Innovative Approaches to Malware Detection in Health Sensor Data Through Machine Learning

Mrs.T. NARESH[1], S. NARENDRA[2]

[1] Assistant Professor of MCA, Dept of MCA, Audisankara College of Engineering and Technology (AUTONOMOUS) Gudur (M), Tirupati (Dt), AP

[2]PG Scholar, Dept of MCA, Audisankara College of Engineering and Technology (AUTONOMOUS) Gudur (M), Tirupati (Dt), AP

ABSTRACT_ Traditional signature-based malware detection approaches are sensitive to small changes in the malware code. Currently, most malware programs are adapted from existing programs. Hence, they share some common patterns but have different signatures. To health sensor data, it is necessary to identify the malware pattern rather than only detect the small changes. However, to detect these health sensor data in malware programs timely, we propose a fast detection strategy to detect the patterns in the code with machine learning-based approaches. In particular, XGBoost, LightGBM and Random Forests will be exploited in order to analyze the code from health sensor dataTerabytes of program with labels, including benign and malware programs, have been collected. The challenges of this task are to select and get the features, modify the three models in order to train and test the dataset, which consists of health sensor data, and evaluate the features and models. When a malware program is detected by one model, its pattern will be broadcast to the other models, which will prevent malware program from intrusion effectively.

1.INTRODUCTION

Malware, sometimes known as malicious software, is programming (code, scripts, active content, and other software) that is intended to collect data that could be exploited or lead to privacy loss, gain unauthorised access to system resources, disrupt or deny operation, and engage in other abusive activities. It is a generic phrase for a wide range of programme code or software that is unfriendly, obtrusive, or bothersome. Malware is classified as software based more on the creator's alleged intentions than any specific qualities. Computer viruses, worms, Trojan horses, spyware, deceptive adware, crime-ware, the majority of rootkits, and other undesired and harmful software are examples of malware.

2.LITERATURE SURVEY

[1]S. Su, Y. Sun, X. Gao, J. Qiu* and Z. Tian*. A Correlation-change based

Feature Selection Method for IoT Equipment Anomaly Detection. Applied Sciences.

In the era of the fourth industrial revolution, there is a growing trend to deploy sensors on industrial equipment, and analyze the industrial equipment's running status according to the sensor data. Thanks to the rapid development of IoT technologies [1], sensor data could be easily fetched from industrial equipment, and analyzed to produce further value for industrial control at the edge of the network or at data centers. Due to the considerable development of deep learning in recent years, a common practice of such analysis is to conduct deep learning [2,3,4]. Such methods select a subset of all fetched sensor data stream as the input features, and generate equipment predictions. As a result, the performance of the learning model was seriously impacted by the features selected, thus feature selection plays a critical role for such methods.

To select an appropriate set of features for the learning model, researchers aim to select the most relevant features to the prediction model to improve the prediction performance, or to select the most informative features to conduct data reduction. Unfortunately, both kinds of methods have intrinsic drawbacks when applied in the online scenarios. The former kind of methods seriously depends on predefined evaluation criteria, such as feature relevance metrics [5] or a predefined learning model [6]. Thus, such method are limited to certain dataset, and are not suitable for online scenarios which involve dynamical and unsupervised feature selection. The later kind of methods right fits in the online scenarios. However, data reduction mainly aims to improve the efficiency (but not accuracy) of the prediction model, which is not the most concerning factor of online industrial equipment status analysis.

To relieve the dependency of predefined evaluation criteria, researchers switch to select the features which can indicate the online sensor data's characters, such as features which are smoothest on the graph [7], or the features with highest clusterability [8,9]. In this paper, we focus on the features with correlation changes such as smoothness and clusterability, which are important characters for traditional pattern recognition fields like image processing and voice recognition [7,8,9]. We believe that correlation changes can significantly pinpoint status changes in industrial environment. As far as we know, this is the first work focusing on correlation changes for online feature selection.

2.X. Yu, Z. Tian, J. Qiu, F. Jiang. A Data Leakage Prevention Method Based on the Reduction of Confidential and Context Terms for Smart Mobile Devices. Wireless Communications and Mobile Computing, https://doi.org/10.1155/2018/5823439.

