| Volume 02 | Issue 06 | November 2026 | ISSN: 3139-1737

International Journal of Advances and Applications in Convergent Studies & Transdisciplinary (IJ-AACST)

Bridging Disciplines, Bridging Knowledge

Aligned with the Vision of NEP 2020 and UGC Guidelines

Website: https://aacstij.org || Email: journaleditor@aacstij.org

A Synthetic Dataset Engineering Exercise for Cyber Fraud Modelling

DOI: https://doi.org/10.5281/zenodo.24102227

Parag Bhalchandra[1]*

*Corresponding Author: [1] School of Computational Sciences, SRTM University, Nanded – 431 606, Maharashtra, India

srtmun.parag@gmail.com

ORCID iD: 0000-0001-5422-9423

Abstract

Many emerging research domains including data sciences, artificial intelligence suffer from a fundamental limitation of real-world datasets due to numerous reasons. A possible solution is in terms of synthetic data generation. However no generic model exists to guide researchers about planned synthetic data generation. The study introduces several important conceptual contributions for synthetic dataset generation and demonstrates its perspective through a case study of cyber-fraud vector synthetic dataset for cybercrime risk modelling. A central conclusion emerging from this research is that datasets must no longer be viewed as static observational artefacts, but as evolving knowledge systems.

Keywords: Synthetic data generation, dataset evolution, data engineering, cyber fraud modelling.

1. Introduction

Many emerging research problems are constrained by the inherent incompleteness of real-world datasets [1]. These limitations arise from restricted access, privacy constraints [2], domain specificity, high collection costs, temporal unavailability, and the absence of important attributes overlooked during initial data design. Consequently, datasets often provide only a partial representation of the underlying phenomenon [3], leading to scientifically valid yet incomplete conclusions that fail to capture hidden behavioural, social, or contextual influences. Synthetic data generation offers a promising solution by systematically identifying and bridging these gaps. Recent studies [4] demonstrate that privacy-preserving synthetic datasets can safely substitute sensitive real-world data while retaining essential statistical and structural characteristics. Rather than serving merely as replacements, synthetic datasets transform static observational data into dynamic, design-driven research assets that support multidimensional analysis, robustness testing, and exploration of latent relationships [4]. This philosophy aligns with the well-known statement, “All models are wrong, but some are useful” [18], emphasizing that datasets is only partial approximations of reality and should evolve through multidimensional enrichment to improve scientific inquiry. The growing demands for data, stricter privacy regulations, and the need for reproducible and scalable experimentation have accelerated interest in synthetic data generation. Modern AI and machine learning models require large, diverse, and well-annotated datasets [12], while

© 2026 The Author(s). This article is published by the AACST in the IJ-AACST under the Creative Commons Attribution 4.0 International (CC BY 4.0) License. The authors retain copyright and grant AACST a non-exclusive license to publish, distribute, and archive the work.

regulatory frameworks often restrict the sharing of sensitive information [7]. Synthetic data addresses these challenges by enabling privacy-preserving, shareable datasets that facilitate open science, cross-institutional collaboration, and ethical AI development [1]. It must also be domain-aware and privacy-safe process for sensitive applications [9]. Furthermore, researchers can create controlled experimental settings by varying specific attributes, amplifying rare events, and introducing edge cases that are difficult to obtain from real-world data [5]. Recent advances in generative AI and diffusion models have further improved the realism and utility of synthetic datasets for applications such as anomaly detection, decision support, and real-time analytics. Overall, the field is shifting from passive data consumption to active data construction, where datasets are viewed as evolving, adaptive systems rather than fixed research inputs. This transition strengthens scientific dataset generation through structured methodologies instead of ad-hoc scripting [8], positioning synthetic data as a strategic instrument for expanding analytical capabilities across domains. These developments strongly support the cross-domain and multidimensional attribute expansion principles underlying the proposed 360° synthetic dataset philosophy [14].

Motivated by these observations, the main contributions of this work is identification of gaps in real-world datasets, understanding synthetic data generation, a comprehensive validation methodology for generated synthetic data and a cyber fraud vector dataset as a case study to support the practical applicability of our perspective.

