Received: 21 September 2024 | Revised: 5 November 2024 | Accepted: 21 November 2024 | Published online: 30 December 2024

RESEARCH ARTICLE

Journal of Computational and Cognitive Engineering
2024, Vol. 00(00) 1–11
DOI: 10.47852/bonviewJCCE42024395

BON VIEW PUBLISHING

VAE-CNN: Deep Learning on Small Sample Dataset Improves Hydrogen Yield Prediction in Co-gasification

Thavavel Vaiyapuri[1,*], Walaa H. Elashmawi[2,3], Shridevi S[4] and William Asiedu[5]

[1]*College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Saudi Arabia*

[2]*Faculty of Computers and Informatics, Suez Canal University, Egypt*

[3]*Faculty of Computer Science, Misr International University, Egypt*

[4]*Centre for Advanced Data Science, Vellore Institute of Technology, India*

[5]*Akenten Appiah-Menka University of Skills Training and Entrepreneurial Development, Ghana*

Abstract: Renewable and ecologically friendly energy sources have piqued the attention of researchers due to the increasing usage of fossil fuels and the looming problem of global warming. One potential solution that might pave the way for sustainable growth is biohydrogen, which could significantly reduce reliance on fossil fuels. The co-gasification process is garnering attention as a promising method for the production of hydrogen from plastic waste and biomass. In this context, optimizing the process is crucial for improving and predicting biohydrogen production. The recent advancements in deep learning models have opened up promising new possibilities. Training these models with small in situ samples, however, results in poor accuracy. Hence, this research is the first of its kind to explore the possibility of using a variational autoencoder (VAE) to provide high-quality synthetic data and aid the identification of process parameters that can improve hydrogen generation in co-gasification. VAE-augmented training set is utilized to guide one-dimensional convolutional neural network (1D-CNN) to accurately capture the relationship between hydrogen production and the process parameters. The efficiency of VAE and 1D-CNN is verified by comprehensive comparison evaluations with different data augmentation (DA) schemes and regression models. The experimental findings demonstrate that the proposed VAE network significantly improves prediction performance by generating data that is more realistic in comparison to other DA schemes. With the synthetic data from VAE, 1D-CNN was able to optimize the co-gasification process for increased hydrogen production, with a 32% improvement in maximum error and a 7% improvement in root mean squared error.

Keywords: thermochemical conversion, biomass energy conversion, deep learning, adversarial networks, SHAP framework

1. Introduction

The accelerating growth of the global population and intensified industrialization have led to a dramatic rise in energy demand. This demand has, for decades, been met primarily through fossil fuel combustion, which emits large amounts of carbon dioxide (CO_2) into the atmosphere [1]. These CO_2 emissions have contributed significantly to global warming, leading to environmental issues such as melting ice caps, rising sea levels, ocean acidification, and more frequent extreme weather events. With CO_2 concentrations reaching alarming levels, there is an urgent need to adopt cleaner energy sources to mitigate these environmental impacts.

Simultaneously, urbanization and improved living standards have changed consumption patterns, increasing waste generation across residential and industrial sectors. This waste, often non-biodegradable, poses severe environmental and health risks, impacting ecosystems and public health [2]. By 2030, global production of electronic and electrical waste is expected to reach millions of tons. This underscores the urgent need for sustainable strategies that tackle both waste management and energy demands [3]. In response to these challenges, this research proposes a technology paradigm that uses waste materials to generate clean, green energy. This approach offers a sustainable alternative to fossil fuels while addressing the global waste crisis.

Thermochemical conversion methods, particularly co-gasification, are promising solutions to these issues [4]. Co-gasification converts waste materials into hydrogen-rich gas, creating a valuable pathway for clean energy production. However, maximizing hydrogen output from complex feedstocks requires optimizing numerous operational parameters. Achieving this is challenging due to the high dimensionality and variability of co-gasification conditions. Traditional methods, such as thermodynamic and kinetic modeling, attempt to capture these dynamics. Yet, they often fall short because of the complexity of gasification processes and their reliance on simplifying assumptions [4].

***Corresponding author:** Thavavel Vaiyapuri, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Saudi Arabia. Email: t.thangam@psau.edu.sa

Recent advancements in artificial intelligence (AI) offer new possibilities for overcoming these challenges [5]. Machine learning (ML) has shown great potential in bioenergy optimization, although its effectiveness depends heavily on the availability of large, high-quality datasets [6]. Experimental data collection for co-gasification models is often costly and labor-intensive, limiting data availability. This has motivated research into using synthetic data to expand datasets without extensive physical trials. The deep learning (DL), particularly with generative models like variational autoencoders (VAEs) and Generative Adversarial Networks (GANs), can now generate synthetic data closely resembling real datasets [7, 8]. This study utilizes VAEs, which are particularly suitable for producing continuous, structured data like hydrogen yield, maintaining feature coherence necessary for effective co-gasification modeling. VAEs are preferable to GANs due to their stable training process and ability to avoid issues such as mode collapse. This stability makes VAEs a powerful tool for generating synthetic data that enhances predictive modeling, ultimately supporting optimized co-gasification processes for sustainable energy solutions.

Our research pioneers the use of VAE to address this persistent data limitation in co-gasification optimization. While the focus is on enhancing the co-gasification process, the proposed model offers adaptability and potential applicability across other domains. This paper seeks to determine whether high-quality synthetic data can be generated to improve co-gasification process optimization. The primary contributions of this study include:

1) Developing a DL model utilizing VAE to generate synthetic tabular data, enhancing prediction accuracy for hydrogen production in biomass-plastic co-gasification.
2) Conducting an in-depth analysis evaluating the efficacy of VAE-generated synthetic data compared to conventional data augmentation (DA) methods.
3) Validating the effectiveness of augmented training sets on predictive accuracy, comparing a one-dimensional convolutional neural network (1D-CNN) model with other ML models for hydrogen yield prediction.

2. Literature Review

With a rising demand for sustainable energy, research into advanced thermochemical processes like co-pyrolysis and co-gasification has accelerated. Integrating AI into these processes shows promise for optimizing chemical conversions and boosting hydrogen yield, especially from waste. AI-driven models, particularly ML and DL, have been increasingly used to predict and enhance yield outcomes. For instance, a study [9] using artificial neural networks (ANN) demonstrated the potential of radial basis functions and multilayer perceptrons (MLP) in modeling hydrogen production from co-gasification of rubber and plastic waste.

Building on this, recent ML efforts emphasize ensemble methods to overcome individual algorithm limitations and improve hydrogen production accuracy. For instance, Devasahayam and Albijanic [10] explore tree-based ensembles effective in capturing complex interactions in biomass-plastic co-gasification, while Ajorloo et al. [11] use statistical modeling and response surface methodology to analyze variable effects, optimizing co-gasification parameters.

A comprehensive investigation into the use of ML for predicting hydrogen yield from biomass and plastic co-gasification was conducted by Khan et al. [12], whose findings demonstrate that ANN outperform other models in predictive accuracy. Extending this line of research, Devasahayam [8] examined the capabilities of advanced DL architectures, testing various models to gain deeper insights into hydrogen production processes. Additionally, Ramos et al. [13] reviewed recent advancements in gasification and co-gasification, focusing on waste-to-energy conversion technologies and their potential applications.

Another body of research underscores the importance of ML in optimizing pyrolysis processes for biomass and waste plastics, which is essential for sustainable conversion technologies [14]. For example, Mishra et al. [15] studied hydrogen production from biomass-plastic co-pyrolysis and co-gasification, while Block et al. [16] reviewed literature on co-pyrogasification, analyzing product distribution, conditions, feedstock, and synergistic effects.

