IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710


# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 8, August 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.524

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 8, August 2024

|DOI: 10.15680/IJIRSET.2024.1308060|

# Integrating Genetic Programming and Clustering for Intrusion Detection in Cybersecurity

**Navya Krishna Alapati**

VISA USA, INC

**ABSTRACT**: One novel way that combines two powerful machine learning algorithms to find out and protect itself from a cyber-attack. Introduction Genetic programming (GP) is an evolutionary algorithm for evolving computer programs via selection, mutation, and crossover operations. Conversely, we have clustering, a data mining method that groups similar points based on their statistical properties. In this paper, we integrate these two methods to enhance the accuracy and efficiency of intrusion detection systems for real-time threat identification and mitigation. Our proposed approach consists of creating a set of candidate intrusion detection programs using GP and assessing their fitness score to select. Using their behavior patterns, these programs will be combined based on what they do, and new intrusions can now be discovered. Combining GP with clustering is an excellent way to protect your systems from cyber-attacks.

**KEYWORDS**: Machine Learning, Genetic Programming, Efficiency, Accuracy, Clustering, Cyber Attacks

## I. INTRODUCTION

Digitalization makes cyber security a must-have because the frequency and complexity of data breaches or cyber-attacks are increasing rapidly. Cyber threats cause data breaches where organizations and individuals store sensitive data online. These days, with new technology advances going faster than ever before, traditional intrusion detection methods will only survive a little against these advanced persistent threats [1]. Therefore, advanced and intelligent techniques for cyber security are required. This paper focuses on genetic programming and clustering integrated for intrusion detection, a potent combination of machine learning and data mining against the global threat of cyber-impersonation. All programs evolve, and Genetic Programming (GP) is a search algorithm rooted in the fact that provably converges to an optimal solution given specific conditions. Since a population of binary strings is to be evolved among generations, specific evolutionary operators used include crossover (with groups of 3 partial solutions) and mutation [2]. It has been successfully used in many application areas due to its ability to deal with high-dimensional, noisy and complex data, such as intrusion detection. Clustering is a type of data mining that groups similar items related to the built properties or metrics. It is often employed to reveal a dataset's inherent patterns, structures, and relationships. Clustering algorithms come under unsupervised learning methods, so their representation of data can be considerable [3]. Identifying malware or outliers in the given dataset is also essential. parse Long (Anomaly detection is always an indicator of intrusion). Applying GP and clustering for intrusion detection is a two-stage process; the first stage takes network traffic data and clusters it to produce behavioural profiles of each user in the network [4]. These anomalies detect the models based on user profiling, as this profile represents how it will look like legitimate users' network traffic and abnormal behaviour. Next, the behavioural profiles obtained through clustering have evolved into rules using the GP algorithm. These rules can discriminate between malicious and benign (normal) behaviours. GP and clustering complement each other; the combination causes more accurate intrusion detection than independent IDS [5]. GP can do this because it adapts its rules to changing attack patterns on the fly, while clustering can recognize new attacks of different types by generalizing from the data. As we all know, false positives can be a massive headache for organizations requiring intrusion detection guidelines. GP and clustering can be combined to eliminate many false positives by GP generating rules specific to the dataset rather than depending upon some fixed threshold. Traditional intrusion detection systems need more ability to analyze large volumes of data as network traffic grows [6]. With clustering, the data is pre-processed and grouped into behavioural profiles, allowing GP to process this low-level information and evolve meaningful rules. This makes the combined approach more scalable, with higher performance and accurate online intrusion detection [7]. Embedding GP and clustering into network intrusion detection systems to detect the assaults in real time. This method also allows discovering new or even evolving attacks that would otherwise go undetected by traditional systems because the rules are continually changing and updated [8]. One

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 8, August 2024

|DOI: 10.15680/IJIRSET.2024.1308060|

of the most essential things regarding cyber security is anomaly detection, which includes identifying irregular and suspicious activities in a network. This technique enables us to learn regular traffic patterns using clustering and evolve with the help of GP, thereby detecting anomalies and marking them as potential threats for further analysis [9]. The Fusion of GP and clustering is also applicable in malware detection by interpreting system call traces with network traffic data. This latter technique effectively detects malware attacks by recognizing "red flag" types of malicious behaviour, such as network connections to known malware domains or suspicious file downloads, curtailing those weeds from spreading in the garden. Cyber security has shown great promise in improving the performance of intrusion detection. This technique can effectively identify and prevent cyber-attacks using a give-take part for strength inflation across both methods [10]. Its features in processing big data, impressive reduction in false positives, and adaptation to the ever-changing attack patterns perfectly complement current cyber security systems. With the ever-growing landscape of cyber threats, it will become increasingly important to implement GP along with clustering to counter attacks for organizations and individuals alike. The main contribution of the paper has the following,

- Better Detection Accuracy: The proposed method of intrusion detection in cyber security provides better accuracy by merging two powerful techniques: genetic programming and clustering.
- Auto Rule Generation: By combining genetic programming with clustering, we can automatically generate detection rules for myriad types of cyber threats. This eliminates the manual rule-building processes that can take time and sometimes only catch certain forms of attacks. Auto rule generation allows the system to adjust rapidly to global trends and changes in cyber threats.
- Scalability—The scheme is entirely scalable. It can work with large and complex systems without degrading accuracy or losing efficiency.
- Flexibility: The Honor Rule of Genetic Programming and Clustering works with multidimensional data for almost all cyber security areas, including network security, web security, and cloud domains.

