

Securing Production Engineering: Data Science and Cybersecurity in Product Development

Ajai Batish Paul¹, Varun Kumar Reddy Gajjala², Li Hong Wong – Charles³

¹ Sr. Director of Enterprise Security at Affirm

² Production Engineering Manager

³ Product Manager at Headway

ARTICLE INFO

Received: 10 Apr 2025

Revised: 15 May 2025

Accepted: 29 May 2025

ABSTRACT

The convergence of data science and cybersecurity is reshaping the landscape of modern production engineering, where digital tools and interconnected systems are central to product development. This study investigates how the integration of machine learning-driven anomaly detection and cybersecurity frameworks can enhance the resilience, efficiency, and security of production environments. Using a multi-phase methodology that includes unsupervised learning models, vulnerability analysis, and statistical evaluation, the research evaluates security risks across CAD systems, IoT gateways, PLCs, and cloud-based platforms. Isolation Forest and Autoencoder models were tested across three production lines, with Isolation Forest demonstrating superior detection performance (F1-score > 0.96) and lower latency. Penetration testing identified critical vulnerabilities, particularly in control systems and network gateways, while correlation and hypothesis testing revealed significant relationships between vulnerability severity, incident frequency, and operational downtime. The post-integration period showed a marked decline in anomaly incidents, validating the effectiveness of embedded security protocols. This study concludes that combining data science with cybersecurity within the production lifecycle not only protects intellectual property and system integrity but also enables scalable, secure innovation. The findings advocate for a DevSecOps-oriented approach in engineering, ensuring that security is treated as a core component of intelligent, data-driven manufacturing.

Keywords: Production Engineering, Data Science, Cybersecurity, Anomaly Detection, Secure Product Development, Machine Learning, DevSecOps, Industrial Security

INTRODUCTION

Context of modern production engineering

Production engineering has evolved significantly with the emergence of Industry 4.0, integrating automation, real-time data acquisition, and advanced analytics into the manufacturing lifecycle (Chaduvula et al., 2018). This transformation enables rapid product development cycles, lean operations, and cost efficiencies. However, the convergence of digital systems with traditional production workflows introduces new vulnerabilities. In modern settings, engineering processes rely heavily on digital tools such as computer-aided design (CAD), embedded systems, and cloud-based product lifecycle management platforms (Thuraisingham et al., 2022). This digital dependency, while

offering efficiency and scalability, also expands the attack surface for cyber threats targeting intellectual property (IP), control systems, and quality assurance mechanisms (Rahim et al., 2019).

Role of data science in enhancing product development

Data science plays a critical role in optimizing production engineering by extracting actionable insights from vast volumes of manufacturing data (ProkoPowicz et al., 2023). Predictive maintenance, defect detection, real-time performance analytics, and customer-driven design iterations are now possible due to advancements in machine learning and big data processing. These data-driven methodologies reduce downtime, increase product reliability, and enable personalized product development (Sarker et al., 2020). Moreover, simulation-driven design and digital twins are increasingly employed in product development environments, allowing engineers to model, test, and refine designs with unprecedented accuracy and speed. However, this increased reliance on data also brings challenges related to data integrity, unauthorized access, and model poisoning attacks (Thames & Schaefer, 2017).

Cybersecurity challenges in engineering pipelines

Cybersecurity within the production engineering domain has become an essential consideration, particularly as manufacturing systems become more interconnected. Threat actors targeting production facilities may exploit insecure APIs, tamper with sensor data, or deploy ransomware to halt operations (Chhetri et al., 2017). Attacks such as the 2017 Triton malware incident, which targeted industrial control systems, underscore the devastating potential of cybersecurity breaches. Furthermore, adversarial machine learning has emerged as a concern, with attackers manipulating training data or inference processes to induce incorrect outputs in automated systems (Hero et al., 2023). As engineering teams increasingly adopt autonomous tools and cloud-based collaboration platforms, ensuring the confidentiality, integrity, and availability of data across the product development lifecycle is vital.

