Novel AI Method for Prediction of Cyber Security Attacks through Analysis of Windows Event ID Using Ants Colony Optimization & ANN - EFBPA

S D Jadhav
Research Scholar,
Shri Guru Gobind Singhji Institute of Engineering & Technology, Nanded, India.

Dr. B R Bombade
Associate Professor,
Shri Guru Gobind Singhji Institute of Engineering & Technology, Nanded, India.

Abstract *—Cyber attacker tries to exploit the loop holes of various sub-systems of cyber security. Logs plays a crucial role in identification of anomalous behaviour. Windows event ID logs records all the events occurring in the system. Analysis of Windows event ID logs helps to alerts the system administrator about possible cyber-security attacks in real time. But analysis of such logs is itself is a huge challenge as big organization have thousands of machines and collectively, they generate sometime Peta - bytes of logs each hour or a day. Hence, knowing the seriousness of the issue to compile such huge amount of log data, in our proposed research, at first, we have done extensive analysis of various log samples of existing cyber security attacks and then applied the concept of Ant – Colony optimization technique to optimize the logs as per our experiment requirements. Once logs are optimized, then we identified and segregated all such cyber security attacks & their patterns. Later we applied the Artificial Neural Network based Error Feedback Propagation Algorithm (ANN-EFBPA) to train the model through its successful identification of cyber-attacks and generated errors. Different to different test conditions and datasets, helped the system to get trained effectively and helped proposed novel system to get evolved and matured. Our novel experiment has proven that, the result generated is 98% accurate for correct identification of cyber security threats through Windows event ID log analysis.*

Keyword *— Cyber Security Attacks, Cyber Incident Analysis, System Log, Security Logs, Log Analysis, Event ID, Anomaly, Incident of Compromise, Artificial Intelligence, Genetic Algorithm, Ant Colony Optimization, Feedback Analysis system.*

I. Introduction

With the development of Computer system in the early nineties, there was another phenomenal rapid growth of internet followed across the globe. Initially everyone was not aware of the cyber security threats. Once in a while there was news in the public domain about the cyber security attacks and the magnitude of the damage and its impact across the globe was at small scale. But with the rapid growth of internet and mobile technology, the risk of cyber security has reached to the doors of common man. World has already lost billions of dollars' worth currency, data and information across the globe. All countries including USA, UK, Canada, Japan, Australia, India & other countries are victims of cyber economic crime [1, 2, 3 and 4]. Dark net based lone wolf hackers and cyber syndicates takes contract to destroy an individual or organization. Terrorist and anti-nationals are using cyber space as a platform to launch an attack against a nation [5, 6].

Every connected system in the internet and intranet generates logs. Such as; Windows Event Logs, Operating Systems Logs, Random Memory Access (RAM) access logs, browser logs, IDS logs, IPS Logs, Firewall Logs, Windows Security event logs, Setup logs, Access logs, Audit Logs, Windows Event logs, Server logs and other such logs. Sometime there exist IoT logs, Perimeter device logs, proxy logs, etc. In any cyber-crime incident, it is observed that the investigator may need to analyze different types of logs and

there is no uniformity in the format of all such logs. The logs format is designed by the developers based on the operating principle of respective device; hence their format may be of CSV, JSON, PDF, XML, HTML, etc.

A. Introduction to Ant Colony Optimization (ACO)

The working of Genetic Algorithm (GA) based Ant Colony Optimization algorithm is based on behavior of ants. In the world of ants, when a food source is located, ant moves back from food source location to their colony and during that process it lays down a special chemical called Pheromone, which helps them to trace back the path up to food. Food Trail Pheromone is sensed by all the ants and then they started to move towards the food. Once they reach and collect the food then while moving back, every ant must find out the shortest path from food source to its colony therefore, each ant spray their Pheromone liquid while heading back towards the colony with hope to find out the shortest path. This creates multiple paths. But it is necessary to select only a single path which has minimum distance. Therefore, ant identifies such path whose Food Pheromone trail intensity is high. This is because, various environmental factors like surface temperature or heat, wind, rain, and human activities sometime destroys the Pheromone trail. Considering such activities, one can say that only shortest path will have strong intensity of Food Trail Pheromones. This is because, greater the distance then lower is the chance that any Pheromones track will survive from heat, humidity, rain, wind, dust, etc. atmospheric conditions. This is demonstrated in Figure 1(a), where ants intelligently discard all the longest paths where Food Pheromone trail is weakest and selects the shortest path through its intelligence [7].

