PREDICTING BIOCHEMICAL METHANE POTENTIAL OF AGRICULTURAL RESIDUES USING MACHINE LEARNING: A RANDOM FOREST REGRESSION MODEL LEVERAGING BIOMASS CHARACTERISTICS

Daniel O. Fasheun[1,2*], Folorunsho B. Omage[3,4], & Viridiana S. Ferreira-Leitão[1,2]

[1] *Laboratório de Biocatálise, Instituto Nacional de Tecnologia (INT), Rio de Janeiro, RJ, Brazil,.*
[2] *Departamento de Bioquímica, Instituto de Química, Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, RJ, Brazil.*
[3] *Department of Organic Chemistry, Institute of Chemistry, University of Campinas (UNICAMP), Campinas, SP, Brazil*
[4] *Computational Biology Research Group, Embrapa Digital Agriculture, Campinas, São Paulo, Brazil*
** Corresponding author's email address: daniel.fasheun@int.gov.br*

ABSTRACT

This study presents a random forest regression machine learning model developed to predict the biochemical methane potential (BMP) of various agricultural feedstocks using cellulose, hemicellulose, and lignin content as independent variables. The model demonstrated strong performance on an independent test dataset, achieving R^2 score, RMSE and MAE of 0.84, 284.31 Nm³ CH_4/ton fresh mass (FM), and 171.66 Nm³ CH_4/ton FM, respectively, confirming the model's accuracy and consistency. SHAP analysis revealed that hemicellulose content is the most significant predictor of BMP, with lignin and cellulose also contributing notably. These findings align with existing literature on the possible recalcitrant nature of different feedstocks and suggest that lower levels of these predictor components enhance BMP. The model's capacity to predict BMP based on biomass characteristics without extensive experimental trials highlights its potential to optimize anaerobic digestion systems and improve the efficient utilization of agricultural residues as renewable energy sources.

Keywords: Biochemical methane potential. Machine learning. Lignocellulose. Biogas. Anaerobic digestion.

1 INTRODUCTION

Biogas production through anaerobic digestion has emerged as a promising renewable energy solution, converting diverse organic feedstocks into methane-rich biogas.[1,2] To quantify the biodegradability of these feedstocks, numerous studies have conducted experimental Biochemical Methane Potential (BMP) tests under anaerobic digestion conditions.[3,4] BMP tests determine the maximum methane generation from a single substrate on a laboratory scale [5], and international inter-laboratory studies have optimized protocols and defined best practices.[3] However, these tests are complex and time-consuming[6], necessitating accurate and timely BMP predictions for the design, optimization, and management of efficient biogas production systems.

Recent research emphasizes the significant impact of lignocellulosic composition—specifically the proportions of cellulose, hemicellulose, and lignin—on the biomethane production potential of various feedstocks, including agricultural residues, energy crops, lignocellulosic biomass, manure, and slurries.[2,5] In a recent study, predictive models were developed to estimate the BMP of lignocellulosic feedstocks based their chemical composition, using machine learning models augmented with Generative Adversarial Network (GAN) data.[6] However, our study presents a robust random forest regression model designed to predict BMP for a wide range of agricultural biogas feedstocks, offering insights into the influence of cellulose, hemicellulose, and lignin content on biomethane production. The model's performance, evaluated using key metrics and SHAP analysis, contributes to optimizing anaerobic digestion systems and the efficient use of diverse agricultural feedstocks.

2 MATERIAL & METHODS

The dataset utilized in this study was sourced from the comprehensive database developed by Lallement et al. (2023)[5], featuring detailed biomass characteristics and corresponding BMP values. This study focused on three key predictors: cellulose, hemicellulose, and lignin contents, with BMP as the target variable. To ensure consistency in BMP predictions across various feedstocks, units were standardized from normal cubic meters of methane per ton of volatile solids (Nm³ CH_4/ton VS) to per ton of fresh mass (Nm³ CH_4/ton FM), using the specific VS content of each sample. Similarly, cellulose, hemicellulose, and lignin measurements were converted from grams per 100 grams of dry matter (g/100g DM) to grams per 100 grams of fresh mass (g/100g FM) to account for moisture content variations. During preprocessing, four outliers were removed from an initial 131 data points, resulting in 127 data points for model development. Descriptive statistics indicated a wide range of cellulose (2.3 to 53 g/100g FM), hemicellulose (1.2 to 21.4 g/100g FM), and lignin (2.3 to 39.7 g/100g FM) contents and BMP values (171 to 3010 Nm³ CH_4/ton FM).

