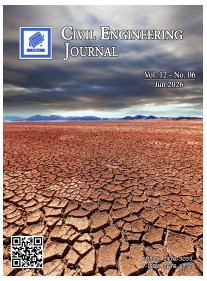
Available online at www.CivileJournal.org

Civil Engineering Journal

(E-ISSN: 2476-3055; ISSN: 2676-6957)

Vol. 12, No. 06, June, 2026



Evaluation of Synthetic Data Generation Methods for ANN-Based Bamboo Property Prediction

Nurwin Adam G. Muhammad [1, 2], Jerson N. Orejudos [3], Mary Joanne C. Aniñon [4*], Lessandro Estelito O. Garciano [4], Anjaneth M. Gonzales [5]

[1] *Department of Civil Engineering, Western Mindanao State University, Zamboanga City 7000, Philippines.*

[2] *Graduate School, Mindanao State University–Iligan Institute of Technology, Philippines, Iligan City 9200, Philippines.*

[3] *Department of Civil Engineering, Mindanao State University–Iligan Institute of Technology, Iligan City 9200, Philippines.*

[4] *Department of Civil Engineering, Gokongwei College of Engineering, De La Salle University, Manila 1004, Philippines.*

[5] *School of Graduate Studies, Mapua University, Manila 1002, Philippines.*

Received 05 January 2026; Revised 13 May 2026; Accepted 19 May 2026; Published 01 June 2026

Abstract

Data-driven modeling in bamboo research is hindered by the limited availability of openly accessible experimental datasets, as most studies report only summary statistics. This study evaluates whether synthetic data can reliably support data-driven modeling of bamboo mechanical properties. Three synthetic data generation methods – parametric Monte Carlo simulation (PMCS), parametric bootstrapping (PB), and Gaussian copula (GC) – were used to generate datasets based on published statistical descriptors of *Bambusa blumeana* for multiple sample sizes ($N = 1{,}000;\ 10{,}000;\ 100{,}000$). Artificial neural network (ANN) models were developed using each dataset, and both statistical fidelity and predictive performance were assessed. Results indicate that PMCS provides the highest statistical consistency with target distributions, while GC generally yields lower prediction errors. PB demonstrates intermediate performance. Both PMCS and GC exhibit the lowest relative errors, indicating that the means and standard deviations of the generated datasets closely match the target values reported in the literature. Feature importance analysis identifies density and cross-sectional area as the most influential predictors across all methods. Despite differences in statistical fidelity, ANN predictive performance remains comparable. These findings demonstrate that synthetic data can serve as a reliable alternative for experimental datasets and highlight the feasibility of developing ANN-based predictive models using published statistical descriptors.

Keywords: Synthetic Data Generation; Monte Carlo Simulation; Bootstrapping; Gaussian Copula; Artificial Neural Network.

1. Introduction

Bamboo continues to attract interest as a construction material due to its sustainability and its favorable mechanical properties [1, 2]; however, a comprehensive understanding of its behavior remains limited by the scarcity of extensive experimental datasets. Mechanical characterization of bamboo species such as *Bambusa blumeana* depends on extensive laboratory testing, but access to raw measurements is often limited by institutional policies, confidentiality agreements, and the high cost of destructive testing. Most published studies provide only summary statistics rather than full experimental records [3, 4]. This scarcity of openly accessible datasets restricts opportunities for secondary analysis, model development, and data-driven advances in bamboo engineering.

* Corresponding author: mary_joanne_aninon@dlsu.edu.ph

https://doi.org/10.28991/CEJ-2026-012-06-09

Recent studies have demonstrated the increasing use of machine learning (ML) methods, particularly artificial neural networks (ANNs), to predict bamboo mechanical properties from physical and geometric characteristics. Correal et al. [5] employed machine learning (ML) techniques to infer mechanical properties and structural grading of *Guadua angustifolia Kunth* bamboo by combining datasets from multiple laboratories, thereby improving statistical robustness. Similarly, Buachart et al. [6] developed ANN models to estimate the characteristic and allowable compressive strengths of *Dendrocalamus Sericeus* bamboo, demonstrating strong predictive performance compared to traditional regression approaches. Furthermore, Mallik et al. [7] explored the prediction of tensile strength of bamboo using various machine learning techniques, including ANN, extreme learning machine (ELM), and support vector regression (SVR). These studies demonstrate the capability of ML methods to capture nonlinear relationship in bamboo materials.

More recent work further highlights the growing adoption of data-driven approaches in bamboo research. Pilapil et al. [8] investigated the use of computer vision and regression models to predict the compressive and tensile capacity of *Bambusa blumeana* bamboo, identifying key physical indicators such as linear mass and wall thickness. Similarly, Ramful & Casseem [9] applied deep neural networks (DNNs) to predict mechanical properties of bamboo using a dataset of 600 samples, demonstrating the potential of artificial intelligence (AI) in modeling bio-based materials. The application of such approaches has also extended to engineered bamboo, such as bamboo-wood composites (BWC) [10, 11], where ANN models have been used to predict modulus of elasticity (MOE) and modulus of rupture (MOR) with high accuracy.

However, a common characteristic in these studies is their reliance on actual datasets, either generated through laboratory experiments or obtained from existing databases. Even when the datasets are limited in size, the availability of raw measurements remains a prerequisite for model development. In practice, however, such datasets are rarely publicly accessible. Most published studies report only statistical descriptors, such as means, standard deviations, and coefficient of variations, without providing underlying raw data. This limitation restricts reproducibility and prevents the direct application of ML techniques in situations where only summary statistics are available.

Synthetic data offers a practical response to these constraints. By generating statistically equivalent datasets from published descriptors such as mean, standard deviation, coefficient of variation, and assumed probability distributions, researchers can recreate material behavior even in the absence of raw experimental data. This approach is particularly relevant in bamboo research, where access to complete experimental dataset is often limited despite the availability of summary statistics, and where experimental testing is both time-consuming and costly. Synthetic datasets have gained increasing importance across scientific fields because they provide a controllable alternative to real measurements [12, 13], allow researchers to scale datasets without additional testing [14–16], and maintain essential statistical characteristics while concealing original observations [13, 14].

In structural materials research, synthetic data plays a growing role in supporting predictive modeling. Artificial neural networks (ANNs) are now widely used to estimate mechanical properties based on physical properties (5–10), but these models are highly sensitive to the size and quality of the training dataset. When only a limited number of physical tests are available, ANN models may suffer from underfitting, poor generalization, and unstable predictions [15, 17, 18]. Synthetic data generation techniques help address these issues, yet their effectiveness depends strongly on the statistical approach used to construct the synthetic datasets [13, 14, 16].

Several statistical frameworks exist for synthetic data generation. Parametric Monte Carlo simulation (MCS) relies on assumed probability distributions and generates data using estimated parameters [19]. Nonparametric techniques, such as bootstrap sampling, recreate variability without imposing strict distributional assumptions [20]. Gaussian copula provides a way to construct multivariate datasets by modeling marginal distributions independently from their dependence structure [21]. Although these approaches are well established in the statistical literature, their application in bamboo material modelling, particularly in conjunction with machine learning (ML), remains limited.

More importantly, no existing study has systematically evaluated how different synthetic data generation techniques influence the predictive performance of ANN models for bamboo materials. This gap is significant because bamboo exhibits natural variability, and different statistical techniques may capture its behavior with varying degrees of fidelity. Furthermore, existing ML studies continue to depend on available datasets, even when limited, and do not address scenarios where only statistical descriptors are available. As a result, it remains unclear whether reliable predictive relationships between mechanical properties and physical properties can be established in the absence of raw experimental data.

Addressing this limitation is essential for advancing data-driven bamboo research. A systematic comparison of synthetic data generation techniques can provide insights into their ability to reproduce statistical behavior while supporting accurate machine learning predictions. Such an approach enables the development of predictive models and reliability assessment without requiring additional experimental programs.

This study addresses this gap by conducting a comparative evaluation of three synthetic data generation methods methods: (a) parametric Monte Carlo simulation, (b) parametric bootstrapping, and (c) Gaussian copula, using multiple

sample sizes ($N = 1{,}000; 10{,}000;$ and $100{,}000$). The objectives are to generate synthetic datasets based on published statistical descriptors of *Bambusa blumeana* bamboo, develop ANN models trained on these datasets, assess both the statistical fidelity of the synthetic data and the predictive performance of the models, and determine whether meaningful relationships between material properties and physical or geometric characteristics can be established without access to raw experimental data. By focusing on data-scarce conditions, this study provides a practical framework for data-driven modelling in bamboo engineering and contributes to improving reproducibility and accessibility in material research.

2. Materials and Methods

This study generated synthetic datasets using three distinct methods and evaluated three sample sizes ($N = 1{,}000; 10{,}000;$ and $100{,}000$), representing small, moderate, and large dataset conditions, respectively. Each synthetic dataset was assessed using selected statistical fidelity metrics. Separate ANN models were trained on the generated datasets, and their predictive performance was evaluated based on established accuracy metrics. Finally, the synthetic data generation method that demonstrated the best overall performance across these evaluations is identified.

2.1. Data Collection

Descriptive statistics for *Bambusa blumeana*, including the mean, standard deviation, and coefficient of variation, were collected from the literature and were summarized in Tables 1 and 2. For variables where the probability distribution was not specified, a normal distribution was assumed to facilitate synthetic data generation. These statistics served as the basis for generating synthetic datasets, enabling the development of predictive models despite the limited availability of complete experimental records. Additionally, Table 3 presents the correlations between the input and output variables as reported in the literature [4, 22, 23]. Due to the limited availability of data for *Bambusa blumeana*, some correlations were adapted from studies on *Phyllostachys edulis*, while others were assumed based on engineering judgment.

Table 1. Descriptive statistics of output variables

Output Variables	Mean	Standard Deviation (SD)	Coefficient of Variation (COV)	Probability Distribution	Source
Compressive strength F_c, (MPa)	64.67	13.62	0.21	Normal	Cacanando et al. [3]
Tensile strength F_t, (MPa)	110.81	31.30	0.28	Lognormal	
Bending strength F_m, (MPa)	88.15	20.40	0.23	Normal	
Shear strength F_v, (MPa)	10.84	2.65	0.24	Normal	
Modulus of Elasticity E_M, (MPa)	20,000	4,283.56	0.21	Normal	

Table 2. Descriptive statistics of input variables

Input Variables	Mean	Standard Deviation (SD)	Coefficient of Variation (COV)	Probability Distribution	Source
Moisture Content MC, (%)	10.9000	0.0080	0.0735	Normal*	Panti et al. [4]
Density D at $MC = 12\%$, (kg/mm^3)	721.610	100.1700	0.1400	Normal*	Cacanando et al. [3]
Cross-sectional Area A, (m^2)	2059.99	377.4500	0.1800	Normal*	Panti et al. [4]
Wall Thickness T, (mm)	8.2100	1.5900	0.1950	Normal*	Cacanando et al. [3]
Outer Diameter OD, (mm)	97.3200	9.3200	0.0960	Normal*	Cacanando et al. [3]

**Assumption*

Table 3. Correlation between input and output variables from literature [4, 22, 23]

Variables	F_c	F_t	F_m	F_v	E_m
OD	0.0593	0.2300	0.4900	0.2300	0.2280
T	0.0130	0.6000	0.3940	0.0600	0.2200
A	0.0410	0.3000	0.3940	0.0300	0.2200
MC	0.2652	0.2652	0.2652	0.2300	0.2652
D	0.4860	0.1500	0.5600	0.1500	0.7300

2.2. Synthetic Data Generation

Synthetic data refers to artificially generated datasets that are not obtained from direct measurements or real-world observations. They are useful when actual measurements are unavailable, limited, or restricted due to privacy or security concerns, providing a flexible and safe alternative for analysis.

In bamboo research, comprehensive experimental datasets are often scarce or rarely publicly accessible, with only descriptive statistics such as mean, standard deviation, coefficient of variation, and assumed probability distribution available. Synthetic data generation allows researchers to reconstruct representative datasets that mimic real material behavior without accessing original measurements.

2.2.1. Parametric Monte Carlo Simulation

Monte Carlo simulation is a statistical technique used to model uncertainty by repeatedly sampling random values from specified probability distributions [24]. By generating a large number of random realizations, the method approximates the expected behavior of a system and quantifies its variability using standard statistical measures [24]. In a parametric framework, Monte Carlo simulation relies on probability distributions defined by known parameters such as the mean, standard deviation, and coefficient of variation. This makes the approach suitable when raw experimental data are unavailable, but descriptive statistics can be obtained from literature or limited experimental testing.

In this study, a parametric Monte Carlo simulation (PMCS) was employed to generate synthetic datasets based on reported statistical descriptors of bamboo physical and mechanical properties. Multiple sample sizes were considered ($N = 1{,}000; 10{,}000;$ and $100{,}000$) to examine how dataset size influenced model performance. A fixed random seed was applied to ensure reproducibility because, without this control, each simulation run would generate different synthetic values, leading to variations in means, variances, and distribution shapes.

It is important to emphasize that this study aimed to develop a surrogate predictive model rather than to establish causal relationships. Accordingly, it was assumed that a latent nonlinear relationship existed between the input and output variables for modeling purposes. The input variables were generated according to the available descriptive statistics and assumed probability distributions. Because the reported correlations between the input variables and the target output were generally weak, independently generating the output variable would result in synthetic datasets with no meaningful functional dependency, leading to poor predictive performance of the ANN. To address this limitation, the output variable was constructed as a weighted summation of the input variables, where the weights were derived from normalized correlation coefficients reported in the literature, as presented in Table 3. A stochastic noise term was incorporated to represent unobserved variability and modeling uncertainty. The resulting output was subsequently scaled to match its published mean and standard deviation, yielding the final synthetic response variable. Figure 1 illustrates a flowchart summarizing the parametric Monte Carlo simulation procedure.