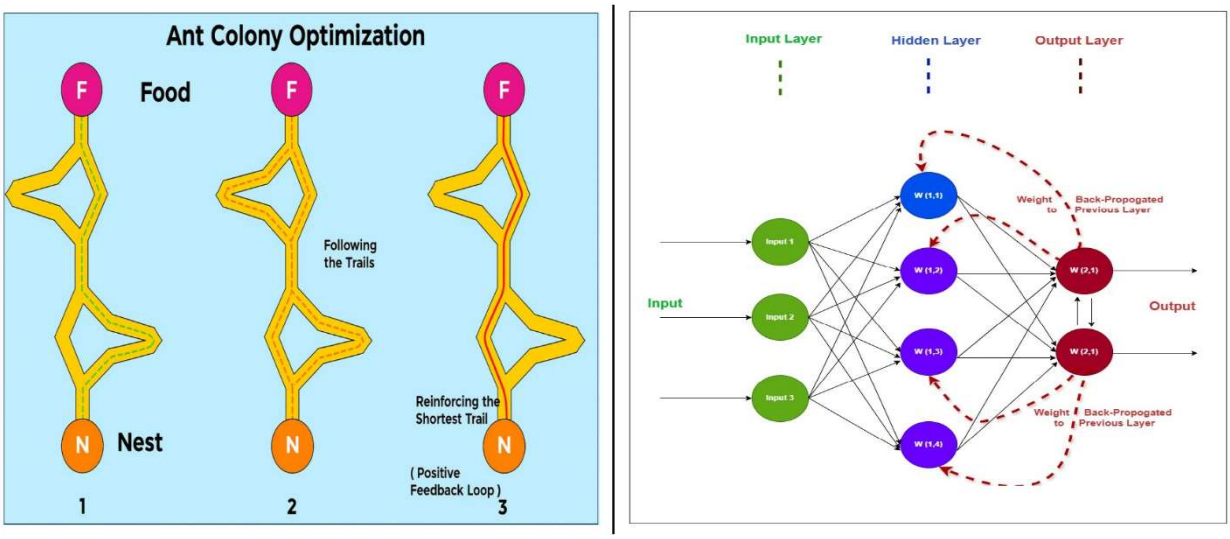

Figure 1(a) Figure 1(b)

Figure 1 (a): GA-ACO Technique for finding the shortest path using Food Trail Pheromones.
1 (b): Formal Depiction of ANN Algorithm based EFBPA.

Based on above logic and behavior of ants, we can assume that there exist number of sub-types of available logs within the Windows Event ID category (Audit log, Application logs, Security logs, etc.) and out of same, we can assume there are certain set of logs which can help us to define or detect the anomalous behavior of the system that may lead to any incident of compromise. For example, the windows Event logs have unique ID. There could be more than 500+ event log IDs in Windows Operation System and each Event ID is unique and has its own purpose and meaning. Almost every Operating system like Linux, etc. have their own set of unique event ID's and logs stored at respective locations. The Windows Event Id's are

normally stored at location C:\Windows\System32\winevt\Logs. Table 1.0 provides a glimpse of few of such Windows Event ID's.

Please note that, there are conditions where an event ID is treated as anomaly or security alert and some time if condition does not meet the required criteria, then the same event ID will be treated as normal event. For example- Event ID 4625 and 4740. If authentic user has forgotten the password and tries multiple times to logon to his system, then the system generates EID – 4625. If any cyber-criminal tries to launch brute-force attack or dictionary attack then in such scenario also, the system generates EID 4625 and it is further followed with generation of EID 4740.

In our proposed novel method, we will be treating all such Event IDs as Pheromones that will lead the trail of actions done by both authentic user and cyber-criminal. Few of the Event IDs from Security domain logs will be considered as most important Pheromones that has ability to provide shortest path to trace or track the cyber-criminal. Such Security Event IDs are marked in our dataset as (Anomaly). The greater the pheromone marks in any given set of logs then there is higher possibility of risk identification and possible cyber-security attack. Hence, the Ant Colony Optimization is best suitable for our problem statement to first optimize the logs in such a way that it sorts all the most important Event ID's that can act as pheromones (Anomaly) from the pool of various general types of Windows Event ID's and then further we proposed to apply the Artificial Neural Network (ANN) based Error Feedback Propagation Algorithm (EFBPA).