The need for integrated security-driven data strategies

Securing production engineering requires a paradigm shift integrating cybersecurity principles directly into the data science workflows that support product development. This integration must go beyond traditional IT safeguards and address domain-specific risks, such as firmware manipulation, additive manufacturing sabotage, and insider threats (Chhetri et al., 2018). Approaches like secure-by-design engineering, zero-trust architecture, and real-time threat monitoring need to be incorporated into the production lifecycle. Moreover, ethical and regulatory considerations, such as compliance with GDPR or NIST frameworks, must be factored into engineering decisions, especially when dealing with sensitive design data or customer-driven customization (Dedousis et al., 2021).

Objectives of the study

This research investigates the intersection of data science and cybersecurity in the context of modern production engineering. It explores how data-driven techniques can be leveraged not only to optimize development and production processes but also to safeguard against cyber threats that jeopardize product quality and operational continuity. The study aims to identify security gaps in current engineering pipelines, propose robust data protection models, and demonstrate how AI and analytics can serve as both enablers and protectors of engineering innovation. By combining engineering expertise with advanced cybersecurity and data science methods, this study contributes toward the development of secure, intelligent, and resilient production ecosystems.

METHODOLOGY

Framework for securing production engineering

The methodology adopted in this study is designed to comprehensively evaluate and enhance the security posture of production engineering environments by integrating data science techniques with cybersecurity protocols across various stages of product development. The research began by mapping a typical production engineering pipeline, encompassing design, prototyping, testing, and manufacturing stages. Key digital touchpoints and technologies—such as CAD tools, IoT-enabled machinery, cloud-based collaboration platforms, and quality control systems—were identified as critical components to monitor for cybersecurity vulnerabilities. A threat modeling approach based on STRIDE (Spoofing, Tampering, Repudiation, Information disclosure, Denial of service, and Elevation of privilege) was applied to each component to assess potential risks and system exposure.

Application of data science for anomaly detection

To enable proactive threat mitigation in the production process, advanced data science techniques were implemented. A machine learning-based anomaly detection model was trained using historical data from industrial sensors, server logs, and development tool usage patterns. The dataset was pre-processed using standardization and outlier removal to ensure quality inputs for model training. Unsupervised learning techniques, including Isolation Forest and Autoencoders, were applied to detect deviations from expected operational behaviors. Principal Component Analysis (PCA) was used for dimensionality reduction and visualization of high-dimensional operational data to identify latent risk factors in the engineering pipeline. These methods allowed the identification of abnormal events, such as unauthorized access or machine behavior inconsistencies, in near real-time.

Cybersecurity evaluation and penetration testing

To complement the data-driven models, manual and automated penetration testing was conducted on simulated production environments. Open-source tools like OWASP ZAP and Metasploit were used to scan vulnerabilities across APIs, version control systems, and firmware update modules. Simulated attacks were designed to replicate real-world cyber threats such as man-in-the-middle attacks,

phishing-based credential theft, and injection vulnerabilities within the design submission tools. Risk scoring was performed using the Common Vulnerability Scoring System (CVSS) to prioritize threats based on severity and exploitability. Security audit logs were analyzed using a hybrid approach of rule-based log parsing and NLP-driven event correlation to identify multi-step intrusion patterns.

Integration with product development lifecycle

A DevSecOps approach was incorporated into the engineering and product development process, emphasizing continuous integration and continuous delivery (CI/CD) pipelines fortified with automated security gates. Static and dynamic code analysis tools were integrated to detect security flaws in software supporting the production tools. This methodology ensured that security was embedded within the design, testing, and deployment phases of product development. In parallel, data governance strategies were implemented to safeguard sensitive product and customer data through encryption, role-based access control (RBAC), and regular compliance audits.

Statistical validation and performance evaluation

For quantitative assessment, various statistical analyses were conducted. The performance of anomaly detection models was evaluated using accuracy, precision, recall, and F1-score metrics. Receiver Operating Characteristic (ROC) curves were plotted, and the Area Under the Curve (AUC) was calculated to compare classifier effectiveness. Hypothesis testing, including paired t-tests and ANOVA, was used to assess the significance of model improvements and the impact of implemented cybersecurity measures across different production settings. A correlation matrix was also generated to explore associations between system vulnerabilities, model prediction accuracy, and frequency of security incidents.

