A New Framework for Anomaly Detection in NSL-KDD Dataset using Hybrid Neuro-Weighted Genetic Algorithm

[1]**Muneeshwari P, & [2]Kishanthini M**
[1]*Department of Information Technology, PSNA College of Engineering and Technology, Dindigul, Tamilnadu, India.*
[2]*Department of Computer Science and Engineering, Amrita College of Engineering and Technology, Nagercoil, Tamilnadu, India.*

****Corresponding Author: radhamunishapcse@gmail.com**

Received: *05.01.2020,*
Revised: *08.02.2020,*
Accepted: *17.03.2020,*
Published: *30.03.2020*

DOI:
10.53409/mnaa.jcsit1105

Abstract: There are an increasing number of security threats to the Internet and computer networks. For new kinds of attacks constantly emerging, a major challenge is the development of versatile and innovative security-oriented approaches. Anomaly-based network intrusion detection techniques are in this sense a valuable tool for defending target devices and networks from malicious activities. With testing dataset, this work was able to use the NSL-KDD data collection, the binary and multiclass problems. With that inspiration, data mining techniques are used to offer an automated platform for network attack detection. The system is based on the Hybrid Genetic Neuro-Weighted Algorithm (HNWGA).In this weighted genetic algorithm is used for the selection of features and in this work a neuro-genetic fuzzy classification algorithm has been proposed which is used to identify malicious users by classifying user behaviors. The main benefit of this proposed framework is that it reduces the attacks by highly accurate detection of intruders and minimizes false positives. The evaluation of the performance is performed in NSL-KDD dataset. The experimental result shows of that the proposed work attains better accuracy when compared to previous methods. Such type of IDS systems are used in the identification and response to malicious traffic / activities to improve extremely accuracy.

Keywords: Data mining, Hybrid Genetic Neuro-Weighted Algorithm, neuro-genetic fuzzy classification and NSL-KDD dataset.

I. INTRODUCTION

Network intrusion detection (IDS), by detecting malicious users, are useful for providing protection to allow only legitimate users and detach malicious users further. The use of IDSs[1] effectively meets safety criteria such as confidentiality, non-repudiation and authentication. Both servers or on network nodes, IDSs can be installed. Innocence and external threats are marked. Due to their membership in businesses and associations, network users misuse their privileges and certificates issued by internally targeted individuals.

This form of malicious intrusion would allow network resources to be leveraged through network services to be disrupted. The external user should then define attacks on the basis of external user anomalies. In this case, network usage habits for these attackers are monitored for a defined length, such that attackers can be differentiated from legitimate users. The literature already contains a wide range of methods and tools for intrusion detection. Until then, most current tools rely on the analysis of a benchmark dataset on specified types of attacks and do not conduct smart, soft calculation-based analysis[2]. New techniques must therefore be put forward, that could analyze all types of attacks effectively by applying intelligent soft-computing techniques.

However, numerous researchers utilize methods for selection of features to increase classification algorithms and IDS[3] performance. With certain cases, knowledge collection or information processing is used to make final decisions. In addition, the values in the dataset come from various attributes and their data types are also not standardized. In order to handle this issue , it is important for each attribute to be normalized and emphasized based upon its importance in the classification process.

This research successfully recognize intrusion into networks through a new neuro-genetic hybrid architecture called HNWGA. The proposed HNWGA includes various components and, above all, feature selection and assault classification. Such components have been developed using smart approaches and thus a new genetic algorithm (GFCA) and a smart classification algorithm (GFCA) have been proposed in the present paper to improve detection accuracy. These algorithms include a new genetic algorithm based upon weighted genetic algorithms (WGA). The proposed WGA is here used to define the optimum number of features which are ideal for classifying intruders successfully. The rest of the paper is structured as follows: In section 2 the related work of IDS is discussed. The proposed IDS detection mechanism using HNGA is described in section 3. The experimental results and discussion is discussed in section 4. The conclusion and future work is given in section 5.

