|  |  |
| --- | --- |
| cetlogo ***CHEMICAL ENGINEERING TRANSACTIONS***  ***VOL. xxx, 2025*** | A publication of  aidiclogo_grande |
| The Italian Association  of Chemical Engineering  Online at www.cetjournal.it |
| Guest Editors: Fabrizio Bezzo, Flavio Manenti, Gabriele Pannocchia, Almerinda di Benedetto  Copyright © 2025, AIDIC Servizi S.r.l. **ISBN** 979-12-81206-17-5; **ISSN** 2283-9216 | |

Enhancing Torrefaction Process Modeling with Data Augmentation: A Study on Agri-Food Residues

Bartolomeo Cosenzaa, Michele Micciob, Ahmad Adeel Arshadb,c

a Dipartimento di Ingegneria Civile e Industriale, Università di Pisa, Largo Lucio Lazzarino, 56122 Pisa (PI), Italy, [bartolomeo.cosenza@unipi.it](mailto:bartolomeo.cosenza@unipi.it)

b Dipartimento di Ingegneria Industriale (DIIn), Università degli Studi di Salerno, via Giovanni Paolo II 132, 84084 Fisciano (SA), Italy, [mmiccio@unisa.it](mailto:mmiccio@unisa.it)

c Department of Science, Technology and Society, University School for Advanced Studies IUSS Pavia, Palazzo del Broletto, Piazza della Vittoria 15, 27100 Pavia (PV), Italy, [arshad.ahmad@iusspavia.it](mailto:arshad.ahmad@iusspavia.it)

Agri-Food residues can pose environmental challenges, but through proper valorization, they can support circular economy principles by providing alternatives to fossil fuels. This study focuses on modeling the torrefaction process of an industrial hazelnut processing waste (i.e., roasted cuticles) to produce high-density solid biofuels. An Artificial Neural Network (ANN) model is proposed. The work leverages the innovative approach of data augmentation through Gaussian noise addition to a relatively scarce experimental dataset. The expanded dataset greatly improved the Neural Network prediction performance: the method achieved considerably lower mean square error compared to the model output based on experimental data only. This approach offers several benefits: it enables accurate modeling with limited experimental data, reduces the need for extensive experimental work, lowers costs, and improves process optimization capabilities, presently in torrefaction processing and, more in general, in resource utilization and waste management.

* 1. Introduction

Torrefaction is a thermal process that heats organic material to temperatures between 200 and 300°C in an oxygen-free or low-oxygen environment (Panwar and Divyangkumar, 2024). This process usually produces a commodity solid fuel from biomass feedstocks by increasing its energy density and modifying its chemical and physical properties or, alternatively, new industrial raw materials. The industrial process occurs in batch or continuous reactors, with the fluidized bed reactor being a convenient option (Miccio et al., 2023). In this latter, the continuous movement of inert particles fluidized by an inert gas or steam ensures homogeneity of the reacting system and enhances heat and mass transfer to the biomass particles. During torrefaction, a series of chemical reactions occur, including thermal decomposition of hemicellulose and, partly, cellulose (Brachi et al., 2016). They release volatile substances such as carbon dioxide, carbon monoxide, methane, hydrogen, and other light gases, apart from water vapor. These latter are generally produced to a limited extent, with carbon dioxide being the prevailing one; therefore, they can be hardly used directly as fuels (Thengane et al., 2022). The solid product of torrefaction, frequently referred to as biochar (Khairy et al., 2024), becomes enriched with carbon, acquiring a higher calorific value compared to the starting material. Thanks to its porous structure, biochar finds use not only as a fuel but also as a fertilizer and soil amendment, as it can retain moisture and nutrients, improving soil quality and water retention capacity (Hu et al., 2022). Its absorptive properties also make it useful in applications such as water purification or the removal of air contaminants (Lin et al., 2023). Recently, the torrefaction process has been increasingly applied in the management of agricultural and food waste. Materials such as fruit shells, straw, prunings, or plant fibers, which often represent a large quantity of difficult-to-manage waste, are particularly suitable for this treatment due to their hemicellulose and cellulose content, which during torrefaction transform into high-calorific-value biochar (Chen et al., 2023). Similarly, relatively dry food waste such as fruit or vegetable scraps and expired products can be processed (Miccio et al., 2024), thus reducing their volume and preventing the release of unpleasant odors. All in all, torrefaction addresses sustainable waste management and energy recovery, reducing the need for landfills and minimizing greenhouse gas emissions from the anaerobic decomposition of these materials.