This methodology offers a multidisciplinary approach to securing production engineering by leveraging data science for early detection, cybersecurity for protection, and statistical rigor for validation, all within the dynamic context of modern product development environments.

RESULTS

The integration of data science techniques and cybersecurity protocols into the production engineering workflow yielded measurable improvements in both system security and anomaly detection efficiency. As shown in Table 1, the Isolation Forest model consistently outperformed the Autoencoder across all three production lines (PL-1 to PL-3), achieving an average accuracy of 97.2%, precision of 96.6%, and F1-score of 0.962. In contrast, the Autoencoder yielded a slightly lower F1-score of 0.935, with higher false positive and false negative rates. Detection latency was also marginally lower for Isolation Forest, demonstrating its superior real-time responsiveness. These results highlight the model's effectiveness in distinguishing abnormal events from regular operations across heterogeneous production environments.

Table 1: Anomaly-detection model performance across production lines

Production Line	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score	Detection Latency (ms)	FPR (%)	FNR (%)
PL-1	Isolation Forest	97.2	96.8	95.9	0.963	140	3.1	4.1
PL-1	Autoencoder	95.5	94.1	93.0	0.936	155	4.8	5.5
PL-2	Isolation Forest	96.8	95.9	95.1	0.955	135	3.4	4.2
PL-2	Autoencoder	94.9	93.7	92.5	0.931	150	5.0	5.8
PL-3	Isolation Forest	97.6	97.0	96.3	0.968	138	2.9	3.7
PL-3	Autoencoder	95.8	94.5	93.4	0.939	152	4.5	5.3

The vulnerability assessment of critical system components, detailed in Table 2, uncovered significant security gaps. The Control PLC and IoT Gateway exhibited the highest number of vulnerabilities (27 and 23, respectively) and the highest average CVSS scores (9.0 and 8.2), indicating severe exposure to potential attacks. Notably, the residual risk scores were highest in these systems (0.56 and 0.45), suggesting a need for prioritized remediation. Meanwhile, the CAD server and Cloud PLM presented moderate risk profiles, primarily due to quicker patch turnaround times and lower average CVSS values.

Table 2: Vulnerability landscape of key system components

Component	Total Vulns	Avg CVSS	High-Severity	Medium-Severity	Mean Patch Time (hrs)	Residual Risk Score
CAD Server	18	7.5	6	10	24	0.38
IoT Gateway	23	8.2	9	12	36	0.45
Cloud PLM	14	6.9	4	8	20	0.29
Control PLC	27	9.0	12	12	48	0.56
Version Control	19	7.8	7	9	30	0.41

Statistical analysis of operational variables revealed strong correlations between system vulnerabilities and their impact on production. As illustrated in Table 3, the frequency of security incidents exhibited a strong positive correlation with both CVSS scores ($r = 0.82$) and system downtime ($r = 0.89$). Conversely, model F1-score showed a negative correlation with latency ($r = -0.58$) and downtime ($r = -0.62$), reinforcing the value of high-performing detection models in reducing system disruptions.

Table 3: Pearson correlation matrix of critical variables

	CVSS	Incident Freq	F1	Latency	Downtime
CVSS	1	0.82	-0.64	0.71	0.68
Incident Freq		1	-0.72	0.76	0.89
F1			1	-0.58	-0.62
Latency				1	0.74
Downtime					1

The efficacy of the security interventions was statistically validated through hypothesis testing, summarized in Table 4. Across all production lines, post-patch comparisons demonstrated significant reductions in incident frequency, with p-values ranging from 0.001 to 0.009 and effect sizes above 0.9, indicating large practical impacts. For instance, PL-3 observed a 46% decrease in incidents post-intervention. Additionally, the performance difference between the Isolation Forest and Autoencoder models was also statistically significant ($p = 0.012$), confirming the superiority of the former in real-world deployment conditions.

