| 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.