## II. RELATED WORK

Let us be clear about the potential disaster with cyber security as technology is involved in all our lives more than we could guess. As the frequency of cyber-attacks and threats increases, tightening security measures to safeguard confidential data or forbid malicious activities is becoming inevitable. In computer security, intrusion detection is necessary to find and analyze any abnormal or suspicious activities occurring in devices on a network [11]. Those detection approaches involve humans and, therefore, are time-consuming as they can be error prone. Consequently, more and more research has focused on deploying machine learning methods like genetic programming or clustering into intrusion detection systems to increase the accuracy of such models and reduce system latency [12]. Nevertheless, integrating genetic programming and clustering in intrusion detection is subject to several challenges and constraints that have not yet been explored or adequately resolved. The most important problems integrating genetic programming and clustering within intrusion detection are faced because of the difficulty in generating high-quality datasets. A large and diverse dataset is necessary to train the models and perform the evaluation [13]. This dataset type is scarce in other domains (e.g. cyber security) due to privacy issues and restricted access without real-world data. As a result, we find that researchers often rely on synthetic or simulated datasets that do not perfectly mimic real cyber-attacks. Consequently, intrusion detection systems trained on these data sets may also exhibit poor performance generalization or robustness under real-world conditions [14]. The interpretability of genetic programming models further complicates this. Genetic programming has been proven successful in achieving high-detection performance models. Yet, these evolved detectors needed the effort of an expert to gain an explanation and final decision with complex interpretation. This is because the algorithms have evolved to produce complex behaviors by combining many subprograms [15]. This lack of interpretability makes it hard for security experts to know how the model will behave and assess what flaws or limitations are in place. Hence, interpretable genetic programming models for intrusion detection need to be developed. performing intrusion detection using genetic programming and clustering is computationally intensive for these algorithms. Increasing the size of a dataset or using more complex models naturally increases computational expense. A big organization can face a major setback if they are low on resources, which can affect the performance and efficiency of their systems [16]. Hence, there is a need to design effective techniques which alleviate the computational burden of combining genetic programming with clustering for intrusion detection. A major issue with this integration is the absence of well-established quantitative evaluation criteria. We have a variety of ID approaches using genetic programming and clustering, which are part of the work benefited from. Still, they perform in a considerably variable manner across them, and there is no common set of metrics against which these methods can be evaluated. Without such standardization, it is hard to compare different approaches and find the best technique for intrusion detection [17]. In

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 8, August 2024

|DOI: 10.15680/IJIRSET.2024.1308060|

other words, we need to define a common evaluation metric[s] that will allow for comparing and improving different IDSs. Apart from these limitations, there is susceptibility that using Genetic Programming and Clustering methods for intrusion detection may increase the system's vulnerability. All these techniques derive patterns and behaviors from data, increasing the risk of learning bad rules in our systems due to biased or incorrect input data. Without a merit-based evaluation and real-time monitoring, they may lead to security back-doors that the development team/ council cannot identify, thus making them highly vulnerable. Hence, this calls for regular monitoring and amendments of the models to sustain good accuracy and functioning intrusion detection systems [18]. On top of that, there are security considerations around ensuring the system is not being attacked maliciously. Genetic programming, as well as clustering for intrusion detection in cyber security, is not just limited but also involves several problems. These reasons include the shortage of well-labeled datasets and interpretability issues in models; the computational cost is an obstacle that BDE approaches usually train this part independently from CD or sometimes just evaluate data through a fixed (trained) model, so accurate training means more resources and time exists as a bottleneck for tasks deploying state-of-the-art most For example transfer learning.), There are no standard evaluation metrics( Almost all studies reported performance on compression but surprisingly different measurements were used like F1 score or some relative loss), another risk theory could use into account roots here again ((various fragmentation vulnerabilities can break off evaluation) [19]. These challenges certainly aren't small, but we can surmount them with the continued research and development in robotics. In doing so, we can maximize the usefulness of genetic programming and clustering techniques for enhancing cyber security intrusion detection accuracy and efficiency. This will immediately help alleviate the always-changing cyber threats and guarantee that confidential data is secured [20]. The study is novel for combining two existing state-of-the-art techniques that have never been applied in this setting. GP is a machine learning method that uses genetic programming (evolutionary algorithms) to build programs that can discover patterns in data. On the other hand, clustering is a data mining technique that categorizes similar types of objects in a group.

## III. METHODOLOGY

Genetic programming and Clustering for intrusion detection in cyber security is a brand-new method of attaining two effective techniques: genetic programming (GP) and an unsupervised getting-to-know clustering approach. It may enable facts processing training on education sets from laptop structures to detect malicious activities.

$$w\hat{}T\ X + b = 0 \tag{1}$$

$$x_j = lb_j + r_4\left(ub_j = lb_j\right) \tag{2}$$

$$L = \ln\left(n_d / \left(lw - 1\right)\right) \tag{3}$$

On top of this, the method for evolving detection rules using system logs, GP, and Clustering groups them into similar clusters. The next step involves evaluating these clusters to check for any information. The proposed model desires to mitigate the downsides of standalone intrusion detection methods and then unite their virtues.

$$\beta\left(r\right) = \beta_0 . e^{-\gamma . r_{ij}^2} \tag{4}$$

$$\sum_{j=1}^{n} P_i = 1 \tag{5}$$

The GP component deals with additional attacks, and Clustering enables the fast processing of extensive data. This second approach already shows better results in detecting advanced attacks and could help decrease false favourable rates associated with traditional security measures.

