USING COMBINATION OF FUZZY SET AND GRAVITATIONAL ALGORITHM FOR IMPROVING INTRUSION DETECTION

Amin Dastanpour [1] and Raja Azlina Raja Mahmood [2]

[1] Computer Department, Kerman Institute of Higher Education, Kerman, Iran
[2] Department of Communication Technology and Network, Faculty of Computer Science and Information Technology, University Putra Malaysia

ABSTRACT

An intrusion detection system (IDS) is a tool used by administrators to protect networks from unknown activities. In signature-based systems, the detection of attacks relies on predefined patterns or behaviours associated with known threats, triggering an alert upon identification of a match. Conversely, anomaly detection systems initiate their process by establishing a baseline profile that reflects the normal operational behaviour of the system or network. These systems possess the capability to identify previously unrecognized attacks, rendering them more effective than their signature-based counterparts. Nevertheless, anomaly-based IDS must consider numerous characteristics when pinpointing attacks Despite these difficulties, machine learning techniques have demonstrated a strong ability to achieve highly accurate anomaly detection and have been employed to identify attacks over the past few decades. Intrusion detection systems are widely used methods to maintain network security. In this paper, the proposed IDS employs machine learning approaches, namely FUZZY are initially applied, followed by optimization algorithms such as Gravitational Search Algorithm (GSA) to determine the optimal subset of detection features. Comparison study on the performance of the FUZZY and FUZZY-GSA models using KDD dataset with selected optimal 27 total features, shows that the proposed model achieves the highest detection rate with the lowest false alarm rate. The highest detection rate for FUZZY-GSA on the KDD dataset is 98.94% in comparison to other recognition algorithm. In summary, the proposed FUZZY-GSA model attains the highest attack recognition percentage with the lowest false positive rate in KDD dataset.

KEYWORDS

Fuzzy, Gravitational Search Algorithm (GSA), Intrusion Detection System (IDS), Security, Networks

1. INTRODUCTION

An intrusion detection system (IDS) serves as a critical resource for network administrators aiming to safeguard their networks against unidentified activities. Intrusion prevention systems (IPS) are regarded as more sophisticated iterations of IDS, as they not only monitor network traffic but also scrutinize system activities for potential threats[1]. The primary distinction lies in the fact that intrusion prevention systems operate in real-time, enabling them to actively thwart or obstruct identified intrusions. Consequently, IDS has emerged as a vital element within security frameworks, facilitating the identification of threats prior to their potential to inflict extensive harm.[2] Furthermore, cyber attackers are increasingly employing shortened URLs to obscure the true destination of links from users. Overall, an IDS functions as a mechanism that assists administrators in defending networks against harmful activities[3].

Machine learning aims to learn, identify, and adapt to evolving conditions over time, thereby enhancing its performance on specific tasks. Beyond intelligently recognizing new attack patterns, these algorithms can also filter out irrelevant and redundant data, retaining only the most important features to streamline and optimize the detection process. [4]. This fundamental task in heuristic data mining is a widely used method for analyzing statistical data across various disciplines such as machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.[5] One significant challenge in the field of machine learning is classification. Within the context of machine learning, classification is categorized as a form of supervised learning[6]. In systems that rely on signature-based methods, the patterns of attacks or the behaviours exhibited by attackers are modelled, prompting the system to generate an alert upon detecting a match [7]. However, a notable limitation of this approach is its reliance on previously identified attacks, necessitating frequent updates to attack signatures[8]. On the other hand, anomaly detection systems need to create a baseline profile that reflects the typical behavior of the system or network before detecting any irregularities[9]. If the activity subsequently diverges from the established norm, it is classified as an intrusion. Anomaly detection systems possess the capability to identify previously unrecognized attacks, rendering them more effective than signature-based systems. [10]. In anomaly-based intrusion detection systems (IDS), numerous features must be considered when pinpointing particular attacks. The system is required to manage a substantial amount of network traffic, coupled with a highly uneven distribution of data, which complicates the differentiation between normal and abnormal behaviour[11]. Given the extensive volume of traffic, a greater quantity of data must be analysed to establish accurate patterns.[12]. There exists a potential for an elevated false positive rate if not addressed appropriately. Nevertheless, in spite of these obstacles, machine learning methodologies have demonstrated their capability to yield highly precise outcomes in the realm of anomaly detection, leading to their application in identifying attacks over the past several decades. [13].

Thus, the main objectives of the study are as follows: 1) to increase the attack detection rate, 2) to reduce the false detection rate, and 3) to find an optimal model for intrusion detection system using the proposed algorithm in KDD dataset. The following are the structure of the paper. Section 2 provides reviews on selected IDSes, Section 3 provides the adopted methodology, Section 4 discusses the results and finally, Section 5 summarizes the finding and presents some future works.

2. RELATED WORKS

Today, many people rely on the Internet for communication. As a result, they anticipate a secure network or communication channel, particularly when exchanging sensitive or confidential information [14]. In recent years, numerous research efforts have focused on network security to guarantee the protection and dependability of data during transmission and storage. One widely adopted method to maintain network security is the Intrusion Detection System (IDS) [15]. The subsequent paragraphs explore some of the published work on IDSes.

