

Machine Learning Prediction of Higher Heating Value in Heterogeneous Biomass: An Evaluation of Synthetic Data Augmentation Strategies

Constance Nakato Nakimuli ^{1, ✉}, Simon Kawuma ², David K. Okot ³, Fred Kaggwa ⁴, Johan De Greef ⁵, Maarten Vanierschot ^{6, 7}

1. Department of Information Technology, Mbarara University of Science and Technology, Mbarara, UGA

2. Department of Software Engineering, Mbarara University of Science and Technology, Mbarara, UGA

3. Department of Physics, Mbarara University of Science and Technology, Mbarara, UGA

4. Department of Computer Science, Mbarara University of Science and Technology, Mbarara, UGA

5. Department of Materials Engineering - Group T Leuven Campus, KU Leuven, Leuven, BEL

6. Department of Mechanical Engineering - Group T Leuven Campus, KU Leuven, Leuven, BEL

7. Materials Science, Innovation and Modelling (MaSIM) Research Focus Area, North-West University, Mmabatho, ZAF

Received: May 20, 2026 | Review began: June 07, 2026 | Review ended: June 16, 2026 | Published: June 24, 2026

© Copyright 2026

This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Machine learning and generative artificial intelligence are increasingly used to predict biomass properties in resource-constrained settings where experimental determination is impractical. Experimental determination of the higher heating value (HHV) via bomb calorimetry remains costly, time-intensive, and largely inaccessible in resource-constrained regions, motivating the development of data-driven alternatives for biomass feedstock screening. This study is the first to evaluate synthetic data augmentation for biomass HHV prediction, moving beyond the visual comparisons used in prior studies toward a quantitative assessment of synthetic data quality. This study evaluates the performance of five machine learning models and examines the impact of synthetic data augmentation on their predictive performance. The models include Support Vector Regression, Bayesian Ridge, Extremely Randomized Trees (Extra Trees), Gradient Boosting, and Extreme Gradient Boosting (XGBoost). All models are trained on a heterogeneous dataset of 315 biomass samples across six feedstock categories obtained from the literature. XGBoost achieved the highest test-set performance, with an R^2 of 0.81 and a root mean square error of 0.94 MJ/kg. The study also evaluates synthetic data generated for biomass HHV prediction using conditional Generative Adversarial Networks (cGAN), Gaussian Copula, and Bootstrap with noise, assessed across fidelity, utility, and privacy metrics. The cGAN demonstrated the most balanced performance, with a Hellinger distance of 0.17, a pairwise correlation difference of 0.09, and a memorization rate of 3.40%. The results establish that augmentation utility is model-dependent. The most notable gains were observed under cGAN augmentation for Gradient Boosting ($\Delta R^2 = +0.028$) and XGBoost ($\Delta R^2 = +0.019$), with XGBoost reaching an R^2 of 0.83. Extra Trees showed consistent improvement under Gaussian Copula augmentation, with the mean absolute error reduced from 0.93 to 0.90 MJ/kg. However, performance deteriorated across all models when synthetic sample sizes exceeded 25-100 samples.

Categories: Ensemble Learning, Predictive Analytics, Machine Learning (ML)

Keywords: machine learning, generative adversarial networks, synthetic data augmentation, higher heating value prediction, agricultural waste

How to cite this article:

Nakimuli C, Kawuma S, Okot D K, et al. (June 24, 2026) Machine Learning Prediction of Higher Heating Value in Heterogeneous Biomass: An Evaluation of Synthetic Data Augmentation Strategies. Cureus J Comput Sci 3 : es44389-026-00133-8. DOI <https://doi.org/10.7759/s44389-026-00133-8>

Introduction

The global shift to low-carbon energy has positioned biomass as a key feedstock in circular economy strategies. Particularly, agricultural residues and other lignocellulosic materials are suitable for thermochemical conversion processes, including combustion, pyrolysis, gasification, and densification [1,2]. The higher heating value (HHV) is the primary parameter governing feedstock selection and process optimization in these systems [3]. However, its experimental determination via bomb calorimetry is resource-intensive and technically demanding. Bomb calorimeters are particularly scarce in sub-Saharan Africa, where they are commercially operated at a high cost per sample [2]. These constraints motivate the development of predictive models that estimate HHV from proximate and ultimate analysis data, which are cheaper, faster, and more widely available, thereby supporting feedstock selection in resource-constrained settings [4,5].

Machine learning (ML) has demonstrated consistent improvements over empirical correlations across diverse biomass types [6,7]. Random Forest and Extra Trees address this challenge by building multiple decision trees on random subsets of data and aggregating their outputs, making them robust to noise and overfitting [2,8,9]. However, Random Forest and Extra Trees treat each tree independently, missing sequential patterns in the prediction error. XGBoost and Gradient Boosting overcome this limitation by building trees sequentially, with each tree fitted to the residual errors of the previous one, and have consistently ranked among the top-performing models in HHV studies [2,8,10]. This sequential approach requires careful tuning, particularly when training data are limited. Support Vector Regression (SVR) takes a different approach by finding a prediction boundary in a mathematically transformed space that tolerates small errors, making it well suited to smaller datasets where other models may overfit [2,11]. Bayesian Ridge Regression offers a regularized linear baseline that is robust to multicollinearity and performs reliably on small datasets.

Several studies have applied these models to biomass HHV prediction. Afolabi et al. [6] tested multiple models on 227 samples from four biomass classes and reported that Random Forest performed best, with a mean absolute error (MAE) of 1.01 MJ/kg. Similarly, Adeleke et al. [2] trained multiple ML models on 200 samples using combined proximate and ultimate analysis features, with XGBoost achieving the highest test R^2 of 0.73. Brandić et al. [7] evaluated ML models across two biomass categories using proximate analysis inputs, with the best-performing model achieving an R^2 of 0.96. Timilsina et al. [11] evaluated Extra Trees, Random Forest, XGBoost, and SVR on waste-derived fuels, with Extra Trees achieving the best performance. Collectively, these studies confirm that ML models can accurately predict HHV; however, training sets rarely exceed a few hundred samples, limiting generalization across heterogeneous multiclass feedstocks [2,6].

However, despite this progress, small effective training sizes remain a key limitation for heterogeneous multiclass biomass datasets [2,6]. Adeleke et al. [2] explicitly recommended exploring Generative Adversarial Networks (GANs) for HHV prediction. A GAN consists of two neural networks: a generator that produces synthetic samples and a discriminator that distinguishes them from real data. This adversarial competition progressively drives the generation of more realistic outputs. GANs have been applied in biomass for structural composition prediction [12], gasification modelling [13], and energy production modelling [14]. Conditional GANs (cGANs) condition the generation process on class labels, enabling targeted synthesis for each feedstock category. Simpler alternatives also exist, including the Gaussian Copula, which generates samples by modelling the joint dependence structure between input variables while preserving inter-variable correlations. Bootstrap with noise resamples observed data with replacement and introduces small perturbations, providing a computationally simple but less statistically rigorous alternative to deep generative approaches.

To the authors' knowledge, no prior study has systematically evaluated synthetic data augmentation for biomass HHV prediction across heterogeneous multiclass feedstocks. GANs have been applied in related biomass contexts, but none of these studies targeted HHV prediction. Okolie et al. [12] used a GAN-based approach for structural composition prediction. Lv et al. [13] applied a GAN for gasification performance prediction. Zhang et al. [14] applied a GAN combined with Model-Agnostic Meta-Learning (GAN-MAML) to predict process-specific outputs across biomass conversion domains, including incineration, pyrolysis, composting, and anaerobic digestion. Augmentation efforts most closely related to HHV prediction have been limited to noise injection on single-feedstock datasets [15], which does not

How to cite this article:

Nakimuli C, Kawuma S, Okot D K, et al. (June 24, 2026) Machine Learning Prediction of Higher Heating Value in Heterogeneous Biomass: An Evaluation of Synthetic Data Augmentation Strategies. *Cureus J Comput Sci* 3 : es44389-026-00133-8. DOI <https://doi.org/10.7759/s44389-026-00133-8>

adequately address multiclass compositional diversity. This study addresses this gap through a systematic evaluation of cGAN, Gaussian Copula, and bootstrap methods with noise augmentation for biomass HHV prediction, assessed under a unified fidelity-utility-privacy framework [16,17]. Existing studies also typically rely on visual comparisons of data distributions [12], which do not provide a comprehensive evaluation of synthetic data quality.

