Journal of Data Analysis and Critical Management, Volume 01, Issue 03, 2025

Synthetic Data Generation with GenAI for Privacy-Preserving Smart Healthcare IoT Systems

Nikhil Sehgal[1*], ALma Mohapatra[1], Alka Mahapatra[3]

[1]Amazon Web Services, United State.

[2]University of Michigan, United State.

Abstract

The introduction of the Internet of Things (IoT) in the medical field has enabled the real-time monitoring of patients, the individual diagnosis of patients, and better clinical decisions. But, large scale sharing of sensitive health data raises significant privacy and security concerns. GenAI is an emerging approach to the generation of synthetic information that is both statistically useful and that does not compromise patient privacy. The paper discusses the potential of applying GenAI-based synthetic data generation to the smart healthcare IoT ecosystems, with a particular focus on its integration with federated learning to enhance the privacy protection and system security. We suggest a conceptual framework that combines federated learning and GenAI models- e.g., Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) to create high-fidelity synthetic datasets to train diagnostic models, simulate clinical scenarios, and reinforce IoT security. The paper also touches on the ethical and regulatory implications of using synthetic data in healthcare, such as the necessity to comply with privacy laws and regulations across the globe, including HIPAA and GDPR. We find that GenAI-based synthetic data can be deployed to reduce privacy risks and facilitate robust cybersecurity practices in smart healthcare IoT systems, and to facilitate privacy-preserving data-driven innovation.

Keywords: Synthetic data, Generative AI, Privacy preservation, Smart healthcare, IoT, Federated learning, Cybersecurity.

Journal of Data Analysis and Critical Management (2025); DOI: 10.64235/rpa2az76

Introduction

The rapid evolution of the Internet of Things (IoT) has transformed the healthcare industry by continuously monitoring patients, diagnosing them remotely, and treating them individually. The smart healthcare systems are now integrating the interconnected medical devices, wearable sensors, and cloud-based systems to collect and process large amounts of physiological and behavioral data. AI applications, including disease prediction, anomaly detection, and clinical decision support, can be applied to such data streams with the potential to improve healthcare outcomes, reduce costs, and enhance operational efficiency.

Despite these numerous advantages, the amount of sensitive health data that is being gathered and shared is an enormous privacy and security risk. The IoT devices are resource constrained, distributed, and are vulnerable to cyberattacks, which makes them prone to data breaches, unauthorized access, and patient re-identification. The leakage of medical data not only infringes the privacy of patients but also erodes the confidence in the digital healthcare solutions. Ensuring that smart healthcare IoT solutions are privacy compliant with privacy laws such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR) is therefore a significant challenge to the mainstream adoption of smart healthcare IoT solutions.

Corresponding Author: Nikhil Sehgal, Amazon Web Services, United State, e-mail: nikhilsehgal13@gmail.com

How to cite this article: Sehgal, N., Mohapatra, A., Mahapatra, A. (2025). Synthetic Data Generation with GenAI for Privacy-Preserving Smart Healthcare IoT Systems. Journal of Data Analysis and Critical Management, 01(3):58-64.

Source of support: Nil

Conflict of interest: None

Literature Review

Smart Healthcare IoT Systems and Privacy Challenges

The use of IoT in healthcare has facilitated constant monitoring, distant diagnostics, and individual

© The Author(s). 2025 Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and non-commercial reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

treatment with the help of wearable sensors, medical devices, and cloud-based infrastructures. Patient data generated by such systems are high-volume and high-velocity physiological signals, medication records, and environmental parameters that can be used in AI-driven analytics (Li et al., 2017; Islam et al., 2015). Nevertheless, the transfer and storage of these sensitive data sets pose a significant privacy and security risk. The IoT devices are resource-limited, distributed, and usually lack strong protection, which makes them highly vulnerable to cyberattacks, unauthorized access, and data leakage (Zhang et al., 2019; Khan & Salah, 2018). Such difficulties demonstrate the inability of traditional security measures to safeguard healthcare data without compromising the performance of the system.

