// ACADEMIC / 2026 / NOV

Synthetic tabular augmentation for peak particle velocity prediction in rock blasting

Paulo Lopes
https://linktr.ee/pftlopes

Abstract

Synthetic augmentation can alter the bias and variance of a blasting-vibration predictor, but its benefit must be evaluated on independent real observations. This study examines a small peak particle velocity benchmark comparing Gaussian Copula, CTGAN, TVAE and CopulaGAN generators. The underlying database contained 88 records and was reduced to 80 modelling rows; 56 were used for training and 24 for real hold-out testing. The selected Gaussian Copula produced 537 retained synthetic records, giving 593 rows in the hybrid training table. It had the strongest reported marginal and pairwise resemblance, although the combined distance and rank summaries are descriptive rather than inferential tests. On the fixed real test set, hybrid training reduced Random Forest RMSE from 3.44 to 2.79 mm/s and XGBoost RMSE from 3.38 to 2.93 mm/s. These changes correspond to reductions of 18.9% and 13.3%. Real-only Ridge remained the strongest reported test predictor, with RMSE = 2.05 mm/s and R^2 = 0.89. The evidence therefore supports model-specific gains within one split, rather than a universal advantage for augmentation. Interpretation is limited by pre-split cleaning, incomplete documentation of generator refitting within cross-validation and the absence of repeated independent real tests. The discussion connects these limits to charge-distance physics, log-scale retransformation, synthetic-data weighting and upper-tail vibration risk, and defines a validation design that can distinguish useful regularisation from optimistic evaluation.

Keywords: peak particle velocity; blast vibration; synthetic data; Gaussian copula; CTGAN; TVAE; Random Forest; XGBoost; validation

1 Introduction

Peak particle velocity is an established descriptor of blast-induced ground motion. Its observed value depends on explosive charge, source-to-receiver distance, confinement, rock structure and wave propagation. Small field databases make it difficult to distinguish stable physical relationships from sampling effects. Hudaverdi's multivariate analysis illustrates the use of blast-design and vibration variables within an empirical modelling framework [1]. Recent machine-learning studies extend the model family and the range of inputs considered [2]. Their reported scores remain specific to their sites, targets, measurement practices and validation designs.

Synthetic tabular generation offers a way to sample an estimated joint distribution of inputs and response. Copulas provide an explicit dependence representation, while CTGAN and TVAE use neural generation to accommodate heterogeneous column distributions [3,4]. The appeal is practical: a fitted generator can produce many plausible training combinations from few observed blasts. The limitation is equally fundamental: generated targets are inferred from the same real observations and do not supply independent measurements of unobserved geological regimes.

Synthetic-data evaluation should consequently distinguish resemblance, predictive utility, coverage and memorisation [5,6]. A generator can match column histograms while changing the conditional PPV response. It can support a particular learner while worsening another. It can also smooth a sparse empirical distribution in a way that reduces apparent cross-validation variance without improving performance on independent blasts. Leakage and selection bias are especially consequential when the real calibration set contains only dozens of observations [7,8].

2 Materials and methods

2.1 Dataset and provenance

The original database contained 88 literature records and was reduced to 80 modelling rows following checks for missingness, duplicates, unit consistency, collinearity and Tukey-type outliers [9]. The source identified as Hudaverdi (2012) concerns blast-induced ground vibration [1]; it should not be confused with the same author's fragmentation publications. The retained table has eight predictors and one PPV target after removal of scaled distance.

The eight predictors are S/B, H/B, B/D, T/B, U/B, powder factor, maximum charge per delay and distance. Table 2 supplies their definitions and units. The data record should distinguish a physical blast from a monitoring record. Multiple receivers observing the same blast are not necessarily independent examples; a row-wise split can share charge, geology and source conditions between training and testing. The summary does not establish the blast-to-record mapping, so independent-blast generalisation remains untested.

Table 1 retains the original contextual summary of PPV studies attributed to Yan et al. [2]. It is a literature inventory, not a common benchmark. Parameter strings and abbreviated study labels in that source summary include repeated symbols and uncertain counts; these are retained transparently rather than interpreted as a harmonised set of independent predictors. Reported R^2 values cannot be ranked across the rows without matched targets, splits and measurement domains.

Table 1. Historical PPV modelling inventory retained from the experimental manuscript and attributed there to Yan et al. [2]. Study names are abbreviated source labels, not complete citations to each underlying experiment. Repeated symbols and parameter counts follow the inventory; the R^2 values are contextual and are not comparable evidence of model superiority.

Study label	Parameters	Count	ML algorithm	R^2
Qiu	HR, HD, B, S, Q, CL, R, BI, E, PR, Pv, VoD, DoE	13	(WOA-, GWO-, BO-) XGBoost	0.9757
Zhang	Q, R, St, B, H, S, PF, HN	8	PSO-XGBoost	0.968
Nguyen	N, Q, PF, St, B, R, S, H, Δt	9	XGBoost	0.952
Amiri	Q, R	2	KNN; ANN	0.88
Armaghani	Q, R	2	ANN	0.987
Ghoraba	Q, R	2	ANFIS	0.952
Hasanipanah	Q, R	2	CART	0.95
Khandelwal	Q, R	2	SVM	0.96
Khandelwal	Q, R	2	CART	0.92
Arzu	Q, R, Sv, S, B	5	ANFIS	1.00

Study label	Parameters	Count	ML algorithm	R^2
Monjezi	Q, R, S, HD	4	ANN	0.9493
Nguyen	Q, R, S, f, Q, R	5	(SpaSO-, SaISO-, MFO-) ELM	0.99
Monjezi	Q, R, B/S, UCS, RW, HN	6	MLPNN	0.954
Nguyen	B, S, f, PF, Q, R	6	(MRFO-, HGS-, AO-) SONIA	0.896
Yang	B, S, St, S, Q, R	5	FFA-SVR	0.992
Zhou	HD, PF, S, Q, R	5	FS-RF	0.903
Hosseini	HN, B, H, HR, Q, R	6	BH-LSTM	0.996
Fissha	Q, R	2	PSO-DRVM	0.917

Table 2. Blast-design, charge, distance and vibration variables. Scaled distance is shown for physical interpretation but is excluded from the nine-variable generation table. Q/delay denotes charge assigned to a delay interval in kilograms; it is not an explosive mass flow rate.

