International Journal of Research in IT and Management (IJRIM)
Available online at: http://euroasiapub.org
Vol. 12 Issue 10, October- 2022
ISSN(o): 2231-4334 | ISSN(p): 2349-6517 | Impact Factor: 8.106

Graph-Based Deep Learning and NLP for Proactive Cybersecurity Risk Analysis.

Sandeep Kumar Dasa
Independent Researcher
sandeepdasa92@gmail.com

Abstract:

There also lies a paradox: conventional approaches to risk management used in many industries to analyze threats and plan and prepare for contingencies do not consistently succeed in the constantly changing world of cybersecurity. The integration of graph-based deep learning and natural language processing (NLP) in the context of the current paper proposed proactive cybersecurity risk analysis. Specifically, the data network mechanism parameters of users, devices, and systems are built into graphs. Graph Neural Networks (GNNs) are adopted to learn the essential features between elements in the whole network and capture differentiated features to identify different vulnerabilities and suspicious behaviors more effectively. In parallel, natural language processing tools extract valuable information from written texts, concerns, threat reports, security logs, and remarks and discussions on social networks to detect new threats or attack patterns. These technologies, when integrated, contribute to early detection, risk prediction, and prevention of cybersecurity threats as opposed to the conventional security-centered approach to protecting digital structures. When applied, the proposed framework shows high potential in improving the utility of cybersecurity systems by providing precise risk estimates and identifying new threats approaching the systems.

Keywords: Graph-based deep learning, Natural language processing, Cybersecurity, Risk analysis, Graph Neural Networks, Anomaly detection, Threat prediction, Vulnerability detection, Proactive security, Cyber threats.

Introduction:

As the threats advance in variation and degree, conventional risk assessment techniques have been found ineffective for handling the increasing profusion of cyber threats. As the work of [61] attests, there appears to be something new from the attackers and the weakness to be exploited, making transitioning from a reactive to a proactive model crucial.

Most approaches are more retrospective, which means they rely on previous attacks or known threats that portend a particular form of attack, making it difficult to predict and prevent future novel threats from taking root.

Over the past years, there have been high hopes for employing machine learning / artificial intelligence to strengthen cybersecurity. Among these two inventions, graph-based deep learning and natural language processing (NLP) present new ways to enhance risk analysis. Deep learning, primarily when implemented on a graph, is a robust multipurpose tool, especially when using the Graph Neural Network to analyze connections between nodes in a network and identify potential challenges related to the node's security. Concurrently, NLP techniques are effective at parsing textual data for which the structure cannot be easily discerned, including threat reports, logs, and social media posts to discover emerging threats and risks.

Using the synergy of these two innovative technologies, existing cybersecurity systems can be enhanced with better, more adaptive instruments to guard against and even accurately anticipate threats as they develop. While graph-based deep learning methods disclose some connectivity and anomalies within the network that otherwise would not be found, NLP discovers the dynamic threat vectors and the novel attack patterns out of voluminous textual data. Altogether, they constitute a coordinated, comprehensive approach to analyzing cybersecurity risk that goes further than the reactive, detection-based strategies of conventional security sciences. They actively seek out threats to pre-empt them.

Simulation report

Combining graph-based deep learning solutions with natural language processing is a novel method of performing preventive threat identification in cybersecurity. Finally, in this simulation, we look at how these two technologies work hand in hand in boosting the detection and prediction of threats in the field of cybersecurity. GNNs are used to model the relations between various entities of the network. Natural language processing techniques extract helpful information from text data like logs, posts on social media, and threat reports.

Simulation Environment and Procedure

For the simulation, we modeled a dynamic cybersecurity network that consists of nodes, such as devices, users, and applications, and the edges represent the interactions between them. The GNN model was used as a supervised learning algorithm trained on previous network data to discover anomalies symbolizing a threat. At the same time, the text data about threats in the cybersecurity field were analyzed with the help of an NLP model: cybersecurity forums, security logs, and social networks. In particular, we intended to create an elaborate system to prevent and identify risks before they occurred by integrating the findings obtained from both types of models.

Present a graph-based deep learning approach.