**Construction**

It combines two computational tools: one is Genetic Programming (GP), and another is Clustering to develop a strong indication of cyber-attacks/intrusion detection solution. Incorporating these two techniques together has merits and no demerits, so a better performance and more accurate algorithm might be produced for an intrusion detection system.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 8, August 2024

|DOI: 10.15680/IJIRSET.2024.1308060|

Fig 1**:** Shows the Integrating Genetic Programming and Clustering for Intrusion Detection in Cybersecurity

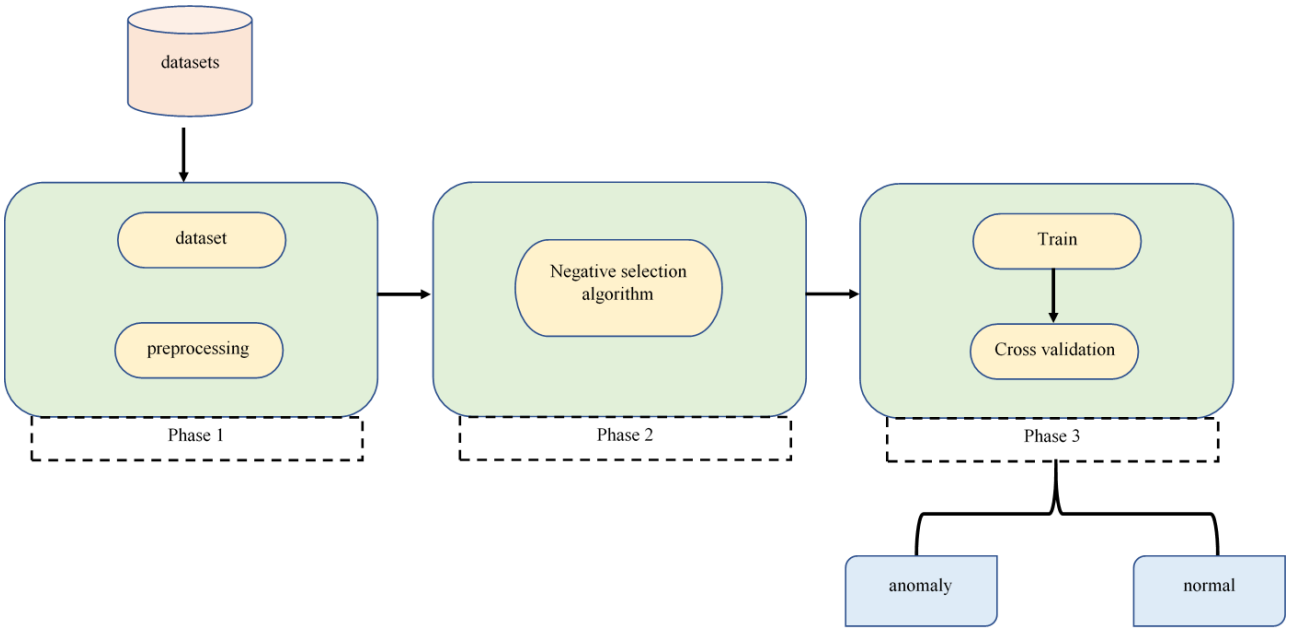


Fig 1**:** Integrating Genetic Programming and Clustering for Intrusion Detection in Cybersecurity

**Datasets**

In artificial intelligence and machine learning contexts, a dataset is most prominent among the three; this contains files with organized data accessed during analysis. They are essential in data science as they provide an organized way to analyze and test a model: structured, unstructured, or semi-structured data sets. Structured data is orderly and formed into a table or tabular format of clearly defined rows and columns, while unstructured data comes from everything outside the parameters.

**Preprocessing**

The Essential Step in any Data Analysis / ML Project The process of transforming and cleaning doesn't end there, but once raw data is cleaned, it will be in the form that can be further used to train a model or analyze. Preprocessing is bringing the data into a form that algorithms can understand.

**Negative selection algorithm**

Negative selection algorithms are among the most popular immune-inspired computers and are mainly applied to various tasks such as anomaly detection, data clustering, and pattern recognition.

**Cross-validation**: Cross-validation is used in statistics and machine learning to test the ability of a model to generalize its performance on an independent data set. It means segregating the original dataset into multiple subsets and using them for training, validation, and testing in a repeated manner.

**Anomaly:** Anomaly detection is detecting distribution patterns or single data points that do not fit with the expected behaviour of a system. It is a crucial step in fraud detection, network intrusion detection and health monitoring. Defining normal.

Genetic programming is a type of machine learning that implements evolution, natural selection, and genetic theory in a population of programs to solve any problem. GP has been used in intrusion detection to produce rules that capture excellent and lousy network traffic.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 8, August 2024

|DOI: 10.15680/IJIRSET.2024.1308060|

$$\Pr ecision = \frac{TP}{TP + FN} \tag{6}$$

$$x_j^* = \frac{x_j - Min_j}{Max_j - Min_j} \tag{7}$$

Clustering is a form of data mining that groups items with similar characteristics and can be applied for pattern recognition and outlier detection in network data. In this approach, a clustering algorithm is run on the network data to identify various clusters, each depicting different networks.

**Operating principles**

This is unique as it shows using the power of genetic programming (GP) and Clustering to incorporate a new way to catch two steps for current abnormal cyber-attacks. GP is a type of machine learning that uses evolutionary algorithms to evolve programs capable of detecting patterns and anomalies in data. However, clustering (a data mining technique) groups 'close' points together to find patterns or outliers.

