Imperial College London

Department of Earth Science and Engineering

MSc in Environmental Data Science and Machine Learning

Independent Research Project

Final Report

Copula-VAE Hybrid Synthesis for Small-Sample Geoscience Data

By

Xiangxin Zhao

Email: xz2821@ic.ac.uk

GitHub username: esemsc-xz2821

Repository: https://github.com/ese-ada-lovelace-2024/irp-xz2821

Supervisors:

Dr. Paulo Lopes

Dr. Pablo Brito Parada

August 2025

AI Acknowledgement Statement

Tool name: Chat GPT-o3

Provider: Open AI

URL: https://chatgpt.com/share/68951c0c-8528-8003-b92f-c384372a0d80

Brief: Used to study concepts and theory.

Statement: I used OpenAI's ChatGPT-o3 (https://chat.openai.com/) to study mathematical concepts of Copula and Tabular Variational Autoencoder (TVAE). This generative AI tool supported my learning process, but the submitted work is my own and it reflects my own understanding and effort as I declared by signing the Academic Integrity Declaration.

Table of Content

Abstract

We study tabular data generation for blasting–fragmentation in a small-sample and complex setting. Eight public datasets were unified and selected to 12 features; three sets fully matched and were merged (262 rows). We compare Gaussian, Student-t and Clayton copulas with Tabular Variational Autoencoder (TVAE) variants (Vanilla, annealed TVAE, TVAE with a Gaussian Mixture Model prior). To avoid leakage, marginals and scalers are estimated on training data only, with random splits. Meanwhile, Copulas are assessed by Akaike Information Criterion (AIC) per observation and log-likelihood per observation. Generative fidelity and utility are measured by maximum mean discrepancy (MMD), Fréchet Autoencoder Distance (FAED) and Train-on-Synthetic-Test-on-Real (TSTR)/Train-on-Real-Test-on-Synthetic (TRTS) for predicting the Fragment Index with XGBoost.

Despite appearing advantages on the full dataset, the Gaussian copula degrades sharply when trained on 80% data, while Clayton remains stable. Furthermore, neither copula-only nor copula-mixed training improves downstream performance with clear evidence. TVAE-GMM attains the best results (mean $R^2 \approx 0.736$, RMSE ≈ 1.698, MMD ≈ 0.160, FAED ≈ 0.436), outperforming Vanilla and annealed TVAE ($R^2 \approx 0.646$, RMSE≈1.967, MMD≈0.164). In this 12-dimensional, n=262 scenario, simple copulas struggle to capture higher-order dependence, whereas a GMM prior enables VAE to model data distribution reasonably in our circumstance. We release a reproducible evaluation process and discuss implications for small-sample geoscience modelling.

1.Introduction

Blasting-induced fragmentation affects downstream crushing efficiency, energy consumption, and plant throughput in mining. Reliable modelling of fragment-size distributions therefore has direct practical value for industrial design and operation. [1] We study tabular data generation to characterise and augment the small and heterogeneous datasets used for predicting tasks such as the Fragment Index prediction.

Publicly available data in this area are limited and inconsistent. We combined eight sources into a common schema of twelve features by unifying names and units and leaving unconvertible fields missing. Three sources fully matched the schema and were merged to a table with 262 rows. The original datasets show cross-site heterogeneity, unit and definition mismatches, missing values, and heavy-tailed marginals. Under these conditions, high-dimensional dependence is hard to learn, and direct evaluation can be misleading if information leaks across splits.

Two families of generators are often used for tabular data. Copulas separate marginals from dependence and offer interpretability, yet high-parameter copulas can be brittle with small samples and may not capture tail dependence. Deep tabular generators such as the Tabular Variational Autoencoder aim to learn higher dimensional data structure. In fragmentation prediction, previous studies mainly report high in-sample accuracy on single-site, small sample size datasets—using ensembles or neural networks—but rarely evaluate cross-site generalisation or task-based utility beyond marginal checks [2,3,4,5]. Similar small-sample patterns appear in related blasting analytics (e.g., vibration modelling), reinforcing the need for leak-resistant and generalised evaluation [6]. These prediction papers matter for synthetic data because they define the downstream task and expose the core limitation: high accuracy is achievable in single-site, small-n settings, yet cross-site generalisation and the effect of augmentation are rarely tested. Our study addresses this gap by learning distributional generators and evaluating them through utility tests and fidelity tests.

We adopt a reproducible, leak-resistant protocol. Marginals and scalers are estimated on the training split only, and splits are random. For copulas (Gaussian, Student-t, Clayton), model selection uses AIC per observation and log-likelihood per observation. [7,8] Generative fidelity is measured with Maximum Mean Discrepancy (MMD) using different data combination and with Fréchet Autoencoder Distance (FAED). [9,10] Utility is assessed via TSTR and TRTS for predicting the Fragment Index with XGBoost. We report multi-seed means with standard deviation. We do not conduct explicit memorisation tests in this study and treat this as a limitation.

Our results preview is as follows: Although the Gaussian copula appears competitive when fitted to all records, its performance deteriorates sharply under an 80% training split, while Clayton is more stable. [11] Furthermore, there is no strong evidence that copula or copulaaugmented training improves downstream accuracy. Among neural generators, a TVAE with a Gaussian Mixture Model (GMM) [13] prior consistently achieves the best fidelity and utility (e.g., mean $R^2 \approx 0.736$, RMSE≈1.698, MMD≈0.160, FAED≈0.436), outperforming vanilla and annealed TVAE.

This paper contributes: (i) a harmonised small-n baseline for blasting–fragmentation with twelve features and n = 262; (ii) a leak-resistant, comparable evaluation protocol spanning likelihood-based, distributional, and task-based criteria; and (iii) evidence that a TVAE with GMM prior offers superior fidelity and utility in this 12-dimensional, multi-source setting. Our claims are scoped to this dataset and protocol; broader generalisation will require validation on additional sites and alternative dependence models (e.g., vine or regularised Gaussian copulas).

2.Data Preprocessing

2.1 Sources and Scope

We merged eight public datasets to a common schema and found three that fully matched the Golden-12 feature set relevant to Fragment Index (FI) modelling. These are:

- hudaverdi_2010_full (n = 110)
- hudaverdi_2012 (n = 62)
- kulatilake_2010 (n = 90)

The three sources were vertically concatenated with a provenance column dataset. A unique row key 'uid', which combined dataset name and blast id, avoids accidental index collisions across sources. The merged table is saved as 'fi_pool.xlsx' (262 rows, 12 columns before trimming).

We use these three sources as the 'FI pool' for all subsequent modelling and evaluation. Other sheets (e.g., aler_1996, trivedi_2014_flyrock, sharma_2017) were cleaned under the same conventions but are outside the FI pool because key variables are missing or incompatible.