2. Literature Reviews

The surveyed literature highlights that modern research increasingly depends on both real and synthetic data. However, real-world datasets are often inaccessible because of privacy laws, organizational restrictions, commercial ownership, institutional confidentiality, or national security concerns. In domains such as cyber fraud, medical diagnosis, disaster prediction, student analytics, and sports analysis, researchers frequently understand the problem but lack sufficient data for experimentation [9]. Synthetic data generation (SDG) addresses this challenge by creating artificial datasets that replicate the structure, behaviour, and constraints of real domains [2]. Rather than fabricating evidence, SDG provides controlled environments for testing hypotheses, algorithms, dashboards, and decision-support systems before real-world deployment. Existing studies further confirm that synthetic data enhances experimentation, robustness, and predictive performance in data-scarce settings [13]. A second motivation for SDG arises during early-stage research, where no suitable dataset exists for validating new concepts. This challenge became evident while developing a cyber fraud vector model requiring attributes such as fraud intensity, device change, geo-distance, failed login attempts, social engineering vulnerability, and fraud typology. Synthetic datasets bridge this gap by enabling the transition from conceptual models to algorithmic validation. Importantly, SDG is not a process of randomly generating records but a scientific methodology that integrates domain knowledge, mathematical modelling, statistical distributions, behavioural rules, simulation logic, and validation. Contemporary literature also emphasizes that synthetic datasets must be evaluated through fidelity, utility, privacy, and domain alignment rather than visual similarity alone. A third justification is that real-world datasets are only partial representations of reality, created for specific objectives and limited by available observations. Consequently, an initial dataset often captures only a narrow view of a phenomenon, producing conclusions that may be statistically valid but epistemically incomplete. When new attributes are introduced, the evolved dataset can produce different analytical outcomes,

including changes in feature importance, clustering, classification boundaries, and predictive accuracy. Such changes are a natural consequence of expanding the observational space rather than inconsistencies in the original analysis. This perspective also highlights the importance of dimensional sensitivity analysis, where conclusions evolve as additional dimensions are incorporated. For instance, a cyber fraud dataset containing only technical variables may identify fraud as a technological issue. However, the inclusion of demographic, behavioural, social, and economic attributes may reveal cyber fraud as a socio-technical phenomenon [7]. Thus, datasets should be viewed as evolving knowledge structures rather than static repositories. Our structured synthetic dataset construction and systematic dataset engineering are consistent with the concepts reported in [6]. Probabilistic synthetic data generation using Variational Auto encoders (VAEs) supports the proposed generation layer [11], while recent AI-driven generation architectures and representation learning approaches further strengthen the framework for producing realistic synthetic datasets [16]. In addition, foundational work on Generative Adversarial Networks (GANs) provides the theoretical basis for many modern synthetic data generation engines employed within the proposed methodology [10].

3. Understanding Synthetic Data Engineering

The proposed work introduces the concept of dataset evolution, where synthetic data generation expands the attribute boundary of an existing dataset instead of replacing it. The objective is to create a multidimensional experimental environment that incorporates hidden social, behavioural, demographic, geographic, and temporal factors. Formally, an original dataset may be represented as D1 and proposed synthetic by D2. The difference between D1 and D2 should be interpreted as expansion of attributes and knowledge gained through dimensional expansion.

A synthetic dataset should be designed from the outset as a purpose-bound experimental instrument rather than a finished data product. Its generation must be anchored in an explicit research objective and analytical intention, whether classification, prediction, anomaly detection, simulation, or decision support, so that it serves as a controlled environment for testing hypotheses, validating models, and simulating domain behaviour; this reflects the logic of simulation-driven synthetic environments and controlled experimentation systems [17]. The dataset's shape and scale, its rows and columns, should be optimized according to the research aim, algorithmic requirements, class distribution, and domain variability, with smaller datasets supporting conceptual validation and larger datasets enabling deep learning, temporal modelling, and system-level simulation.