For model development, the independent variables were normalized, and the dataset was split into training (70%) and test sets (30%). The training set was further divided into training and validation sets (80% and 20%, respectively). The random forest regression algorithm, implemented using Python and libraries such as pandas, NumPy, and scikit-learn, was employed to train the model. A 5-fold cross-validation technique assessed the model's learning curve and potential overfitting or underfitting. Additionally, Shapley Additive Explanations (SHAP) analysis[8] provided insights into the importance of input variables in the model's predictions. The model's accuracy was evaluated using the coefficient of determination (R^2), Root Mean Squared Error

(RMSE), and Mean Absolute Error (MAE) on both validation and independent test datasets, ensuring robust predictive performance and generalization capabilities.

3 RESULTS & DISCUSSION

The random forest regression model developed in this study showed strong performance in predicting BMP of various feedstocks based on their cellulose, hemicellulose, and lignin content (Figure 1). On the validation set, the model achieved an R^2 score of 0.88, indicating that it could explain 88% of the variation in BMP values. The RMSE was 242.90 Nm³ CH_4/ton FM, and the MAE was 192.63 Nm³ CH_4/ton FM. These metrics suggest that while the model captures the overall trend of BMP variation accurately, there remains a moderate level of error in individual predictions.

To ensure the model's generalizability, its performance was also evaluated on an independent test set. The results revealed an R^2 score of 0.84, which, while slightly lower than the validation set score, still indicates strong predictive capability. The RMSE and MAE on the test set were 284.31 Nm³ CH_4/t FM and 171.66 Nm³ CH_4/t FM, respectively. These values are comparable to those from the validation set, underscoring the model's consistency and reliability across different datasets.

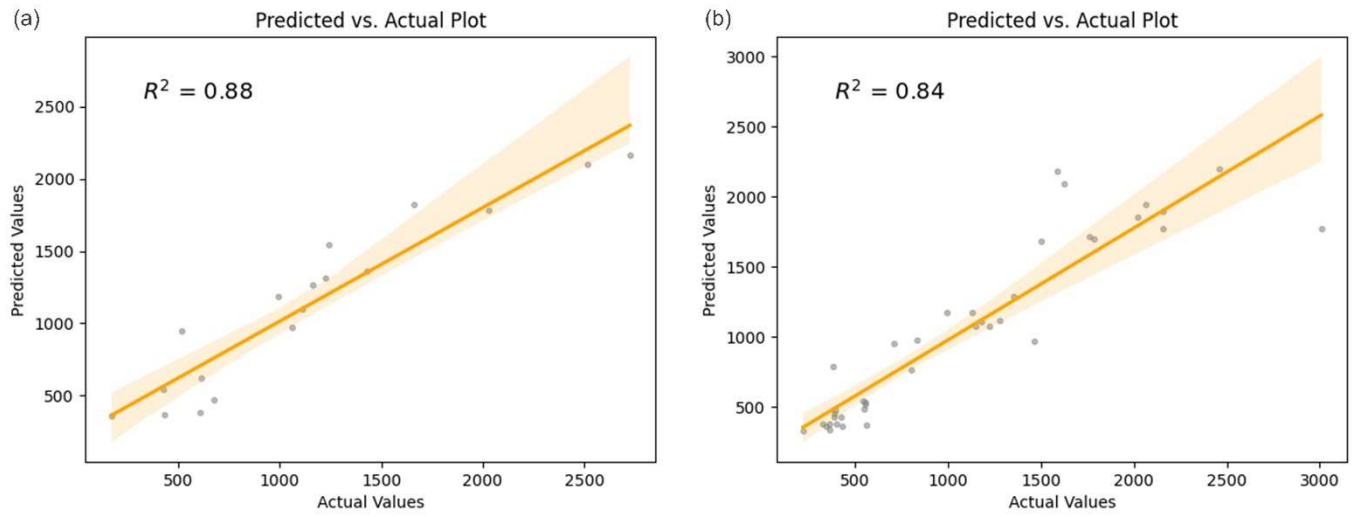

Figure 1: The random forest regression model's R^2 performance on (a) validation and (b) test datasets.

