*School of Information Technology Engineering, VIT - Vellore*

Network Information and Security

ITE4001

**J-Component Project**

Zeroeth Review

1. ***Title of the project***

Encrypted Network Traffic Analysis using unsupervised learning methods

2, 3. ***Name and registration number of the students***

1. Aradhita Bhandari – 18BIT0313
2. Disha Bhaglal – 18BIT0468

4. ***Problem Definition***

Most modern applications generate encrypted traffic in order to protect the contents of all the communications between clients and servers. Ideally, this encrypted traffic would be legitimate, however it may contain anomalies or be malicious. Various research has been performed on the use of supervised machine learning models for the analysis of encrypted network traffic. However, supervised learning algorithms require carefully labelled data, and depend on those labels for training. Networks generate a huge amount of unlabelled data, that could potentially be used for unsupervised learning. In this project, we aim to implement unsupervised learning algorithms for encrypted network traffic and to compare the results against the supervised learning algorithms. For this purpose, we will be implementing the K-means clustering and DBscan algorithms. We will use random forest and AdaBoost for comparison as they have given some of the best results in existing research. Once we have identified the two best performing algorithms, we will attempt to improve their results by employing various techniques.

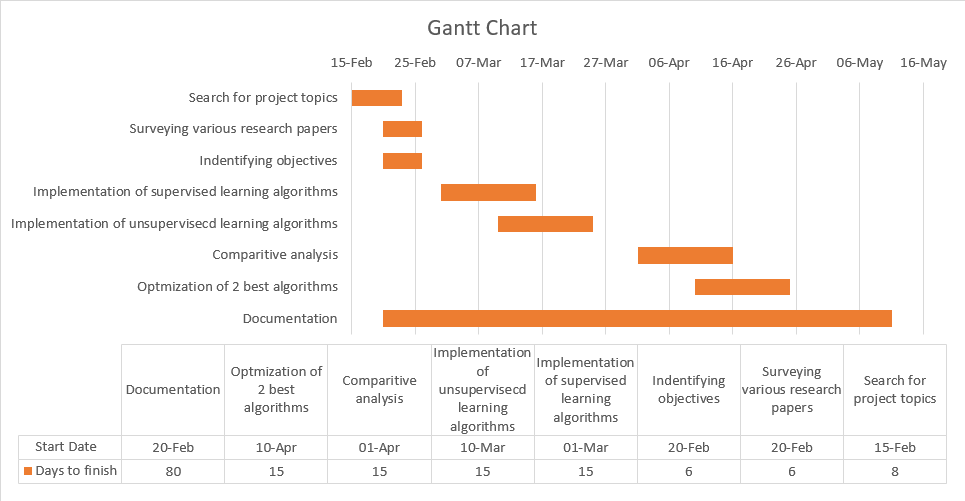
5. ***List of objectives to be achieved***

1. Developing a highly efficient technique using machine learning to identify malicious activity in encrypted network traffic
2. Determining effectiveness of novel approach of unsupervised algorithms for encrypted network traffic fingerprinting
3. Performing a comparative analysis between the chosen unsupervised and supervised learning algorithms for encrypted network traffic analysis
4. Using various techniques in an effort to improve the results of the two best performing networks

6. ***Expected outcomes and timeline***

A technique for encrypted traffic analysis that employs unsupervised learning so that it can make use of real-world unlabelled data and efficiently detect abnormal or malicious activity

*Timeline depicted via Gantt Chart:*



7.  ***References***

[1] Akbari, I., Salahuddin, M. A., Ven, L., Limam, N., Boutaba, R., Mathieu, B., Moteau, S., & Tuffin, S. (2021). A Look Behind the Curtain: Traffic Classification in an Increasingly Encrypted Web. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, *5*(1), 1–26. <https://doi.org/10.1145/3447382>

[2] Bagui, S., Fang, X., Kalaimannan, E., Bagui, S. C., & Sheehan, J. (2017). Comparison of machine-learning algorithms for classification of VPN network traffic flow using time-related features. *Journal of Cyber Security Technology*, *1*(2), 108–126. <https://doi.org/10.1080/23742917.2017.1321891>

[3] Barut, O., Grohotolski, M., DiLeo, C., Luo, Y., Li, P., & Zhang, T. (2020a). Machine Learning Based Malware Detection on Encrypted Traffic: A Comprehensive Performance Study. *7th International Conference on Networking, Systems and Security*, 45–55. <https://doi.org/10.1145/3428363.3428365>

[4] Cai, X., Nithyanand, R., Wang, T., Johnson, R., & Goldberg, I. (2014). A Systematic Approach to Developing and Evaluating Website Fingerprinting Defenses. Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, 227–238. <https://doi.org/10.1145/2660267.2660362>

[5] Cherubin, G., Hayes, J., & Juarez, M. (2017). Website Fingerprinting Defenses at the Application Layer. *Proceedings on Privacy Enhancing Technologies, 2017*(2), 186–203. <https://doi.org/10.1515/popets-2017-0023>

[6] Leroux, S., Bohez, S., Maenhaut, P.-J., Meheus, N., Simoens, P., & Dhoedt, B. (2018). Fingerprinting encrypted network traffic types using machine learning. *NOMS 2018 - 2018 IEEE/IFIP Network Operations and Management Symposium*, 1–5. <https://doi.org/10.1109/noms.2018.8406218>

[7] Limthong, K., & Tawsook, T. (2012). Network traffic anomaly detection using machine learning approaches. 2012 IEEE Network Operations and Management Symposium, 1–4. <https://doi.org/10.1109/noms.2012.6211951>

[8] Moustafa, N., & Slay, J. (2015a). UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). *2015 Military Communications and Information Systems Conference (MilCIS)*, 1–6. <https://doi.org/10.1109/milcis.2015.7348942>

[9] Moustafa, N., & Slay, J. (2016). The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set. *Information Security Journal: A Global Perspective, 25*(1–3), 18–31. <https://doi.org/10.1080/19393555.2015.1125974>

[10] Msadek, N., Soua, R., & Engel, T. (2019). IoT Device Fingerprinting: Machine Learning based Encrypted Traffic Analysis. *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, 1–8. <https://doi.org/10.1109/wcnc.2019.8885429>

[11] Muliukha, V. A., Laboshin, L. U., Lukashin, A. A., & Nashivochnikov, N. V. (2020). Analysis and Classification of Encrypted Network Traffic Using Machine Learning. *2020 XXIII International Conference on Soft Computing and Measurements* (SCM), 194–197. <https://doi.org/10.1109/scm50615.2020.9198811>

[12] Perera, P., Tian, Y.-C., Fidge, C., & Kelly, W. (2017a). A Comparison of Supervised Machine Learning Algorithms for Classification of Communications Network Traffic. *Neural Information Processing*, 445–454. <https://doi.org/10.1007/978-3-319-70087-8_47>

[13] Velan, P., Čermák, M., Čeleda, P., & Drašar, M. (2015). A survey of methods for encrypted traffic classification and analysis. *International Journal of Network Management, 25*(5), 355–374. <https://doi.org/10.1002/nem.1901>