The dataset must also be governed by domain logic and statistical realism. Domain constraints should encode physical, behavioural, and contextual boundaries, such as positive transaction amounts or bounded risk scores in case of cyber fraud modelling. They must follow a domain-aware roadmap [5]. Each attribute should follow a realistic distribution, such as normal, log-normal, Poisson, Bernoulli, or categorical, based on empirical evidence or theoretical justification, so that statistical fidelity is preserved and implausible patterns are avoided. Because real-world phenomena are interdependent, synthetic data must also capture inter-variable dependencies, interactions, and causal or correlational structures. For example, fraud risk may be influenced by device changes and login anomalies, converting isolated variables into an interconnected system, as supported by dependency-modelling theory [15]. The target variable should be explicitly defined and, where necessary, derived through domain-specific formulas or weighted composite scoring schemes, thereby operationalizing the research

objective for supervised learning, ranking, or decision modelling. Ultimately, the dataset must be validated across four dimensions like fidelity to statistical reality, utility for analysis, privacy and safety against sensitive-data replication, and adherence to domain validity. Such validation ensures that the synthetic data are not merely structurally valid but also scientifically credible and suitable for the intended research purpose.

The preceding discussion recasts synthetic data generation as evolutionary. There must be dataset evolution loop which allows a dataset to grow iteratively by incorporating additional attributes, especially variables drawn from other disciplines. There must be correlation to denote the dataset at iteration, newly added attributes, and the resulting analytical outcomes or improvements in the insights. A synthetic dataset should not be judged only by its size or visual appearance. Its research value depends on whether it preserves statistical authenticity, supports experimental usefulness, avoids disclosure of private information, follows domain logic, and regulates bias. Therefore, the general quality of a synthetic dataset can be represented through a synthetic data quality index (SDQI), expressed as follows:

$$SDQI = \alpha Fd + \beta Ut + \gamma Ps + \delta Dv + \lambda Bc \quad (1)$$

$$\alpha + \beta + \gamma + \delta + \lambda = 1 \quad (2)$$

Here, SDQI represents the composite quality score of the generated synthetic dataset. The parameters $\alpha, \beta, \gamma, \delta$ *and* λ *denote the relative importance assigned to each quality dimension.* The sum of all weights is constrained to one so that the index remains normalized and interpretable. The meaning of various symbols is shown in table 1.

Symbol	Meaning	Research interpretation
Fd	Fidelity score	Degree to which the synthetic data statistically resembles the intended domain.
Ut	Utility score	Usefulness of synthetic data for ML, analytics, simulation, dashboard testing or hypothesis validation.
Ps	Privacy/Safety score	Ability of the dataset to avoid replication of real individuals, sensitive cases or confidential records.
Dv	Domain validity score	Extent to which the dataset follows expert rules, physical constraints and contextual logic.
Bc	Bias-control score	Degree to which the dataset avoids unfair or unrealistic skew in demographic, behavioural or outcome variables.

Table 1: Symbols in SDQI

A high-fidelity dataset may still be unsuitable if it violates domain rules. Similarly, a dataset that performs well in machine learning may still be unsafe if privacy protection is weak. Therefore, SDQI provides a balanced quality view rather than a single-dimensional accuracy view.

$$D\{syn\}^{*} = \arg\max SDQI(D\{syn\}) \quad (3)$$

Equation (3) indicates that the preferred synthetic dataset is the one that maximizes the SDQI across all candidate synthetic datasets generated during experimentation. This revels that a weak synthetic dataset is one that only fills rows and columns. A strong synthetic dataset is one that captures behaviour, context, causality, uncertainty, and decision relevance. Now our proposed SDGF is demonstrated for a case study of cyber fraud

vector model.

4. Case Study: Synthetic data generation for Cyber Fraud Vector Model

4.1 Research Context: Cyber fraud is a complex behavioural, technological, financial, and social phenomenon. Real fraud datasets are difficult to obtain because they involve banking information, police records, victim details, IP addresses, mobile numbers, financial losses, and legal proceedings. Therefore, a synthetic cyber fraud vector dataset can help researchers test fraud risk models, AI classifiers, anomaly detection systems, dashboards, and preventive decision-support frameworks.