2.2 Feature Schema and Harmonisation

We standardised column names to a Golden-12 schema:

site, blast_id, spacing_over_burden (S/B), benchheight_over_burden (H/B), burden_over_diameter (B/D), stemming_over_burden (T/B), powder_factor (kg·m^{-3}), youngs_modulus_gpa, fragment_median_m (X50, m), fragmentation_index (FI), group, mesh_area_m2.

Table 1: Nomenclature of Rock Blasting

Symbol/Name	Definition	Unit
S (Spacing)	Center-to-center spacing between adjacent blastholes in a row	m
B (Burden)	Distance from a blasthole to the free face (normal to face)	m
H (Bench height)	Vertical height of the bench being blasted	m

T (Stemming length)	Length of inert stemming material at hole collar	m
D (Hole diameter)	Drillhole diameter	m
A (Mesh area)	Blast pattern cell area	m^2
S/B	Spacing-to-burden ratio	-
H/B	Bench-height to burden	-
T/B	Stemming to burden	-
B/D	Burden to diameter	-
X50 (fragment_median_m)	Median fragment size (D50)	m
PF (powder factor)	Explosive mass per rock volume	kg·m^{-3}
E (Young's modulus)	Elastic modulus of rock mass	GPa
FI (Fragmentation Index)	Dimensionless fragmentation score used in sources	-
ssB, ss50	Specific surfaces	m^{-1}

Harmonisation rules applied consistently across sheets [12]:

- Unit alignment & derived quantities
    - Where mesh area $A = S \times B$ and the ratio $S/B = r$ were available, we recovered spacing and burden as

$$B = \sqrt{A/r}\,, S = \sqrt{A \cdot r}$$

    - Where specific surface was reported, we used $ssB = 6/B$ and $ss50 = 6/X50$ to compute $B = 6/ssB$, $S = (S/B) \cdot B$, $X50 = 6/ss50$.
- Categorical alignment
    - Where a numeric group label existed (e.g., group ∈ {1,2,3}), we mapped it to text bands by FI: FI<4, 4≤ $FI \leq 8$, FI>8.
- Placeholders
    - Columns not reported by a source were left as NA (we do not impute globally).

For modelling on the FI pool, we dropped obvious meta/redundant columns (site, blast_id, mesh_area_m2), retaining the six core predictors with full coverage: FI (target), X50, PF, H/B, B/D, T/B.

2.3 Quality control, Missingness, and Plausibility Checks

For the six core variables across the three FI sources (fragmentation_index, fragment_median_m, powder_factor, H/B, B/D, T/B), missingness is 0% in 'hudaverdi_2010_full', 'hudaverdi_2012', and 'kulatilake_2010'.

Given this complete coverage we do not perform global imputation. Variables outside the core set (e.g., youngs_modulus_gpa) may be absent in some sources and are excluded from the FI pool baseline. We verified unit conversions and back-calculations by construction and performed spot checks for impossible values.

2.4 Source Heterogeneity (Descriptive Statistics)

Per-source summaries highlight distributional and scale differences:

- Fragmentation Index (FI): medians 3.30 (hudaverdi_2010_full), 5.11 (hudaverdi_2012), 3.59 (kulatilake_2010); means: 4.40, 5.41, 4.73.
- Fragment median X50 (m): means/medians 0.299/0.230, 0.205/0.170, 0.316/0.245 for the three sources respectively (right-tailed; skew≈1.21–2.57).
- Powder factor (kg·m^{-3}): means 0.539, 0.599, 0.537 with consistent right skew (skew≈1.16–1.39).
- Design ratios: B/D centers are similar (means ≈ 26.7–27.4); H/B is higher in hudaverdi_2012 (mean 3.76 vs 3.34/3.46); T/B shows pronounced right tails across sources (skew≈2.62–2.72).

These shifts indicate cross-source heterogeneity in both marginals (location/shape) and dependence.

2.5 Splitting Strategy and Leakage Control

Subsequent modelling uses random splits so that training/test partitions can simulate the real-world condition of data scarcity problem. All scalers and any distributional estimates are fit on training data only. We will talk more about this section in Copula section.

2.6 Normalisation and encoding

All features used in modelling are numeric and on physical or ratio scales. We standardise predictors using training-only means/standard deviations when required by a method. No target transformation is applied to FI. We do not perform winsorisation or global outlier clipping; extreme but physically plausible ratios (notably T/B) are retained and handled by the models.

2.7 Figures and Interpretation

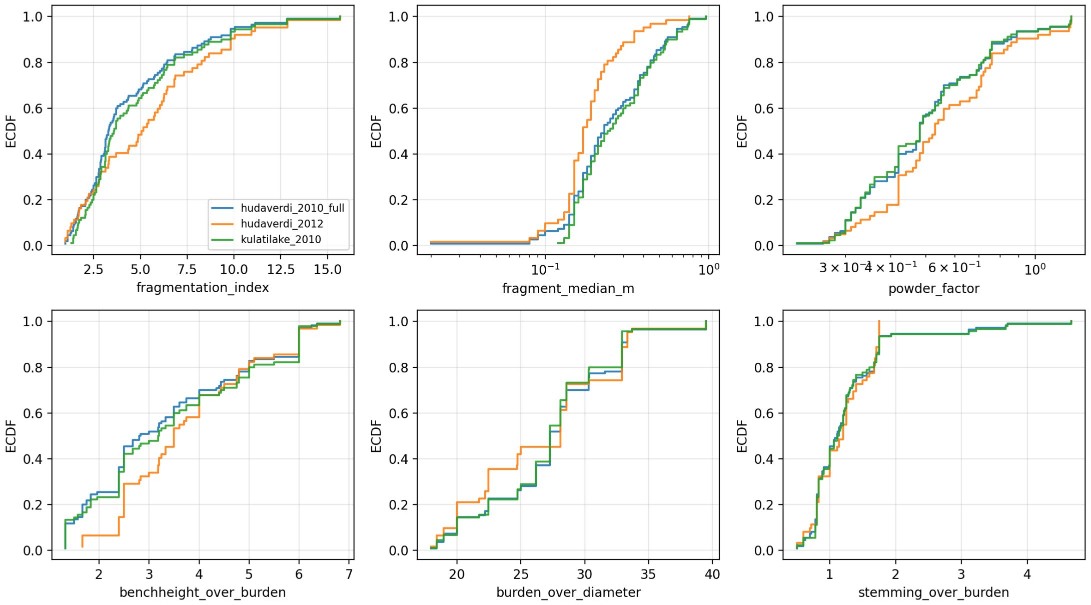

Figure 1. ECDFs by source, six panels for six core parameters

We visualise per-source Empirical Cumulative Distribution Functions (ECDF) of FI, X50, PF, H/B, B/D, and T/B (log-x for X50 and PF). ECDFs are bin-free step estimators of the underlying CDF and converge uniformly to the true distribution as sample size grows, making them suitable for small-sample, heavy-tailed comparisons. In our data, hudaverdi_2012 shows a higher FI level and smaller X50, while PF is right-skewed and higher on average than in the other sources. These patterns substantiate the presence of distribution shift across sites.