With the development of Internet and information technology, smart mobile devices appear in our daily lives, and the problem of information leakage on smart mobile devices will follow which has become more and more serious [1, 2]. All kinds of private or sensitive information, such as intellectual property and financial data, might be distributed to unauthorized entity intentionally or accidentally. And that it is impossible to prevent from spreading once the confidential information has leaked.

According to survey reports [3, 4], most of the threats to information security are caused by internal data leakage. These internal threats consist of approximate 29% private or sensitive accidental data leakage, approximate 16% theft of intellectual property, and approximate 15% other thefts including customer information, and financial data. Further, the consensus of approximate 67% organizations shows that the damage caused from internal threats is more serious than those form outside.

Although laws and regulations have been passed to punish various behaviors of intentional data leakage, it is still hard to prevent data leakage effectively. Confidential data can be easily disguised by rephrasing confidential contents or embedding confidential contents in nonconfidential contents [5, 6]. In order to avoid the problems arising from data leakage, lots of software and hardware solutions have been developed which are discussed in the following chapter.

In this paper, we present CBDLP, a data leakage prevention model based on confidential terms and their context terms, which can detect the rephrased confidential contents effectively. In CBDLP, a graph structure with confidential terms and their context involved is adopted to represent documents of the same class, and then the confidentiality score of the document to be detected is calculated to justify whether confidential contents is involved or not. Based on the attribute reduction method from rough set theory, we further propose a pruning method. According to the importance of the confidential terms and their context, the graph structure of each cluster is updated after pruning. The motivation of the paper is to develop a solution which can prevent intentional or accidental data leakage from insider

effectively. As mixed-confidential documents are very common, it is very important to accurately detect the documents containing confidential contents even when most of the confidential contents have been rephrased.

[3]Y. Sun, M. Li, S. Su, Z. Tian, W. Shi, M. Han. Secure Data Sharing Framework via Hierarchical Greedy Embedding in Darknets. ACM/Springer Mobile Networks an

Geometric routing, which combines greedy embedding and greedy forwarding, is a promising approach for efficient data sharing in darknets. However, the security of data sharing using geometric routing in darknets is still an issue that has not been fully studied. In this paper, we propose a Secure Data Sharing framework (SeDS) for future darknets via hierarchical greedy embedding. SeDS adopts a hierarchical topology and uses a set of secure nodes to protect the whole topology. To support geometric routing in the hierarchical topology, a two-level bit-string prefix embedding approach (Prefix-T) is first proposed, and then a greedy forwarding strategy and a data mapping approach are combined with Prefix-T for data sharing. SeDS guarantees that the publication or request of a data item can always pass through the corresponding secure node, such that security strategies can be performed. The experimental results show that SeDS provides scalable and efficient end-to-end communication and data sharing.

3.PROPOSED SYSTEM

In this project, we mainly focus on static code analysis. The early static code analysis methods mainly include feature matching or broad-spectrum signature scanning. Feature matching simply uses feature string matching to complete the detection, while the broad-spectrum scanning scans the feature code and uses masked bytes to divide the sections that need to be compared and those that do not need to be compared. Since both methods need to get malware samples and extract features before they can be detected, the hysteresis problem is serious. Furthermore, with the development of malware technology, malware begins to deform in the transmission process in order to avoid being found and killed, and there is a sudden increase in the number of malware variants. The shape of the variants changes a lot so that it is difficult to extract a piece of code as a malware signature.