Our review shows that most studies emphasize the role of operating conditions in hydrogen production. Despite AI's potential to improve predictive accuracy, data scarcity remains a challenge [17, 18]. This study proposes a VAE model to generate synthetic data resembling real-world data, enhancing DL models for more accurate, generalizable predictions in co-gasification. This approach helps bridge data gaps, expanding AI's applicability in bioenergy research.

3. Research Methodology

This section outlines the proposed model, which combines a VAE with a 1D-CNN regressor for synthetic data generation and prediction, as shown in Figure 1. Observational data are initially preprocessed and then input into the VAE to generate synthetic data, which is evaluated for quality using statistical measures. The high-quality synthetic data then used to train the 1D-CNN regression model to predict hydrogen production. Detailed descriptions of each component are provided in the following subsections.

3.1. Synthetic data generation using VAE

An autoencoder (AE) is an unsupervised neural network that encodes the input x into a low-dimensional vector z and subsequently reconstructs the original data, ensuring that the output d(z) closely approximates the input [19]. VAE is an advanced deep generative architecture that expands the concept of AEs to produce novel synthetic data derived from in-situ observation [20], as seen in Figure 2. The encoder part of the VAE compresses the actual input to understand the latent probability distributions, while the decoder creates or replicates the learnt probability distribution to reconstruct the original input. Few recent studies [21, 22] utilized VAE for the generation of synthetic 1D data. Driven by the results of their investigation, this study incorporates the VAE to produce synthetic data for predicting hydrogen production in co-gasification process. The equation below outlines the process from computational perspective,

$$G(x) = V(\text{Decode}(\text{Encode}(x)) \tag{1}$$

In Equation (1), $G(x)$ denotes the generated synthetic data, x signifies the in situ observations, and V is the VAE function that processes x to produce (x). Furthermore, the Encode() function serves as the encoder of the VAE model by learning the latent distribution from the available in situ observations, while the Decode() function acts as the decoder of VAE by analyzing the latent distributions to produce the synthetic data for co-gasification process. To generate new high-quality data points, a constraint is applied to the learning of the latent space, ensuring that it represents the latent characteristic as a probability distribution. The encoding concept is delineated as follows [23],

Figure 1
Experimental framework designed for assessing the proposed model VAE-CNN for hydrogen yield prediction

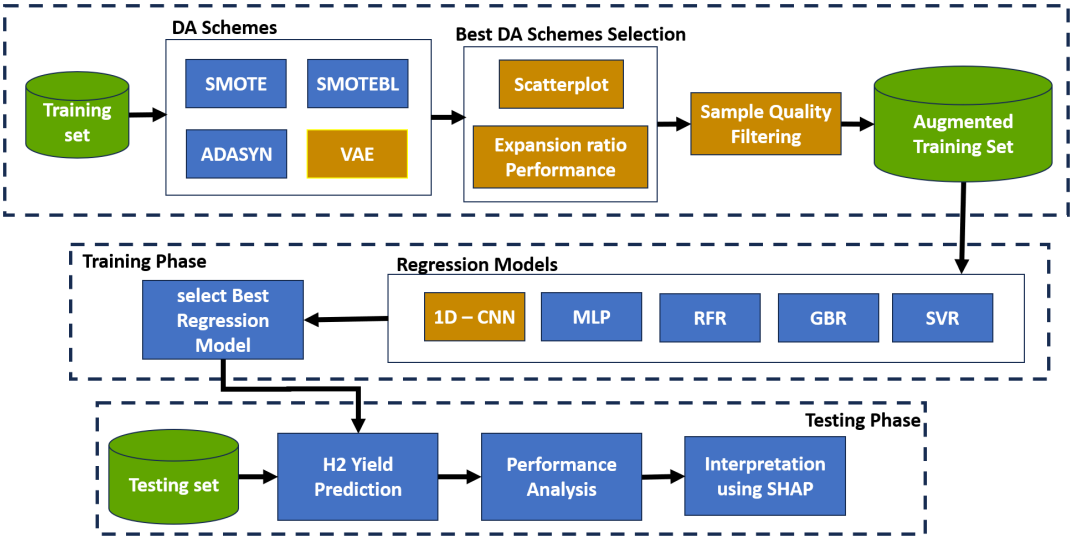
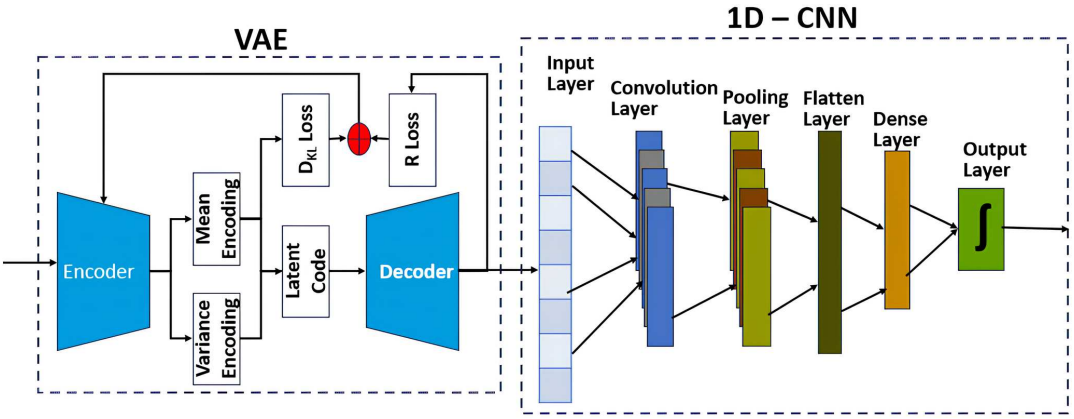


Figure 2
The proposed VAE-CNN model schematic design with its control flow



$$VAE\ loss = R\ Loss + KL\ Loss \tag{5}$$

$$z = \sigma(x) * N(0,1) + \mu(x) \tag{2}$$

Here $\sigma(x)$ and $\mu(x)$ represent the standard deviation and mean of the in situ observations, respectively. The VAE normalizes the encoding distribution during training to make the latent space suitable for synthetic data production. The input data may be used to produce a large number of new data points with comparable properties by sampling points from latent space and decoding them. To generate data closer to the in situ observations, in this study, the loss function is structured with two components: the reconstruction term, which enhances the efficiency of the encoding–decoding process, and the regularization term, which ensures the regularity of the latent space. The initial term is defined as the MSE, while the subsequent term represents the Kullback–Leibler divergence (KL) [22], which quantifies the disparity between the two distributions. The loss function structured in this study is mathematically represented as follows:

$$R\ loss = ||x - decode(z)||^2 \tag{3}$$

$$KL\ loss = KL[N(\mu x, \sigma x), N(0,1)] \tag{4}$$

In this context, N (0,1) denotes the standard normal distribution across the in situ observations, while $N(\mu_x, \sigma_x)$ represents the learning latent distribution adopted by the encoder network. The VAE simultaneously learns the mean and covariance of the latent distribution, with the key aim to reconstruct the output while minimizing reconstruction errors. Doing so, the loss function guides the VAE designed in this study to generate synthetic samples closer to the actual data used for training.