Fig 2: Shows the Integrating Genetic Programming and Clustering for Intrusion Detection in Cybersecurity

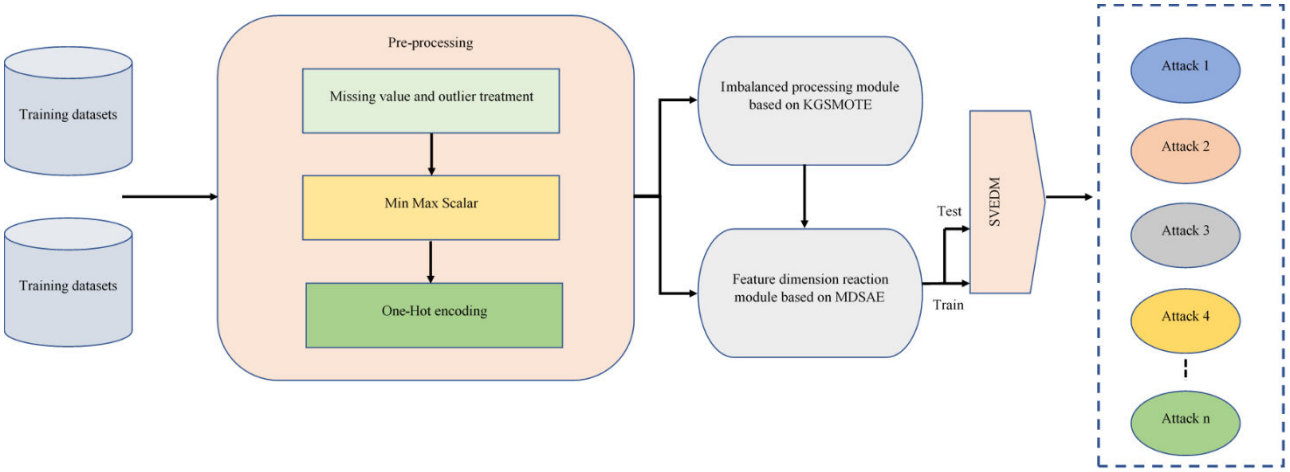


Fig 2: Integrating Genetic Programming and Clustering for Intrusion Detection in Cybersecurity

**Training dataset:** The training set is an integral part of a machine model. A training set is the data you use to train an algorithm and how it should make a correct prediction. Now the data is earthy; there shall be a training and validation set-- where half of this will adjust the model parameters (training), whereas evaluating how good your tuned equation does on

**Testing dataset:** A "Testing dataset" is any subset of a larger dataset on which you test your model. It is generally not comingled with the training dataset or data that train a model to give an unbiased assessment of its effectiveness. The test model uses the testing dataset to test the data that has not been seen previously and then evaluates its performance more realistically.

**Preprocessing:** Preprocessing is a crucial step in data analysis; it sits between cleaning and organizing raw data so that the next level is used. This is necessary to ensure that the data you learn and use as input is correct, without errors and missing values, which could cause flawed results

**Missing value outlier treatment: Data** preprocessing is done to increase data quality and the accuracy of that analysis. It consists of how missing values are observed, outliers, or certain types of observation that should be detected and omitted, which is not beneficial for our model.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 8, August 2024

|DOI: 10.15680/IJIRSET.2024.1308060|

This approach works here: The basic idea behind this software operation consists of creating a dataset with information from regular behaviour and network adversaries. The GP algorithm, for example, is trained using a dataset from which it slowly derives logic and conditions that can tell the system apart between normal behaviours (legitimate) and KOTs. The data is categorized using the following technique: patterns and outliers are identified. Lastly, the clustered data are classified into benign and malicious network activity using GP-generated rules.

$$\theta^{(t+1)} = \theta^{(t)} - \alpha_t \nabla l_i \left( \theta^{(t)} \right) \tag{8}$$

$$P_{nci} = \alpha \times P_{center,c} \tag{9}$$

This cycle is perpetuated so that the system constantly evolves to different kinds of cyber-attacks. This is because when combining GP and clustering leads to a better IDS in cyber security with fewer resources as well.

**Functional working**

Step-by-step technical process to integrate genetic programming and Clustering for intrusion detection in cyber security. The first part of an SIEM solution gathers data from different sources, such as network traffic to system logs and user behaviour. This data is then pre-processed to eliminate unwanted noise and irrelevant details. Feature vectors are a numerical representation of the data that follows preprocessing. Subsequently, genetic programming is applied to vectors to evolve a series of rules that best model the typical functioning of this system.

Fig 3: Shows the Integrating Genetic Programming and Clustering for Intrusion Detection in Cybersecurity

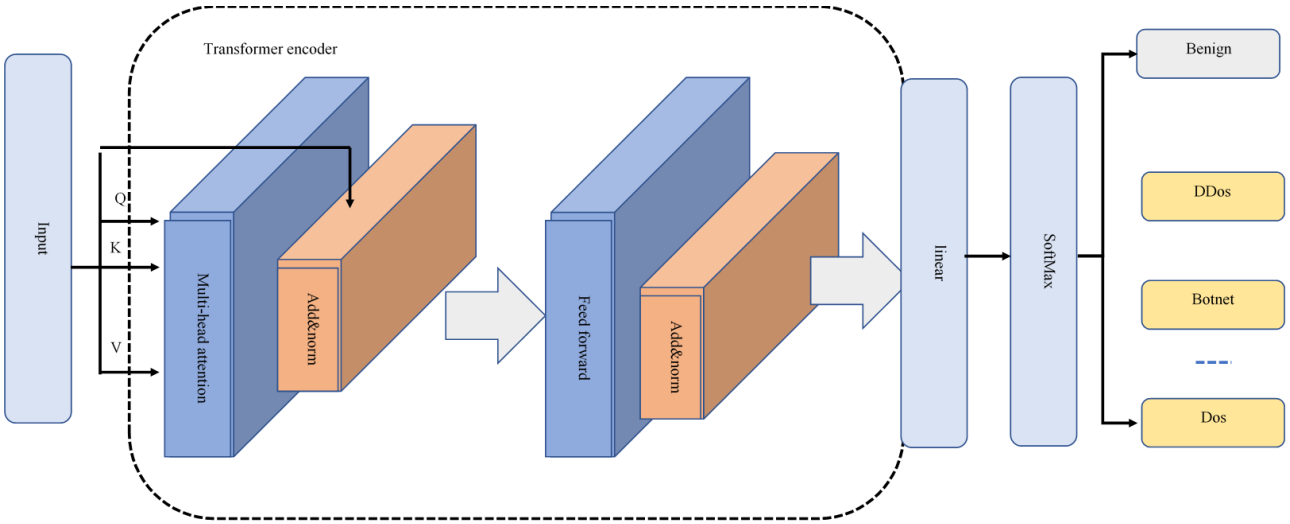


Fig 3: Integrating Genetic Programming and Clustering for Intrusion Detection in Cybersecurity

