Evaluation of the Thermogravimetric Profile of Hybrid Cellulose Acetate Membranes using Machine Learning Approaches[☆]

Avaliação do perfil termogravimétrico de membranas híbridas de acetato de celulose empregando abordagens de aprendizado de máquina

Filipi França dos Santos[1,†], Kelly Cristine da Silveira[1], Daniella Herdi Cariello[1], Gesiane Mendonça Ferreira[1], Guilherme de Melo Baptista Domingues[1], Mônica Calixto Andrade[1]

[1]*State University of Rio de Janeiro, Polytechnic Institute, Nova Friburgo, RJ, Brazil*

[†]**Corresponding author:** filipi.santos@iprj.uerj.br

Abstract

Thermogravimetric analysis (TGA) is a characterization technique routinely used in materials science. In this particular case, TGA determines the variation of weight with temperature. The thermogravimetric analysis of cellulose acetate (CA) hybrid membranes can provide very similar results, despite their different chemical composition. The present study uses machine learning algorithms to try to correlate data from thermogravimetric analyses with variations in chemical composition. Experimental points relating to temperature and weight from these analyses were treated in different ways and used to estimate the composition of the membranes. The Extra-Trees Classifier, Random Forest, Decision Tree, and K-Nearest Neighbors (KNN) algorithms were applied to this data and then evaluated using a confusion and accuracy matrix. The decision tree-based algorithms demonstrated a superior capacity for estimating the composition, albeit with negligible disparities in the thermogravimetric profile. The Extra-Trees Classifier algorithm, in particular, stood out for its ability to estimate composition in all tests, achieving 90% accuracy.

Keywords

Cellulose acetate hybrid membranes • Thermogravimetric analysis (TG) • Machine learning

Resumo

A análise termogravimétrica (TGA) é uma técnica de caracterização rotineiramente utilizada na ciência dos materiais. Neste caso particular, a TGA determina a variação de massa com a temperatura. A análise termogravimétrica das membranas híbridas de acetato de celulose (CA) pode fornecer resultados muito semelhantes, apesar de sua composição química diferente. O presente estudo utiliza algoritmos de aprendizado de máquina para tentar correlacionar dados de análises termogravimétricas com variações na composição química. Pontos experimentais relacionados à temperatura e massa dessas análises foram tratados de diferentes maneiras e utilizados para estimar a composição das membranas. Os algoritmos *Extra-Trees Classifier*, *Random Forest*, *Decision*

[☆] This article is an extended version of the work presented at the Joint XXV ENMC National Meeting on Computational Modeling, XIII ECTM Meeting on Science and Technology of Materials, 9th MCSul South Conference on Computational Modeling and IX SEMENGO Seminar and Workshop on Ocean Engineering, held in webinar mode, from October 19th to 21th, 2022

Tree e *K-Nearest Neighbors* (KNN) foram aplicados a esses dados e, em seguida, avaliados usando uma matriz de confusão e de acurácia. Os algoritmos baseados em árvore de decisão mostraram habilidade superior na estimativa da composição, com diferenças menores no perfil termogravimétrico. O algoritmo *Extra-Trees Classifier*, em particular, destacou-se por sua habilidade em estimar a composição em todos os testes, atingindo 90% de acurácia.

Palavras-chave

Membranas híbridas de acetato de celulose • Análise termogravimétrica (TG) • Aprendizado de máquina

1 Introduction

Materials science is an interdisciplinary field that integrates chemistry, physics, and various fields of engineering. However, in recent years, a new area of expertise has emerged: machine learning. As a subfield of artificial intelligence, machine learning aims to equip computers with the ability to learn and solve problems through data [1]. Researchers have since been exploring the potential of this intersection between materials science and machine learning, particularly in the development of advanced materials. One parameter that evidences this trend is the growth in publications linking the two fields in the last decade. From 2014 to 2019, the number of publications per year increased from fewer than 100 to nearly 700 [2].

Regarding the real impact that machine learning has had on the field of materials science, it has been instrumental in addressing significant challenges, such as synthesizing and processing unprecedented materials with innovative properties [2]. With numerous options for compositions and synthesis methods, achieving an optimized structure has always been a significant challenge. The experimental process of producing a material is costly, making it impractical to synthesize and characterize it indiscriminately. Finding an optimal composition and synthesis method to create a material often leads to many unsuccessful attempts, as observed in "The Open Membrane Database," which features numerous unsuitable materials for the desired application [3]. However, what was once a challenge is now accessible and advantageous with the aid of machine learning algorithms.