This study advances the field through three novel contributions, each addressing a distinct gap in the existing literature. First, it provides the first systematic evaluation of synthetic data augmentation specifically for biomass HHV prediction across heterogeneous multiclass feedstocks, a task not previously addressed by GAN-based approaches [12-14] or noise injection methods [15] within the biomass domain. Second, synthetic data quality is assessed using a rigorous quantitative fidelity-utility-privacy framework [16,17], extending beyond the visual distributional comparisons used in prior studies [12] toward a comprehensive multidimensional evaluation. Third, the utility of augmentation is demonstrated to be model-dependent and constrained by data volume, with an empirically identified optimal range of 25-100 synthetic samples across all evaluated models and strategies. This finding has not been previously reported for heterogeneous multiclass biomass datasets.

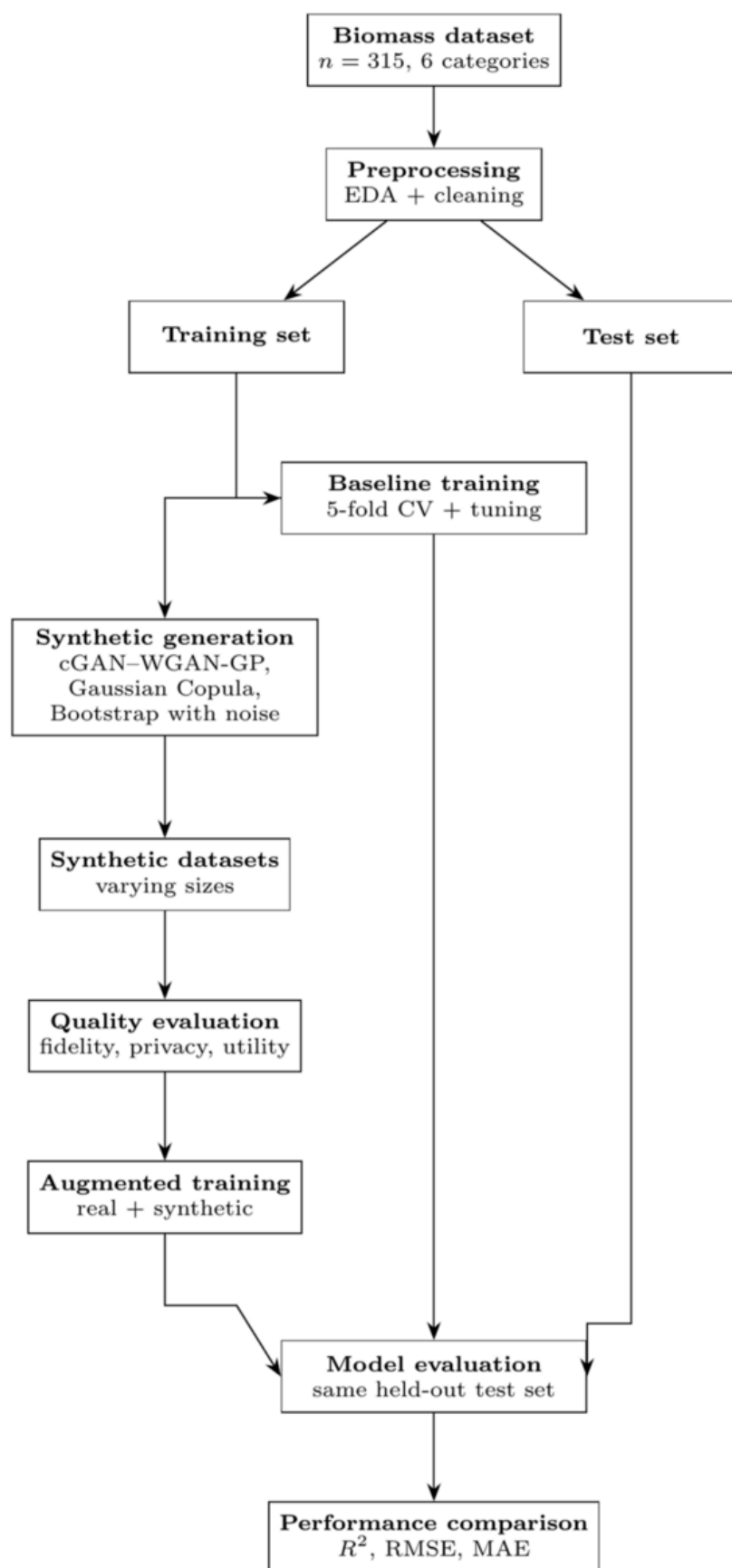
Materials And Methods

Dataset and data preprocessing

A dataset comprising 315 biomass samples was compiled from the literature. The samples were classified according to the TNO [18] taxonomy into six categories: untreated wood (n = 50), grass/plant (n = 60), husk/shell/pit (n = 64), organic residue/product (n = 50), manure (n = 41), and straw/stalk/cob/ear (n = 50). Each sample was characterised using proximate analysis (ash content, volatile matter, and fixed carbon) and ultimate analysis (carbon, hydrogen, oxygen, nitrogen, and sulphur) on a dry basis. The dataset is publicly available on Mendeley Data at <https://doi.org/10.17632/5xswr8yp34.3>. The overall workflow of the study, including preprocessing, model training, synthetic data generation, and evaluation, is presented in Figure 1. The dataset contained missing values only for sulphur (11 samples, 3.5%) and nitrogen (2 samples, 0.6%), which were imputed as zero to reflect their typically negligible or unreported concentrations in biomass [5,18]. Elemental closure analysis confirmed that 8 of the 11 affected samples summed to approximately 100% when sulphur was set to zero, supporting the validity of this imputation approach. The imputed values are clearly indicated in bold in the deposited dataset. Compositional variables were subsequently scaled to maintain mass closure by normalising each feature category to 100%. The cleaned dataset was then partitioned into training and test subsets using an 80:20 ratio to prevent data leakage [19]. Features were Min-Max scaled to the range of 0-1, with the scaler fitted exclusively on the training data to ensure that the test set remained unseen during feature transformation. Feedstock category labels were converted into binary indicator variables through one-hot encoding and used to stratify the train-test split and condition synthetic data generation.

How to cite this article:

Nakimuli C, Kawuma S, Okot D K, et al. (June 24, 2026) Machine Learning Prediction of Higher Heating Value in Heterogeneous Biomass: An Evaluation of Synthetic Data Augmentation Strategies. *Cureus J Comput Sci* 3 : es44389-026-00133-8. DOI <https://doi.org/10.7759/s44389-026-00133-8>



How to cite this article:

Nakimuli C, Kawuma S, Okot D K, et al. (June 24, 2026) Machine Learning Prediction of Higher Heating Value in Heterogeneous Biomass: An Evaluation of Synthetic Data Augmentation Strategies. Cureus J Comput Sci 3 : es44389-026-00133-8. DOI <https://doi.org/10.7759/s44389-026-00133-8>

FIGURE 1: Experimental workflow for biomass HHV prediction, incorporating five baseline regression models (SVR, Bayesian Ridge, Extra Trees, Gradient Boosting, and XGBoost) trained on real and synthetically augmented data generated via cGAN-WGAN-GP, Gaussian Copula, and Bootstrap with noise

cGAN, Conditional Generative Adversarial Network; CV, Cross-Validation; EDA, Exploratory Data Analysis; HHV, Higher Heating Value; MAE, Mean Absolute Error; RMSE, Root Mean Square Error; SVR, Support Vector Regression; WGAN-GP, Wasserstein Generative Adversarial Network with Gradient Penalty

Baseline model development

Five ML models were implemented: SVR, Bayesian Ridge, Extra Trees, Gradient Boosting, and XGBoost. These models were selected for their suitability for nonlinear regression tasks and their prior application in biomass HHV prediction [2,6,8]. Collectively, they represent a diverse range of learning paradigms, spanning kernel-based, probabilistic, and ensemble approaches, enabling a comprehensive comparison of predictive performance across heterogeneous biomass feedstocks.