The torrefaction process is not easily described by traditional mathematical models, as it involves several complex and also nonlinear phenomena such as a series-parallel network of organic degradation reactions and changes in the physical properties of the solid material. In addition, the composition of the feedstock (moisture, proximate and elemental analysis), the highly variable size distribution and shape of the biomass particles, and the process variables (temperature, heating rate, inert gas flow rate, particle residence time, etc.) play an important role and interact each other in ways that are difficult to predict using deterministic models. Given these fundamental challenges in developing accurate first-principles models, researchers have explored alternative methods to better understand and predict torrefaction behavior.

* 1. Torrefaction modeling by Machine Learning

Machine learning (ML) has emerged as a promising tool in this context, offering a data-driven approach to understanding and predicting torrefaction behavior. Unlike traditional methods, ML can capture complex relationships between variables without requiring explicit mathematical formulations of the underlying physical and chemical phenomena, learning directly from experimental data to identify patterns and correlations that might be overlooked by conventional approaches. For instance, Su and Jiang (2024) developed ML models specifically designed to predict biochar properties derived from lignocellulosic biomass torrefaction. Oladosu et al. (2024) integrated experimental optimization with ML techniques to predict energy conversion efficiency in the torrefaction of Bambara Groundnut Shells. Furthermore, Naveed et al. (2024) contributed valuable insights into the prediction and optimization of torrefied biomass quality using various ML algorithms.

However, ML application is not free from critical problems. Limited data availability represents one of the main challenges (Nijman et al., 2022). As mentioned, the torrefaction process depends on multiple variables and, in the case of a small dataset, the model's ability to generalize and make accurate predictions is severely compromised. Traditional ML models, especially neural networks, require large and diverse datasets to effectively learn patterns and relationships between input variables and outcomes. Data scarcity, e.g., when they are costly to collect, can cause the model to not adequately capture these relationships, causing overfitting or poor performance on new and unseen scenarios. Additionally, if the data is not representative or contains noise, the results may be inaccurate. To address the challenge of limited data, data augmentation techniques can be employed to artificially expand the dataset. This involves generating modified versions of the existing data, e.g., by adding Gaussian noise (Dong et al., 2023; Dou et al., 2022; Valenti et al., 2016). This approach introduces slight variations to the original data, mimicking real-world uncertainties and measurement errors. By exposing the neural network to a wider variety of scenarios, data augmentation improves its ability to generalize, enabling better performance when the model encounters new, unseen data (Rebuffi et al., 2021).

While data augmentation has been widely used in image processing and signal analysis, its application to chemical process data is relatively recent. Zhang et al. (2021) investigated data augmentation and transfer learning strategies for reaction prediction in low chemical data regimes. In bioprocess applications, Wei et al. (2022) applied data augmentation techniques combined with machine learning for control strategy development in bio-polymerization processes. More recently, Balhorn et al. (2023) extended data augmentation methods to chemical process flowsheeting, presenting a systematic framework for machine learning applications.

* 1. Discussion
     1. Experimental

A bench-scale, specially instrumented, batch-operated torrefaction reactor (Figure 1A) enabled the collection of experimental data needed for subsequent ML applications. The experimental apparatus and the test procedure are described in detail elsewhere (Miccio et al., 2024).