Paper No 1 based on [16]: With the continuous growth of computer networks, network intrusion detection has become a crucial element in ensuring security within these systems. Its primary tools include managing network traffic and analysing user behaviour. One effective way to implement such systems is through classification techniques, which assist in recognizing patterns within large datasets. By applying data mining methods and assigning binary labels (normal packet or abnormal packet) along with relevant data features for identifying anomalies, the effectiveness of the intrusion detection system improves, thereby enhancing overall network security. The model discussed in this article focuses on the support vector machine algorithm for feature selection and explores how machine learning algorithms affect the accuracy and

efficiency of intrusion detection. The findings demonstrate that using this algorithm significantly boosts both the precision and correct identification of alerts compared to previous approaches.

Paper No 2 based on [17]: An intrusion can be characterized as a series of actions aimed at undermining the integrity, confidentiality, or availability of a resource. Intrusion detection is categorized into two primary types: exploitative intrusion detection and anomalous intrusion detection. The former relies on established attack patterns that take advantage of vulnerabilities within system and application software to identify intrusions. These patterns are predetermined and help match user behaviour with recognized intrusion signatures. In contrast, anomaly intrusion detection aims to spot deviations from established normal usage patterns. These normal patterns are established through statistical analysis of system characteristics, and any significant variance from these patterns is identified as a potential intrusion. In a Distributed Intrusion Detection System (DIDS), conventional intrusion detection methods are integrated into intelligent agents and spread across large networks. Within this distributed setup, IDS agents can communicate with each other or with a central server. By deploying these collaborative agents throughout the network, incident analysts, network operations teams, and security personnel gain a thorough view of network activities. This distributed monitoring allows for the early detection of planned and coordinated attacks, enabling network administrators to implement proactive defences. Additionally, DIDS plays a vital role in controlling the spread of worms, improving network monitoring and incident analysis, and tracking various types of attacks. It is also crucial for identifying emerging threats from unauthorized users, backdoor attackers, and hackers at multiple geographically diverse locations. In the context of DIDS, it is important that each IDS component remains lightweight and accurate. The use of data mining techniques in intrusion detection first emerged through audit data mining, which aimed to create automated models for detecting intrusions. Different data mining algorithms are applied to audit data to develop models that effectively capture both intrusive behaviours and normal operational activities.

Paper No 3 based on [18]: This research investigates the optimal number of clusters as a critical factor influencing clustering efficiency. In this article, we introduce a novel approach for quantifying C opt in centroid-based clustering. Initially, we present a new clustering validity index, termed fRisk, which is grounded in fuzzy set theory. This index serves to normalize and aggregate local risks associated with actions such as splitting or merging data within clusters. fRisk leverages local distribution information from the dataset to encapsulate the global characteristics of the clustering process through a risk degree metric. Utilizing the theoretically established property of monotonous reduction in fRisk, we propose a new algorithm, fRisk4bA, designed to ascertain C. This algorithm incorporates the well-established L-method as a supplementary tool to identify C opt from the fRisk (C) graph. The method's stable convergence trend is theoretically validated, and numerical experiments are conducted to further support our findings. These experiments demonstrate the method's high reliability and stability, as well as its sensitivity in distinguishing and merging clusters in high-density regions, even in the presence of noise within the datasets, highlighting the strengths of the proposed approach.

Paper No 4 based on [19]: Fuzzy logic serves as a mechanism for pattern recognition and is applicable in the domain of intrusion detection classification. Additionally, it possesses selflearning and adaptive capabilities. When provided with system audit data and network data packets, fuzzy logic can derive a normal user model or system characteristics, enabling it to differentiate between attack patterns and anomalous activities. Probabilistic Neural Networks (PNN), a form of artificial fuzzy logic introduced by Dr. Donald F. Specht in 1989, is characterized by its straightforward structure, rapid convergence, and extensive applicability. This parallel algorithm utilizes an estimation method grounded in Bayesian classification rules and the Parzen window probability density function for effective pattern classification. In practical scenarios, particularly in addressing classification challenges, PNN offers the advantage of linear

learning algorithms, which complement nonlinear learning algorithms while preserving their high accuracy. This study applies probabilistic fuzzy logic to network intrusion classification, proposing a model for this purpose. Experiments were conducted using the KDD Cup dataset to enhance the efficacy of network intrusion classification. By exploring the principles of probabilistic fuzzy logic, a network intrusion classification model is developed to address both binary and multi-class classification issues in intrusion detection. The experiments on the KDD Cup 99 dataset demonstrate that the proposed model exhibits superior performance, achieving a higher rate of intrusion detection.