With the advancement of machine learning in the field of materials science, sophisticated algorithms facilitate simulations focused on material composition, molecular structure, property prediction, and material characterization [2]. This possibility has led to a growing number of studies that use machine learning algorithms to predict the final properties of materials even before their synthesis [4][5][6]. Not only has the process of producing unprecedented materials benefited from this computational tool, but also the characterization of materials. Morgan and Jacobs [2] discuss the impact that machine learning has had on traditional characterization techniques such as scanning electron microscopy (SEM) and X-ray diffraction, assisting in the analysis, processing, and classification of experimental data. In this context, thermogravimetric analysis (TGA), an extensively employed technique in material characterization, appears to be underutilized in the literature. TGA is a method of analyzing thermal degradation curves that relates analysis temperature points and material weight loss, presenting thermal degradation profiles for different materials.

Cellulose acetate membranes are materials of significant importance in industrial applications and the scientific community, encompassing various uses, including seawater desalination, wastewater treatment, and the incorporation of technologies in the pharmaceutical industry. Using cellulose acetate instead of other polymers for membrane synthesis can bring about improved characteristics in the material, such as low toxicity, low production cost, and biodegradability. Chemical modifications in cellulose acetate membranes, making them hybrid, have altered crucial properties such as anti-scaling characteristics and increased permeability, overcoming some of the challenges encountered in their application [7][8].

This study aims to investigate the data obtained from thermal analyses of hybrid cellulose acetate membranes synthesized by Ferreira [9] by employing machine learning algorithms. The objective is to develop a classification model for the thermogravimetric profile of both pure cellulose acetate membranes and their hybrid counterparts, which exhibit limited differentiation in thermal behavior as a result of variations in the final composition of the material. This preliminary investigation has significant potential for application in the synthesis of cellulose acetate-derived materials, as the ability of machine learning algorithms to classify and identify subtle differences in the thermogravimetric profile of these materials is compared. Random Forest, Extra-Trees Classifier, Decision Tree, and K-Nearest Neighbor are the algorithms compared in this study, owing to their high capacity for supervised classification.

2 Materials and methods

2.1 Experimental data

The experimental data used in this study were obtained from the synthesis and characterization study by Ferreira [10], which focused on the production of new polymeric structures with a fixed composition of 95% (w/w) cellulose acetate and 3.5% (w/w) available for chemical modification with organometallic precursors, namely tetraethyl orthosilicate (TEOS) and 3-aminopropyltriethoxysilane (APTES). The titanium isopropoxide precursor (TiPOT) was fixed in all hybrid membranes, representing 1.5% (w/w) of the final composition. Table 1 presents a summary of the chemical composition explored by Ferreira [10], highlighting the percentage variation of the organometallics TEOS and APTES, which were incorporated into the membrane matrix to produce the silica portion. Details of the synthesis and characterization protocols can be found in previous works [7, 10]. The membranes were thermally characterized using thermogravimetric analysis in the temperature range of 30°C to 600°C and a heating rate (β) of 5°C/min. The analysis was performed with a continuous flow of 20 mL/min of nitrogen and an approximate weight of 6 mg. This technique was carried out in the Biomaterials Laboratory, IPRJ/UERJ, using the Simultaneous Thermal Analyzer STA 6000, Perkin Elmer. Weight loss-temperature points were generated for each composition.

Table 1: Composition of the evaluated cellulose acetate hybrid membranes

Composition	AC – PURO	B 100/0/30	B 75/25/30	B 50/50/30
TEOS (%)	0	100	75	50
APTES (%)	0	0	25	50
TiPOT (%)	0	30	30	30

2.2 Organization of experimental data

To evaluate the effectiveness of machine learning algorithms in identifying subtle differences in the thermogravimetric profile and classifying them, a total of 5,292 data points were obtained from thermal analysis by recording the scanning temperature ratio and weight loss for pure cellulose acetate membranes and the new hybrid membranes, as shown in Table 1. The experimental data were sorted and grouped based on the temperature range to which each membrane was subjected: 30 to 600 ºC, 30 to 450 °C, and 250 to 450 °C.