Name	Parameter	Description	Unit
S/B	Spacing-to-burden ratio	Ratio of spacing between blastholes to burden	-
H/B	Bench-height-to-burden ratio	Ratio of bench height to burden	-
B/D	Burden-to-hole-diameter ratio	Ratio of burden to blasthole diameter	-
T/B	Stemming-to-burden ratio	Ratio of stemming length to burden	-
U/B	Subdrilling-to-burden ratio	Ratio of subdrilling depth to burden	-
PF	Powder factor	Explosive mass per unit rock volume	kg/m^3
Q/delay	Charge weight per delay	Maximum explosive charge detonated per delay interval	kg
d(m)	Blast distance	Distance from blast source to monitoring point	m
Sd(m)	Scaled distance	Distance normalised by charge weight	m/kg^0.5
PPV(mm/s)	Peak particle velocity	Maximum vibration velocity at the monitoring point	mm/s

2.2 Charge distance scaling

The retained square-root scaled-distance relation is:

$$S_d = \frac{d}{\sqrt{Q/\text{delay}}} \qquad (1)$$

Its units are metres per square root kilogram when charge is in kilograms and distance in metres. A conventional empirical relation expresses PPV as:

$$\text{PPV} = K(S_d)^n \qquad (2)$$

The coefficients depend on the fitted domain, and the exponent is ordinarily negative for attenuation with increasing scaled distance. The coefficient K changes numerically when the units of scaled

distance or PPV change. Equations (1) and (2) provide an engineering baseline; they do not establish a universal attenuation law for every geology, delay pattern or propagation path [1,10].

Scaled distance was removed from the machine-learning feature list because it is derived from charge and distance. This avoids an explicitly redundant input, but does not remove possible nonlinear dependence between the retained predictors. A model fitted on the individual variables can learn a charge-distance response without being constrained to obey attenuation everywhere. Checking that response at feasible fixed-design combinations is therefore useful, especially when the generator samples outside the observed ranges.

2.3 Partitioning and preprocessing chronology

The cleaned table was divided into 56 real training rows and 24 held-out real test rows, corresponding to 70/30. Generator fitting followed this partition [9]. Figure 1 preserves the original chronology, which places real-data preprocessing before train-test splitting. Deterministic unit conversion can safely precede a split, and exact duplicate identity may need to be established globally to prevent copies crossing the boundary. By contrast, data-fitted outlier thresholds or distribution-dependent selection rules should be learned inside the training boundary [7].

The exclusion of eight rows represents 9.1% of the original database. The summary does not identify which exclusions were due to verified recording errors and which were statistical extremes, or where the Tukey thresholds were fitted. A large or unusual PPV observation may be a valid engineering event. Its removal can make the prediction problem easier while reducing the relevance of evaluation to extreme vibration. The reported results are therefore conditional on the cleaned population, not an estimate for every original blast condition.

2.4 Generators and synthetic sample counts

Gaussian Copula, CTGAN, TVAE and CopulaGAN were implemented through the SDV ecosystem [3,4,11]. Correct metadata are essential because column types affect transforms, likelihoods and postprocessing. B/D had six distinct observed levels and was represented as categorical during synthesis. This choice preserves the observed design levels, but may restrict generation to them. It does not establish that B/D is inherently categorical in every mine, and it requires a documented encoding when the downstream regressor consumes the column.

The Gaussian copula estimates marginals and a transformed dependence structure. Its relative simplicity can be advantageous with 56 real training rows, although Gaussian-copula tail independence and correlation-based dependence remain restrictive [12]. CTGAN uses conditional adversarial modelling and mode-specific transformations, and TVAE uses a variational latent representation [3]. CopulaGAN combines marginal transformation with adversarial generation [11]. Model labels alone do not determine the full implementation: library versions, marginal choices, architecture, random seeds and postprocessing affect the output.