II. RELATED WORK

In literature there are a few works aimed at detecting network intrusion. In [4], an efficient, reliable classifier is developed to evaluate a visit to a network as normal and not to the ant-colony algorithm and the support vector machine (SVM). In [5] a technical approach used by PCA to choose the sub-set of SVM, the classification feature and the selection of feature subsets based on their own values was proposed. [6], proposed algorithms such as Efficient Data Adapted Decision Tree(EDADT), Hybrid IDS, Semi-supervised methods and Hopping Period Alignment and Adjustment (HOPERAA) varying algorithms respectively. In [7], the new hybrid intrusion detection method is proposed, consisting of a C4.5 decision tree algorithm and a decomposition structure anomaly model, integrated hierarchically. Next, for desiccated subsets, SVM models are created. Two new methods for hybrid intrusive detection are reported in the study presented in [8]: one is based on gravitational search, and one on a mix of GS and GSPSO. In [9], suggested a deep learning approach on recurrent neural networks (RNN-IDS) for intrusion detection. Still be extremely cautious to reduce the training time with the use of GPU acceleration, prevent explosions and disappearances of gradients and study LSTM's classification performance in the field of the bidirectional RNN algorithm. In [10], a hybrid intrusion detection model of several levels, which uses a vector and extreme learning machine to enhance the effectiveness of the detection of external and internal attacks. Despite a large number of existing works, only few types of attacks and other kinds of attacks were taken into consideration by most of the existing systems. In this work , a new IDS is therefore proposed that considers all kinds of known dataset attacks as well as feature selection to enhance the precision of classification.

III. PROPOSED METHODOLOGY

HNWGA-based IDS consists of two primary components, the feature-selection framework for the optimum number of features and the data classification framework using the selected functions. Currently, the necessary data is obtained from the NSL-KDD dataset. The proposed algorithm of WGA function selection will be included in this context in order to choose the appropriate number of features from 41, and the proposed HNWGA algorithm can then be used to classify the data set effectively and the design diagram is shown in Figure 1. In the final judgment of the network data received from IDs by the fluctuating rules manager and the dynamic rule base, this Judgment Manager assists.

3.1. Input NSL-KDD dataset and preprocessing

Various statistical analysis showed the inherent drawbacks in the KDD cup 99 dataset which affect the accuracy of many researchers' IDS detection systems. Data set NSLKDD is its predecessor's refined version. The complete KDD dataset contains an essential record. A collection of files for the researchers can be downloaded. In [11], the details of the attributes are indicated: the name, description of the attributes and sample data. Pre-processing data set: The initial pre-processing steps are as follows:

- ➢ Dataset upload: The planned data set is uploaded for use in the data mining process in this step.
- ➢ Features for extraction: the desired features are chosen from a data set up in step 1. A special feature, a subset of properties or all features may be included in the extraction basis.
- ➢ Feature roles determination: The feature roles are indicated in this stage. "Roles" identify a specific feature identification number and determine whether it is regular, special, labeled, etc.

- Conversion of nominal data to numerical: nominal data must be converted to numerical data.
- Standardization: Feature values based on Z-transformation are normalized in this step.

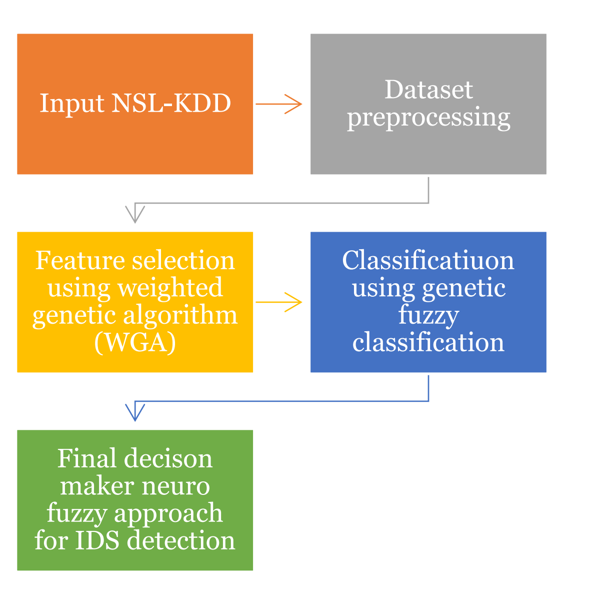
Input NSL-KDD
Dataset preprocessing
Feature selection using weighted genetic algorithm (WGA)
Classificatiuon using genetic fuzzy classification
Final decison maker neuro fuzzy approach for IDS detection

Figure 1. The architecture diagram of HNWGA based IDS.