Paper No 5 based on [20]: In this study, the researchers introduce two novel clustering algorithms designed for identifying network fraud and intrusions: the Enhanced Competitive Learning Network (ECLN) and the Supervised Enhanced Competitive Learning Network (SECLN). ECLN operates as an unsupervised clustering algorithm, integrating new rules into the conventional Competitive Learning Fuzzy Network (SCLN). Within the ECLN framework, neurons are trained to accurately represent the data centre using a newly established rewardpunishment update mechanism, which effectively mitigates the instability issues found in SCLN. Conversely, SECLN acts as the supervised version of ECLN. This algorithm utilizes a new supervised update rule that takes advantage of data labels to improve the training process, resulting in better clustering performance. SECLN is adaptable, capable of processing both labelled and unlabelled datasets, and shows considerable resilience against missing labels or delays. Furthermore, SECLN possesses self-reconstruction abilities, making it entirely independent of the initial number of clusters. The experimental evaluations performed on both research datasets and real-world data pertaining to fraud and network intrusion detection indicate that both ICLN and SICLN demonstrate strong performance, with SICLN surpassing traditional unsupervised clustering algorithms.

Table 1 shows the comparison table of the studied systems. The IDSes have the ability to detect various types of malicious network activities and attacks on computer systems. The existing body of literature has produced numerous anomaly detection systems that utilize a range of machine learning methods. These methods serve as classifiers, assessing whether incoming internet traffic is harmless or potentially harmful. Although this paper does not provide a systematic review of the different machine learning strategies in intrusion detection, the comparative analysis of existing studies highlights the need for additional research to enhance the development of intrusion detection systems that leverage machine learning approaches.

Table1. Comparison Study of Selected Intrusion Detection Systems

Research work 5 [20]		Research work 4 [19]	Research work 3 [18]	Research work 2 [17]	Research work 1 [16]	Article
Supervised Improved Competitive Learning Network (SICLN)	Improving Competitive Learning Network (ICLN)	Probabilistic Fuzzy (PNN)	Artificial Fuzzy and Fuzzy Clustering (FC-FUZZY)	Fuzzy Classification (FR2)	Feature selection based on linear correlation (LCFS)	Algorithm name
✓	✓	✓	✓	✓	✓	High detection rate
			✓			Low false alarm
					✓	Number of features

Although most of the methods provide high detection rate, yet only few of the methods are low in false detection. Some of the methods focus on small number of features for pattern detection with the detection rate in these methods is low. Conversely, certain algorithms exhibit a high detection rate; however, this coverage encompasses numerous features for pattern detection. While some fuzzy methods demonstrate elevated detection rates, they also suffer from a significant false detection rate, and the stability of detection requires enhancement. In essence, the fuzzy algorithm achieves a high detection rate, yet this comes with a considerable incidence of false detections. Although integrating GSA with other IDS algorithms can optimize pattern detection, resulting in a low false detection rate and a high detection rate, these methods are hindered by either a low detection rate or an excessive number of features utilized.

3. METHODOLOGY

The information produced by intrusion detection systems undergoes meticulous analysis, which constitutes the primary function of any IDS, to identify potential attacks or intrusions. Machine learning methodologies can autonomously develop the necessary model utilizing a training dataset. By employing machine learning approaches for intrusion detection, it is possible to automatically construct a model derived from a training dataset composed of data samples characterized by a specific set of features (attributes) and corresponding labels. The incoming packets are analysed and classified according to the values of the features to generate both training and testing datasets. Various machine learning techniques are used and built the intrusion detection system, in this case, fuzzy and GSA algorithms. Figure 1 shows the overall proposed intelligent detection model of IDS.

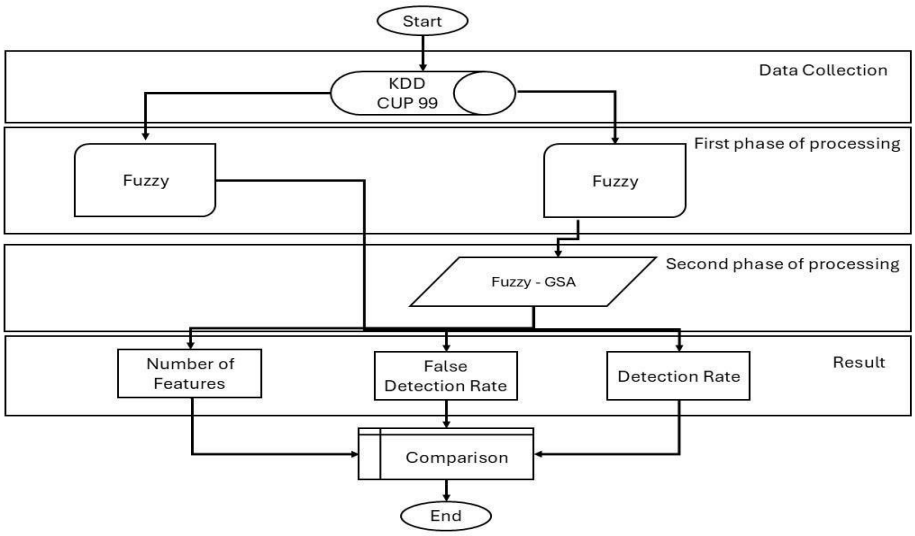

Figure 1. Proposed Intelligent Detection Model of IDS.