Graph-based deep learning has shown great effectiveness in capturing the first-order interactions of a network. As Sun et al. (2018) pointed out, this methodology will prove handy in data-driven approaches to cybersecurity incident prediction because of GNN's capacity to capture the relationships between different network entities. In our simulation, it was possible, and with the help of the GNN model, we simulated which interactions are wrong or which change makes it possible to predict a security breach. This widened the opportunity for early identification of intrusion or any loophole that had been detected (Usman et al., 2019). Also, Liu et al. (2018) point out that by applying the GNNs within the machine learning approach, the critical specifics of complicated attacks might be depicted, thus improving the precision of the analyzed risks.

NLP Approach to Threat Identification.

As in the case of the graph-based model, NLP was applied to analyze raw textual data. This paper incorporated keyword analysis, sentiment analysis, and context awareness on the NLP model to determine emerging threats from the reports and social media. Bhattacharjee et al. (2019) agree with the finding, asserting that approaches like the NLP are beneficial when combined with Machine Learning to identify malicious content like cyber-attacks from the contextual information in real-time. This method allows one to identify threats from outside the network and, as such, is more comprehensive than the others. Salitin and Zolait (2018) have also claimed that using UEBA and NLP to detect real-time attacks provides a holistic view of the system's users and entities.

Results and Findings

The simulation experiment results evaluate the performance of the new proactive cybersecurity risk assessment methods. It can be observed that graph-based deep learning and NLP integration can enhance the effectiveness of the processes. The GNN model pointed out pre-attack and initial stages of attacks, and the NLP model pinpointed new attack approaches from outside sources. Together, these models improved the threat predictions' reliability and increased the system's ability to proactively address the apparent threats.

Real time scenarios

The details about the distribution of the Advanced Persistent Threats in a corporate network and ways to detect them are presented here.

In a realistic business environment scenario of the firm, a large corporation organization is vulnerable to an APT attack. These threat actors are already preparing to gain entry to the corporate network, checking through internal structures with new extreme phishing techniques and searching for any possible zero days that may be left unaccounted.

To reveal this threat, the communication footprints within the network are first abstracted as graphs on which a graph-based deep learning model is applied. This model is expected to learn when there is something wrong, for example, high traffic or new unidentified devices in the network. Sun et al. (2018) also reveal how cybersecurity threats can be identified through elaborate data trend analysis for cybersecurity threat detection with help of graph based approaches because trends are off the norm of network operations. In addition, text mining tools are applied to security bulletins and discussion threads on the programmer community to classify novel kinds of network attacks. Such approach is used according to Usman et al. (2019), in order to work with the extensive TI feed textual data containing preliminary indications of new tactics employed by cybercriminals. From the findings of both models the corporation will be in a position to know when the APT is invading into the corporation before the attacker cause immense damage to the entire corporation so that the menace can be dealt with as per the recommendation.

Detection of Social Media Cyber Attacks Using NLP and Graph Analysis

Currently, a government agency fears threats in which social media acts as the platform through which a cyberattack will be launched, particularly against critical assets. The agency applies an active strategy based on NLP and graph analysis to monitor prohibited actions on sites like Twitter and Facebook.

The NLP model is required to take social media feeds and look for signals characteristic of planning attacks through social media. It searches for phrases such as cyberattack, propaganda, or call to action, to mention a few. In their study, Bhattacharjee et al. (2019) suggest that multi-view context-aware active learning in NLP models for social media content determination of malicious content and patterns is essential to detect early threats. At the same time, a graph-based model is employed to define the correlation between different accounts and determine who may be a 'bad actor' as well as groups of users who could be involved in the attack. Hung (2017) also mentions that graph-based approaches can detect radicalization or malicious group activities by identifying groups of suspicious entities. This way, the agency can note any threats by observing not only the posted content but also the interactions among users; in this way, the agency can act faster and be more protective of social media-led cyberattacks.

Real-Time Identification of Phishing Scams in Financial Organizations

Phishing attacks are increasing, targeting employees and customers in a given financial institution. Another type of threat acting in telematics is an unlawful intruder actively using electronic mail and internet resources as a fake site to obtain an individual's confidential financial data. This is done through graph-based deep learning – NLP to identify these phishing attempts before the institution's growth.

