|  |  |
| --- | --- |
| 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: David Bogle, Flavio Manenti, Piero Salatino  Copyright © 2025, AIDIC Servizi S.r.l. **ISBN** 979-12-81206-21-2; **ISSN** 2283-9216 | |

Machine Learning Ensemble Strategy for HTC Reactor Modelling Under Data Scarcity

Bartolomeo Cosenzaa, Antonio Piconeb, Alessandro Concasc, Federico Atzoric, Antonio Messineob, Chiara Corradod, Maurizio Volpeb\*

aDepartment of Civil and Industrial Engineering, University of Pisa, Largo Lucio Lazzarino, 56122, Pisa, Italy

bDepartment of Engineering and Architecture, University of Enna, Kore, cittadella universitaria, 94100, Enna, Italy

cDepartment of Mechanical, Chemical and Materials Engineering, University of Cagliari, Via Marengo 2, 09123 Cagliari, Italy

dDepartment of Biomedicine, Neurosciences and Advanced Diagnostics, University of Palermo, Via Divisi 83, 90133, Palermo, Italia

\*Corresponding author: [maurizio.volpe@unikore.it](mailto:maurizio.volpe@unikore.it)

This study investigates an adaptive ensemble machine learning approach to predict HTC reactor output parameters under severe data scarcity conditions. Three complementary learning algorithms were selected: Neural Networks (NN), capable of capturing complex non-linear relationships but sensitive to data scarcity; Support Vector Regression (SVR), known for its robustness in high-dimensional spaces and generalization capability with limited samples; and TreeBagger (Random Forest), which provides stable predictions through ensemble averaging of decision trees. These models were integrated through a hybrid strategy that dynamically combines model predictions using optimized weights determined through systematic grid search.

Results demonstrate significant improvements in critical parameters. Where combined predictions do not improve upon the best single model, the approach automatically defaults to the superior individual predictor, ensuring robust performance across all parameters. This adaptive framework proves particularly valuable for HTC process optimization under data scarcity, offering a practical solution that maximizes information extraction from limited experimental data.

* 1. Introduction

The global demand for clean energy and the increasing agro-industrial waste production have driven research towards sustainable energy recovery strategies. In this difficult context, the development of efficient conversion technologies for waste valorisation has become crucial. Hydrothermal carbonization (HTC) represents an effective technology for converting wet agro-wastes into hydrochars - carbon-rich, energy-dense materials with multiple potential applications (Picone et al., 2021; Volpe, M. et al. 2020). The process operates under subcritical conditions in sealed reactors with water present as both reaction medium and reactant. Unlike conventional thermochemical processes such as torrefaction and pyrolysis (Volpe, R., et al., 2016), HTC doesn't require feedstock pre-drying, making it more energy and cost-efficient, especially for high-moisture biomass processing (Yu et al., 2024, Volpe, M. et al., 2023) and even in municipal sludge treatment (Zhi et al., 2024). In HTC processes traditional mathematical modeling faces significant challenges due to the complexity of simultaneous reactions, the high number of variables involved (temperature, pressure, residence time, biomass composition), and the non-linear relationships between process parameters and final product properties. Machine learning has emerged as a powerful alternative approach by leveraging experimental data to identify complex patterns and relationships that are difficult to capture through conventional mathematical equations. This has led to increasing adoption of ML techniques in HTC research and optimization. Djandja et al. (2023) made a significant contribution by studying the prediction of hydrochar fuel properties obtained from co-hydrothermal carbonization, specifically focusing on the combined use of sewage sludge and lignocellulosic biomass, demonstrating the effectiveness of this approach in optimizing final product properties. A complete and systematic overview of machine learning integration in hydrothermal biomass treatment was provided by Zhang et al. (2023), highlighting emerging trends and future challenges in this field. Bacci di Capaci et al., (2023) proposed an integrated approach to the hydrothermal carbonization of sewage sludge. More recently significant progress has also been made in optimizing related processes, as demonstrated by Ma et al. (2024), who developed an interpretable ensemble prediction model for anaerobic digestion performance of HTC wastewater, addressing one of the most critical aspects of by-product management. Leng et al. (2024) took a further step forward using machine learning to simultaneously optimize operating parameters and biomass mixing recipes, opening new possibilities for precise process control. Katongtung et al. (2024)'s work on multifaceted predictive modeling highlighted the versatility of the machine learning approach in HTC; Zhang et al. (2024) further expanded these concepts by developing an intelligent system for multi-target prediction of hydrochar properties, demonstrating the possibility of simultaneously predicting multiple product characteristics. This intense research activity demonstrates the growing interest of the scientific community in applying machine learning to HTC and its potential for improving the understanding and optimization of this important biomass conversion process.