4.2 Unit of Observation: One row represents one cyber fraud event or suspicious digital transaction.

4.3 Proposed Attributes: we can propose 17 attributes as below in table 2,

Attribute	Type	Explanation
Case_ID	String	Unique case identifier
User_Age	Integer	Age of victim/user
Gender	Category	Male/Female/Other/Unknown
District	Category	Location of incident
Platform	Category	UPI, banking app, social media, e-commerce
Fraud_Type	Category	Phishing, OTP fraud, investment scam, loan app fraud
Transaction_Amount	Float	Amount involved
Failed_Login_Attempts	Integer	Number of suspicious failed logins
Device_Change_Count	Integer	Device switching behaviour
Geo_Distance	Float	Distance between usual and suspicious location
Cross_Border_Flag	Boolean	Whether transaction involves cross-border signal
Time_Anomaly_Score	Float	Unusual transaction time measure
Social_Engineering_Score	Float	Human manipulation intensity
Financial_Loss	Float	Loss amount
Recovery_Status	Category	Full, partial, none
CFV_Score	Float	Composite fraud vector risk
Risk_Level	Category	Low, Medium, High, Critical

Table 2: Attributes in CFV

4.4 Attribute Thinking: The key idea is that cyber fraud is not caused by a single variable. It emerges from the interaction of behavioural, financial, spatial, technological, and social features. Therefore, attributes must represent different fraud dimensions. For example, a transaction of ₹5,000 may not be risky if it occurs from a known device, known location, and normal transaction hour. But the same ₹5,000 may become suspicious if it occurs after midnight, from a new device, after multiple failed logins, with cross-border routing. Thus, synthetic data must model interaction effects, not only isolated variables.

4.5 Dataset Shape: Recommended dataset of 10,000 to 50,000 rows, 20 to 35 attributes, 4 risk classes and class imbalance included. The suggested class distribution is as below in table 3,

Risk Level	Approximate Percentage
Low	45%
Medium	30%
High	18%
Critical	7%

Table 3: Class distribution in CFV

This imbalance reflects real-world fraud detection, where critical events are fewer but highly important.

4.6 Generation Logic: A simple generation pipeline may be understood as generate demographic attributes, generate platform and fraud type, generate transaction amount using skewed distribution, generate behavioural anomaly variables, generate spatial anomaly variables, compute CFV score, assign risk level, inject noise, validate domain constraints and run ML experiments.

4.7 CFV Score Formula: Computation of CFV can take shape like

$$CFV=0.20A+0.20D+0.20G+0.15F+0.15T+0.10SCFV = 0.20A + 0.20D + 0.20G + 0.15F + 0.15T + 0.10SCFV=0.20A+0.20D+0.20G+0.15F+0.15T+0.10S$$

Here A is amount anomaly, D is device change intensity, G is geo-distance anomaly, F is failed login anomaly, T is time anomaly and S is social engineering score

4.8 Experimentation Plan: The synthetic CFV dataset can support Random Forest classification, Logistic Regression baseline, XGBoost fraud risk prediction; K-Means clustering of fraud typologies, SHAP based feature importance, ROC-AUC evaluation, confusion matrix analysis, district-wise fraud dashboard and demographic vulnerability analysis. This highlights applicability of SDG similar to real static datasets.

4.9 Expected Research scope: This case study proves that synthetic data can help formulate, validate, and demonstrate a new fraud-risk model even before police, bank, or cyber-cell datasets become available. It also enables explainable AI experimentation by showing which variables most influence fraud risk.

The CFV case study models cyber fraud as a multidimensional socio-technical phenomenon. Fraud risk is not treated as a single technical anomaly but as a composite effect of transaction behaviour, device irregularity, geographic deviation, failed authentication, time anomaly and behavioural vulnerability.

4.10 Cyber Fraud Vector Score: We can propose a CFV score as seen in below equations and table 4,

$$CFV = w1A + w2D + w3G + w4F + w5T + w6B \quad (4)$$

$$\Sigma\{i=1\}^{6} wi = 1, \quad wi \geq 0 \quad (5)$$

Symbol	Meaning	Possible measurement idea
A	Transaction amount anomaly	Normalized deviation of amount from usual user or population behaviour.
D	Device change score	Frequency or severity of device switching during suspicious activity.
G	Geo-distance score	Distance between usual user location and suspicious transaction location.
F	Failed login score	Number or intensity of failed authentication attempts.
T	Time anomaly score	Unusual timing of transaction compared with normal behavioural pattern.
B	Behavioural vulnerability score	Exposure to social engineering, low digital literacy, trust-based response or risky behaviour.
wi	Expert/model-defined weight	Relative contribution of each factor to cyber fraud risk.

Table 4: Class distribution in CFV

Equation (4) creates a weighted composite risk score. The weights can be assigned through expert judgment, statistical learning, feature importance, or policy priority. Equation (5) normalizes the contribution of all components so that the CFV score remains interpretable.