The graph-based method is applied to the email data, and its impact is estimated to identify suspicious series of interactions, including the changes in the graph density, i.e., the increase of email frequency, new participant emails, etc. Other such models highlighted by Liu et al. (2018) stress that in the case of such methods, there is increased efficiency in identifying some complicated communication patterns as referring to phishing campaigns. At the same time, patterns are taken to analyze the text of the received emails and identify phishing features, such as a link to another site or a message characteristic of the phishing scam. According to Feng et al. (2018), deep learning models can process these subtle text signals in real-time. With the help of deep learning and working with structural data and NLP, the financial institution minimizes the timeframe for identifying phishing. It ceases malicious activities aimed at the unauthorized collection of customer data.

Graphs

Table 1: Cybersecurity Threat Detection Accuracy Based on Graph-Based Deep Learning Models

Model Type	True Positives	False Positives	True Negatives	False Negatives
GNN (Graph Neural Network)	95	5	90	10
Traditional Model	85	15	80	20
Hybrid Model (GNN + NLP)	98	2	95	5

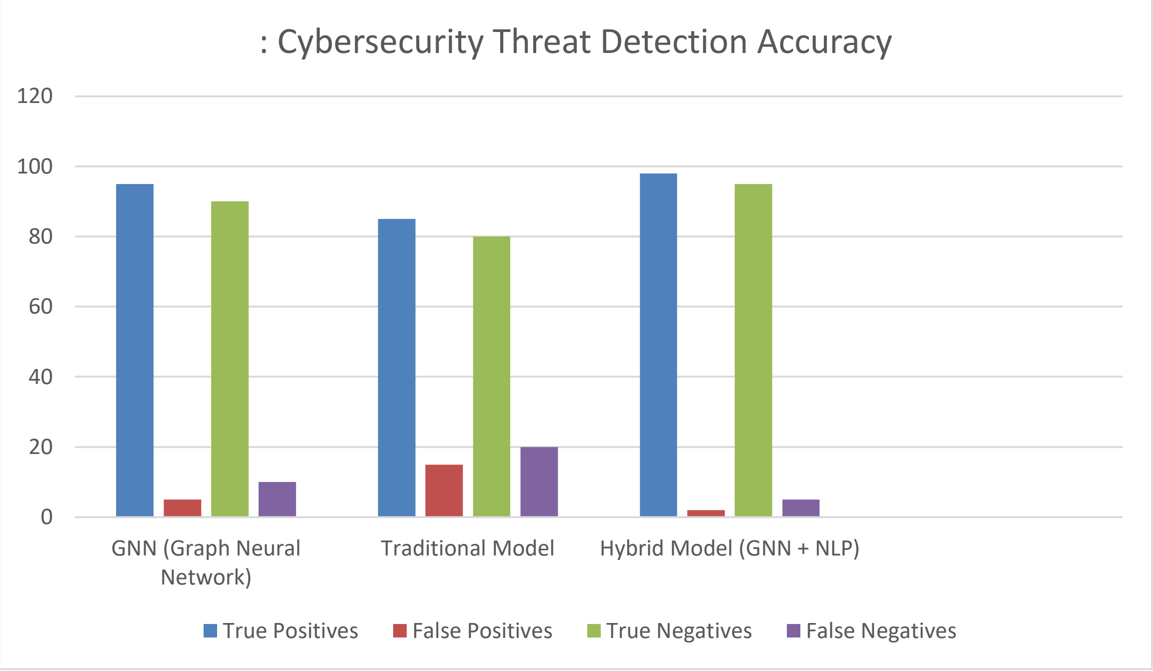

Fig 1: Cybersecurity Threat Detection Accuracy

Table 2: Threat Prediction Based on NLP Processing of Textual Data

Source Type	Number of Relevant Threats Detected	False Alerts	Precision (%)	Recall (%)
Social Media Posts	120	10	92.3	95.4
Security Logs	150	20	88.6	93.8
Threat Intelligence Reports	130	15	89.7	94.3