* 1. Ensemble learning

Ensemble learning emerged from the fundamental observation that in complex decision-making scenarios, combining multiple perspectives often leads to better outcomes than relying on a single viewpoint (Rane et al., 2024). This principle is particularly valuable in contexts like HTC process modeling, where experimental data scarcity presents significant challenges due to the resource-intensive and time-consuming nature of reactor experiments. Under such limited data conditions, individual models often struggle to generalize effectively, exhibiting high variance or bias. Ensemble methods address this limitation by integrating diverse modeling approaches, effectively extracting maximum information from the available experimental data while reducing prediction uncertainty. This approach creates a robust methodology for modeling complex thermochemical processes where collecting additional experimental data may be prohibitively expensive or time-consuming. While other techniques such as data augmentation exist to address data scarcity (Cosenza et al., 2024) the ensemble approach provides a direct way to maximize the utility of limited experimental datasets through model diversity and optimized combination strategies (Satoła and Satoła, 2024). The approach gained significant traction as researchers recognized that different models, each with their own strengths and biases, could complement each other to produce more reliable and robust predictions. The methodology builds on the statistical principle that aggregating multiple independent estimates can reduce variance and improve accuracy, similar to how averaging multiple measurements can provide a more precise estimate than a single measurement. Grounded in the principle of collective decision-making, the theoretical underpinnings of ensemble learning have solidified its status as a cornerstone of modern machine learning.

A comprehensive survey by Dong et al. (2020) provided a vital framework for understanding ensemble methods and their core methodologies. Building on that research, Mahajan et al. (2023) showcased the approach’s usefulness in disease prediction and medical diagnostics, underscoring its versatility. The versatility of ensemble learning has led to its adoption across numerous fields. Wang et al. (2024) utilized ensemble learning for parameter analysis in sorption-enhanced steam methane reforming; Dai and Allman (2024) developed a boosted approach for model predictive control in modular facilities; Jaiyeoba et al. (2024) proposed a model to classify skin diseases using stacking ensemble machine learning techniques. This remarkable progression from theoretical foundations to diverse practical applications demonstrates ensemble learning's exceptional adaptability. Its successful implementation across healthcare, agricultural monitoring, industrial processes, chemical engineering, environmental management, and transportation systems underscores its robustness in handling complex, multi-variable problems. The methodology's ability to consistently deliver superior performance across such varied domains has cemented its position as an indispensable tool in modern data analysis and prediction tasks.

* 1. Materials and methods

The hybrid ensemble learning developed by the authors integrates weighted average (Zamani et al., 2023) and selective ensemble (Liu et al., 2023) techniques, exploiting the strengths of both methods. Weighted averaging consolidates outputs from models by applying a weighted mean, where each weight is tuned to lessen error. On the other hand, selective ensemble chooses the best-performing model for each individual prediction. This hybrid strategy automatically determines which method to use: if the optimal weighted combination yields better results than the best single model, it is employed; otherwise, the top-performing model is selected. Such an adaptive approach allows for the benefits of combination when advantageous, while retaining the robustness of model selection as an alternative. The success of these ensemble strategies relies on three crucial elements working in harmony: the diversity among the individual models, which ensures different perspectives on the problem; the baseline competence of each model, ensuring that each contributor performs better than random chance; and the effectiveness of the combination strategy (Brownlee, 2021). These elements together enable ensemble methods to integrate various types of models - from decision trees to neural networks and support vector machines - creating robust predictive systems that consistently outperform individual models, particularly in complex scenarios where single models might struggle to capture all aspects of the underlying relationships. The methodology employs a systematic grid search to find optimal weight combinations for each output variable, leveraging the complementary strengths of individual models through error compensation.