We further need to go for normalization of CFV components as CFV variables may exist on different scales, so each component should be normalized before aggregation.

$$X\{norm\} = (X - X\{min\}) / (X\{max\} - X\{min\}) \quad (6)$$

Equation (6) converts every raw component into a comparable range between 0 and 1. This prevents high-scale variables such as transaction amount or geographic distance from dominating smaller-scale behavioural variables. Further, a risk-level categorization can be expressed as

$$Risk(CFV) = \{Low, CFV < 30; Medium, 30 \leq CFV < 60; High, 60 \leq CFV < 80; Critical, CFV \geq 80\} \quad (7)$$

Equation (7) converts the continuous CFV score into interpretable risk classes. These classes support supervised learning, risk dashboards, alert generation and policy decision-making.

4.11 CFV Dataset Evolution: The original cyber fraud dataset may initially contain only technical variables. The proposed synthetic strategy expands the dataset by adding social, demographic and behavioural dimensions.

$$D\{CFV\} \wedge \{(1)\} = \{A, D, G, F, T\} \quad (8)$$

$$D\{CFV\} \wedge \{(2)\} = D\{CFV\} \wedge \{(1)\} \cup \{B, S, E, L, R\} \quad (9)$$

Here, B may denote behavioural vulnerability, S social engineering exposure, E education/digital literacy, L locality or region, and R recovery or reporting behaviour.

The movement from Equation (8) to Equation (9) represents dataset evolution from a technology-centric dataset to a socio-technical fraud intelligence dataset. Experimental use of CFV score can be demonstrated as

$$Y\{fraud\} = f(A, D, G, F, T, B) \quad (10)$$

$$P(Fraud = 1 \mid X) = \sigma(\theta 0 + \theta 1A + \theta 2D + \theta 3G + \theta 4F + \theta 5T + \theta 6B) \quad (11)$$

Equation (10) represents the general predictive relationship, where fraud outcome is modelled as a function of technical and behavioural attributes. Equation (11) represents a logistic modelling view, wherein σ denotes the sigmoid function. This can be used for binary fraud classification or as a baseline before applying Random Forest, XGBoost or deep learning models.

The CFV case study as discussed above shows how synthetic data can support a new cyber fraud modelling framework. The example proves that synthetic data can be used across very different domains if a proper roadmap is followed.

5. Conclusion

This work proposed a domain-aware synthetic data engineering roadmap and systematically guides researchers through scientific narrations. Through the CFV case study, we also demonstrated that the same methodological foundation can be effectively extended across diverse domains, thereby establishing the generalizability and domain-agnostic nature of the proposed approach. A central conclusion emerging from this research is that datasets must no longer be viewed as static observational artefacts, but as evolving knowledge systems. As new dimensions, attributes, and disciplinary perspectives are introduced, datasets evolve, interpretations shift, and deeper insights emerge. This perspective fundamentally reframes synthetic data generation from ad-hoc scripting into a rigorous scientific pipeline capable of supporting robustness testing, multidimensional experimentation, and adaptive knowledge discovery.

10. References

1. Jordon, J., et al. Synthetic Data: What, Why and How? Royal Society survey.
2. Kaabachi, B., et al. Scoping review of privacy and utility metrics in synthetic health data.
3. Adams, T., et al. Fidelity, privacy and utility trade-off in synthetic data.
4. Jordon, J., Yoon, J., & van der Schaar, M. (2019). *PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees*. International Conference on Learning Representations (ICLR).
5. Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). *Modeling Tabular Data using Conditional GAN*. Advances in Neural Information Processing Systems (NeurIPS).
6. Patki, N., Wedge, R., & Veeramachaneni, K. (2016). *The Synthetic Data Vault*. IEEE International Conference on Data Science and Advanced Analytics (DSAA). [Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., & Saleem, K. (2020). *Generation and Evaluation of Synthetic Patient Data*. BMC Medical Research Methodology, 20(108).
7. El Emam, K., Mosquera, L., & Hoptroff, R. (2020). *Practical Synthetic Data Generation*. O'Reilly Media.

8. Yoon, J., Drumright, L. N., & van der Schaar, M. (2020). *Anonymization Through Data Synthesis Using GANs*. IEEE Journal of Biomedical and Health Informatics