3.2. Feature Selection Using Weighted Genetic Algorithm

To The input data is converted into binary formats and used for generation formation in the genetic algorithm process. In this model, the proposed algorithm for each iteration is taken from two chromosomes pertaining to two datasets, and they are assessed for physical accuracy, while the algorithm selects two parents when accommodate. If one of them is not suitable, the following dataset record will be considered for the next parent. This process is repeated with the application of fitness values, the choice of two parents, cross-sectional operations and mutations before fitness is assessed. The function set is made up of all attributes chosen by the weighted genetic algorithm. The initial population was created through transforming the data values to binary values. The number of 1s is created for each individual based on their original values, for various characteristics in sub-sets. The suggested weighted average fitness assessment function has been used to determine the weighted average accuracy and number one. Variations of the current generations of different chromosome groups are analyzed using the fusion and mutation operators. Crossover is performed in the latter part of the chromosome by means of uniform crossover operation and mutation. The selection is done by selecting a tournament in which the algorithm selects chromosome subsets for the whole population. The working principle flow diagram of WGA is shown in Figure 1. Table 1 provides pseudo code for the selection of features using weighted genetic algorithms.

The initial population was created through transforming the data values to binary values. The number of 1s is created for each individual based on their original values, for various characteristics in sub-sets. The suggested weighted average fitness assessment function has been used to determine the weighted average accuracy and number one. Variations of the current generations of different chromosome groups are analyzed using the fusion and mutation operators. Crossover is performed in the latter part of the chromosome by means of uniform crossover operation and mutation. The selection is done by selecting a tournament in which the algorithm selects chromosome subsets for the whole population. The working principle flow diagram of WGA is shown in Figure 1. Table 1 provides pseudo code

for the selection of features using weighted genetic algorithms.

Table 1. The pseudo code of feature selection using weighted genetic algorithms

```
Input:   NSL-KDD features (F) (41 features), maximum number of iterations (max_generations), number of records (population size taken from the dataset), crossover probability ($\mathcal{C}\wp$), mutation probability ($\mathcal{M}\wp$).
Output:  Optimal number of selected features
1.  Initialize the chromosome population consisting of 41 attributes
2.  Convert each value of the attributes into binary so that chromosomes can be either '0' or '1.'
3.  Initialise the weights $w_1 = 0.6$ and $w_2 = 0.4$ to the chromosomes ($C_i$)
4.  Do for every single chromosome ($C_i$)
5.  Calculate uniform crossover on $C_i$ with a $\mathcal{C}\wp$ probability.
6.  Calculate mutation operator with a probability of $\mathcal{M}\wp$ to the last bits of the $C_i$.
7.  Evaluate the weighted average fitness evaluation function $\mathcal{F}(i) = \frac{\left[w_1 * accuracy(i) + w_2 * \left(\frac{1}{number\ of\ ones}\right)\right]}{(w_1 + w_2)}$
8.  Condition check $\mathcal{F}(i) > threshold$ then $\mathcal{F}(i) = \mathcal{F}(i)\ Feature\ set$
9.  Using tournament selection from Feature set and choose the top best chromosomes in the new population as the optimal features.
10. Repeat the steps from 3 to 9 until the stop criterion is met
Produce the results of an optimal selection of features.
```

The input data is converted into binary formats and used for generation formation in the genetic algorithm process. In this model, the proposed algorithm for each iteration is taken from two chromosomes pertaining to two datasets, and they are assessed for physical accuracy, while the algorithm selects two parents when accommodate. If one of them is not suitable, the following dataset record will be considered for the next parent. This process is repeated with the application of fitness values, the choice of two parents, cross-sectional operations and mutations before fitness is assessed. The function set is made up of all attributes chosen by the weighted genetic algorithm.