The ensemble learning was implemented in MATLAB code presents a sophisticated machine learning system designed for predicting and analyzing reactor characteristics through a comparative and ensemble approach. At its core, the system implements three distinct base models, each chosen for their specific strengths in handling complex regression tasks:

* The Neural Network (NN) architecture features a single hidden layer with three neurons, employing the Levenberg-Marquardt algorithm for efficient training over 100 epochs.
* The Support Vector Regression (SVR) component utilizes a Radial Basis Function kernel with automatic scaling optimization, handling each output variable independently to ensure specialized prediction capability.
* The Random Forest, implemented through MATLAB's TreeBagger (TB) with 10 decision trees and optimized leaf size, adds robust ensemble-based prediction to the mix.

The system's innovative aspect lies in its hybrid ensemble methodology, which dynamically evaluates and combines these models. Rather than applying a one-size-fits-all approach, the ensemble analyzes performance metrics for each output variable independently, using grid search optimization to determine whether a single model or a weighted combination would yield better results. The algorithm allows the system to leverage the strengths of each model where they perform best, with weights explored in 0.2 increments to find optimal combinations. The prediction targets encompass crucial reactor parameters including solid, liquid, and gas fractions, volatile matter, ash content, fixed carbon, higher heating value, and energy yield. Physical consistency is maintained through carefully implemented constraints, ensuring that relevant parameters sum to 100%. Performance evaluation is conducted using Mean Absolute Percentage Error (MAPE), with results presented through comprehensive visualization tools including comparative plots and detailed performance metric tables. This systematic approach enables both accurate predictions and insightful analysis of model behavior across different reactor characteristics.