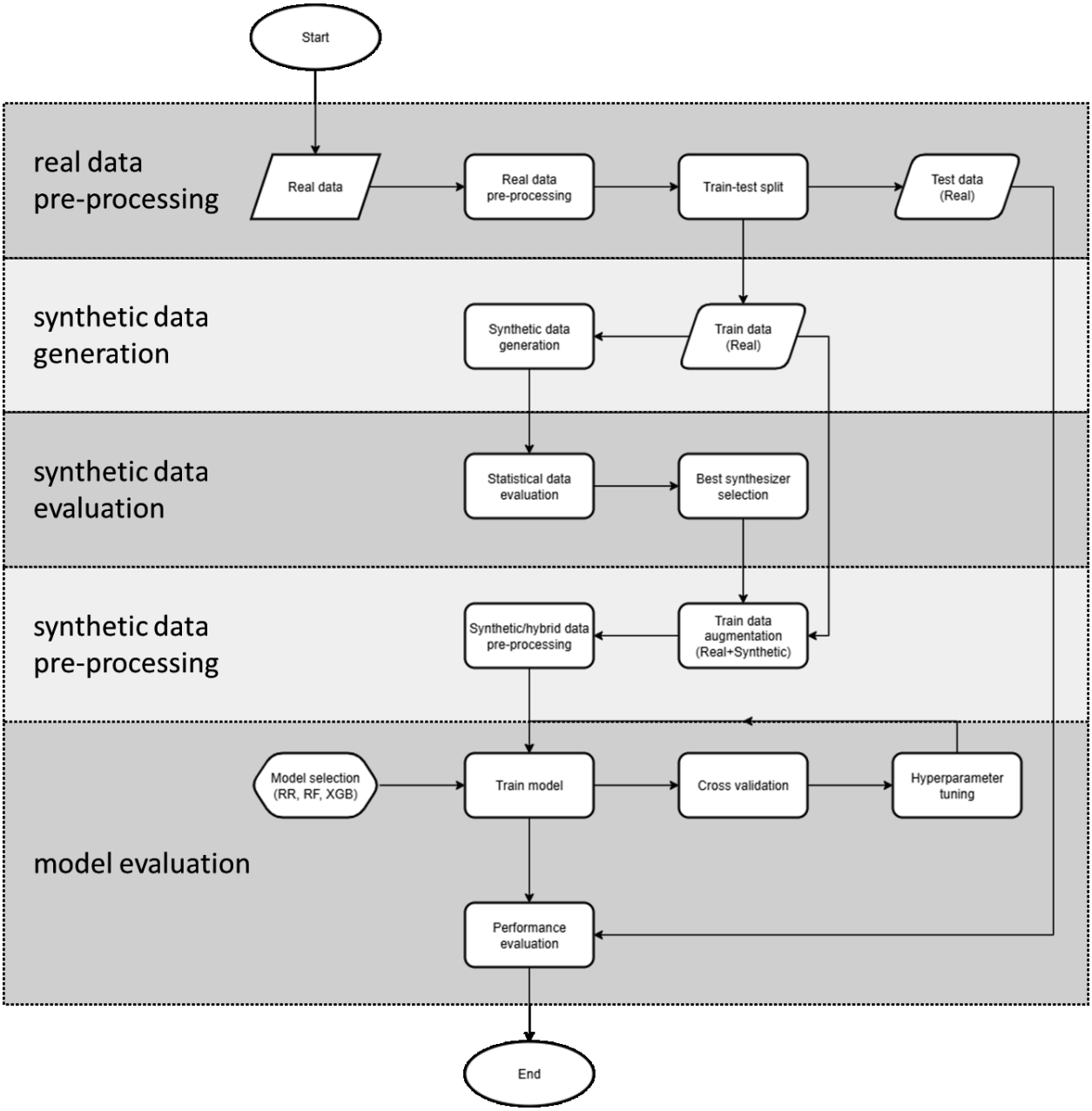


Figure 1. Original synthesis and evaluation workflow. The diagram retains its historical preprocessing order and all original stages. Section 2.3 distinguishes deterministic cleaning from data-fitted filtering; Section 2.7 explains the need to refit generators within real-data validation folds.

Neural generators were trained for 300 epochs in the source experiments. A plateau in training loss motivated this value, but does not establish optimal held-out fidelity or lack of memorisation. Minimum-maximum enforcement was disabled, allowing generated values beyond observed extrema. Positive-value constraints were imposed by rejection sampling, followed by the reported duplicate, missingness and Tukey-type filters [9]. Positivity is a necessary validity rule under these definitions, not a complete physical constraint on blast geometry and charge.

The selected workflow initially sampled 560 records, ten times the real training count, and retained 537 after filtering. The hybrid table contained 56 real plus 537 synthetic records, or 593 rows. The retention fraction is approximately 95.9%, and synthetic records account for approximately 90.6% of

the hybrid table. These calculated fractions characterise the implemented training distribution. They do not imply that information increased tenfold or that the unweighted hybrid model placed equal total influence on real and synthetic evidence.

The source mentions a favourable five-to-ten-fold sampling range, but the main summary does not provide a complete multiplier grid or repeated comparisons. Consequently, tenfold sampling is treated as the implemented setting, not an established optimum. Filtering also changes the sampled distribution; its thresholds and retained counts should be recorded separately for each synthesizer and validation fold.

2.5 Statistical resemblance and descriptive ranking

Evaluation combined histograms, SDV validity and structure diagnostics, column-shape and pair-trend scores, the Kolmogorov-Smirnov statistic, Jensen-Shannon summaries and Wasserstein distance [5,13]. Higher SDV similarity values indicate closer agreement under their particular diagnostics, while smaller distance values indicate closer distributional resemblance. These directions must be aligned explicitly before aggregation.

The KS statistic compares empirical cumulative distributions; it does not identify which tail contributes most to a discrepancy. Jensen-Shannon results require a declared probability estimator, common histogram bins or smoothing rule, logarithm base and distinction between divergence and its square-root distance. Wasserstein distance retains the units of its input unless a normalisation is applied. The benchmark's small numerical Wasserstein summaries therefore require the corresponding scaling specification before being interpreted in engineering units.

The source aggregates metric values and assigns average ranks described as Friedman-style. Because metric scales, normalisations and column-level dependencies are not fully specified, the combined value is an internal descriptive score rather than a universal statistical error. Average ranks can be a useful sensitivity check, but correlated columns of one small table do not constitute independent benchmark datasets for a Friedman significance test. No p-values or statistically significant superiority are inferred from this ranking [14].

2.6 Downstream predictors and target scale

Ridge Regression, Random Forest and XGBoost represent regularised linear, bagged-tree and boosted-tree models. Three regimes were compared: train on real and test on real, train on synthetic and test on real, and train on hybrid and test on real, abbreviated TRTR, TSTR and THTR. All three use the same 24 real hold-out rows in the reported experiment. Only Gaussian Copula was retained for the main downstream comparison after the resemblance assessment.

The source reports a log transformation of PPV and feature standardisation fitted on the original real training data. The same transform parameters were applied to synthetic, hybrid and test rows. Table 3 lists the initial search spaces and Table 4 the selected values. The latter include settings outside the initial coarse grid, consistent with a refinement stage described by the source, but the full refined search specification is needed for exact reproduction [9].

Table 3. Initial coarse hyperparameter ranges retained from the project. These are starting ranges, not a complete record of the subsequent refinement search. A dash indicates a parameter not used by that learner.

Model	Alpha	Trees	Max depth	Min split	Learning rate
Ridge	logspace(-3, 3, 50)	-	-	-	-
Random Forest	-	[100, 300, 500]	[None, 10, 20, 30]	[2, 5, 10]	-
XGBoost	-	[100, 300, 500]	[3, 6, 9]	-	[0.01, 0.1, 0.2]

If PPV is predicted on a logarithmic scale, exponentiating the predicted log value does not generally give the conditional arithmetic mean in mm/s. Retransformation bias matters for RMSE and high-tail predictions; a training-derived smearing or distributional correction is one established option [15]. The source labels the final error metrics in mm/s, but does not specify every inverse-transform detail. The retained scores are interpreted on the reported physical scale without asserting an unrecorded correction.

2.7 Cross validation and uncertainty

Cross-validation can assess sensitivity to real-data partitions only when the complete fitted pipeline respects each fold boundary. A generator fitted to all 56 real training rows before fold evaluation can transmit information about a real validation fold into the synthetic training records. Cross-validation performed within an already generated synthetic table instead estimates prediction within that fitted synthetic distribution. Neither design is equivalent to refitting the generator on each real training fold [7,8].

The reported CV summaries do not establish the fold-wise generator chronology. Their mean and standard deviation are therefore retained as descriptive diagnostics of the recorded workflow. Independent evaluation should refit scaling, marginals, generator, filters and learner tuning inside each real fold, with grouped splitting where multiple observations share a physical blast [16]. Repeated outer partitions and paired errors on real records are needed to quantify uncertainty in augmentation gains. Standard deviations over synthetic rows cannot substitute for that design.