Table 1. Glimpse of few Windows Event ID's and their definition.

Windows Event ID	Details	Windows Event ID	Details
529	Logon Failure - Unknown user name or bad password	644	User Account Locked Out
531	Logon Failure - Account currently disabled	676	Authentication Ticket Request Failed
576	Special privileges assigned to new logon	4625	Failed Logon attempts
627	Change Password Attempt	4740	User Account Lockout
629	User Account Disabled	4723	An attempt was made to change an account's password
630	User Account Deleted	1100	The event logging service has shut down

B. Introduction to ANN - Error Back Propagation Algorithm (ANN-EFBPA)

The Artificial Neural Network based ANN-EFBPA algorithm consist of single layer or multiple hidden layers. Its formal depiction is shown in Figure 1(b). Actual ANN-EFBPA depiction may vary with number of inputs, hidden layers it consists, function it is using, etc. The Error Feed Back propagation is a method of feedback that involves the computation of the error gradient at each layer of the network. The error gradient is then used to update the neural network's parameters. Back propagation is widely used in AI and deep neural networks due to its efficiency, accuracy and faster machine learning approach and have only three major steps involved Forward pass, backward pass and Weight update as per given function [8, 9]. The ANN-EFBPA is already used in various applications of network, cyber security and for other complex issues in the field of computer science and engineering [10, 11 and 12]. The concept of ANN is well explained in scripts [13, 14].

II. Materials and Methods

Proposed Novel Model (ACO and ANN-EFFPA)

A. Dataset Details: We optioned for Kaggle listed Cyber-Security Datasets. One of them is dataset of Los Alamos National Laboratory (LANL). The Dataset is of 12-GB size and contains 1,648,275,307 events, total 12,425 users, 17,684 computers and 62,974 processes and is available in chunks and in CSV, JSON, etc. formats. Similarly, we have created our own local lab dataset, which consist of 10 million Microsoft Windows Events IDs collected from individual desktop systems of our research institute. Both the collectively dataset has required contents and are of size that are enough to conduct our proposed experimentation [15, 16, 17, 18 19 and 20]. Both the datasets need to be optimized first and then to be used for further experimentation.

B. Setting up the Overall Generic Process Flow: The overall generic process flow is divided into two main parts, Part A and Part B. This is well depicted in figure 2(a) and figure 2(b) respectively.

Part A – GA (ACO) Part: As per figure 2(a), initially the logs are collected then it is necessary to do the pre-processing. It involves, checking the size of file, format of file, whether file is properly received or is corrupt or unable to process, etc. If any issues are raised, then it should be brought to the notice of System administrator. Now, consider a scenario where all required input log files are received properly, and now system is ready to process it. Since multiple log files are received from multiple sub-systems within the organization, therefore it is necessary to first set the sequence and index of log file processing. The log processing policy selection is a dynamic action and on ground the cyber-crime investigator has to take such decisions based on type of cyber-attack occurred. Such Collected- Sequenced data set of logs is of tremendous size and contains many redundant fields, which are needed to be optimized. Hence, once log sequencing is done, now the next step is to get the logs optimized.

Now, in the first stage of proposed novel system, we have used Ant Colony optimization algorithm to optimize the log data set. The rules of optimization are to first select the user defined set of Windows operating system event ID's that are marked initially as either anomaly or incident of compromise (IOC) or normal event. Such marking are initially done by user based on certain conditions like; for example, when event is triggered in windows event manager in which system authorization level is changed by anyone or any access right violation is registered or any warning message, error message is registered or any warning alert message is registered or any multiple failed-login attempts are registered in operating system or attempt to install any unauthorized software where the machine is not authorized for same or sudden user access level changes are done by any attacker or any third-party unauthorized software is installed and such software further making attempt to access any drive/folder/files/API/ ports/network systems/any machine outside network, etc. Now, based on Optimization rule, once logs are optimized the task of GA-ACO algorithm is completed. Now the output of Part A is provided as input to Part B.

Part B – ANN (EFBPA) Part: As per figure 2(b), the ANN based Error Feedback Propagation Algorithm (ANN-EFBPA) may consist of one-to-many hidden layers (selection of number of hidden layers is as per operational requirements). In Part B, the output of GA-ACO based optimized logs are taken as input to ANN-EFBPA algorithm. Such optimized logs are first processed through the initial layer of weight adjustment. The initial layer weight adjustment is based on the input provided by user which is nothing but the conditions and terms under which any log events is considered as normal log event or anomaly or Incident of

Compromise. Based on same the weight adjustment of respective node is done. Once initial layer node weight adjust is done then the next consecutive log is selected for processing.