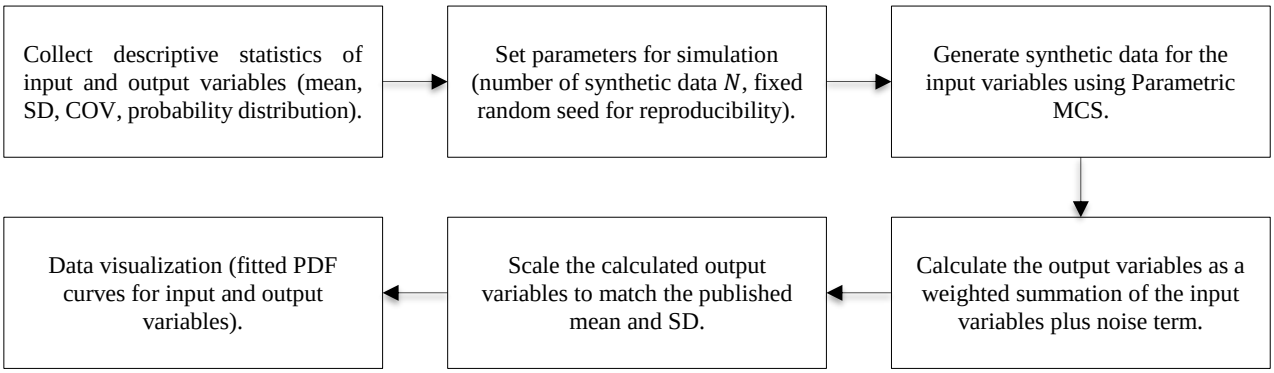


Figure 1. Flowchart for parametric Monte Carlo simulation

2.2.2. Parametric Bootstrapping

Bootstrapping is a resampling technique that creates new datasets by repeatedly drawing values from an existing sample with replacement. Each draw returns the selected value to the pool, making every sample independent and allowing the creation of resampled datasets that may exceed the size of the original [20, 25]. Bootstrapping is widely used to approximate sampling distributions, estimate uncertainty, and support data-driven modeling when the true underlying distribution is unknown or when the available sample is small [20, 25].

In classical applications, bootstrapping requires an existing dataset from which values can be resampled. In this study, however, real measurements are not available. To address this limitation, a parametric bootstrapping (PB)

approach was adopted. This variant first generates a set of pseudo-samples using descriptive statistics reported in the literature, including the mean, variance, coefficient of variation, and assumed probability distributions for each variable.

The procedure began by compiling all available descriptive statistics for the bamboo mechanical and physical properties. The simulation parameters were then specified, including the target synthetic dataset sizes ($N = 1{,}000; 10{,}000;$ and $100{,}000$) and a fixed random seed to ensure reproducibility. A pseudo-sample size, denoted as $pseudoN$, was also defined. Since no established guideline exists for selecting $pseudoN$ relative to N, this study adopted a ratio of 50 as an initial benchmark, subject to future refinement.

Parametric Monte Carlo simulation was first employed to generate the pseudo-samples based on the reported statistics. These pseudo-samples then served as the source dataset for the bootstrapping process. Resampling with replacement was performed repeatedly until the desired synthetic dataset size was reached for each value of N. Figure 2 shows a flowchart summarizing the parametric bootstrapping procedure.

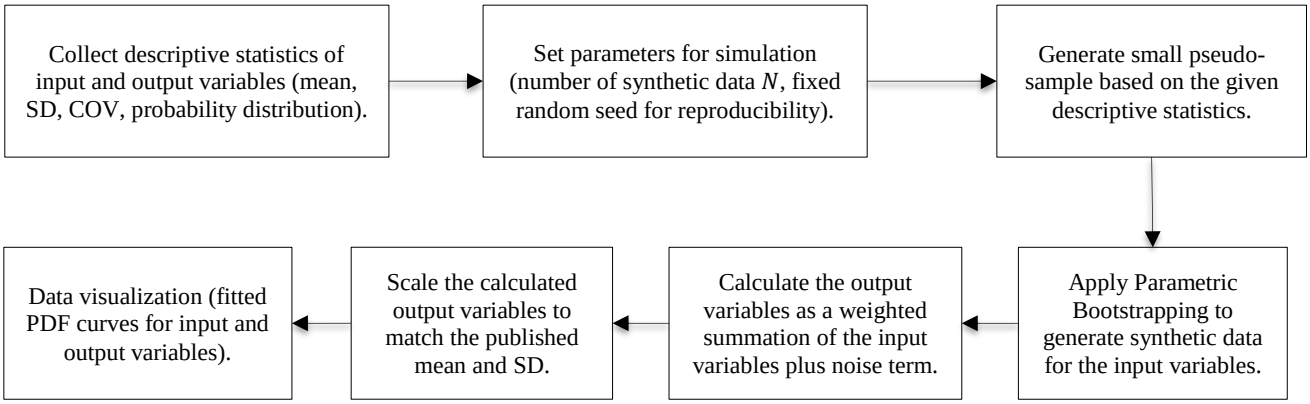


Figure 2. Flowchart for parametric bootstrapping

2.2.3. Gaussian Copula

A copula is a function that links univariate marginal distributions to form a multivariate joint distribution [21]. Sklar's theorem formalizes this separation of marginals and dependence [26]. In the Gaussian copula, dependence is represented through a correlation matrix in a latent normal space, while each variable retains its own marginal distribution [27]. Each observed variable is transformed to a uniform probability scale using the fitted cumulative distribution function (CDF), the dependence is modeled via a Gaussian correlation structure, and samples are mapped back through inverse CDFs [27].

The Gaussian copula (GC) framework requires a correlation matrix describing the dependence between input variables. However, due to the absence of empirical input-input correlation data for *Bambusa blumeana*, the present implementation assumed statistical independence among the input variables. Consequently, the correlation matrix was initialized as an identity matrix, where all off-diagonal elements are zero. This assumption provided a neutral baseline while preserving the ability to incorporate empirically derived correlations in future studies by updating the matrix entries.

To generate synthetic data, a set of multivariate normal random variables was first drawn from a latent Gaussian space using the specified correlation matrix. This was accomplished using a zero-mean multivariate normal distribution, producing an $N \times num\_input$ matrix of latent variables, where N is the number of synthetic data ($N = 1{,}000; 10{,}000;$ and $100{,}000$), and num_input is the number of input variables.

The latent normal samples were then transformed into independent uniform random variables on the interval [0 1] by applying the standard normal CDF. This conversion ensured that the dependent structure defined in the Gaussian space was preserved while converting the data to a common probabilistic scale.

Finally, the synthetic input variables were obtained by mapping the uniform samples to their respective marginal distributions using inverse CDFs. For variables assumed to follow normal distributions, the inverse normal CDF was applied using the specified mean and standard deviation. For lognormally distributed variables, the corresponding lognormal inverse CDF was used after converting the reported mean and standard deviation to lognormal parameters. This step produced synthetic input data that matched the prescribed marginal statistics while maintaining the dependence structure imposed by the Gaussian copula. As in the Monte Carlo simulation framework, the input variables were generated first. The output variable was subsequently calculated following the procedures described in section 2.2.1. Figure 3 presents a flowchart summarizing the Gaussian copula procedure.

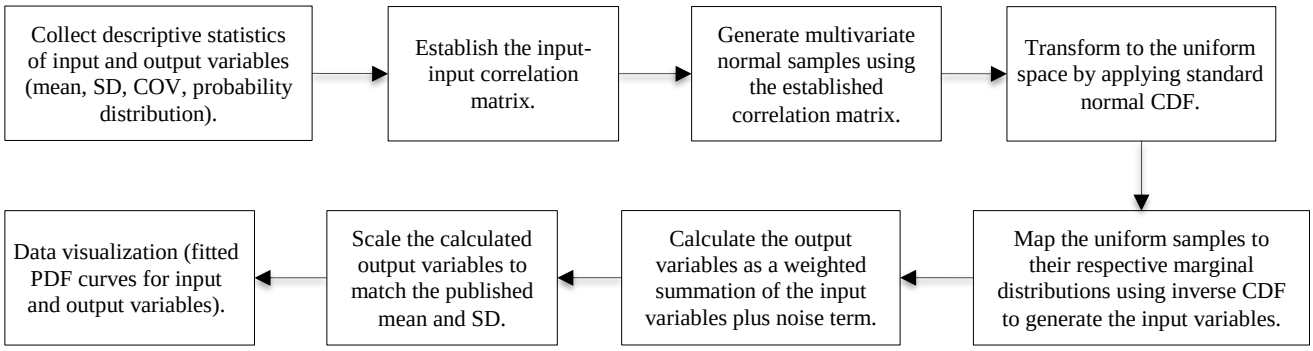


Figure 3. Flowchart for the Gaussian copula

All generated datasets were visualized to assess whether the resampled data reproduce the expected statistical behavior. All generation processes were conducted in MATLAB R2025b [28].

2.3. Data Preparation

Data preparation was conducted before ANN model development to ensure the quality, consistency, and suitability of all synthetic datasets. Each dataset was initially visualized using histograms with fitted probability density functions (PDFs) to inspect distribution shapes and verify whether the simulated data conformed to the assumed statistical behavior. Correlation analysis was then performed using Pearson correlation heatmaps to examine relationships among input variables and detect potential multicollinearity.

All datasets were normalized to improve training stability and accelerate convergence. The normalized datasets are subsequently partitioned into training (70%), validation (15%), and testing (15%) subsets. Random sampling was employed to ensure that each subset preserved the statistical characteristics of the original synthetic dataset. The training set was used to optimize network weights, the validation set to tune hyperparameters and prevent overfitting, and the testing set to assess model generalization.

2.4. Artificial Neural Network

A fully connected feedforward artificial neural network (ANN) was developed to predict the mechanical properties of *Bambusa blumeana* using five physical input variables. Rather than adopting a multi-output ANN, individual ANN models were constructed for each output variable. Preliminary trials indicated that networks with multiple output neurons exhibited unstable training behavior and degraded predictive performance due to the high level of noise and weak inter-output dependency in the synthetic data. Consequently, each output variable was modeled independently to ensure stable training and reliable performance evaluation.

For each output variable, three ANN models were developed corresponding to the three synthetic data generation techniques: parametric Monte Carlo simulation, parametric bootstrapping, and Gaussian copula. Given that five mechanical output variables were considered, this results in a total of 15 ANN models. Furthermore, since three sample sizes ($N = 1{,}000; 10{,}000;$ and $100{,}000$) were examined for each synthetic data generation method, the complete analysis involves 45 ANN models. This modeling framework enables a systematic assessment of the influence of both data generation technique and sample size on ANN predictive performance.

A grid search strategy was employed to determine an optimal ANN configuration. The hyperparameters investigated include various hidden-layer architectures, consisting of single and double hidden layers with different numbers of neurons, as well as alternative training algorithms. Specifically, the Levenberg–Marquardt algorithm (*trainlm*), scaled conjugate gradient method (*trainscg*), and Bayesian regularization approach (*trainbr*) were evaluated. The hyperbolic tangent sigmoid transfer function (*tansig*) was adopted for the hidden layers to capture nonlinear relationships, while a linear transfer function (*purelin*) was used in the output layer to accommodate continuous-valued predictions. For all models, the mean squared error (MSE) served as the loss function. The maximum number of training epochs was set to 200, with early stopping controlled by validation performance to mitigate overfitting.

Each candidate network was trained using MATLAB's *fitnet* function, and its predictive performance was evaluated using multiple metrics, including Mean Squared Error (MSE), Root Mean Squared Error ($RMSE$), Mean Absolute Error (MAE), Mean Absolute Percentage Error ($MAPE$), correlation coefficient (R), and coefficient of determination (R^2). Model performance was assessed separately for the training, validation, and testing subsets to evaluate generalization capability, training stability, and potential overfitting. The MSE was used as the primary criterion to rank candidate models and select the optimal baseline ANN architecture.

Following model training, regression plots were generated for the training, validation, testing, and overall datasets. These plots provide a visual assessment of agreement between predicted and target values and serve as a qualitative tool to examine model accuracy, generalization capability, and potential bias before proceeding to feature importance analysis.

2.5. Evaluation Metrics

2.5.1. Synthetic Data Fidelity Metrics

The fidelity of the synthetic datasets was evaluated by comparing their statistical characteristics with reference values extracted from published experimental studies. For each variable, the mean, standard deviation (SD), and coefficient of variation (COV) are computed and compared with the corresponding reported statistics to assess agreement in central tendency and dispersion.

To quantify these differences, the relative error (RE) is calculated for each statistical descriptor as presented in Equation 1.

$$RE(\%) = \left|\frac{S_{syn}-S_{ref}}{S_{syn}}\right| \times 100 \quad (1)$$

where S_{syn} denotes the statistic obtained from the synthetic dataset and S_{ref} represents the reference value reported in the literature. Smaller relative error values indicate closer agreement between the synthetic and reference statistics, thereby reflecting higher data fidelity.

In addition to summary statistics, distributional consistency was examined. When reference probability distributions were available, goodness-of-fit tests, including the Kolmogorov-Smirnov (KS) and Anderson-Darling (AD) tests, were employed to evaluate whether the synthetic data follow the same assumed underlying distributions. Visual inspections using fitted PDF plots were also performed to support the quantitative assessments.

For variables lacking reported probability distributions in the literature, the distributions were assumed based on commonly adopted models in related studies. In such cases, the fidelity evaluation does not represent a direct comparison against experimental distributions but rather an assessment of how well the generated synthetic data conformed to the stated modeling assumptions. This distinction was acknowledged to ensure transparency regarding the limitations of the available reference information.

These statistical and distributional evaluations provided an essential validation step, ensuring that the synthetic datasets reasonably reproduce the reported descriptive characteristics before being used for ANN model development.

2.5.2. ANN Model Performance Metrics

The predictive capability of the ANN models was evaluated using several standard error metrics, namely the Mean Squared Error (MSE), Root Mean Squared Error ($RMSE$), Mean Absolute Error (MAE), Mean Absolute Percentage Error ($MAPE$), correlation coefficient (R), and coefficient of determination (R^2) as expressed in Equations 2 to 7. These indicators quantify the discrepancy between the ANN-predicted values and the corresponding synthetic dataset values. For Equations 2 to 5, smaller numerical values indicate better predictive accuracy, whereas for Equations 6 and 7, values closer to unity signify stronger model performance. Using multiple evaluation metrics provides a more robust assessment of model performance by capturing different aspects of prediction error (Equations 5 and 6). Detailed interpretations of these evaluation metrics can be found in the authors' previous works [18, 29].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \hat{Z}_i)^2 \quad (2)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Z_i-\hat{Z}_i)^2}{n}} \quad (3)$$

$$MAE = \sqrt{\frac{\sum_{i=1}^{n}(Z_i-\hat{Z}_i)}{n}} \quad (4)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Z_i-\hat{Z}_i}{Z_i}\right) \quad (5)$$

$$R = \frac{\sum_{i=1}^{n}(Z_i-\bar{Z})(\hat{Z}_i-\bar{\hat{Z}})}{\sum_{i=1}^{n}(Z_i-\bar{Z})^2\sum_{i=1}^{n}(\hat{Z}_i-\bar{\hat{Z}})^2} \quad (6)$$

$$R^2 = \left(\frac{\sum_{i=1}^{n}(Z_i-\bar{Z})(\hat{Z}_i-\bar{\hat{Z}})}{\sum_{i=1}^{n}(Z_i-\bar{Z})^2\sum_{i=1}^{n}(\hat{Z}_i-\bar{\hat{Z}})^2}\right)^2 \quad (7)$$

where, Z_i is the synthetic (target) value; $\hat{Z}_i$ is the output (predicted) value; $\bar{Z}$ and $\bar{\hat{Z}}$ are the mean values of the target and predicted datasets, respectively, and n is the total number.