Immagine che contiene testo, diagramma, design

Descrizione generata automaticamente Immagine che contiene verdura

Descrizione generata automaticamente

*A) B)*

*Figure 1: The lab-scale torrefaction experiments: A) fluidized bed reactor; B) hazelnut cuticles before and after torrefaction.*

The selected Agri-Food residues have been cuticles generated as a waste of industrial roasting of three different hazelnut feedstocks, i.e., Mortarella, Grimaldi and Prodal. As an example, they are shown in Figure 1B before and after the torrefaction test #17 with 2-4 mm Grimaldi particles at the temperature of 300 °C.

The experimental campaign included 29 tests. The residence time of the biomass batch in the fluidized bed has always been 5 min. The biomass particle size has always been 2-4 mm.

The main torrefaction process variables have been: 1. temperature (at three levels, i.e., 200°C, 250°C, 300°C); 2. sample mass loaded into the batch reactor (1.5, 2.5, 3.0 and 4.3 g). The other test parameters have been those related to the selected biomass composition: 3. moisture content (varying from 0 to 11.6 % wt); 4. Fixed Carbon, 5. Volatile Matter and 6. Ash (as provided by Proximate Analysis of the raw feedstock); 7. Carbon, 8. Hydrogen, 9. Nitrogen and 10. Oxygen (as provided by Ultimate Analysis of the raw feedstock). The experimental results provided the observed data for subsequent ML applications. The first one is: 1. the mass yield (on a dry basis) as defined by the equation (Miccio et al., 2024):

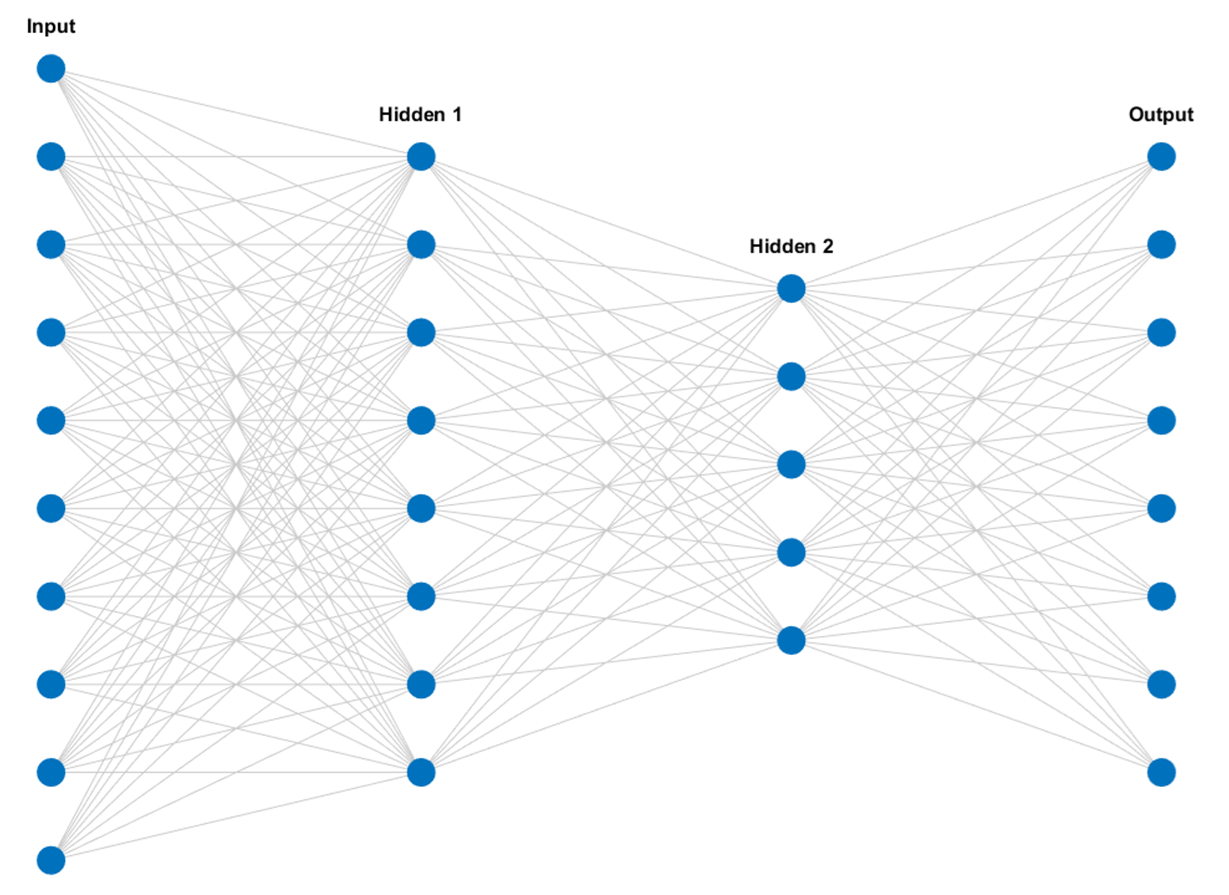
MY (%, db) = (mass of torrefied solids)/(mass of dry sample) 100 (1)