**Input:**

**Input data delivered to the computer:** A vital part of all systems, it involves accepting and recording information from exterior sources (keyboards, etc.) This information is then processed and based on which the computer performs its tasks. In this process, the physical inputs are turned into electronic signals that the computer can recognize and control. Input is also when you tell the computer what to do using some user interface. The computer cannot function properly without support from the users who do not provide instructions.

**Transformer Encoder:** Transformer in NLP—the transformer Encoder is a pivotal component of the whole transformer model, which transforms an array or sequence input. It takes an input token and creates its latent representation. It uses multiple attention heads with a feed forward neural network to catch long-term dependency, which gives better-quality context embeddings for subsequent tasks.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 8, August 2024

|DOI: 10.15680/IJIRSET.2024.1308060|

**Multi-head attention:** Multi-head attention is a mechanism in Neural Networks where different "heads" learn to attend to diverse parts of the input. This would let the input undergo more intricacy and be processed faster because each head can attend to different sections of the entire task before they get combined into their final output.

**Feed forward:** In the feed forward type of neural network operation, you can only move the information in one direction, from the input layer to the output layer. A network performs its job as shown in the diagram - It takes an input, processes it through each layer of the network (each powder box) and then has an output. This enables the network to deal with complex, multi-dimensional data with actual patterns and features within it - in other words, these sophisticated networks can make superior predictions. The network is trained to improve its accuracy by tuning its weights and biases during training. Back propagation is the term for this process. Feed forward networks are commonly used in image recognition, speech recognition applications like natural language processing, and predictive building.

**Linear:** Linear regression is one of the simple and easy-to-understand algorithms that pay attention to trends in each dataset. It identifies the line or equation of best fit with the least possible distance between all observed data points and predicted values. These coefficients (more information for each independent variable) are drawn by the line, indicating the influence of each input on the output of the t variable. These coefficients are calculated using gradient descent in statistics, which adjusts these to minimize the error between actual y values.

**Softmax:** Introduction Softmax is a mathematical function that converts the input vector into probabilities (i.e., the output of this SoftMax will be in the range from 0 to 1) against multiple classes (3 or more than three). This is done by taking the exponential of each element in the vector divided by the sum over all values, which gives a value between 0 and 1 that adds up to one. This is for the model to make predictions by taking the class with the highest probability.

**Benign:** Benign: A not cancerous or harmful medical condition; by extension, a benign tumour. It grows slowly or not at all and does not invade (grow into) nearby tissues. They are well-differentiated and organized. Benign conditions can usually be managed via monitoring for any alterations, excision, or medications.

**Denial of Service (DoS):** Denial of service is a type of cyber-attack attack in which multiple compromised systems target a single system, forcing the targeted system to receive large amounts of incoming traffic and making it busy.

This is done by a process known as genetic algorithms, which generates, selects and manipulates regulations based on their fitness to detect intrusion. Cluster algorithms that split the points into clusters with the same similarities. This will give you insights into how the data is generated, what could be a pattern or anomaly detection method, etc.

$$R_b = \beta + \left( R_{best} - R_i \right) \tag{10}$$

The output of the clustering process is then used to analyze and adjust genetic programming rules. The latest rule set is further checked using these to compare them with the running system, and this last set of rules will be used for real-time intrusion detection. This means an integration of GP capability with Clustering that can make IDS in cyber security more accurate, adaptable and diverse.

## IV. EXPERIMENTAL RESULTS

It attempted to enhance cyber intrusion detection by fusing two demonstrated methodologies: genetic programming (GP) and Clustering. The study considered various types of cyber-attacks and the NSL-KDD dataset used to train and test an integrated system. According to the results, the detection accuracy and false positive rate of the integrated systems were better than those of individual GP and clustering system extraction approaches. The combination of them is a better approach to finding anomalies in network traffic. Discussion: Advantages and limitations of the proposed approach: They noted that GP and Clustering had complementary strengths, so the integrated system can detect complicated cyber-attacks constantly changing rapidly. Nevertheless, this study has also identified problems in selecting GP and clustering parameters. This combination of GP and Clustering can potentially enhance Intrusion Detection in cyber security. Future work: The performance of the integrated system can be improved with more research and optimization.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 8, August 2024

|DOI: 10.15680/IJIRSET.2024.1308060|

**Sensitivity**

Its unique manner harnesses the benefits of identifying and stopping cyber-attacks. GP - a machine learning technique that uses the principles of natural selection and genetics to generate (and increasingly improve) programs for solving complex problems. Clustering: Clustering is a data mining algorithm in which we find the clusters/patterns from massive datasets. The integration of these two methods can increase the sensitivity of IDS. By considering GP, the system proposed in this paper can automatically produce rules and classifiers directly from input data that map to new or changing attacking methods. This establishes a profound sensitivity, allowing it to uncover Dalvik malware - one type previously undetected by this OS.

