International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Botnet Attack Detection in the Network with SBLSTM Classification

J. Aruna[1*], S. Prayla Shyry[2]

Submitted: 19/01/2024 **Revised**: 28/02/2024 **Accepted**: 05/03/2024

Abstract: With the emerging growth of Information Technology, criminals are utilizing cyberspace to perform various software-oriented Cybercrimes that are very difficult to analyze. Cyber infrastructure is highly malicious and directly impacts the current performance of the system. Cybercrime occurs in the form of intuitions and other trips. Various devices are connected over the internet of things (IoT) platform to make the evaluation process more effective. The presented system is focused on creating an efficient framework adaptable to various kinds of environments, the creator of Cyber-attacks. In the presented study, staked boosted long short-term memory (LSTM) encoder algorithm is utilized to detect BOTENET Attacks. The proposed work aims to provide a solution with optimized detection Framework that discovers The Attack coming over the network. In order to attend high accuracy, the proposed approach needs to be an effective validate in terms of performance accuracy, and proactive security techniques are combined using machine learning and deep learning algorithm.

Keywords: *Information technology (IT), cybercrimes, Botnet attacks, SBLSTM, deep learning.*

1. Introduction

Wireless sensor networks are almost applied in every field towards developing 5G communication and effective internet of things environment. Wireless sensor networks need to be improved shortly to handle all the issues coming over the network in an automated way without any manual interruptions. The presented system is focused on taking a vital problem called the lifetime benefits of sensor notes through the cluster head selection technique. A machine learning algorithm for employed over here, such as a support vector machine as an existing system and Gaussian regression process and proposed system. By applying the algorithms to the cluster head selection mechanism during the face of the energy degradation process, the presented system performs civil during the critical time. It handles the lifespan problem of the wireless sensor network. The system needs to be improved by combining multiple machine learning hybrid algorithms to handle the situation more efficiently and employ the error-free communication model.

1.1 Botnet Threats:

BOTNET is one of the severe conventional forms of the virus are worms that affect the core content of the distributed anal of services connected systems in the form of spam data cyber warfare exploiting the resources, fraudulent emails, and botnet attacks. Various existing studies have manipulated the BOTNET detection frame using machine learning.

1.2 Distributed denial of service (DDoS) :

BOTNET is connected with distributor Daniel of services attack frameworks collect the traffic information and produce an enormous number of sources flood that create the massive number of participants. The destructive power of BotNet attacks directly impacts the application systems and websites. The distributed services connect with the standard system and further impact the entire network by affecting the single system.

1.3 Spam:

Spam is an unwanted email message containing malicious data in the email links are attachments. These emails also contain advertisement images that often contain back and direct links that direct the complete control into other platforms containing malware attacks. BOTNET attacks are oriented from BOTNET Masters that contain numerals address information for mailing the spam messages to the bulk User ID.

1.4 Stealing information:

BOTNET Masters are control mechanisms operated by pre-programmed methods that command the boat net to Grab the user data from a compromised host. Few systems have less authentication hold to have the data storage. The raw data directly impacts the system key login data and further Steal private information. Most of the credential data is stored as user-accessible messages. Messages easily collect username passwords from social media networks. Further, the BOTS can extract the user information before the web browser encrypts it.

[1] *Department of Computer science, Sathyabama University, Chennai – 600119 – Tamilnadu -India, Corresponding Author Email: arunajaisankar@gmail.com*

2 Department of computer Science, Sathyabama University, Chennai - 600119 –Tamilnadu- India, Email: praylashyry@gmail.com

1.5 Exploiting resources:

In wireless network topology, few compromised systems allow to execution of illegal actions. Fake IDs were used on Twitter and Facebook to create fake votes and raise the number of machines to allocate proper decisions.

The problem statement that evaluation of BOTNET detection using Michel learning and deep learning is present here. Machine learning and deep learning algorithm individually impact Malware attack detection. Various existing articles discuss on Malware attack detection using impacted botnets are discussed here.