The other ones are those characterizing the composition of torrefied solids: 2. Fixed Carbon, 3. Volatile Matter and 4. Ash (as provided by Proximate Analysis of the torrefied solids); 5. Carbon, 6. Hydrogen, 7. Nitrogen and 8. Oxygen (as provided by Ultimate Analysis of the torrefied solids).

The experimental plan was structured to systematically investigate the influence of the key process variables (Negi et al., 2020) on the torrefaction of different hazelnut cuticles. Temperature levels (200°C, 250°C, 300°C) were selected to cover the typical torrefaction range, while sample mass variations (1.5 - 4.3 g) allowed investigation of potential mass transfer effects. Three different hazelnut cultivars (Mortarella, PRODAL, Grimaldi) with their inherent variations in moisture content (0 - 11.6% wt) and compositional characteristics were tested to account for feedstock variability in industrial applications.

* + 1. Methodological Approach

The Artificial Neural Network (ANN) consists of an input layer followed by two hidden layers with 8 and 5 neurons respectively (Figure 2). The number of neurons in the input (10) and output (8) layers was determined by the process variables and target parameters, respectively. The two hidden layers architecture with 8 and 5 neurons provides sufficient model capacity for the complexity of the torrefaction process modeling. The hidden layers use hyperbolic tangent (tansig) activation functions while the output layer employs a linear function (purelin). The optimization algorithm employs the Levenberg-Marquardt backpropagation (trainlm) for training. The entire neural network was implemented in MATLAB (2023b), using its Deep Learning Toolbox™. This architecture, while not necessarily the optimal one, represents a practical compromise between different requirements: model capacity to capture the non-linear relationships in the torrefaction process, computational efficiency in the training phase, negligible overfitting despite the limited size of the original experimental dataset and need of maintaining a good generalization ability.



*Figure 2: The Artificial Neural Network (ANN) structure.*

The original dataset consisted of 29 samples as obtained from the 29 experimental tests. Starting from this, the authors exploited the modeling approach by adopting data augmentation techniques. Adding Gaussian noise to the original data, both input and output, substantially expanded the available dataset while maintaining the fundamental physical relationships among measured experimental points. The slight variations introduced by noise addition can prevent overfitting to the experimental values, although it's important to note that any improved performance metrics should be interpreted primarily as an indication of better generalization within the bounds of experimental conditions, rather than as evidence of the model capturing new physical relationships. The augmentation procedure was applied to the original dataset (29 samples) before any data splitting. Both datasets, i.e., the original and the augmented one, were divided into training (70%), validation (15%), and test subsets (15%). This approach ensured consistent representation of both experimental and augmented data across all subsets while maintaining the integrity of the following cross-validation process and the reliability of the model evaluation process.