3.1 IMPLEMENTATION

1. Data Collection: Collect sufficient malware code samples and legitimate software samples.

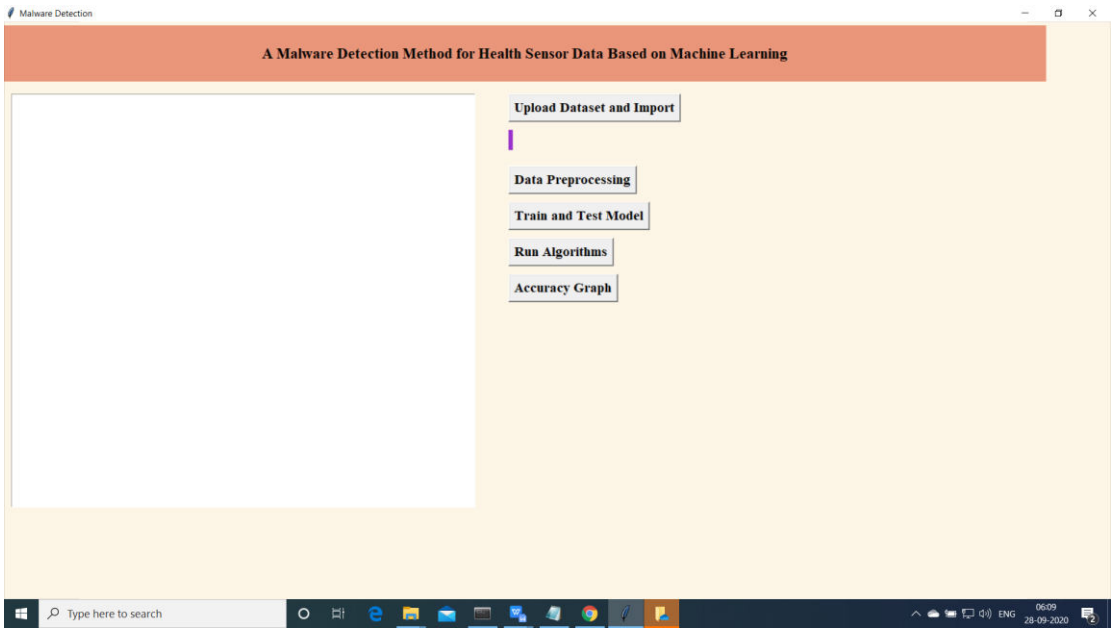
2. Data Preporcessing:Perform effective data processing on the sample and extract the features.

3. **Train and Test Modelling:** Split the data into train and test data Train will be used for trainging the model and Test data to check the performance

4. Feature Selection: Further select the main features for classification.

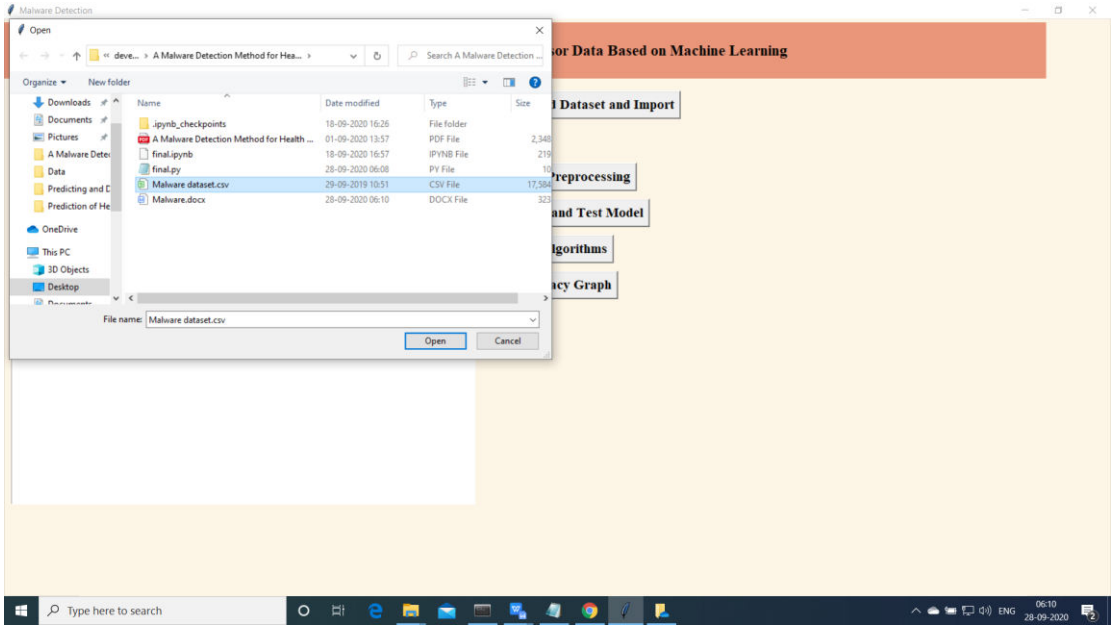
5. Modelling: SVM Navie bayes, Random Forest XGbooost, lightgbm. Combine the training using machine learning algorithms and establish a classification model.