3 Results

3.1 Resemblance of the generated distributions

Figure 2 shows the original nine-variable real-versus-synthetic histograms for all four generators. Gaussian Copula visually retains much of the observed shape, while the neural alternatives show variable-specific mismatches. Some blast-design fields have few design levels or a narrow empirical range; smoothing them can produce a visually continuous density without preserving the original discreteness. Histograms should therefore be read alongside metadata and numerical diagnostics rather than treated as direct tests of physical adequacy.

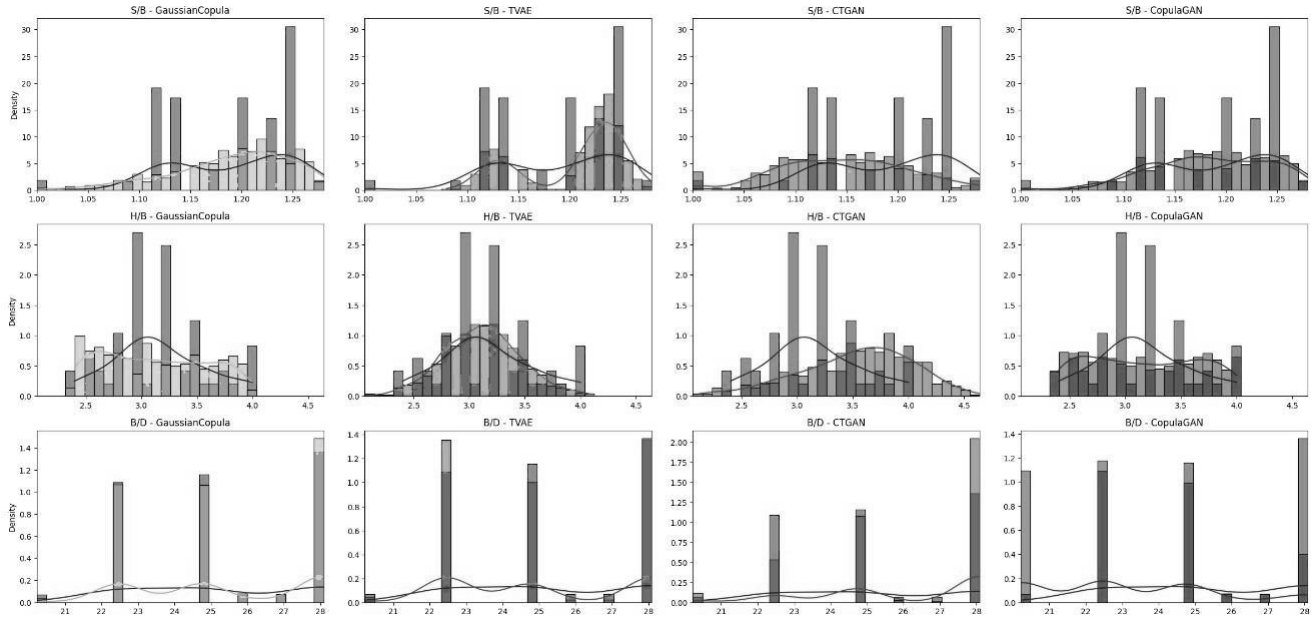


Figure 2a. Original distribution panels for S/B, H/B and B/D. All four synthesizers are retained in their original columns. The full original figure is presented in three consecutive parts for legibility.

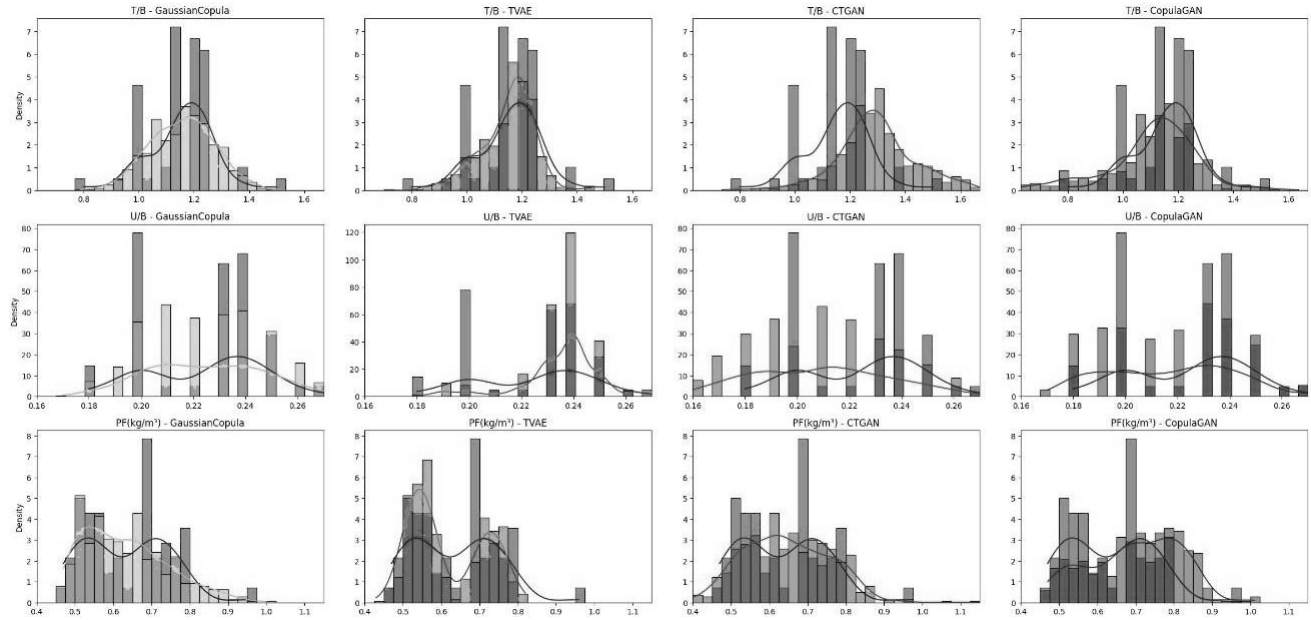


Figure 2b. Original distribution panels for T/B, U/B and powder factor. Continuation of the same original figure; no panel or numerical content is omitted.