SVR

SVR was implemented using a radial basis function (RBF) kernel to model nonlinear relationships between the input features and HHV. Here, $\mathbf{w}$ is the regression coefficient vector, C is the regularization parameter, ε is the error tolerance that defines the insensitive loss tube, ξ_i and ξ_i^* are slack variables representing deviations, $\mathbf{x}_i$ is the input vector, $K(\cdot)$ is the kernel function, γ is the kernel parameter, $\mathbf{y}_i$ is the target value, and $\mathbf{b}$ is the bias term. The model optimizes a trade-off between model complexity and prediction error, as expressed in Eq. (1), subject to the constraints in Eq. (2). The nonlinear mapping is defined by the RBF kernel function (Eq. 3).

$$\min \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i (\xi_i + \xi_i^*) \quad (1)$$

$$\text{subject to } \begin{cases} \mathbf{y}_i - \mathbf{w}^T \mathbf{x}_i - \mathbf{b} \leq \varepsilon + \xi_i \\ \mathbf{w}^T \mathbf{x}_i + \mathbf{b} - \mathbf{y}_i \leq \varepsilon + \xi_i^* \end{cases} \quad (2)$$

$$K(\mathbf{x}_i, \mathbf{x}) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}\|^2) \quad (3)$$

Bayesian Ridge Regression

Bayesian Ridge Regression was used as a regularized linear regression model to ensure stable parameter estimation in the presence of multicollinearity and limited data. The model estimates its parameters by optimizing the function presented in Eq. (4):

$$\mathbf{w} = \arg \min_{\mathbf{w}} (\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \lambda \|\mathbf{w}\|^2), \quad \lambda \sim \Gamma(\lambda_1; \lambda_2), \quad \alpha \sim \Gamma(\alpha_1; \alpha_2)$$

where $\mathbf{w}$ is the vector of regression coefficients, $\mathbf{X}$ is the input feature matrix, $\mathbf{y}$ is the target vector, λ is the regularization parameter governed by a Gamma prior with shape λ_1 and rate λ_2 , and α is the noise precision governed by a Gamma prior with shape α_1 and rate α_2 .

Extra Trees

The Extra Trees Regressor was used as an ensemble learning method based on multiple randomized decision trees. The model reduces variance through averaging predictions from independently constructed trees, as shown in Eq. 5:

$$\hat{\mathbf{y}}(\mathbf{x}) = \frac{1}{T} \sum_{k=1}^T T_k(\mathbf{x}) \quad (5)$$

where $\hat{\mathbf{y}}(\mathbf{x})$ is the predicted output, T is the total number of trees, and $T_k(\mathbf{x})$ is the prediction of the decision tree.

How to cite this article:

Nakimuli C, Kawuma S, Okot D K, et al. (June 24, 2026) Machine Learning Prediction of Higher Heating Value in Heterogeneous Biomass: An Evaluation of Synthetic Data Augmentation Strategies. Cureus J Comput Sci 3 : es44389-026-00133-8. DOI <https://doi.org/10.7759/s44389-026-00133-8>

Gradient Boosting Regression

Gradient Boosting Regression was used to iteratively improve prediction accuracy by sequentially adding weak learners that minimize residual errors, where each weak learner is trained based on the residual errors committed by its predecessors. The final model is presented as the sum of all weak learners, as per Eq. 6:

$$\hat{y}(x) = \sum_{k=1}^K T_k(x) \quad (6)$$

where K is the total number of boosting iterations, and $T_k(x)$ represents the contribution of the k^{th} weak learner to the final prediction.

XGBoost

XGBoost was implemented as a regularized gradient boosting framework that optimizes a combined objective function that balances predictive loss and model complexity. The objective function is defined in Eq. 7:

$$J = \sum_{i=1}^n L(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(T_k) \quad (7)$$

where $L(y_i, \hat{y}_i)$ is the loss function measuring prediction error, and $\Omega(T_k)$ is the regularization term defined in Eq. 8:

$$\Omega(T_k) = N + \frac{1}{2} \lambda \sum_{j=1}^N w_j^2 \quad (8)$$

where λ is the regularization parameter, N is the number of leaf nodes, and w_j are the leaf weights.

Model training and evaluation

These ML models were implemented in Python using the scikit-learn library, with hyperparameter optimization performed via RandomizedSearchCV [20]. This technique randomly samples candidate configurations from predefined search ranges, evaluating 15-20 configurations per model across combinations of regularization strength, tree depth, learning rate, and ensemble size, as detailed in Table 7. This approach was adopted because it efficiently identifies well-performing hyperparameter configurations by sampling broadly across the search space rather than testing every possible combination [20]. The search was conducted exclusively on the training set, with the held-out test set reserved solely for final performance evaluation. Each candidate configuration was evaluated using stratified 5-fold cross-validation repeated 10 times, with R^2 used as the scoring metric. In stratified 5-fold cross-validation, the training set is partitioned into five equal folds while preserving the distribution of feedstock categories across each fold. At each iteration, the model is trained on four folds and validated on the held-out fifth fold, ensuring that every sample contributes to validation exactly once per repetition. Repeating this procedure 10 times with different random partitions reduces variance in the performance estimate and improves the reliability of hyperparameter selection under limited-data conditions. The search ranges explored for each model are presented in Table 7. Following hyperparameter selection, each model was refitted on the full training set and evaluated on the held-out test set using R^2 , RMSE, and MAE as performance metrics. To further assess generalization stability, the two highest-ranked models based on cross-validation performance were additionally evaluated across 10 random seed-based train-test splits.

How to cite this article:

Model	Parameter	Search Range/Candidate Values	Selected Value
SVR (RBF)	Kernel type	RBF	RBF
	Regularization strength (C)	10, 30, 50, 100, 200	100
	Kernel width (γ)	scale, auto, 0.01, 0.03, 0.05, 0.10	0.03
	Error tolerance (ϵ)	0.01, 0.05, 0.1, 0.2	0.05
Bayesian Ridge	Noise precision priors (α_1, α_2)	1e-7, 1e-6, 1e-5	(1e-7, 1e-6)
	Weight regularization priors (λ_1, λ_2)	1e-7, 1e-6, 1e-5	(1e-5, 1e-7)
Extra Trees	Number of trees (M)	200, 300, 500	500
	Maximum tree depth	5, 6, 7, 8, 10	7
	Minimum samples per leaf	2, 4, 6, 8	2
	Minimum samples to split	5, 10, 15	10
	Feature subset size	sqrt, log2, 0.6	0.6
Gradient Boost	Number of boosting stages (K)	100, 200, 300	300
	Learning rate	0.01, 0.03, 0.05, 0.1	0.01
	Maximum tree depth	2, 3, 4	4
	Subsample ratio	0.7, 0.8, 0.9	0.7
	Minimum samples per leaf	2, 4, 6	2
XGBoost	Number of trees (K)	100, 200, 300	200
	Maximum tree depth	2, 3, 4	4
	Learning rate	0.01, 0.03, 0.05, 0.1	0.05
	Subsample ratio	0.6, 0.7, 0.8	0.8
	Feature sampling ratio	0.6, 0.7, 0.8	0.8
	L1 regularization (α)	0.1, 0.5, 1.0, 2.0	0.5

How to cite this article:

Nakimuli C, Kawuma S, Okot D K, et al. (June 24, 2026) Machine Learning Prediction of Higher Heating Value in Heterogeneous Biomass: An Evaluation of Synthetic Data Augmentation Strategies. Cureus J Comput Sci 3 : es44389-026-00133-8. DOI <https://doi.org/10.7759/s44389-026-00133-8>

	L2 regularization (λ)	2.0, 5.0, 10.0	5
--	---------------------------------	----------------	---

TABLE 1: Hyperparameter tuning configuration and optimal parameter selection for the implemented machine learning models
RBF, Radial Basis Function; SVR, Support Vector Regression

Synthetic data generation methods

Three synthetic data generation approaches were considered, including a conditional Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP), Gaussian Copula, and Bootstrap with noise. These methods represent deep generative, probabilistic, and resampling-based approaches, respectively.

Synthetic data generation using a cGAN

A cGAN based on the WGAN-GP framework was implemented with physics-informed loss terms (mass closure, correlation preservation, and HHV-driver alignment) following [12,14,21-25]. The generator G maps a latent vector and condition to synthetic samples $\tilde{x} = G(z, c)$, where c is the one-hot-encoded biomass category. The discriminator D distinguishes between real and generated samples conditioned on c .