The KDD99 dataset was created in 1999 by combining individual TCP packets into TCP connections, based on the DARPA98 network traffic dataset [21]. This dataset served as the benchmark in the International Knowledge Discovery and Data Mining Tools Competition, and it is the most widely utilized dataset in the field of intrusion detection [22]. Each TCP connection has41 features with a label which specifies the status of a connections either being normal, or a specific attack type[23]. There are 38 numeric features and 3 symbolic features, falling into the following four categories [24]. The first category is the basic features: Nine basic features were used to describe each individual TCP connection. The second category is the content features:

Thirteen features related to domain knowledge were utilized to signify suspicious behaviour that exhibited no sequential patterns within the network traffic. The third category is the time-based traffic features: Nine features were utilized to encapsulate the relationships from the previous two seconds that shared either the same destination host or the same service as the ongoing connection. Finally, the fourth category is the host based traffic features: Ten features were developed utilizing a window of 100 connections to the same host rather than a time window, as slow scan attacks can span a significantly longer duration.

The training dataset comprises 4,940,000 data instances, which include normal network traffic as well as 24 distinct attacks. The testing dataset consists of 311,029 data instances, encompassing a total of 38 attacks, 14 of which are not represented in the training dataset [25]. Due to the excessively large size of the training set, an alternative training set comprising 10% of the data is often utilized [26]. KDD99 dataset is complete and hence one of the best datasets in investigating one's IDS performance. This dataset contains 24 types of attacks, and they can be categorized in four groups DOS, R2L, U2R and probing, with details as follows [22]:

DOS: denial of service. This type of attack is used for understanding the behaviour of the user. This kind of attack needs to spend some memory and computing resources[27].

R2L: unauthorized access from a remote machine. This type of attack sends some packets in the network in order to achieve accessibility in the network as a local and known user[28].

U2R: unauthorized access to local super user (root) privileges. U2R attacks are considered as the type of attacks, in which the system is accessible to the attacker and can exploit the vulnerabilities for gaining the main permissions.

Probing: surveillance and other probing: This type of attack scans a network to collect data about the host that has been targeted[29].

The complete type of attacks of KDD Cup 99 dataset are shown in Table 2, categorized into four types of attacks (DOS, R2L, U2R and Prob)[30]. A complete listing of the set of features defined for the connection records is given in the three tables below, namely Table 2, Table 3 and Table 4.

Table 2. Complete type of attack in KDD Cup 99 dataset

Name of attack	**Type of attack**
Back	Dos
Buffer_overflow	U2R
ftp_write	R2L
guess_passwd	R2L
Imap	R2L
Ipsweep	Probe
Land	Dos
Loadmodule	U2R
Multihop	R2L
Neptune	Dos
Nmap	Prob

Name of attack	Type of attack
Perl	U2R
Phf	R2L
Pod	Dos
Portsweep	Probe
Rootkit	U2R
Satan	Probe
Smurf	Dos
Spy	R2L
Teardrop	Dos
Warezclient	R2L
Warezmaster	R2L

Table 3. Basic features of individual TCP connections.

Feature name	Descriptio n	Type
duration	length (number of seconds) of the connection	Continuous
protocol_type	type of the protocol, e.g. tcp, udp, etc.	Discrete
service	network service on the destination, e.g., http, telnet, etc.	Discrete
src_bytes	number of data bytes from source to destination	Continuous
dst_bytes	number of data bytes from destination to source	Continuous
flag	normal or error status of the connection	Discrete
land	1 if connection is from/to the same host/port; 0 otherwise	Discrete
wrong_fragmen t	number of ``wrong" fragments	Continuous
urgent	number of urgent packets	Continuous

Table 4. Content features within a connection suggested by domain knowledge.

Feature name	Description	Type
hot	number of ``hot'' indicators	Continuous
num_failed_logins	number of failed login attempts	Continuous
logged_in	1 if successfully logged in; 0 otherwise	Discrete
num_compromised	number of ``compromised'' conditions	Continuous
root_shell	1 if root shell is obtained; 0 otherwise	Discrete
su_attempted	1 if ``su	Discrete
	root'' command attempted; 0 otherwise	
num_root	number of ``root'' accesses	Continuous
num_file_creations	number of file creation operations	Continuous
num_shells	number of shell prompts	Continuous
num_access_files	number of operations on access control files	Continuous
num_outbound_cmds	number of outbound commands in an ftp session	Continuous
is_hot_login	1 if the login belongs to the ``hot'' list; 0 otherwise	Discrete
is_guest_login	1 if the login is a ``guest''login; 0 otherwise	Discrete

The inference system in question comprises a collection of fuzzy IF-THEN rules endowed with learning capabilities, enabling the approximation of nonlinear functions. Consequently, fuzzy logic is recognized as a universal estimator. The application of fuzzy modelling spans a wide array of practical domains, particularly in control and predictive tasks. This innovative inference system incorporates universal approximation techniques to effectively represent highly nonlinear functions. An adaptive fuzzy system is structured as a network of interconnected nodes, each linked by directional connections. Each node is defined by a node function that possesses fixed yet adjustable parameters. The learning or training phase within a fuzzy system involves the process of calibrating these parameters to ensure a proper fit with the training data. Fuzzy Sugeno models are integrated within the framework of adaptive systems to enhance learning and adaptability. In this fuzzy architecture, the first and fourth layers serve as adaptive layers. The parameters subject to modification include the initial parameters in the first layer and the subsequent parameters in the fourth layer. The learning objective is to fine-tune all adjustable parameters so that the fuzzy output aligns with the training data. This training process specifically adjusts the parameters associated with the membership function. Consequently, model validation becomes essential to confirm the efficacy of the fuzzy inference system, utilizing a test dataset.