The author proposes an algorithm to detect IoT attacks that help create variances of challenging patterns. A hacking mechanism enables small changes in the existing taxes to create new boats that spread the malware. Developers utilize AI technology to safeguard the system from external attacks. Deep learning and machine learning algorithms are created to make systems efficiently handle the various boards coming over the network. The existing Framework is considered for attack detection using stacked boosted long short-term memory systems tuning the features of the raw data.

2. Related Works:

Various existing frameworks are analyzed through machine learning models and various neural principles.

N. Agarwal et al. the author provides a device to make a less security system that exploits the classes between the significant section issues in computer infrastructure. Infrastructure leakages are one of the main problems. So many computers are compromised to provide confidential data in the system. The predictive analysis model provides major trends to detect distributed denial of services (DDoS) attacks.

X. A. EI. Houda et al. the author presented a system using explain able AI-based intrusion detection model. Critical decisions are formulated by machine learning and deep learning models using deep neural networks. Multiple explainable AI mechanisms are combined with distributed neural network architecture to offer more deep attractive decision models in-depth system model shows the benefit of explainable AI.

Ahmed et al. The author utilizes the machine learning-based algorithm to enable attack detection. In the presented approach, a classification algorithm, namely the Logistic regression Gaussian decision tree algorithm random forest algorithm, K-nearest neighbor algorithm, and extremely gradient-boosted algorithm for Malware detection in IoT network traffic, is being discussed. A real heterogeneous dataset is considered for selecting impacted features in the raw data.

A..Hekmali et al. presented a real-time application of AI in cyber security to provide an isolated network for detecting distributed denial of service attacks (DDoS) are developed. Malicious and Furious attacks are developed in the traffic network being analyzed.

H. Somya et al. the author developed and exploded the benefit of a supervised machine learning classification model by adaptive tuning the hyper parameters to regulate the regularize the boat access in the network to innovative ideas are utilized in the network security to make the Optimization process better than the existing state of the art approaches.

R. F. Hayat et al. the author propose the distributed Anil of services enabled IoT protection Framework using blockchain technique. The proposed idea is evaluated using a hypervisor caliber tool to measure with the three benchmark applications. Despite detecting botnet attacks, various data sets are comparatively valuators with the block chain algorithm.

Abjith d.et al., the author presented an intrusion analysis model implemented using a machine learning algorithm in which parallel and distributed approaches utilize d the presented technique is enabled to design ideas distribution through an online streaming method.

XiaoKang Zhou et al. the author proposed a system to detect inconsistencies like vibration extended short-term memory model for an intelligent anomaly detection model for a massive dataset. Furthermore, in handling unstructured data, the difficulties may remain the same.

Tran Hoang Hai et al. The author proposes day novel architecture for an intrusion detection system in a network Framework. The proposed architecture is displayed using Spark stream real-time analysis of RAW processing.

3. Methodology:

The proposed methodology focused on utilizing the CIC IDS data set. Canadian Institute of Cyber Security(CIC)-intrusion detection system(IDS) data set consists of unstructured data with the recording of various innovative City enabled device nature. The pattern of the device needs to be measured concerning the analysis provided with the supervised network. The proposed approach considers the SBLSTM classification technique in which the stack booster LSTM encoder model is developed to detect BOTNET attacks. The proposed system focuses on developing a unique model differentiated from the existing Framework on feature selection with significant value in improving the accuracy and handling of training data. The Attacks are detected using classification techniques which can be measured in terms of performance using accuracy.

4. Dataset Description:

Data set a description CICIDS2017 data set concepts of both meaning behavior and details of new Malware attack present such as a brute force FTP attack, brute force SSH attack, Distributed Denied of services attack (DDoS), Heart bleeds attack with an attack infiltration—DoS attack, etc. The dataset is formulated concerning the IP address of a source system and destination system connected topology protocol details number of rooters connected over the network specific configuration of network operating system and contains 18 network flow features of the captured network traffic.