The WGAN-GP formulation was used to improve training and avoid mode collapse. It uses the Wasserstein distance with a gradient penalty [21,22]. The discriminator loss is defined as:

$$L_D = \mathbb{E}[D(\tilde{x}, c)] - \mathbb{E}[D(x, c)] + \lambda_{GP} \mathbb{E}[(\|\nabla D(\hat{x}, c)\|_2 - 1)^2]$$
 (9)

where D is the discriminator, $\tilde{x}$ is the generated sample, x is the real sample, c is the biomass category condition, λ_{GP} is the gradient penalty coefficient, $\hat{x}$ is an interpolated sample between real and generated data, and ∇ is the gradient operator.

The generator was optimised using:

$$L_G = L_{adv} + \lambda_{fm} L_{fm} + \lambda_{pair} L_{pair} + \lambda_{moment} L_{moment} + \lambda_{driver} L_{driver}$$
 (10)

where L_{adv} is the adversarial loss, L_{fm} is the feature matching loss, L_{pair} enforces correlation preservation, L_{moment} aligns feature distributions, and L_{driver} preserves relationships with the HHV.

The architecture used in this study is illustrated in Table 2. The generator and discriminator were implemented as fully connected layers with Layer Normalization, LeakyReLU activation, and dropout. Synthetic datasets of increasing sizes were generated, and the optimal size was selected at the point of diminishing improvement in fidelity metrics [17]. After training, generated samples were inverse-transformed and adjusted to satisfy compositional constraints.

Component	Parameter	Value
Generator	Latent dimension	64
	Hidden layer size	384
	Dropout	0.08
Discriminator	Hidden layers (h1-h2)	384-192
	Dropout	0.08
Training	Generator learning rate	2×10^{-5}
	Discriminator learning rate	6×10^{-5}
	Batch size	32
	Epochs	8,000
	Critic updates (n_critic)	5
	Gradient penalty (λ_{GP})	10.0
Loss weights	λ_{fm} (feature matching)	1.5
	λ_{pair} (correlation)	0.30
	λ_{moment} (moments)	0.20
	λ_{driver} (HHV drivers)	0.35
	$\lambda_{closure}$ (proximate)	0.001
Input	Joint input dimension	10
	Category conditioning	6 one-hot classes

TABLE 2: Architecture and training configuration of the conditional WGAN-GP, including generator and discriminator layer specifications, optimiser settings, loss weight assignments, and input dimensionality

HHV, Higher Heating Value; WGAN-GP, Wasserstein Generative Adversarial Network with Gradient Penalty

How to cite this article:

Baseline augmentation methods

Two baseline methods, namely the Gaussian Copula and Bootstrap with noise, were used for comparison. The Gaussian Copula models dependencies between variables using a parametric correlation structure. Bootstrap with noise generates synthetic samples by resampling the training data with replacement and adding small perturbations.

Synthetic data evaluation

Quality assessment of the synthetic data was conducted using a multidimensional framework based on fidelity, utility-related similarity, and privacy risk, as recommended in the synthetic tabular data literature [16].

Fidelity

Fidelity evaluates how well synthetic data replicates the statistical properties of the real dataset and was assessed using multiple metrics, including the Hellinger distance, which measures similarity between feature distributions, and the pairwise correlation difference (PCD), which assesses preservation of inter-feature relationships [26,27]. The Wasserstein distance was also computed to measure differences in distributional shape and support across features. The Jensen-Shannon (JS) divergence was used as a symmetric and bounded measure of distributional similarity, ensuring consistent comparison between real and synthetic data distributions.

$$JS(P\|Q) = \frac{1}{2}KL(P\|M) + \frac{1}{2}KL(Q\|M), \quad M = \frac{1}{2}(P + Q) \quad (11)$$

where P and Q denote the probability distributions of real and synthetic data, respectively, M is their average distribution, and KL denotes the Kullback-Leibler divergence.

Privacy

Privacy was evaluated using the distance to closest record (DCR), memorization rate, and classifier-based distinguishability. DCR is defined as the average distance between synthetic data points and their nearest real counterparts [16]. Memorization was evaluated as the proportion of synthetic samples that closely replicate individual training records. Lower values for both indicate a reduced risk of reproducing original data points. The area under the receiver operating characteristic curve (AUC-ROC) was used to assess how well a classifier can distinguish between synthetic samples and real data points [16]. An AUC-ROC score of 0.5 indicates that the classifier fails to discriminate between the two datasets, meaning that the synthetic data are statistically indistinguishable from the real data [16]. The higher the AUC-ROC score, the greater the distinguishability of the synthetic data.

Utility

Utility was evaluated by assessing the impact of synthetic data augmentation on downstream predictive performance. The augmented datasets were used to train ML models, and performance was compared with models trained on real data using metrics including R^2 , RMSE, and MAE. To ensure a fair comparison, augmented models used identical hyperparameters to the baseline model, with performance evaluated on the same held-out test set. To assess whether the choice of augmentation strategy significantly affected model performance, a Friedman test was conducted comparing test R^2 values across four conditions: models trained on real data only and those trained on data augmented by cGAN, Gaussian Copula, and Bootstrap with noise, evaluated across five ML models. The Friedman test was selected as it is a non-parametric test suitable for comparing multiple conditions across the same set of subjects without assuming a normal distribution.

Results And Discussion

Experimental setup

This section presents the results of the descriptive analysis, baseline model performance, synthetic data generation and quality evaluation, and downstream augmentation utility. The proposed experimental framework was implemented in Python 3.7 using scikit-learn 1.7.2, XGBoost 3.2.0, NumPy 2.2.6, Pandas 2.3.3, SHAP 0.49.1, Matplotlib, Seaborn, SciPy, and

How to cite this article:

Nakimuli C, Kawuma S, Okot D K, et al. (June 24, 2026) Machine Learning Prediction of Higher Heating Value in Heterogeneous Biomass: An Evaluation of Synthetic Data Augmentation Strategies. Cureus J Comput Sci 3 : es44389-026-00133-8. DOI <https://doi.org/10.7759/s44389-026-00133-8>

the Synthetic Data Vault package. The modelling was conducted on a heterogeneous dataset of 315 biomass samples spanning six feedstock categories, characterized by eight proximate and ultimate analysis features, including ash, volatile matter, fixed carbon, carbon, hydrogen, oxygen, nitrogen, and sulphur, with HHV as the prediction target. The dataset was partitioned into training and test subsets using an 80:20 ratio. Five ML models were trained and evaluated, namely SVR, Bayesian Ridge, Extra Trees, Gradient Boosting, and XGBoost. Model performance was assessed using R^2 , RMSE, and MAE. The cGAN was implemented using PyTorch with the WGAN-GP framework and trained exclusively on the training subset, generating 700 synthetic samples, of which 695 were retained after quantile filtering. Gaussian Copula and Bootstrap with noise were similarly fitted on the training subset only to prevent data leakage, and synthetic data quality was assessed across fidelity, privacy, and utility dimensions. Augmented datasets of varying sizes, ranging from 25 to more than 400 synthetic samples, were evaluated, with the optimal augmentation size selected based on cross-validation performance. All computations were executed on the Vlaams Supercomputer Centrum Tier-2 High-Performance Computing infrastructure at KU Leuven, Belgium, using a single node with 8 CPU cores and 32 GB RAM, with a fixed random seed of 7 applied throughout to ensure reproducibility.

Descriptive statistics

The HHV ranged between 4.22 and 25.29 MJ/kg, with a mean value of 17.75 ± 2.45 MJ/kg. Among the compositional variables, ash exhibited the highest variability, with a standard deviation of 9.52% by weight on a dry basis, while carbon showed relatively lower dispersion (standard deviation = 6.50% on a dry basis). These descriptive statistics are presented in Table 3.

Correlation analysis revealed the relationships between compositional variables and HHV. Carbon demonstrated a strong positive association with HHV ($r = 0.79$), followed by hydrogen ($r = 0.64$), whereas ash exhibited a strong negative relationship ($r = -0.67$). Volatile matter showed a moderate positive correlation ($r = 0.55$), reinforcing its contribution to combustible fractions. These trends are consistent with established thermochemical theory, whereby carbon and hydrogen serve as primary energy carriers, while ash represents inert material that reduces calorific value [4,28]. Comparable patterns have been reported in recent data-driven HHV prediction studies, which consistently identify carbon, hydrogen, and ash as dominant predictors [2,6,8,11]. Other elements, including nitrogen and sulphur, showed weak negative associations.