In similar fashion one by one all Windows event logs are processed and the respective node weight adjustment is done by ANN-EFBPA algorithm. During this process, the ANN-EFBPA algorithm compares the earlier node weight adjustment with the next consecutive node weight adjustment. Such comparison helps the ANN-EFBPA to learn and quantify the error generated out of the comparison. The generated error is feed-backed to the system itself and this helps to adjust/re-adjust the respective node weight. This process is repeated till the End of the log file is reached. Finally, it evaluates the complete training and store the final result in the database.

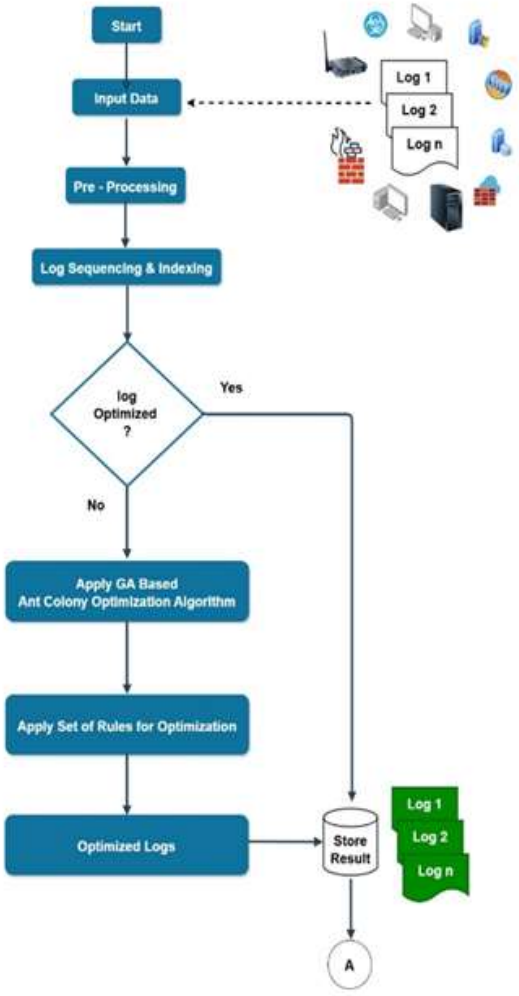

Figure 2(a)

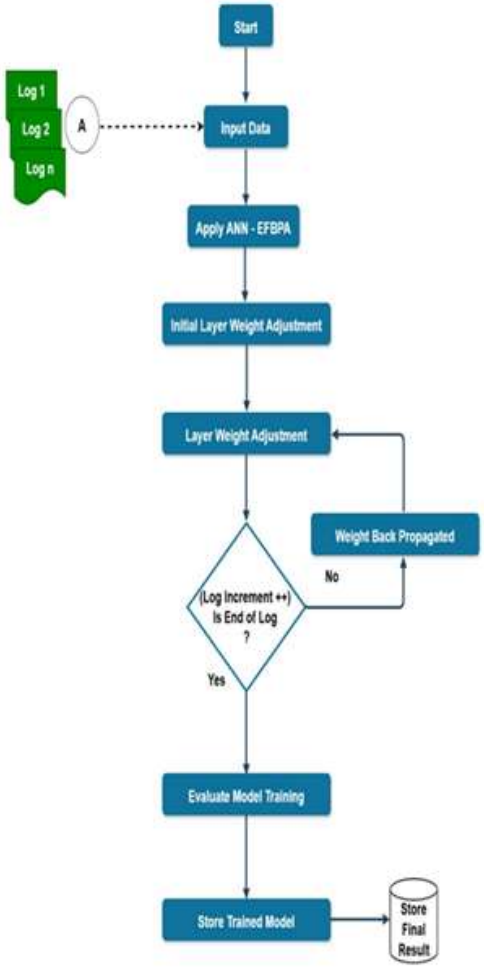

Figure 2(b)

Figure 2(a): Proposed Research Process Flow Chart Part-A (showing application of GA-ACO used to optimize the Logs as per novel-dynamic set of rules for optimization).
2(b): Proposed Research Process Flow Chart Part-B (showing application of ANN–EFBPA used to train the model through Error Feedback method to identify the Cyber-Security Attacks).