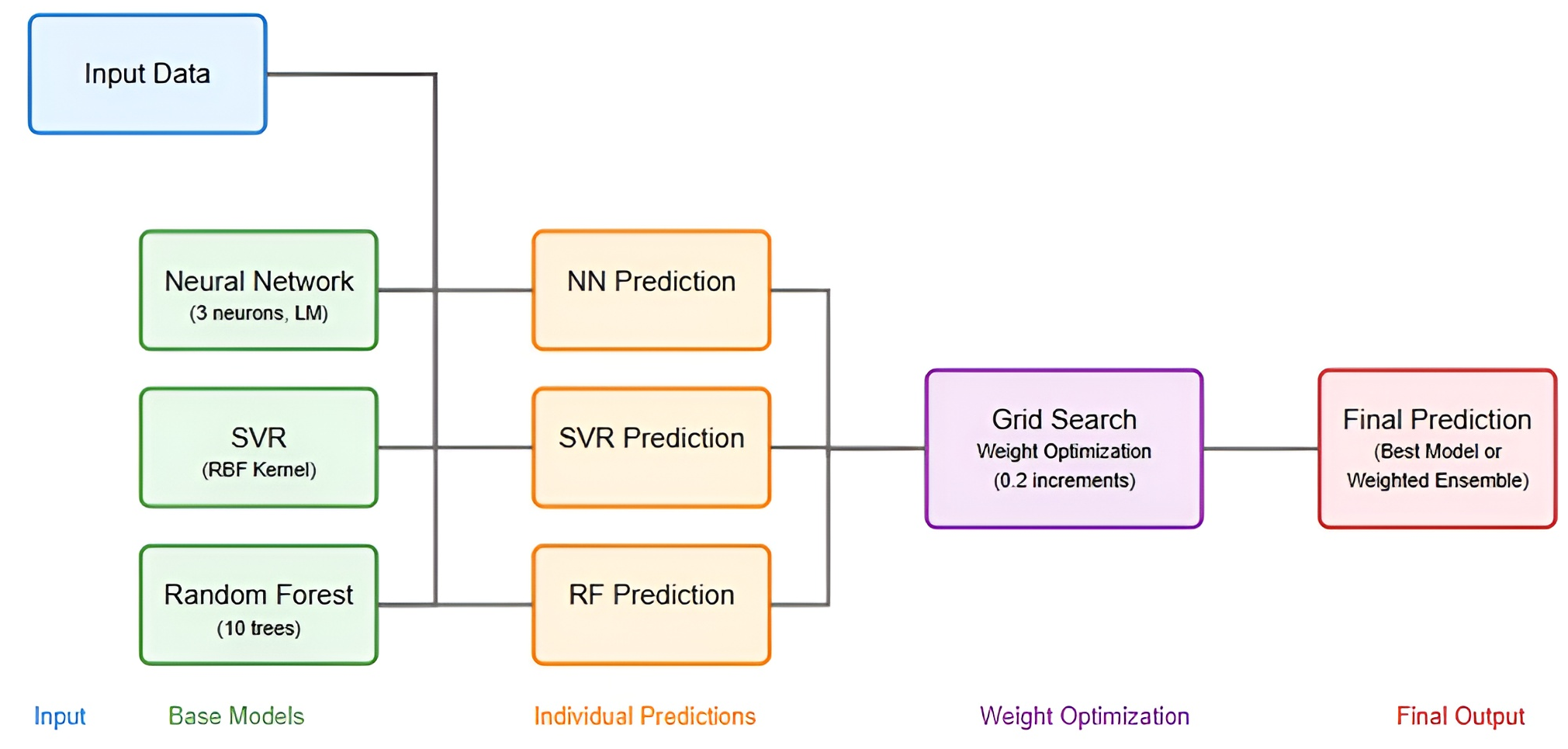


Figure 1: Schematic representation of the hybrid ensemble learning system architecture.

* 1. Results

The results are presented through two complementary visualizations that demonstrate the ensemble's effectiveness. Figure 2 shows time series comparisons across five test samples for each output variable (Solid, Liquid, Gas, VM, ASH, FC, HHV, and EY). The plots reveal how the hybrid ensemble (magenta line) generally tracks the actual values (black line) more closely than individual models. In the Solid fraction predictions, the ensemble effectively captures the sharp increase between samples 3 and 4, while individual models show more deviation. For Liquid predictions, the ensemble successfully follows the trend of the actual values, particularly in tracking the peak at sample 2. In the Gas predictions, all models show reasonable tracking of the declining trend, but the ensemble maintains the most consistent accuracy across all samples.

The VM (Volatile Matter) predictions show interesting behavior where SVR (blue dashed line) tends to overestimate the values, while the ensemble maintains closer alignment with actual measurements. The ASH content predictions present a challenging case with high variability, particularly evident in the sharp dip at sample 2, where all models, including the ensemble, struggle to capture the extreme variation. For FC (Fixed Carbon), the ensemble effectively tracks the general declining trend, though some deviations are observed in the middle samples.

**Immagine che contiene diagramma, Diagramma, linea, testo

Il contenuto generato dall'IA potrebbe non essere corretto.**

Figure 2: Time series comparison of actual values versus predictions from different models (Neural Network, SVR, TreeBagger, and Hybrid Ensemble) across eight reactor output variables.

Figure 3 shows the MAPE comparison across models, where lower values indicate better performance. The bar chart clearly illustrates how the ensemble (magenta bars) consistently achieves lower error rates across most output variables. This improvement is particularly evident for Gas predictions, where the ensemble achieves a MAPE of 3.4% compared to TreeBagger (6.1%), SVR (14.3%), and Neural Network (20.6%). For VM, the ensemble achieves a MAPE of 1.8%, performing better than the individual models. The ASH predictions show higher error rates across all models, with MAPE values above 29%. The HHV (Higher Heating Value) predictions show the ensemble achieving a MAPE of around 2.2%, notably better than the individual models. The Energy Yield (EY) predictions also demonstrate the ensemble's superior performance, with consistently lower error rates compared to individual models. The MAPE comparison reveals that Neural Networks tend to perform less consistently across different variables, showing particularly high error rates in Gas predictions (20.6%) and ASH content (approximately 36%). SVR shows more stable performance but still generally higher error rates than the ensemble. TreeBagger often performs as the second-best model, suggesting its robustness in handling this type of data. It should be emphasized that the limited visualization of a five-sample test set does not fully capture the ensemble's superiority, which is clearly demonstrated through improved MAPE metrics. While individual models exhibit variable performance across different parameter regions, the hybrid ensemble maintains more consistent accuracy by employing optimized weighted combinations that counteract the specific weaknesses of each model. This systematic reduction in error is statistically evident, but a visualization encompassing the complete dataset would be required to fully appreciate the ensemble's capacity to deliver reliable predictions across the entire parameter space.