4.1 Preprocessing:

The preprocessing of the dataset is attained by removing the unwanted error data present in it, unwanted spaces, and junk values in it. The nearest value computation is formulated with the help of a matrix and multi-dimensional array technique. The length of the raw data needs to zero score transformation function. The normalization approaches considered are zeros normalization, expressed by the formula below.

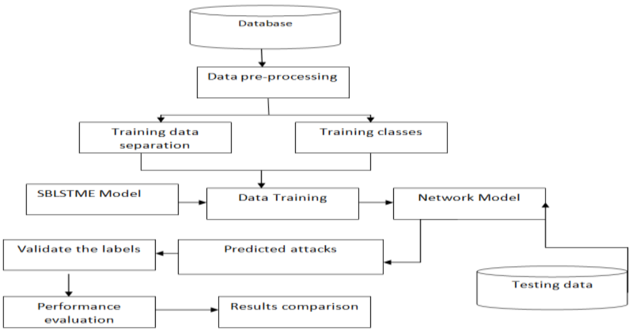

Fig 1. Block diagram of the proposed method:

$$X_i' = \frac{X_i - mean\ (X)}{SD(X)} \qquad (1)$$

Eqn(1) shows the preprocessing steps with zero score normalization. The data set is split into training data at 70%, with an equivalent random distribution of test data at 30% considered.

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N} \left(X_i - mean(X)\right)^2} \qquad (2)$$

Eqn. (2) shows the standard deviation of the input X. the proposed approach needs to be created and trained using the training data set. The boosted classifier model random user sampling technique algorithm is implemented to specify a particular parameter. The training model needs to be fed to sleep mode to conjugate the sweep parameters. This model produces a learner with the best setting.

4.2 Dimensionality reduction:

The raw dataset is switched to diamonds, not a reduction process of the feature selection process. Here irrelevant features are removed from the data set. The presented approach considers particles from the Optimization technique to Naturally stimulate the data into required optimized values and unwanted data. The optimization process regulates the unstructured data by formulating the data to a unique pattern that contains an evolutionary process considering the best value of position and population. The particles from the Optimization process divide the data into optimum levels providing vibration data set using the formula below.

$$V_i(t+1) - V_i(t) = c_1 r_1\left(P_i - X_i(t)\right) + c_2 r_2\left(P_g - X_i(t)\right) \qquad (3)$$

$$x_i(t+1) = X_i(t) + V_i(t+1) \qquad (4)$$

Where $X_i = (X_i^1, X_i^2 \ldots\ldots, X_i^D)$ act as the input

The particles from the generator by the given data is a named

new data. The new data is the best-suited data ever for making the classification process. When the data is optimized, it is fetched into the classification model using the SBLSTM algorithm.

The particles from Optimization sweep better into Optimization to converge all the related data in the given pattern.

$$V_i(t+1) = \alpha V_i(t) + \frac{\alpha}{2} V_i(t-1) + \frac{\alpha(1-\alpha)}{6} V_i(t-2) + \frac{\alpha(1-\alpha)(2-\alpha)}{24} V_i(t-3) + c_1 r_1\left(P_i - X_i(t)\right) + c_2 r_2\left(P_g - X_i(t)\right) \qquad (5)$$

$$D^{\alpha}[v_t] = \frac{1}{T^{\alpha}} \sum_{k=0}^{r} \frac{(-1^k)\Gamma(\alpha+1)v(t-kT)}{\Gamma(k+1)\Gamma(\alpha-k+1)} \qquad (6)$$

4.3 Classification using the SBLSTM algorithm:

Classification is analyzing the data from the given basic information after the preprocessing technique feature selection need to be done. Classification decors the data pattern from the training pattern and analyze efficient pattern recognition to evaluate the data class. The process of encoding at training consists of two parts encoder and Decoder. Encoder is used to analyze the input data present in the hidden Network and Decoder is used to reconstruct the data after the hidden presentation.