4.RESULTS AND DISCUSSION

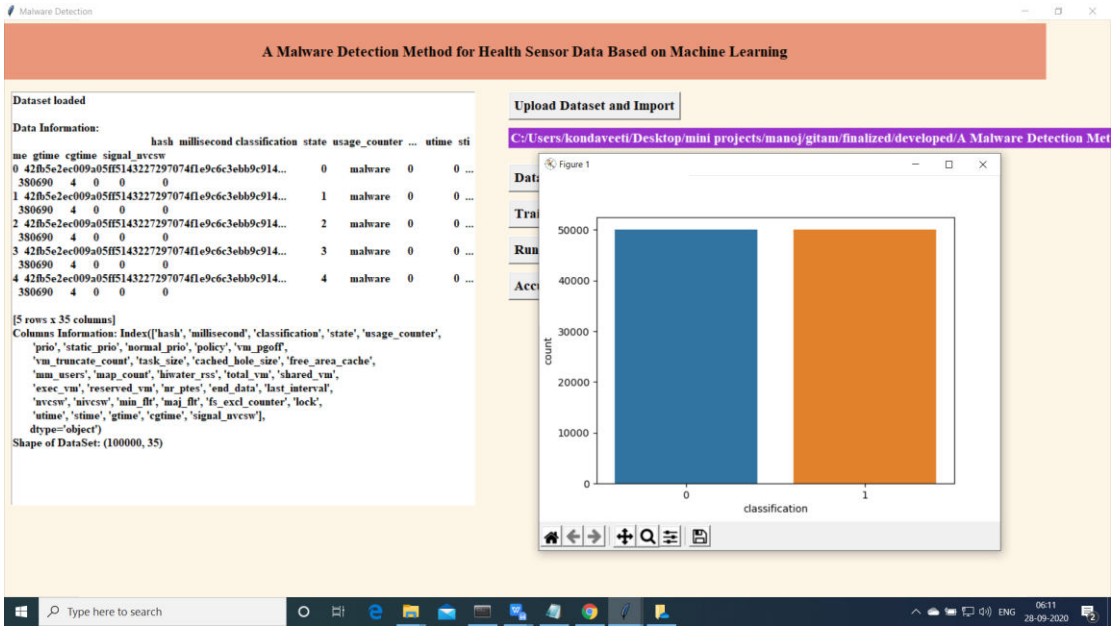


Above screen will be opened.

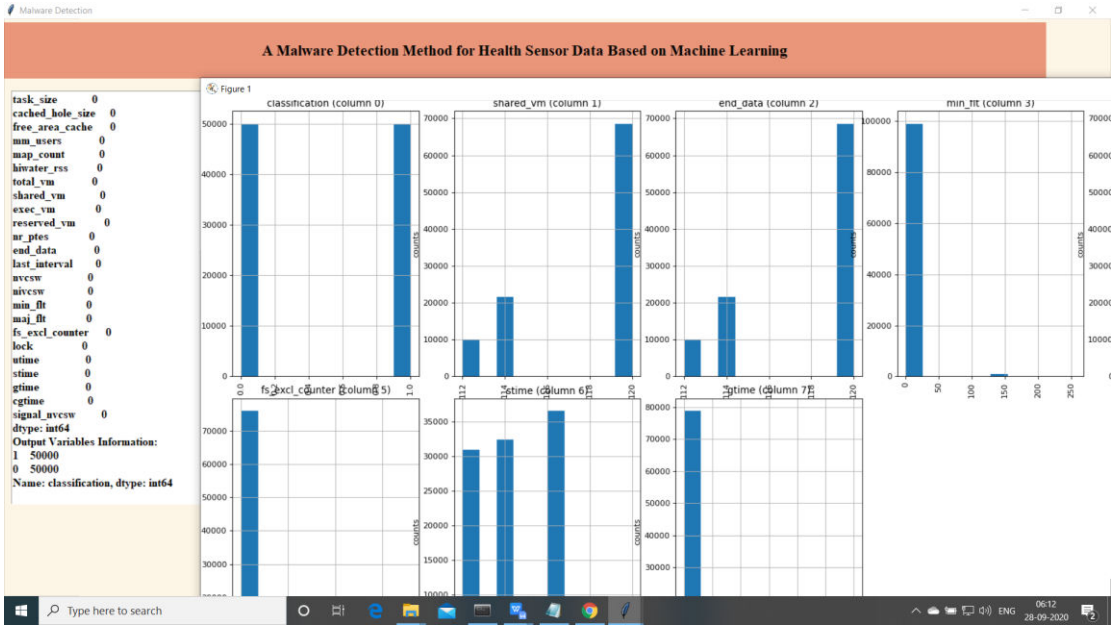
1. Now click on “Upload data and import”