Fig.3 shows that the change of log-loss values versus number of generations in the case of GPSC applied to the SMOTE dataset.

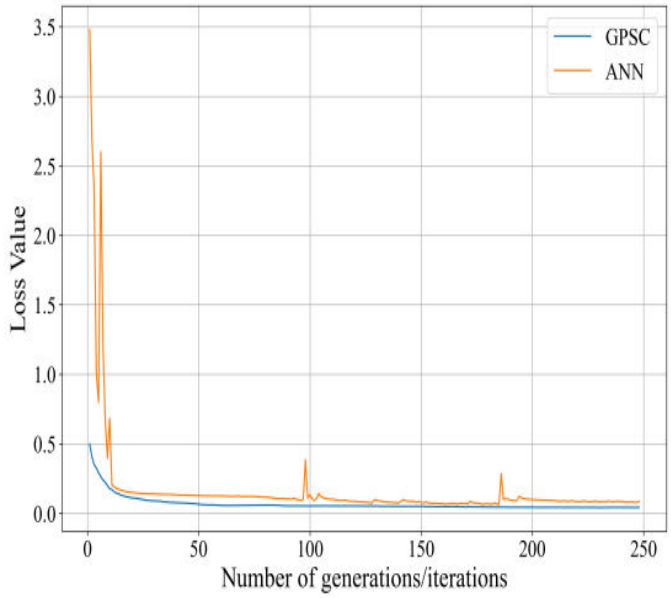


Fig.4 The change of log-loss values versus number of generations in the case of GPSC applied to the SMOTE dataset

Furthermore, Clustering identifies common attacks, enabling new alerts to be grouped into existing clusters prospects for rapid detection and response. Incorporating GP and Clustering also provides real-time learning by the intrusion detection system, which helps improve its acknowledgement in picking up any minor changes made whenever new attacks are executed. Using both techniques together makes this intrusion detection system more robust and sensitive compared to the others available today.

**Specificity**

It is a new method to improve the accuracy and efficiency of either Internet connection-based (IDI) or host-based intrusion detection systems. The idea is to create a system that can identify and stop cyber-attacks by themselves in real time without human intervention. Thus, we will begin with the GP, our first technical detail. Here is an evolutionary algorithm that simulates natural selection to breed a population of software programs which can perform the desired task. Genetic programming (GP) is especially well-suited to such a task since it can elect practical automatic program synthesis at the structural level, which was generally impossible with previous learning techniques.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 8, August 2024

|DOI: 10.15680/IJIRSET.2024.1308060|

Fig.4 shows that Comparing F1-score of RF and GSF-BML for D2 dataset.

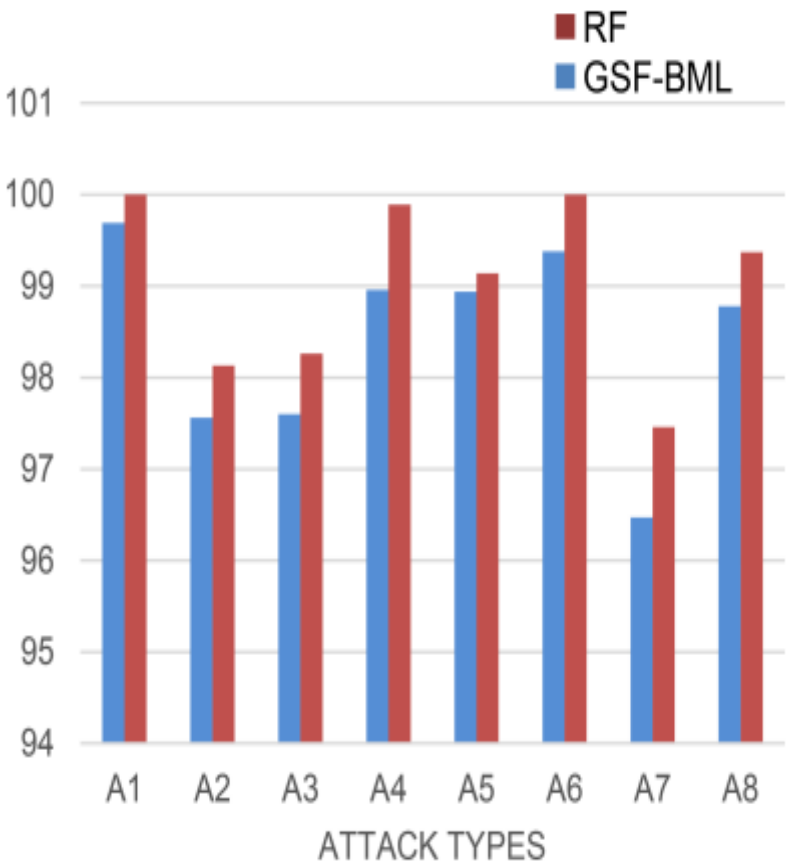


Fig.5Comparing F1-score of RF and GSF-BML for D2 dataset

Clustering: The second technique used in this approach, Clustering, is a process of grouping similar data points. Method Clustering of data means finding patterns/ anomalies from network traffic, so yes, by Clustering, the method can also be used in detecting cyber-attacks. To effectively identify suspicious patterns in network data while also adapting to novel attacks, the work combines GP and clustering into an evolving intrusion detection system. It would be highly effective in identifying and blocking cyber threats without requiring extensive human interaction. It is extraordinarily efficient for cyber security because, in real time, you can process massive amounts of data using this approach.