III. Results and Discussions

For experimentation, we used Python as it has rich set of libraries that are suitable for AI based work. For experimentation we need high-end computational server system with higher RAM memory. We used both our institutes research lab facilities and Google Co-labs, for our experiment. Figure 3(a) provides the glimpse of very first Graphical User Interface (GUI) of proposed novel system. The GUI consist of 6 tabs and one display window. The six tabs facilitate user to upload the logs, then application of proposed algorithm over logs, showing exact all anomalies and Incident of compromise (IOC) and then another tab is for showing the resultant graph generated out of it as an output. Finally, if user want to exit then he can press the exit tab. If you observe figure 3(a), then we can find out that user has successfully uploaded a log file and its details (File name & total number of logs) are shown in its display window along with file upload successfully message.

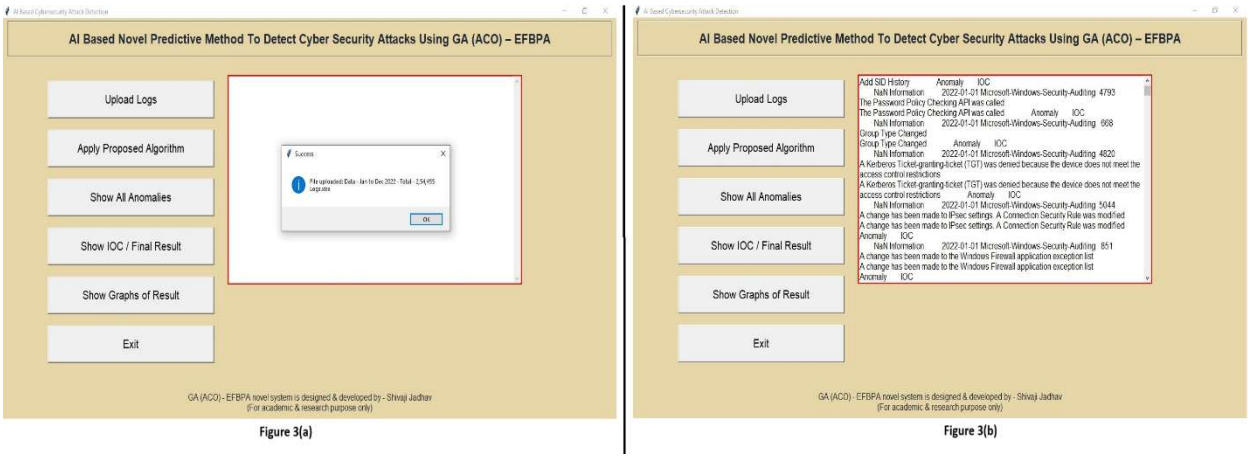

Figure 3 (a): GUI Window showcasing pop-up message highlighting successful uploading of logs.
3 (b): GUI showcasing the list of all the Anomalies and IOC found in given log dataset.

Since the logs are uploaded successfully, now apply the proposed algorithm through pressing the second tab "Apply Proposed Algorithm". Since the algorithm consist of application of GA- ACO and ANN-EFBPA therefore it will take some time and memory. The amount of time required for compilation of logs by the proposed algorithm is directly proportional to the underlying systems computational power and memory installed. We have used a server class machine of our local research lab and also used the online Google Co-lab facility provided by Google Inc. Since our dataset is quite large therefore, the system needs some time to compile it. Post successful compilation of all logs as per proposed algorithm, the system highlights success message. Further, from the perspective of system/network administrator it is now necessary for him to check all such logs those are highlighting the anomalous behavior. For the same, the administrator can click on the tab "Show All Anomalies". The proposed novel system then generates the result in GUI based window showcasing all the anomalies found in the given log dataset. This is shown in figure 3(b).

Finding anomaly in given log is a matter of concern. But, kindly note that, not all listed anomalies are considered as a threat or potential future Incident of Compromise. This is because sometime anomalies can be suspicious at first but only after close observation, one can conclude that the respective anomaly is a real cyber threat or not. Further, as a system/network administrator it is also needed to check all the Incident of Compromise (IOC's) the system has found in given log dataset. For the same the user should use the tab "Show IOC/Final Result". This will cause the system to display in GUI window the list of all the Incident of Compromise (IOC) it found in given log dataset. This is shown in figure 3(b).