The initial population was created through transforming the data values to binary values. The number of 1s is created for each individual based on their original values, for various characteristics in sub-sets. The suggested weighted average fitness assessment function has been used to determine the weighted average accuracy and number one. Variations of the current generations of different chromosome groups are analyzed using the fusion and mutation operators. Crossover is performed in the latter part of the chromosome by means of uniform crossover operation and mutation. The selection is done by selecting a tournament in which the algorithm selects chromosome subsets for the whole population. The working principle flow diagram of WGA is shown in Figure 1. Table 1 provides pseudo code for the selection of features using weighted genetic algorithms.

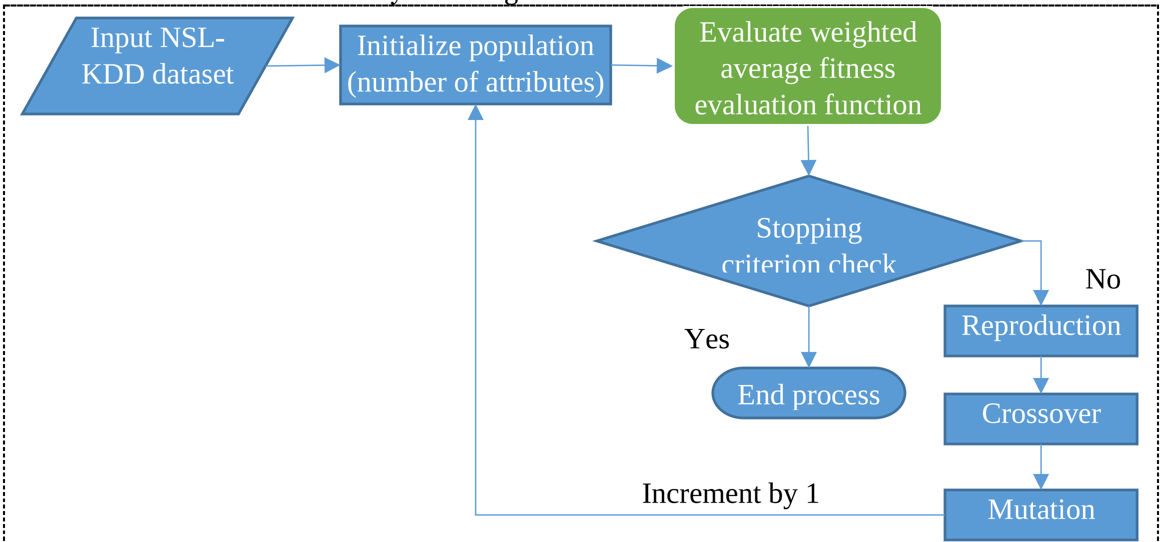

Figure 1. The architecture diagram of HNWGA based IDS.

3.3. *Hybrid Neuro-weighted genetic fuzzy classification for IDS detection*

In this work a new algorithm with fuzzyrules was proposed and evaluated with the benchmark data collection, called the neuroweighted genetic fuzzy classification algorithm (HNWGA). In this classification algorithm, one input layer, one output layer, and two hidden layers were used for back propagation neural networks (BPNNs). As something of an activation function for neural network modeling, the exponential function was being used. Weight modification is often done by means of genetic algorithms with fuzzy rules and the fuzzy rules are used for ultimate choice. The proposed IDS algorithm, HNWGA, is shown in Table 2.

Table 2. The pseudo code of feature selection using weighted genetic algorithms

```
Input: NSL-KDD with records $R_i , i = 1, 2 \dots, n$, with attributes $A_j, j = 1, 2 \dots, m$, Optimal number of selected features, weights for chromosome $x$ are $w_1 = 0.6$ and $w_2 = 0.4$.
Output: The anomaly detection results with attack types
From the KDD cup dataset select some amount of records from the total records at random.
Train data set utilizing neural networks with optimum features in back propagation
Initialize population size, binary form attributes, probability of crossover and probability of mutation.
Read genetic algorithm fitness function as $\mathbb{F}(x) = \frac{(w_1*no.of\ zeros)+(w_2*no.of\ ones)}{(w_1+w_2)}$
For $i = 1\ to\ n$ do
Start substituting the initial BPNN class labels with the training that was conducted by means of first labeling.
By applying union operation enhance new data samples into the training data set.
Repeat steps 14 and15 until the stopiing criterion is obtained
Generate fuzzy rules by applying trapezoidal membership to Training data [12].
Apply fuzzy rules to make weight adjustment decisions
Apply activation function to hidden layers output
Perform the process of defuzzification.
Load the attack types which are classified.
Form rules and store them in fuzzy rule base for testing.
```

Read rules from the base of Fuzzy rules and apply to test data.