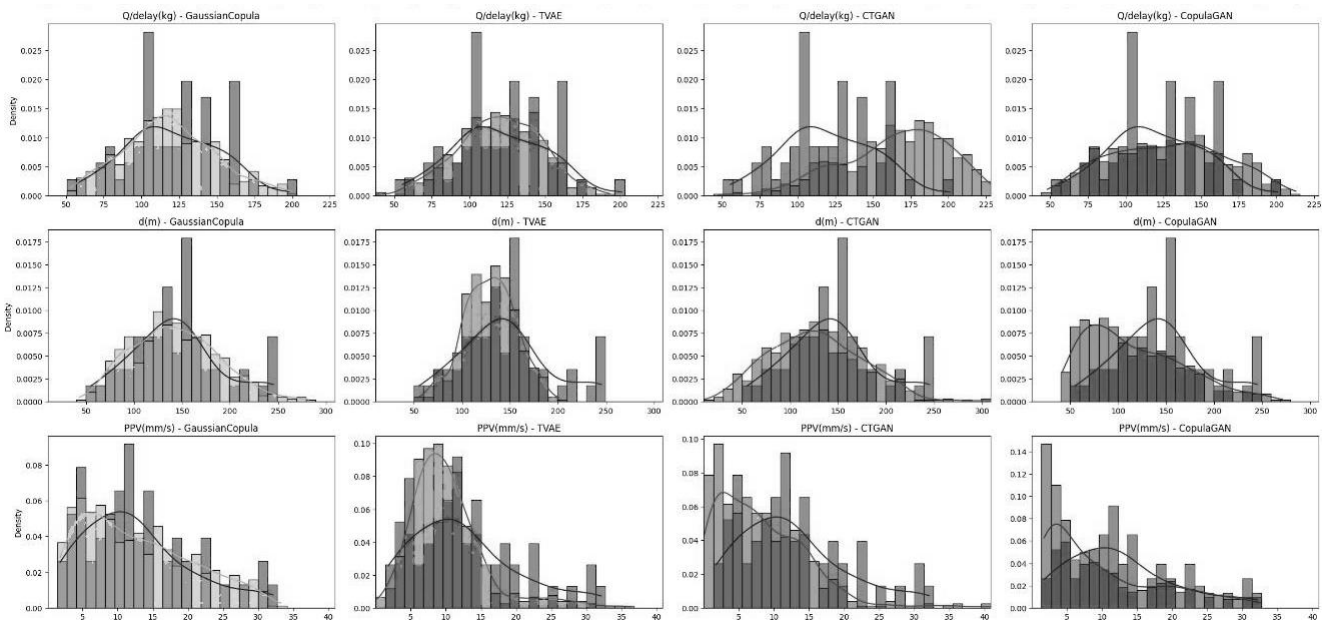


Figure 2c. Original distribution panels for charge per delay, distance and PPV. Continuation of the same original figure; no panel or numerical content is omitted.

The reported SDV scores favour Gaussian Copula, with data validity = 1.00, Column Shapes = 0.86 and Column Pair Trends = 0.86. TVAE achieved 0.82 and 0.83 for the latter two properties, while the GAN-based models had lower column-shape scores. A validity score of one indicates compliance with the evaluated validity rules, not validation of every physically meaningful variable combination.

The combined descriptive scores were 0.21 for Gaussian Copula, 0.23 for TVAE, 0.28 for CopulaGAN and 0.34 for CTGAN. Average rank values were 1.70, 1.96, 2.93 and 3.41 in the same order. Figure 3 retains the individual metric panels and both aggregates. The matching ordering supports the original generator choice descriptively, while the limitations of aggregation in Section 2.5 remain applicable.