5. Rusboost Algorithm

RUSboost is an algorithm to handle them in a balanced data format in a large data set. The boosting technique is nothing but a random user sampling technique that removes the majority of samples to remove the unwanted class of data while sampling. The adaptive boosting technique formulated here boosts the process into different classes. The modified techniques are employed here to analyze the data randomly. It is very similar to the smooth boost process, where the input data is formulated to handle the pattern understood by the classification algorithm. The proposed model considers the input data and makes feature extraction effective. The model is trained using a training dataset in the initial states. The training dataset is performance measured after the classification process. The better-performing model is considered for further analysis of test data.

$$h_n = f(W_1 X_n + b_1) \qquad (7)$$

$$\widehat{X_n} = g(W_2 h_n + b_2) \qquad (8)$$

$$\emptyset(\Theta) = \underset{\theta,\theta^l}{arg\ min} \frac{1}{n}\sum_{i=1}^{n} L(x^i, \hat{x}^i) \qquad (9)$$

where

$$L(X, \hat{X}) = \left\|X - \hat{X}\right\|^2$$

$$Y_i - f(\textstyle\sum_{i=1}^{n} w_i . x_i + b) \qquad (10)$$

The input data x= [x1, x2, x3… . Xn]

Each data is represented by the weights w=w1, w2, w3… .Wn

The sigmoid function is formulated to handle the intermediate data.

Eq(10) shows the hidden layer equation.

Eq(9) shows the weight matrix, and xi shows the feature vector. The soft-max layer is used for normalization; thus, the equation below provides tuning up the prediction process using SBLSTM Technique.

$$f(x) = \frac{1}{1 + e^{-(x_i w_i)}} \qquad (11)$$

$$P(y = j|x) = \frac{e^{x^T w_j}}{\sum_{k=1}^{K} e^{x^T w_i}} \qquad (12)$$

The equation holds the performance of the output layer irrespective of selection using neuron modeling.

The selected features of the CIC-IDS dataset are fed into the input layer. Reconfigurable hidden layer and classification output layer. These data are applied with the sigmoid activation function, and the results are integrated with the functionality of the sigmoid layer and provide a predicted class accordingly.

Table 1 Performance metrics using SBLSTME

Class	Total number of data	Correctly Predicted	Wrongly predicted
NORMAL	3883	3864	19
BOT	666	600	66
DDOS	9661	9151	510
PORTSCAN	8000	7985	15

Table 2 Data Classification Report of Testing data:

Metrics	Training	Testing
Accuracy	97.1	97.25
Precision	97.2	100
Recall	90.2	90.06
F measure	93.57	94.78

Table 3 Accuracy comparison of proposed and existing methods

Algorithm	Accuracy
QBSA-IELM[15]	94.55
RF[16]	90.3
REP Tree [17]	93.56
SBLSTME	97.25

Table 2 shows the classified report for each class in the dataset taken for testing data. From this table, it is observed that for the testing the prediction of BOTattack is wrongly classified for 66 samples from the total data and standard samples.

5.1 Performance measure:

The percentage system investigates the performance of and analysis of detecting the botnet attacks concerning time. Usually the performance of the system is measured using actual positive valid negative false positive false negative rate measured by the confusion Matrix. By declaring the amount of actual value expected from the presented system and the predicted value of time from the given system need to be validation. The detection performance is measured using accuracy, Precision recall and score.

5.1.1 Accuracy :

Accuracy is formulated with a number of correctly predicted values concerning input values. The expression is given below.