2.3 Development and comparison of classification models

The Python programming language was used to implement the machine learning algorithms and assess the experimental data points for the temperature-loss weight relationship for each membrane. For the data analysis, the following supervised machine learning algorithms were employed:

- Decision Tree, Random Forest, and Extra-Trees Classifier are based on decision tree. The primary differences are the number of trees generated and the level of randomness in the creation of nodes, which are parameters used to split and classify objects. In Decision Tree, only one tree is produced, in Random Forest, several trees (with some degree of randomness in the nodes) are created, and in Extra-Trees Classifier, several trees (with a high degree of randomness in the nodes) are generated [11][12][13].
- K-Nearest Neighbors (KNN): This model organizes points or objects in a specific space and classifies them based on their proximity. [14].

The classification models were evaluated based on the determination of accuracy, which is defined by Eq. (1). The accuracy of the model represents the ratio of the sum of all temperature-weight points classified correctly (TWPC) to the sum of all temperature-weight points (both correctly (TWPC) and incorrectly (TWPI) classified). Each of the six hybrid membranes studied has its respective TWPC and TWPI. As this is a multiclass accuracy, the confusion matrix was used to observe possible influences of specific classes on the final accuracy of the model.

$$accuracy = \frac{\sum_{i=1}^{n} TWPC}{\sum_{i=1}^{n} TWPC + \sum_{i=1}^{n} TWPI} \tag{1}$$

The training and performance testing of the models followed the steps outlined below:

(i) Treatment and organization of the experimental data;
(ii) Random division of temperature-weight points into training points (70%) and test points (30%);
(iii) Training the models using the training temperature-weight points group;
(iv) Classification of the points reserved for testing their respective membranes;
(v) Evaluation of the models by calculating accuracy, based on the prediction of the classification and the real classification of the thermogravimetric profile;
(vi) Calculation of the confusion matrix to analyze multiclass classifications.

3 Results and discussion

To identify the unique thermodegradation profiles for each evaluated composition of cellulose acetate hybrid membranes, machine learning approaches were employed. The dataset of temperature and weight loss values provided by [9] for each membrane was considered intrinsic characteristics of the respective membranes. This dataset underwent various treatments and was subsequently organized into three temperature range groups for evaluation in machine learning algorithms. This section presents the results of the implementation of the classification models and their respective validation.

The thermogravimetric profile of pure cellulose acetate and hybrid membranes displays a similar and well-known profile from the literature [9], as depicted in Fig. 1. The thermogravimetry (TG) curve for each studied membrane shows three characteristic stages of weight loss, with the last one being the most intense, as it is closely related to the final degradation of the polymer chain. In the first stage, below 100 °C, the loss is related to residual water or volatiles from the material production process. In the second stage, at approximately 125 to 225 °C, there is a loss of functional groups and breakage of important bonds for the polymer structure. This temperature range is the primary degradation stage of the cellulose structure. In the range of 290 °C to 360 °C, there is an intense loss of compound weight, referred to as the third stage, where the sample is in an advanced degradation state, and most of the weight loss is attributed to charred fragments of the chain main.

Initially, the experimental data was evaluated without treatment, i.e., all points provided by the thermal analysis equipment were used (raw data). The points were randomly divided into test data (30%) and training data (70%). The information from this stage of organization of the experimental data is presented in Fig. 2, where the points reserved for training and testing are displayed in their respective temperature ranges of analysis of the cellulose acetate samples.

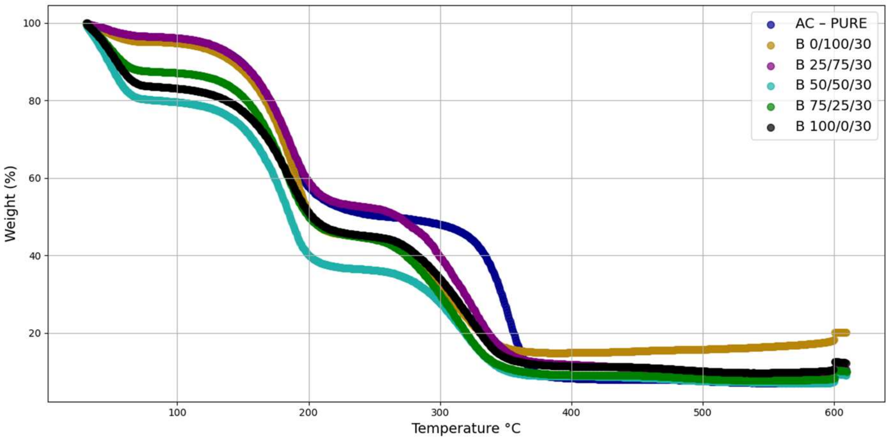


Figure 1: TG curves for the pure cellulose acetate membrane and for the hybrid cellulose acetate membranes.