Table 1 outlines the structure of the VAE designed for generating synthetic data in the proposed system. The VAE architecture begins with an input layer of shape (n, 4), which is processed through two encoder layers, each increasing dimensionality from (n, 4) to (n, 16) using ReLU activation functions. The encoder then outputs mean and variance layers, both with shape (n, 2), which define the latent space by learning the distribution from the in situ observations. This latent representation is then decoded through two layers, reducing dimensionality back to the original shape (n, 4). The decoder layers also use ReLU activation, ensuring nonlinearity and effective feature extraction.

This structured network allows the VAE to produce high-quality synthetic data that closely resembles real data, enhancing

Table 1
VAE network structure design in the proposed model

Layer	Input shape	Output shape	Activation
Input	(n, 4)	(n, 4)	-
Encoder-1	(n, 4)	(n, 8)	ReLU
Encoder-1	(n, 8)	(n, 16)	ReLU
Mean	(n, 16)	(n, 2)	-
Variance	(n, 16)	(n, 2)	-
Latent code	(n, 16)	(n, 2)	-
Decoder-1	(n, 16)	(n, 8)	ReLU
Decoder-2	(n, 8)	(n, 4)	ReLU
Output	(n, 4)	(n, 4)	

predictive accuracy for the co-gasification process. Upon completion of training, the encoder generates means and covariances, which are used to sample new latent vectors. These vectors are subsequently processed through the decoder to generate new synthetic samples.

3.2. Hydrogen yield prediction using 1D-CNN

A CNN is a specialized form of feed-forward neural network within DL [24]. It autonomously learns features through the optimization of filters, also known as kernels. The CNN processes input data by applying a sequence of convolution and pooling layers. The convolution operation is an essential step in the process of extracting useful characteristics from input data, usually images. Convolution scans the input data using trainable filters, called kernels. These kernels systematically traverse the input, calculating a weighted sum of pixel values within their designated receptive fields. The core idea is that these learned kernel functions act as feature detectors to identify underlying patterns. In the convolution procedure, the kernel's weights are uniformly distributed over the whole input, considerably decreasing the number of parameters in comparison to dense or fully connected layers. This characteristic of parameter sharing enables CNN to learn hierarchical representations of features and makes them computationally proficient. The convolution procedure, in essence, progressively converts the input image into a collection of feature maps. Each feature map indicates the response generated by a specific kernel to a particular feature present in the input data. Multiple convolutional layers are stacked to allow the network learn progressively more complicated and abstract features, which facilitates the recognition of intricate patterns within the data. The convolution process can be described mathematically as [24],

$$C[i] = \sum_{n=0}^{k-1} x[i+n] \cdot K[n] \tag{6}$$

In this context, x refers to the input feature map, K and k represent the filter and dimension of the filter, respectively. While $C[i]$ indicates the value of the feature map output at position $[i]$. This study employs a 1D-CNN, designed for processing one-dimensional sequences of data. The 1D CNN first learns and captures the significant features from the sequences of input data. Subsequently, it delineates the intrinsic features of the input sequence data. A CNN model typically has the following layers: input, convolutional, pooling, dense, and output as shown in Figure 2. Hence, it is imperative to meticulously arrange convolutional layers in a logical manner and fine-tune the width and depth of the deep neural network (DNN) when working with intricate co-gasification process data to achieve highest possible outcome.

Table 2
1D-CNN structure design in the proposed model

Layer	Input shape	Output shape	Activation
Input	(n, 4,1)	(n, 4,1)	
Conv-1D	(n, 4,1)	(n, 4,8)	ReLU
Up-sampling	(n, 4,8)	(n, 8,8)	
Conv-1D	(n, 8,8)	(n, 8,16)	ReLU
Up-sampling	(n, 8,16)	(n, 16,16)	
Dropout	(n, 16,16)	(n, 16,16)	
Flatten	(n, 16,16)	(n, 256)	
Dense	(n, 256)	(n, 1)	ReLU

The 1D-CNN designed for the proposed system, as detailed in Table 2, begins with an input layer of shape (n, 4,1), followed by two Conv-1D layers with output shapes (n, 4,8) and (n, 8,16), respectively, each using ReLU activation for capturing nonlinear features. These convolutional layers are interspersed with upsampling layers, expanding the dimensionality to (n, 8,8) and (n, 16,16), preserving spatial relationships in the data. To prevent overfitting, a dropout layer with the same shape (n, 16,16) is included. The network then utilizes a flatten layer to reshape the data to (n, 256), followed by a dense layer with output shape (n, 1) using ReLU activation to complete the model. This structured architecture, with ReLU activations throughout, enables the 1D-CNN to effectively process and extract complex features, optimizing predictive accuracy by minimizing MSE loss through gradient descent during training [25].

4. Experimental Design and Setup

4.1. Data description

A previous study on co-gasification of waste plastic and rubber has compiled and published the data used in this work, as mentioned in Devasahayam [8], Ayodele et al. [9], and Chin et al. [26]. The dataset is comprised of thirty separate experiments that were conducted using a central composite design. The independent variables included parameters such as gasification temperature, amount of plastic in the mixture, size of the rubber seed shell (RSS) biomass particles, and HDPE particles. Whereas, one of the dependent variables that has been measured is the amount of hydrogen produced during the co-gasification process of waste plastic and rubber. A thermogravimetric analyzer coupled with a mass spectrometer has been used to carry out the investigations.

Figure 3 [26] illustrates correlation among the variables in the dataset used for the co-gasification study. Most correlations are close to zero, indicating weak or negligible linear relationships between variables such as temperature, RSS size, HDPE particle size, plastic content, and hydrogen (H_2) yield. Notably, RSS size and HDPE show a slightly negative correlation with H_2, while plastic content has a minor positive correlation (0.13) with H_2. These results suggest that, individually, these variables have minimal direct linear impact on hydrogen yield within the observed range.

Table 3 provides a statistical summary, highlighting key details: the particle sizes for both HDPE and biomass RSS vary from 0.13 to 0.63 mm, and the plastic content in the mixture ranges from

Figure 3
Correlation between independent and dependent variables in the study dataset

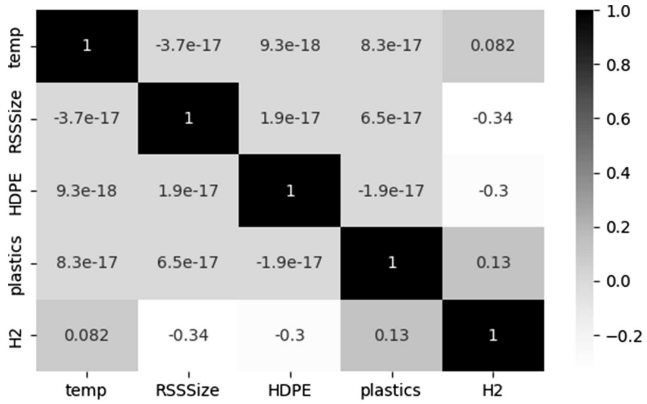


Table 3
Descriptive statistics of the study dataset

	Temp	RSSSize	HDPE	plastics	H2
Count	30.00	30.00	30.00	30.00	30.00
Mean	700.00	0.38	0.38	20.00	44.71
std	90.97	0.11	0.11	9.10	3.65
min	500.00	0.12	0.12	0.00	38.57
25%	600.00	0.25	0.25	10.00	42.21
50%	700.00	0.38	0.38	20.00	45.64
75%	800.00	0.50	0.50	30.00	47.33
max	900.00	0.62	0.62	40.00	50.12

0% to 40%. The gasification temperature spans 500 °C to 900 °C, with a mean of 700 °C. For hydrogen (H_2) yield, the mean value is 44.71, with values ranging from 38.57 to 50.12. The descriptive statistics, including standard deviation, minimum, and maximum values for each variable, offer insights into the variability and distribution within the dataset, which is critical for understanding parameter influences on hydrogen production outcomes.