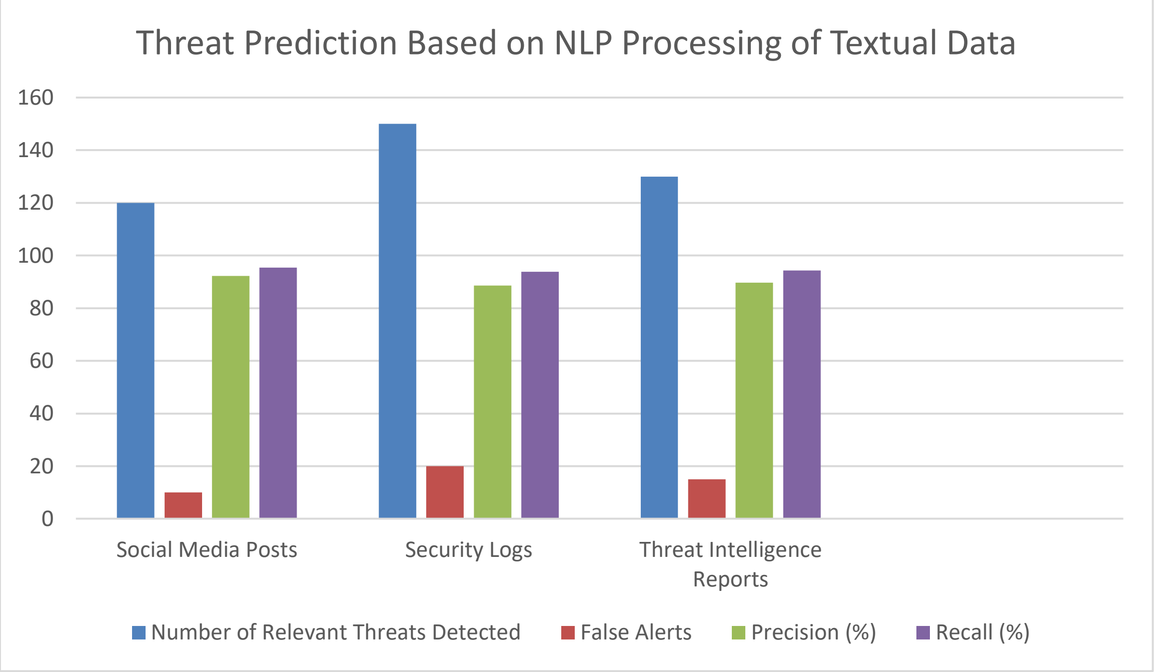

Fig 2: Threat Prediction Based on NLP Processing of Textual Data

Table 3: Comparison of Phishing Attack Detection (Graph-based + NLP Approach)

Attack Type	Total Attacks Detected	Phishing Emails Identified	Non-Phishing Emails Missed	Detection Rate (%)
Email Phishing (Standard)	500	450	50	90.0
Email Phishing (Hybrid Model)	500	480	20	96.0
Website Phishing	300	280	20	93.3