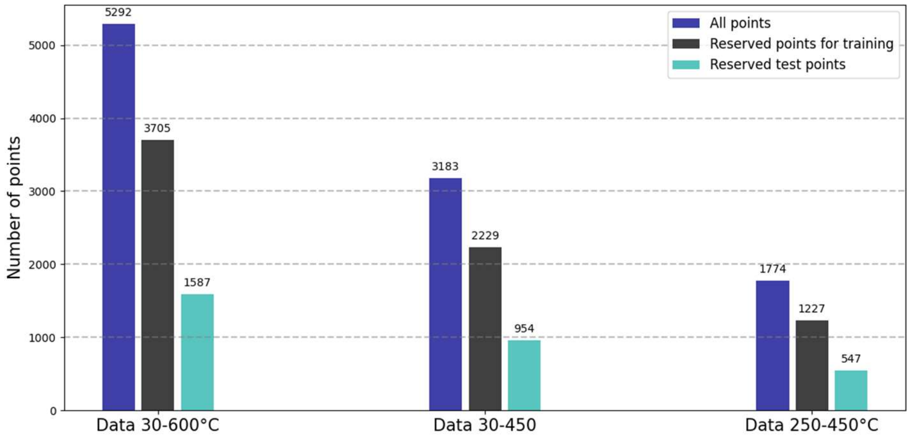


Figure 2: Organization of experimental data according to organized temperature ranges and number of points reserved for training and testing.

In order to improve data analysis, different treatments were applied to the experimental data. For the first treatment, all experimental points in the range of 30 to 600 °C were used for all membranes. However, to facilitate the observation of the most relevant points, an algorithm was used to capture temperature/weight points with a reduction of at least 1% of the total weight for each sample analyzed. After observing the thermogravimetric profile of each membrane, the dataset related to the temperature analysis range of 30 to 450 °C was selected. The resulting data was then divided into training and testing data sets, as shown in Fig. 2.

The third data treatment was performed on the most important degradation range for cellulose acetate membranes, which occurs between 250 to 450 °C. In this range, the third stage of degradation occurs. The selected data points for observation of temperature/weight were distributed to facilitate the visualization of important points. Due to the limited temperature range, a new algorithm was used to select points where the weight variation is above 0.5%, as demonstrated in Fig. 4. After observing the thermogravimetric profile of each membrane, the dataset related to the temperature analysis range of 250 to 450 °C was selected and divided into training and testing data sets, as shown in Fig. 2.

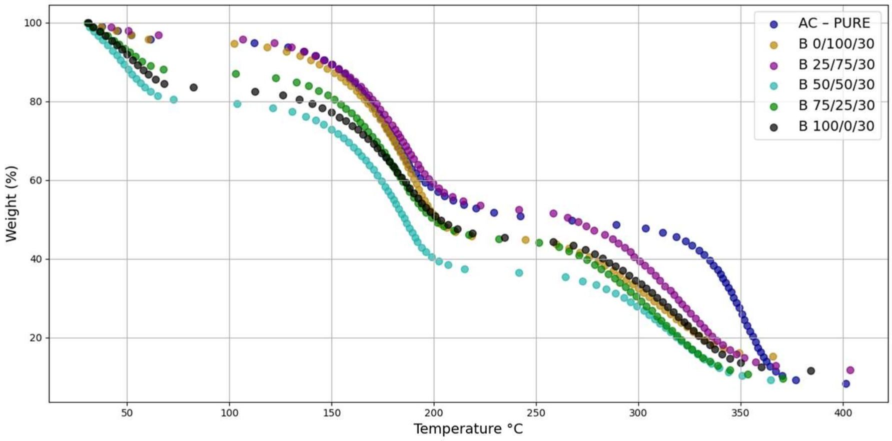


Figure 3: TG curves for the pure cellulose acetate membrane and for the hybrid cellulose acetate membranes, only with points with weight loss equal to or greater than 1%, in the temperature range from 30 to 450°C.