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (13)$$

5.1.2 Precision:

Precision is defined as obtained predicted result through OGSVM's model concerning true positive and false positive values.

$$precision = \frac{TP}{TP+FP} \qquad (14)$$

5.1.3 Recall :

A recall is the parameter utilized for calculating the sensitivity of optimistic prediction with respect to all true positive and false negative rate.

$$recall = \frac{TP}{TP+FN} \qquad (15)$$

F measure is defined by resilient balance of precision and recall values estimated.

$$Fmeasure\ = 2 * \left(\frac{1}{\frac{1}{presicion_p}+\frac{1}{Recall\_R}}\right) \qquad (16)$$

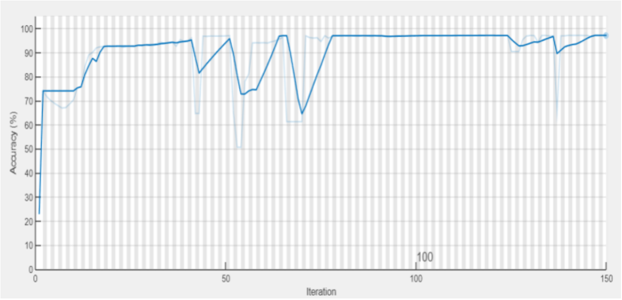

Fig 2 accuracy graph

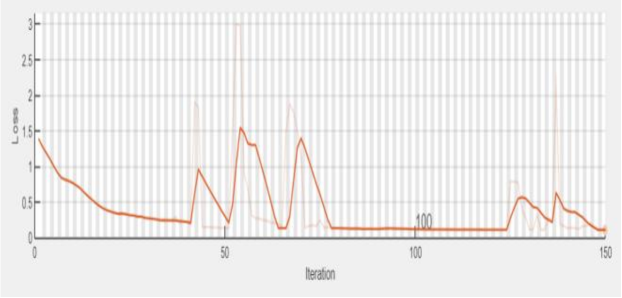

Fig 3 loss graph

Fig 6 The confusion matrices and obtained detection rates are depicted in Fig. 6. The confusion matrices show the number of data categorized as either malicious or benign. Figures 2 and 3 depict the model's accuracy and loss during training with the proposed classifier. We note that the highest actual positive rate (TPR) is achieved.