Please note that, at initial stage all IOCs are first considered as anomaly. But when any anomaly become a practical threat to the system then it is treated as IOC. The condition for an anomaly to become IOC varies

sometimes. We have to set each time a threshold under which an anomaly can be considered as IOC. This depends on the global threat vector parameters that evolves constantly with respect to updates and upgrades in technology platforms including hardware, software and firm-wares. This is because cyber criminals always take advantages of this changing technology platforms and the loopholes exist in them. The moment when one loopholes is fixed by the manufacturer / developer, the next moment cybercriminal find outs and exploits another loophole.

Further, by using the tab "Show Graphs of Result" will plot the graphs that gives the overall statistics of total number of logs marked as anomaly and total number of logs marked as Incident of compromise in month-wise manner. This again helped us to understand that not all anomalies will be treated as IOC. This is well depicted in figure 3(c) where month-wise statistics of various anomalies and IOCs for two consecutive years 2022 and 2023 respectively are provided. Similarly, figure 3(d) provides details of month-wise statistics of various anomalies and IOCs for consecutive year 2024 respectively.

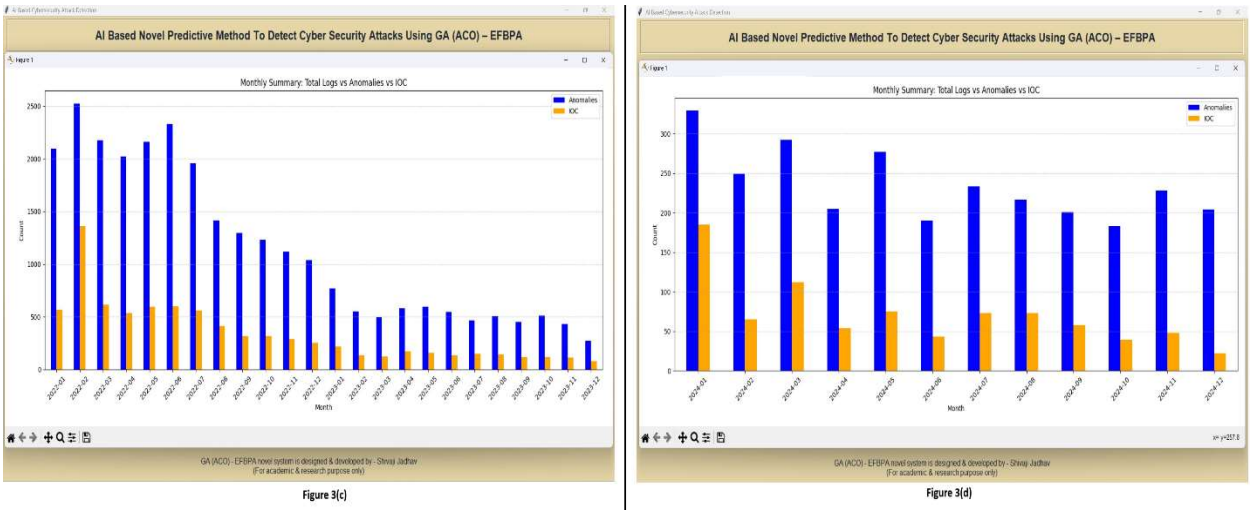

Figure 3(c). Graphs with Anomalies and IOCs found in given log dataset (for year 2022 & 2023).
3(d). Graphs with Anomalies and IOCs found in given log dataset (for year 2024).

We have done numerous such experimentation with various standard dataset like Kaggle listed Los Alamos National Laboratory (LANL), our own local research lab generated log datasets and other such available datasets. Based on all figures and tables, it is observed that, the proposed novel algorithm and implemented system correctly identifies and marks each such anomaly and IOC found in given log datasets. We can call this marking as Pheromone based marking of GA-ACO, and the correctness of such markings will help to adjust the weight of the nodes of ANN-EFBPA algorithm and this will be a learning cycle for the ANN-EFBPA algorithm.