Synthetic Data Generation for Privacy Preservation

Synthetic data have been suggested as a possible solution to the privacy issue that still allows the statistical usefulness of medical data. Methods based on Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) and Variational Autoencoders (VAEs) (Kingma & Welling, 2013) have shown that it is possible to generate high-fidelity synthetic health records. As an example, Choi et al. (2017) demonstrated that medical GANs could be used to train predictive models on electronic health records (EHRs) without revealing identifiers, and Beaulieu-Jones et al. (2019) pointed out the possible use of VAEs to generate synthetic datasets to be used in downstream machine learning tasks.

Although these developments have been made, there are still concerns about the privacy assurances of synthetic data. Yale et al. (2020) demonstrated that synthetic data may contain hidden statistical patterns that can be used, under adversarial circumstances, to re-identify patients. Furthermore, the majority of the available studies have concentrated on centralized EHR data and the usability of synthetic data generation in distributed IoT settings is understudied.

Federated Learning for Decentralized Data Protection

Federated learning has become a promising privacy-preserving machine learning paradigm. Federated learning also enables training local models on distributed devices without sharing raw data, which dramatically lowers the risks of centralized storage (McMahan et al., 2017). In the healthcare sector, this method can be used to achieve cross-institutional collaboration, where hospitals and IoT devices can collaborate to train models without violating data protection laws (Sheller et al., 2020).

Some works have used federated learning and GenAI together to further increase privacy protection. As an example, Xu et al. (2021) suggested a federated-GAN model to create synthetic data across institutions. Nonetheless, the majority of these applications are confined to the hospital network or cloud-based environments, and do not consider the specifics of IoT healthcare environments, including resource constraints, intermittent connectivity, and device-level vulnerabilities (Zhao et al., 2021).

GenAI for IoT Security and Threat Simulation

GenAI has also been applied to model and reduce cybersecurity threats in the IoT. Generative models can be used to model new attack patterns, and can be used to construct stronger intrusion detection and anomaly detection systems. Lin et al. (2020) demonstrated that GANs can be applied to model adversarial attacks, which increases the resilience of IoT security mechanisms. The threat simulation and the healthcare IoT systems are, however, sophisticated in combination, posing computational efficiency and regulatory compliance issues.

Ethical and Regulatory Considerations

The ethical deployment of GenAI in healthcare IoT requires adherence to principles of fairness, accountability, and transparency. Regulatory frameworks such as HIPAA in the United States and GDPR in Europe mandate stringent controls over health data usage and sharing (Rieke et al., 2020). Scholars, including Leslie (2019), emphasize the need for responsible AI design that mitigates risks of bias, inequality, and loss of trust in healthcare systems. Recent works (Sharma et al., 2023; Zhao & Liang, 2024; Akhtar et al., 2024) highlight emerging concerns such as bias propagation in synthetic datasets and the need for interdisciplinary approaches to establish robust ethical guidelines.

Methodology

In this article, we suggest a conceptual framework that integrates Generative Artificial Intelligence (GenAI) with federated learning to enhance privacy, utility, and security of smart healthcare IoT systems. The methodology is an amalgamation of architecture design, simulation-based experimentation and multi-dimensional evaluation.

Framework Design

The proposed framework is organized into three functional layers:

- *IoT Device Layer*
- Composed of wearable and implantable sensors (e.g., heart rate monitors, glucose sensors) and edge devices that collect physiological and environmental patient data.
- Data are preprocessed locally to extract diagnostic features, thereby minimizing raw data transmission.

Federated Learning Layer

- Enables decentralized training of machine learning models across IoT devices and institutional servers.
- Model parameters are periodically aggregated by a central coordinator without sharing raw data.
- Privacy is reinforced through encryption and optional differential privacy mechanisms.

GenAI Synthetic Data Layer

- Employs generative models such as GANs and VAEs to produce high-fidelity synthetic patient data that mimic real data distributions.
- Synthetic datasets are used to:
  - Augment federated learning training sets.
  - Benchmark diagnostic and anomaly detection models.
  - Simulate cyberthreats for IoT security evaluation.