How to cite this article:

Nakimuli C, Kawuma S, Okot D K, et al. (June 24, 2026) Machine Learning Prediction of Higher Heating Value in Heterogeneous Biomass: An Evaluation of Synthetic Data Augmentation Strategies. *Cureus J Comput Sci* 3 : es44389-026-00133-8. DOI <https://doi.org/10.7759/s44389-026-00133-8>

Variable	Unit	Mean	Std. Dev.	Min	Q1	Median	Q3	Max
Ash	% (dry basis)	7.81	9.52	0.01	1.92	4.25	10.50	73.52
Volatile Matter	% (dry basis)	75.78	9.35	21.33	72.00	77.70	82.00	93.60
Fixed Carbon	% (dry basis)	17.28	6.54	0.67	13.99	16.81	19.96	50.91
Carbon	% (dry basis)	44.77	6.50	12.96	41.70	46.30	49.03	57.00
Hydrogen	% (dry basis)	5.63	0.83	1.45	5.34	5.80	6.10	7.60
Oxygen	% (dry basis)	39.74	6.52	10.08	37.08	41.36	43.80	54.00
Nitrogen	% (dry basis)	1.05	1.20	0.00	0.23	0.56	1.54	7.22
Sulphur	% (dry basis)	0.15	0.27	0.00	0.00	0.02	0.19	2.00
HHV	MJ/kg	17.75	2.45	4.22	16.40	18.11	19.38	25.29

TABLE 3: Descriptive statistics of proximate analysis, ultimate analysis, and higher heating value (HHV) for the biomass dataset (n = 315), reported on a dry basis

Results from Table 4 show that XGBoost achieved the highest baseline test performance ($R^2 = 0.81$; $RMSE = 0.94$ MJ/kg), followed by SVR ($R^2 = 0.77$), Gradient Boosting ($R^2 = 0.76$), Extra Trees ($R^2 = 0.71$), and Bayesian Ridge ($R^2 = 0.70$). The cross-validation R^2 values ranged from 0.56 to 0.61, which were lower than the test performance, reflecting the heterogeneity of the dataset. SVR and Bayesian Ridge showed the smallest training-cross-validation gaps (0.06 and 0.08, respectively), consistent with their regularization mechanisms. Ensemble models exhibited larger gaps (XGBoost = 0.36; Gradient Boosting = 0.29; Extra Trees = 0.19), indicating greater sensitivity to the training data partition. Nevertheless, XGBoost remained the strongest performer on unseen data. These findings are consistent with previous studies reporting strong performance of ensemble models for HHV prediction in heterogeneous biomass systems. Adeleke et al. [2] reported an XGBoost test performance of $R^2 = 0.73$, while artificial neural network, SVR, and random forest models have reported R^2 values between 0.70 and 0.98 depending on dataset size and evaluation strategy [2,4,6-8,29]. These baseline results provide the reference against which the effect of synthetic data augmentation is evaluated in the following section.

Model	Train R ²	CV R ² (mean ± SD)	Test R ²	RMSE (MJ/kg)	MAE (MJ/kg)	Gap
SVR (RBF)	0.66	0.60 ± 0.13	0.77	1.03	0.82	0.06
Bayesian Ridge	0.69	0.61 ± 0.13	0.70	1.18	0.91	0.08
XGBoost	0.93	0.57 ± 0.11	0.81	0.94	0.73	0.36
Gradient Boosting	0.88	0.59 ± 0.11	0.76	1.06	0.85	0.29
Extra Trees	0.75	0.56 ± 0.10	0.71	1.16	0.93	0.19

TABLE 4: Predictive performance of machine learning models for higher heating value estimation based on proximate and ultimate analysis features. CV results are reported as mean ± standard deviation across five stratified folds

CV, Cross-Validation; MAE, Mean Absolute Error; RBF, Radial Basis Function; RMSE, Root Mean Square Error; SVR, Support Vector Regression

Generation and evaluation of synthetic data

The optimal cGAN configuration presented in Table 2 generated 700 synthetic samples, of which 695 were retained after quantile filtering. Table 5 presents the comparative performance of the cGAN, Bootstrap with noise, and Gaussian Copula across multiple fidelity and privacy metrics. Consistent with recent ML literature, synthetic data quality is best interpreted using a multidimensional framework, as fidelity, privacy, and distinguishability often reflect competing objectives rather than a single measure of performance [16,17].

Metric	cGAN	Bootstrap and Noise	Gaussian Copula
Hellinger	0.17	0.27	0.24
PCD	0.09	0.19	0.12
AUC-ROC	0.70	0.36	0.32
Wasserstein (mean)	0.90	1.03	0.98
JS divergence (category)	0.03	0.12	0.08
DCR	3.82	1.03	4.04
Memorization (%)	3.40	61.76	2.63

TABLE 5: Fidelity and privacy evaluation of synthetic data generation methods. Real-vs-Real Hellinger distance = 0.22, computed between two random halves of the training set as a distributional baseline

Bold indicates the best value for each metric. For Hellinger, PCD, Wasserstein, JS divergence, and memorization, lower values indicate better fidelity. For DCR, higher values indicate lower privacy risk.

JS, Jensen–Shannon; PCD, Pairwise Correlation Difference; AUC-ROC, Area Under the Receiver Operating Characteristic Curve; DCR, Distance to Closest Record; cGAN, Conditional Generative Adversarial Networks

The cGAN-generated synthetic data had the lowest Hellinger distance (0.17), outperforming Gaussian Copula (0.24) and Bootstrap (0.27). Two random splits of the real data had a Hellinger distance of 0.22 between them, providing a natural variability baseline. Synthetic data from cGAN fell below this threshold, indicating that synthetic samples were representative of any random sub-sample of the training data. The cGAN-generated synthetic data still demonstrated superior performance across the remaining fidelity metrics. It achieved the lowest JS divergence (0.03) and pairwise correlation difference (0.09), indicating that both marginal distributions and inter-feature relationships were preserved, as illustrated in Figure 2 and Figure 3. Similarly, cGAN-generated data had the lowest mean Wasserstein distance (0.90), followed by Gaussian Copula (0.98) and Bootstrap (1.03). Data generated by Bootstrap exhibited the highest memorization rate of 61.76%, suggesting that distributional similarity is achieved through replication of training records rather than genuine synthesis.

Figure 2. Kernel density estimation of real and cGAN synthetic feature distributions

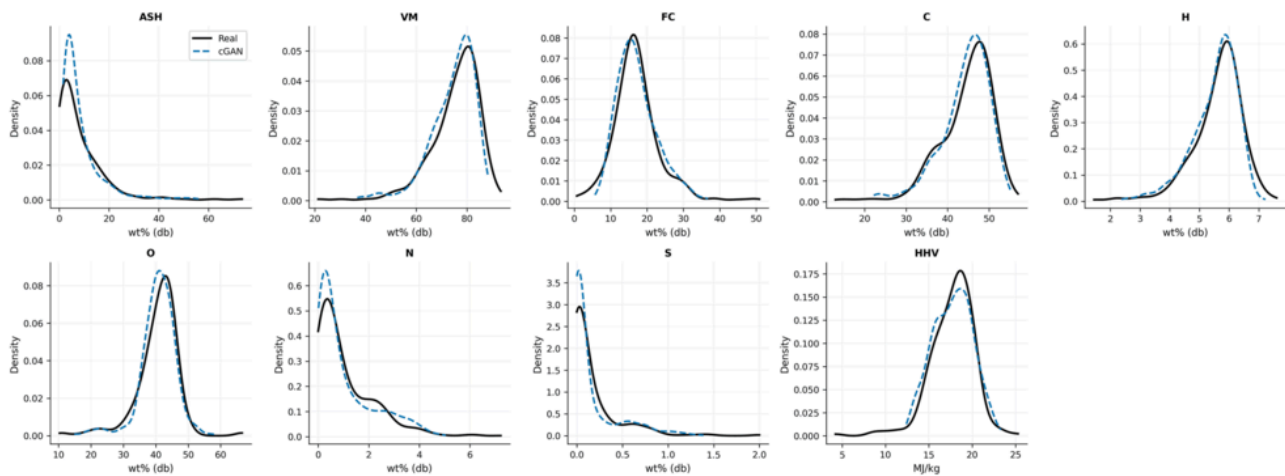


FIGURE 2: Kernel density estimation comparing real (solid black) and cGAN-generated synthetic (dashed blue) distributions across proximate analysis (ASH, VM, FC), ultimate analysis (C, H, O, N, S), and HHV features

cGAN, Conditional Generative Adversarial Network; HHV, Higher Heating Value; VM, Volatile Matter; FC, Fixed Carbon; C, Carbon; H, Hydrogen; O, Oxygen; N, Nitrogen; S, Sulphur