2.6. Feature Importance Analysis

The contribution of individual input variables was further investigated using permutation feature importance (PFI) analysis on the trained baseline ANN. In this approach, each input feature was permuted independently while keeping all others unchanged, and the resulting increase in prediction error was recorded. Features that caused larger increases in MSE were considered more influential. The importance scores were normalized to facilitate comparison across variables and were used to rank the input features according to their relative contribution to model performance.

Based on the PFI results, a reduced set of important input features was identified using a threshold defined as a fraction of the maximum importance score. To ensure model robustness, a minimum number of input variables was retained even if fewer features exceeded the threshold. The dataset was then reduced to include only the selected features, and the inputs were renormalized using the same scaling procedure.

Because feature reduction alters the learning characteristics of the ANN, a second grid search was performed using only the selected input variables. The same range of hidden-layer configurations and training algorithms was explored, and model selection was again based on the MSE values. The best-performing network from this second grid search was retrained and designated as the feature-selected ANN model.

The predictive performance of the feature-selected ANN was evaluated using the same performance metrics as the baseline model. Finally, the performance of the baseline and feature-selected ANN models was compared directly to determine whether feature selection improved prediction accuracy, reduced model complexity, or enhanced generalization capability.

3. Results and Discussion

3.1. Descriptive Statistics of the Synthetic Data

This section presents the descriptive statistics of the synthetic datasets generated using the three data generation techniques described in Section 2.2, namely parametric Monte Carlo simulation, parametric bootstrapping, and Gaussian copula. The summary statistics for all input and output variables are reported in Tables 4 to 6 for sample sizes of $N = 1{,}000$, $10{,}000$, and $100{,}000$, respectively. For each variable and sample size, the mean, standard deviation (SD), coefficient of variation (COV), and the identified probability distributions based on Kolmogorov–Smirnov (KS) and Anderson–Darling (AD) tests were provided.

Table 4. Descriptive statistics of the synthetic data generated using a parametric Monte Carlo simulation

Variables	Sample Size	Mean	Standard Deviation	Coefficient of Variation	Probability Distribution
OD	1000	97.3220	9.0644	0.0931	Lognormal
	10,000	97.4068	9.3199	0.0957	Lognormal
	100,000	97.3250	9.3176	0.0957	Lognormal
T	1000	8.1515	1.5741	0.1931	Lognormal
	10,000	8.2132	1.6050	0.1954	Lognormal
	100,000	8.2042	1.5849	0.1932	Lognormal
A	1000	2070.1789	372.9705	0.1802	Lognormal
	10,000	2052.1697	377.0838	0.1837	Lognormal
	100,000	2059.2324	376.8837	0.1830	Lognormal
MC	1000	10.9001	0.0085	0.0008	Normal
	10,000	10.9000	0.0080	0.0007	Lognormal
	100,000	10.9000	0.0080	0.0007	Normal
D	1000	721.7981	101.3098	0.1404	Normal
	10,000	721.1785	99.0444	0.1373	Normal
	100,000	721.7400	100.3522	0.1390	Normal
F_c	1000	64.6700	13.6132	0.2105	Normal
	10,000	64.6700	13.6193	0.2106	Normal
	100,000	64.6700	13.6199	0.2106	Normal
F_t	1000	110.8100	31.2843	0.2823	Normal
	10,000	110.8100	31.2984	0.2825	Lognormal
	100,000	110.8100	31.2998	0.2825	Lognormal
F_m	1000	88.1500	20.3898	0.2313	Lognormal
	10,000	88.1500	20.3990	0.2314	Lognormal
	100,000	88.1500	20.3999	0.2314	Lognormal
F_v	1000	10.8673	2.6581	0.2446	Normal
	10,000	10.8400	2.6499	0.2445	Normal
	100,000	10.8400	2.6500	0.2445	Normal
E_m	1000	20000.00	4281.4177	0.2141	Normal
	10,000	20000.00	4283.3458	0.2142	Normal
	100,000	20000.00	4283.5386	0.2142	Normal

Table 5. Descriptive statistics of the synthetic data generated using parametric bootstrapping

Variables	Sample Size	Mean	Standard Deviation	Coefficient of Variation	Probability Distribution
OD	1000	96.9569	12.0466	0.1242	Lognormal
	10,000	96.9680	8.2637	0.0852	Lognormal
	100,000	96.8991	9.2696	0.0957	Lognormal
T	1000	7.2863	1.4173	0.1945	Normal
	10,000	8.0501	1.5344	0.1906	Normal
	100,000	8.1953	1.5905	0.1941	Lognormal
A	1000	2005.6306	443.5258	0.2211	Lognormal
	10,000	2067.2313	377.8618	0.1828	Lognormal
	100,000	2061.5699	387.4268	0.1879	Lognormal
MC	1000	10.8993	0.0084	0.0008	Normal
	10,000	10.8994	0.0080	0.0007	Normal
	100,000	10.9001	0.0083	0.0008	Normal
D	1000	721.4622	118.1085	0.1637	Lognormal
	10,000	709.7559	96.4808	0.1359	Normal
	100,000	722.9818	97.8668	0.1354	Normal
F_c	1000	64.6700	13.6132	0.2105	Lognormal
	10,000	64.6700	13.6193	0.2106	Normal
	100,000	64.6700	13.6199	0.2106	Normal
F_t	1000	110.8100	31.2843	0.2823	Lognormal
	10,000	110.8100	31.2984	0.2825	Normal
	100,000	110.8100	31.2998	0.2825	Lognormal
F_m	1000	88.1500	20.3898	0.2313	Lognormal
	10,000	88.1500	20.3990	0.2314	Normal
	100,000	88.1500	20.3999	0.2314	Lognormal
F_v	1000	10.8400	2.6487	0.2443	Lognormal
	10,000	10.8400	2.6499	0.2445	Normal
	100,000	10.8400	2.6500	0.2445	Normal
E_m	1000	20000.00	4281.4177	0.2141	Lognormal
	10,000	20000.00	4283.3458	0.2142	Normal
	100,000	20000.00	4283.5386	0.2142	Normal

Table 6. Descriptive statistics of the synthetic data generated using a Gaussian copula

Variables	Sample Size	Mean	Standard Deviation	Coefficient of Variation	Probability Distribution
OD	1000	97.4965	9.8829	0.1014	Lognormal
	10,000	97.3574	9.3176	0.0957	Lognormal
	100,000	97.3212	9.3656	0.0962	Lognormal
T	1000	8.2167	1.6133	0.1963	Lognormal
	10,000	8.1997	1.5679	0.1912	Lognormal
	100,000	8.2125	1.5913	0.1938	Lognormal
A	1000	2059.2868	365.7562	0.1776	Lognormal
	10,000	2063.5389	378.0173	0.1832	Lognormal
	100,000	2060.2643	377.4241	0.1832	Lognormal
MC	1000	10.8997	0.0080	0.0007	Normal
	10,000	10.9000	0.0081	0.0007	Lognormal
	100,000	10.9000	0.0080	0.0007	Lognormal

D	1000	724.7213	99.0417	0.1367	Normal
	10,000	719.5658	100.8523	0.1402	Normal
	100,000	721.5850	100.0092	0.1386	Normal
F_c	1000	64.6700	13.6132	0.2105	Normal
	10,000	64.6700	13.6193	0.2106	Normal
	100,000	64.6700	13.6199	0.2106	Normal
F_t	1000	110.8100	31.2843	0.2823	Lognormal
	10,000	110.8100	31.2984	0.2825	Lognormal
	100,000	110.8100	31.2998	0.2825	Lognormal
F_m	1000	88.1500	20.3898	0.2313	Lognormal
	10,000	88.1500	20.3990	0.2314	Lognormal
	100,000	88.1500	20.3999	0.2314	Lognormal
F_v	1000	10.8400	2.6487	0.2443	Normal
	10,000	10.8400	2.6499	0.2445	Normal
	100,000	10.8400	2.6499	0.2445	Normal
E_m	1000	20000.00	4281.4177	0.2141	Normal
	10,000	20000.00	4283.3458	0.2142	Normal
	100,000	20000.00	4283.5386	0.2142	Normal

For both input and output variables, the computed means, standard deviations, and COVs exhibited consistent numerical behavior as the sample size increased across all three generation techniques. Minor fluctuations were observed at smaller sample sizes ($N = 1{,}000$), while progressively more stable statistics were obtained for $N = 10{,}000$and $N = 100{,}000$. This trend was apparent in physical input variables, indicating reduced sampling variability with increasing dataset size.

The standard deviations and corresponding coefficients of variation (COVs) did not show a strictly consistent decrease with increasing sample size. In several variables, both the standard deviation and COV increased from $N = 1{,}000$ to $N = 10{,}000$. However, the differences between the statistics obtained for $N = 10{,}000$ and $N = 100{,}000$ were noticeably smaller than those between $N = 1{,}000$ and $N = 10{,}000$. This behavior indicated that while small sample sizes may exhibit greater numerical fluctuation, larger sample sizes led to more stable estimates of dispersion.

When comparing the input variables across the three synthetic data generation techniques, the estimated means, standard deviations, and COVs remained closely aligned for all sample sizes. Parametric Monte Carlo simulation and Gaussian copula yielded nearly identical summary statistics, with only negligible differences observed. The parametric bootstrap approach showed slightly larger variation in summary statistics at smaller sample sizes, though these differences diminished as the sample size increased. For the output variables, similar behavior was observed across the three generation techniques. The estimated means remained stable across different sample sizes. The corresponding standard deviations and COVs showed comparable trends, with very slight variability.

Across all variables, sample sizes, and data generation techniques, the synthetic datasets displayed consistent and numerically stable descriptive statistics. Increasing the sample size generally improved the stability of the computed means, standard deviations, and COVs. Differences between the three generation methods were most noticeable at smaller sample sizes and diminished as N increases.

Overall, the presented statistics demonstrate consistent behavior for both input and output variables across all synthetic data generation techniques. These results provided a comprehensive descriptive overview of the dataset generated and established a foundation for the statistical fidelity assessment presented in the subsequent section.

3.2. Statistical Fidelity of Synthetic Data

The statistical fidelity of the synthetic datasets was evaluated by examining both the distributional behavior and the summary statistics of the generated variables. The reference experimental statistics reported in Tables 1 and 2 defined

the target mean, standard deviation, coefficient of variation (COV), and assumed probability distribution for each physical and mechanical property. The fidelity assessment, therefore, focused on (a) agreement between assumed and empirically identified probability distributions and (b) relative errors in summary statistics across different synthetic data generation techniques and sample sizes.

3.2.1. Probability Distribution Agreement

The synthetic datasets were generated by explicitly specifying probability distributions for each variable based on published literature or reasonable assumptions where references were unavailable. However, when the generated data were subjected to goodness-of-fit testing using the Kolmogorov–Smirnov (KS) and Anderson–Darling (AD) tests, the empirically identified distributions did not always coincide with the initially assumed distributions, as summarized in Figure 4.

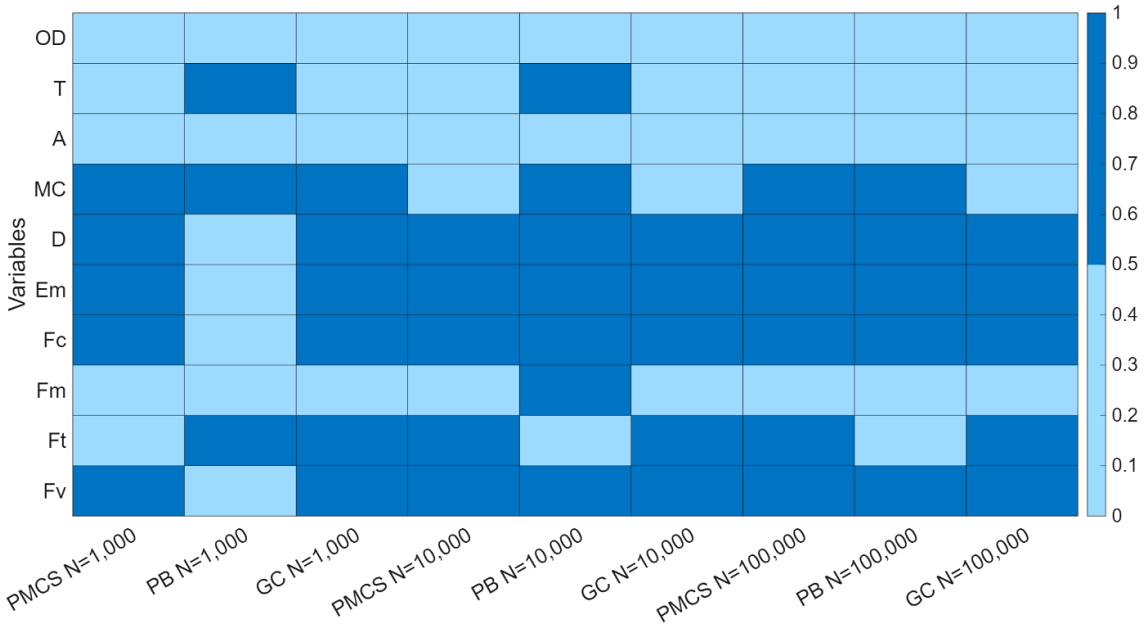


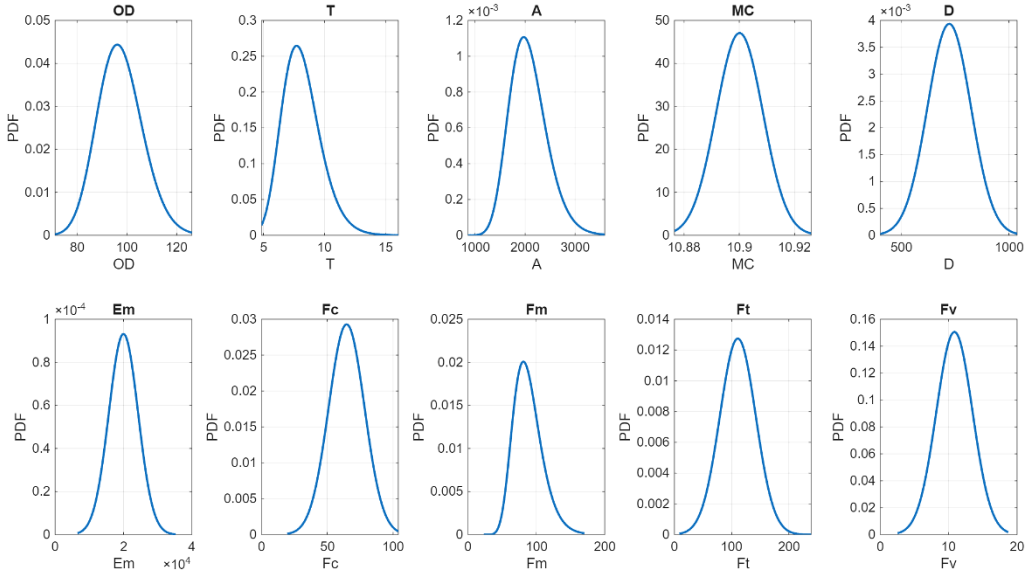
Figure 4. Agreement between the assumed probability distributions and those identified from the synthetic data using goodness-of-fit tests, where light blue indicates disagreement and dark blue indicates agreement