The authors decided to generate an augmented dataset that should have been more than one order of magnitude larger than the original one. Therefore, the number of original samples was expanded 20 times by adding Gaussian noise to the original data, while maintaining specific noise parameters for each type of variable to ensure physical consistency. In total, therefore, 580 (=29×20) new samples were generated. These new samples were concatenated with the original 29 ones, bringing the total augmented dataset to 609 samples. The noise parameters were selected based on measurement uncertainties of laboratory equipment (e.g., ±0.05 g for cuticle sample mass), and typical analytical errors in lab analysis equipment, i.e., 1.75% for moisture content, [1, 1, 0.25 %] for proximate analysis (FC, VM, ASH), and [1, 0.5, 0.1, 1 %] for ultimate analysis (C, H, N, O). Notably, a constrained noise was applied to compositional data to preserve the congruence check (100% sum) in both proximate and ultimate analyses, thus reflecting the physical reality. For proximate analysis, the perturbed values were rescaled to sum to 100%. For ultimate analysis, the components were adjusted while conserving the value of the perturbed ash content from proximate analysis, thus guaranteeing the sum to 100%.

To ensure a fair comparison, the Neural Network architecture for the augmented dataset was identical to that previously adopted for the original 29 samples (see Figure 2). The ANN model performance was evaluated using a 5-fold cross-validation approach. For each fold, the data were normalized using z-score standardization and split into training (70%), validation (15%), and testing (15%) subsets. This cross-validation strategy was applied to both the original dataset (29 samples) and the augmented one (609 samples) to assess the model performance across different data partitions.

Limitations of the Authors’ approach include: potential amplification of systematic measurement errors in proximate and ultimate analyses, assumption of Gaussian error distribution for process variables that might not fully reflect real process variations, and inability to introduce new physical relationships beyond those captured in the original 29 experimental tests. Nevertheless, the augmented dataset maintains the fundamental characteristics and physical constraints of torrefaction measurements, particularly the compositional sum constraints and the conservation of the ash content between proximate and ultimate analyses.

Immagine che contiene testo, diagramma, Diagramma, Carattere

Descrizione generata automaticamente

Figure 3: Comparison of predictions vs. real values: Original dataset vs. Augmented dataset in ANN modeling.

* + 1. Results

The performance comparison between the original and augmented ANN models is carried out through the coefficient of determination R², which indicates how well the predicted values align with the actual values on the parity line (y=x). Figure 3 shows the results from the 5-fold cross-validation procedure, comparing ANN predictions versus real values for both the original and augmented datasets. Each point represents a prediction made when that sample was used in the test set during cross-validation, thus including predictions for all samples from both the original dataset (29 samples) as blue dots and the augmented one (609 samples) as red crosses. The Neural Network model applied to the original dataset reveals poor predictive capability, with points scattered far from the parity line and most output variables exhibiting negative R² values. Only Nitrogen and Carbon show positive R² values, with Nitrogen reaching a moderately positive value (0.49) and Carbon showing a marginal positive value (0.03). All in all, substantial improvements in modeling accuracy are disclosed by data augmentation techniques, with the predictions from the Neural Network model applied to the augmented dataset exhibiting consistently positive R² values. The Mass Yield, for instance, improved dramatically from R² = -2.60 to 0.52 (indicating that about 52% of the points align well with the ideal prediction line), while Fixed Carbon (FC) shifted from -1.33 to 0.39. This systematic improvement across all outputs demonstrates that the augmented model better captures the underlying relationships in the torrefaction process.

The analysis of the individual prediction Mean Squared Errors (MSE) in Figure 4a provides further evidence of the augmentation strategy's effectiveness. Here the MSE was first calculated between the ANN model predictions and the experimental data using the original dataset only; then, the MSE was computed for the predictions using the augmented dataset. The most striking improvements are observed in Mass Yield and VM, where the original model predictions showed extremely high errors (335.12 and 246.76, respectively). The augmented model reduced these values by approximately 77% for Mass Yield (to 78.52) and 71% for VM (to 72.51). Similar substantial improvements are seen in the prediction of Carbon, where MSE decreased from 36.70 to 16.10, and Oxygen, dropped from 60.52 to 19.17. The ASH prediction showed particularly notable improvement, with MSE reducing from 45.86 to 4.61, representing a remarkable 90% improvement.