Immagine che contiene schermata, linea, Diagramma, diagramma

Il contenuto generato dall'IA potrebbe non essere corretto.

Figure 3: MAPE (Mean Absolute Percentage Error) comparison across different models (Neural Network, SVR, TreeBagger, and Ensemble) for eight reactor output variables.

* 1. Conclusions

The ensemble learning approach proves particularly valuable in reactor applications, addressing several critical challenges in process modeling and optimization. When dealing with reactor experiments that are both costly and time-intensive, data availability often becomes a limiting factor. The ensemble method effectively maximizes the utility of limited experimental data by integrating diverse modeling perspectives. The approach is especially powerful for capturing the complex, non-linear relationships inherent in reactor behavior, where different models excel at capturing distinct aspects of the system dynamics. The robust nature of ensemble learning emerges from its ability to mitigate the limitations and biases of individual models. By implementing a sophisticated grid search mechanism for weight optimization, the system adapts to the specific characteristics of each output variable, ensuring reliable predictions across all reactor parameters. This adaptive capability is particularly evident in the results, where the ensemble consistently outperforms individual models, achieving notable improvements in critical parameters like gas production and volatile matter content. The hybrid ensemble's ability to automatically adjust model weights based on performance makes it particularly suitable for industrial applications where prediction reliability is crucial. The proposed ensemble strategy not only improves prediction accuracy but also provides insights into the relative strengths of different modeling approaches for specific reactor parameters.

The success of this approach suggests its potential application in other complex chemical processes where experimental data is limited and process relationships are highly non-linear. Future work could focus on incorporating additional models and exploring dynamic weight adjustment strategies to further enhance prediction accuracy.

Acknowledgments

Authors gratefully acknowledge support of the "SiciliAn MicronanOTecH Research And Innovation Center - SAMOTHRACE" project (MUR, PNRR-M4C2, ECS\_00000022).

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

References

Bacci di Capaci, R., Tasca, A. L., Gori, R., Vitolo, S., Puccini, M., & Pannocchia, G., 2023, An Integrated Approach to the Hydrothermal Carbonization of Sewage Sludge: Simulation, Modeling, and Life Cycle Assessment. Chem. Engineering, 7(3), 44.

Brownlee, J., 2021, Ensemble learning algorithms with Python: Make better predictions with bagging, boosting, and stacking. Machine Learning Mastery.

Cosenza, B., Miccio, M., Pannocchia, G., & Sammarco, I. (2024). Advanced Simulation Model for Studying Biofuel-producing Microalgae Populations. *Chemical Engineering Transactions*, *109*, 559-564.

Dai, Y., & Allman, A., 2024, Boosted Ensemble Learning for Model Predictive Control with Reconfiguration of Modular Facilities. In Computer Aided Chemical Engineering, 53, 1933-1938.

Djandja, O. S., Kang, S., Huang, Z., Li, J., Feng, J., Tan, Z., Lougou, B. G., 2023, Machine learning prediction of fuel properties of hydrochar from co-hydrothermal carbonization of sewage sludge and lignocellulosic biomass. Energy, 271, 126968.

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q., 2020, A survey on ensemble learning. Frontiers of Computer Science, 14, 241-258.

Jaiyeoba, O., Ogbuju, E., Yomi, O. T., & Oladipo, F., 2024, Development of a Model to Classify Skin Diseases using Stacking Ensemble Machine Learning Techniques. Journal of Computing Theories and Applications, 2(1), 22-38.

Katongtung, T., Prasertpong, P., Sukpancharoen, S., Sinthupinyo, S., & Tippayawong, N., 2024, Predictive modeling for multifaceted hydrothermal carbonization of biomass. Journal of Environmental Chemical Engineering, 12(5), 114071.