Figure 3. Pearson correlation structure of real and cGAN synthetic features including HHV

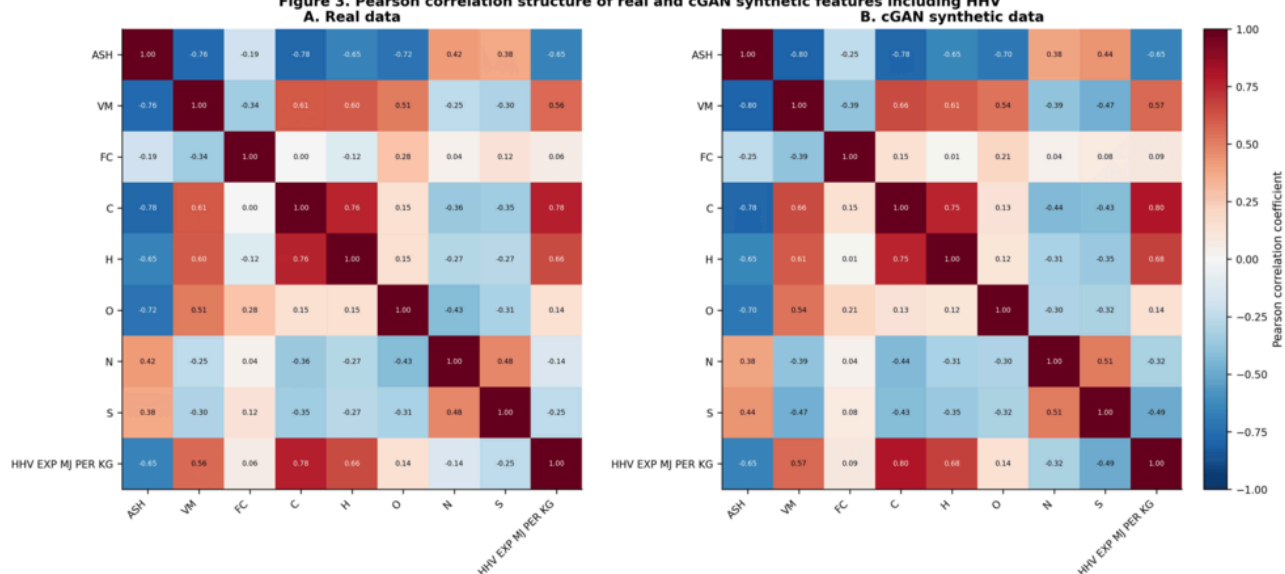


FIGURE 3: Pearson correlation matrices for real data (Panel A) and cGAN-generated synthetic data (Panel B) across all input features and HHV, where colour intensity reflects the strength and direction of the correlation coefficients

cGAN, Conditional Generative Adversarial Network; HHV, Higher Heating Value; VM, Volatile Matter; FC, Fixed Carbon; C, Carbon; H, Hydrogen; O, Oxygen; N, Nitrogen; S, Sulphur

From a privacy perspective, the cGAN achieved a DCR of 3.82 and a memorization rate of 3.40%, indicating that synthetic samples remain sufficiently distinct from real records. In contrast, data generated by Bootstrap exhibited a lower DCR (1.03) and substantially higher memorization (61.76%), reflecting a tendency to reproduce training data. The Gaussian

How to cite this article:

Nakimuli C, Kawuma S, Okot D K, et al. (June 24, 2026) Machine Learning Prediction of Higher Heating Value in Heterogeneous Biomass: An Evaluation of Synthetic Data Augmentation Strategies. Cureus J Comput Sci 3 : es44389-026-00133-8. DOI <https://doi.org/10.7759/s44389-026-00133-8>

Copula achieved the strongest privacy profile (DCR = 4.04; memorization = 2.63%), though this privacy advantage came at the cost of reduced fidelity, as reflected in the higher Hellinger and PCD values, illustrating the inherent fidelity-privacy trade-off in synthetic data generation [17]. The cGAN-generated synthetic data achieved an AUC-ROC of 0.70, which reflects the limited training set size ($n = 252$ across 6 categories) rather than poor fidelity, as demonstrated in [16,17]. The AUC-ROC values for synthetic data generated by Bootstrap (0.36) and Gaussian Copula (0.32) were below the recommended threshold of 0.5 [16]. However, these values may indicate distributional collapse, where the classifier encounters two clearly separated distributions and effectively reverses its predictions rather than reflecting superior fidelity [13]. Overall, the cGAN-generated data demonstrated the best balance across all evaluated metrics.

Downstream augmentation utility

Augmentation improvements were modest and model-dependent, as shown in Table 6 and Figure 4. Under cGAN augmentation, Gradient Boosting recorded the highest ΔR^2 (+0.028), while XGBoost achieved the highest post-augmentation test performance ($R^2 = 0.83$, MAE = 0.72 MJ/kg) using 25 synthetic samples. Gradient Boosting also benefited under Gaussian Copula augmentation ($\Delta R^2 = +0.019$), while Extra Trees responded most strongly to Gaussian Copula augmentation ($\Delta R^2 = +0.013$, MAE 0.93 to 0.90 MJ/kg) at 100 samples. SVR showed negligible or negative response across all strategies, and Bayesian Ridge improved only marginally. While the observed gains are modest in absolute magnitude, a Friedman test was conducted to determine whether the choice of augmentation strategy significantly affected model performance. The test compared test R^2 values across four conditions: models trained on real data only, cGAN, Gaussian Copula, and Bootstrap with noise. The test confirmed statistically significant differences across the four conditions ($\chi^2 = 10.68$, $p = 0.014$). cGAN achieved the highest mean rank (3.6) across all models, confirming it as the most consistently effective augmentation strategy, while Bootstrap with noise ranked lowest (1.0), reducing performance across all model families. These outcomes directly mirror the fidelity rankings in Table 5, confirming that augmentation utility is governed by synthetic data quality. cGAN, which achieved the lowest Hellinger distance (0.17) and memorization rate (3.40%), generated statistically distinct samples that enriched the training distribution and produced consistent performance gains. Bootstrap with noise, which exhibited the highest memorization rate (61.76%), introduced redundant rather than novel training variation and reduced performance across all five models [16,17]. These findings are consistent with synthetic tabular data literature, confirming that augmentation utility is determined by synthetic data quality rather than volume alone [13,17]. The observed gains over baseline demonstrate that cGAN augmentation improves HHV prediction accuracy without requiring additional experimental data collection, representing a practically significant advantage in settings where each biomass characterization measurement carries substantial cost. In the broader context of biomass HHV prediction, the post-augmentation XGBoost ($R^2 = 0.83$; MAE = 0.72 MJ/kg) falls within the performance range reported in recent studies [2,6-8], confirming that cGAN augmentation is a viable strategy for improving HHV prediction accuracy in heterogeneous multiclass biomass datasets, with direct practical relevance for feedstock screening in resource-constrained settings where expanding experimental datasets is prohibitively expensive. Multi-seed evaluation of individual model augmentation gains is recommended as future work.