Presentation of final result, showing the IDS with its attack types.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In NSL-KDD 1999 cup data set was used in this work to evaluate the feature selection algorithm and classification algorithm developed for the development of an IDS containing five classes (probe, U2R, R2L, DoS, and normal). In this section, the proposed HNWGA quality is assessed using traditional methods such as SVM-IDS [4], GSPSO-IDS [8], and RNN-IDS [9] with certain parameters such as accuracy, f-measure, accuracy and recall.

Precision: It represents the proportion of positive samples correctly classified to the total number of positive samples predicted as shown in equation (1):

$$Precision = \frac{TP}{FP+TP} \qquad (1)$$

Recall: a classifier's recall represents the positive samples correctly classified to the total number of positive samples and is estimated as in equation (2):

$$Recall = \frac{TP}{TP+FN} \qquad (2)$$

F-measure: this is also referred to as F 1-score, and is the harmonic mean of precision and recall as in equation (3):

$$F - measure = \frac{2*(Recall * Precision)}{(Recall + Precision)} \qquad (3)$$

Accuracy: This is one of the most commonly used classification performance measures and is defined as a ratio between the correctly classified samples and the total number of samples as in equation (4):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (4)$$

Where true positive (TP) samples are properly classified as normal, false positive (FP) samples are incorrectly classified as abnormal, true negative (TN) samples are properly classified as abnormal, and false negatives (FN) are incorrectly classified as normal.

4.1. Precision Rate comparison

From the above Figure 2, the graph explains the comparison of precision for the number of records in the specified datasets. Such methods as SVM-IDS, GSPSO-IDS, RNN-IDS and HNWGA are executed. When it increases the number of records according to the precision value. From this graph, it is learned that, due to optimal feature selection technique.

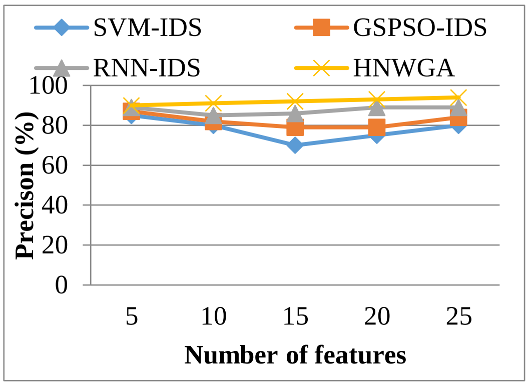

Figure 2. Representation of Precision Comparison

The proposed HNWGA provides 94% higher precision than the previous methods that produce better results in attack detection. The numerical results of Precision Comparison is shown in Table 3.

Table 3. The numerical results of Precision Comparison

No.of features	SVM-IDS	GSPSO-IDS	RNN-IDS	HNWGA
5	85	87	89	90
10	80	82	85	91
15	70	79	86	92
20	75	79	89	93
25	80	84	89	94

4.2. Recall comparison

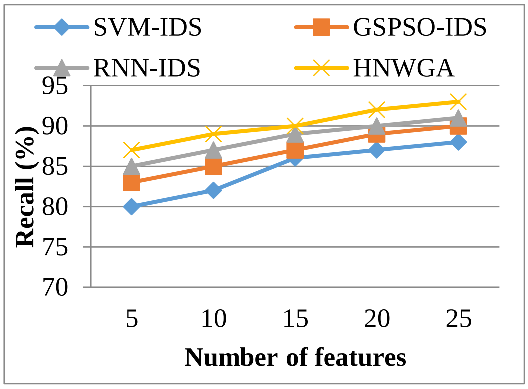

Figure 3. Representation of Recall Comparison

The graph explains from the above Figure 3 that the recall comparison for the number of records in the specified datasets. Such methods as SVM-IDS, GSPSO-IDS, RNN-IDS and HNWGA are executed. Increasing the number of images also increases the corresponding recall value. It is learned from this graph that the proposed HNWGA provides 93% higher recall than previous methods. The reason for this is that the WGA produces the optimal features that will improve the results of attack detection. The numerical results of Recall Comparison is shown in Table 4.