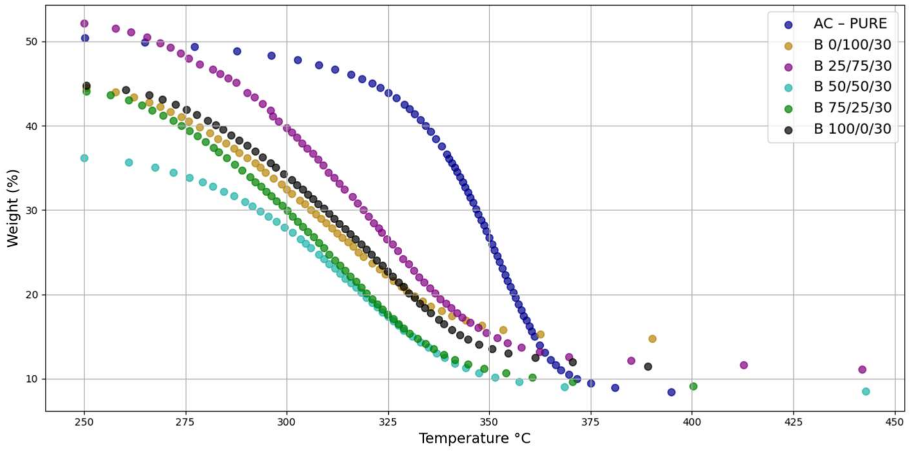


Figure 4: TG curves for pure cellulose acetate membrane and for hybrid cellulose acetate membranes, only with points with weight loss equal to or greater than 0.5% in the temperature range 250 to 450 °C.

3.1 Application and validation of machine learning algorithms

The machine learning algorithms selected for comparison in this study, namely Random Forest, Extra-Trees Classifier, Decision Tree, and K-Nearest Neighbor, were applied to the three treated and organized datasets. For better visualization, the three groups are presented in Figure 5, which highlights the evaluated temperature range.

All temperature-weight points obtained from the equipment for each membrane were considered as intrinsic features. The points reserved for training were used to adjust the classification model, where the algorithm was employed to identify patterns related to each thermogravimetric profile of a specific chemical composition.

With the implemented models, the set of points reserved for testing was used to determine the accuracy of each algorithm in classifying the experimental points (temperature-weight data) from the thermogravimetric analysis of cellulose acetate membranes. The accuracy of each algorithm was measured by the number of temperature-weight points that the model correctly related to a membrane composition. Details on prediction errors and correct predictions are presented in the confusion matrix shown in Figure 6. The Extra-Trees Classifier model is highlighted due to its superior accuracy in all data groups. The matrix reveals the greatest difficulty in classifying the B 50/50/30 and B 75/25/30 membranes, which can be better understood by evaluating the degradation profile of these membranes in relation to temperature, where a great similarity in thermal behavior is observed for these two samples. However, even with a lower classification accuracy, the algorithm correctly classified approximately 80% of the B 50/50/30 and B 75/25/30 membranes.