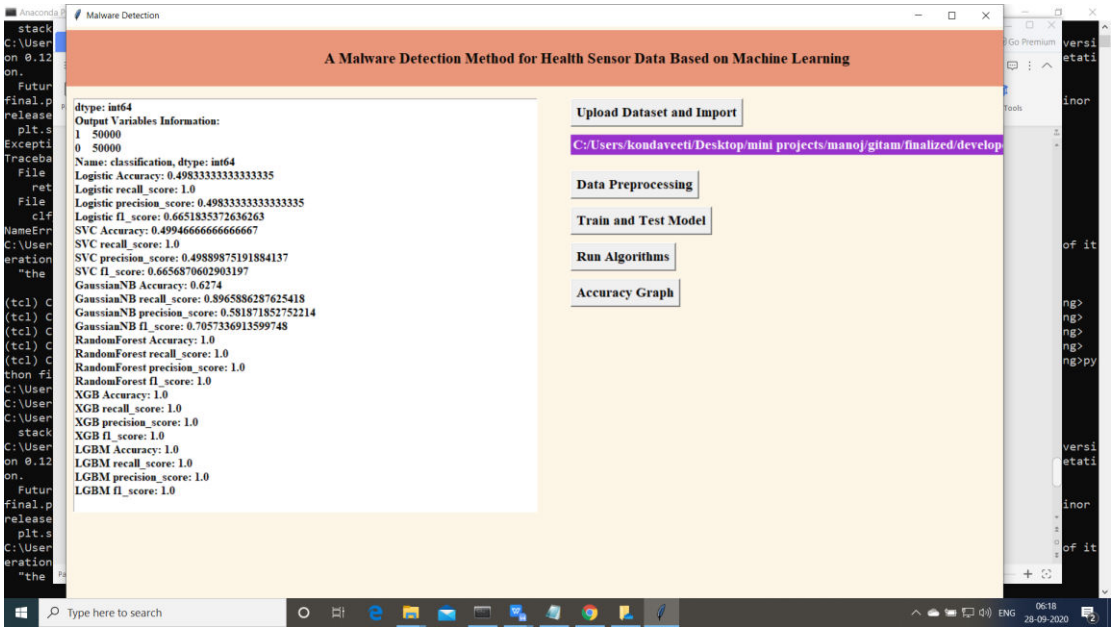
Upload the data and read the basic data information will be shown on the screen