Leng, L., Zhou, J., Zhang, W., Chen, J., Wu, Z., Xu, D., ... & Li, H., 2024, Machine-learning-aided hydrochar production through hydrothermal carbonization of biomass by engineering operating parameters and/or biomass mixture recipes. Energy, 288, 129854.

Liu, S., Qin, H., Liu, G., Xu, Y., Zhu, X., & Qi, X., 2023, Runoff forecasting of machine learning Model based on selective ensemble. Water Resources Management, 37(11), 4459-4473.

Ma, Z., Wang, R., Song, G., Zhang, K., Zhao, Z., & Wang, J, 2024, Interpretable ensemble prediction for anaerobic digestion performance of hydrothermal carbonization wastewater. Science of The Total Environment, 908, 168279.

Mahajan, P., Uddin, S., Hajati, F., & Moni, M. A., 2023, Ensemble learning for disease prediction: A review. In Healthcare, 11(12), 1808)

Picone, A.; Volpe, M.; Giustra, M.G.; Di Bella, G.; Messineo, A., 2021 Hydrothermal carbonization of lemon peel waste: Preliminary results on the effects of temperature during process water recirculation, 2021, Applied System Innovation, 4(1), 19.

Rane, N., Choudhary, S. P., & Rane, J., 2024, Ensemble deep learning and machine learning: applications, opportunities, challenges, and future directions. Studies in Medical and Health Sciences, 1(2), 18-41.

Satoła, A., & Satoła, K., 2024, Performance comparison of machine learning models used for predicting subclinical mastitis in dairy cows: Bagging, boosting, stacking, and super-learner ensembles versus single machine learning models. Journal of Dairy Science, 107(6), 3959-3972.

Wang, L., Li, H., Du, C., & Hong, W., 2024, Sorption-enhanced steam methane reforming parameter analysis and performance prediction of ensemble learning methods using improved drag model. Advanced Powder Technology, 35(8), 104576.

Yu, S., He, J., Zhang, Z., Sun, Z., Xie, M., Xu, Y., ... & Zhou, H., 2024, Towards negative emissions: Hydrothermal carbonization of biomass for sustainable carbon materials. Advanced Materials, 36(18), 2307412.

Volpe, M., Fiori, L., Merzari, F., Messineo, A.; Andreottola, G., 2020, Hydrothermal carbonization as an efficient tool for sewage sludge valorization and phosphorous recovery, Chemical Engineering Transaction, 80, 199-204.

Volpe, M., Codignole Luz, F., Saha, N., Reza, M.T., Chelang’at Mosonik, and Messineo, A., 2023, Enhancement of energy and combustion properties of hydrochar via citric acid catalysed secondary char production, Biomass Conv. Bioref., 13, 10527–10538.

Volpe, R., Messineo, S., Volpe, M., Messineo, A., 2016, Catalytic effect of char for tar cracking in pyrolysis of citrus wastes, design of a novel experimental set up and first results, Chemical Engineering Transactions, 50, 181-186.

Zamani, M. G., Nikoo, M. R., Jahanshahi, S., Barzegar, R., & Meydani, A., 2023, Forecasting water quality variable using deep learning and weighted averaging ensemble models. Environmental Science and Pollution Research, 30(59), 124316-124340.

Zhang, W., Zhou, J., Liu, Q., Xu, Z., Peng, H., Leng, L., & Li, H., 2024, A novel intelligent system based on machine learning for hydrochar multi-target prediction from the hydrothermal carbonization of biomass. Biochar, 6(1), 19.

Zhang, W., Chen, Q., Chen, J., Xu, D., Zhan, H., Peng, H., ... & Li, H., 2023, Machine learning for hydrothermal treatment of biomass: A review. Bioresource Technology, 370, 128547.

Zhi, Y., Xu, D., Jiang, G., Yang, W., Chen, Z., Duan, P., & Zhang, J., 2024, A review of hydrothermal carbonization of municipal sludge: Process conditions, physicochemical properties, methods coupling, energy balances and life cycle analyses. Fuel Processing Technology, 254, 107943.