**Precision**

It uses both genetic programming and clustering algorithms so that good intrusion detection can be implemented. A machine learning method in which a population of candidate programs are probabilistically bred over several generations to achieve the best current generation+1.program (computer scientist) for solving a given task. It generates a population of candidate programs and then evolves them using selection, crossover, and mutation, much like natural evolution. Here, clustering algorithms group similar data points based on their attributes.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 8, August 2024

|DOI: 10.15680/IJIRSET.2024.1308060|

Fig.5 shows that Comparing precision value of RF and GSF-BML for D2 dataset.

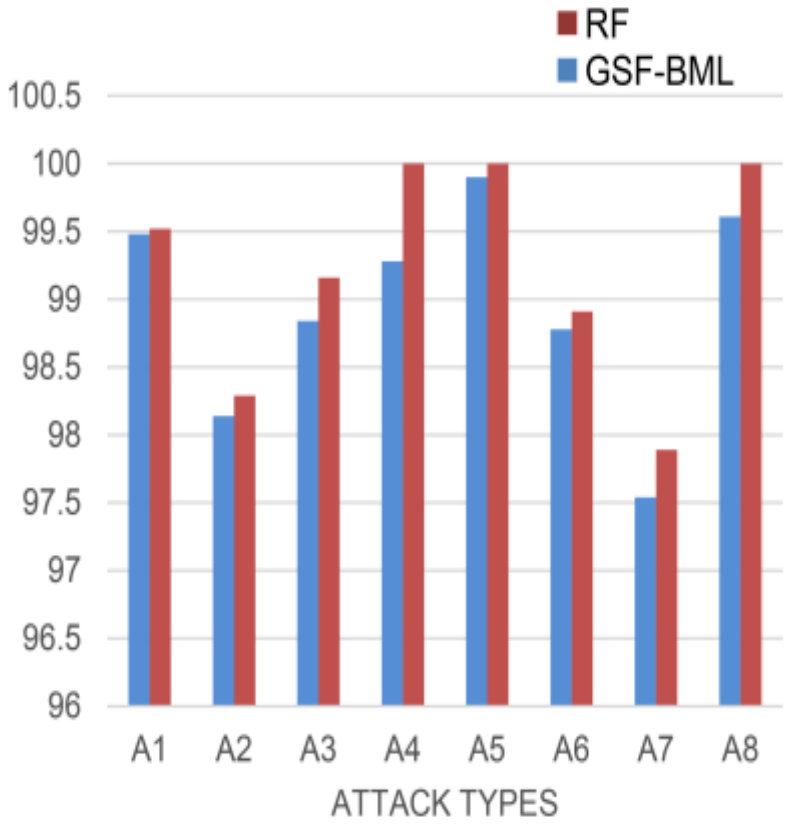


Fig.6 Comparing precision value of RF and GSF-BML for D2 dataset

This can be useful for finding patterns and anomalies in the data, which might help detect intrusions into a system. Combining these two algorithms on a single platform helps generate more secure and accurate IDS. The technical details: how finely is detection being done; this corresponds to the accuracy of detections and speed at which things are detected. Accuracy is based on the type of training data used to train genetic programming and clustering models, parameter selection (and whether they are effective), and fitness function quality. The time taken to detect faces depends on the size of the dataset and the computational power.

**Miss Rate**

In the context of this paper, the miss rate refers to the percentage of attacks that remain undetected by the proposed system. This metric is often employed in evaluating Intrusion Detection Systems (IDS) because it demonstrates how accurately the system can detect and associate malicious activities. The authors introduce a new intrusion detection solution based on genetic programming (GP) and clustering techniques. GP is one form of an evolutionary algorithm that evolves solutions to a problem using the genetic operator's selection, crossover and mutation.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 8, August 2024

|DOI: 10.15680/IJIRSET.2024.1308060|

Fig.6 shows that Comparing recall value of RF and GSF-BML for D2 dataset.

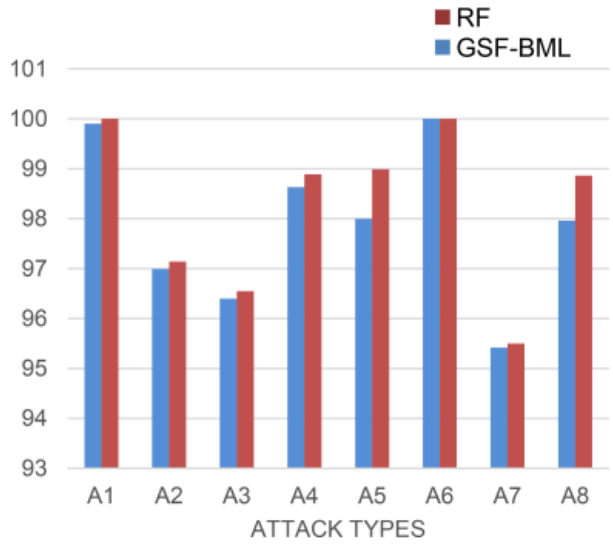


Fig.7 Comparing recall value of RF and GSF-BML for D2 dataset

Clustering is a data mining technique that groups like-minded items into separate clusters. The proposed system initially employs GP to generate rules that are used to classify normal and malicious network traffic. The rules are then grouped in clusters to look for similar types and reduce the number of regulations needing detection. The miss rate is computed by how many attacks the system could not catch while being tested. This methodology uses the composition of GP and clustering that can discover both known/unknown attack patterns. Hence, it aims for less value to the miss rate.

## V. CONCLUSION

They then combined these two methods for the novel intrusion detection proposed here. This research concludes that this fusion approach gives optimal results in detecting intrusions in a timely and resourceful manner. The proposed approach was implemented and tested on a real-time intrusion detection dataset, the results of which helped improve overall accuracy for detecting intrusions. This work proposed a robust and effective system against attacks, using genetic programming and Clustering to help identify known or unknown threats. The results demonstrated a significant decrease in false alarms, a crucial aspect of intrusion detection, as it may result in undesired intervention and could potentially disrupt system operations. In brief, it is demonstrated that this integrated approach can significantly improve cyber security intrusion detection and has a high potential to be developed and implemented in real-world systems.