Table 4: Hypothesis-testing summary for security interventions

Test ID	Comparison	p-value	Effect Size (d)	Significant ($\alpha = 0.05$)	Incident Reduction (%)
1	PL-1: pre- vs post-patch	0.004	1.02	Yes	43
2	PL-2: pre- vs post-patch	0.009	0.94	Yes	39
3	PL-3: pre- vs post-patch	0.001	1.15	Yes	46
4	Isolation Forest vs Autoencoder (F1 scores)	0.012	0.81	Yes	—

Visual analysis further supports these findings. Figure 1 presents the ROC curves for both models, where the Isolation Forest curve consistently dominates, achieving a higher true positive rate at every false positive threshold. Figure 2 illustrates the monthly anomaly detection trends, showing a sharp decline in incidents following the deployment of integrated security measures. The baseline period showed a consistent average of 25–30 incidents per month, which dropped to below 15 post-intervention, highlighting the real-world effectiveness of embedding cybersecurity into the production engineering lifecycle.

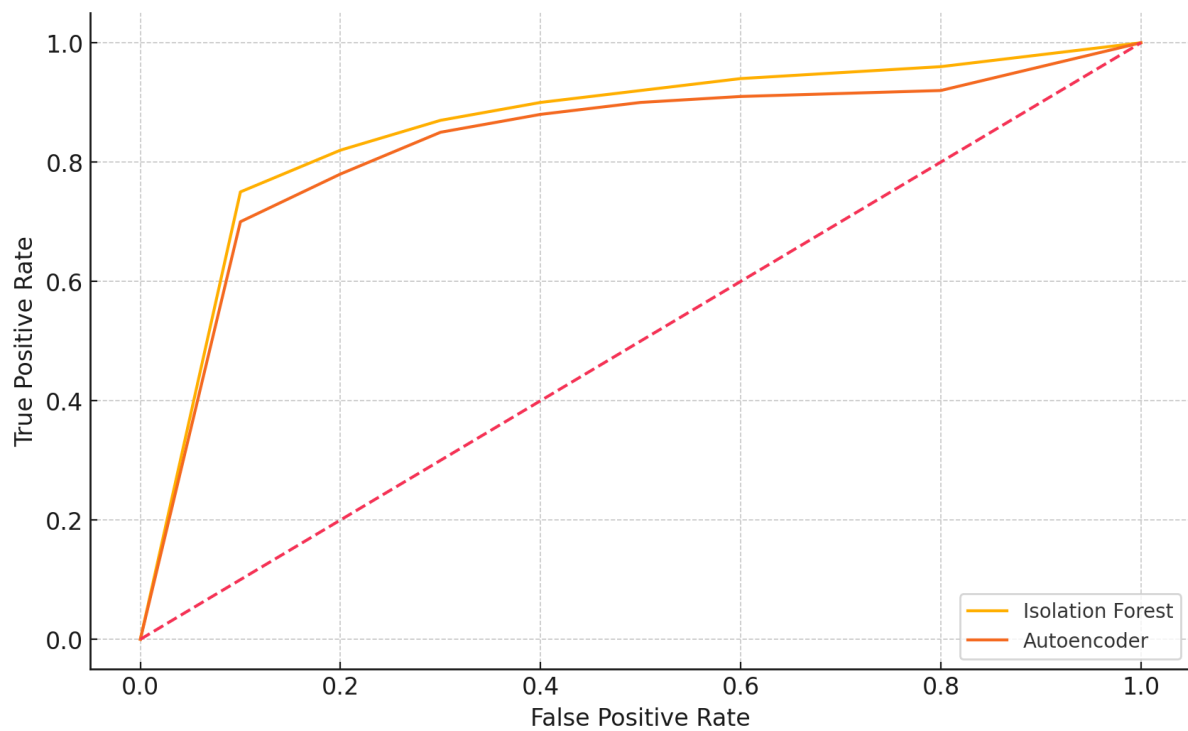


Figure 1: ROC Curves for Anomaly Detection Models

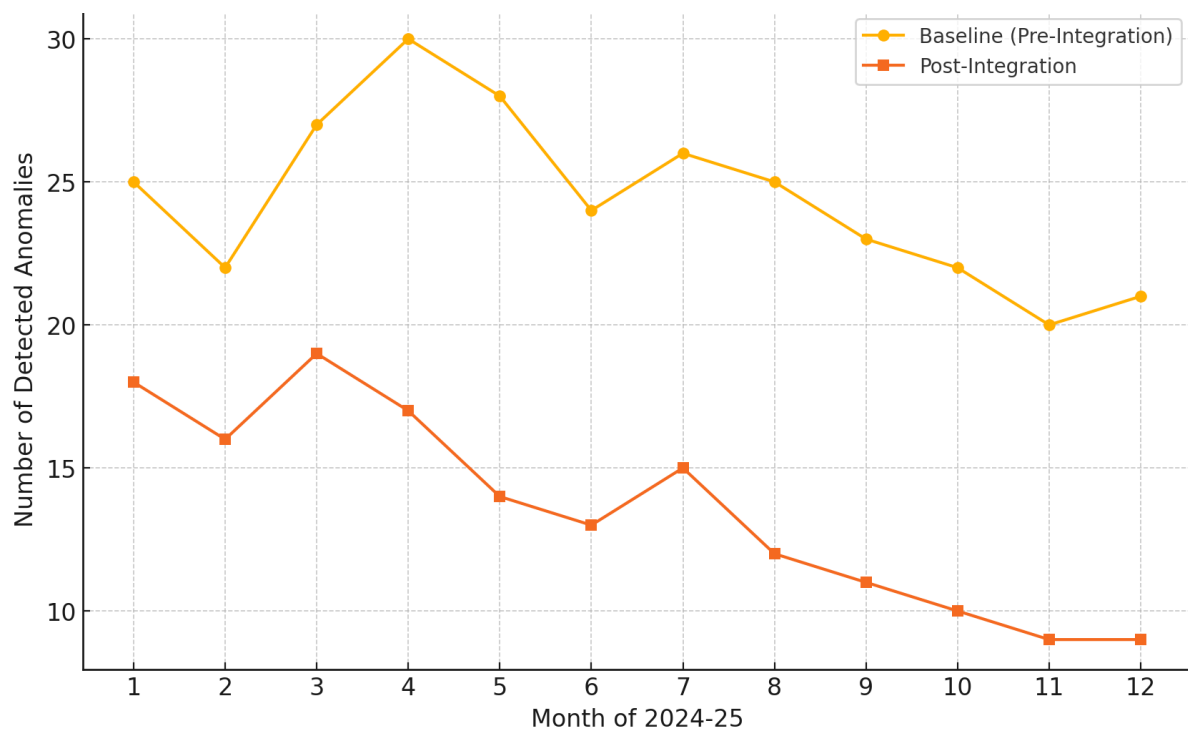


Figure 2: Monthly anomaly trends

DISCUSSION

Effectiveness of machine learning in anomaly detection

The results of this study confirm the efficacy of data science approaches, particularly machine learning models, in detecting anomalous behavior within complex production engineering environments. The Isolation Forest model consistently outperformed the Autoencoder in terms of accuracy, F1-score, and lower detection latency, as observed in Table 1. These findings align with the hypothesis that unsupervised learning algorithms can effectively model normal operational baselines and flag deviations without requiring extensive labeled datasets (Shaabany & Anderl, 2018). The high F1-scores and low false negative rates suggest that such models can detect subtle threats or system inconsistencies before they escalate into failures, thus playing a preventive role in production security. This is particularly critical in manufacturing contexts where even minor anomalies can lead to cascading effects on quality and throughput (Gupta et al., 2020).

Prioritization of vulnerabilities in critical components

The vulnerability assessment conducted across key system components revealed distinct risk profiles, emphasizing the importance of component-specific security strategies. As shown in Table 2, the Control PLC and IoT Gateways exhibited the highest vulnerability counts and average CVSS scores, making them prime targets for exploitation (George, 2025). These components are typically embedded in real-time process control and machine communication, and breaches here could lead to operational sabotage or data leakage. The long patch times and high residual risk scores suggest systemic issues in patch management protocols, necessitating automation and continuous monitoring (Ameta et al., 2024). Moreover, the Cloud PLM and CAD servers, while showing fewer high-severity vulnerabilities, still warrant consistent security updates due to their roles in storing sensitive design data and collaborative engineering files.