How to cite this article:

Nakimuli C, Kawuma S, Okot D K, et al. (June 24, 2026) Machine Learning Prediction of Higher Heating Value in Heterogeneous Biomass: An Evaluation of Synthetic Data Augmentation Strategies. *Cureus J Comput Sci* 3 : es44389-026-00133-8. DOI <https://doi.org/10.7759/s44389-026-00133-8>

Model	Base R^2	cGAN ΔR^2	Bootstrap ΔR^2	Copula ΔR^2	Base MAE	cGAN MAE	Bootstrap MAE	Copula MAE
SVR	0.77	-0.0004	-0.0103	-0.0038	0.82	0.83	0.83	0.82
Bayesian Ridge	0.70	+0.0095	-0.0117	+0.0023	0.91	0.91	0.92	0.90
Extra Trees	0.71	+0.0050	-0.0040	+0.0131	0.93	0.92	0.94	0.90
Gradient Boost	0.76	+0.0278	-0.0077	+0.0187	0.85	0.82	0.87	0.82
XGBoost	0.81	+0.0192	-0.0192	-0.0023	0.73	0.72	0.79	0.71

TABLE 6: Test-set predictive performance of machine learning models under three augmentation strategies at the optimal augmentation size. ΔR^2 values indicate changes relative to the baseline. RMSE and MAE are reported in MJ/kg

cGAN, Conditional Generative Adversarial Networks; MAE, Mean Absolute Error; SVR, Support Vector Regression; RMSE, Root Mean Square Error

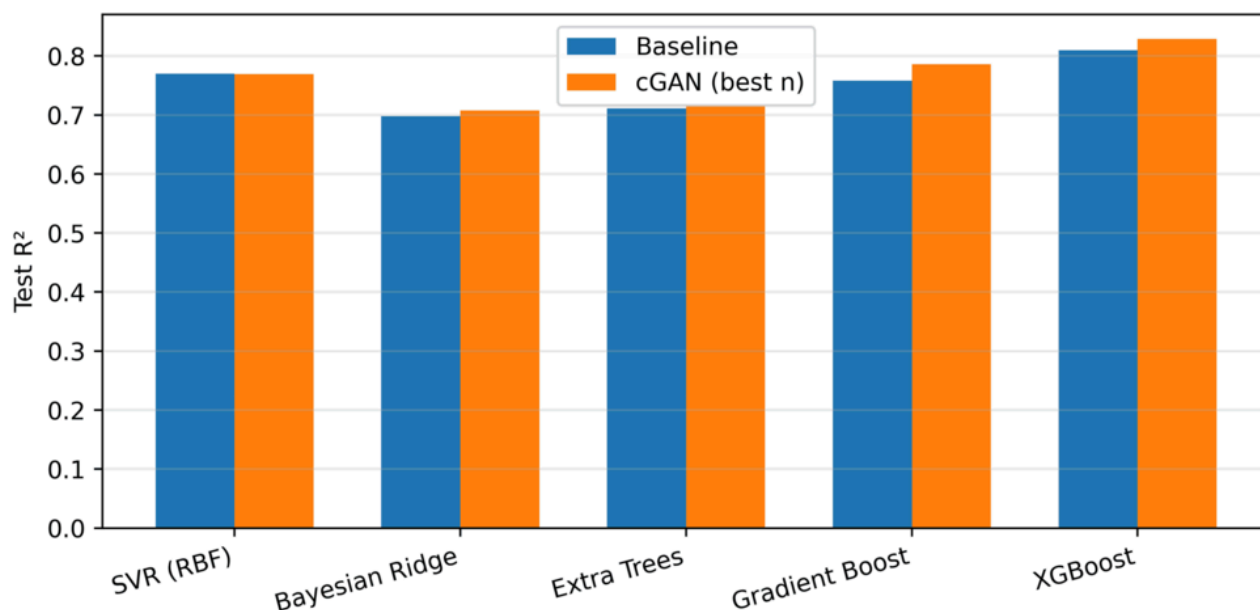


FIGURE 4: Test R^2 performance comparison of baseline and cGAN-augmented machine learning models for biomass HHV prediction

Blue bars show baseline model performance trained on real data only. Orange bars show performance after cGAN augmentation at the optimal synthetic sample size.

cGAN, Conditional Generative Adversarial Networks; HHV, Higher Heating Value; RBF, Radial Basis Function; SVR, Support Vector Regression

How to cite this article:

Nakimuli C, Kawuma S, Okot D K, et al. (June 24, 2026) Machine Learning Prediction of Higher Heating Value in Heterogeneous Biomass: An Evaluation of Synthetic Data Augmentation Strategies. Cureus J Comput Sci 3 : es44389-026-00133-8. DOI <https://doi.org/10.7759/s44389-026-00133-8>

The selection of augmentation size was guided by the cross-validation performance (Table 7). All models showed reduced performance when augmentation exceeded 100 synthetic samples. For example, Gradient Boosting declined from a peak cross-validation R^2 of 0.49 at 25 samples to 0.40 at 400 samples. XGBoost also decreased from 0.53 at 25-50 samples to 0.46 at 400 samples. This establishes an optimal range of 25-100 synthetic samples, which corresponds to approximately 10-40% of the real training set. This is consistent with findings in synthetic tabular data that synthetic samples should supplement rather than dominate the real training distribution [13,17]. These findings suggest that augmentation volume should be carefully controlled under small and heterogeneous data conditions.

Model	25	50	100	200	400
SVR (RBF)	0.62	0.61	0.62	0.55	0.48
Bayesian Ridge	0.65	0.64	0.64	0.64	0.63
Extra Trees	0.55	0.53	0.55	0.51	0.47
Gradient Boosting	0.49	0.46	0.47	0.42	0.40
XGBoost	0.53	0.53	0.49	0.52	0.46

TABLE 7: Cross-validation R^2 by model and synthetic sample size (cGAN augmentation)

cGAN, Conditional Generative Adversarial Networks; RBF, Radial Basis Function; SVR, Support Vector Regression

Conclusions

Bomb calorimetry remains costly and inaccessible in resource-constrained regions, motivating ML alternatives for HHV prediction from proximate and ultimate analysis data. This study presents the first quantitative evaluation of fidelity, utility, and privacy in synthetic data augmentation for biomass HHV prediction across heterogeneous multiclass feedstocks. Five ML models were trained on a dataset of 315 biomass samples spanning six feedstock categories, with XGBoost achieving the strongest baseline performance ($R^2 = 0.81$; RMSE = 0.94 MJ/kg). Three synthetic data generation methods were subsequently assessed, with cGAN demonstrating the most balanced performance across fidelity and privacy metrics, achieving a Hellinger distance of 0.17 and a memorization rate of 3.40%. Downstream augmentation experiments revealed modest, model-dependent gains, with improvements observed for Gradient Boosting ($\Delta R^2 = +0.028$) and XGBoost ($\Delta R^2 = +0.019$; post-augmentation $R^2 = 0.83$) within an optimal range of 25-100 synthetic samples, beyond which performance deteriorated across all models. SVR and Bootstrap with noise showed no consistent benefit. The study is limited by its reliance on literature-compiled data, the relatively small dataset constraining cGAN training across six categories, and the absence of external validation against independently collected measurements. Future work should expand feedstock diversity, explore SMOGN for feedstock-specific augmentation, and pursue external validation to confirm model generalization beyond the training distribution.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

How to cite this article:

Nakimuli C, Kawuma S, Okot D K, et al. (June 24, 2026) Machine Learning Prediction of Higher Heating Value in Heterogeneous Biomass: An Evaluation of Synthetic Data Augmentation Strategies. Cureus J Comput Sci 3 : es44389-026-00133-8. DOI <https://doi.org/10.7759/s44389-026-00133-8>

Concept and design: Constance Nakato Nakimuli, Simon Kawuma, David K. Okot, Fred Kaggwa, Johan De Greef, Maarten Vanierschot

Acquisition, analysis, or interpretation of data: Constance Nakato Nakimuli

Drafting of the manuscript: Constance Nakato Nakimuli

Critical review of the manuscript for important intellectual content: Constance Nakato Nakimuli, Simon Kawuma, David K. Okot, Fred Kaggwa, Johan De Greef, Maarten Vanierschot

Supervision: Simon Kawuma, David K. Okot, Fred Kaggwa, Johan De Greef, Maarten Vanierschot

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** Constance Nakato Nakimuli reports that this research was supported by a scholarship from the Mbarara University of Science and Technology Inter-University Cooperation Partnership (MUST-IUC), known as the University as Facilitator Community Based Sustainable Solutions (UCoBS) project, funded by the Flemish Interuniversity Council (VLIR-UOS), Belgium. The funders had no role in study design, data collection, analysis, interpretation, or the decision to submit for publication. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The data (and/or code) supporting this study are openly available in Mendeley Data at <https://doi.org/10.17632/5xswr8yp34.3>, reference 10.17632/5xswr8yp34.3.