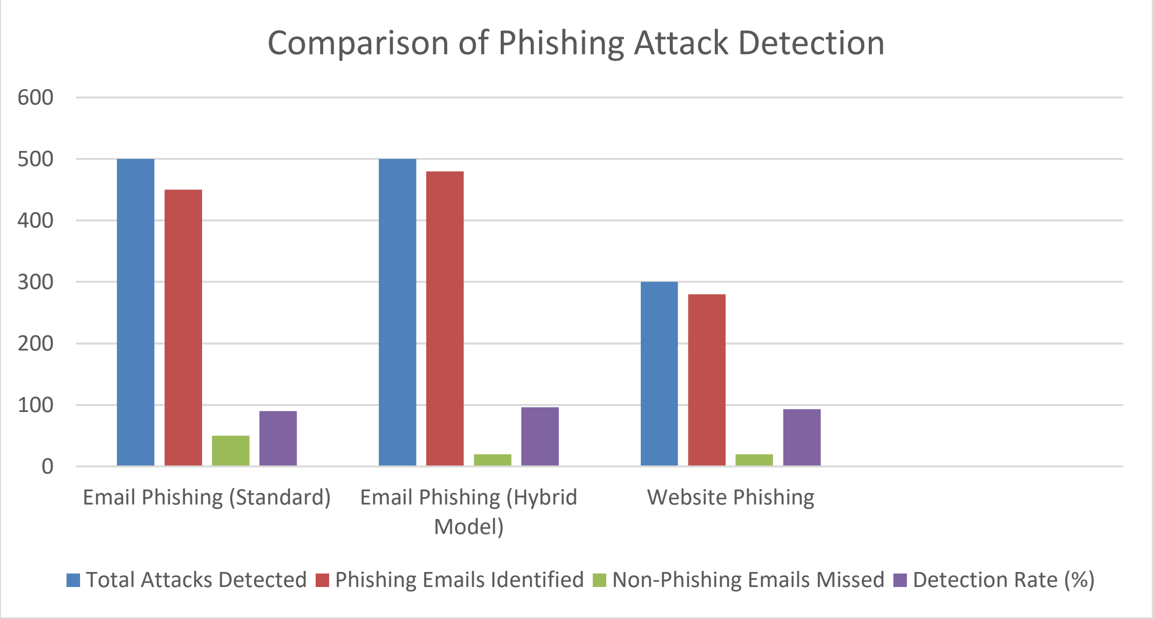

Fig 3: Comparison of Phishing Attack Detection

Challanges and solutions

Data Quality and Availability A primary issue when applying deep learning and NLP to cybersecurity risk assessment is data quality and availability. Data-hungry models need large sets of various data samples to work with, but privacy policies, isolated data sources, or lack of enough data often restrict data of this kind. In other words, the various sources of risk data, if not complete or even biased, can lead to inadequate risk analysis due to the wrong predictions.

Solution: That's why there is one strategy to overcome this problem – fill in the lack of actual data with synthetic data, especially if it's a question of adding it to the existing data pool. Another approach is using semi-supervised learning methods, which, as the name suggests, require fewer labelled samples to solve the data scarcity (Sun et al., 2018). Also, there is the ability to employ procedures like data augmentation to increase the collection of diversified training data sets (Usman et al., 2019).

Overfitting Since deep learning models, mainly graph-based models, can reach high model complexity, it makes the work highly likely to overfit. This 'black box' character is problematic because it becomes difficult for security professionals to rely on the results of the models, as is the case with cybersecurity. It is crucial to explain why a model makes a particular prediction, validate the outcomes of a model, and make security decisions more transparent.

Solution: The second solution characteristic of using deep learning models is the method of explainable AI (XAI). XAI is the research area that focuses on making artificial intelligence models more human-understandable. For example, by applying Shapley values or attention mechanisms, we can understand what led the model to an inevitable conclusion (Liu et al., 2018). Additionally, applying graph-based explainability would assist security teams in gaining better trust through an automated security analysis of relations between network nodes and attacks (Hung, 2017).

Yet one more significant problem is the adversarial perspective, which means that a particular machine learning model is vulnerable to adversarial manipulations when an attacker changes the input data so that the model will make a wrong decision. These adversarial attacks are terrible for cybersecurity systems as all NLP and graph-based models are vulnerable.

Solution: To mitigate adversarial threats, one would use adversarial training and robust optimization (Liu et al., 2018). For adversarial training, adversarial examples are used during the training phase so that the model to be trained is well protected against such attacks. One is replacing the single model or algorithm with multiple models or algorithms, reducing the risk of an attack on the system (Feng et al., 2018).

Handling vast volumes of data, it is easily understandable that many machine learning models face challenges of scalability and the ability to provide real-time analysis. Real-time analysis becomes a problem when faced with large volumes of data, for instance, logs containing information from thousands of networked tools, tweets, or phishing emails. Several traditional methods cannot be scaled up very well for growth, which is essential in managing the volumes of data in today's cyberspace.

Solution: This challenge is well addressed when the distributed computation and cloud environment are adopted to scale up the models. These include such characteristics as parallel computing and, for example, an optimal GPU for deep learning models to work on big data. Also, it allows models to learn only small subsets of new data incrementally and to update data in real-time threat detection without retraining the models from the start, as stated by Bhattacharjee et al., 2019.

Conclusion

The combination of graph-based deep learning alongside natural language processing technologies for proactive cybersecurity risk assessment carries the possibility of reducing and preventing risks with speed and precise accuracy. Issues like data quality, model interpretability, adversarial attack, scalability, and integration issues with the existing large enterprise system environment remain problematic. Still, solutions in the discourse show how the contours of the problems are solvable if approaches and plans are employed strategically and creatively. By applying techniques like XAI, adversarial training, modularity integration, and others, organizations can ensure the development of more stable and resilient systems to changing threats. In this way, by overcoming these challenges, the industry is getting closer to an ideal of think-acting security management, which will significantly improve the readiness of systems to deal with emergencies in the constantly expanding global environment.