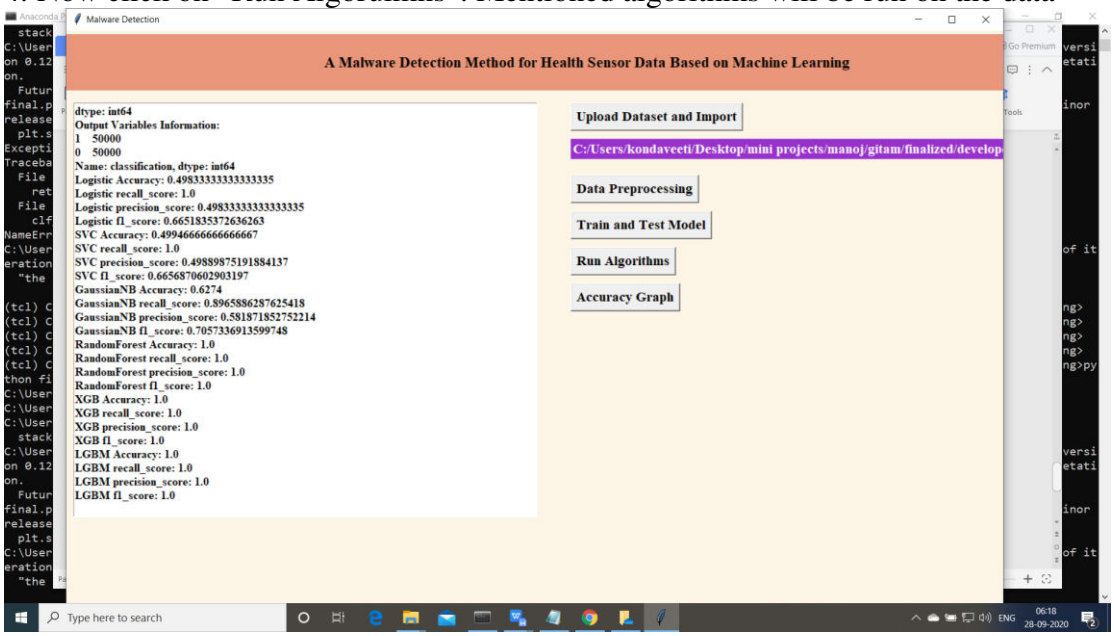
2. Now click on preprocessing. Basic preprocessing will be done



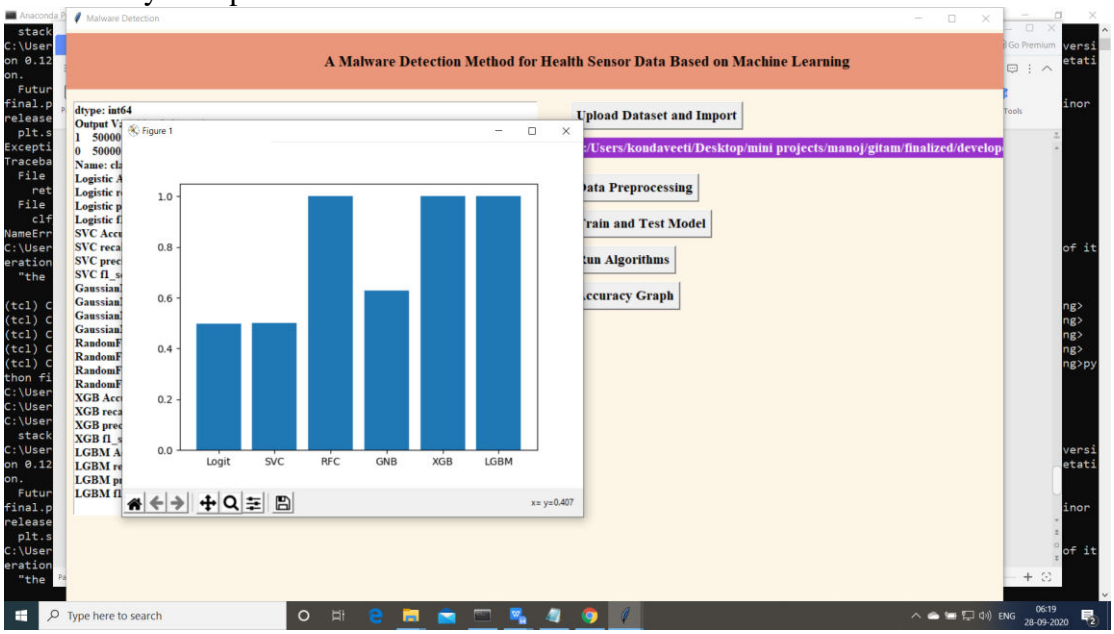
3. Now click on “Train and Test model”. split the data into train and test and traain will be used for training and to tets the performace we are using test data