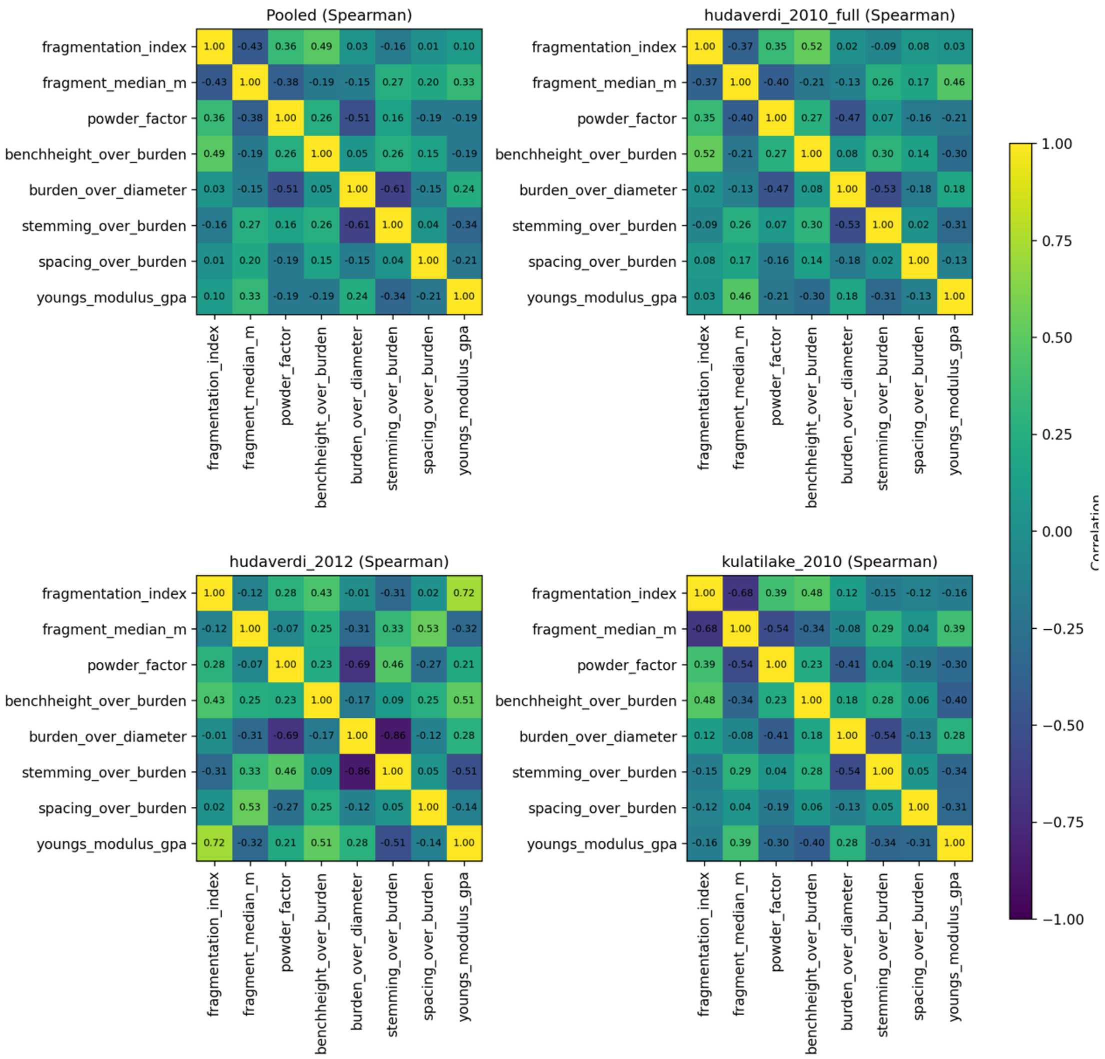

Figure 2. Blasting parameters correlation heatmaps: pooled vs. per-source

We report Spearman correlations for the core variables. Pooled correlations differ from source-specific ones (e.g., pairs involving PF, X50, and FI), demonstrating that dependence structure varies by source. This variability foreshadows why high-parameter dependence models may become unstable when trained on fewer records, and motivates flexible generators (e.g., mixture-prior VAEs) in later sections.

This chapter documents a repeatable path from heterogeneous public sheets to a clean, leak-resistant baseline of n=262. It provides quantitative evidence of cross-source distributional shift and variable dependence by source, which directly motivates the evaluation and model choices in Methods and Results.

3.Methodology

3.1 Problem Setup

We study tabular data generation for blasting–fragmentation under a small-sample, heterogeneous setting. The FI pool introduced in data preprocessing step comprises a dataset with 262 rows drawn from three sources (hudaverdi_2010_full, hudaverdi_2012, kulatilake_2010). The target is the Fragmentation Index (FI). Core predictors' inputs with complete coverage are fragment median size X50 (m), powder factor (PF, kg·m^{-3}), and four design ratios (H/B, B/D, T/B, S/B). All variables are continuous and reported in consistent units after harmonisation.

Our goal is twofold: (i) to model multivariate dependence and generate synthetic data that match the training distribution (fidelity), and (ii) to assess whether synthetic data can support a downstream task (utility) without harming generalisation. We therefore evaluate along four axes:

- Copula model selection via AIC per observation and log-likelihood per observation on train dataset.
- Distributional fidelity via MMD and FAED computed against the training distribution.
- Task utility via TSTR/TRTS protocols using XGBoost to predict FI.
- Robustness is assessed by seeded resampling of a fixed checkpoint on a fixed train/test split: for each seed we resample synthetic data and retrain the downstream regressor, then summarise metrics as mean and standard deviation across N_{seed}.

Unless stated otherwise, all transformations, marginals, and model fits are estimated on the training split only. The held-out test split is never used for selection or tuning.

3.2 Splits & Leakage Control

To respect cross-site heterogeneity and avoid optimistic estimates from information leakage, we adopt train/test random splits:

- Hold-out protocol: For all evaluation run, we perform an 80/20 random split with shuffling to form the global training and test sets. Hence, we induce unbalanced group distribution to imitate the real-world data scarcity and heterogeneity.
- Train-only fitting and selection: All preprocessing statistics, marginal transforms, model fits and hyperparameter choices are estimated on the training split only. In particular:

- Copulas: We report AIC per observation [7] on train set. For generalisation, we compute log-likelihood per observation by applying the train-fitted marginals and copula parameters to the train set and evaluating the copula density. No quantities are re-estimated on test.
- Generators: TVAE variants are trained only on train. When synthetic data are used (TSTR), the downstream regressor is trained on synthetic and evaluated on the real test set.

- Inner validation: Early stopping and hyperparameter search, when used, operate on a validation fold which is inside the training split.
- Feature visibility: The provenance label dataset is used solely for splitting, diagnostics and analysis. It is excluded from model inputs.
- Reporting: We use a random 80/20 train/test split. For robustness to sampling randomness, we fix the trained weights of each generator and repeat synthetic sampling N_{seed} times with different seeds; the downstream XGBoost is retrained for each seed. We report mean and standard deviation across seeds for each metric (R^2, RMSE, MMD).