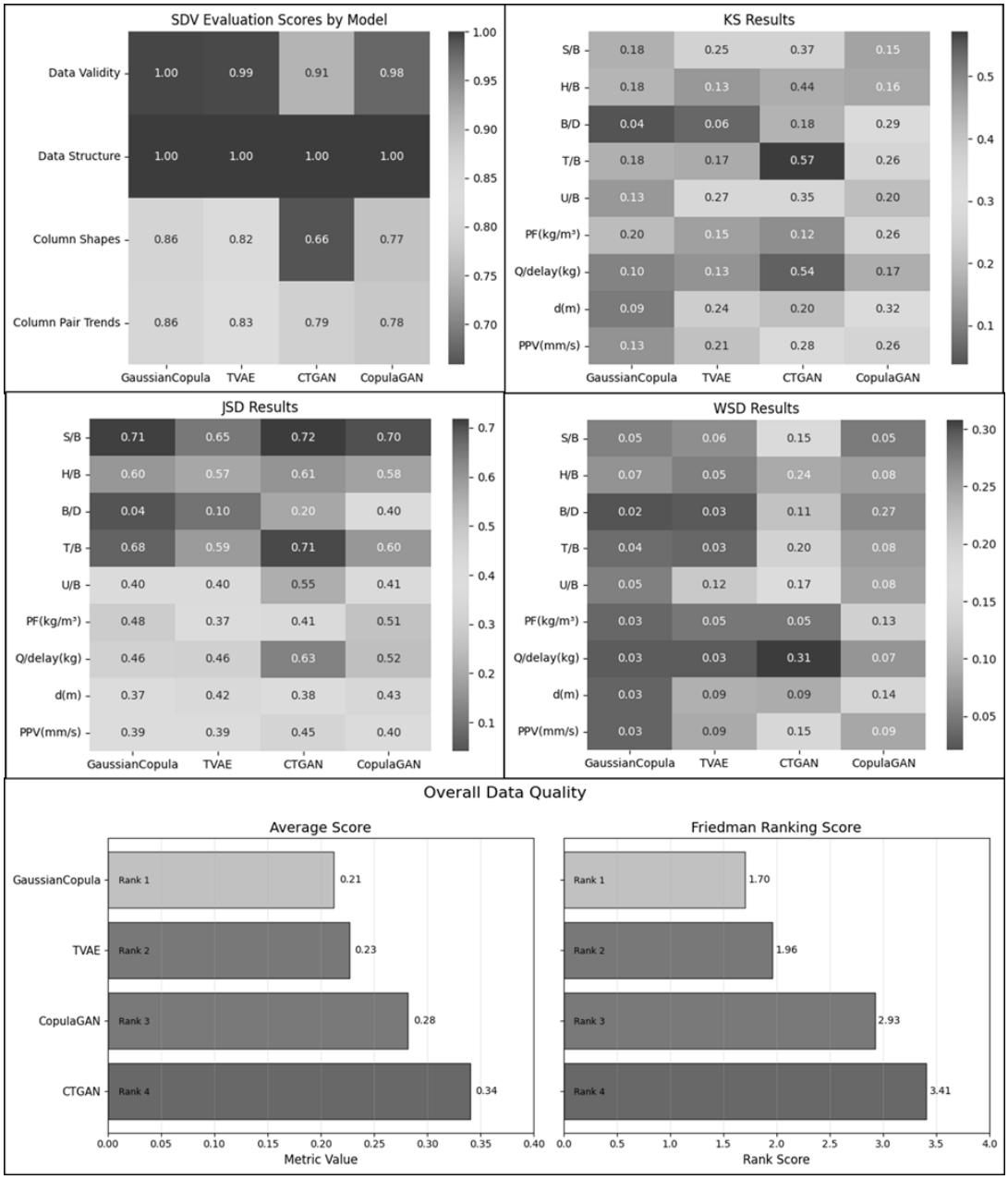


Figure 3. Original SDV, KS, Jensen-Shannon and Wasserstein summaries and aggregate rankings. All numerical annotations are retained. The combined values and Friedman-style averages describe the benchmark; they are not independent-data significance tests.

3.2 Selected hyperparameters

Ridge alpha increased from 4.71 under real-only training to 14.56 for synthetic-only and 10.99 for hybrid training. Random Forest selected maximum depth 5 on real data and 15 on synthetic and hybrid data. The selected XGBoost learning rates also differed. Table 4 preserves the full reported settings.

Table 4. Reported selected hyperparameters for each learner and training regime. Settings outside Table 3 reflect the reported refinement step; the complete search history is not given in the principal summary.

Model	Training	Alpha	Trees	Max depth	Min split	Learning rate
Ridge	Real	4.71	-	-	-	-
Ridge	Synthetic	14.56	-	-	-	-
Ridge	Hybrid	10.99	-	-	-	-
Random Forest	Real	-	300	5	2	-
Random Forest	Synthetic	-	300	15	4	-
Random Forest	Hybrid	-	400	15	4	-
XGBoost	Real	-	500	3	-	0.01
XGBoost	Synthetic	-	100	3	-	0.09
XGBoost	Hybrid	-	100	3	-	0.10

These changes establish that the fitted configurations differed across regimes. They do not prove that synthetic data introduced more genuine nonlinear information. Effective training size, fitted distribution, tuning noise and the search procedure can all change selected settings. Matched tuning budgets and fold-wise generation are needed to attribute any improvement to augmentation rather than unequal model selection.

3.3 Reported cross validation summaries

Figure 4 shows the reported mean CV R² increasing from 0.69 to 0.77 for Ridge, from 0.51 to 0.71 for Random Forest and from 0.44 to 0.75 for XGBoost when moving from real-only to hybrid training. Error-score standard deviations generally decreased. These are internally favourable summaries, but their interpretation is conditional on the unverified generator-refitting chronology in Section 2.7.

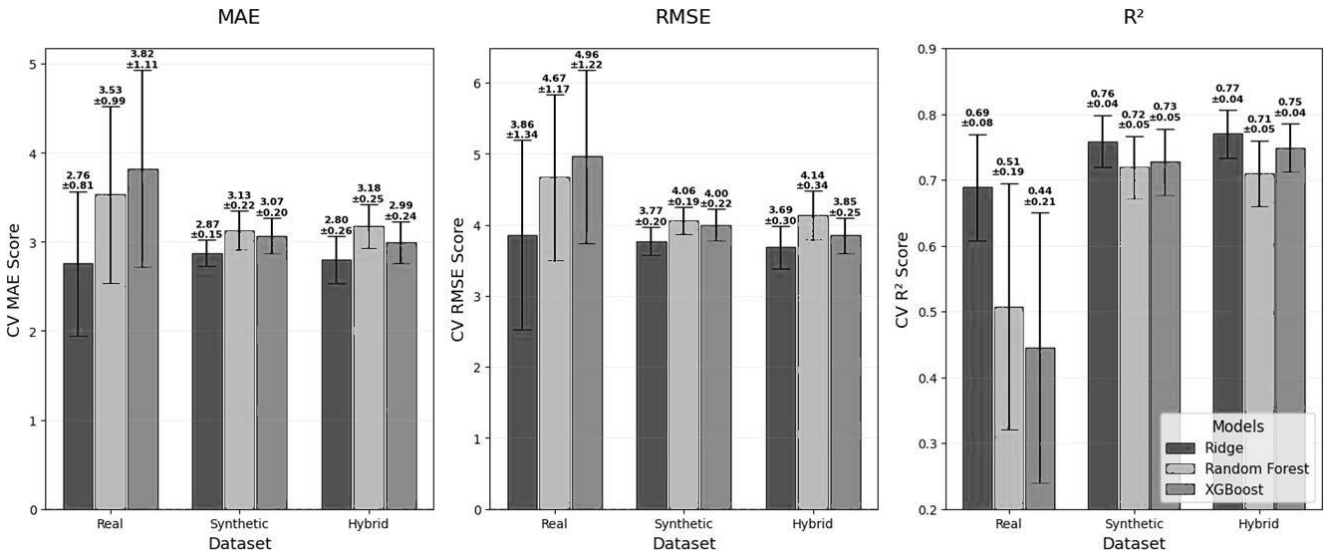


Figure 4. Original cross-validation means and reported standard deviations for real, synthetic and hybrid training. Model positions, bars and all values are retained. Reduced spread within this workflow is not, by itself, evidence of reduced uncertainty across independent physical blasts.

3.4 Performance on the fixed real hold out set

Random Forest's hybrid RMSE was 2.79 mm/s versus 3.44 mm/s for real-only training, an arithmetic reduction of 18.9%. Its R^2 increased from 0.69 to 0.79. XGBoost's hybrid RMSE was 2.93 mm/s versus 3.38 mm/s, a 13.3% reduction, with R^2 increasing from 0.70 to 0.77. These comparisons use the same real evaluation split and therefore provide the most relevant reported evidence of model-specific augmentation utility.