4.2. Traditional DA schemes for comparison

This subsection presents traditional DA methods [19]—SMOTE, SMOTEBL, and ADASYN—selected for comparison in this study as shown in Figure 1, along with rationale for their inclusion:

SMOTE (Synthetic Minority Over-sampling Technique): Although primarily used for classification, it can be adapted for regression tasks by generating synthetic samples along the line between neighboring data points in the feature space. This method serves as a foundational comparison for improving predictive accuracy in regression.

SMOTEBL (SMOTE Borderline): a variant of SMOTE, focuses on generating synthetic samples near regions of high variability or challenging data boundaries. For regression, this approach aims to enhance data diversity in critical regions, allowing for more accurate modeling of complex relationships in co-gasification parameters.

ADASYN (Adaptive Synthetic Sampling): adapts the generation of synthetic samples based on data density, adding more samples where data is sparse. In regression, this approach helps to reinforce underrepresented areas in the feature space, improving model robustness and predictive accuracy in hydrogen yield estimation.

These methods provide benchmark DA techniques, facilitating a comprehensive comparison with the proposed VAE-based data generation to identify the most effective approach for enhancing predictive accuracy in the co-gasification dataset.

4.3. ML regression models for comparison

This subsection discusses the ML regression models chosen for comparison in this study as shown in Figure 1. These models offer varied and effective approaches to predictive modeling in complex datasets, making them highly suitable for hydrogen yield prediction in co-gasification processes. Brief notes on these models [12, 27] are as follows:

Support Vector Regression (SVR): is adept at managing nonlinear relationships through the use of kernel functions, making it ideal for capturing intricate interactions among variables. It controls error margins around a decision boundary, which enhances accuracy in hydrogen yield predictions, particularly in datasets with limited samples or outliers.

Random Forest Regression (RFR): As an ensemble method, RFR builds and combines multiple decision trees for a more robust prediction. This approach captures interactions among multiple variables effectively and minimizes the risk of overfitting, especially useful in datasets with high variance and nonlinear relationships, as seen in co-gasification data.

Gradient Boosting Regression (GBR): is another ensemble technique, iteratively refines model accuracy by addressing errors from previous iterations. Its adaptive learning approach optimizes predictive performance, making it well-suited for datasets with complex dependencies and enabling refined predictions for hydrogen yield.

Multilayer Perceptron (MLP): is a neural network-based model that processes data through multiple layers of neurons, allowing it to learn complex patterns. Its ability to model intricate data distributions makes MLP valuable for uncovering hidden patterns in the data, essential for generating nuanced predictions.

These models were selected to provide a comprehensive analysis of different regression models, covering a spectrum of linear, ensemble, boosting, and neural network approaches. This variety supports a balanced comparison of each model's strengths and limitations, tailored to address the complexities of predicting hydrogen production in co-gasification processes.

4.4. Evaluation metrics

This subsection outlines the metrics employed to evaluate the proposed model for hydrogen yield prediction, along with the rationale for their selection. Root mean squared error (MSE) quantifies the average squared differences between predicted and actual values, placing greater emphasis on larger errors, which supports the identification of models capable of delivering stable, accurate predictions. Maximum error (ME) measures the largest deviation observed, providing insight into the model's worst-case performance—a critical factor for assessing reliability in sensitive applications. Additionally, R-squared (R^2) is utilized to gauge the model's explanatory power by showing how well it captures variance within the data, thus indicating its effectiveness in reflecting underlying data patterns. To further assess robustness, standard deviations across various datasets are considered, offering a measure of the model's consistency and generalizability. Collectively, these

metrics provide a well-rounded evaluation of the model's accuracy, reliability, and applicability in hydrogen yield prediction.

4.5. Experimental framework

This study makes a two-pronged contribution. First, it compares the effectiveness of four DA methods—SMOTE, SMOTE Border Line (SMOTEBL), ADASYN, and VAE—for generating synthetic data from limited in-situ observations of the co-gasification process. Second, it evaluates the predictive accuracy of the proposed 1D-CNN model for hydrogen yield prediction against four other regression models (RFR, GBR, SVR, and MLP) using the augmented training dataset. The experimental framework is structured into two parts:

4.5.1. Data augmentation process

1) Four DA methods are applied to the training dataset to increase sample size.
2) The effectiveness of each DA scheme is assessed using scatterplot analysis and performance across different expansion ratios to identify the most suitable augmentation method.
3) A quality filtering step is then performed on the selected DA method's synthetic samples. Samples with high reconstruction errors or flagged as outliers are excluded, ensuring that only high-quality data are included in the augmented training set. This step reduces noise and enhances the reliability of the synthetic data.

4.5.2. Model training and evaluation

1) The augmented dataset is used to train five regression models—1D-CNN, RFR, GBR, SVR, and MLP—with fivefold cross-validation to optimize hyperparameters.
2) The effectiveness of each model is assessed on both the original and augmented datasets to evaluate the impact of DA on predictive accuracy. Key performance metrics such as MSE, ME, and R^2 are used to compare model performance.
3) Finally, the best regression model is selected based on these performance criteria for hydrogen yield prediction.

This experimental process facilitates a comprehensive evaluation of the impact of DA, particularly VAE, on enhancing the training set and improving model accuracy. It demonstrates the advantages of quality-filtered synthetic data in robust hydrogen yield prediction, contributing to the reliability and effectiveness of the proposed approach for practical applications.

4.6. Implementation details

The proposed model and the ML models examined in this study were implemented using the sklearn package in Python [28]. Additionally, Python libraries such as matplotlib, seaborn, and statsmodels were employed for exploratory data analysis to examine relationships between the target variable and predictors. The study was conducted within the Jupyter notebook interface on the Google Colaboratory platform [29], offering an interactive environment that eliminates the need for local system setup. All experiments in this study were performed using this platform.

4.7. Hyperparameter tuning

This study employs two distinct DNNs: a VAE for synthetic data generation and a 1D-CNN for regression. Given the limited data samples, specific strategies were implemented to enable the VAE to produce high-quality synthetic data despite constraints typically requiring larger datasets. To address this, the VAE's hyperparameters—learning rate, batch size, and latent space dimensions—were carefully tuned to maximize the model's capacity to capture essential patterns with fewer samples. Regularization techniques and early stopping were also applied to prevent overfitting, supporting stable and robust training.

While VAEs are generally optimized with extensive data, research suggests that, with targeted hyperparameter tuning and model configuration aligned to the dataset structure, VAEs can still perform effectively on smaller datasets [30]. Thus, a grid search method, using original in-situ observations, guided the selection of hyperparameters to ensure optimal performance. For the 1D-CNN, additional parameters such as dropout rates and weight decay were also tuned, and both networks employed the Adam optimizer to enhance the optimization process. Table 4 provides a summary of the final hyperparameter values determined for each model.