4. Now click on “Run Algoruimns”. Mentioned algorithms will be run on the data



5. Accuracy Comparision for all the models



Navie bayes algorithm is performed better

Extension is Navie Bayes and perfromed well compare to other algorithms

5.CONCLUSION

With the increasing complexity of malware codes concealed in health sensor data the application of machine learning algorithms in the detection of malicious code has been increasingly valued by the

academic community and numerous security vendors. Based on the theory of machine learning, this paper combines the advantages of different models and discusses the static code analysis based on different machine learning algorithms and different code features. This work can provide referential value for the future design and implementation of malware detection technology for machine learning .However, this area still belongs to the developmental stage. There are still many future tasks and challenges and they are summarized below. 1. Lack of valuable data: A machine learning algorithm often requires tens of thousands of data [5] to be trained in order to get an effective model. The acquisition of these basic data often requires manual operations and the speed cannot be guaranteed .

REFERENCES

[1] L. Wu, X. Du, W. Wang, B. Lin, "An Out-of-band Authentication Scheme for Internet of Things Using Blockchain Technology," in Proc. of IEEE ICNC 2018, Maui, Hawaii, USA, March 2018.

[2] M. Shen, B. Ma, L. Zhu, R. Mijumbi, X. Du, and J. Hu, "Cloud-Based Approximate Constrained Shortest Distance Queries over Encrypted Graphs with Privacy Protection", IEEE Transactions on Information Forensics & Security, Volume: 13, Issue: 4, Page(s): 940 – 953, April 2018, DOI: 10.1109/TIFS.2017.2774451.

[3] P. Dong, X. Du, H. Zhang, and T. Xu, "A Detection Method for a Novel DDoS Attack against SDN Controllers by Vast New Low-Traffic Flows," in Proc. of the IEEE ICC 2016, Kuala Lumpur, Malaysia, 2016. [4] Z. Tian, Y. Cui, L. An, S. Su, X. Yin, L. Yin and X. Cui. A Real-Time Correlation of Host-Level Events in Cyber Range Service for Smart Campus. IEEE Access. vol. 6, pp. 35355-35364, 2018. DOI: 10.1109/ACCESS.2018.2846590.

[5] Q. Tan, Y. Gao, J. Shi, X. Wang, B. Fang, and Z. Tian. Towards a Comprehensive Insight into the Eclipse Attacks of Tor Hidden Services. IEEE Internet of Things Journal. 2018. DOI: 10.1109/JIOT.2018.2846624. [6] Z. Wang, C. Liu, J. Qiu, Z. Tian, C., Y. Dong, S. Su Automatically Traceback RDP-based Targeted Ransomware Attacks. Wireless Communications and Mobile Computing. 2018. https://doi.org/10.1155/2018/7943586.

[7] L. Xiao, Y. Li, X. Huang, X. Du, "Cloud-based Malware Detection Game for Mobile Devices with Offloading", IEEE Transactions on Mobile Computing, Volume: 16, Issue: 10, Pages: 2742 –

2750, Oct. 2017. DOI: 10.1109/TMC.2017.2687918.

[8] https://en.wikipedia.org/wiki/Malware_analysis

[9] Z. Tian, W. Shi, Y. Wang, C. Zhu, X. Du, et al., "Real-Time Lateral Movement Detection Based on Evidence Reasoning Network for Edge Computing Environment", IEEE Transactions on Industrial Informatics, Volume: 15, Issue: 7, Page(s): 4285 – 4294, March 2019.

[10]L. Xiao, X. Wan, C. Dai, X. Du, X. Chen, M. Guizani, "Security in mobile edge caching with reinforcement learning", IEEE Wireless Communications Volume: 25, Issue: 3, pp. 116-122, June 2018, DOI: 10.1109/MWC.2018.1700291.

Author Profiles

Mrs. T. NARESH currently he has working Assistant Professor in Audisankara College of Engineering & Technology Gudur(M), Tirupati (DT), he is done M.Tech from Jawaharlal Nehru Technological university , Ananthapur in 2014.

S. NARENDRA is pursuing MCA from Audisankara college of Engineering &Technology (AUTONOMOUS), Gudur, Affiliated to JNTUA in 2024. Andhra Pradesh, India.