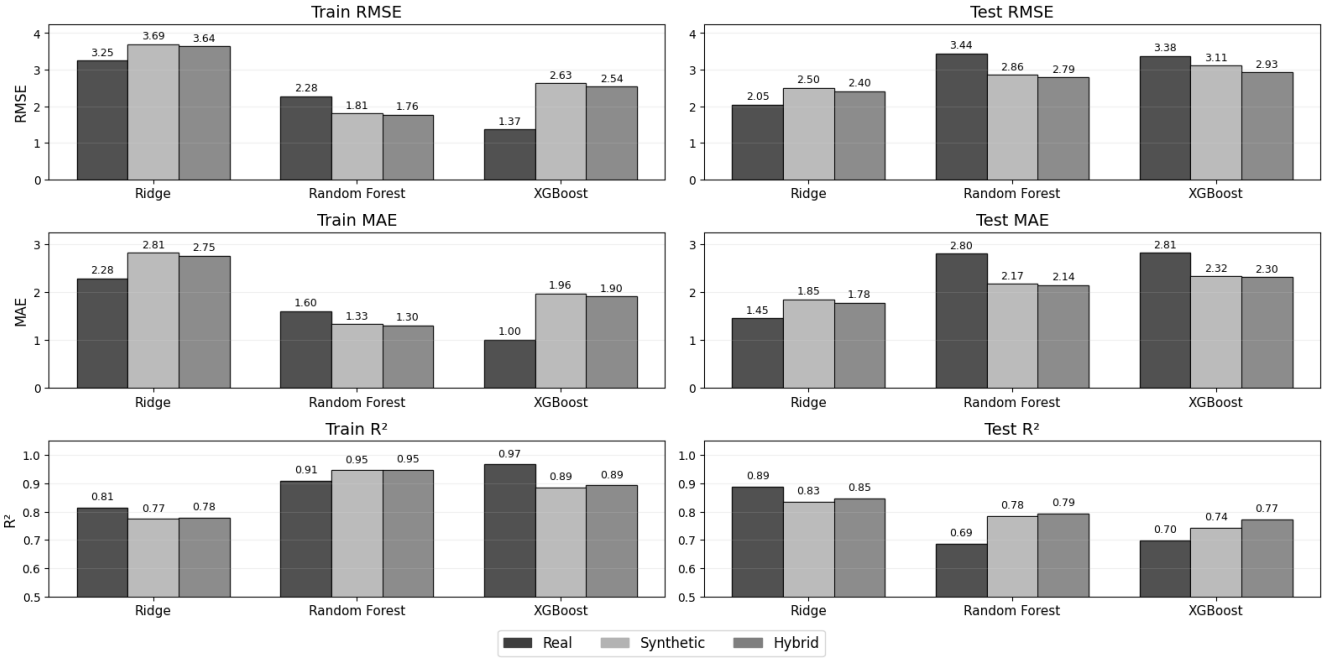


Figure 5. Original training and real-test performance for the three learners under all training regimes. Every panel and numerical annotation is retained. Real, synthetic and hybrid refer to training data; all right-hand test panels use real hold-out observations.

Table 5. Real hold-out metrics transcribed from the numerical annotations in Figure 5 to support direct comparison. RMSE and MAE are in mm/s as reported. The transcription adds no new model run, extra precision or confidence interval.

Learner	Training	RMSE	MAE	R^2
Ridge	Real	2.05	1.45	0.89
Ridge	Synthetic	2.50	1.85	0.83
Ridge	Hybrid	2.40	1.78	0.85
Random Forest	Real	3.44	2.80	0.69
Random Forest	Synthetic	2.86	2.17	0.78
Random Forest	Hybrid	2.79	2.14	0.79
XGBoost	Real	3.38	2.81	0.70
XGBoost	Synthetic	3.11	2.32	0.74
XGBoost	Hybrid	2.93	2.30	0.77

Ridge performed best on this hold-out set when trained on real data alone: RMSE = 2.05 mm/s, MAE = 1.45 mm/s and R^2 = 0.89. Its synthetic-only and hybrid RMSE values increased to 2.50 and 2.40 mm/s. The hybrid Ridge RMSE is approximately 17.1% higher than the real-only value. Moreover, the real-only Ridge error is lower than that of either hybrid tree model. The finding is consequently an

improvement within the tree-model families, not superiority of the augmented workflow over the strongest reported baseline.

4 Discussion

4.1 Generator simplicity in a small calibration set

Gaussian Copula provides an explicit low-complexity dependence representation relative to the neural generators evaluated here. With only 56 real training rows, that inductive bias may offer a useful bias-variance trade-off. However, the comparison does not establish a general law that copulas outperform neural synthesis in geoscience. The outcome also depends on marginal distributions, design-level metadata, optimisation settings, filtering and the particular dependency structure [3,5,12].

The weaker GAN resemblance is compatible with estimation or optimisation difficulties, but the figures do not diagnose mode collapse, unstable adversarial dynamics or a particular training failure. Such explanations require seed-level loss trajectories, coverage analysis and sample-level diagnostics. Likewise, a competitive TVAE score does not establish that its latent distribution captured every operational regime. The observed ordering should guide the next controlled comparison rather than replace it.

4.2 What the augmentation gain may represent

Sampling from the fitted copula changes the density of training combinations encountered by a learner. Tree ensembles may benefit when that smoothing reduces sensitivity to a few real rows. The generator may also impose a dependency structure that approximates useful regularisation. These are plausible interpretations of the observed Random Forest and XGBoost hold-out gains, rather than demonstrated mechanisms.

The predominance of synthetic rows makes weighting consequential. In the unweighted 593-row hybrid table, over nine tenths of records are synthetic. This can pull a learner toward errors in the fitted generator even when the real observations are retained. A controlled study should compare sample counts and explicit real-versus-synthetic weights, keeping downstream tuning and the real evaluation set fixed. A simpler resampling or regularisation baseline would help determine whether a generative model adds benefit beyond smoothing the small training set.

4.3 Ridge and the physical baseline

The strongest real-test result belongs to real-only Ridge. This is consistent with a test subset adequately represented by a regularised trend in the selected transformed features; it does not establish that the physical PPV process is linear. The contrasting CV ranking could reflect split composition, tuning variability or the synthetic-validation dependence discussed earlier. More folds on generated rows cannot resolve that disagreement.

An empirical scaled-distance fit and a simple log-linear charge-distance model would provide useful additional baselines. They should be fitted and assessed on exactly the same real splits as the machine-learning models. Their coefficients and residuals would help identify whether augmentation improves a relationship beyond charge and distance, or mainly reproduces the dominant attenuation trend. No new empirical-baseline score is inferred from the present results.