The learning curve analysis demonstrated the model's robustness and capacity to generalize (Figure 2). The training score increased steadily with more training examples, indicating effective learning of data patterns. The validation score also showed strong performance, plateauing after an initial increase, which suggests that the model had reached optimal performance. The minimal gap between training and validation scores indicates a lack of overfitting, further validating the model's reliability in making accurate predictions on new data.

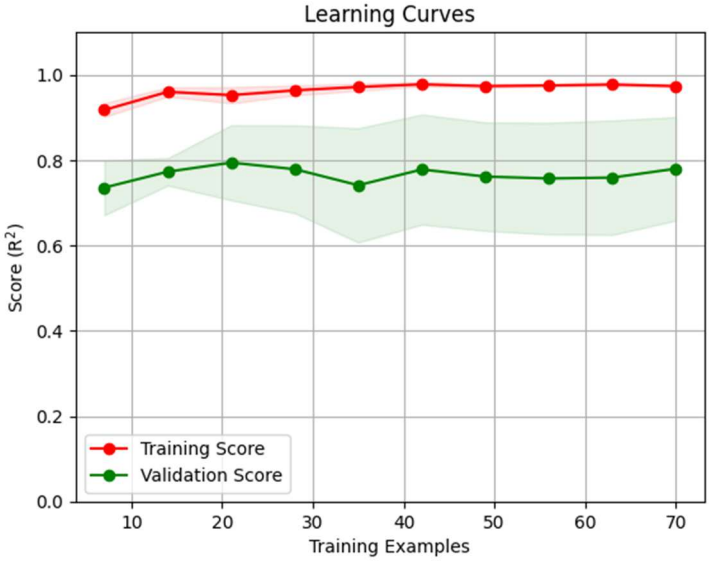

Figure 2: Learning curves of the random forest regression model on the training set and cross-validation

SHAP analysis was employed to interpret the model's predictions and assess the impact of each input variable (Figure 3). Aggregating SHAP values across the dataset provided insights into the overall significance of the input variables. Hemicellulose was identified as the most significant contributor to BMP prediction, with lignin and cellulose also playing important roles.

Biochemically, hemicellulose works with cellulose and lignin to provide structural integrity to plant cell walls, which resist degradation. [9,10] The SHAP values indicated that lower hemicellulose content typically led to higher BMP values. Similarly, lower levels of lignin and cellulose were associated with higher BMP, though to a lesser extent than hemicellulose. This suggests that low contents of these components in agricultural biomass could favor biodegradation and thus enhance BMP. Conversely, high levels of these components create barriers that hinder enzymes from accessing cellulose, thereby reducing BMP. The model effectively accounted for these intricate biochemical relationships, enabling accurate BMP predictions for various agricultural feedstocks. This highlights the potential of the model as a reliable tool for optimizing biomass utilization in methane production.