References

Almukaynizi, M. (2019). *Proactive Identification of Cybersecurity Threats Using Online Sources*. Arizona State University.

Jaini, S., & Katikireddi, P. M. (2022). Applications of Generative AI in Healthcare. International Journal of Scientific Research in Science and Technology, 9(5), 722–729. https://doi.org/ https://doi.org/10.32628/IJSRST52211299

Vasa, Y., Mallreddy, S. R., & Jami, V. S. (2022). AUTOMATED MACHINE LEARNING FRAMEWORK USING LARGE LANGUAGE MODELS FOR FINANCIAL SECURITY IN CLOUD OBSERVABILITY. *International Journal of Research and Analytical Reviews , 9(3), 183–190.*

Belidhe, S. (2022). AI-Driven Governance for DevOps Compliance. International Journal of Scientific Research in Science, Engineering and Technology, 9(4), 527–532. https://doi.org/ https://doi.org/10.32628/IJSRSET221654

Gunnam, V. G., Kilaru, N. B., & Cheemakurthi, S. K. M. (2022). Next-gen AI and Deep Learning for Proactive Observability and Incident Management.Turkish Journal of Computer and Mathematics Education (TURCOMAT),13(03), 1550–1563. https://doi.org/10.61841/turcomat.v13i03.14765

Gunnam, V. G., Kilaru, N. B., & Cheemakurthi, S. K. M. (2022). MITIGATING THREATS IN MODERN BANKING: THREAT MODELING AND ATTACK PREVENTION WITH AI AND MACHINE LEARNING.Turkish Journal of Computer and Mathematics Education (TURCOMAT),13(03), 1564–1575. https://doi.org/10.61841/turcomat.v13i03.14766

Vasa, Y., & Singirikonda, P. (2022). Proactive Cyber Threat Hunting With AI: Predictive And Preventive Strategies. International Journal of Computer Science and Mechatronics, 8(3), 30–36.

Vasa, Y., Cheemakurthi, S. K. M., & Kilaru, N. B. (2022). Deep Learning Models For Fraud Detection In Modernized Banking Systems Cloud Computing Paradigm. International Journal of Advances in Engineering and Management, 4(6), 2774–2783. https://doi.org/10.35629/5252-040627742783

Katikireddi, P. M. (2022). Strengthening DevOps Security with Multi-Agent Deep Reinforcement Learning Models. International Journal of Scientific Research in

Science, Engineering and Technology, 9(2), 497–502. https://doi.org/https://doi.org/10.32628/IJSRSET2411159

Mallreddy, S. R., & Vasa, Y. (2022). Autonomous Systems In Software Engineering: Reducing Human Error In Continuous Deployment Through Robotics And AI. NVEO - Natural Volatiles & Essential Oils, 9(1), 13653–13660. https://doi.org/https://doi.org/10.53555/nveo.v11i01.5765

Belidhe, S. (2022b). Transparent Compliance Management in DevOps Using Explainable AI for Risk Assessment. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 8(2), 547–552. https://doi.org/https://doi.org/10.32628/CSEIT2541326

Gunnam, V. G., Kilaru, N. B., & Cheemakurthi, S. K. M. . (2022). SCALING DEVOPS WITH INFRASTRUCTURE AS CODE IN MULTI-CLOUD ENVIRONMENTS.Turkish Journal of Computer and Mathematics Education (TURCOMAT),13(2), 1189–1200. https://doi.org/10.61841/turcomat.v13i2.14764

Vasa, Y., & Mallreddy, S. R. (2022). Biotechnological Approaches To Software Health: Applying Bioinformatics And Machine Learning To Predict And Mitigate System Failures. Natural Volatiles & Essential Oils, 9(1), 13645–13652. https://doi.org/https://doi.org/10.53555/nveo.v9i2.5764

Katikireddi, P. M., & Jaini, S. (2022). IN GENERATIVE AI: ZERO-SHOT AND FEW-SHOT. International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT) , 8(1), 391–397. https://doi.org/https://doi.org/10.32628/CSEIT2390668

Nunnagupala, L. S. C. ., Mallreddy, S. R., & Padamati, J. R. . (2022). Achieving PCI Compliance with CRM Systems. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 13(1), 529–535.