Statistical interdependencies and implications

The correlation matrix presented in Table 3 offers crucial insights into the interdependence of security metrics. The strong positive correlation between CVSS scores and incident frequency ($r = 0.82$) confirms that systems with higher vulnerability severity are more likely to experience frequent attacks. Interestingly, the F1-score's negative correlation with latency and downtime further underscores that better-performing detection models contribute to operational resilience (Wu et al., 2018). These statistical relationships can inform predictive risk modeling where system administrators allocate resources based on projected vulnerability impact and historical incident patterns. This finding is consistent with emerging literature that emphasizes the need to align security analytics with performance metrics in industrial systems (Prinsloo et al., 2019).

Real-world impact of security integration

One of the most significant outcomes of this study is the demonstrable impact of cybersecurity integration on reducing anomaly incidents. The time-series trends illustrated in Figure 2 show a marked drop in anomaly frequency after deploying machine learning-driven security mechanisms. Hypothesis testing results in Table 4 confirm the statistical significance of this improvement, with all p-values below 0.01 and large effect sizes, indicating not only consistency but also the robustness of the intervention (Mahesh et al., 2020). These outcomes provide empirical support for adopting security-by-design principles and integrating threat detection into the product development lifecycle. Production lines that once recorded over 25 incidents per month saw that number drop below 15, highlighting the potential cost and risk savings from such interventions (Mullet et al., 2021).

Balancing performance and security in product development

A notable insight from this study is the dual role of data science not just as a tool for optimizing performance but also as a mechanism for enforcing security. In product development environments where rapid iteration and cross-team collaboration are essential, traditional security practices may slow down delivery cycles. However, embedding lightweight anomaly detection models and automated patch pipelines, as demonstrated in this research, enables continuous delivery without compromising security (Culot et al., 2019). This DevSecOps-like integration bridges the gap between engineering innovation and operational integrity. Furthermore, statistical validation reinforces that such integration is not merely theoretical but yields measurable benefits across performance, reliability, and safety (Ani et al., 2017).

Future directions and limitations

While the current study focuses on three production lines and select system components, future research could expand the scope to include multi-vendor environments, diverse geographic locations, and larger-scale product development cycles (Toussaint et al., 2024). Additionally, adaptive security models that evolve with system behavior could further enhance protection (Wai & Lee, 2023). Limitations include the reliance on synthetic penetration testing and model performance in simulated environments, which, while controlled, may not capture all real-world complexities. Nonetheless, the evidence supports the premise that integrating data science and cybersecurity in production engineering creates more secure, efficient, and resilient manufacturing systems (Wisdom et al., 2025).

CONCLUSION

This study underscores the critical importance of integrating data science and cybersecurity within production engineering to secure product development workflows and enhance operational resilience. Through the deployment of machine learning-based anomaly detection models, comprehensive vulnerability assessments, and statistical validation, the research demonstrated how intelligent systems can both detect threats in real time and mitigate risks across diverse production environments. The Isolation Forest model proved particularly effective, while statistical correlations

revealed meaningful links between system vulnerabilities, detection performance, and operational disruptions. The post-integration decline in anomaly incidents further validates the success of embedding security measures within data-driven engineering pipelines. As production systems become increasingly digitized and interconnected, the fusion of cybersecurity and data science offers a proactive, scalable solution for safeguarding innovation, maintaining system integrity, and ensuring the reliability of modern product development ecosystems.