This protocol enforces a clean separation between training-visible information and held-out evaluation, while explicitly testing robustness to cross-source distributional differences.

3.3 Preprocessing

We model the FI pool with Fragmentation Index as the target and five fully covered predictors. Column names and units are harmonised in previous step. Also, the six core variables have 0% missingness across the three sources, so no global imputation is performed.

For neural generators (TVAE variants), continuous columns are z-scored using training-set statistics only (per-column value computed on train, then reused for inverse-transforming synthetic samples to physical units). Tree-based downstream models (XGBoost) are trained on the original physical scales.

3.4 Copula Models

Copulas are functions that enable us to separate the marginal distributions from the dependency structure of a given multivariate distribution. The core theorem behind Copula is Sklar's theorem, which was introduced by Abe Sklar in 1959 [11]:

Consider a d-dimensional CDF, F, with marginals$F_1, \dots, F_d$. Then there exists a copula, C, such that

$$F(x_1, \dots, x_d) = C\big(F_1(x_1), \dots, F_d(x_d)\big) \quad (1)$$

For all $x_i \in [-\infty, \infty]$ and $i = 1, \dots, d$.

Basing on Sklar's theorem, we can model the margins and dependence separately via Copula C. We fit three multivariate copulas that separate marginals from dependence: Gaussian**,** Student-t (degrees of freedom $\upsilon \in 3,4,6,8,12,20$ selected by log-likelihood), and the Clayton Archimedean copula (lower-tail dependence). Models are trained on the joint of FI and the five numeric predictors. All metadata are excluded.

It is also important to clarify the concept of Marginal transform (Probability Integral Transform): Let $X \in \mathbb{R}^{n \times d}$ be the training matrix with columns X_j. We fit 'QuantileTransformer' on the training split to approximate each marginal CDF $\widehat{F_j}$ and map features to pseudo-observations

$$U_j = \widehat{F_j}(X) \in (0,1).$$

For numerical stability, we clip U to $[\varepsilon, 1 - \varepsilon]$ with $\varepsilon = 10^{-3}$.

Let's introduce more details of how Copula achieves the dependence estimation:

For Gaussian / Student-t Copula, the important step is calculating the Kendall's τ. It is a metric to measure the ordinal correlation of the dataset. From the (train) pseudo-observations, we compute Kendall's τ matrix $\hat{\tau}$ (average-rank convention) and map it to a Gaussian-copula correlation [14]:

$$\hat{\rho}_{jk} = sin\left(\frac{\pi}{2} \cdot \hat{\tau}_{jk}\right) \quad (2)$$

Hence, the correlation matrix is calculated from Kendall's τ holding the copula-invariant feature of τ.

To stabilise small-sample estimates, we apply shrinkage toward identity:

$$\rho \leftarrow (1 - \lambda)\rho + \lambda I, I = 0.05 \quad (3)$$

and, if needed, a positive-definite repair by flooring eigenvalues (10^{-6}) when the Cholesky factorisation [15] fails, which is used to check the positive definite property. For Student-t, we use the same ρ and a fixed $\nu = 8$. And for Clayton, we estimate the single parameter $\theta > 0$ by maximum likelihood on the same uniform distribution.

Talk about the Copula selection, for Gaussian Copula as instance, we write $z = \Phi^{-1}(u)$ for the probit transform. The Gaussian-copula log-density for one row is

$$logc_{\rho}(u) = -\frac{1}{2}log|\rho| - \frac{1}{2}\mathrm{z}^T(\rho^{-1} - I)z. \ (4)$$

(Analogous closed forms are used for Student-t and Clayton.)

We decide to use AIC for model selection. Akaike Information Criterion (AIC), known as AIC, is a statistical method for measuring the quality of a model for the same data set, with lower AIC scores indicating better quality models. Hence, we can report AIC per observation by[16]:

$$\frac{AIC}{n_{train}} = \frac{2k - 2log\,\hat{L}_{train}}{n_{train}} \ (5)$$

with k the number of copula parameters (Gaussian: $k = \frac{d(d-1)}{2}$; Student-t: $k + 1$; Clayton: $k = 1$). We also report $log\,\hat{L}_{train} = \frac{log\,\hat{L}_{train}}{n_{train}}$. Due to time limitation, out-of-sample log-likelihood is not computed to protect test dataset from the leakage of downstream test. Generalisation is assessed later via fidelity (MMD/FAED) and task utility (TSTR/TRTS).

Finally, we can sample by inverse marginal transform. From the fitted copula, we draw m pseudo-observations $u^{(1)}, \dots, u^{(m)}$ and map back to physical space using the inverse of the train-fitted marginals:

$$x_j^{(s)} = \hat{F}_j^{-1}\left(u_j^{(s)}\right), s = 1, \dots, m; j = 1, \dots, d \ (6)$$

These synthetic tables are used for the Copula-only and Mixed training regimes in downstream experiments. Hence, there is no test information used in fitting or sampling.

