selection, such as training artificial neural networks by tuning the weights and biases between the layers.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Contributions

Karrar Mohsin Alwan: Conceptualization, Methodology, Software, Validation, Investigation, Writing - original draft, Writing - review & editing.

Ahmed H. AbuEl-Atta: Supervision, Writing - review & editing.

Hala Helmy Zayed: Supervision, Writing - review & editing.

References

[1] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection for discrete sequences: A survey”, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24, No. 5, pp. 823–839, 2012.

[2] S. Agrawal and J. Agrawal, “Survey on anomaly detection using data mining techniques”, *Procedia Computer Science*, Vol. 60, No. 1, pp. 708–713, 2015.

[3] V. Balamurugan and R. Saravanan, “Enhanced intrusion detection and prevention system on cloud environment using hybrid classification and OTS generation”, *Cluster Computing*, Vol. 22, No. 6, pp. 13027-13039, 2019.

[4] A. S. K. Pathan, The State of the Art in Intrusion

Prevention and Detection. *CRC Press, Taylor & Francis Group*, 2014.

[5] B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms", *Applied Soft Computing*, Vol. 18, pp. 261–276, 2014.

[6] N. C. Long, P. Meesad, and H. Unger, "A highly accurate firefly based algorithm for heart disease prediction", *Expert Systems with Applications*, Vol. 42, No. 21, pp. 8221–8231, 2015.

[7] C. H. Cheng, T. L. Chen, and L. Y. Wei, "A hybrid model based on rough sets theory and genetic algorithms for stock price forecasting", *Information Sciences*, Vol. 180, No. 9, pp. 1610–1629, 2010.

[8] P. Moradi and M. Rostami, "Integration of graph clustering with ant colony optimization for feature selection", *Knowledge-Based Systems*, Vol. 84, pp. 144–161, 2015.

[9] H. Alazzam, A. Sharieh, and K. E. Sabri, "A feature selection algorithm for intrusion detection system based on Pigeon Inspired Optimizer", *Expert Systems with Appications.*, Vol. 148, p. 113249, 2020.

[10] S. Hosseini and B. M. H. Zade, "New Hybrid Method for Attack Detection Using Combination of Evolutionary Algorithms, SVM, and ANN", *Computer Networks*, 2020, https://doi.org/10.1016/j.comnet. 2020.107168.

[11] M. H. Kamarudin, C. Maple, and T. Watson, "Hybrid feature selection technique for intrusion detection system", *International Journal of High Performance Computing and Networking*, Vol. 13, No. 2, pp. 232–240, 2019.

[12] B. Selvakumar and K. Muneeswaran, "Firefly algorithm based feature selection for network intrusion detection", *Computers & Security*, Vol. 81, pp. 148–155, 2019.

[13] M. Mazini, B. Shirazi, and I. Mahdavi, "Anomaly network-based intrusion detection system using a reliable hybrid artificial bee colony and AdaBoost algorithms", *Journal of King Saud University-Computer and Information Sciences*, Vol. 31, No. 4, pp. 541–553, 2019.

[14] B. N. Kumar, M. S. V. S. B. Raju, and B. V. Vardhan, "Enhancing the performance of an intrusion detection system through multi-linear dimensionality reduction and Multi-class SVM", *International Journal of Intelligent Engineering and Systems*, Vol. 11, No. 1, pp. 181–192, 2018.

[15] I. S. Thaseen and C. A. Kumar, "Intrusion detection model using fusion of chi-square feature selection and multi class SVM", *Journal of King Saud University-Computer and Information Sciences*, Vol. 29, No. 4, pp. 462–472, 2017.

[16] K. S. Desale and R. Ade, "Genetic algorithm based feature selection approach for effective intrusion detection system", In: *Proc. of International Conf. On Computer Communication and Informatics*, pp. 1–6, 2015.

[17] A. J. Malik and F. A. Khan, "A hybrid technique using multi-objective particle swarm optimization and random forests for probe attacks detection in a network", In: *Proc. of IEEE International Conf. on Systems, Man, and Cybernetics*, pp. 2473–2478, 2013.

[18] M. Schiezaro and H. Pedrini, "Data feature selection based on Artificial Bee Colony algorithm", EURASIP Journal on Image and Video Processing, Vol. 2013, No. 1, pp. 47-54, 2013.

[19] S. Srinoy, "Intrusion detection model based on particle swarm optimization and support vector machine", *In 2007 IEEE Symposium on Computational Intelligence in Security and Defense Applications*, pp. 186-192. IEEE, 2007.

[20] X. S. Yang, "Firefly algorithms for multimodal optimization", *In International symposium on stochastic algorithms*, pp. 169-178, 2009.

[21] L. Zhang, L. Shan, and J. Wang, "Optimal feature selection using distance-based discrete firefly algorithm with mutual information criterion", *Neural Computing and Applications*, Vol. 28, No. 9, pp. 2795-2808, 2016.

[22] A. H. Gandomi, X. S. Yang, S. Talatahari, and A. H. Alavi, "Firefly algorithm with chaos", *Communications in Nonlinear Science and Numerical Simulation*, Vol. 18, No. 1, pp. 89–98, 2013.

[23] X. S. Yang, "Firefly algorithm, Levy flights and global optimization", *In Research and Development in Intelligent Systems*, Vol. XXVI, pp. 135–146, 2010.

[24] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set", *In 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, pp. 1–6, 2009.

[25] K. A. Taher, B. M. Y. Jisan, and M. M. Rahman, "Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection", In: *Proc. of International Conf. On* Robotics, Electrical and Signal Processing Tec*hniques*, pp. 643–646, 2019.

[26] R. F. Najeeb and B. N. Dhannoon, "A feature selection approach using binary firefly

algorithm for network intrusion detection system", *ARPN Journal of Engineering and Applied Sciences*, Vol. 13, No. 6, pp. 2347–2352, 2018.

[27] F. Kuang, W. Xu, and S. Zhang, "A novel hybrid KPCA and SVM with GA model for intrusion detection", *Applied Soft Computing*, Vol. 18, pp. 178–184, 2014.