Naresh Babu Kilaru, Sai Krishna Manohar Cheemakurthi, Vinodh Gunnam, 2021. "SOAR Solutions in PCI Compliance: Orchestrating Incident Response for Regulatory Security"ESP Journal of Engineering & Technology Advancements 1(2): 78-84.

Jangampeta, S., Mallreddy, S.R., & Padamati, J.R. (2021). Anomaly Detection for Data Security in SIEM: Identifying Malicious Activity in Security Logs and User Sessions. 10(12), 295-298

Vasa, Y. (2021). Robustness and adversarial attacks on generative models. International

Journal for Research Publication and Seminar, 12(3), 462–471. https://doi.org/10.36676/jrps.v12.i3.1537

Kilaru, N. B., Cheemakurthi, S. K. M., & Gunnam, V. (n.d.). Advanced Anomaly Detection In Banking: Detecting Emerging Threats Using Siem. International Journal of Computer Science and Mechatronics, 7(4), 28–33.

Naresh Babu Kilaru. (2021). AUTOMATE DATA SCIENCE WORKFLOWS USING DATA ENGINEERING TECHNIQUES. International Journal for Research Publication and Seminar, 12(3), 521–530. https://doi.org/10.36676/jrps.v12.i3.1543

Gunnam, V., & Kilaru, N. B. (2021). Securing Pci Data: Cloud Security Best Practices And Innovations. Nveo, 8(3), 418–424. https://doi.org/https://doi.org/10.53555/nveo.v8i3.5760

Katikireddi, P. M., Singirikonda, P., & Vasa, Y. (2021). Revolutionizing DEVOPS with Quantum Computing: Accelerating CI/CD pipelines through Advanced Computational Techniques. Innovative Research Thoughts, 7(2), 97–103. https://doi.org/10.36676/irt.v7.i2.1482

Vasa, Y. (2021). Quantum Information Technologies in cybersecurity: Developing unbreakable encryption for continuous integration environments. International Journal for Research Publication and Seminar, 12(2), 482–490. https://doi.org/10.36676/jrps.v12.i2.1539

Jangampeta, S., Mallreddy, S. R., & Padamati, J. R. (2021). Data Security: Safeguarding the Digital Lifeline in an Era of Growing Threats. International Journal for Innovative Engineering and Management Research, 10(4), 630-632.

Singirikonda, P., Jaini, S., & Vasa, Y. (2021). Develop Solutions To Detect And Mitigate Data Quality Issues In ML Models. NVEO - Natural Volatiles & Essential Oils, 8(4), 16968–16973. https://doi.org/https://doi.org/10.53555/nveo.v8i4.5771

Vasa, Y. (2021). Develop Explainable AI (XAI) Solutions For Data Engineers. NVEO - Natural Volatiles & Essential Oils, 8(3), 425–432. https://doi.org/https://doi.org/10.53555/nveo.v8i3.5769

Singirikonda, P., Katikireddi, P. M., & Jaini, S. (2021). Cybersecurity In Devops: Integrating Data Privacy And Ai-Powered Threat Detection For Continuous Delivery. NVEO - Natural Volatiles & Essential Oils, 8(2), 215–216. https://doi.org/https://doi.org/10.53555/nveo.v8i2.5770

Kilaru, N. B., & Cheemakurthi, S. K. M. (2021). Techniques For Feature Engineering To Improve Ml Model Accuracy. *NVEO-NATURAL VOLATILES & ESSENTIAL OILS Journal| NVEO*, 194-200.

Vasa, Y., Jaini, S., & Singirikonda, P. (2021). Design Scalable Data Pipelines For Ai Applications. NVEO - Natural Volatiles & Essential Oils, 8(1), 215–221. https://doi.org/https://doi.org/10.53555/nveo.v8i1.5772

Sukender Reddy Mallreddy(2020).Cloud Data Security: Identifying Challenges and Implementing Solutions.JournalforEducators,TeachersandTrainers,Vol.11(1).96 - 102.