Table 4. The numerical results of Recall Comparison

No.of features	SVM-IDS	GSPSO-IDS	RNN-IDS	HNWGA
5	80	83	85	87
10	82	85	87	89
15	86	87	89	90
20	87	89	90	92
25	88	90	91	93

4.3. *F-measure Rate comparison*

From the above Figure 4, the graph explains the comparison of the f-measure for the number of images in specified datasets. Such methods as SVM-IDS, GSPSO-IDS, RNN-IDS and HNWGA are executed. When the number of data is increased and correspondingly the f-measure value is raised. It is learned from this graph that the proposed HNWGA provides 96 per cent higher f-measurement than previous methods.

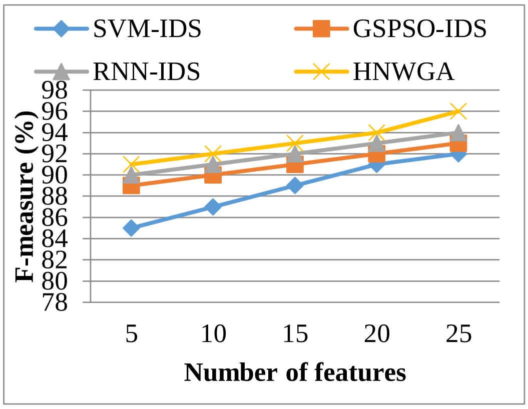

Figure 4. Representation of F-measure Comparison

Thus the proposed HNWGA algorithm is greater in terms of better results of attack detection than the existing algorithms. The reason Neuro fuzzy's parameter is optimized with genetic algorithm which will enhance the results of attack detection. The numerical results of F-measure Comparison is shown in Table 5.

Table 5. The numerical results of F-measure Comparison

No.of features	SVM-IDS	GSPSO-IDS	RNN-IDS	HNWGA
5	85	89	90	91
10	87	90	91	92
15	89	91	92	93
20	91	92	93	94
25	92	93	94	96

4.4. *Accuracy comparison*

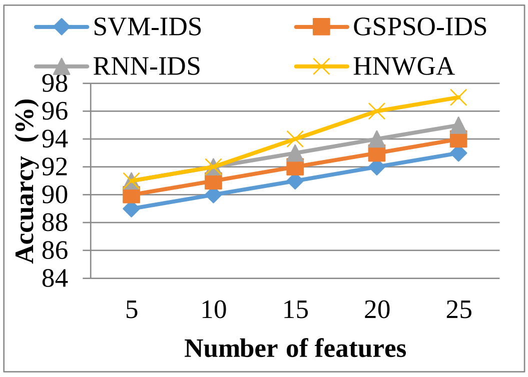

Figure 5. Representation of Accuracy Comparison

From the above Figure 5, the graph explains the comparison of processing time for the number of images in the specified datasets. Such methods as SVM-IDS, GSPSO-IDS, RNN-IDS and HNWGA are executed. It is learned from this graph that the proposed HNWGA algorithm is higher than the existing algorithms in terms of better template matching results with a high precision rate of 97%. The reason is that existing approaches also have a low success rate which has a high likelihood of causing misdetection of emerging changes. This is due to the use of the most important features selected by the feature selection algorithm which uses intelligent agents for decision making and, in addition, the use of fuzzy rules and genetic algorithm in the classification algorithm increases the classification accuracy resulting in an increase in in intrusion detection accuracy. The numerical results of Accuracy Comparison is shown in Table 5.