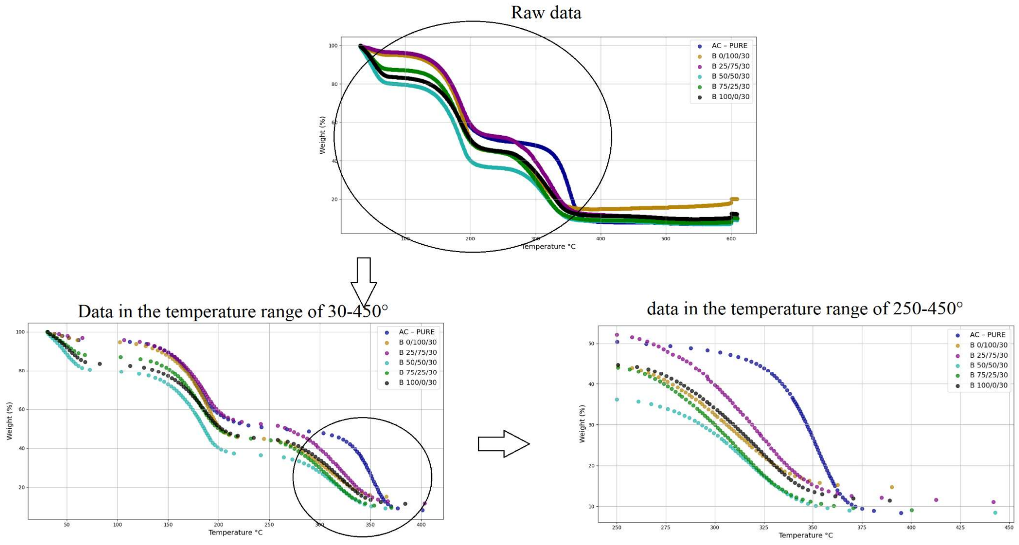


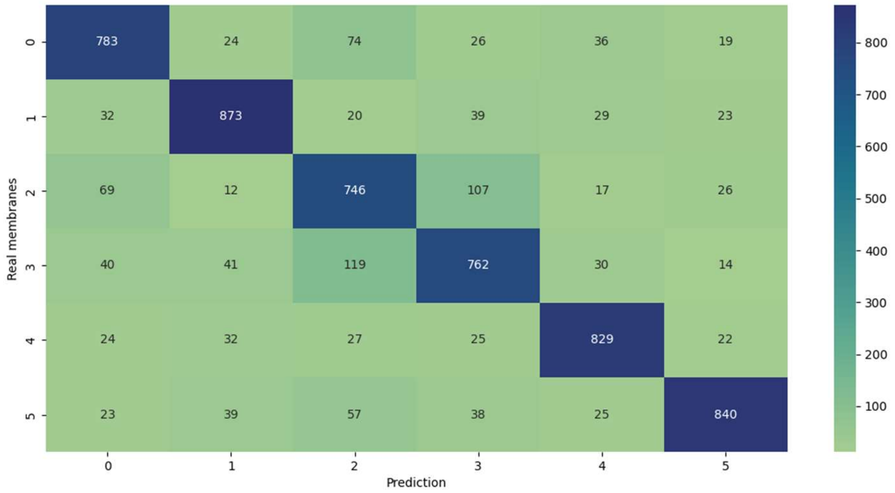
Figure 5: Treatment and organization of experimental data into three groups, according to the temperature range to which the samples were exposed.

The percentage of prediction successes and errors were also determined to calculate the accuracy of the algorithms in classifying the thermal degradation profiles of the membranes. Additionally, the impact of data treatment on the accuracy of the algorithms was evaluated. The results are presented in Figure 7.

The data points organized in the 250 to 450 °C range obtained the highest accuracy in most tests, making it the most favorable range for model implementation. The Extra-Trees Classifier algorithm achieved the highest accuracy in this range, with a precision of 90% after model training. The Random Forest algorithm obtained the second-best result, with 84% accuracy. The Decision Tree algorithm achieved 84% accuracy, which is its second-best result.

When considering the raw data contained between 30 and 600 °C, it was observed that this range is the second most efficient for algorithm application. The Extra-Trees Classifier model was the most accurate, with 87% accuracy, followed by Random Forest, with 83% accuracy. The Decision Tree model obtained its best result in this temperature range, with 81% accuracy.

Finally, the least efficient range for analyzing and classifying the thermal degradation behavior of the membranes was found to be the range containing points between 30 to 450 °C. Despite this, the Extra-Trees Classifier algorithm still delivered the best model, with 84% accuracy, followed by Random Forest (78%) and Decision Tree (75%).



Real membranes \ Prediction	0	1	2	3	4	5
0	783	24	74	26	36	19
1	32	873	20	39	29	23
2	69	12	746	107	17	26
3	40	41	119	762	30	14
4	24	32	27	25	829	22
5	23	39	57	38	25	840

Figure 6: Confusion matrix showing predictions of the Extra-Trees Classifier model when fed with data treated between 250 and 450 °C. AC – PURE (0), B 0/100/30 (1), B 25/75/30 (2), B 50/50/30 (3), B 75/25/30 (4), B 100/0/30 (5).