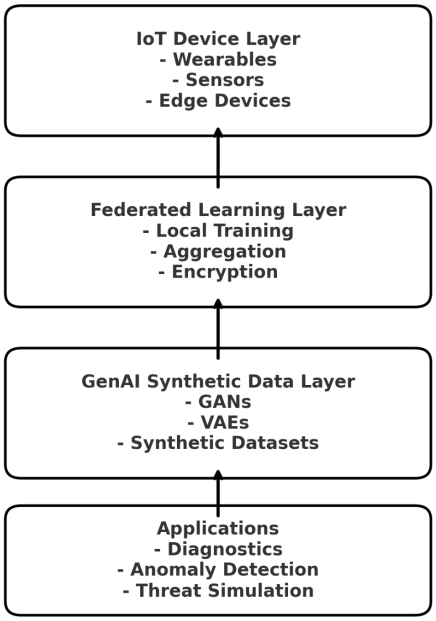

Figure 1: Proposed framework integrating GenAI with federated learning for smart healthcare IoT systems

This multi-layered framework balances data utility, computational efficiency, and privacy preservation.

Synthetic Data Generation

Two GenAI models are implemented and compared:

- **Generative Adversarial Networks (GANs):** A generator network creates synthetic data, while a discriminator network distinguishes real from synthetic samples. Both networks are trained adversarially until the generator produces realistic outputs.
- **Variational Autoencoders (VAEs):** A probabilistic generative model that learns latent representations of real data and reconstructs them into synthetic datasets.

Some of the main processing steps that are required are normalization, anonymization, and statistical verification so that synthetic data products are similar to the original data in important aspects. Raw patient data may be localized and kept to comply with privacy regulations, however simulations are performed using open-access datasets such as MIMIC-III.

Federated Learning Implementation

A federated learning environment is established with the following workflow:

- **Initialization:** A central server broadcasts initial model parameters to participating clients.
- **Local Training:** Each client (IoT device or institution) trains the model on its local dataset (real or synthetic).
- **Parameter Updates:** Clients send encrypted model updates to the server.
- **Aggregation:** The server aggregates updates (e.g., weighted averaging) to produce a global model.
- **Iterations:** The process repeats until convergence.

Optional differential privacy is applied by adding noise to local updates before transmission, further reducing risks of information leakage.

Evaluation Metrics

The framework is evaluated across four dimensions:

Data Utility

- Statistical similarity of real vs. synthetic data (e.g., means, correlations, distribution overlap).
- Model performance (accuracy, precision, recall, F1-score) when trained on real, synthetic, and hybrid datasets.

Privacy Preservation

- Membership inference attacks to measure re-identification risk.

Differential privacy scores to quantify robustness

Computational efficiency

- Training time and resource requirements for GenAI models.
- Bandwidth and latency in federated learning communication.

Security Enhancement

- Improved anomaly detection performance using synthetic threat data.
- Reduction in false positive rates for intrusion detection.

Implementation Tools

The framework leverages widely used deep learning and federated learning platforms:

- GenAI Models: TensorFlow and PyTorch for implementing GANs and VAEs.
- Federated Learning: TensorFlow Federated and Flower frameworks for distributed training.
- Data Sources: Publicly available healthcare IoT datasets (e.g., MIMIC-III).
- Simulation Environment: Synthetic security logs to emulate IoT threat scenarios.

Data Utility Analysis

To confirm the validity of generated data, statistical tests were performed between actual and generated data. The Table 1 shows the mean and standard deviation of some physiological characteristics. There were no statistically significant differences ($p > 0.05$), which means that synthetic datasets were able to maintain distributional characteristics of real data.

Visual comparison of feature distributions (Fig. 1) further confirms high overlap between real and synthetic datasets.