This test dataset plays a critical role in assessing the generalizability of the fuzzy inference system's outcomes. Figure 2 illustrates the flowchart of an IDS employing fuzzy logic.

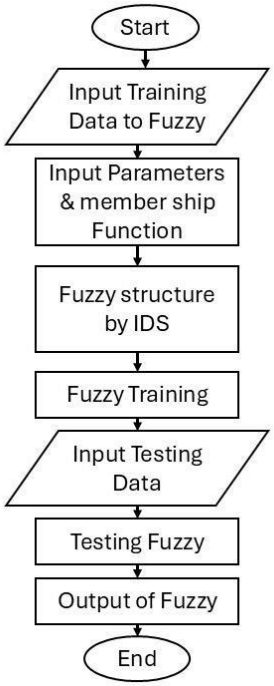

Figure 2. An overall flowchart of Fuzzy

In this research, we used GSA to optimize the fuzzy results. The parameters and settings of GSA are shown in Table 5. GSA creates a random factor to calculate the mass value of each factor for optimizing the fuzzy model.

Table 5. Parameters Algorithm Search Gravitational

Values Algorithm Search Gravitational	Parameter Algorithm Search Gravitational
20	Number Representative
100	Maximum Number Repetition
1	Review Elite
1	" Power" R
1	Index Function Test
Maximum	Objective function
Binary	Algorithm type Search Gravitational

4. RESULT AND DISCUSSION

Our proposed solution focuses on anomaly detection system that utilizes machine learning, Fuzzy and GSA in particular. We evaluated the proposed algorithm with Fuzzy algorithm. Table 6 illustrates the results of each algorithm using KDD dataset, detailing the detection rate and the incidence of false positives.

Table 6. Evaluation of the proposed model

KDD CUP 99 database			Algorithm
Number of features 41	False detection rate 0.02%	Detection percentage 98.93 %	**Fuzzy**
27	0.007 %	98.94 %	**Fuzzy - GSA**

In Figure 3, the column depicts the detection rate represented as a percentage, while the rows denote the number of features from the KDD dataset. The FUZZY method applied to feature number 23 of the KDD dataset produces the lowest detection result; in contrast, employing feature number 4 leads to the highest detection rate. A substantial increase in detection rates is apparent between feature numbers 4 and 23, with an increase of 19.47%, which underscores the FUZZY false detection process across the KDD dataset for each feature (see Figure 4). Conversely, the highest false detection rate for FUZZY within the KDD dataset is recorded as 0. Additionally, a significant decrease in false detection rates is noted between feature numbers 23 and 40, with a reduction of 0.02 indicating the optimal result for false detection at feature number 23 in the KDD dataset.