Table 4
Tuned hyperparameters for the proposed VAE-CNN

Parameters	VAE	1D-CNN
Batch size	32	32
Learning rate	0.001	0.001
Optimizer	Adam	Adam
Epochs	15	15
Activation function	ReLU	ReLU
Loss function	MSE+KLD	MSE
Dropout	-	0.01

5. Results and Discussion

This section presents the experimental results that support the study objectives and demonstrate the advantages of the proposed model at each stage of the research process. The interpretation of these results provides a foundation for the research findings. Initially, we evaluate the effectiveness of the VAE by comparing the quality of its synthetic data with that generated by three other DA methods. Subsequently, we assess the 1D-CNN's prediction accuracy against four selected ML regression models using the VAE-augmented training dataset. To ensure reliable model performance evaluations, diverse training, and testing datasets are used to reduce biases related to overfitting or underfitting. The augmented dataset is divided using stratified sampling, following the standard 80:20 split ratio, with 80% allocated for training and 20% for testing.

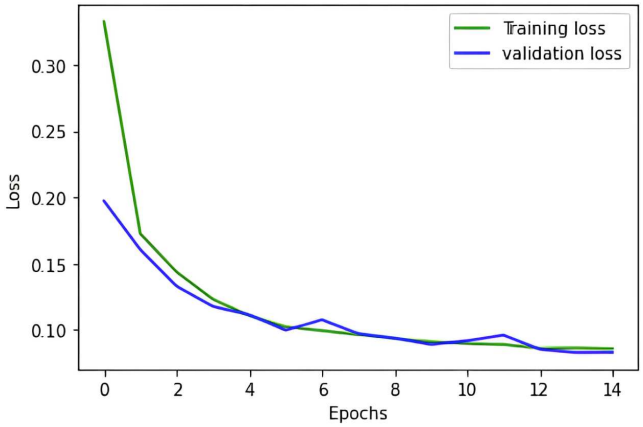
5.1. Performance comparison of VAE

This section aims to evaluate the effectiveness of using a VAE for DA. To achieve this, we consider three alternative DA strategies: ADASYN, SMOTEBL, and SMOTE. To assess and compare performance, two sets of experiments are conducted, as detailed below.

5.1.1. Scatterplot analysis

In this experiment, we design a VAE network with the specified hyperparameter values and use in situ observations from the "Data Description" section to evaluate the quality of the synthetic tabular data generated by the network. Figure 4 provides a visualization of the loss values for both the training and validation sets, highlighting the model's performance and stability during the training process.

Figure 4
Learning curve of VAE in DA process



The training loss and validation loss are indicated by blue and green curves, respectively. Details of the specific loss functions used are provided in the "Research Methodology" section. In Figure 4, as the number of iterations increases, a significant reduction in both training and validation loss values is observed. This decline indicates that the model is progressively learning and adjusting its parameters to fit the data more accurately. Over time, the losses stabilize, suggesting that the model has reached convergence and is no longer significantly improving with additional training cycles. To improve the clarity and interpretability of these training results, the x-axis in Figure 4 has been set to represent epochs rather than iterations, as this provides a more comprehensive view of model progress across full passes through the dataset.

Upon completing the training phase, the quality of the synthetic data samples was assessed against the original data using a scatter plot with a regression line. This approach is a well-established technique for visually and statistically evaluating the correlation between generated and original data samples, making it a fundamental strategy in assessing generative model efficacy in data-driven studies [23]. The degree of correlation, measured by the correlation coefficient, provides insight into the similarity between synthetic data (G) and original data (O), with stronger correlations indicating greater similarity as follows.

$$\varnothing_{G,O} = \sum_{i=1}^{F} 1 - \frac{|G_i - O_i|}{2} \tag{7}$$

A successful DA model produces synthetic data that clusters closely around the regression line with minimal error. In an ideal scenario, synthetic data points align perfectly with in situ observations along the diagonal [1:1] regression line.

Figure 5 uses a scatter plot and regression line to compare the quality of synthetic data generated by the VAE with that of the original data and other DA methods. Here, the red line represents the ideal [1:1] regression line, where data points from synthetic samples would align with original samples if identical. The blue points denote the distribution of in-situ samples, while the orange points represent synthetic data generated by the VAE and alternative DA methods.

Upon examining Figure 5, it becomes evident that the VAE model-generated samples cluster more closely around the [1:1] regression line than those from the other DA methods. This close clustering indicates a high degree of correlation between the VAE-generated samples and the original data, suggesting that the VAE has effectively learned the underlying patterns in the data. The improved alignment with the regression line highlights the VAE's superior performance in capturing essential features and maintaining the original dataset's distribution, outperforming the alternative DA methods in generating high-quality, representative synthetic data. This finding underscores the VAE's efficacy in DA for this study.

5.1.2. Performance against expansion ratio

In this experiment, DA was applied iteratively, with augmentation factors ranging from 1 to 20. Each increment factor represents the volume of additional data integrated into the training dataset. Starting with an initial augmentation doubling the training set size, subsequent iterations expanded the dataset to three times its original size, continuing up to a total of 20 iterations. Each iteration thus contributes an expanded training set aimed at improving the model's predictive performance.

Figure 6 presents the performance comparison across different DA strategies as the augmentation ratio increases. Here, the 1D-CNN regressor trained on the original dataset without DA serves as a baseline for performance comparison. As illustrated, all four DA strategies lead to substantial improvements over the baseline model, with the regressor performance consistently enhanced by DA as the augmentation ratio rises. Notably, the model's performance improves markedly with each increase in the augmentation factor, particularly up to an expansion ratio of eight.

Beyond an expansion ratio of eight, however, performance begins to decline, likely due to model overfitting as the regressor becomes excessively tuned to the augmented training data. This pattern suggests an optimal augmentation ratio range that maximizes performance without inducing overfitting. Among the DA techniques, ADASYN and VAE exhibit greater stability and consistently superior results compared to SMOTE, particularly as the augmentation ratio increases.

The results indicate that VAE is particularly effective in optimizing regressor performance. This advantage may stem from the unique structure of the VAE, where the generator and latent-space components enable the model to learn intricate features from the original dataset and produce synthetic samples closely aligned with the original data distribution. This feature allows the VAE-augmented dataset to enhance model generalizability, supporting robust performance across varied expansion ratios.

5.2. Performance comparison of 1D-CNN

The purpose of this experiment is to validate the research findings through a comparative analysis of the effectiveness of the 1D-CNN model versus four selected ML models: SVR, RFR, GBR, and MLP. As outlined in the "Hyperparameter Tuning" section, cross-validation and grid search were performed on each algorithm to identify optimal hyperparameters, enhancing model generalization and ensuring a fair performance comparison. The prediction performance metrics for all five regression models, both without DA and with VAE-based DA, are summarized in Table 5, with results visually represented as bar charts in Figure 7 for clearer interpretation.

The application of VAE DA leads to marked improvements in prediction accuracy across all models. Without DA, the models show varying R^2 values, from 0.41 for MLP to 0.76 for SVR, indicating a moderate fit between predictions and actual values. With VAE DA, R^2 scores increase significantly for each model, reflecting stronger alignment with actual outcomes. The 1D-CNN model achieves the highest R^2 score of 0.97, followed by GBR at 0.88 and RFR at 0.87, demonstrating the effectiveness of VAE DA in boosting predictive accuracy.