Table 5. The numerical results of Accuarcy Comparison

No.of features	SVM-IDS	GSPSO-IDS	RNN-IDS	HNWGA
5	89	90	91	91
10	90	91	92	92
15	91	92	93	94
20	92	93	94	96
25	93	94	95	97

V. CONCLUSION AND FUTURE WORK

A new HNWGA framework for an IDS was suggested in this work. Towards this end , a new algorithm called WGA has been suggested for the selection of features which will enhance detection accuracy, network efficiency, and optimal selection of features. Furthermore, there has been a special proposal on a new classification algorithm called GFCA, that also aims to improve the accuracy of intrusion detection. Experiments carried out in this paper show that the GFCA increases the accuracy of classification and that the classification time when selected features are used. Ultimately, the work proposed named HNWGA was tested and its performance analyzed by means of a precise analysis and comparable with SVM-IDS, GSPSO-IDS, RNN-IDS in constant environments. Experimental studies in this study have shown that the proposed algorithm of classifying results is more accurate than the other three classifiers. The key benefit of the proposed IDS system is that the identification is more precise, false positive and deduction times are of. Future research can be carried out in this way to help handle uncertainty by using the adaptive fuzzy inference model.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No animals/humans were used for studies that are basis of this research.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The authors confirm that the data supporting the findings of this research are available within the article.

FUNDING

None.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

The authors would like to thank their present employer for providing support while carrying out this research work.

REFERENCES

[1]. Ghorbani AA, Lu W, Tavallaee M. Network intrusion detection and prevention: concepts and techniques. Springer Science & Business Media; 2009 Oct 10.

[2]. Liao HJ, Lin CH, Lin YC, Tung KY. Intrusion detection system: A comprehensive review. Journal of Network and Computer Applications. 2013 Jan 1;36(1):16-24.

[3]. Ganapathy S, Kulothungan K, Muthurajkumar S, Vijayalakshmi M, Yogesh P, Kannan A. Intelligent feature selection and classification techniques for intrusion detection in networks: a survey. EURASIP Journal on Wireless Communications and Networking. 2013 Dec 1;2013(1):271.

[4]. Li Y, Xia J, Zhang S, Yan J, Ai X, Dai K. An efficient intrusion detection system based on support vector machines and gradually feature removal method. Expert Systems with Applications. 2012 Jan 1;39(1):424-30.

[5]. Ahmad I, Hussain M, Alghamdi A, Alelaiwi A. Enhancing SVM performance in intrusion detection using optimal feature subset selection based on genetic principal components. Neural computing and applications. 2014 Jun 1;24(7-8):1671-82.

[6]. Nadiammai GV, Hemalatha MJ. Effective approach toward Intrusion Detection System using data mining techniques. Egyptian Informatics Journal. 2014 Mar 1;15(1):37-50.

[7]. Kim G, Lee S, Kim S. A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. Expert Systems with Applications. 2014 Mar 1;41(4):1690-700.

[8]. Dash T. A study on intrusion detection using neural networks trained with evolutionary algorithms. Soft Computing. 2017 May 1;21(10):2687-700.

[9]. Yin C, Zhu Y, Fei J, He X. A deep learning approach for intrusion detection using recurrent neural networks. IEEE Access. 2017 Oct 12;5:21954-61.

[10]. Al-Yaseen WL, Othman ZA, Nazri MZ. Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system. Expert Systems with Applications. 2017 Jan 1;67:296-303.

[11]. L. Dhanabal and S. P. Shantharajah, “A Study on NSLKDD Dataset for Intrusion Detection System Based on Classification Algorithms,” vol. 4, no. 6, pp. 446–452, 2015.

[12]. Palanivel K. Fuzzy commercial traveler problem of trapezoidal membership functions within the sort of α optimum solution using ranking technique. Afrika Matematika. 2016 Mar 1;27(1-2):263-77.

Cite this article as: Muneeshwari P, Kishanthini M. A New Framework for Anomaly Detection in NSL-KDD Dataset using Hybrid Neuro-Weighted Genetic Algorithm. J. Comput. Sci. Intell. Technol. 2020; 1(1): 29-36. ©JCSIT, MNAA PUB WORLD, 2020.