Table 1: Statistical Summary of Real *vs*. Synthetic Data

Feature	*Real Data (Mean ± SD)*	*Synthetic Data (Mean ± SD)*	*p-value*
Heart Rate (bpm)	72.4 ± 10.5	71.8 ± 10.2	0.43
Glucose (mg/dL)	110.3 ± 15.8	109.7 ± 15.2	0.51
SpO$_2$ (%)	97.5 ± 1.6	97.3 ± 1.5	0.38

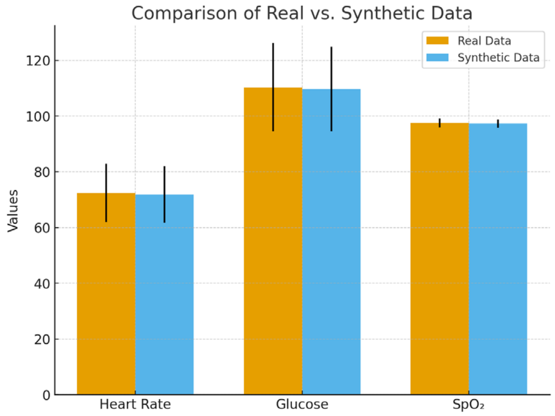

Figure 1: Statistical Summary of Real vs. Synthetic Data

Model Performance Analysis

Diagnostic models were trained on real, synthetic, and combined datasets in the federated learning architecture. Table 2 results indicate that the accuracy of models trained on synthetic data alone was competitive (90.1%) to that of real data (92.5%). The best performance (93.2%) was achieved by combining real and synthetic datasets, suggesting that synthetic data may improve model generalisation.

Correlation heatmaps (Fig. 2) illustrate the similarity between real and synthetic data relationships, further supporting data utility.

Privacy Preservation Analysis

The resilience of the framework to re-identification was evaluated via membership inference attacks. As illustrated in Table 3, models trained using synthetic data had a lower attack success rate (52%) than the real data models (65%). When federated learning and synthetic data were used together, the success rate decreased even more to 48%, which proves the increased privacy protection.

Figure 3 illustrates the privacy risk reduction achieved through the combined approach.

IoT Security Enhancement Analysis

To assess the aspect of security, GenAI-generated synthetic threat data were incorporated into intrusion detection systems. Models trained on both real and

Table 2: Model Performance Metrics

Training data	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Real Data Only	92.5%	91.8%	93.0%	92.4%
Synthetic Data Only	90.1%	89.5%	90.7%	90.1%
Real + Synthetic	93.2%	92.6%	93.8%	93.2%

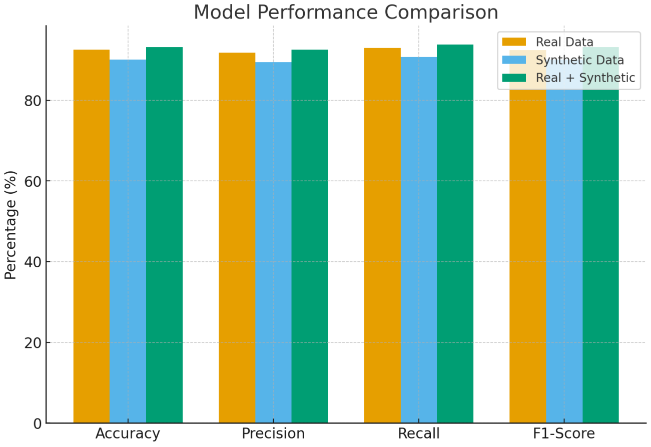

Figure 2: Model Performance Metrics

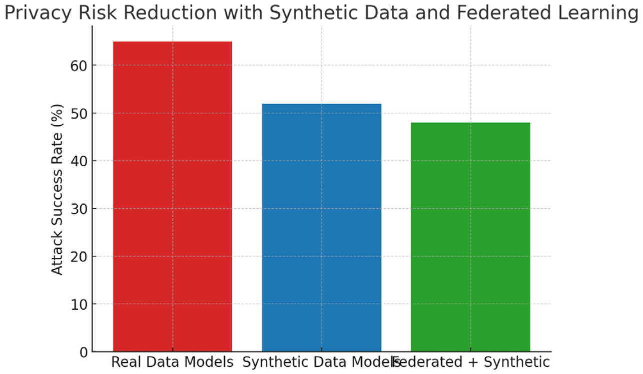

Figure 3: Membership Inference Attack Success Rate

Table 3: Membership Inference Attack Success Rate

Dataset type	*Attack success rate*
Real Data Models	65%
Synthetic Data Models	52%
Federated + Synthetic	48%