This apparent discrepancy does not indicate an error in the data generation process. Instead, it reflects several well-established statistical and methodological factors. First, goodness-of-fit tests evaluate how well a candidate distribution fits a finite sample, not how the data were generated. Even when data are drawn from a known distribution, KS and AD tests may favor an alternative distribution, particularly when multiple distributions exhibit similar shapes over the observed data range. This effect was more pronounced for moderately skewed variables or variables with limited variance, where normal and lognormal distributions may appear statistically interchangeable.

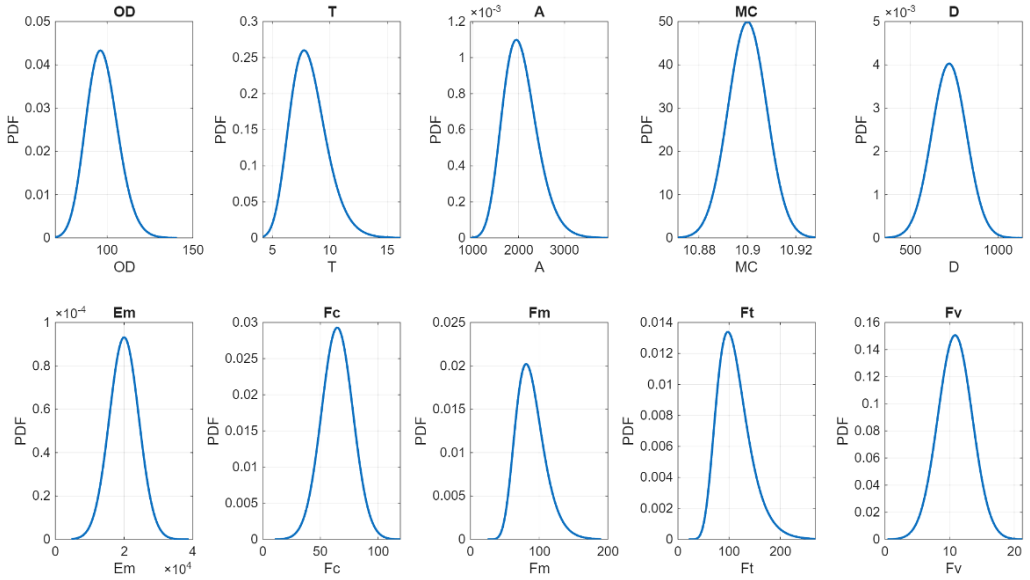
Second, some variables were generated using assumed probability distributions due to the absence of definitive experimental evidence. In these cases, the goodness-of-fit tests effectively evaluate the consistency of the synthetic data against the assumption itself rather than against verified experimental distributions. Consequently, disagreement in Figure 4 should be interpreted as sensitivity to distributional assumptions rather than a failure of the synthetic data to reproduce the prescribed statistical structure.

Third, sampling variability and transformation effects also contribute to distributional mismatch. In parametric bootstrapping, resampling with replacement from finite pseudo-samples can introduce irregularities in tails and local clustering, particularly at smaller sample sizes. In the Gaussian copula approach, dependence enforcement in the latent normal space can slightly moderate extreme values, affecting tail behavior when mapped back to the marginal distributions.

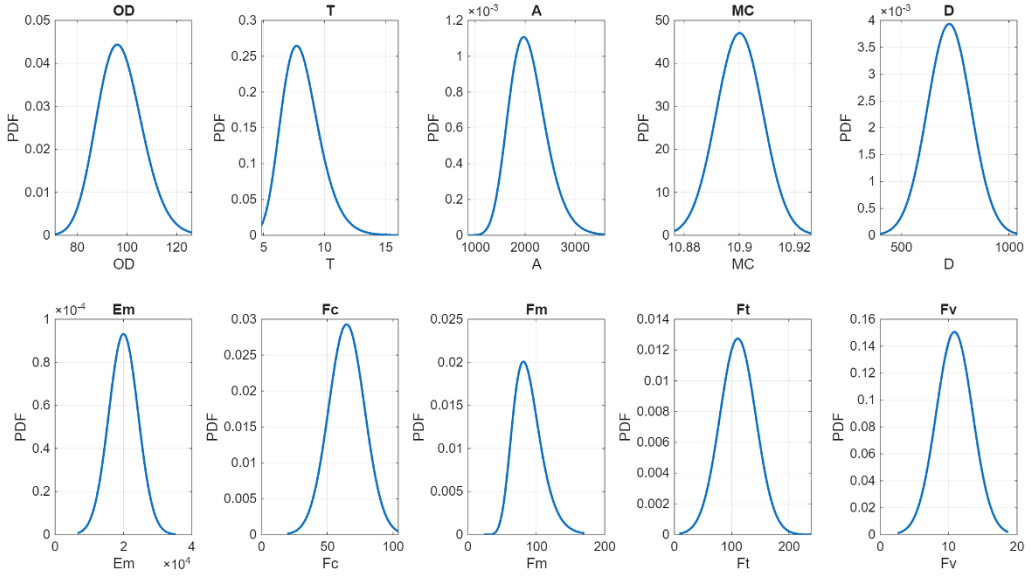
Despite these differences, visual inspections on the PDF plots in Figures 5 to 7 indicated that all three data generation techniques reproduced acceptable overall shape, symmetry, and skewness expected from the target distributions. Normal variables remain approximately symmetric, while lognormal variables consistently exhibit right-skewed behavior. Thus, while strict distributional classification may vary under KS and AD testing, the marginal distributional behavior remains physically consistent with the intended assumptions.



(a)



(b)



(c)

Figure 5. Probability density functions of synthetic input and output variables using parametric Monte Carlo simulation: (a) $N = 1000$; (b) $N = 10,000$; and (c) $N = 100,000$

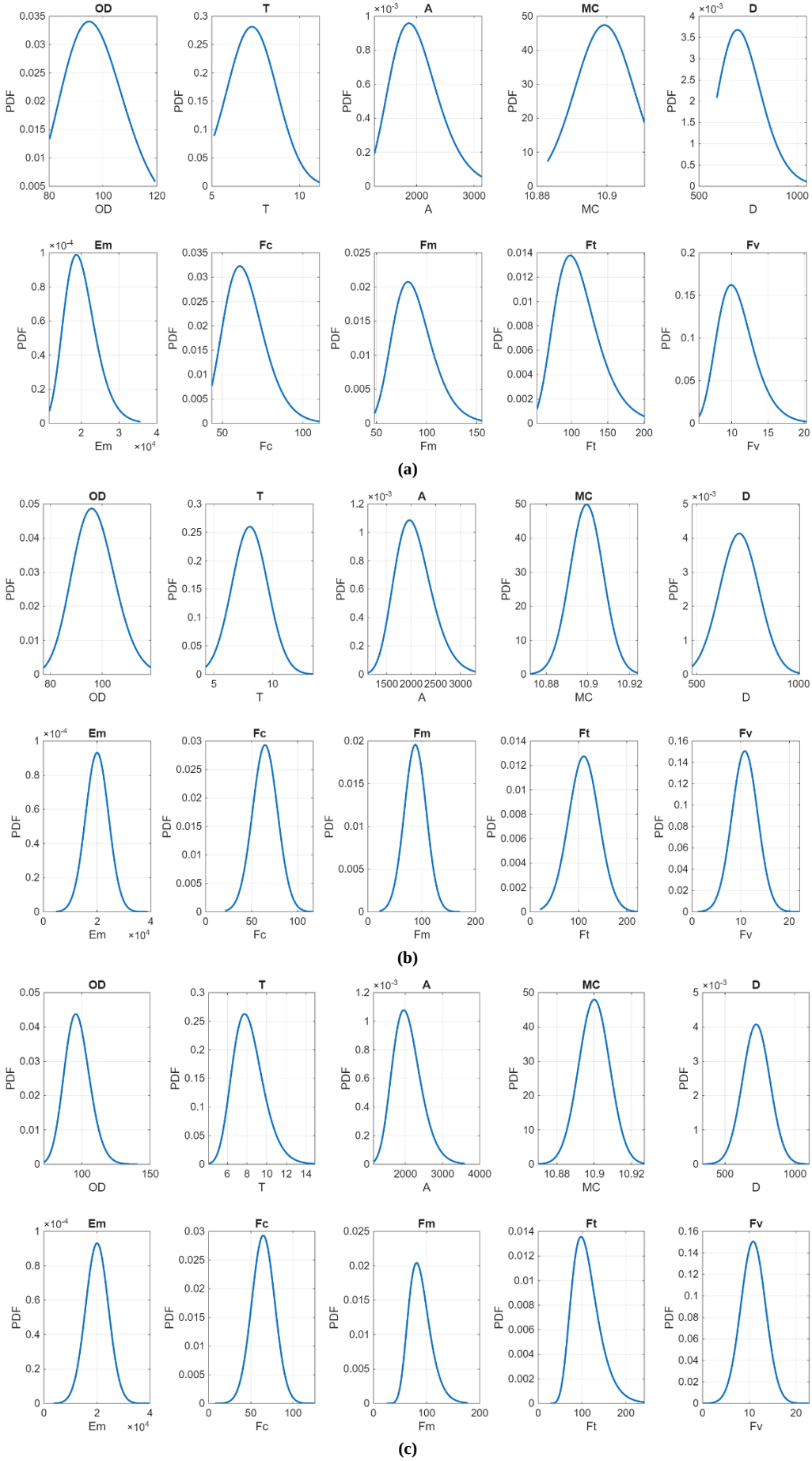


Figure 6. Probability density functions of synthetic input and output variables using parametric bootstrapping: (a) $N = 1000$; (b) $N = 10,000$; and (c) $N = 100,000$

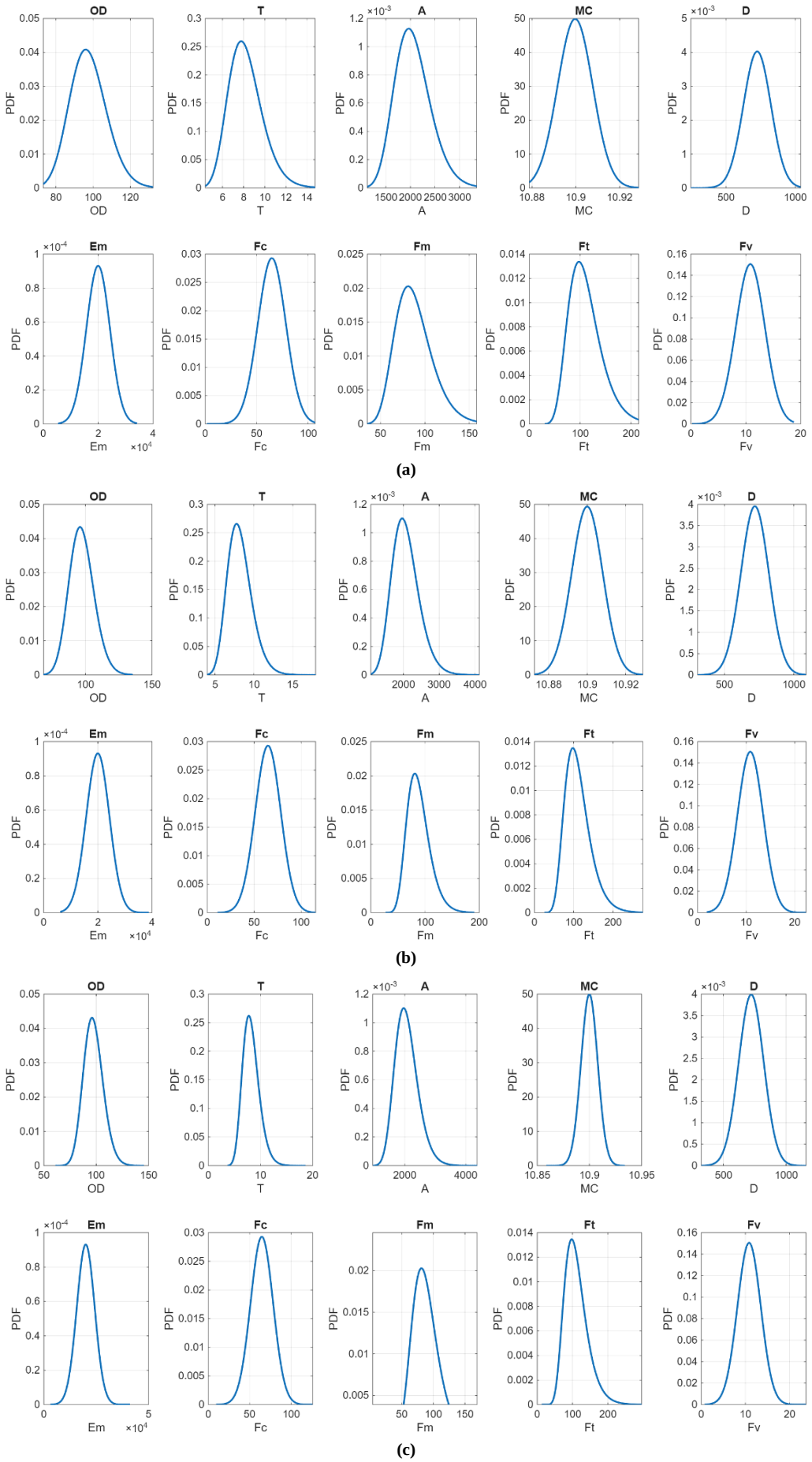


Figure 7. Probability density functions of synthetic input and output variables using Gaussian copula: (a) $N = 1000$; (b) $N = 10,000$; and (c) $N = 100,000$

3.2.2. Correlation Matrices

Figures 8 to 10 present the correlation matrices obtained from the synthetic datasets. Across all data generation methods and sample sizes, consistent trends were observed. In particular, D and A exhibited moderate to relatively high correlations with the output variable, whereas the remaining input variables showed weak or negligible correlations.