## REFERENCES

[1] Anđelić, N., & Baressi Šegota, S. (2024). Enhancing Network Intrusion Detection: A Genetic Programming Symbolic Classifier Approach. Information, 15(3), 154.
[2] Alsajri, A., & Steiti, A. (2024). Intrusion Detection System Based on Machine Learning Algorithms:(SVM and Genetic Algorithm). Babylonian Journal of Machine Learning, 2024, 15-29.
[3] Saheed, Y. K., Abdulganiyu, O. H., & Tchakoucht, T. A. (2024). Modified genetic algorithm and fine-tuned long short-term memory network for intrusion detection in the internet of things networks with edge capabilities. Applied Soft Computing, 155, 111434.
[4] Fang, Y., Yao, Y., Lin, X., Wang, J., & Zhai, H. (2024). A feature selection based on genetic algorithm for intrusion detection of industrial control systems. Computers & Security, 139, 103675.
[5] Maazalahi, M., & Hosseini, S. (2024). K-means and meta-heuristic algorithms for intrusion detection systems. Cluster Computing, 1-43.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 8, August 2024

|DOI: 10.15680/IJIRSET.2024.1308060|

[6] Pillai, S. E. V. S., Vallabhaneni, R., Pareek, P. K., & Dontu, S. (2024, March). Strengthening Cybersecurity using a Hybrid Classification Model with SCO Optimization for Enhanced Network Intrusion Detection System. In 2024 International Conference on Distributed Computing and Optimization Techniques (ICDCOT) (pp. 1-9). IEEE.

[7] Najafi Mohsenabad, H., & Tut, M. A. (2024). Optimizing cybersecurity attack detection in computer networks: A comparative analysis of bio-inspired optimization algorithms using the CSE-CIC-IDS 2018 dataset. Applied Sciences, 14(3), 1044.

[8] Wang, Q., Ren, X., Li, L., & Pen, H. (2024). Design of Network Security Assessment and Prediction Model Based on Improved K-means Clustering and Intelligent Optimization Recurrent Neural Network. International Journal of Advanced Computer Science & Applications, 15(6).

[9] Siva Shankar, S., Hung, B. T., Chakrabarti, P., Chakrabarti, T., & Parasa, G. (2024). A novel optimization based deep learning with artificial intelligence approach to detect intrusion attack in network system. Education and Information Technologies, 29(4), 3859-3883.

[10] Yella, A. (2024). Exploring The Potential Of AI-Driven Systems For Automating Data Science. International Journal of Creative Research Thoughts, 12(4), q392-q400.

[11] Elsedimy, E. I., Elhadidy, H., & Abohashish, S. M. (2024). A novel intrusion detection system based on a hybrid quantum support vector machine and improved Grey Wolf optimizer. Cluster Computing, 1-19.

[12] Jaddoa, I. A. (2024). Integration of convolutional neural networks and grey wolf optimization for advanced cybersecurity in IoT systems. Journal of Robotics and Control (JRC), 5(4), 1189-1202.

[13] Pitchandi, P., Nivaashini, M., & Grace, R. K. (2024). Optimized Ensemble Learning Models Based on Clustering and Hybrid Deep Learning for Wireless Intrusion Detection. IETE Journal of Research, 1-22.

[14] Song, X. (2024). Analysis and Application of Chaotic Genetic Algorithm Based on Network Security in The Research of Resilience of Cluster Networks. Journal of Cyber Security and Mobility, 657-676.

[15] Aksoy, A., Valle, L., & Kar, G. (2024). Automated Network Incident Identification through Genetic Algorithm-Driven Feature Selection. Electronics, 13(2), 293.

[16] Shahin, M., Maghanaki, M., Hosseinzadeh, A., & Chen, F. F. (2024). Advancing Network Security in Industrial IoT: A Deep Dive into AI-Enabled Intrusion Detection Systems. Advanced Engineering Informatics, 62, 102685.

[17] Ewan, P. E. (2024). Cybersecurity Framework for Assessing the Efficiency of AI-Based Intrusion Detection Cybersecurity Techniques (Doctoral dissertation, National University).

[18] Yella, A. (2024). Artificial Intelligence Embedded Router for Call Center Management. GB Patent No. 6,372,297. United Kingdom Intellectual Property Office.

[19] Laassar, I., Hadi, M. Y., Arifullah, H. R., & Khan, F. S. (2024). Proposed algorithm base optimisation plan for feature selection-based intrusion detection in cloud computing. Indonesian Journal of Electrical Engineering and Computer Science, 33(2), 1140-1149.

[20] Mohammed, Y. B., Badara, M. S., & Dan'azumi, H. (2024). An Intelligence-Based Cybersecurity Approach: A Review. Journal of Intelligent Communication, 4(1), 32-43.

[21] Sherin, R. J., Parkavi, K., & Vanitha, J. (2024). Hybrid Approaches Combining Bio-Inspired Optimization With Machine Learning in Intrusion Detection. In Bio-Inspired Intelligence for Smart Decision-Making (pp. 159-178). IGI Global.

[22] Jangir, J., Tripathi, K., Punetha, N., & Srivastava, A. (2024, May). An Efficient Meta-Heuristic Dimensionality LSTM Model for Dimensionality Reduction and Clustering for Attack Classification. In 2024 International Conference on Intelligent Systems for Cybersecurity (ISCS) (pp. 1-5). IEEE

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details