Figure 5
Performance comparison of VAE against other DA schemes using scatterplot

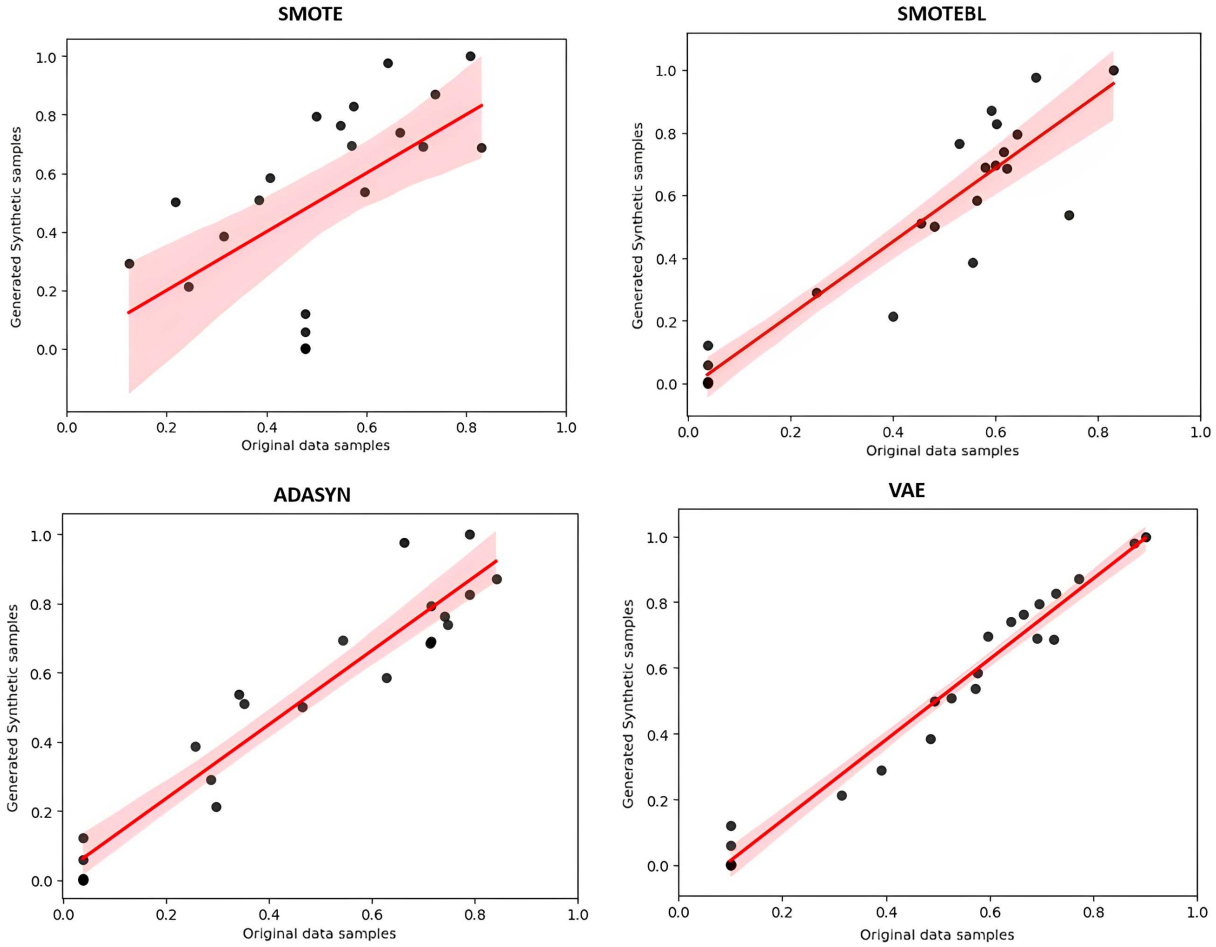
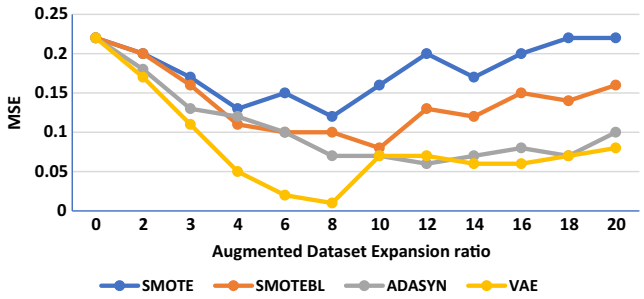


Figure 6
Performance comparison of different DA schemes against expansion ratio



Similarly, MSE and ME values, representing prediction error rates, decrease substantially with VAE DA. For example, SVR's MSE decreases from 0.18 to 0.09, and ME from 0.27 to 0.15, indicating that VAE DA effectively reduces prediction errors. The 1D-CNN model shows the lowest MSE (0.06) and ME (0.11), highlighting its superior accuracy and reduced error rate for hydrogen yield prediction.

Figure 7 illustrates that VAE DA has the most pronounced impact on MLP and 1D-CNN. These DL models benefit more from synthetic samples than traditional ML models (SVR, RFR, and GBR) because they can capture complex, nonlinear relationships in larger, more diverse datasets. While traditional models generally perform well with smaller datasets, MLP and 1D-CNN leverage VAE-augmented data to achieve improved accuracy and adaptability.

Table 5
Comparison of prediction performance for hydrogen yield without DA and with VAE DA

	Hydrogen yield prediction without DA			Hydrogen yield prediction using VAE DA		
Regression models	R^2	MSE	ME	R^2	MSE	ME
SVR	0.76	0.18	0.27	0.84	0.09	0.15
RFR	0.66	0.22	0.35	0.87	0.12	0.20
GBR	0.56	0.24	0.33	0.88	0.10	0.18
MLP	0.41	0.27	0.43	0.91	0.11	0.18
Proposed	0.58	0.13	0.67	0.97	0.06	0.11

Figure 7
Visual comparison of prediction performance for hydrogen yield without DA (top) and with VAE DA (bottom)

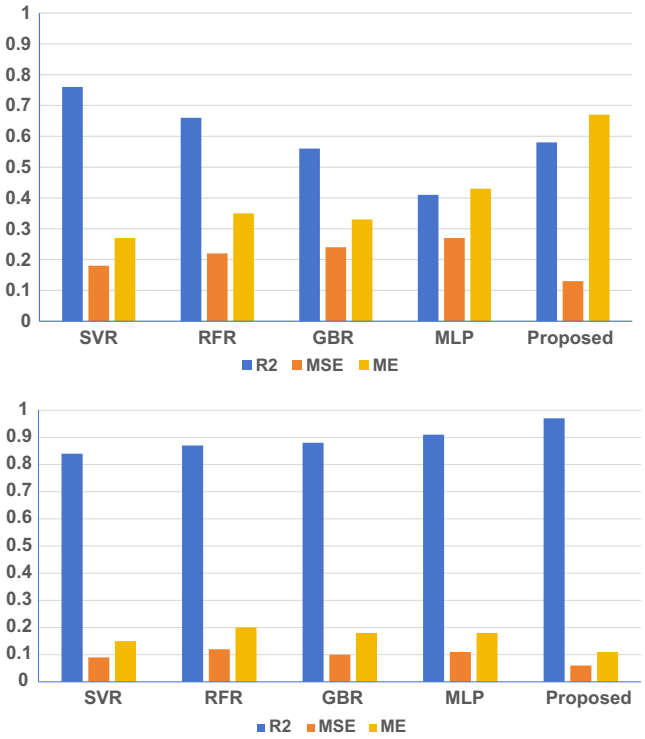
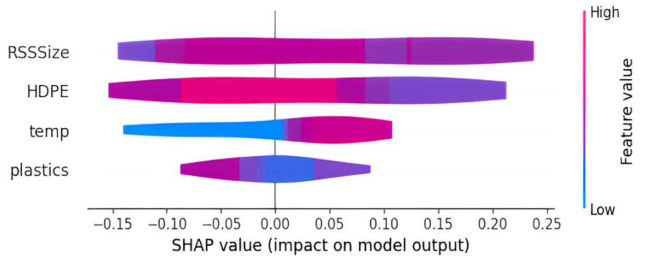


Figure 8
Summary plot depicting the feature importance for the proposed VAE-CNN model



Additionally, MLP and 1D-CNN can handle high-dimensional data and intricate feature representations, allowing them to make the most of the diverse synthetic data generated by VAE. This enhances their generalization capability, leading to greater accuracy gains than simpler models. By filling feature space gaps, VAE DA provides MLP and 1D-CNN with the necessary data diversity to generalize effectively, resulting in substantial improvements in R^2, MSE, and ME.