The overall MSE comparison in Figure 4b provides compelling evidence of the global improvement achieved through data augmentation. The total MSE for the original model was 869.82, while the augmented model achieved a total MSE of 259.42, representing an overall improvement of 70.18%.

Immagine che contiene testo, schermata, diagramma, Diagramma

Descrizione generata automaticamenteImmagine che contiene testo, schermata, Diagramma, diagramma

Descrizione generata automaticamente

*a) b)*

Figure 4: Bar chart comparison of the Mean Squared Error (MSE) between ANN predictions and experimental data; a) individual MSE for each output variable with the original and the augmented dataset. b) overall MSE.

* 1. Conclusions

The availability of an augmented neural network model represents a significative step forward in the field of biomass torrefaction, where multiple physical and chemical transformations take place simultaneously. The collection of experimental data, including key process parameters and detailed compositional analyses, provided the initial basis for the modeling approach. In general, the results validate the effectiveness of the chosen augmentation approach in addressing the limitations of the original dataset while maintaining physical consistency in the predictions. The data augmentation strategy, carefully designed to preserve physical constraints and relationships within the experimental measurements, demonstrates how limited datasets from complex thermochemical processes can be effectively expanded while maintaining their fundamental characteristics. This methodological approach proves particularly valuable for torrefaction, where extensive experimental campaigns are both costly and time-consuming due to the complexity of the process, the variability in the feedstocks and the sophisticated analytical techniques required.

While the Authors’ approach to data augmentation shows promising results, future work should explore complementary techniques like transfer learning, especially for industrial scale-up where process conditions may considerably differ from laboratory settings. The effectiveness of the Authors’ ANN model might be reduced when dealing with substantially different feedstocks or operating conditions outside the experimental range. Integration with physics-based models could potentially address some of these limitations.

A possible industrial applicability of the Authors’ ANN model will require large-scale data collection/analysis and further testing/validation under actual industrial operating conditions. This is strongly welcome in the broader context of sustainable process industry operation and renewable energy production, where torrefied biomass plays an increasingly important role as a high-quality renewable solid fuel.

Acknowledgments

National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.3, Call for tender No.1561 of 11.10.2022 by Ministero Università e Ricerca (MUR); funded by the European Union, NextGenerationEU:  Award No.: Project code PE0000021, Concession Decree by Ministero Università e Ricerca, CUP: I53C22001450006, Project title “Network 4 Energy Sustainable Transition – NEST”.

References

Balhorn, L. S., Hirtreiter, E., Luderer, L., & Schweidtmann, A. M. (2023). Data augmentation for machine learning of chemical process flowsheets. In Computer Aided Chemical Engineering 52, 2011-2016).

Brachi, P., Miccio, F., Miccio, M., & Ruoppolo, G. (2016). Torrefaction of Tomato Peel Residues in a Fluidized Bed of Inert Particles and a Fixed-Bed Reactor. Energy and Fuels, 30(6), 4858–4868.

Chen, J., Lu, Z., Jian, J., Bao, Z., Cai, J., & Yao, S. (2023). Effect of torrefaction on yield, reactivity and physicochemical properties of pyrolyzed char from three major biomass constituents. Journal of Analytical and Applied Pyrolysis, 173, 106104.

Dong, H., Wang, J., Wu, X., Zhou, M., & Lü, J. (2023). Gaussian noise data augmentation-based delay prediction for high-speed railways. IEEE Intelligent Transportation Systems Magazine, 15(6), 8-18.

Dou, H. X., Lu, X. S., Wang, C., Shen, H. Z., Zhuo, Y. W., & Deng, L. J. (2022). PatchMask: A data augmentation strategy with Gaussian noise in hyperspectral images. Remote Sensing, 14(24), 6308.

Głód, K., Lasek, J. A., Supernok, K., Pawłowski, P., Fryza, R., & Zuwała, J. (2023). Torrefaction as a way to increase the waste energy potential. Energy, 285, 128606.