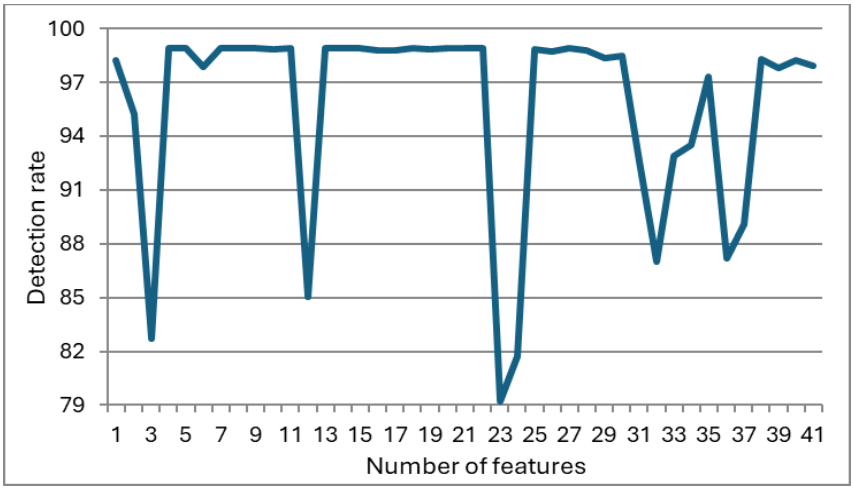

Figure 3. Result Fuzzy detection in Collection Data KDD

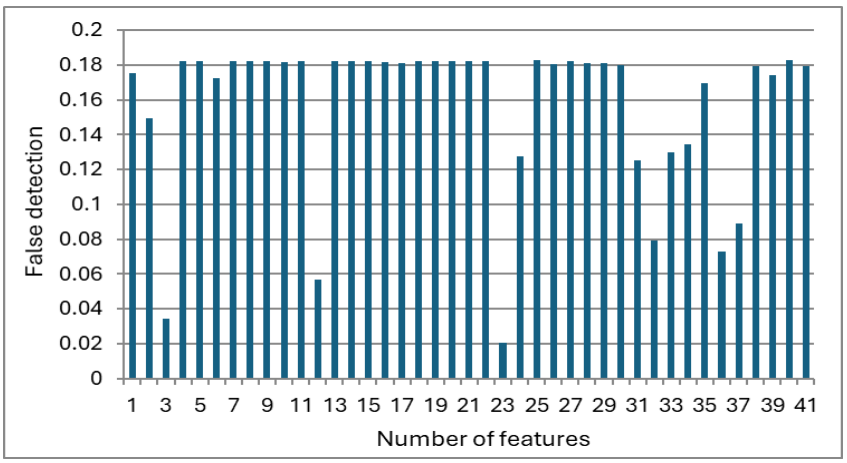

Figure4. Process Diagnosis False Fuzzy In Collection Data KDD

Figure 5 shows the FUZZY-GSA evaluation results, in particular the FUZZY-GSA recognition results on KDD dataset based on different features are shown. The columns represent the recognition rate expressed as a percentage, while the rows denote the number of features in the KDD dataset. This dataset contains 41 features. The minimum observed recognition rate is 88.22% and the highest for FUZZY-GSA on KDD dataset is 98.94%. FUZZY-GSA achieves the minimum recognition result with feature number 1 but achieves the best result with total 27 features. Although this model also performs well on other features, its performance is as good as 27 features (98%). There is a significant change in the recognition rate between features number 1 and 18, which shows an increase of 1.76%. The role of GSA in improving the Fuzzy recognition rate on the KDD dataset is significant. Here, GSA plays an optimization role and optimizes the Fuzzy classification performance. Figure 6 shows the Fuzzy recognition rate results after optimization by GSA. The system successfully attained a recognition rate of 98.94% using 27 features. This indicates that the system can achieve such high detection rate using only 27 features, and that GSA is capable of minimizing the number of features to the lowest possible count.

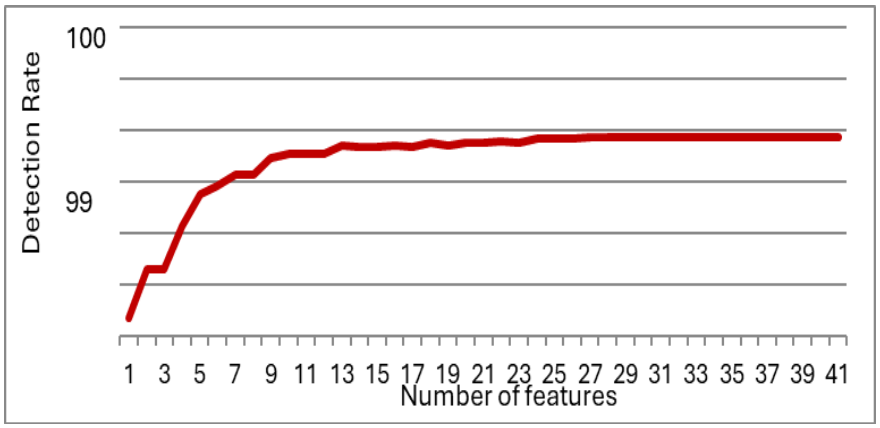

Figure5. Result Diagnosis Fuzzy - GSA in KDD Dataset

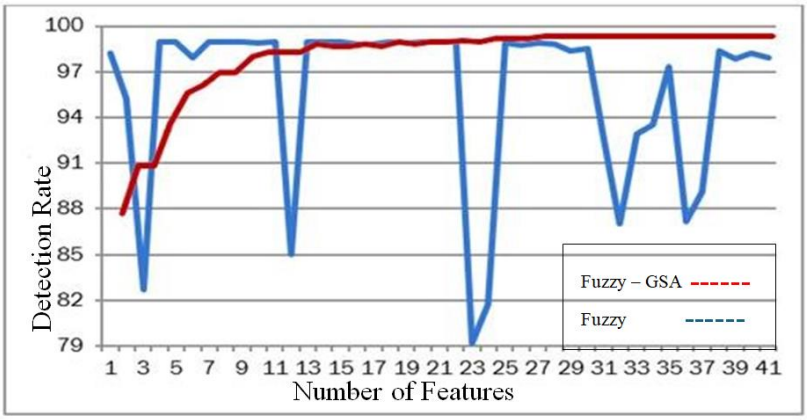

Figure 6. Result Rate Diagnosis Fuzzy With Fuzzy - GSAIn Collection Data KDD

The main limitation with previous methods in intrusion detection systems is their low detection rate combined with a high rate of false detections. To overcome the limitations of IDS, this study utilizes FUZZY-GSA model to improve the performance of intrusion detection systems by increasing the detection rate and reducing false detections. The results from other research, along with the outcomes from the FUZZY-GSA applied in this study, are shown in Table 7, which

compares FUZZY-GSA with various detection algorithms from other researchers using the KDD dataset across each feature. The column represents the detection rate as a percentage, while the rows indicate the number of features in the KDD dataset. Furthermore, Table 7 illustrates the performance of FUZZY-GSA in relation to other algorithms from different researchers concerning the false detection rate within the KDD dataset.