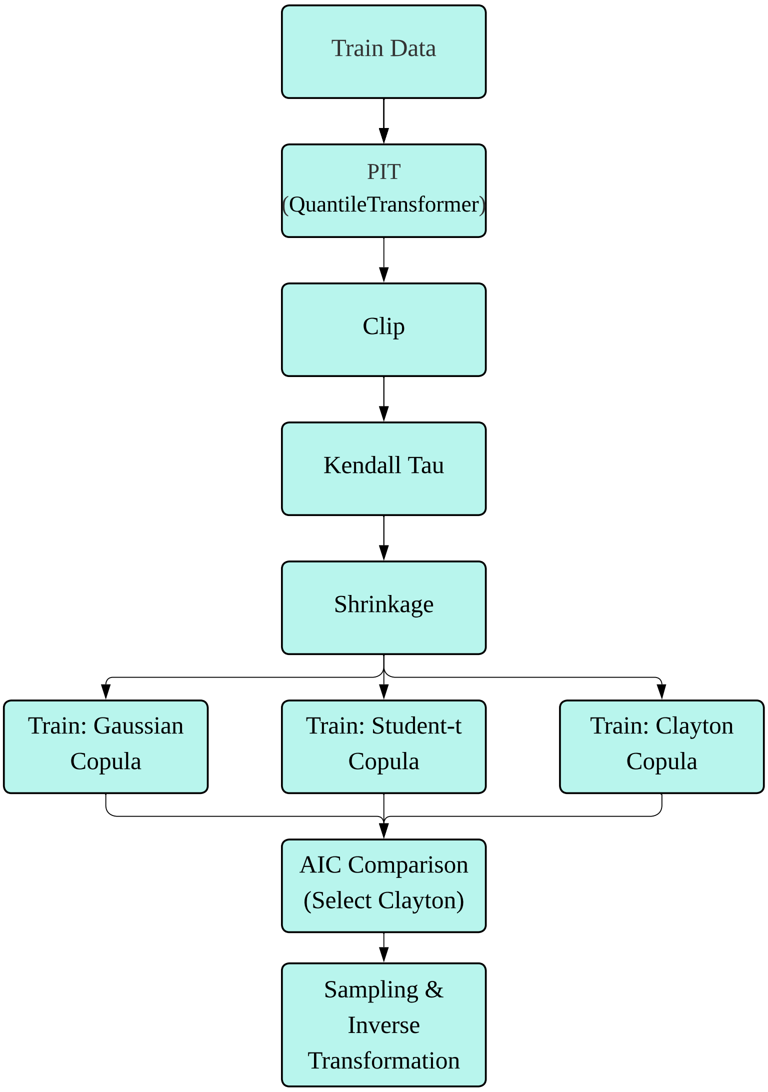

Figure 3. The complete process of copula

Rationale: PIT removes marginal shapes while holding the dependence. Hence, the dependence relation is learned on a common interval [0, 1] scale. Also, the clip prevents the inverse normal distribution $\phi^{-1}(u)$ from numerically blowing up. The Kendall's correlation & Pearson's correlation mapping generates a rank-based, heavy-tail-robust estimator. The shrinkage and Positive-Definite repairs ensure correlation matrix ρ is well-conditioned so that $log|\rho|$, ρ^{-1} and quadratic forms are stable.

3.5 TVAE Variants

We model high-order dependence in a small-sample, 12-dimensional blasting–fragmentation table to generate realistic, task-useful synthetic data. Three generators are studied in increasing expressive power: Vanilla TVAE (standard normal prior) [17,18], Annealing TVAE (Kullback-Leibler weight annealing to mitigate posterior collapse) [19], and TVAE-GMM (latent Gaussian-mixture prior to capture multi-modal structure) [20]. This section specifies representations, objectives, training, and sampling. All evaluation metrics (utility, fidelity, robustness) are defined in later section.

Data representation and preprocessing:

- Columns: Only numeric predictors and the target (Fragmentation Index, FI) are used.
- Scaling: Continuous columns are z-scored using statistics fitted on the training split only. Synthetic samples are inverse transformed back to physical units before downstream use.
- Training sources:
  - Real-only: train on the real training split.
  - Copula-only: train on samples drawn from the best copula model.
  - Mixed: train on a convex mix of real and copula samples with ratio α (e.g., 1:1, 1:2, 1:4, 1:8).

All models in the section are variants of Variational Autoencoder (VAE). A Variational Autoencoder (VAE) posits a latent-variable generative model with prior $p(z)$ and decoder $p_\theta(x|z)$. [21] Given a z-scored record $x \in \mathbb{R}^d$ (standardised with training data only), an encoder defines an approximate posterior $q_\phi(z|x)$ and a decoder models the conditional distribution of x given the latent $z \in \mathbb{R}^\ell$. Exact inference is intractable, so training maximises the evidence lower bound (ELBO):

$$log\, p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)}[log\, p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p(z)) \quad (7)$$

summed/averaged over a minibatch. Gradients flow through the reparameterisation $z = \mu_\phi(x) + \sigma_\phi(x) \odot \varepsilon$ with $\varepsilon \sim N(o, I)$.

TVAE is tabular VAE focusing on tabular data generation. Classical VAEs target images/audio with convolutional decoders and domain-specific likelihoods. Tabular data differ: features have heterogeneous scales, interactions are not spatially local, and column-wise interpretability matters. The TVAE adapts the VAE to this structure via three choices:

1. Column-wise likelihoods in a standardised space. [18] After z-scoring (train-only), the decoder uses a factorised Gaussian likelihood per column. Conditional independence holds given z. The cross-feature dependence is captured through the shared latent rather than an

explicit covariance.

2. Latent priors matched to structure. A unimodal standard normal prior is simple but may underfit heterogeneous, multi-regime tables. A Gaussian-mixture prior provides multi-modal capacity while keeping the decoder simple and numerically stable. [20]
3. Optimisation safeguards. Small-n training is prone to posterior collapse. A β-VAE objective with annealing lets the model first learn faithful reconstructions, then tighten the latent to the prior. [19]

After training, samples are decoded in the standardised space and then inverse z-scored back to physical units for downstream use. The next subsection formalises these model classes and objectives.

We learn a low-dimensional latent representation that captures higher-order dependence in the standardised tabular space. All losses are computed in this space. Synthetic samples are later mapped back to physical units via the inverse z-score.

Let's declare the notation first: $x \in \mathbb{R}^d: z - scored\ data;$ $z \in \mathbb{R}^\ell: latent;$ $\ell: latent\ dimension;$ $\ell: latent\ dimension;$ $\theta, \phi: decoder/encoder\ parameters;$ $\beta_t: KL\ weight\ at\ step\ t;$ $K: number\ of\ mixture\ components.$

The encoder outputs a diagonal-Gaussian posterior:

$$q_\Phi(z|x) = \mathcal{N}\left(\mu_\phi(x), diag\ \sigma_\phi^2(x)\right) \quad (8)$$

The decoder models continuous columns with a factorised Gaussian likelihood:

$$p_\Theta(x|z) = \prod_{j=1}^d \mathcal{N}\left(x_j; \hat{\mu}_{\theta,j}(x), \widehat{\sigma}_{\theta,j}^2(z)\right) \quad (9)$$

In practice we predict $log\ \hat{\sigma}^2$ and pass it through softplus\lower bounds to keep variances positive and numerically stable.

For a minibatch $\{x^{(i)}\}_{i=1}^B$, we maximise ELBO with β annealing as:

$$\mathcal{L}_\beta(\theta, \phi) = \frac{1}{B}\sum_{i=1}^B \mathbb{E}_{q_\phi\left(z|x^{(i)}\right)}[log\ p_\theta(x^{(i)}|z)] - \beta \cdot D_{KL}(q_\phi(z|x^{(i)})||p(z)) \quad (10)$$

The three variants differ only in the latent prior $p(z)$ and the schedule for β:

(a) Vanilla TVAE (standard normal prior)

$$p(z) = \mathcal{N}(0, I) \quad (11)$$

The KL admits a closed form:

$$D_{KL}(\mathcal{N}(\mu, diag\ \sigma^2)||\mathcal{N}(0, I)) = \frac{1}{2}\sum_{j=1}^\ell\left(\mu_j^2 + \sigma_j^2 - log\sigma_j^2 - 1\right) \quad (12)$$

(b) Annealed TVAE

Same prior $p(z) = \mathcal{N}(0, I)$, but β increases monotonically to mitigate early posterior collapse:

- Linear warm-up with horizon T_w and cap β_{max}:

$$\beta_t = min(\beta_{max}, \beta_{max} \cdot t/T_w)$$

- Sigmoid schedule with centre t_0 and slope κ:

$$\beta_t = \frac{\beta_{max}}{1 + exp\big(-\kappa(t - t_0)\big)}$$

(c) TVAE-GMM (Gaussian-mixture prior)

$$p(z) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(z;\ \mu_k, \Sigma_k), \sum_k \pi_k = 1, \Sigma_k \ is\ diagonal. \ (13)$$

Mixture parameters $\{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}$ are learned jointly with θ, ϕ by back-propagation. The KL is estimated by Monte-Carlo:

$$D_{KL}(q_\phi(z|x)||p(z)) = \mathbb{E}_{q_\phi}[log\ q_\Phi(z|x) - log\ p(z)] \ (14)$$

With a log-sum-up implementation for

$$log\ p(z) = log \sum_{k=1}^{K} \pi_k \, \mathcal{N}(z; \mu_k, \Sigma_k) \ (15)$$

ensuring numerical stability.