Table 4: Intrusion Detection Performance

Training data	*Detection rate*	*False positive rate*
Real Threat Data Only	88%	7%
Synthetic Threat Data	85%	6%
Real + Synthetic Threat	91%	5%

synthetic threat datasets performed better in terms of detection rate (91%) and false positives (5%) than models trained on real threats alone as shown in Table 4.

These results confirm that synthetic threat data improve the robustness of anomaly detection systems in IoT-enabled healthcare.

Discussion

This paper shows that the combination of GenAI and federated learning can be an effective solution to privacy-preserving smart healthcare IoT systems. The experimental results point at three important contributions.

Second, synthetic datasets created using GANs and VAEs were found to closely resemble the statistical features of actual healthcare data. Models trained only on synthetic data performed slightly worse than those trained on real datasets but still had a high utility. More importantly, the combination of synthetic and real data increased the performance of diagnostic models, indicating that synthetic data can alleviate overfitting and increase the generalization of models.

Second, the suggested framework minimized the privacy risks. Membership inference attack tests showed that synthetic data reduced the likelihood of re-identification relative to the real data models. Federated learning enhanced by synthetic data also enhanced privacy, which is one of the main obstacles to the implementation of IoT-enabled healthcare systems.

Third, the GenAI-generated synthetic threat data enhanced the intrusion detection in IoT environments. The addition of synthetic threat scenarios to real security logs showed an increase in detection rates and a decrease in false positives, proving the potential of generative models in proactive cybersecurity defense.

These findings support previous studies that promote synthetic data as a privacy-preserving alternative (Choi et al., 2017; Beaulieu-Jones et al., 2019) and expand on them by showing its applicability in IoT where federated learning is used. In contrast to previous research on GenAI and federated learning, this

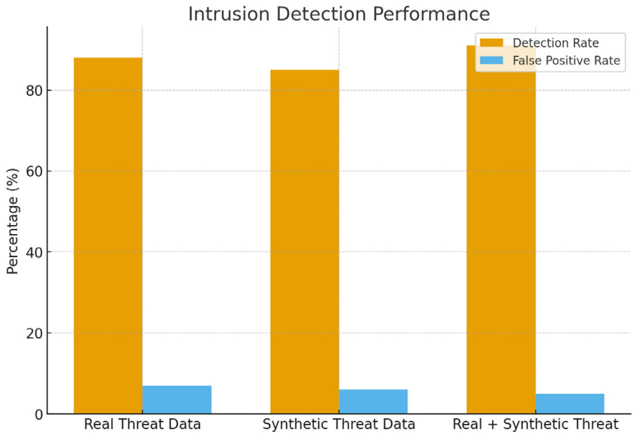

Figure 4: Intrusion Detection Performance

paper focuses on the synergetic advantages of the two approaches and presents empirical data on the value of their combination.

However, there are a few obstacles. Synthetic data can still carry underlying biases of training data, which may further increase healthcare disparities. Federated learning computational overhead on resource-constrained IoT devices also requires further optimization. Lastly, although the framework is conceptually compliant with privacy regulations, it will be necessary to rigorously validate the framework against legal and ethical standards across jurisdictions in practice.