Hu, H., Zhang, J., Wang, T., & Wang, P. (2022). Adsorption of toxic metal ion in agricultural wastewater by torrefaction biochar from bamboo shoot shell. Journal of Cleaner Production, 338, 130558.

Khairy, M., Amer, M., Ibrahim, M., Ookawara, S., Sekiguchi, H., & Elwardany, A. (2024). The influence of torrefaction on the biochar characteristics produced from sesame stalks and bean husk. Biomass Conversion and Biorefinery, 14(15), 17127-17148.

Lin, S. L., Zhang, H., Chen, W. H., Song, M., & Kwon, E. E. (2023). Low-temperature biochar production from torrefaction for wastewater treatment: a review. Bioresource technology, 129588.

Miccio, M., Brachi, P., Ruoppolo, G., Miccio, F., Chirone, R., & Tauleigne, B. (2023). Fluidized Bed Design and Process Calculations for the Continuous Torrefaction of Tomato Peels with Solid Product Separation. Chemical Engineering Transactions, 99, 49-54.

Miccio, M., Fraganza, M., Cascone, G., Pataro, G., Rrucaj, E., Ferrari, G., Brachi, P., Migliaccio, R., Favot, M., Marchi, F., Le Nouaille, E., 2024, Final Upcycling of Post-treatment Agri-food Residues by Torrefaction, Chemical Engineering Transactions, 109, 373-378.

Naveed, M. H., Gul, J., Khan, M. N. A., Naqvi, S. R., Štěpanec, L., & Ali, I. (2024). Torrefied biomass quality prediction and optimization using machine learning algorithms. Chem. Eng. Journal Advances, 19, 100620.

Nijman, S. W. J., Leeuwenberg, A. M., Beekers, I., Verkouter, I., Jacobs, J. J. L., Bots, M. L., & Debray, T. P. A. (2022). Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. Journal of Clinical Epidemiology, 142, 218-229.

Oladosu, K. O., Babalola, S. A., Ajao, R. K., & Erinosho, M. F. (2024). Torrefaction of Bambara Groundnut Shell: experimental optimization and prediction of the energy conversion efficiency using statistical and machine learning approaches. International Journal of Ambient Energy, 45(1), 2277309.

Panwar, N. L., & Divyangkumar, N. (2024). An overview of recent advancements in biomass torrefaction. Environment, Development and Sustainability, 1-48.

Rebuffi, S. A., Gowal, S., Calian, D. A., Stimberg, F., Wiles, O., & Mann, T. A. (2021). Data augmentation can improve robustness. Advances in Neural Information Processing Systems, 34, 29935-29948.

Su, G., & Jiang, P. (2024). Machine learning models for predicting biochar properties from lignocellulosic biomass torrefaction. Bioresource Technology, 399, 130519.

Thengane, S. K., Kung, K. S., Gomez-Barea, A., & Ghoniem, A. F. (2022). Advances in biomass torrefaction: Parameters, models, reactors, applications, deployment, and market. Progress in Energy and Combustion Science, 93. https://doi.org/10.1016/J.PECS.2022.101040.

Valenti, D., Denaro, G., Conversano, F., Brunet, C., Bonanno, A., Basilone, G., Mazzola S., & Spagnolo, B. (2016). The role of noise on the steady-state distributions of phytoplankton populations. Journal of Statistical Mechanics, 2016, 054044.

Wei, S., Chen, Z., Arumugasamy, S. K., & Chew, I. M. L. (2022). Data augmentation and machine learning techniques for control strategy development in bio-polymerization process. Environmental Science and Ecotechnology, 11, 100172.

Zhang, Y., Wang, L., Wang, X., Zhang, C., Ge, J., Tang, J. & Duan, H. (2021). Data augmentation and transfer learning strategies for reaction prediction in low chemical data regimes. Organic Chemistry Frontiers, 8(7), 1415-1423.