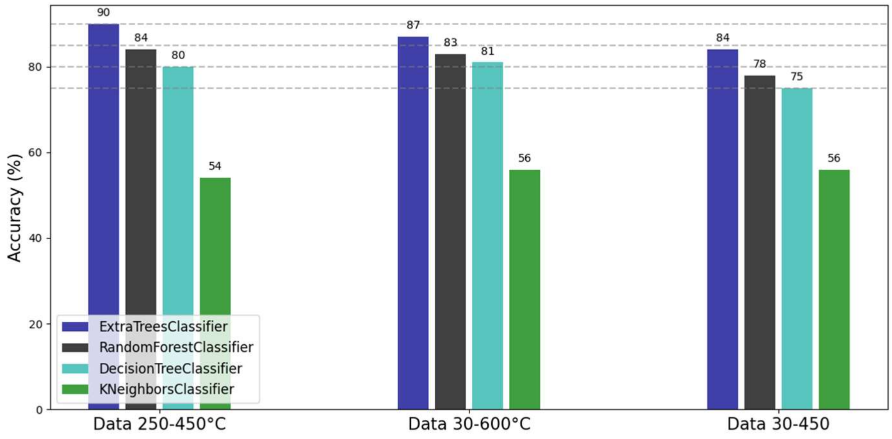


Figure 7: Accuracy of machine learning models in classifying temperature/weight points in specific temperature ranges.

Comparison of accuracy shows that the Extra-Trees Classifier algorithm was the most efficient in learning patterns and classifying profiles according to the evaluated composition. This model achieved an accuracy of 90% when fed with data in the range of 250 to 450 °C. The fact that this treatment achieved the best accuracy may be related to the associated thermal degradation stage, the third, considered to be the stage with the greatest impact on weight loss for cellulose acetate-derived compounds. The Random Forest algorithm was the second-best for TG data analysis of the membranes, achieving its best accuracy when fed with data from 250 to 450 °C. The Decision Tree algorithm was the third-best, and unlike the Extra-Trees Classifier and Random Forest algorithms, achieved higher accuracy when fed with raw data (temperature range of 30 to 600 °C). The KNN model, which is capable of classifying objects according to their location in a space, and which would initially seem interesting for TG points on a Cartesian plane, did not obtain good accuracy in the classification of hybrid membranes. This is probably due to different membranes having very similar temperature-weight points, which can cause confusion in the KNN clustering classification.

Observing the influence of chemical composition on the thermal decomposition profile of synthesized membranes and especially verifying the effect of the insertion of organometallic compounds in an efficient and automated way can optimize the synthesis work for new compositions, facilitating the verification of each batch according to the desired theoretical composition. This preliminary study shows a promising approach for the application of machine learning methods in the process of classifying cellulose acetate membranes, as well as their hybrid derivatives, which present an almost indistinguishable thermogravimetric profile.

4 Conclusions

This study applies machine learning approaches to an experimental dataset that investigates the compositional variation of hybrid cellulose acetate membranes resulting from the incorporation of organometallic compounds, with a focus on enhancing their physical and chemical properties. By utilizing different algorithms, it becomes possible to assess and identify the thermogravimetric profile for similar thermogravimetry curves, but with different compositions. This dataset knowledge enables, particularly for decision tree-based algorithms, a classification model for this thermal profile in accordance with the intrinsic thermal decomposition curve of the cellulose acetate membranes, whether pure or hybrid. The decision tree models prove to be highly efficient. With the Decision Tree, Random Forest, and Extra-Trees Classifier algorithms, the models could accurately classify 84%, 87%, and 90% of the test points, respectively. Among the explored dataset, the Extra-Trees Classifier model is remarkable, demonstrating a promising path for evaluating and classifying future thermal analyses, enabling the identification of compositional variations in a faster and more efficient way.

Acknowledgements

This work was carried out with the support of the Coordination for the Improvement of Higher Education Personnel (CAPES) Brazil, funding code 001. The authors also thank the financial support of the Carlos Chagas Filho Foundation for Research Support of the State of Rio de Janeiro (FAPERJ) and the National Council for Scientific and Technological Development (CNPq).