Sometime, the algorithm may identify and mark any simple log event as anomaly or IOC incorrectly, which is false identification of a simple log event. Such actions can be measured as False Acceptance Rate (FAR) and it is well represented in Figure 4(a). It consists of 10 iterations for log compilation and with each Iteration Cycle (IC) the system gets trained and matured. Here the false acceptance rate at first iteration cycle is approx. 4.0% and with each successive iteration cycle the false acceptance rates falls up to 0.001% till the tenth iteration cycle. This proves that the algorithm is getting enough training and correct set of iteration cycles with appropriate examples. Similarly, the accuracy of the proposed system of GA (ACO) & ANN-EFBPA gets slowly improved in stepwise manner. From 1st iteration cycle (IC) till up to the 10th iteration cycle the accuracy of the system has reached approximately up to 98%. This is well depicted in the graph shown in figure 4(b), along with the comparison of proposed novel model with formal GA(ACO) model without application of ANN-EFBPA algorithm.

Sr. No.	Number of Logs	False Acceptance Rate (FAR)
1	10,000	4.0%
2	1,000,00	4.0%
3	5,000,00	3.0%
4	10,000,00	3.0%
5	40,000,00	1.3%
6	80,000,00	0.80%
7	100,000,00	0.11%
8	10,00,00,000	0.07%
9	1,64,82,75,307	0.02%
10	3,50,40,55,208	0.001%

Figure 4(a)

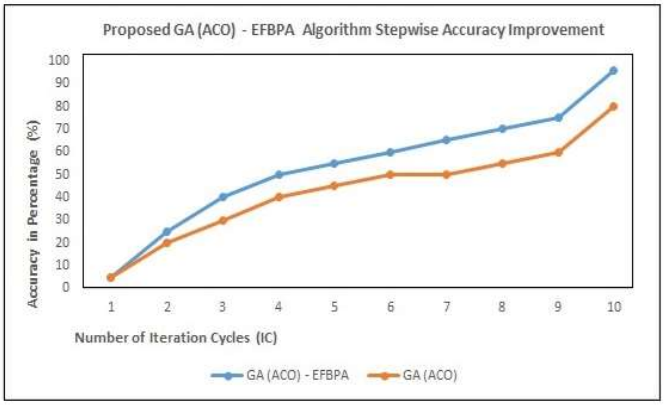

Figure 4(b)

Figure 4(a). Number of Logs and Proposed Systems False Acceptance Rate (FAR)
4(b). Graph showing step-wise Improvement in accuracy (98%) of proposed novel system

Please note that, the primary focus of our research was to compile system logs in our proposed novel GA (ACO) & ANN-EFBPA system. In this paper, we have presented our experimentation using the Windows Event ID's logs generated by Windows Operating System. But we have also tried this experiment with few other machine-based logs that have different formats and parameters. This include but not limited to IDS, IPS, Firewall, and others. But due to research paper number of pages' limitations, it is not possible for us to quote all the results under one research paper. The provided tables and figures sufficiently help us to completely understand the developed novel system and its functionalities.

IV. Conclusion

Cyber security is a challenge to internet world. Researchers across the globe are trying to find out various solutions to counter cyber-attacks. Logs provides crucial help to identify the cyber-attacks. Windows Event ID logs can play crucial role to identify and mitigate the cyber risks. Therefore, by using the Windows Event ID logs, in this paper, we are proposing a novel method for identification and prediction of cyber security attack. The model first uses the Genetic Algorithm based Ant Colony Optimization technique and later ANN based Error Feed Forward Propagation Algorithm to solve the problem statement. The Ant Colony Optimization technique tries to successfully optimize the Windows Event ID logs in such a way that it discards all such logs which are not directly related to security incident events. Such selection of interested logs is dependent on various conditions or parameters, such as the Windows Event ID's that are marked as suspicious in previous cyber-attack scenarios helps the ant colony optimization model to tune up its optimization parameters. Therefore, the Ant colony optimization algorithm first narrow down the size of given log data by its unique optimization method.

Optimized logs are then provided as input to ANN-EFBPA algorithm. The Error Feedback propagation algorithm adjust its backward directed node weight based on the identification or marking of anomalies and IOCs by ant colony algorithms. This helps the ANN-EFBPA to calculates and quantify the error magnitude and uses it as like a learning curve. This helps the ANN-EFBPA algorithm to improve itself through each learning cycle. The proposed novel system was designed and developed in Python programming language and tested against various online available standard datasets and our own local research lab generated log datasets. Through our experimentation, it is proved that the proposed novel model is highly accurate and provides 98% accuracy for correct identification of anomalies and IOC in given Windows Event ID log datasets.