Table 7. Result Rate Diagnosis and false diagnosis rate w ith Number OptimalFeatures

Number of features	False detection rate	Detection percentage	Algorithm
30	1.05	97	**SICLN**
34	2.03	97	**ICLN**
32	0.83	97	**PNN**
41	0.41	96	**FC-FUZZY**
31	1.03	98	**FR2**
29	0.64	90	**LCFS**
41	0.02	98.93	**FUZZY**
27	0.007	98.94	**FUZZY – GSA**

One of the most evident challenges in analysing intrusion detection systems is the number of detection features. An increased quantity of features requires a significant amount of computing resources. To improve the number of these features, the optimization method known as GSA was applied to enhance the systems' performance. However, through the optimization of the intrusion detection system, the optimal number of features was determined. Utilizing fewer detection features offers advantages, such as requiring less computer memory and improving the systems' detection speed. As a result, machine learning techniques like FUZZY are initially used as intrusion detection systems, followed by the application of optimization methods such as GSA to identify the ideal number of detection features. Subsequently, optimization model known as FUZZY-GSA has been developed and assessed in this study. The main goal of the proposed model is to improve the detection rate through optimal pattern recognition. By comparing the results of the FUZZY-GSA model used in this research, the detection rate outcomes with the optimal number of features derived from a total of 27 features are displayed in Table 6. The study demonstrates that the model has shown improvement in both positive detection rate and false detection rate. In specific, it yields a detection rate of approximately 98.94% with a reduced number of features, as well as reducing the false detection rate to
0.007%.

5. CONCLUSIONS AND FUTURE WORK

In summary, FUZZY-GSA demonstrates the highest detection capability while maintaining the lowest false detection rate on the KDD dataset in comparison to other published works. When comparing FUZZY-GSA with other researchers' recognition algorithms on the KDD dataset across each feature, FUZZY-GSA achieves a recognition rate of 98.94%, marking it as the best result for recognition rate, alongside a false positive rate of 0.007%, which is the best result for false positives when using a total of 27 features. FUZZY-GSA attains the highest recognition percentage, representing the best result for recognition rate on the KDD dataset, while also achieving the lowest percentage for false positives in this dataset. For future works, we intend to focus on improving the original methods. This include proposing and using new models and new machine learning algorithm for IDS, by hybridizing new machine learning and optimization algorithms. We also aim to test the proposed machine learning model on other real-world datasets.

REFERENCES

[1] G. Prethija and J. Katiravan, "Machine learning and deep learning approaches for intrusion detection: a comparative study," in Inventive Communication and Computational Technologies: Proceedings of ICICCT 2021, 2022, pp. 75-95.

[2] A. Thakkar and R. Lohiya, "A review on machine learning and deep learning perspectives of IDS for IoT: recent updates, security issues, and challenges," Archives of Computational Methods in Engineering, vol. 28, pp. 3211-3243, 2021.

[3] A. Thakkar and R. Lohiya, "A review of the advancement in intrusion detection datasets," Procedia Computer Science, vol. 167, pp. 636-645, 2020.

[4] Y. Lu, S. Chai, Y. Suo, F. Yao, and C. Zhang, "Intrusion detection for Industrial Internet of Things based on deep learning," Neurocomputing, vol. 564, p. 126886, 2024.

[5] I. F. Kilincer, F. Ertam, and A. Sengur, "Machine learning methods for cyber security intrusion detection: Datasets and comparative study," Computer Networks, vol. 188, p. 107840, 2021.

[6] H. Bostani and M. Sheikhan, "Hybrid of binary gravitational search algorithm and mutual information for feature selection in intrusion detection systems," Soft computing, vol. 21, pp. 2307-2324, 2017.

[7] L. Jin, R. Fan, X. Han, and X. Cui, "IGSA-SAC: a novel approach for intrusion detection using improved gravitational search algorithm and soft actor-critic," Frontiers in Computer Science, vol. 7, p. 1574211, 2025.

[8] A. Dastanpour and A. Farizani, "Improving Intrusion Detection System Using The Combination Of Neural Network And Genetic Algorithm," Available at SSRN 5046099, 2024.

[9] H. V. Vo, H. P. Du, and H. N. Nguyen, "Apelid: Enhancing real-time intrusion detection with augmented wgan and parallel ensemble learning," Computers & Security, vol. 136, p. 103567, 2024.

[10] Z. Ahmad, A. Shahid Khan, C. Wai Shiang, J. Abdullah, and F. Ahmad, "Network intrusion detection system: A systematic study of machine learning and deep learning approaches," Transactions on Emerging Telecommunications Technologies, vol. 32, p. e4150, 2021.

[11] H. Ji, D. Kim, D. Shin, and D. Shin, "A study on comparison of KDD CUP 99 and NSL-KDD using artificial neural network," in Advances in Computer Science and Ubiquitous Computing: CSA-CUTE 17, 2018, pp. 452-457.

[12] Z. Sun, G. An, Y. Yang, and Y. Liu, "Optimized machine learning enabled intrusion detection 2 system for internet of medical things," Franklin Open, vol. 6, p. 100056, 2024.