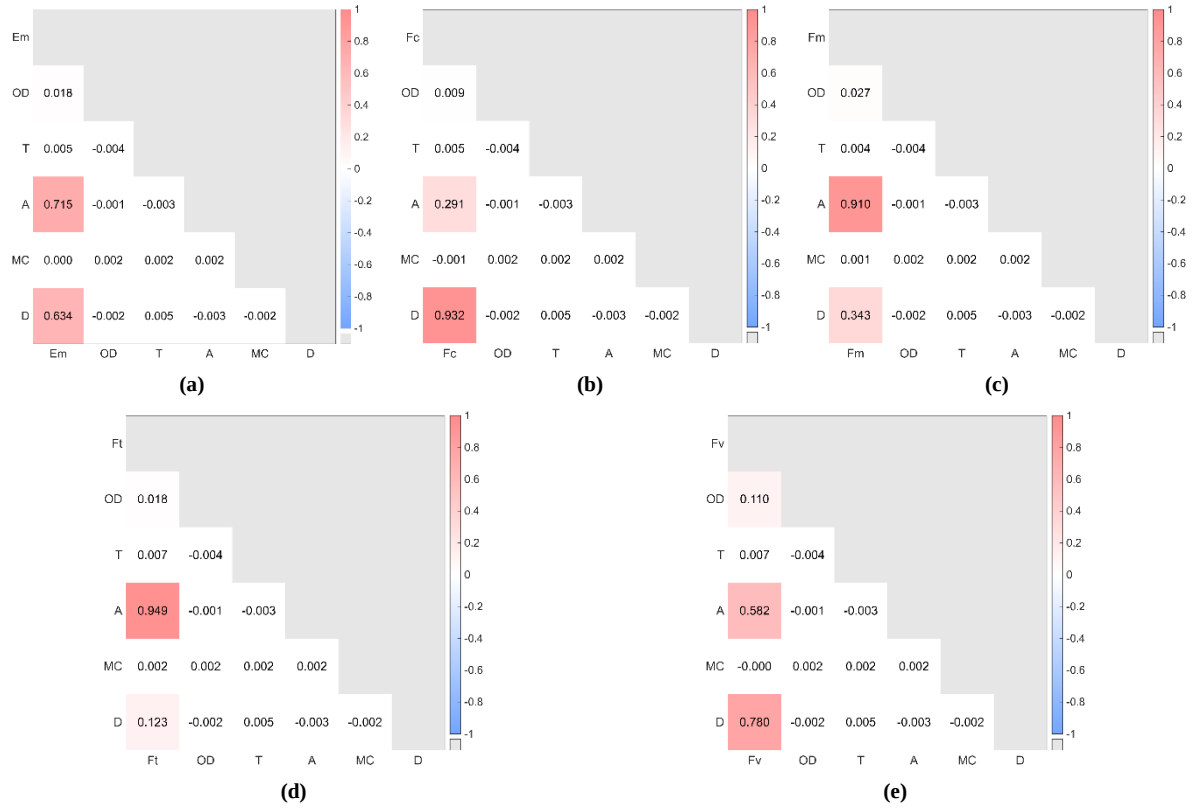


Figure 8. Correlation matrix for $N = 100,000$ using PMCS: (a) E_m; (b) F_c; (c) F_m; (d) F_t; and (e) F_v

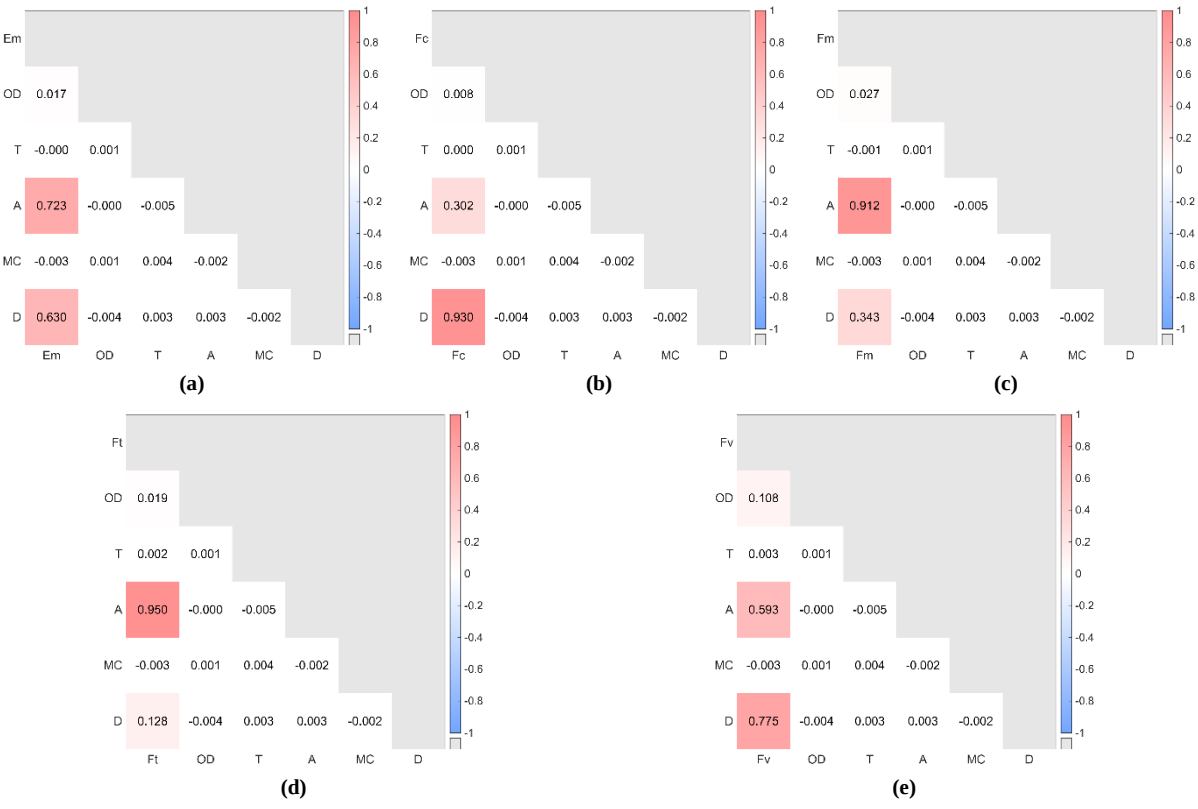


Figure 9. Correlation matrix for $N = 100,000$ using PB: (a) E_m; (b) F_c; (c) F_m; (d) F_t; and (e) F_v

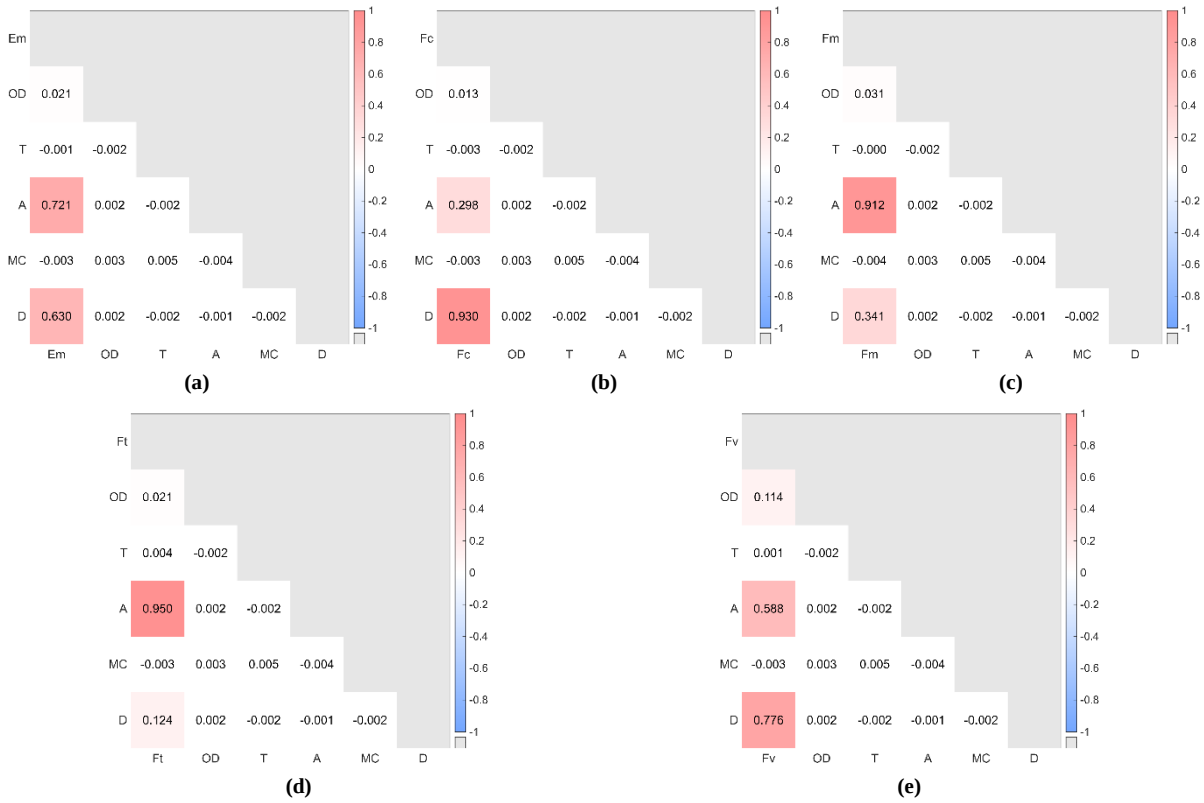


Figure 10. Correlation matrix for $N = 100,000$ using GC: (a) E_m; (b) F_c; (c) F_m; (d) F_t; and (e) F_v

It is noted that, based on the correlation values obtained from the literature as shown in Table 3, density (D), wall thickness (T), and outer diameter (OD) generally exhibit the highest correlations with the mechanical properties, with cross-sectional area (A) following these variables. However, in the synthetic datasets, cross-sectional area appears as one of the dominant correlated variables alongside density. Since cross-sectional area is a function of wall thickness, and outer diameter, this shift suggests that geometric effects may be implicitly captured through A in the generated datasets. If experimental datasets become available, it would be valuable to examine the correlations among A, T, and OD, and to assess whether including all these variables simultaneously in machine learning models may introduce multicollinearity.

3.2.3. Relative Error Analysis of Summary Statistics

The fidelity of the synthetic datasets was further quantified using relative errors (RE) for the mean, standard deviation, and coefficient of variation. Heatmaps were constructed to visualize the relative errors across synthetic data generation methods and sample sizes for each statistic, as shown in Figure 11.

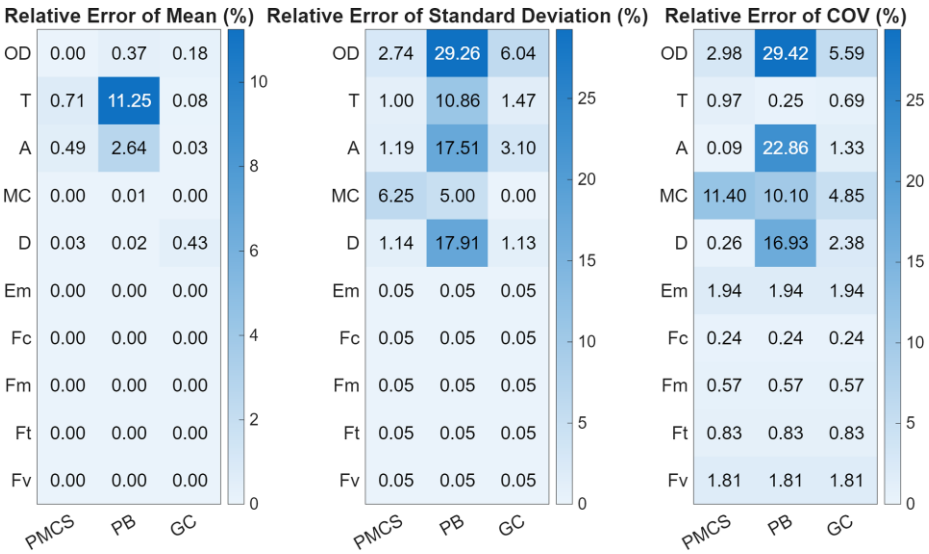


Figure 11. Comparison of relative errors across synthetic data generation techniques for $N = 100,000$

An important observation was that the relative error values remained identical across sample sizes ($N = 1{,}000; 10{,}000;$ and $100{,}000$) for a given data generation method and variable. This behavior is expected and directly linked to the data generation framework employed in this study. In all three synthetic data generation techniques, the target means and standard deviations were explicitly imposed during generation or scaling. As a result, increasing the sample size improves the numerical stability of the estimates but does systematically shift their expected values. Consequently, once a sufficiently large sample is reached, the computed statistics converge to the prescribed values, yielding nearly identical relative errors across different N.