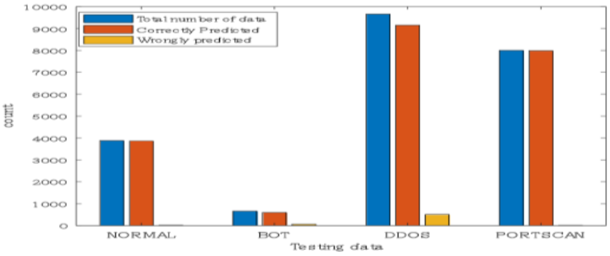

Fig. 4 Count of testing data

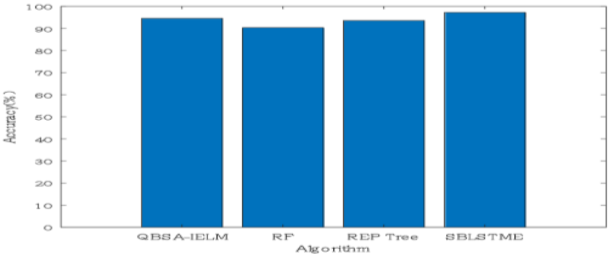

Fig. 5 Accuracy comparison with existing algorithms

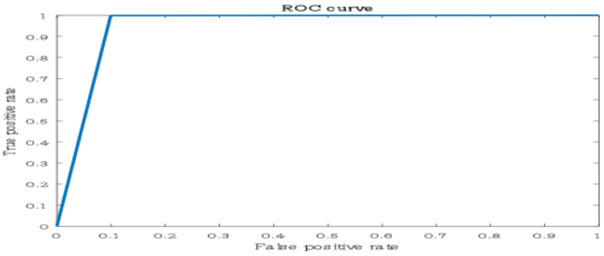

Fig. 6 ROC curve of the proposed model

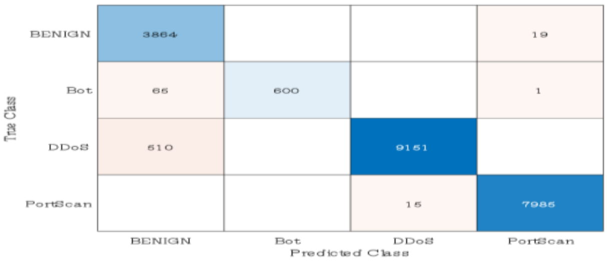

Fig 7. Confusion matrix

The Receiver Operating Characteristic (ROC) curve looks like this: The FPR is plotted along the ROC's x-axis, and the TPR is plotted along its y-axis. A ROC curve is created by plotting the TPR as a function of the FPR for several different cut-off values. A particular FPR and TPR combination corresponds to a distinct judgment criterion at each position along the ROC curve. As the criterion for categorization is changed, a new point on the ROC with a distinct False Alert Rate (FAR) and True Positive Rate (TPR) is chosen as the starting point. The point in the upper left-hand quadrant of the graph represents where the ROC curve of a test would be if it had flawless classification (no crossover between the two distributions) (100 percent specificity, 100 percent sensitivity). a representation of the ROC Curve can be found in Figure 7.

5.2 Advantages of the proposed algorithm:

- Using the signature database, IDS ensures quick and effective detection of known anomalies with a low risk of raising false alarms.
- It analyzes different types of attacks, identifies malicious content patterns, and helps administrators tune, organize and implement adequate controls.
- It helps the company maintain regulatory compliance and meet security regulations, providing greater visibility across the entire network.

5.3 Applications:

- The proposed IDS model monitors network traffic and can detect suspicious activities.
- An IDS is a hardware device or software application that uses known intrusion signatures to detect and analyze both inbound and outbound networks.
- Monitor and evaluate threats, catch intruders and take action in real-time.
- Maintain users' privacy as IPS only records the network activity when it finds an activity that matches the known malicious activities.
- Detect and foil OS fingerprinting attempts that hackers use to find out the OS of the target system to launch specific exploits.

6. Conclusion:

Using the classification technique known as SBLSTME, we discuss how to boost the effectiveness of an attack detection system in this paper. The IDS has dealt with some problems. These IDS cannot quickly identify the new type of attack, just like other supervised learning techniques. Users' behaviors change over time, but if a new attack is found in the testing data, it is detected as normal data. The proposed classifier provides the highest detection accuracy in resolving these issues. The feature selection algorithm is used to generate the results. The paper has looked at various classifications for intrusion detection and determined how accurate they are. With the CICIDS2017 dataset, SBLSTME produces higher accuracy—roughly 97.25 percent—according to the necessary parameter.

References

[1] Guerra-Manzanares, H. Bahsi and S. Nõmm, “Hybrid Feature Selection Models for Machine Learning Based Botnet Detection in IoT Networks,” International Conference on Cyberworlds (CW), Kyoto, Japan, pp. 324-327, 2019.

[2] U. Rehman, R. A. Naqvi, A. Rehman, A. Paul, M. T. Sadiq, and D. Hussain, “A trustworthy SIoT aware mechanism as an enabler for citizen services in smart cities,” Electronics, vol. 9, no. 6, p. 918, 2020.

[3] S. M. Istiaque, M. T. Tahmid, A. I. Khan, Z. A. Hassan and S. Waheed, "Artificial Intelligence Based Cybersecurity: Two-Step Suitability Test," 2021 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI), 2021, pp. 1-6, doi: 10.1109/SOLI54607.2021.9672437.

[4] K. N. Karaca and A. Çetin, "Botnet Attack Detection Using Convolutional Neural Networks in the IoT Environment," 2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), 2021, pp. 1-6, doi: 10.1109/INISTA52262. 2021.9548445.

[5] N. Agarwal, A. Q. Md, V. T, P. K and A. K. Sivaraman, "A Robust Pipeline Approach for DDoS Classification using Machine Learning," 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT), 2022, pp. 1621-1627, doi: 10.1109/ICICICT54557.2022.9917596.

[6] Z. A. El Houda, B. Brik and S. -M. Senouci, "A Novel IoT-Based Explainable Deep Learning Framework for Intrusion Detection Systems," in IEEE Internet of Things Magazine, vol. 5, no. 2, pp. 20-23, June 2022, doi: 10.1109/IOTM.005.2200028.