[13] N. Mishra and S. Pandya, "Internet of things applications, security challenges, attacks, intrusion detection, and future visions: A systematic review," IEEE Access, vol. 9, pp. 59353-59377, 2021.

[14] Y. Jiang, Z. Jia, W. Liu, L. Yang, and H. Liu, "Security identification of abnormal access to network databases based on GSA-SVM algorithm and deep features," in Fifth International Conference on Computer Communication and Network Security (CCNS 2024), 2024, pp. 438-443.

[15] S. Muneer, U. Farooq, A. Athar, M. Ahsan Raza, T. M. Ghazal, and S. Sakib, "A Critical Review of Artificial Intelligence Based Approaches in Intrusion Detection: A Comprehensive Analysis," Journal of Engineering, vol. 2024, p. 3909173, 2024.

[16] F. Amiri, M. R. Yousefi, C. Lucas, A. Shakery, and N. Yazdani, "Mutual informationbased feature selection for intrusion detection systems," Journal of network and computer applications, vol. 34, pp. 1184-1199, 2011.

[17] M. Moorthy and S. Sathiyabama, "Hybrid fuzzy based intrusion detection system for wireless local area networks," Eur. J. Sci. Res, vol. 53, pp. 431-446, 2011.

[18] S. D. Nguyen, V. S. T. Nguyen, and N. T. Pham, "Determination of the optimal number of clusters: A fuzzy-set based method," IEEE Transactions on Fuzzy Systems, vol. 30, pp. 3514-3526, 2021.

[19] T. Sree Kala and A. Christy, "HFFPNN classifier: a hybrid approach for intrusion detection based OPSO and hybridization of feed forward neural network (FFNN) and probabilistic neural network (PNN)," Multimedia Tools and Applications, vol. 80, pp. 6457-6478, 2021.

[20] K. SUPARNA and S. HALIMUNNISA, "Deep Learning Anti-Fraud Model for Internet Loan Where We Are Going," International Journal of Information Technology and Computer Engineering, vol. 12, pp. 269-277, 2024.

[21] A. M. Amine and Y. I. Khamlichi, "Optimization of Intrusion Detection with Deep Learning: A Study Based on the KDD Cup 99 Database," International Journal of Safety & Security Engineering, vol. 14, 2024.

[22] X. Zeng, "Unmasking Intruders: An In-Depth Analysis of Anomaly Detection Using the KDD Cup 1999 Dataset," in 2024 3rd International Conference on Artificial Intelligence and Computer Information Technology (AICIT), 2024, pp. 1-4.
[23] H. G. Kayacik, A. N. Zincir-Heywood, and M. I. Heywood, "Selecting features for intrusion detection: A feature relevance analysis on KDD 99 intrusion detection datasets," in Proceedings of the third annual conference on privacy, security and trust, 2005.
[24] S. Kumar, S. Gupta, and S. Arora, "A comparative simulation of normalization methods for machine learning-based intrusion detection systems using KDD Cup'99 dataset," Journal of Intelligent & Fuzzy Systems, vol. 42, pp. 1749-1766, 2022.
[25] J. J. Tanimu, M. Hamada, P. Robert, and A. Mahendran, "Network Intrusion Detection System Using Deep Learning Method with KDD Cup'99 Dataset," in 2022 IEEE 15th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC), 2022, pp. 251-255.
[26] A. Shehadeh, H. ALTaweel, and A. Qusef, "Analysis of Data Mining Techniques on KDD-Cup'99, NSL-KDD and UNSW-NB15 Datasets for Intrusion Detection," in 2023 24th International Arab Conference on Information Technology (ACIT), 2023, pp. 1-6.
[27] D. Dwivedi, A. Bhushan, A. K. Singh, and Snehlata, "Improving Network Security with Gradient Boosting from KDD Cup Dataset," SN Computer Science, vol. 5, p. 877, 2024.
[28] M. Zhang, "Effectiveness Evaluation of Random Forest, Naive Bayes, and Support Vector Machine Models for KDDCUP99 Anomaly Detection Based on K-means Clustering," in ITM Web of Conferences, 2025, p. 04010.
[29] Y. Han, "Research on Optimization of Network Intrusion," 2025.
[30] D. D. Protić, "Review of KDD Cup '99, NSL-KDD and Kyoto 2006+ datasets," Vojnotehnički glasnik/Military Technical Courier, vol. 66, pp. 580-596, 2018.

AUTHORS

Amin Dastanpour received his B.Sc. degree in Computer Science from the Kerman University of Iran and M.Sc. degree in Information Technology from University Putra Malaysia. Ph.D Degree in Advance Information Technology Special in Network Security from University Technology Malaysia. he is a vice chancellor of Research in Kerman university, and Faculty member of computer Science in Kerman University at Iran. his research interests include Intrusion Detection System, network security and machine learning.

Raja Azlina received her B.Sc. degree in Computer Science from the University of Michigan and M.Sc. degree in Software Engineering from University Technology Malaysia. She is a faculty member of the Faculty of Computer Science and Information Technology, University Putra Malaysia, and is currently pursuing her Ph.D. in network security. Her research interests include wireless network, network security and machine learning.