After training, we draw $z{\sim}p(z)$ (standard normal for Vanilla/Annealed; mixture sampling for GMM), decode to obtain x' in the standardised space, and apply the inverse z-score to recover physical units and the original column schema. In the z-scored space, per-column Gaussian outputs generate a well-scaled, stable objective for small-sample tabular learning. Importantly, correlations are not discarded: they are represented through the shared latent z rather than explicit observation covariances. This keeps optimisation simple and robust while allowing the latent prior—especially the GMM prior—to encode multi-modal structure.

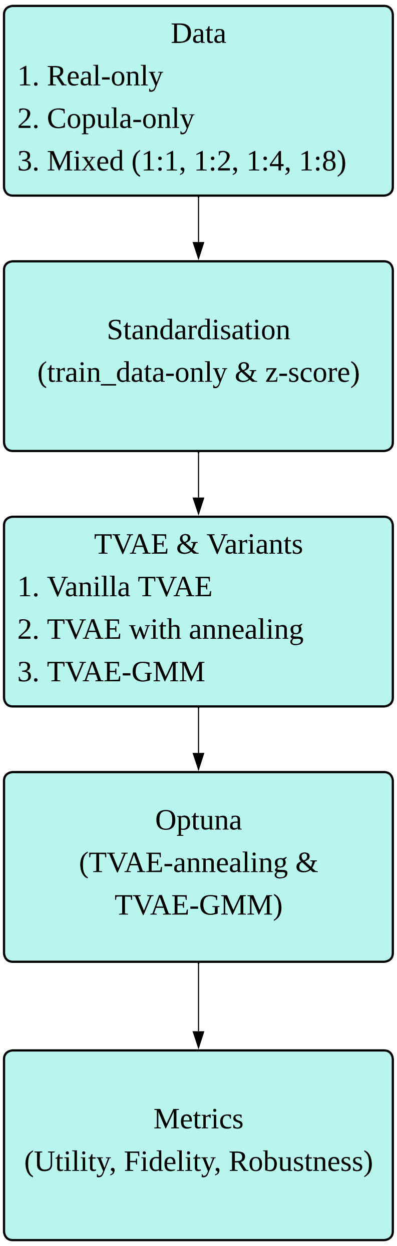

Figure 4. TVAE Process

3.6 Evaluation & Metrics

We evaluate generators along three orthogonal axes:

- **Utility** (task performance): does training a predictor on synthetic data transfer to the real held-out set?
- **Fidelity** (distributional similarity): how close are synthetic and real distributions?
- **Robustness** (stability): how sensitive are metrics to sampling/learner randomness?

All evaluation uses the pre-saved split: `train_set.csv` and `test_set.csv`. Hyper-parameter selection is confined to the training split via a train-internal validation. The test split is never used for tuning. For each TVAE variant, we assess three types of training data as previous statement: real-only, copula-only (Clayton) and mixed data (e.g., 1:1, 1:2, 1:4, 1:8 for real: copula). For the final check of robustness, we run different seeds. For seed s, we set seed value first, then sample a synthetic table to fit the downstream model. After computing all metrics, we report mean with SD across seeds.

To check the utility of synthetic data, we decide to use TSTR and TRTS [22]. Let $\left(X_{train}^{real}, y_{train}^{real}\right)$ and $\left(X_{test}^{real}, y_{test}^{real}\right)$ refer to the real train/test splits. Let $\left(X_{syn}, y_{syn}\right)$ be a synthetic table sampled from a generator trained under the previous structure. For TSTR, we fit a fixed-hyperparameter XGBoost regressor on $\left(X_{syn}, y_{syn}\right)$ and evaluate on $\left(X_{train}^{real}, y_{train}^{real}\right)$. We do the same thing with exchanged data for TRTS. The metrics of utility check are R^2 and RMSE which define as:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y})^2}{\sum_i (y_i - \bar{y})^2} \quad (16)$$

and

$$RMSE = \sqrt{\frac{1}{n} \sum_i (y_i - \hat{y})^2} \quad (17)$$

The R^2 measures how well the model explain the variance of the data, and RMSE display the accuracy.

For the fidelity part, we decide to check Max Mean Discrepancy (MMD) and Fréchet Autoencoder Distance (FAED). Given samples $X = \{x_i\}_{i=1}^{n}$ (real) and $Y = \{y_j\}_{j=1}^{m}$ (synthetic), the squared MMD with kernel k is

$$MMD^2(X, Y) = \frac{1}{n(n-1)} \sum_{i \neq i'} k(x_i, x_{i'}) + \frac{1}{m(m-1)} \sum_{j \neq j'} k((y_j, y_{j'}) - \frac{2}{nm} \sum_{i,j} k\left(x_i, y_j\right) \quad (18)$$

We use RBF multi kernel mixture $k(\cdot,\cdot) = \sum_r exp(-|| \cdot - \cdot ||^2/2\sigma_r^2)$ with bandwidths σ_r set by the median heuristic. RBF kernel is characteristic which mean MMD (P, D) = 0 if and only if P = D. Hence, the MMD can show the real "distance". [10]

The idea behind FAED is widely used for assessing image generation. In our case, we train an autoencoder on real-train data to obtain an embedding $\phi(\cdot)$. Let (μ_r, Σ_r) (real) and (μ_s, Σ_s) (synthetic) be the empirical mean and covariance of $\phi(X_{real})$ and $\phi\left(Y_{syn}\right)$. Define:

$$FAED = \left|\left|\mu_r - \mu_s\right|\right|_2^2 + Tr\left(\Sigma_r + \Sigma_s - 2\left(\Sigma_r^{1/2}\Sigma_s\Sigma_r^{1/2}\right)^{1/2}\right) \quad (19)$$

Lower FAED is better. Hence, MMD compare the original data and synthetic data in original space and kernel space while FAED check the second-order moment similarity.

To verify the robustness of metrics results, we collect S runs' consequences (different seeds). For each metric, we report $M \in R^2, RMSE, MMD, FAED$ such that:

$$\bar{M} = \frac{1}{S}\sum_{s=1}^{S} M^{(s)}, SD(M) = \sqrt{\frac{1}{S-1}\sum_{s=1}^{S}(M^{(s)} - \bar{M})^2} \quad (20)$$

Hence, we can display relatively fair and reproducible results.