This indicates that the relative error metric primarily reflects methodological bias rather than sampling variability. Differences observed between parametric Monte Carlo simulation, parametric bootstrapping, and Gaussian copula sampling therefore arise from the intrinsic properties of each method, such as resampling effects or dependence enforcement, rather than from dataset size alone. Sample size mainly influences smoothness, numerical stability, and tail representation, which are better captured through distributional analysis than through summary statistics.

Despite the observed discrepancies in fitted probability distributions as presented in Section 3.2.1., the relative errors of the summary statistics remain minimal because the data generation procedures explicitly enforce the target mean and standard deviation. As a result, even when distributional forms differ under goodness-of-fit testing, the central statistical properties are preserved, leading to stable and low relative error values.

Overall, all three synthetic data generation techniques demonstrate acceptable statistical fidelity with respect to the target descriptive statistics. Parametric Monte Carlo simulation exhibited the most consistent agreement with assumed marginal distributions, while parametric bootstrapping showed greater irregularity. The Gaussian copula approach achieved strong marginal fidelity while additionally preserving multivariate dependence.

3.3. ANN Model Performance

The predictive performance of the developed ANN models was evaluated using the metrics discussed in section 2.5.2. Model performance was assessed separately for the training, validation, and testing datasets to ensure robustness and to identify potential overfitting. The ANN models were trained using synthetically generated datasets with random data partitioning to ensure unbiased selection of samples for each learning phase. Hyperparameters, including network architecture and training algorithms, were optimized through grid search, as detailed in Section 2.4.

Tables 7 to 11 summarize the ANN performance metrics for the training, validation, and testing subsets. Overall, the results indicate stable and consistent predictive capability of the ANN models across all data partitions. The close agreement between training and validation errors suggests that the networks can be able to generalize effectively beyond the data used for calibration.

Table 7. Performance of the ANN model for F_c for $N = 100,000$

Phase	Performance Metrics	Parametric Monte Carlo Simulation		Parametric Bootstrapping		Gaussian Copula	
		Without FS	With FS	Without FS	With FS	Without FS	With FS
Training	$RMSE$	2.8994	2.8996	2.9158	2.9159	2.9139	2.9139
	MAE	2.3167	2.3166	2.3255	2.3256	2.3281	2.3281
	MSE	8.4064	8.4077	8.5020	8.5023	8.4906	8.4908
	$MAPE$	3.7787	3.7776	3.7924	3.7912	3.7966	3.7957
	R	0.9771	0.9771	0.9769	0.9769	0.9769	0.9769
	R^2	0.9548	0.9548	0.9543	0.9543	0.9543	0.9543
Validation	$RMSE$	2.8939	2.8928	2.9350	2.9346	2.9076	2.9069
	MAE	2.3074	2.3063	2.3424	2.3420	2.3181	2.3177
	MSE	8.3749	8.3684	8.6142	8.6116	8.4540	8.4502
	$MAPE$	3.7775	3.7741	3.8295	3.8277	3.7788	3.7772
	R	0.9773	0.9773	0.9762	0.9762	0.9766	0.9766
	R^2	0.9552	0.9552	0.9530	0.9531	0.9537	0.9537
Testing	$RMSE$	2.9177	2.9167	2.9175	2.9160	2.9312	2.9301
	MAE	2.3263	2.3254	2.3261	2.3247	2.3367	2.3356
	MSE	8.5132	8.5069	8.5120	8.5030	8.5920	8.5852
	$MAPE$	3.7926	3.7894	3.7908	3.7875	3.8164	3.8139
	R	0.9764	0.9764	0.9766	0.9767	0.9766	0.9767
	R^2	0.9534	0.9534	0.9538	0.9538	0.9538	0.9539

Table 8. Performance of the ANN model for F_t for $N = 100,000$

Phase	Performance Metrics	Parametric Monte Carlo Simulation		Parametric Bootstrapping		Gaussian Copula	
		Without FS	With FS	Without FS	With FS	Without FS	With FS
Training	$RMSE$	9.0048	9.0098	8.9076	8.9136	8.9481	8.9503
	MAE	7.1954	7.1978	7.1044	7.1102	7.1501	7.1507
	MSE	81.0871	81.1760	79.3451	79.4515	80.0678	80.1080
	$MAPE$	7.1477	7.1481	7.0788	7.0817	7.1367	7.1321
	R	0.9579	0.9578	0.9584	0.9583	0.9586	0.9586
	R^2	0.9175	0.9174	0.9185	0.9184	0.9189	0.9188
Validation	$RMSE$	8.9909	8.9877	8.9684	8.9660	8.9292	8.9335
	MAE	7.1683	7.1667	7.1554	7.1518	7.1194	7.1238
	MSE	80.8371	80.7785	80.4315	80.3885	79.7303	79.8067
	$MAPE$	7.1443	7.1432	7.1208	7.1144	7.0844	7.0833
	R	0.9574	0.9574	0.9588	0.9588	0.9577	0.9577
	R^2	0.9166	0.9166	0.9192	0.9192	0.9172	0.9171
Testing	$RMSE$	9.0604	9.0666	8.9140	8.9121	9.0028	8.9992
	MAE	7.2242	7.2306	7.1077	7.1042	7.1767	7.1764
	MSE	82.0916	82.2036	79.4600	79.4254	81.0512	80.9852
	$MAPE$	7.2095	7.2143	7.0617	7.0563	7.1150	7.1129
	R	0.9569	0.9569	0.9591	0.9591	0.9570	0.9570
	R^2	0.9157	0.9156	0.9199	0.9199	0.9158	0.9158

Table 9. Performance of the ANN model for F_m for $N = 100,000$

Phase	Performance Metrics	Parametric Monte Carlo Simulation		Parametric Bootstrapping		Gaussian Copula	
		Without FS	With FS	Without FS	With FS	Without FS	With FS
Training	$RMSE$	4.6409	4.6414	4.5958	4.5963	4.6173	4.6174
	MAE	3.7081	3.7083	3.6656	3.6664	3.6894	3.6893
	MSE	21.5380	21.5427	21.1215	21.1258	21.3194	21.3201
	$MAPE$	4.4627	4.4633	4.4181	4.4189	4.4504	4.4479
	R	0.9739	0.9739	0.9741	0.9741	0.9742	0.9742
	R^2	0.9485	0.9485	0.9489	0.9489	0.9491	0.9491
Validation	$RMSE$	4.6333	4.6317	4.6249	4.6237	4.6082	4.6080
	MAE	3.6941	3.6934	3.6905	3.6884	3.6743	3.6741
	MSE	21.4672	21.4525	21.3900	21.3782	21.2353	21.2341
	$MAPE$	4.4611	4.4614	4.4462	4.4433	4.4233	4.4210
	R	0.9736	0.9736	0.9743	0.9743	0.9737	0.9737
	R^2	0.9479	0.9479	0.9492	0.9492	0.9481	0.9481
Testing	$RMSE$	4.6694	4.6710	4.5973	4.5958	4.6452	4.6433
	MAE	3.7233	3.7248	3.6649	3.6638	3.7029	3.7021
	MSE	21.8029	21.8187	21.1356	21.1216	21.5776	21.5601
	$MAPE$	4.4871	4.4887	4.4075	4.4063	4.4524	4.4495
	R	0.9731	0.9731	0.9747	0.9747	0.9733	0.9733
	R^2	0.9469	0.9469	0.9500	0.9500	0.9473	0.9473

Table 10. Performance of the ANN model for F_v for $N = 100,000$

Phase	Performance Metrics	Parametric Monte Carlo Simulation		Parametric Bootstrapping		Gaussian Copula	
		Without FS	With FS	Without FS	With FS	Without FS	With FS
Training	$RMSE$	0.5197	0.5200	0.5195	0.5196	0.5201	0.5201
	MAE	0.4153	0.4154	0.4143	0.4144	0.4156	0.4155
	MSE	0.2701	0.2704	0.2698	0.2700	0.2705	0.2705
	$MAPE$	4.1247	4.1237	4.1167	4.1181	4.1342	4.1311
	R	0.9807	0.9806	0.9806	0.9806	0.9806	0.9806
	R^2	0.9617	0.9617	0.9616	0.9615	0.9617	0.9617
Validation	$RMSE$	0.5187	0.5186	0.5228	0.5227	0.5189	0.5191
	MAE	0.4136	0.4134	0.4172	0.4170	0.4138	0.4139
	MSE	0.2691	0.2689	0.2733	0.2732	0.2693	0.2695
	$MAPE$	4.1624	4.1576	4.1593	4.1572	4.1009	4.1009
	R	0.9806	0.9806	0.9803	0.9803	0.9803	0.9802
	R^2	0.9617	0.9617	0.9610	0.9610	0.9609	0.9609
Testing	$RMSE$	0.5229	0.5231	0.5196	0.5195	0.5232	0.5229
	MAE	0.4169	0.4171	0.4143	0.4141	0.4171	0.4169
	MSE	0.2734	0.2736	0.2700	0.2699	0.2737	0.2735
	$MAPE$	4.1562	4.1579	4.1083	4.1068	4.1507	4.1445
	R	0.9799	0.9799	0.9807	0.9807	0.9803	0.9803
	R^2	0.9603	0.9602	0.9617	0.9618	0.9609	0.9609

Table 11. Performance of the ANN model for E_m for $N = 100,000$

Phase	Performance Metrics	Parametric Monte Carlo Simulation		Parametric Bootstrapping		Gaussian Copula	
		Without FS	With FS	Without FS	With FS	Without FS	With FS
Training	$RMSE$	1233.47	1233.50	1226.68	1226.76	1230.83	1230.83
	MAE	985.5638	985.5453	978.3438	978.5581	983.4581	983.4022
	MSE	1521460	1521532	1504737	1504941	1514949	1514938
	$MAPE$	5.2036	5.2010	5.1707	5.1711	5.1984	5.1983
	R	0.9578	0.9578	0.9580	0.9580	0.9581	0.9581
	R^2	0.9175	0.9175	0.9177	0.9177	0.9179	0.9179
Validation	$RMSE$	1231.26	1230.67	1234.71	1234.17	1228.43	1228.40
	MAE	981.6467	981.0880	985.2389	984.5868	979.4587	979.4327
	MSE	1515996	1514549	1524521	1523165	1509048	1508963
	$MAPE$	5.2691	5.2580	5.2114	5.2069	5.1691	5.1697
	R	0.9577	0.9577	0.9578	0.9578	0.9571	0.9571
	R^2	0.9172	0.9173	0.9174	0.9175	0.9161	0.9161
Testing	$RMSE$	1241.00	1241.07	1227.12	1226.67	1238.24	1237.91
	MAE	989.4029	989.6101	978.2783	977.9818	987.1693	986.9211
	MSE	1540090	1540255	1505836	1504716	1533228	1532418
	$MAPE$	5.2251	5.2240	5.1632	5.1608	5.2122	5.2111
	R	0.9562	0.9562	0.9586	0.9586	0.9570	0.9570
	R^2	0.9144	0.9144	0.9189	0.9189	0.9158	0.9158

Although ANN models were developed using synthetic datasets with varying sample sizes, only the results corresponding to $N = 100{,}000$ were presented to streamline the presentation. Notably, the performance metrics scores remain within acceptable ranges for all sample sizes considered. Although increasing the number of samples does not consistently lead to a reduction in training errors, this behavior is not unexpected. Once the ANN has sufficiently learned the underlying statistical and nonlinear structure embedded in the synthetic data, additional samples primarily serve to reinforce already learned patterns rather than improve fitting accuracy. This observation indicates model convergence and learning saturation, rather than underfitting or instability. More importantly, the absence of a large discrepancy between training and testing errors further suggests that overfitting was minimal. The final ANN architecture is presented in Table 12.

Table 12. ANN architectures for $N = 100,000$

	Number of Neurons in the Hidden Layer/s						Training Function					
	Without FS			With FS			Without FS			With FS		
	PMCS	PB	GC	PMCS	PB	GC	PMCS	PB	GC	PMCS	PB	GC
F_c	[10]	[10]	[5]	[8]	[8 5]	[8 5]	LM	LM	BR	BR	LM	BR
F_t	[8]	[10]	[8 5]	[8 5]	[8 5]	[8 5]	LM	LM	LM	LM	BR	LM
F_m	[8]	[5]	[8 5]	[8 5]	[8 5]	[8 5]	LM	LM	LM	LM	BR	BR
F_v	[5]	[5]	[8]	[8]	[8]	[8 5]	BR	LM	LM	BR	BR	BR
E_m	[8 5]	[10]	[5]	[8 5]	[8 5]	[8 5]	BR	LM	BR	BR	BR	LM

Regression plots for the training, validation, testing, and overall datasets were presented in Figures 12 to 14. These plots show a strong linear relationship between predicted and target values. The high values of R and R^2 obtained in all datasets confirm the ANN's ability to capture the latent nonlinear relationship between the input variables and the predicted material properties.