Overall, while VAE DA enhances performance across all models, its impact is most notable for 1D-CNN and MLP, underscoring its value for DL models that benefit from rich, varied datasets to model complex patterns accurately.

5.3. Interpretation ability analysis

In practice, when a researcher identifies an ML model that demonstrates an acceptable level of accuracy, the next step is to delve into the prediction process and derive insights to make informed decisions based on expert domain knowledge. The opaque nature of ML models poses difficulties in comprehending how input parameters affect the target during the modeling process. This experimental analysis goes beyond a step in solving the issue by investigating the interpretability of the developed ML models and providing rationale for their predictions, regardless of how complicated they may be [31].

In this direction, the global explanation of the Shapley method for the proposed model is illustrated in Figure 8, which effectively showcases the importance of features. It also emphasizes the positive and negative associations with the target, which are represented by red and blue colors, respectively. The examination of the summary plot in Figure 8 shows that, on average, all process parameters have the greatest influence on hydrogen generation during biomass-plastics co-gasification. It is worth noting that the proposed model identifies HDPE, RSSize, and temp as the primary process factors when compared to plastics. This explanation about the effects of process parameters aligns with the prevailing literature on co-gasification and existing knowledge in the discipline. The detailed examination of data distribution along the X axis reveals that the proposed model successfully captured the combined influence of HDPE and RSSize on hydrogen generation.

5.4. Computational complexity and feasibility

The VAE and 1D-CNN models in this study are computationally intensive due to their structural design and data processing requirements. The computational complexity of the VAE primarily arises from the encoder and decoder architectures, where each layer's operations scale with the input dimensionality. Specifically, the VAE has a complexity of $O(N \times d \times h)$, where N is the number of samples in training dataset, d is the input dimensionality, and h is the number of hidden units per layer. As the dataset size or input dimensionality increases, the memory and processing time required for encoding and decoding grow proportionally, which can be challenging for very large datasets.

Similarly, the 1D-CNN model's complexity depends on the number of layers, with each convolutional layer having a time complexity of $O(N \times d \times f \times k)$, where f is the filter size, and k is the number of filters. As more layers and filters are added, the computational cost increases, which can lead to significant training time and memory requirements, especially as dataset size grows. Although this complexity is manageable for smaller datasets, it may pose challenges for larger ones.

For production environments, these models may be optimized to ensure feasible implementation. Once training is complete, inference complexity is relatively low, making the VAE-CNN models practical for real-time or on-demand predictions in production settings. To reduce training time and memory requirements for larger datasets, several optimization strategies are recommended. For example, dimensionality reduction on VAE inputs can help control feature size, and simplifying the 1D-CNN architecture by reducing layers or filters can retain accuracy while decreasing computational load. Furthermore, deploying the models on distributed or parallel processing systems, such as multiple GPUs or TPUs, can improve scalability and efficiency, enabling faster processing and handling of larger datasets.

These strategies indicate that, while computationally intensive during training, the VAE-CNN framework can be adapted for production with proper resource management and optimization, making it feasible for deployment in production environments.

6. Conclusion

This study emphasizes the potential for synthesizing biomass-plastic samples to overcome the challenges posed by rigorous experimentation and limited availability of data in analyzing the co-gasification process. Following a thorough performance comparison with VAE-generated data, this study proposed to incorporate VAE for the development of synthetic tabular data to achieve improved the prediction accuracy for hydrogen production. Furthermore, we established a novel reconstruction loss for the VAE network within the proposed model by integrating the latent loss and divergence loss. Incorporating this loss allowed the VAE to enhance its ability to understand the relationship between process parameters and produce samples that more closely align with in-situ observations.

After augmenting the modest in situ observations using the VAE network, the experiments included training a 1D-CNN regression model and the chosen four ML models on augmented dataset. The findings confirmed the efficacy of 1D-CNN, demonstrating enhanced predictive performance relative to its equivalents in terms of R2, MSE, and ME. Lastly, the SHAP framework was used to perform an interpretation ability study of the proposed VAE-CNN model. This analysis aimed to shed light on the biomass-plastics co-gasification process and pinpoint the critical components that affect hydrogen production. The findings aligned with the prevailing literature on the co-gasification process and existing knowledge in the discipline.

To the best of our knowledge, this work is the first to use VAE and 1D-CNN in the context of hydrogen production research, with the goal of reducing the required sample size. While our methodology is presented within the framework of hydrogen generation, it is readily adaptable to other application areas where data collection is challenging, or data sensitivity is high.

Acknowledgment

The authors are grateful to Deanship of Research, Prince Sattam bin Abdulaziz University, Al Kharj, Saudi Arabia for their technical and financial support.

Funding Support

This project was sponsored by Prince Sattam bin Abdulaziz University (PSAU) as part of funding for its SDG Roadmap Research Funding Programme project number PSAU-2023-SDG-130.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest in this work.

Data Availability Statement

The data that support the findings of this study are openly available at https://doi.org/10.1002/er.6483, reference number [9].

Author Contribution Statement

Thavavel Vaiyapuri: Conceptualization, Methodology, Formal analysis, Writing – original draft, Supervision, Funding acquisition. **Walaa H. Elashmawi:** Methodology, Software, Resources, Writing – original draft. **Shridevi S:** Software, Validation, Formal analysis, Investigation, Resources, Writing – review & editing, Visualization, Project administration. **William Asiedu:** Validation, Investigation, Writing – review & editing, Visualization, Project administration.

References

[1] Squadrito, G., Maggio, G., & Nicita, A. (2023). The green hydrogen revolution. *Renewable Energy, 216,* 119041. https://doi.org/10.1016/j.renene.2023.119041

[2] Barma, M., & Modibbo, U. M. (2022). Multiobjective mathematical optimization model for municipal solid waste management with economic analysis of reuse/recycling recovered waste materials. *Journal of Computational and Cognitive Engineering, 1*(3), 122–137. https://doi.org/10.47852/bonviewJCCE149145

[3] Nguyen, V. G., Nguyen-Thi, T. X., Nguyen, P. Q. P., Tran, V. D., Ağbulut, Ü., Nguyen, L. H., . . . , & Pham, N. D. K. (2024). Recent advances in hydrogen production from biomass waste with a focus on pyrolysis and gasification. *International Journal of Hydrogen Energy, 54,* 127–160. https://doi.org/10.1016/j.ijhydene.2023.05.049

[4] Ghorbani, B., Zendehboudi, S., Afrouzi, Z. A., & Mohammadzadeh, O. (2024). Efficient hydrogen production via electro-thermochemical process and solid oxide fuel cell: Thermodynamics, economics, optimization, and uncertainty analyses. *Energy Conversion and Management, 307,* 118175. https://doi.org/10.1016/j.enconman.2024.118175