4.Results

We report results on the FI pool (n=262) formed from three sources (110/62/90 rows). The six core variables (FI, X50, PF, H/B, B/D, T/B) have 0% missingness. An 80/20 split is held fixed across experiments. Evaluation follows a leak-resistant protocol: all transforms are train-only; XGBoost uses fixed hyper-parameters; metrics are aggregated as means across seeds. We assess utility with TSTR (TRTS where noted) and fidelity with MMD and FAED. §4.1 reports train AIC per observation for copulas (selection only). §§4.2–4.4 present TVAE results for Vanilla, Annealed, and GMM under Real-only/Copula-only/Mixed training.

4.1 Copula Fitting

Train AIC per observation (lower is better):

- Clayton 2.56 (k=1), Student-t ($\upsilon = 8$) 6.44, Gaussian 8.19 (k=28).

Log likelihood per observation:

- Clayton: −1.276, Student-t: −3.080, Gaussian: −3.962.

With PIT (train-only), ε-clipping (10^{-3}), Kendall's τ and transferred Pearson's correlation ρ, 5% shrinkage and PD repair, Clayton is the most stable with best fitting results on train dataset. The Gaussian copula's larger parameter count, and tail-independence make it fragile at small-n. The performance of Gaussian copula degraded remarkably when the input becomes 80% of original data (train dataset). The sensitivity of data size is a problem of copula. Considering to stability and performance, we decide to use Clayton copula.

4.2 TVAE Baseline

TSTR (real-only): $R^2 \approx 0.434$, RMSE≈2.489

TSTR(Copula-only): $R^2 \approx -0.613 < 0$, RMSE≈4.203

TSTR(Mixed): Utility drops roughly as the copula share increases:

$R^2 \approx 0.289/0.385/0.253/-0.359$ for 1:1/1:2/1:4/1:8

RMSE≈2.791/2.596/2.859/3.857

At row number n=262, column number d=12, a plain TVAE underfits higher-order structure; combining copula samples (which encode mostly low-order monotone dependence) dilutes the signal the VAE needs, so downstream transfer worsens.

4.3 Annealing TVAE

TSTR (real-only): $R^2 \approx 0.646$, RMSE≈1.967

Fidelity: MMD≈0.164

Annealing reduces early KL pressure, improves reconstructions and latent usage, and substantially closes the gap to a usable synthetic surrogate relative to Vanilla TVAE.

4.4 TVAE-GMM

TSTR (real-only): $R^2 \approx 0.736$, RMSE≈1.698

Fidelity: MMD≈0.160, FAED≈0.436

A mixture prior better matches multi-regime behaviour across sites. With the same fatorised decoder, the latent space captures heterogeneity more cleanly. TVAE-GMM is the best overall among VAEs both utility and fidelity in this setting.

Analysis

Under real-only training, TVAE-GMM and Annealed TVAE achieve similar MMD (about 0.16). When FAED is considered, GMM gives the best overall fidelity. FAED has no absolute threshold, but lower is better, and values closer to 0 indicate closer alignment with real data. Qualitatively speaking, TVAE-GMM tracks the real distribution most closely.

All metrics are means across seeds with the same XGBoost settings. Across seeds, GMM maintains the highest mean R^2 and the lowest RMSE among the VAE variants. The dispersion is moderate and consistent with small-n variability. Starting from an 80% train split (around 210 rows), TVAE-GMM can generate 5,000 synthetic records while holding $R^2 \approx 0.7$ on TSTR with high fidelity.

Models trained on copula-only data, or mixed inputs, show negative or degrading TSTR results. This reflects a limitation of copula-based sampling in our setting: it does not capture the higher-order dependencies needed for generating high-utility synthetic data. As the copula fraction increases, R^2 falls and RMSE rises, indicating that copula mixing biases the generator away from the structures required for TSTR.

Overall, for small-sample, cross-site blasting–fragmentation tables, use TVAE-GMM (real-only training) as the default generator. Choose Annealed TVAE when a simpler structure prior is preferred.

5.Discussion

Clayton copula on train outperformed Gaussian/Student-t in AIC/obs. The one-parameter form plus lower-tail dependence is stable at small-n and matches the right-tailed marginals we observe. In contrast, Gaussian relies on a full correlation matrix and is tail-independent, which makes it fragile when n is limited.

TVAE with β-annealing lifted utility and fidelity over the vanilla baseline. Reducing early KL pressure let the model learn reconstructions before regularising the latent, which is critical at small-n. TVAE-GMM was best overall on the real-only recipe (TSTR $R^2 \approx 0.736$, RMSE≈1.698, MMD≈0.160, FAED≈0.436). A multi-modal prior aligns with cross-site heterogeneity; with a simple factorised decoder, the latent absorbs regime differences without over-parameterising observation covariances.

When the sample size is small with moderate column numbers (10-20) and heterogeneous sites, TVAE-GMM will be suitable choice with real-only training. When it comes to tighter computation, annealed TVAE offered a strong utility gain over the vanilla TVAE.

Synthetic data is not a goal but a means. The use case dictates the generator and the metrics:

- Model development/stress-testing: optimise utility (TSTR) under realistic distribution shift. MMD and FAED act as safety checks.
- Distributional mimicry/scenario replay: prioritise fidelity (MMD, FAED, ECDF tails)
- Coverage expansion/rare regimes: bias the sampler to tails or under-represented situations. It is acceptable to have utility trade-offs.

However, there is a problem: to generate the “best” synthetic data, the model must already capture the core information in the real data. If it does, the marginal benefit of synthesis for accuracy may be small. Where synthesis does add value is in generality—introducing controlled diversity and noise, creating stress scenarios, and enabling repeated experimentation without overfitting to the limited train split. Expecting a model trained on a tiny dataset to learn beyond that dataset’s information is unrealistic, and this is where transfer learning and LLM become very potential ideas.

The project has limitations:

- Small-n sample, numeric-only features and missing or sparse properties will affect the stability and generality of results.
- Due to time limitation, we do not add explicit memorisation audit. The privacy and overfitting risk are not quantified.
- For new rock blasting data, we may need to design new data preprocessing pipeline because

of the lack of standardised data collection. Each data has different units and physical quantity.

- More heterogeneous data are needed for stability testing.

6.Conclusion & Future work

On train-only model selection, the Clayton copula has the lowest AIC per observation. Its one-parameter form and lower-tail dependence are stable at small-n; the Gaussian copula's large parameter count, and tail independence make it fragile when n is limited.

Among VAEs, β-annealing improves utility and fidelity over vanilla. TVAE-GMM is best overall on the real-only recipe ($R^2 \approx 0.736$, RMSE$\approx$1.698, MMD$\approx$0.160, FAED$\approx$0.436), aligning a multi-modal prior with cross-site heterogeneity while keeping the decoder simple.

There is no clear evidence to prove that the combination of copula and TVAEs can improve quality of synthetic data. Conversely, in our tests, utility degrades as the copula share increases.