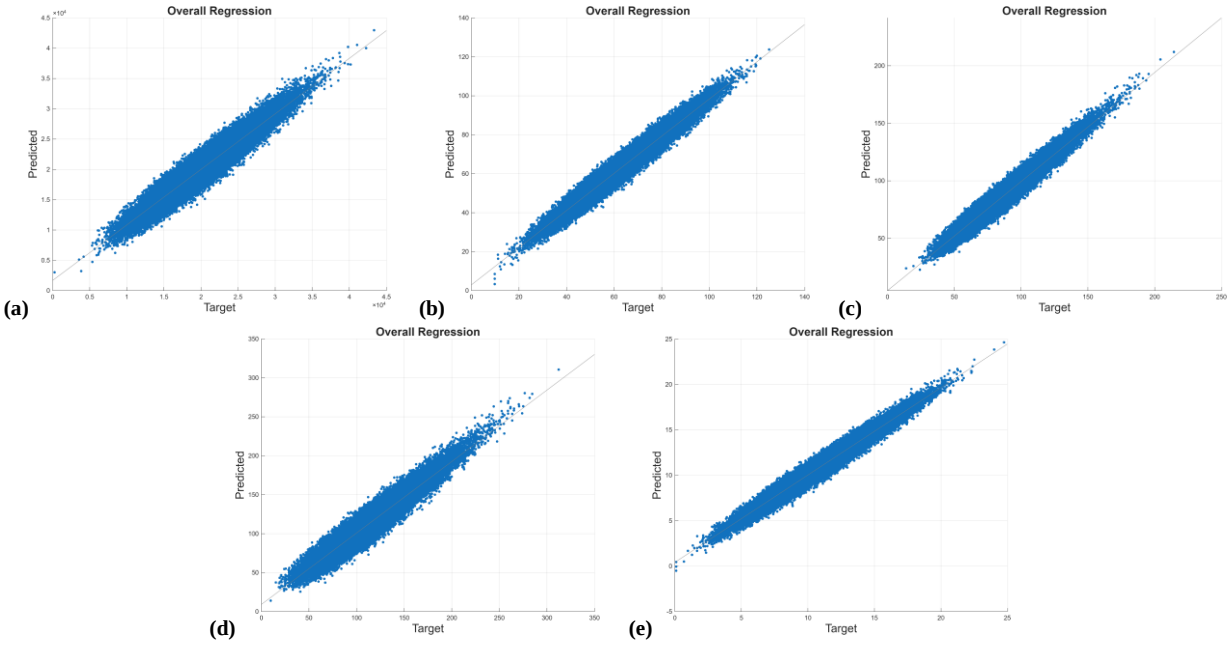


Figure 12. Regression plots for $N = 100,000$ using Parametric Monte Carlo simulation without FS: (a) E_m; (b) F_c; (c) F_m; (d) F_t; (e) F_v

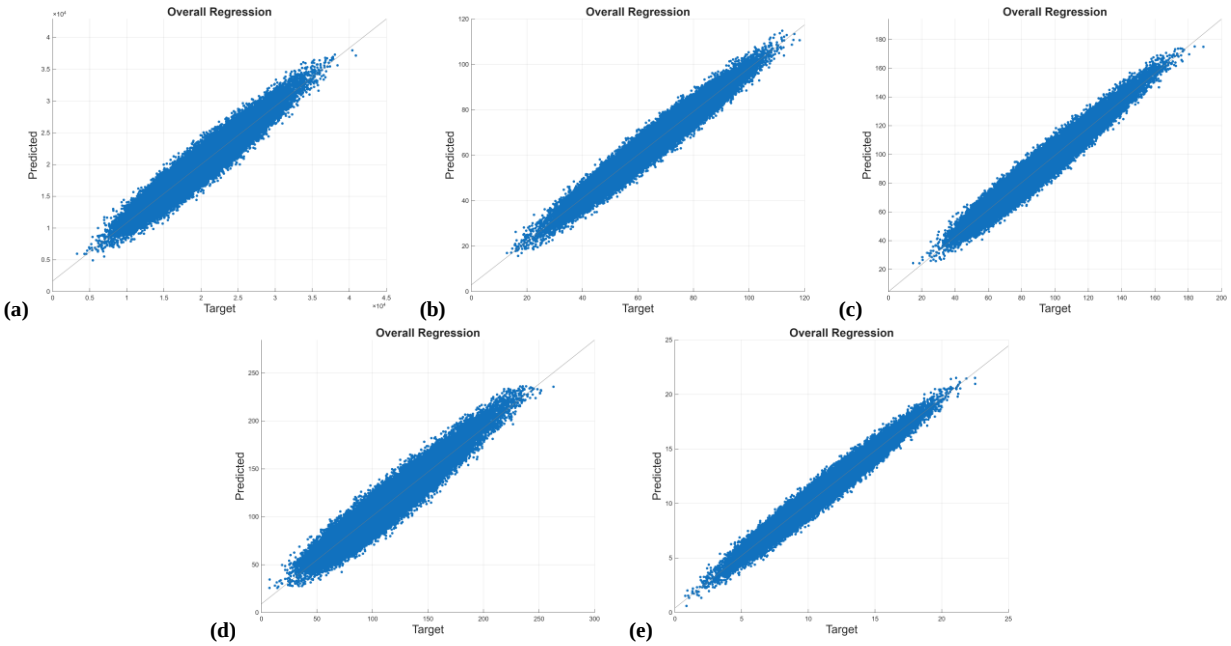


Figure 13. Regression plots for $N = 100,000$ using Parametric bootstrapping without FS: (a) E_m; (b) F_c; (c) F_m; (d) F_t; (e) F_v

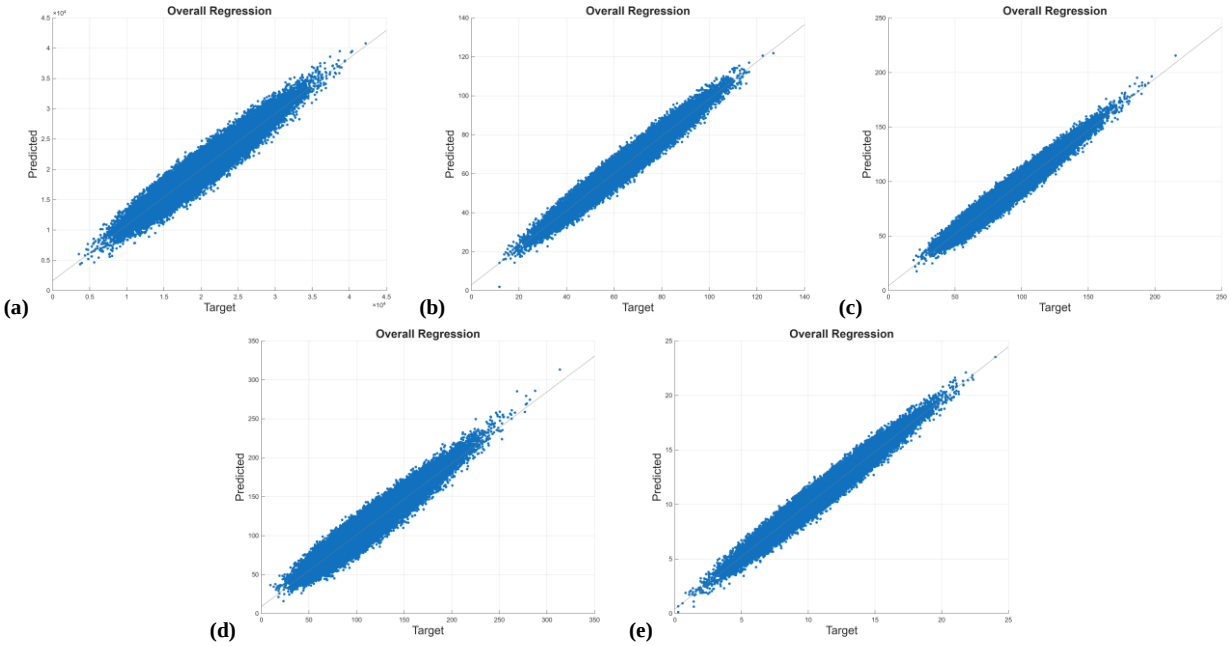


Figure 14. Regression plots for $N = 100,000$ using Gaussian copula without FS: (a) E_m; (b) F_c; (c) F_m; (d) F_t; (e) F_v

3.4. Feature Importance Analysis

Permutation feature importance analysis revealed that, in most cases, fewer than three input variables exceeded the predefined importance threshold. To avoid excessive feature reduction and preserve sufficient physical representation of the system, a minimum of three features was enforced in the feature-selected models.

A comparative analysis between ANN models trained with and without feature selection indicates that feature selection generally led to improved predictive performance. Models incorporating selected features exhibit lower error metrics and marginally higher R^2 values across validation and testing datasets. This improvement suggests that removing weak or redundant input variables helps reduce noise propagation, enhances learning efficiency, and improves generalization.

Table 13 presents the ranking of input variables based on the computed feature importance scores. Across all cases, the most influential predictors were consistently either density (D) or cross-sectional area (A), which agrees with the trends observed in the correlation matrices.

Table 13. Ranking based on feature importance scores using $N = 100,000$

Features (Variables)	Parametric Monte Carlo Simulation					Parametric Bootstrapping					Gaussian Copula				
	F_c	F_t	F_m	F_v	E_m	F_c	F_t	F_m	F_v	E_m	F_c	F_t	F_m	F_v	E_m
OD	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
T	4	4	4	4	4	4	4	4	4	4	4	4	4	4	5
A	2	1	1	2	1	2	1	1	2	1	2	1	1	2	1
MC	5	5	5	4	5	5	5	5	5	5	5	5	5	5	4
D	1	2	2	1	2	1	2	2	1	2	1	2	2	1	2

The physical properties considered in this study were selected based on their established relevance to the mechanical performance of bamboo reported in the literature [5, 30–33]. These parameters describe the material composition, geometric characteristics, and moisture condition of the culm, all of which influence the load transfer mechanism and failure behavior. Interpreting feature importance within these physical properties enables assessment of whether the ANN learns meaningful relationships rather than artifacts introduced by synthetic data generation.

3.4.1. Effects of Density

Density reflects how compact and solid the bamboo material is within the culm wall. A higher density generally indicates a greater concentration of fibers and stronger bonding within the cellular structure. As a result, dense bamboo tends to exhibit improved load-bearing capacity and is less susceptible to localized damage and stress concentrations. The presence of more fiber enhances internal stress redistribution and resistance to crack initiation.

In contrast, lower-density bamboo typically contains a higher proportion of voids and weaker intercellular interfaces, which can reduce its ability to sustain mechanical loading before degradation occurs. Because mechanical strength is closely linked to how effective stresses are distributed throughout the material, density serves as a physically meaningful and reliable predictor of multiple mechanical properties.

3.4.2. Effects of Area

The cross-sectional area governs how applied forces are distributed within the bamboo culm and directly influences its overall load-carrying capacity. A larger cross-sectional area provides more material to resist applied loads, thereby reducing average stress levels and limiting the development of critical stress concentrations.

Since stress is defined as force divided by area, smaller sections experience higher stresses under the same loading conditions, while larger sections can sustain greater loads before reaching failure. Consequently, the cross-sectional area plays a fundamental role in regulating stress magnitude and contributes significantly to the mechanical resistance of bamboo.

3.4.3. Effects of Outer Diameter

Outer diameter describes the overall size of the bamboo culm and affects how material is distributed around the section perimeter. A larger diameter allows applied forces to be spread over a wider perimeter, which can reduce peak stress intensity and promote a more uniform stress distribution. This geometric characteristic also influences the position of fibers relative to the centroid, affecting stress paths and deformation behavior.

However, the influence of outer diameter cannot be considered independently of wall thickness. A culm with a large diameter but very thin walls may still have a limited load-resisting area. Therefore, understanding the combined effects and interdependence of geometric properties such as diameter and thickness is essential for accurately interpreting their influence on mechanical performance.

3.4.4. Effects of Thickness

Wall thickness represents the amount of material between the inner and outer surfaces of the bamboo culm. Variations in thickness directly affect the volume of material available to resist applied loads. Thinner walls tend to concentrate stress closer to the surface, increasing susceptibility to localized damage.

In contrast, thicker walls provide a larger resisting region, allowing stress to be distributed more evenly across the section. This improves the culm's ability to withstand higher loads and delays the onset of failure. As such, wall thickness plays a critical role in determining both stiffness and strength characteristics.

3.4.5. Effects of Moisture Content

Moisture content refers to the amount of water present within the bamboo culm and is known to significantly influence mechanical behavior. Changes in moisture level affect the interaction between fibers and the surrounding matrix, as well as the stiffness of the cell walls.

Lower moisture content generally results in increased stiffness and strength but may also lead to more brittle behavior. Conversely, higher moisture content tends to soften the matrix, reducing the effective load-bearing capacity of fibers and decreasing overall mechanical strength. Given its impact on material stiffness, strength, and deformation response, moisture content is an important parameter in modeling bamboo mechanical performance.

3.5. Comparison with Existing Studies

The present study developed artificial neural network (ANN) models to predict multiple mechanical properties of bamboo culms, including compressive strength, tensile strength, bending strength, shear strength, and modulus of elasticity. To the best of the authors' knowledge, existing studies typically focus on predicting a limited number of mechanical properties (commonly one to three). This may be attributed to the constraints associated with experimental testing, which is both time-consuming and costly. As shown in Tables 14 and 15, most published prediction models rely exclusively on experimentally obtained datasets.

Table 14. Comparative analysis for the prediction of bamboo tensile strength

Sources	Data	Method	R^2	Rank
Present study	Synthetic data ($N = 100{,}000$)	Artificial Neural Network with Feature Selection	0.9534	2
Mallik et al. [7]	Experimental data obtained from Mahzuz et al. [34] ($N = 30$)	Extreme Learning Machine	0.9966	1
	Experimental data obtained from Mahzuz et al. [34] ($N = 30$)	Support Vector Regression	0.8007	3
	Experimental data obtained from Mahzuz et al. [34] ($N = 30$)	Artificial Neural Network	0.6678	4

Table 15. Comparative analysis for the prediction of bamboo compressive strength

Sources	Data	Method	R^2	Rank
Present study	Synthetic data ($N = 100{,}000$)	Artificial Neural Network with Feature Selection	0.9534	1
Pilapil et al. [8]	Experimental data ($N = 38$)	Multiple Linear Regression	0.635	3
Buachart et al. [6]	Experimental data ($N = 39$)	Artificial Neural Network	0.667	2
Tangphadungrat et al. [35]	Experimental data ($N = 111$)	Multiple Linear Regression	0.597	4

In contrast, this study utilizes synthetically generated datasets to develop predictive models. This approach addresses the limitation of data scarcity in bamboo research and enables the simultaneous modelling of multiple mechanical properties. Due to differences in scope and available data, only selected properties, specifically tensile strength and compressive strength, were compared with existing studies.

As presented in Table 14, the dataset used by Mallik et al. [7], obtained from Mahzuz et al. [34], consisted of only 30 experimental samples. This highlights the inherent limitation of experimental data availability and reinforces the motivation for exploring synthetic data generation as an alternative approach. Mallik et al. [7] evaluated multiple machine learning methods for predicting tensile strength, including extreme learning machine (ELM), support vector regression (SVR), and ANN. Their results showed that ELM achieved the highest coefficient of determination ($R^2 = 0.9966$), followed by the ANN model developed in the present study, then SVR, and finally the ANN model reported in Mallik et al. [7].

The ANN model developed in the present study achieved an R^2 value of 0.9534, ranking second among the compared methods. Although slightly lower than the ELM model reported by Mallik et al. [7], the performance remains competitive, particularly considering that the present study relies entirely on synthetic data rather than experimental measurements. It is also important to note that direct comparison is limited by differences in dataset size, input variables, and modelling approaches. The present study employed five input variables – outer diameter, density, wall thickness, moisture content, and cross-sectional area – whereas [7, 34] utilized geometric properties specific to tensile test specimens.