References

1. Abdulyekeen KA, Umar AA, Patah MFA, Daud WMAW: [Torrefaction of biomass: Production of enhanced solid biofuel from municipal solid waste and other types of biomass](#). *Renewable and Sustainable Energy Reviews*. 2021, 150:111436. [10.1016/j.rser.2021.111436](#)
2. Adeleke AA, Adedigba A, Adeshina SA, et al.: [Comparative studies of machine learning models for predicting higher heating values of biomass](#). *Digital Chemical Engineering*. 2024, 12:100159. [10.1016/j.dche.2024.100159](#)
3. Taki M, Rohani A: [Machine learning models for prediction the higher heating value \(HHV\) of municipal solid waste \(MSW\) for waste-to-energy evaluation](#). *Case Studies in Thermal Engineering*. 2022, 31:101823. [10.1016/j.csite.2022.101823](#)
4. Sheng C, Azevedo JLT: [Estimating the higher heating value of biomass fuels from basic analysis data](#). *Biomass and Bioenergy*. 2005, 28:499-507. [10.1016/j.biombioe.2004.11.008](#)
5. Yin CY: [Prediction of higher heating values of biomass from proximate and ultimate analyses](#). *Fuel*. 2011, 90:1128-1132. [10.1016/j.fuel.2010.11.031](#)
6. Afolabi IC, Epelle EI, Gunes B, Güleç F, Okolie JA: [Data-driven machine learning approach for predicting the higher heating value of different biomass classes](#). *Clean Technologies*. 2022, 4:1227-1241. [10.3390/cleantechnol4040075](#)
7. Brandić I, Voća N, Gunjača J, et al.: [Biomass higher heating value prediction machine learning insights into ultimate, proximate, and structural analysis datasets](#). *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*. 2024, 46:2842-2854. [10.1080/15567036.2024.2309303](#)
8. Adeleke O, Jen TC: [Multivariate analysis and soft computing-based prediction of energy potential in heterogenous waste streams](#). *E3S Web of Conferences*. 2025, 647:02001. [10.1051/e3sconf/202564702001](#)

How to cite this article:

Nakimuli C, Kawuma S, Okot D K, et al. (June 24, 2026) Machine Learning Prediction of Higher Heating Value in Heterogeneous Biomass: An Evaluation of Synthetic Data Augmentation Strategies. *Cureus J Comput Sci* 3 : es44389-026-00133-8. DOI <https://doi.org/10.7759/s44389-026-00133-8>

9. Ghiasi MM, Mohammadzadeh O, Zendejboudi S, Fgaier H: [Effective approach to assess higher heating value of biomass from ultimate and proximate analysis](#). The Canadian Journal of Chemical Engineering. 2025, 103:1155-1168. [10.1002/cjce.25287](#)
10. Samadi SH, Ghobadian B, Nosrati M: [Prediction of higher heating value of biomass materials based on proximate analysis using gradient boosted regression trees method](#). Energy Sources, Part A: Recovery, Utilization, and Environmental Effects. 2021, 43:672-681. [10.1080/15567036.2019.1630521](#)
11. Timilsina MS, Sen S, Uprety B, Patel VB, Sharma P, Sheth PN: [Prediction of HHV of fuel by machine learning algorithm: interpretability analysis using Shapley Additive Explanations \(SHAP\)](#). Fuel. 2024, 357:129573. [10.1016/j.fuel.2023.129573](#)
12. Okolie JA: [Can biomass structural composition be predicted from a small dataset using a hybrid deep learning approach?](#). Industrial Crops and Products. 2023, 203:117191. [10.1016/j.indcrop.2023.117191](#)
13. Lv C, Wang W, Ding L, Zhou W, Ma Z, Jia J: [Local data augmentation of biomass gasification performance prediction based on improved generative adversarial network](#). Energy. 2025, 330:136808. [10.1016/j.energy.2025.136808](#)
14. Zhang Y, Hao Y, Fu Y, et al.: [GAN-MAML strategy for biomass energy production: Overcoming small dataset limitations](#). Applied Energy. 2025, 387:125568. [10.1016/j.apenergy.2025.125568](#)
15. Eren B, Uzun S, Ozdemir S: [A robust machine learning framework for predicting the higher heating value of poultry litter using proximate analysis](#). Biomass and Bioenergy. 2026, 211:109153. [10.1016/j.biombioe.2026.109153](#)
16. Hernandez M, Osorio-Marulanda PA, Catalina M, Loinaz L, Epelde G, Aginako N: [Comprehensive evaluation framework for synthetic tabular data in health: fidelity, utility and privacy analysis of generative models with and without privacy guarantees](#). Frontiers in Digital Health. 2025, 7:1576290. [10.3389/fdgth.2025.1576290](#)
17. Chai S, Chadney G, Avery C, Grunewald P, Van Hentenryck P, Donti PL: [Defining 'good': evaluation framework for synthetic smart meter data \[PREPRINT\]](#). arXiv. [10.48550/arXiv.2407.11785](#)
18. Phyllis2. [Database for the physico-chemical composition of \(treated\) lignocellulosic biomass, micro- and macroalgae, various feedstocks for biogas production and biochar](#). (2020). Accessed: March 1, 2025: <https://phyllis.nl/>.
19. Rosenblatt M, Tejaviubulya L, Jiang R, Noble S, Scheinost D: [Data leakage inflates prediction performance in connectome-based machine learning models](#). Nature Communications. 2024, 15:1829. [10.1038/s41467-024-46150-w](#)
20. Bergstra J, Bengio Y: [Random search for hyper-parameter optimization](#). Journal of Machine Learning Research. 2012, 13:281-305.
21. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A: [Improved training of Wasserstein GANs](#). Advances in Neural Information Processing Systems. 2017, 30:1-11.
22. Arjovsky M, Chintala S, Bottou L: [Wasserstein generative adversarial networks](#). Proceedings of the 34th International Conference on Machine Learning. 2017, 214-223.
23. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al.: [Generative adversarial nets](#). Advances in Neural Information Processing Systems. 2014, 1-9.
24. Chen Z, Gao J, Wang W, Yan Z: [Physics-informed generative neural network: an application to troposphere temperature prediction](#). Environmental Research Letters. 2021, 16:065003. [10.1088/1748-9326/abfde9](#)
25. Yang L, Zhang D, Karniadakis GE: [Physics-informed generative adversarial networks for stochastic differential equations](#). SIAM Journal on Scientific Computing. 2020, 42:A292-A317. [10.1137/18m1225409](#)
26. Albano C, Camacho N, Hernández M, Matheus A, Gutiérrez A: [Influence of content and particle size of waste pet bottles on concrete behavior at different w/c ratios](#). Waste Management. 2009, 29:2707-2716. [10.1016/j.wasman.2009.05.007](#)
27. Xia Y, Wang CH, Mabry J, Cheng G: [Advancing retail data science: comprehensive evaluation of synthetic data \[PREPRINT\]](#). arXiv. 2024, [10.48550/arXiv.2406.13130](#)
28. Boumanchar I, Charafeddine K, Chhiti Y, et al.: [Biomass higher heating value prediction from ultimate analysis using multiple regression and genetic programming](#). Biomass Conversion and Biorefinery. 2019, 9:499-509. [10.1007/s13399-019-00386-5](#)
29. Ighalo JO, Igwegbe CA, Adeniyi AG: [Multi-layer perceptron artificial neural network \(MLP-ANN\) prediction of biomass higher heating value \(HHV\) using combined biomass proximate and ultimate analysis data](#). Modeling Earth Systems and Environment. 2022, 8:3177-3191. [10.1007/s40808-021-01276-4](#)

How to cite this article:

Nakimuli C, Kawuma S, Okot D K, et al. (June 24, 2026) Machine Learning Prediction of Higher Heating Value in Heterogeneous Biomass: An Evaluation of Synthetic Data Augmentation Strategies. Cureus J Comput Sci 3 : es44389-026-00133-8. DOI <https://doi.org/10.7759/s44389-026-00133-8>