[5] Sakheta, A., Nayak, R., O'Hara, I., & Ramirez, J. (2023). A review on modelling of thermochemical processing of biomass for biofuels and prospects of artificial intelligence-enhanced approaches. *Bioresource Technology, 386,* 129490. https://doi.org/10.1016/j.biortech.2023.129490

[6] Okolie, J. A. (2023). Can biomass structural composition be predicted from a small dataset using a hybrid deep learning approach? *Industrial Crops and Products, 203,* 117191. https://doi.org/10.1016/j.indcrop.2023.117191

[7] Torky, M., Dahy, G., & Hassanein, A. E. (2023). GH2_MobileNet: Deep learning approach for predicting green hydrogen production from organic waste mixtures. *Applied Soft Computing, 138,* 110215. https://doi.org/10.1016/j.asoc.2023.110215

[8] Devasahayam, S. (2023). Deep learning models in Python for predicting hydrogen production: A comparative study. *Energy, 280,* 128088. https://doi.org/10.1016/j.energy.2023.128088

[9] Ayodele, B. V., Mustapa, S. I., Kanthasamy, R., Zwawi, M., & Cheng, C. K. (2021). Modeling the prediction of hydrogen production by co-gasification of plastic and rubber wastes using machine learning algorithms. *International Journal of Energy Research, 45*(6), 9580–9594. https://doi.org/10.1002/er.6483

[10] Devasahayam, S., & Albijanic, B. (2024). Predicting hydrogen production from co-gasification of biomass and plastics using tree-based machine learning algorithms. *Renewable Energy, 222,* 119883. https://doi.org/10.1016/j.renene.2023.119883

[11] Ajorloo, M., Ghodrat, M., Scott, J., & Strezov, V. (2022). Modelling and statistical analysis of plastic biomass mixture

co-gasification. *Energy, 256,* 124638. https://doi.org/10.1016/j.energy.2022.124638

[12] Khan, M., Naqvi, S. R., Ullah, Z., Taqvi, S. A. A., Khan, M. N. A., Farooq, W., . . . , & Štěpanec, L. (2023). Applications of machine learning in thermochemical conversion of biomass – A review. *Fuel, 332,* 126055. https://doi.org/10.1016/j.fuel.2022.126055

[13] Ramos, A., Monteiro, E., Silva, V., & Rouboa, A. (2018). Co-gasification and recent developments on waste-to-energy conversion: A review. *Renewable and Sustainable Energy Reviews, 81,* 380–398. https://doi.org/10.1016/j.rser.2017.07.025

[14] Wang, Z., Burra, K. G., Lei, T., & Gupta, A. K. (2021). Co-pyrolysis of waste plastic and solid biomass for synergistic production of biofuels and chemicals – A review. *Progress in Energy and Combustion Science, 84,* 100899. https://doi.org/10.1016/j.pecs.2020.100899

[15] Mishra, R., Ong, H. C., & Lin, C. W. (2023). Progress on co-processing of biomass and plastic waste for hydrogen production. *Energy Conversion and Management, 284,* 116983. https://doi.org/10.1016/j.enconman.2023.116983

[16] Block, C., Ephraim, A., Weiss-Hortala, E., Minh, D. P., Nzihou, A., & Vandecasteele, C. (2019). Co-pyrogasification of plastics and biomass, a review. *Waste and Biomass Valorization, 10,* 483–509. https://doi.org/10.1007/s12649-018-0219-8

[17] Fang, B., Jiang, M., Shen, J., & Stenger, B. (2022). Deep generative inpainting with comparative sample augmentation. *Journal of Computational and Cognitive Engineering, 1*(4), 174–180. https://doi.org/10.47852/bonviewJCCE2202319

[18] Jiang, P., Fan, J., Li, L., Wang, C., Tao, S., Ji, T., & Zhu, J. (2024). A hybrid approach combining mechanism-guided data augmentation and machine learning for biomass pyrolysis. *Chemical Engineering Science, 296,* 120227. https://doi.org/10.1016/j.ces.2024.120227

[19] Vaiyapuri, T., & Binbusayyis, A. (2020). Application of deep autoencoder as an one-class classifier for unsupervised network intrusion detection: A comparative evaluation. *Peer J Computer Science, 6,* e327. https://doi.org/10.7717/peerj-cs.327

[20] Kingma, D. P., & Welling, M. (2019). An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning, 12*(4), 307–392. https://doi.org/10.1561/2200000056

[21] Huang, T., Cheng, G., & Huang, K. (2020). Using variational auto encoding in credit card fraud detection. *IEEE Access, 8,* 149841–149853. https://doi.org/10.1109/ACCESS.2020.3015600

[22] Delgado, J. M. D., & Oyedele, L. (2021). Deep learning with small datasets: Using autoencoders to address limited datasets in construction management. *Applied Soft Computing, 112,* 107836. https://doi.org/10.1016/j.asoc.2021.107836

[23] Turénko, D., Khan, A., Hussain, R., & Imran Ali, S. (2020). Oversampling versus variational autoencoders: Employing synthetic data for detection of Heracleum Sosnowskyi in satellite images. In *Information Science and Applications: ICISA 2019,* 399–409. https://doi.org/10.1007/978-981-15-1465-4_40

[24] Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., & Inman, D. J. (2021). 1D convolutional neural networks and applications: A survey. *Mechanical Systems and Signal Processing, 151,* 107398. https://doi.org/10.1016/j.ymssp.2020.107398

[25] Chandra, K., Xie, A., Ragan-Kelley, J., & Meijer, E. (2022). Gradient descent: The ultimate optimizer. In *Proceedings of the 36th International Conference on Neural Information Processing System,* 8214–8225.

[26] Chin, B. L. F., Yusup, S., Al Shoaibi, A., Kannan, P., Srinivasakannan, C., & Sulaiman, S. A. (2015). Optimization study of catalytic co-gasification of rubber seed shell and high density polyethylene waste for hydrogen production using response surface methodology. In P. Ravindra (Ed.), *Advances in bioprocess technology* (pp. 209–223). Springer. https://doi.org/10.1007/978-3-319-17915-5_11

[27] Binbusayyis, A., Alaskar, H., Vaiyapuri, T., & Dinesh, M. J. (2022). An investigation and comparison of machine learning approaches for intrusion detection in IoMT network. *The Journal of Supercomputing, 78*(15), 17403–17422. https://doi.org/10.1007/s11227-022-04568-3

[28] van Rossum, G., & Drake, F. L. (2010). *The Python language reference* (Vol. 561). USA: Python Software Foundation.

[29] Bisong, E. (2019). *Building machine learning and deep learning models on Google cloud platform.* USA: Apress. https://doi.org/10.1007/978-1-4842-4470-8

[30] Moreno-Barea, F. J., Jerez, J. M., & Franco, L. (2020). Improving classification accuracy using data augmentation on small data sets. *Expert Systems with Applications, 161,* 113696. https://doi.org/10.1016/j.eswa.2020.113696

[31] Ahmed, R., Shehab, S. A., Elzeki, O. M., Darwish, A., & Hassanein, A. E. (2024). An explainable AI for green hydrogen production: A deep learning regression model. *International Journal of Hydrogen Energy, 83,* 1226–1242. https://doi.org/10.1016/j.ijhydene.2024.08.064

How to Cite: Vaiyapuri, T., Elashmawi, W. H., S, S., & Asiedu, W. (2024). VAE-CNN: Deep Learning on Small Sample Dataset Improves Hydrogen Yield Prediction in Co-gasification. *Journal of Computational and Cognitive Engineering*. https://doi.org/10.47852/bonviewJCCE42024395