For compressive strength prediction, a similar trend was observed as shown in Table 15. Previous studies relied on relatively small experimental datasets, ranging from 38 to 111 samples, and employed either regression-based or ANN models. In comparison, the present study utilized a substantially larger synthetic dataset, which may contribute to improved model stability and predictive performance.

Among the compared studies, the ANN model developed in this study achieved the highest R^2 value (0.9534), followed by the ANN model of Buachart et al. [6] and regression-based models from Pilapil et al. [8] and Tangphadungrat et al. [35]. Despite differences in methodology and input variables, the results demonstrate that synthetic data-driven models can achieve predictive accuracy comparable to, or exceeding, those based on limited experimental datasets.

It is also noteworthy that, although different studies employed varying input variables, several key predictors were consistently identified. In the present study, density and cross-sectional area emerged as the most influential features across all models. Since cross-sectional are a function of outer diameter and wall thickness, these variables align closely with those commonly used in previous studies. This consistency suggests that the developed ANN models capture physically meaningful relationships rather than effects introduced by the synthetic data generation process.

Nevertheless, it is acknowledged that further validation using experimental datasets with matching input variables is necessary to fully establish the generalizability of the proposed models. Such validation would provide additional confidence in the applicability of synthetic data-driven approaches for bamboo material characterization.

4. Conclusion

This study presents a systematic comparison of three synthetic data generation methods, namely parametric Monte Carlo simulation (PMCS), parametric bootstrapping (PB), and Gaussian copula (GC), in predicting bamboo mechanical properties using artificial neural networks. The evaluation was based on two criteria: statistical fidelity of the generated datasets relative to published reference statistics, and ANN predictive performance.

Based on statistical fidelity, PMCS demonstrates the most consistent agreement with the target marginal statistics and assumed probability distributions, exhibiting minimal irregularity across sample sizes. The GC approach achieves comparable marginal fidelity while additionally providing a structured framework that can accommodate multivariate dependence, although input-input independence was assumed in the present implementation. PB shows slightly greater variability at smaller sample sizes due to finite resampling effects but converges toward the reference statistics as sample size increases.

ANN performance evaluation indicates that all three synthetic data generation methods support stable and accurate prediction of strengths and modulus of elasticity. Across all models, training, validation, and testing metric values are closely aligned, indicating good generalization and minimal overfitting. Although the differences are numerically small, PB consistently produces the lowest testing MSE across most output variables, followed closely by GC and PMCS. The reliability of the ANN model was further examined through feature importance analysis. Density emerges as the most influential predictor for compressive and shear strengths, reflecting the governing role of material resistance. Cross-sectional area dominates the prediction of tensile and bending strength and bending modulus, where section geometry is more critical. Crucially, the ranking of influential features remains consistent across three synthetic data generation methods. This consistency shows that the ANN models rely on physically meaningful input variables and that the choice of synthetic data generation method does not distort the learned relationships among variables.

Based on the two criteria, no single synthetic data generation method dominates. PMCS provides the most statistically stable synthetic datasets, while PB yields marginally superior ANN accuracy. GC consistently occupies an intermediate position. Overall, ANN predictive performance is similar for all three methods, and the small differences observed do not indicate a clear performance advantage. These results indicate that all three synthetic data generation methods are suitable for ANN-based prediction of bamboo mechanical properties when experimental data are limited. PMCS is preferable when statistical consistency and distributional control are prioritized, PB may be favored when marginal improvements in ANN accuracy are desired, and GC offers a flexible framework for future extensions involving explicit dependence modeling. These findings confirm that moderate differences in statistical fidelity do not translate into meaningful degradation of ANN predictive capability and support the use of synthetic data as reliable surrogates for experimental datasets in bamboo engineering research.

While synthetic data has proven effective for ANN-based prediction when experimental information is limited, its application remains dependent on assumptions about statistical descriptors and correlation structure. Thus, the next step must be to train the ANN models using experimentally measured bamboo datasets, where relationships among variables are observed directly rather than prescribed. Evaluating ANN performance using experimental data and comparing it with the models trained on synthetic datasets would help clarify the conditions under which synthetic data can reliably supplement laboratory testing.

5. Declarations

5.1. Author Contributions

Conceptualization, M.J.A. and N.A.M.; methodology, A.G., M.J.A., and N.A.M.; software, A.G., M.J.A., and N.A.M.; validation, J.O. and L.E.G.; formal analysis, A.G., M.J.A., and N.A.M.; investigation, A.G., M.J.A., and N.A.M.; resources, J.O. and L.E.G.; data curation, A.G., M.J.A., and N.A.M.; writing—original draft preparation, A.G., M.J.A., and N.A.M.; writing—review and editing, J.O. and L.E.G.; visualization, A.G., M.J.A., and N.A.M.; supervision, J.O. and L.E.G.; project administration, J.O. and L.E.G.; funding acquisition, J.O. and L.E.G. All authors have read and agreed to the published version of the manuscript.

5.2. Data Availability Statement

The data presented in this study are available in the article.

5.3. Funding and Acknowledgements

The authors would like to thank the Department of Science and Technology – Science Education Institute (DOST-SEI) and Engineering Research and Development for Technology (ERDT) for providing funding support.

5.4. Conflicts of Interest

The authors declare no conflict of interest.

6. References

[1] Aniñon, M. J. C., & Garciano, L. E. O. (2024). Advances in Connection Techniques for Raw Bamboo Structures—A Review. Buildings, 14(4), 1126. doi:10.3390/buildings14041126.

[2] Muhammad, N. A. G., Orejudos, J. N., & Aniñon, M. J. C. (2024). A Compendium of Research, Tools, Structural Analysis, and Design for Bamboo Structures. Buildings, 14(8), 2419. doi:10.3390/buildings14082419.

[3] Cacanando, C. J. D., López, L. F., Atienza, E., & Pradhan, N. P. N. (2025). Experimental characterization of mechanical properties of Bambusa blumeana bamboo poles and determination of design values. Construction and Building Materials, 490, 142498. doi:10.1016/j.conbuildmat.2025.142498.

[4] Panti, C. A. T., Cañete, C. S., Navarra, A. R., Rubinas, K. D., Garciano, L. E. O., & López, L. F. (2024). Establishing the Characteristic Compressive Strength Parallel to Fiber of Four Local Philippine Bamboo Species. Sustainability (Switzerland), 16(9), 3845. doi:10.3390/su16093845.

[5] Correal, J. F., Calvo, A. F., Trujillo, D. J. A., & Echeverry, J. S. (2022). Inference of mechanical properties and structural grades of bamboo by machine learning methods. Construction and Building Materials, 354, 129116. doi:10.1016/j.conbuildmat.2022.129116.

[6] Buachart, C., Hansapinyo, C., Sukontasukkul, P., Zhang, H., Sae-Long, W., Chetchotisak, P., & O'Brien, T. E. (2024). Characteristic and allowable compressive strengths of Dendrocalamus Sericeus bamboo culms with/without node using artificial neural networks. Case Studies in Construction Materials, 20, 2794. doi:10.1016/j.cscm.2023.e02794.

[7] Mallik, M., Dubey, S., & Gupta, D. (2024). Machine learning approach to forecast the tensile strength of bamboo. Journal of Electrical Systems, 20, 1526-1538. doi:10.52783/jes.1456.

[8] Pilapil, R. A. E., Ongpeng, J. M., & Valerio, D. N. (2025). Prediction of Mechanical Properties of Bambusa blumeana Bamboo Culms Using Non-Destructive Indicating Properties. Proceedings of International Exchange and Innovation Conference on Engineering & Sciences (IEICES), 11, 1367–1372. doi:10.5109/7395688.

[9] Ramful, R., & Casseem, M. S. (2023). Prediction of the Mechanical Characteristic of Bamboo Using Deep Neural Network. 2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), 1–5. doi:10.1109/ICECCME57830.2023.10253219.

[10] Su, Z., Jiang, Z., Liang, Y., Wang, B., & Sun, J. (2022). An artificial neural network model for predicting mechanical strength of bamboo-wood composite based on layups configuration. BioResources, 17(2), 3265–3277. doi:10.15376/biores.17.2.3265-3277.

[11] You, G., Wang, B., Li, J., Chen, A., & Sun, J. (2022). The prediction of MOE of bamboo-wood composites by ANN models based on the non-destructive vibration testing. Journal of Building Engineering, 59, 105078. doi:10.1016/j.jobe.2022.105078.

[12] Goyal, M., & Mahmoud, Q. H. (2024). A Systematic Review of Synthetic Data Generation Techniques Using Generative AI. Electronics (Switzerland), 13(17), 3509. doi:10.3390/electronics13173509.

[13] Endres, M., Mannarapotta Venugopal, A., & Tran, T. S. (2022). Synthetic Data Generation: A Comparative Study. Proceedings of the 26th International Database Engineered Applications Symposium, 94–102. doi:10.1145/3548785.3548793.

[14] Pathare, A., Mangrulkar, R., Suvarna, K., Parekh, A., Thakur, G., & Gawade, A. (2023). Comparison of tabular synthetic data generation techniques using propensity and cluster log metric. International Journal of Information Management Data Insights, 3(2), 100177. doi:10.1016/j.jjimei.2023.100177.

[15] White, M., & Rozovskaya, A. (2020). A Comparative Study of Synthetic Data Generation Methods for Grammatical Error Correction. Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications. doi:10.18653/v1/2020.bea-1.21.

[16] Thielen, N., Rachinger, B., Schröder, F., Preitschaft, A., Meier, S., Seidel, R., Reinhardt, A., & Franke, J. (2024). Comparative Study on Different Methods to Generate Synthetic Data for the Classification of THT Solder Joints. 2024 1st International Conference on Production Technologies and Systems for E-Mobility (EPTS), 1–6. doi:10.1109/EPTS61482.2024.10586740.

[17] Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning. MIT press, Cambridge, United States.

[18] Aniñon, M. J. C., & Albiento, E. E. M. (2022). Prediction of 28-day Compressive Strength of Concrete at the Job Site using Artificial Neural Network. Mindanao Journal of Science and Technology, 20(1), 177–205. doi:10.61310/mndjsteect.1121.22.

[19] Rubinstein, R. Y., & Kroese, D. P. (2016). Simulation and the Monte Carlo Method. Wiley Series in Probability and Statistics., Hoboken, United States. doi:10.1002/9781118631980.

[20] Efron, B., & Tibshirani, R. J. (1994). An Introduction to the Bootstrap. Chapman and Hall/CRC, New York, United States. doi:10.1201/9780429246593.

[21] Nelsen, R. B. (2006). An introduction to copulas. Springer New York, United States.

[22] Zhou, Q., Tian, J., Liu, P., & Zhang, H. (2021). Test and prediction of mechanical properties of Moso bamboo. Journal of Engineered Fibers and Fabrics, 16, 15589250211066802. doi:10.1177/15589250211066802.

[23] Liu, P., Zhou, Q., Fu, F., & Li, W. (2021). Effect of bamboo nodes on the mechanical properties of p. Edulis (phyllostachys edulis) bamboo. Forests, 12(10), 1309. doi:10.3390/f12101309.

[24] Kissell, R. L. (2021). Algorithmic Trading. Algorithmic Trading Methods, 23–56, Academic Press, New York, United States. doi:10.1016/b978-0-12-815630-8.00002-8.

[25] Kostanek, J., Karolczak, K., Kuliczkowski, W., & Watala, C. (2024). Bootstrap Method as a Tool for Analyzing Data with Atypical Distributions Deviating from Parametric Assumptions: Critique and Effectiveness Evaluation. Data, 9(8), 95. doi:10.3390/data9080095.

[26] Sklar, M. (1959). N-dimensional distribution functions and their margins. Annales de l'ISUP, 8(3), 229-231. (In French).

[27] Kim, J. M. (2025). Integrating Copula-Based Random Forest and Deep Learning Approaches for Analyzing Heterogeneous Treatment Effects in Survival Analysis. Mathematics, 13(10), 1659. doi:10.3390/math13101659.

[28] The MathWorks Inc. (2025). MATLAB R2025b. The MathWorks Inc, Natick, United States.

[29] Muhammad, N. A., & Orejudos, J. (2025). Machine Learning-Based Prediction of Mechanical Properties of Bambusa Blumeana. Proceedings of the 5th International Symposium on Concrete Structures for the Next Generation (CSN2025), 3-4 March, 2025, Manila, Philippines.

[30] Bahtiar, E. T., Imanullah, A. P., Hermawan, D., Nugroho, N., & Abdurachman. (2019). Structural grading of three sympodial bamboo culms (Hitam, Andong, and Tali) subjected to axial compressive load. Engineering Structures, 181, 233–245. doi:10.1016/j.engstruct.2018.12.026.

[31] Liu, P., Zhou, Q., & Tian, J. (2022). A Two-variable Model for Predicting the Effects of Moisture Content and Density on the Mechanical Properties of Phyllostachys edulis Bamboo. BioResources, 17(1), 400–410. doi:10.15376/biores.17.1.400-410.

[32] Bautista, B. E., Garciano, L. E. O., & Lopez, L. F. (2021). Comparative analysis of shear strength parallel to fiber of different local bamboo species in the philippines. Sustainability (Switzerland), 13(15), 8164. doi:10.3390/su13158164.

[33] Javadian, A., Smith, I. F. C., Saeidi, N., & Hebel, D. E. (2019). Mechanical properties of bamboo through measurement of culm physical properties for composite fabrication of structural concrete reinforcement. Frontiers in Materials, 6, 15. doi:10.3389/fmats.2019.00015.

[34] Mahzuz, H. M. A., Ahmed, M., Dutta, J., & Rose, R. H. (n.d.). Determination of Several Properties of a Bamboo of Bangladesh. Journal of Civil Engineering Research, 3(1), 16. doi:10.5923/j.jce.20130301.02.

[35] Tangphadungrat, P., Hansapinyo, C., Buachart, C., Suwan, T., & Limkatanyu, S. (2023). Analysis of Non-Destructive Indicating Properties for Predicting Compressive Strengths of Dendrocalamus sericeus Munro Bamboo Culms. Materials, 16(4). doi:10.3390/ma16041352.