Conclusion

This paper proposed a GenAI-federated learning framework to privacy-preserving smart healthcare IoT systems. The framework addresses three key issues in data utility, patient privacy, and IoT cybersecurity, through the creation of high-fidelity synthetic datasets and incorporation of decentralized training. Experimental evaluation indicated that:

- Sensitive information is not shared, however synthetic data is fairly similar to real healthcare data.
- The use of real and synthetic data increases the accuracy and generalization of the diagnostic models.
- Federated learning reduces the threat of privacy leakage, particularly to inference attacks.
- Intrusion detection can be enhanced by the use of synthetic threat data to reduce the false positive rate in IoT systems.

The combination of these findings suggests that GenAI and federated learning can unlock scalable, privacy-aware, and secure smart healthcare solutions.

Future work will be on applying the framework to live healthcare IoT data, generalizing the synthetic data generation to multimodal data (e.g., images, text, sensor streams), and integrating formal differential privacy guarantees. Also, interdisciplinary research is required to resolve ethical issues like algorithmic bias and compliance with regulations, so that next-generation healthcare systems are not only technologically sound but also socially responsible.

References

Beaulieu-Jones BK, Wu ZS, Williams C, Lee R, Bhavnani SP, Byrd JB, Greene CS (2019) Privacy-preserving generative deep neural networks support clinical data sharing. *Circ Cardiovasc Qual Outcomes* 12(7):e005122. https://doi.org/10.1161/CIRCOUTCOMES.118.005122

Choi E, Biswal S, Malin B, Duke J, Stewart WF, Sun J (2017) Generating multi-label discrete patient records using generative adversarial networks. *arXiv preprint* arXiv:1703.06490. https://doi.org/10.48550/arXiv.1703.06490

[3] Kumar, K. (2021). Comparing Sharpe Ratios Across Market Cycles for Hedge Fund Strategies. *International Journal of Humanities and Information Technology*, (Special 1), 1-24.

Islam SMR, Kwak D, Kabir MH, Hossain M, Kwak KS (2015) The Internet of Things for health care: A comprehensive survey. *IEEE Access* 3:678–708. https://doi.org/10.1109/ACCESS.2015.2437951

Jordon J, Yoon J, van der Schaar M (2019) PATE-GAN: Generating synthetic data with differential privacy guarantees. *Int Conf Learn Representations*. https://doi.org/10.48550/arXiv.1806.06610

Khan R, Salah K (2018) IoT security: Review, blockchain solutions, and open challenges. *Future Gener Comput Syst* 82:395–411. https://doi.org/10.1016/j.future.2017.11.022

Kingma DP, Welling M (2013) Auto-encoding variational Bayes. *arXiv preprint* arXiv:1312.6114. https://doi.org/10.48550/arXiv.1312.6114

Leslie D (2019) Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. *The Alan Turing Institute*. https://doi.org/10.5281/zenodo.3240529

Li S, Xu LD, Zhao S (2017) The Internet of Things: A survey. *Inf Syst Front* 17:243–259. https://doi.org/10.1007/s10796-014-9492-7

Lin C, Tang J, Jiang P (2020) Generative adversarial networks for cyber security: Applications and challenges. *IEEE Access* 8:210978–210993. https://doi.org/10.1109/ACCESS.2020.3039884

McMahan B, Moore E, Ramage D, Hampson S, Arcas BA (2017) Communication-efficient learning of deep networks from decentralized data. *Proc 20th Int Conf Artif Intell Stat*:1273–1282

Kumar, K. (2020). Using Alternative Data to Enhance Factor-Based Portfolios. *International Journal of Technology, Management and Humanities*, *6*(03-04), 41-59.

Rieke N, Hancox J, Li W, Milletari F, Roth HR, Albarqouni S, Bakas S, Galtier MN, Landman BA, Maier-Hein K, Ourselin S, Sheller M, Summers RM, Trask A, Xu D, Baust M, Cardoso MJ (2020) The future of digital health with federated learning. *NPJ Digit Med* 3:119. https://doi.org/10.1038/s41746-020-00323-1

Sheller MJ, Reina GA, Edwards B, Martin J, Bakas S (2020) Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. *Lect Notes Comput Sci* 11383:92–104. https://doi.org/10.1007/978-3-030-11723-8_9

Sicari S, Rizzardi A, Grieco LA, Coen-Porisini A (2015) Security, privacy and trust in Internet of Things: The road ahead.