References

[1] *Vergara Cobos, Estefania and Cakir, Selcen. 2024. A Review of the Economic Costs of Cyber Incidents. Washington, DC: World Bank.*

[2] *Mihail Antonescu, Ramona Birău, Financial and Non-financial Implications of Cybercrimes in Emerging Countries, Procedia Economics and Finance, Volume 32,2015,Pages 618-621, ISSN 2212-5671, https://doi.org/10.1016/S2212-5671(15)01440-9.*

[3] *Sunny A, A study on financial cyber-crimes, trends, patterns, and its effects in the economy. Allied Academics Addiction & Criminology - Addiction & Criminology (2024) Volume 7, Issue 1. DOI: 10.35841/aara-7.1.186.*

[4] *National Crime Record Bureau (NCRB) Annual Report on Crimes in India 2022. Volume I. Ministry of Home Affairs, Government of India.*

[5] *Amr Adel and Mohammad Norouzifard, Weaponization of the Growing Cybercrimes inside the Dark Net: The Question of Detection and Application, MDPI Journal Big Data Cogn. Comput. 2024, Volume 8, Issue 8, pp91, https://doi.org/10.3390/bdcc8080091.*

[6] *Almukaynizi, M., Grimm, A., Nunes, E., Shakarian, J., & Shakarian, P. (2017, October). Predicting cyber threats through hacker social networks in darkweb and deepweb forums. In Proceedings of the 2017 International Conference of The Computational Social Science Society of the Americas (pp. 1-7).*

[7] *Anand Nayyar et. all., Ant Colony Optimization – Computational Swarm Intelligence technique, 3 rd Interational Conference on Computing for Sustiniable Global Development , 16-18 March 2016, pp 392-398.*

[8] *Xinghuo Yu, M. O. Efe and O. Kaynak, A general backpropagation algorithm for feedforward neural networks learning, in IEEE Transactions on Neural Networks, vol. 13, no. 1, pp. 251-254, Jan. 2002, doi: 10.1109/72.977323.*

[9] *H. Bersini and V. Gorrini, A simplification of the backpropagationthrough time algorithm for optimal neurocontroller, IEEE Trans.Neural Networks, vol. 8, pp. 437–441, Mar. 1997*

[10] *M. Fujimoto, W. Matsuda and T. Mitsunaga, "Detecting attacks leveraging vulnerabilities fixed in MS17-010 from Event Log," 2019 IEEE Conference on Application, Information and Network Security (AINS), Pulau Pinang, Malaysia, 2019, pp. 42-47, doi: 10.1109/AINS47559.2019.8968703.*

[11] *J. Dwyer and T. M. Truta, "Finding anomalies in windows event logs using standard deviation," 9th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing, Austin, TX, USA, 2013, pp. 563-570, doi: 10.4108/icst.collaboratecom.2013.254136.*

[12] *A. Zrelli, C. Nakkach and T. Ezzedine, "Cyber-Security for IoT Applications based on ANN Algorithm," 2022 International Symposium on Networks, Computers and Communications (ISNCC), Shenzhen, China, 2022, pp. 1-5, doi: 10.1109/ISNCC55209.2022.9851715.*

[13] *Introduction to Artificial Neural Systems book by author Jacek Zurada, West Publishing Company USA, ISBN 0-314--93391-3.*

[14] *Artificial Neural Networks: An Introduction to ANN Theory and Practice: 931 (Lecture Notes in Computer Science) by P.J. Braspenning, F. Thuijsman, A.J.M.M. Weijters (Eds), ISBN-10: 9783540594888. Springer-Verlag Berlin and Heidelberg GmbH & Co. K publisher. June 1995.*

[15] *Kaggle Datas et https://www.kaggle.com/discussions/general/335189*

[16] *LANL Dataset availaible on https://csr.lanl.gov/data/cyber1/*

[17] *LANL Dataset availaible on https://csr.lanl.gov/data/2017/*

[18] *LANL Dataset availaible on; https://csr.lanl.gov/data-fence/1744233138/3aTHpiA7DmWNPD1elmhbhGjLzV0=/unified-host-network-dataset-2017/wls.html*

[19] *LANL Dataset availaible on; https://csr.lanl.gov/data-fence/1744233138/3aTHpiA7DmWNPD1elmhbhGjLzV0=/unified-host-network-dataset-2017/netflow.html*

[20] *Local Lab Created Dataset Uploaded on SGGS IET Repository space availaible at https://www.sggs.ac.in/*