A. Ahmed and C. Tjortjis, "Machine Learning based IoT-BotNet Attack Detection Using Real-time Heterogeneous Data," 2022 International Conference on Electrical, Computer and Energy Technologies (ICECET), 2022, pp. 1-6, doi: 10.1109/ICECET55527. 2022.9872817.

[7] Hekmati, E. Grippo and B. Krishnamachari, "Neural Networks for DDoS Attack Detection using an Enhanced Urban IoT Dataset," 2022 International Conference on Computer Communications and Networks (ICCCN), 2022, pp. 1-8, doi: 10.1109/ICCCN54977. 2022.9868942.

[8] H. Somaya and M. Tomader, "Tuning the hyperparameters for supervised machine learning classification, to optimize detection of IoT Botnet," 2022 11th International Symposium on Signal, Image, Video and Communications (ISIVC), 2022, pp. 1-6, doi: 10.1109/ ISIVC54825.2022.9800742.

[9] R. F. Hayat, S. Aurangzeb, M. Aleem, G. Srivastava and J. C. -W. Lin, "ML-DDoS: A Blockchain-Based Multilevel DDoS Mitigation Mechanism for IoT Environments," in IEEE Transactions on Engineering Management, doi: 10.1109/TEM.2022.3170519.

[10] Abhijit D. Jadhav, VidyullathaPellakuri,” Intrusion Detection System Using Machine Learning Techniques for Increasing Accuracy and Distributed & Parallel Approach For Increasing Efficiency”, 978-1-7281-4042-1/19/$31.00 ©2019 IEEE.

[11] Xiaokang Zhou , Member, IEEE, Yiyong Hu , Member, IEEE, Wei Liang , Member, IEEE, Jianhua Ma, Member, IEEE, and Qun Jin , Senior Member, IEEE,” Variational LSTM Enhanced Anomaly Detection for Industrial Big Data”, IEEE transactions on industrial informatics, vol. 17, no. 5, may 2021.

[12] Hai, T. H., &Khiem, N. T. (2020). Architecture for IDS Log Processing using Spark Streaming.2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE). doi:10.1109/icecce49384.2020.9179188.

[13] J. Tillett, R. Rao, and F. Sahin, "Cluster-head identification inad hoc sensor networks using particle swarm optimization," in Proceedings of the ICPWC 2018 - IEEE International Conferenceon Personal Wireless Communications, pp. 201-205, New Delhi,India.

[14] Y. Dong, W. Hu, J. Zhang, M. Chen, W. Liao, and Z. Chen, ''Quantum beetle swarm algorithm optimized extreme learning machine for intrusion detection,'' Quantum Inf. Process., vol. 21, no. 1, pp. 1–26, Jan. 2022.

[15] R. Vinayakumar, M. Alazab, K. Soman, P. Poornachandran, A. AlNemrat, and S. Venkatraman, "Deep learning approach for intelligent intrusion detection system," IEEE Access, vol. 7, 2019.

[16] P. Dahiya and D. K. Srivastava, "Network intrusion detection in big dataset using spark," Procedia computer science, vol. 132, pp. 253–262, 2018.

AUTHORS:

Aruna J. she received her bachelor degree and master degree in computer science and engineering from sathyabama university, Chennai. Currently she is pursuing her Ph.D in sathyabama university, Chennai. Her specializations include network security, cryptography, wireless sensor network.

S.PraylaShyry is currently working as Professor in the Department of Computer Science and Engineering. She acquired her M.E in the field of CSE from Annamalai University and PhD from Sathyabama University in the year 2014 She has actively participated as chairperson in many workshops, conferences. She has also been a reviewer in reputed journals. She has also published more than 50 national, International journals and conferences. Her area of specialization includes cyber security, network security and overlay networks, Artificial Intelligence, Machine Learning. She has also published patents and modelled many products. She has also have granted Projects by the Govt of India.