4.4 Tail behaviour and operational interpretation

RMSE gives large errors greater weight, but an overall RMSE can conceal systematic underprediction at high PPV. Outlier filtering may further reduce the number of demanding evaluation cases. An engineering assessment should therefore examine signed residuals, upper-quantile errors, charge-distance extrapolation and performance by geological or monitoring group. A monotonic decrease with distance at a fixed feasible configuration can be checked as a diagnostic, while recognising that unmodelled path effects can create scatter.

PPV point prediction alone does not establish an acceptable vibration outcome. Frequency content, duration, receiver conditions and measurement definitions are not represented by a single scalar error score. The dataset summary does not resolve whether PPV denotes a selected component maximum or another velocity aggregation. This definition and instrument processing must be harmonised before different sources can be combined. Calibrated prediction intervals on independent real records would express uncertainty more directly than presenting one favourable point estimate.

4.5 Descriptive ranking and validation independence

The agreement of visual and numerical resemblance rankings is useful but limited. Marginal KS, Jensen-Shannon and Wasserstein diagnostics are statistically dependent, especially when calculated on correlated columns from the same 56 records. Averaging their ranks does not create multiple independent experiments [14]. A robust evaluation should retain the individual diagnostics, specify normalisations and show how rankings respond to alternative reasonable metric choices.

The original workflow separates hold-out testing from synthesis, which is an important design feature. Its remaining uncertainty lies in operations surrounding that separation: how data-fitted cleaning used the full table, which data selected the synthesizer and multiplier, how CV regenerated synthetic data, and whether model decisions were revised after inspecting test results. These questions matter because model-selection bias can be similar in magnitude to differences between algorithms [8]. The fixed test gains are promising observations, while causal attribution to improved generalisation requires a more fully nested evaluation.

4.6 Reproducibility and external validation

A stronger protocol would first resolve physical-blast identity and measurement units, then reserve an outer real test set by blast or site. Each inner training fold would fit its own outlier thresholds, metadata-dependent transforms, generator, rejection rules and learner. Generator selection, synthetic multiplier and hybrid weights would be chosen without observing the outer outcomes. Repeating this procedure would produce paired real-test differences with uncertainty tied to independent blasts [7,8,16].

The reproducibility package should contain the original 88 rows with inclusion reasons, the final 80-row mapping, real split identifiers, all generator and learner configurations, software versions, seeds, accepted and rejected sample counts, target inverse-transform code and per-record predictions. Exact matches and nearest-neighbour distances should be compared with a real-hold-out reference to assess memorisation [6]. Synthetic generation does not automatically provide anonymisation or unrestricted rights to the source data [17]; the provenance of any released dataset remains material.

5 Conclusions

Gaussian Copula achieved the strongest reported resemblance in the analysed PPV benchmark and supported lower fixed-test errors for hybrid Random Forest and XGBoost training. The corresponding RMSE reductions of 18.9% and 13.3% are model-specific descriptive results. Real-only Ridge remained the strongest reported real-test model, so augmentation did not establish the best overall predictive performance.

The appropriate next step is independent, fully nested validation that preserves physical-blast identity, tests high-PPV behaviour and documents the complete target scale. The current evidence supports further investigation of copula-based augmentation as a regularisation strategy for small blasting tables. It does not establish a general reduction in uncertainty, transfer to new mines or replacement of vibration monitoring by synthetic observations.

References

[1] Hudaverdi, T. (2012). Application of multivariate analysis for prediction of blast-induced ground vibrations. Soil Dynamics and Earthquake Engineering, 43, 300-308. https://doi.org/10.1016/j.soildyn.2012.08.002

[2] Yan, Y.; Guo, J.; Bao, S.; Fei, H. (2024). Prediction of peak particle velocity using hybrid random forest approach. Scientific Reports, 14, 30793. https://doi.org/10.1038/s41598-024-81218-z

[3] Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. Advances in Neural Information Processing Systems, 32. https://proceedings.neurips.cc/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-Abstract.html

[4] Patki, N.; Wedge, R.; Veeramachaneni, K. (2016). The Synthetic Data Vault. IEEE International Conference on Data Science and Advanced Analytics, 399-410. https://doi.org/10.1109/DSAA.2016.49

[5] Fonseca, J.; Bacao, F. (2023). Tabular and latent space synthetic data generation: A literature review. Journal of Big Data, 10, 115. https://doi.org/10.1186/s40537-023-00792-7

[6] Alaa, A.; Van Breugel, B.; Saveliev, E.S.; van der Schaar, M. (2022). How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models. Proceedings of Machine Learning Research, 162, 290-306. https://proceedings.mlr.press/v162/alaa22a.html

[7] Kapoor, S.; Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns, 4(9), 100804. https://doi.org/10.1016/j.patter.2023.100804

[8] Cawley, G.C.; Talbot, N.L.C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research, 11, 2079-2107. https://jmlr.org/papers/v11/cawley10a.html

[9] Kamel, N.M. (2025). MSc Independent Research Project report. Imperial College London. Project identifier irp-nmk124. Experimental source identified in the accompanying manuscript. Repository recorded as https://github.com/ese-ada-lovelace-2024/irp-nmk124

[10] Duvall, W.I.; Fogelson, D.E. (1962). Review of criteria for estimating damage to residences from blasting vibrations. U.S. Bureau of Mines, Report of Investigations 5968.

[11] DataCebo. Single-table synthesizers. Synthetic Data Vault documentation. https://docs.sdv.dev/sdv/modeling/single-table-synthesizers (accessed 28 October 2026).

[12] Nelsen, R.B. (2006). An Introduction to Copulas. 2nd edition. Springer. https://doi.org/10.1007/0-387-28678-0

[13] DataCebo. Quality Report API. SDMetrics documentation. https://docs.sdv.dev/sdmetrics/data-metrics/quality/quality-report-api (accessed 28 October 2026).

[14] Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, 7, 1-30. https://jmlr.org/papers/v7/demsar06a.html

[15] Duan, N. (1983). Smearing estimate: A nonparametric retransformation method. Journal of the American Statistical Association, 78(383), 605-610. https://doi.org/10.1080/01621459.1983.10478017

[16] Roberts, D.R.; Bahn, V.; Ciuti, S.; et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40(8), 913-929. https://doi.org/10.1111/ecog.02881

[17] Stadler, T.; Oprisanu, B.; Troncoso, C. (2022). Synthetic data - Anonymisation groundhog day. 31st USENIX Security Symposium, 1451-1468. https://www.usenix.org/conference/usenixsecurity22/presentation/stadler