For future work, there are three directions:

- Model extensions:
  There are some potential mature models and new ideas which can be tested.
  - CTGAN/CTAB-GANs [18] for mixed-type tables.
  - Diffusion models for tabular data. [23]
  - Data generation basing on LLM. [24]
- Purpose-aligned generation: conditional sampling and TSTR-aware sampling; tail-focused augmentation; physics-informed constraints (plausible ranges, monotonic relations).
- Learning beyond the small set: transfer learning or pretraining on broader mining/materials data.

In conclusion, we show that for small-n, heterogeneous blasting data, TVAE-GMM (real-only) gives the best synthetic utility and fidelity (TSTR $R^2 \approx 0.736$, RMSE≈1.698; MMD≈0.160, FAED≈0.436), while copula-mixed training degrades performance. β-annealing helps, but a multimodal prior is key. Synthetic data should be purpose-driven (development, stress-testing or distributional replay). In future, we can add test log-likelihood for copulas, memorisation audits, and different data splits, and explore CTGAN/CTAB-GAN, diffusion, and transformer models, plus transfer learning to import structure beyond the tiny dataset.

References

[1] M. P. Roy, R. K. Paswan, M. D. Sarim, S. Kumar, R. R. Jha, and P. K. Singh, “Rock Fragmentation by Blasting—A Review,” *Journal of Mines, Metals & Fuels*, vol. 64, no. 9, pp. 424–431, 2016. [Online].

[2] I. Krop, T. Sasaoka, H. Shimada, and A. Hamanaka, “Optimizing mean fragment size prediction in rock blasting: A synergistic approach combining clustering, hyperparameter tuning, and data augmentation,” Eng, vol. 5, no. 3, pp. 1905–1936, 2024, doi:10.3390/eng5030102.

[3] P. H. S. W. Kulatilake, Q. Wu, T. Hudaverdi, and C. Kuzu, “Mean particle size prediction in rock blast fragmentation using neural networks,” *Engineering Geology*, vol. 114, no. 3–4, pp. 298–311, 2010, doi: 10.1016/j.enggeo.2010.05.008.

[4] S. K. Sharma and P. Rai, “Establishment of blasting design parameters influencing mean fragment size using state-of-art statistical tools and techniques,” *Measurement*, vol. 96, pp. 34–51, 2017, doi: 10.1016/j.measurement.2016.10.047.

[5] Y. Sui, Z. Zhou, R. Zhao, Z. Yang, and Y. Zou, “Open-Pit Bench Blasting Fragmentation Prediction Based on Stacking Integrated Strategy,” *Appl. Sci.*, vol. 15, no. 3, Art. no. 1254, Feb. 2025, doi: 10.3390/app15031254.

[6] S. Hosseini, J. Khatti, B. O. Taiwo, Y. Fissha, K. S. Grover, H. Ikeda, M. Pushkarna, M. Berhanu, and M. Ali, “Assessment of the ground vibration during blasting in mining projects using different computational approaches,” *Scientific Reports*, vol. 13, Art. no. 18582, Oct. 2023, doi: 10.1038/s41598-023-46064-5.

[7] H. Akaike, “Information Theory and an Extension of the Maximum Likelihood Principle,” in *Selected Papers of Hirotugu Akaike*, E. Parzen, K. Tanabe, and G. Kitagawa, Eds. New York, NY, USA: Springer, 1998, pp. 199–213, doi: 10.1007/978-1-4612-1694-0_15.

[8] A. Vehtari, A. Gelman, and J. Gabry, “Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC,” Statistics and Computing, vol. 27, no. 5, pp. 1413–1432, Sep. 2017, doi: 10.1007/s11222-016-9696-4.

[9] D. Herurkar, A. Ali, and A. Dengel, “Evaluating Generative Models for Tabular Data: Novel Metrics and Benchmarking,” arXiv:2504.20900 [cs.LG], Apr. 29, 2025. [Online]. Available: https://arxiv.org/abs/2504.20900

[10] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, “A Kernel Two-Sample Test,” J. Mach. Learn. Res., vol. 13, pp. 723–773, Mar. 2012.

[11] R. B. Nelsen, *An Introduction to Copulas*, 1st ed. New York, NY, USA: Springer, 1999.

[12] United States National Park Service, “Blast Design,” in *National Park Service Handbook for the Storage, Transportation, and Use of Explosives*. Washington, DC, USA: U.S. Dept. of the Interior, National Park Service, 1999, ch. 8, pp. 113–122.

[13] P. A. Apellaniz, J. Parras, and S. Zazo, “An improved tabular data generator with VAE-GMM integration,” in *Proc. 32nd Eur. Signal Process. Conf. (EUSIPCO)*, 2024, doi: 10.23919/EUSIPCO.2024.9

[14] G. A. Fredricks and R. B. Nelsen, “On the relationship between Spearman’s rho and Kendall’s tau for pairs of continuous random variables,” *J. Stat. Plan. Inference*, vol. 137, no. 7, pp. 2143–2150, Jul. 2007, doi: 10.1016/j.jspi.2006.06.045.

[15] “Cholesky Factorization – an overview,” *ScienceDirect Topics (Engineering)*, Elsevier.

[16] H. Bozdogan, “Model selection and Akaike’s Information Criterion (AIC): The general theory and its analytical extensions,” *Psychometrika*, vol. 52, no. 3, pp. 345–370, Sep. 1987, doi: 10.1007/BF02294361.

[17] SDV Developers, “TVAESynthesizer,” *SDV Documentation*, DataCebo, Inc., last updated Aug. 6, 2025.

[18] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, “Modeling Tabular Data using Conditional GAN,” *arXiv* preprint arXiv:1907.00503, Jul. 2019.

[19] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, “Generating Sentences from a Continuous Space,” in *Proc. 20th SIGNLL Conf. Computational Natural Language Learning (CoNLL)*, Berlin, Germany, Aug. 2016, pp. 10–21.

[20] P. A. Apellaniz, J. Parras, and S. Zazo, “An improved tabular data generator with VAE-GMM integration,” in *Proc. 32nd Eur. Signal Process. Conf. (EUSIPCO)*, 2024, doi: 10.23919/EUSIPCO.2024.9

[21] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” *arXiv:1312.6114*, Dec. 2022.

[22] C. Esteban, S. L. Hyland, and G. Rätsch, “Real-valued (medical) time series generation with recurrent conditional GANs,” *arXiv:1706.02633*, Jun. 8, 2017.

[23] J. Zhu, “Synthetic data generation by diffusion models,” *National Science Review*, vol. 11, no. 8, Art. no. nwae276, Aug. 2024, doi: 10.1093/nsr/nwae276.

[24] L. Long, R. Wang, R. Xiao, J. Zhao, X. Ding, G. Chen, and H. Wang, “On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey,” in *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand, 2024, pp. 11065–11082, doi:

10.18653/v1/2024.findings-acl.658.