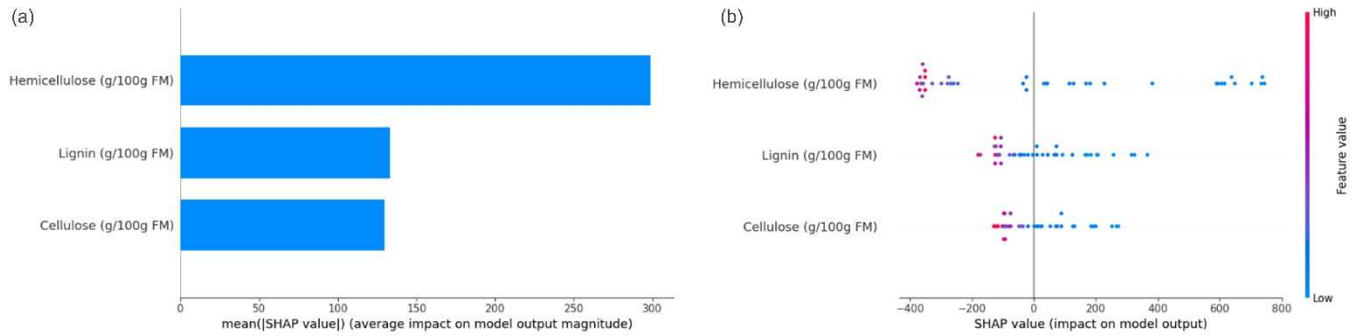

Figure 3: Mean SHAP value plot depicting the average impact of hemicellulose, cellulose and lignin, and (b) SHAP Beeswarm plot showing the individual contributions of hemicellulose, cellulose and lignin to the model's predicted output.

4 CONCLUSION

This study's random forest regression model offers a promising method for predicting the BMP of various agricultural biogas feedstocks. The model demonstrated strong performance on an independent test dataset, with a high R^2 value (0.84), low RMSE (284.31 Nm^3 CH_4/ton FM), and MAE (171.66 Nm^3 CH_4/ton FM), confirming its reliability and applicability. SHAP analysis identified hemicellulose as the most significant factor in BMP prediction, with lignin and cellulose also playing important roles. This insight aligns with existing literature on the recalcitrant nature of biomass with lignin, hemicellulose, cellulose contents. The model's ability to predict BMP based on biomass characteristics without extensive experimental trials saves valuable time and resources, making it a valuable tool for optimizing anaerobic digestion systems and utilizing agricultural residues as renewable energy sources.

REFERENCES

1. Mao C, Feng Y, Wang X, Ren G. Review on research achievements of biogas from anaerobic digestion. Renew Sustain Energy Rev [Internet]. 2015 May;45:540–55. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1364032115001203
2. Archana K, Visckram AS, Senthil Kumar P, Manikandan S, Saravanan A, Natrayan L. A review on recent technological breakthroughs in anaerobic digestion of organic biowaste for biogas generation: Challenges towards sustainable development goals. Fuel [Internet]. 2024;358(PB):130298. Available from: https://doi.org/10.1016/j.fuel.2023.130298
3. Angelidaki I, Alves M, Bolzonella D, Borzacconi L, Campos JL, Guwy AJ, et al. Defining the biomethane potential (BMP) of solid organic wastes and energy crops: A proposed protocol for batch assays. Water Sci Technol. 2009;59(5):927–34.
4. Raposo F, De La Rubia MA, Fernández-Cegrí V, Borja R. Anaerobic digestion of solid organic substrates in batch mode: An overview relating to methane yields and experimental procedures. Renew Sustain Energy Rev [Internet]. 2012;16(1):861–77. Available from: http://dx.doi.org/10.1016/j.rser.2011.09.008
5. Lallement A, Peyrelasse C, Lagnet C, Barakat A, Schraauwers B, Maunas S, et al. A Detailed Database of the Chemical Properties and Methane Potential of Biomasses Covering a Large Range of Common Agricultural Biogas Plant Feedstocks. Waste. 2023;1(1):195–227.
6. Adeleke AA, Okolie JA, Ogbaga CC, Ikubanni PP, Okoye PU, Akande O. Machine Learning Model for the Evaluation of Biomethane Potential Based on the Biochemical Composition of Biomass. Bioenergy Res [Internet]. 2024;17(1):731–43. Available from: https://doi.org/10.1007/s12155-023-10681-9
7. Li Y, Park SY, Zhu J. Solid-state anaerobic digestion for methane production from organic waste. Renew Sustain Energy Rev. 2011;15(1):821–6.
8. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2(1):56–67.
9. Sista Kameshwar AK, Qin W. Understanding the structural and functional properties of carbohydrate esterases with a special focus on hemicellulose deacetylating acetyl xylan esterases. Mycology. 2018;9(4):273–95.
10. Zoghlami A, Paës G. Lignocellulosic Biomass: Understanding Recalcitrance and Predicting Hydrolysis. Front Chem. 2019;7(December).

ACKNOWLEDGEMENTS

This work was supported by the Brazilian National Council for Scientific and Technological Development (CNPq - grant number 306978/2022-9) and Research Support Foundation of the State of Rio de Janeiro (FAPERJ - grant number E-26/210.791/2021).