References

[1] I. H. Sarker, “Machine learning: Algorithms, real-world applications and research directions,” *SN Computer Science*, vol. 2, no. 3, paper no. 160, 2021. Available at: https://doi.org/10.1007/s42979-021-00592-x

[2] D. Morgan and R. Jacobs, “Opportunities and challenges for machine learning in materials science,” *Annual Review of Materials Research*, vol. 50, no. 1, pp. 71-103, 2020. Available at: https://doi.org/10.1146/annurev-matsci-070218-010015

[3] C. L. Ritt, T. Stassin, D. M. Davenport, R. M. DuChanois, I. Nulens, Z. Yang, A. Ben-Zvi, N. Segev-Mark, M. Elimelech, C. Y. Tang, G. Z. Ramon, I. F. J. Vankelecom, R. Verbeke, “The open membrane database: Synthesis–structure–performance relationships of reverse osmosis membranes,” *Journal of Membrane Science*, vol. 641, paper no. 119927, 2022. Available at: https://doi.org/10.1016/j.memsci.2021.119927

[4] Y.-J. Hu, G. Zhao, M. Zhang, B. Bin, T. Del Rose, Q. Zhao, Q. Zu, Y. Chen, X. Sun, M. de Jong, and Q. Liang, "Predicting densities and elastic moduli of SiO_2-based glasses by machine learning," *Npj Computational Materials*, vol. 6, paper no. 25, 2020. Available at: https://doi.org/10.1038/s41524-020-0291-z

[5] H. Khakurel, M. F. N. Taufique, A. Roy, G. Balasubramanian, G. Ouyang, J. Cui, D. D. Johnson, and R. Devanathan, "Machine learning assisted prediction of the Young's modulus of compositionally complex alloys," *Scientific Reports*, vol. 11, paper no. 17149, 2021. Available at: https://doi.org/10.1038/s41598-021-96507-0

[6] K. Low, R. Kobayashi, and E. I. Izgorodina, "The effect of descriptor choice in machine learning models for ionic liquid melting point prediction," *The Journal of Chemical Physics*, vol. 153, no. 10, paper no. 104101, 2020. Available at: https://doi.org/10.1063/5.0016289

[7] M. C. Andrade, J. C. Pereira, N. de Almeida, P. Marques, M. Faria, and M. C. Gonçalves, "Improving hydraulic permeability, mechanical properties, and chemical functionality of cellulose acetate-based membranes by co-polymerization with tetraethyl orthosilicate and 3-(aminopropyl) triethoxysilane," *Carbohydrate Polymers*, vol. 261, paper no. 117813, 2021. Available at: https://doi.org/10.1016/j.carbpol.2021.117813

[8] V. Vatanpour, M. E. Pasaoglu, H. Barzegar, O. O. Teber, R. Kaya, M. Bastug, A. Khataee, I. Koyuncu, "Cellulose acetate in fabrication of polymeric membranes: A review," *Chemosphere*, vol. 295, paper no. 133914, 2022. Available at: https://doi.org/10.1016/j.chemosphere.2022.133914

[9] G. M. Ferreira, "Production and characterization of hybrid cellulose acetate membranes," Master's thesis, Materials Science and Technology Postgraduate Program, State University of Rio de Janeiro, Nova Friburgo, Brazil, 2022. Available at: http://www.bdtd.uerj.br/handle/1/17895

[10] G. M. Ferreira, D. H. da Silva, K. C. Da Silveira, M. C. Gonçalves, and M. C. Andrade, "Evaluation of Thermal Degradation Kinetics of Hybrid Cellulose Acetate Membranes using Isoconversional Methods," *VETOR - Revista de Ciências Exatas e Engenharias*, vol. 32, no. 1, pp. 52-61, 2022. Available at: https://doi.org/10.14295/vetor.v32i1.13766

[11] N. M. Abdulkareem and A. M. Abdulazeez, "Machine learning classification based on Radom Forest Algorithm: A review," *International Journal of Science and Business*, vol. 5, no. 2, pp. 128-142, 2021. Available at: https://ijsab.com/wp-content/uploads/676.pdf

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011. Available at: https://scikit-learn.org/stable/

[13] B. T. Jijo, and A. M. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 1, pp. 20-28, 2021. Available at: https://www.jastt.org/index.php/jasttpath/article/download/65/24

[14] H. A. A. Alfeilat, A. B. A. Hassanat, O. Lasassmeh, A. S. Tarawneh, M. B. Alhasanat, H. S. E. Salman, and V. B. S. Prasath, "Effects of distance measure choice on k-nearest neighbor classifier performance: A review," *Big Data*, vol. 7, no. 4, pp. 221-248, 2019. Available at: https://doi.org/10.1089/big.2018.0175