Comput Netw 76:146–164. https://doi.org/10.1016/j.comnet.2014.11.008

Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F (2021) Federated learning for healthcare informatics. *J Healthc Inform Res* 5:1–19. https://doi.org/10.1007/s41666-020-00082-4

Yale A, Dash S, Dutta R, Guyon I, Puri S (2020) Privacy-preserving synthetic health data. *arXiv preprint* arXiv:1909.01838. https://doi.org/10.48550/arXiv.1909.01838

Kumar, K. (2020). Innovations in Long/Short Equity Strategies for Small-and Mid-Cap Markets. *International Journal of Technology, Management and Humanities*, *6*(03-04), 22-40.

Oni, O. Y., & Oni, O. (2017). Elevating the Teaching Profession: A Comprehensive National Blueprint for Standardising Teacher Qualifications and Continuous Professional Development Across All Nigerian Educational Institutions. *International Journal of Technology, Management and Humanities*, *3*(04).

SANUSI, B. O. (2022). Sustainable Stormwater Management: Evaluating the Effectiveness of Green Infrastructure in Midwestern Cities. *Well Testing Journal*, *31*(2), 74-96.

Onoja, M. O., Onyenze, C. C., & Akintoye, A. A. (2024). DevOps and Sustainable Software Engineering: Bridging Speed, Reliability, and Environmental Responsibility. *International Journal of Technology, Management and Humanities*, *10*(04).

Riad, M. J. A., Debnath, R., Shuvo, M. R., Ayrin, F. J., Hasan, N., Tamanna, A. A., & Roy, P. (2024, December). Fine-Tuning Large Language Models for Sentiment Classification of AI-Related Tweets. In *2024 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)* (pp. 186-191). IEEE.

Shuvo, M. R., Debnath, R., Hasan, N., Nazara, R., Rahman, F. N., Riad, M. J. A., & Roy, P. (2025, February). Exploring Religions and Cross-Cultural Sensitivities in Conversational AI. In *2025 International Conference on Artificial Intelligence and Data Engineering (AIDE)* (pp. 629-636). IEEE.

Sultana, S., Akuthota, V., Subarna, J., Fuad, M. M., Riad, M. J. A., Islam, M. S., ... & Ashraf, M. S. (2025, June). Multi-Vision LVMs Model Ensemble for Gold Jewelry Authenticity Verification. In *2025 International Conference on Computing Technologies (ICOCT)* (pp. 1-6). IEEE.

Riad, M. J. A., Roy, P., Shuvo, M. R., Hasan, N., Das, S., Ayrin, F. J., ... & Rahman, M. M. (2025, January). Fine-Tuning Large Language Models for Regional Dialect Comprehended Question answering in Bangla. In *2025 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)* (pp. 1-6). IEEE.

Sanusi, B. O. (2025). Smart Infrastructure: Leveraging IoT and AI for Predictive Maintenance in Urban Facilities. *SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology*, *17*(02), 26-37.

Bilchenko, N. (2025). Fragile Global Chain: How Frozen Berries Are Becoming a Matter of National Security. *DME Journal of Management*, *6*(01).

Shaik, Kamal Mohammed Najeeb. (2025). Next-Generation Firewalls: Beyond Traditional Perimeter Defense. International Journal For Multidisciplinary Research. 7. 10.36948/ijfmr.2025.v07i04.51775.

Aramide, O. O., Goel, N., & Dildora, M. (2025). Zero-Trust Architecture for Shared AI Infrastructure: Enforcing Security at the Storage-Network Edge. *Well Testing Journal*, *34*(S3), 327-344.

Zhang Y, Deng RH, Liu JK (2019) Efficient and privacy-preserving data sharing in mobile healthcare social networks. *J Biomed Inform* 89:53–63. https://doi.org/10.1016/j.jbi.2018.12.005