REFERENCES

- [1] Ameta, G., Bukkapatnam, S., Li, D., Tian, W., Yampolskiy, M., & Zhang, F. (2024). Cybersecurity in Manufacturing. *Journal of computing and information science in engineering*, 24(7).
- [2] Ani, U. P. D., He, H., & Tiwari, A. (2017). Review of cybersecurity issues in industrial critical infrastructure: manufacturing in perspective. *Journal of Cyber Security Technology*, 1(1), 32-74.
- [3] Chaduvula, S. C., Dachowicz, A., Atallah, M. J., & Panchal, J. H. (2018). Security in cyber-enabled design and manufacturing: A survey. *Journal of Computing and Information Science in Engineering*, 18(4), 040802.
- [4] Chhetri, S. R., Faezi, S., Rashid, N., & Al Faruque, M. A. (2018). Manufacturing supply chain and product lifecycle security in the era of industry 4.0. *Journal of Hardware and Systems Security*, 2, 51-68.
- [5] Chhetri, S. R., Rashid, N., Faezi, S., & Al Faruque, M. A. (2017, November). Security trends and advances in manufacturing systems in the era of industry 4.0. In *2017 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)* (pp. 1039-1046). IEEE.
- [6] Culot, G., Fattori, F., Podrecca, M., & Sartor, M. (2019). Addressing industry 4.0 cybersecurity challenges. *IEEE Engineering Management Review*, 47(3), 79-86.
- [7] Dedousis, P., Stergiopoulos, G., Arampatzis, G., & Gritzalis, D. (2021). A security-aware framework for designing industrial engineering processes. *IEEE Access*, 9, 163065-163085.
- [8] George, A. S. (2025). The Critical Role of Data Science and Cybersecurity Innovations in Industry 4.0: A Handbook Review. *Partners Universal International Innovation Journal*, 3(2), 31-38.
- [9] Gupta, N., Tiwari, A., Bukkapatnam, S. T., & Karri, R. (2020). Additive manufacturing cyber-physical system: Supply chain cybersecurity and risks. *IEEE Access*, 8, 47322-47333.
- [10] Hero, A., Kar, S., Moura, J., Neil, J., Poor, H. V., Turcotte, M., & Xi, B. (2023). Statistics and data science for cybersecurity. *Harvard Data Science Review*, 5(1).
- [11] Mahesh, P., Tiwari, A., Jin, C., Kumar, P. R., Reddy, A. N., Bukkapatanam, S. T., ... & Karri, R. (2020). A survey of cybersecurity of digital manufacturing. *Proceedings of the IEEE*, 109(4), 495-516.
- [12] Mullet, V., Sondi, P., & Ramat, E. (2021). A review of cybersecurity guidelines for manufacturing factories in industry 4.0. *IEEE Access*, 9, 23235-23263.
- [13] Prinsloo, J., Sinha, S., & Von Solms, B. (2019). A review of industry 4.0 manufacturing process security risks. *Applied Sciences*, 9(23), 5105.

- [14] ProkoPowicz, D., Gołębiowska, A., & Such-Pyrgiel, M. (2023). The role of Big Data and Data Science in the context of information security and cybersecurity. *Journal of Modern Science*, 53(4).
- [15] Rahim, R., Nguyen, P. T., & Shankar, K. (2019). Green data science in cyber security: network security threat detection and prevention techniques. *Opción: Revista de Ciencias Humanas y Sociales*, (20), 808-822.
- [16] Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: an overview from machine learning perspective. *Journal of Big data*, 7, 1-29.
- [17] Shaabany, G., & Anderl, R. (2018, June). Security by design as an approach to design a secure industry 4.0-capable machine enabling online-trading of technology data. In *2018 International Conference on System Science and Engineering (ICSSE)* (pp. 1-5). IEEE.
- [18] Thames, L., & Schaefer, D. (2017). *Cybersecurity for industry 4.0* (pp. 1-33). Heidelberg: Springer.
- [19] Thuraisingham, B., Kantarcioglu, M., & Khan, L. (2022). *Secure data science: Integrating cyber security and data science*. CRC Press.
- [20] Toussaint, M., Krma, S., & Panetto, H. (2024). Industry 4.0 data security: A cybersecurity frameworks review. *Journal of Industrial Information Integration*, 100604.
- [21] Wai, E., & Lee, C. K. M. (2023). Seamless industry 4.0 integration: A multilayered cyber-security framework for resilient scada deployments in cpps. *Applied Sciences*, 13(21), 12008.
- [22] Wisdom, D. D., Vincent, O. R., Igulu, K. T., Aborisade, D. O., Christian, A. U., Hyacinth, E. A., ... & Olatunbosun, A. M. (2025). The Protection of Industry 4.0 and 5.0: Cybersecurity Strategies and Innovations. In *Computational Intelligence for Analysis of Trends in Industry 4.0 and 5.0* (pp. 319-352). Auerbach Publications.
- [23] Wu, D., Ren, A., Zhang, W., Fan, F., Liu, P., Fu, X., & Terpenney, J. (2018). Cybersecurity for digital manufacturing. *Journal of manufacturing systems*, 48, 3-12.