In-Between Randomization Assisted Machine Learning Performance Analysis for Naturally-Sensitized Solar Cells

Hisham A. Maddah*
Department of Chemical Engineering,
Faculty of Engineering—Rabigh
King Abdulaziz University, Jeddah
21589, Saudi Arabia

*Corresponding author.
hmaddah@kau.edu.sa

***Abstract*— The utilization of free, renewable, and available solar energy became a focus research area in recent years. Sustainable natural photosensitizers in dye-sensitized solar cells (DSSCs) are among the hot researched topics in the scientific community. Herein, various naturally-sensitized-photoanode-based DSSCs were studied *via* statistical and machine learning analysis to investigate the possibility to achieve relatively high PCEs in naturally-sensitized DSSCs. Studied photosensitizer (dye) characteristics included chemical structure and bandgap which were correlated to the literature reported PCEs. Input parameters used in models classification training were: the number of π-bonds (PI), the number of anchoring groups (X), HOMO(H)-LUMO(L), and bandgap energy (BG), with only 2 responses regarding the statistical possibility to achieve high PCEs (Yes/No). Both training/testing (80/20)% datasets were carefully chosen to identify the dye controlling parameters responsible for increasing PCEs. The built trained classification models (decision trees) were tested and showed high prediction accuracy. The idea here is to check whether a certain dye and its correlated PCE would achieve below or above the average PCE. Thus, this allowed us to classify the problem according to the selected parameters so that the dyes can be correlated to their BGs and the other parameters. This work shows the potential of applying statistical analysis to natural sensitizers for enhanced charge injection (current density) for renewable, cost-effective, and sustainable energy production.**

Keywords— performance, statistical analysis, dye solar cells, machine learning, natural sensitizers.

I. INTRODUCTION

Natural dyes extracted from different natural sources (e.g. anthocyanin, carotenoid, etc.) have been previously proposed to be used as sensitizers in dye-sensitized solar cells (DSSCs) due to their low cost and environmental friendliness [1]. However, the power conversion efficiency (PCE) from naturally-sensitized DSSCs is typically in the average range (<0.05–1.7%) [2] requiring a thorough understanding of the role of pigment's molecular structure, electronic properties, anchoring groups, and conjugated double bonds or free π-electrons for improved PCE from enhanced carriers transport and decreased recombination [3]–[5]. Photosensitization is the basis for designing efficient DSSCs capable of initiating electron injection and charge transfer from dye molecules to the semiconductor [6]–[8].

A photosensitizer is considered efficient when it fulfills these requirements [9]: (i) intense visible-light absorption, (ii) strong chemisorption onto the semiconductor surface, (iii) fast electron injection into the semiconductor CB, and (iv) involve several =O or –OH groups to anchor dye molecules onto semiconductor surface. There are different kinds of natural and environmental-friendly photosensitizers [1], which can be extracted from light-harvesting complexes of anthocyanin, carotenoid, and chlorophyll biomolecules [10] emerged as an attempt to substitute the expensive and toxic [11] metal-based ruthenium polypyridyl dyes [12].

Since the beginning of technology, machines (computers) in many trials have been used to learn specific patterns for data classification and decision making [13]. Classification algorithm distributes variously mixed datasets into categories by constructing a model *via* supervised learning the relation between input attributes and an output-dependent parameter [14]. A common classification algorithm is known as "Decision trees" is well-known for its ability to categorize datasets. The method consists of a tree with internal nodes used as categories allowing building classification models from observations of attributes or predictors (terminal nodes) to reach conclusions based on categorized responses [15]. Decision trees work great with redundant attributes, provide good results in presence of data noise, classify small datasets easily, give high accuracy with minimum nodes or features. [14]. PCE prediction of DSSCs was earlier studied *via* multi-learner ensembles based on clustering and modeling approaches for achieving high accuracy and generalization. The L-SVM-KNN-WMA based achieved high accuracy >91% for PCE prediction [16]. Im et al. applied gradient-boosted regression trees (GBRT) to predict structural heat of formation and bandgap for lead-free perovskites [17]. Prediction of dye adsorption on titania and absorption capabilities was previously studied *via* various classification methods which accurately indicated spectral shifts in 70–80% of inspected photosensitizers [18].

In this work, we statistically investigate the possibility to achieve high PCEs in naturally-sensitized-photoanode-based DSSCs using decision trees machine learning [16], [19], [20] of dye structural, electronic, and molecular properties. An earlier introduced concept in our work [20], called "in-between randomization", was then applied for an expansion of datasets information from 27 natural sensitizers. Models building algorithms were carried out *via* 4 (dye) predictors including the number of π-bonds (PI), the number of anchoring groups (X), HOMO(H)-LUMO(L), and bandgap energy (BG), with 2 responses for PCEs >1.82% (Yes/No).

978-1-6654-7826-7/21/$31.00 ©2021 IEEE

A "parameters importance" analysis was conducted to find the prime factors and controlling variables that would enhance dye absorption abilities for maximum performance. The idea here is to check whether a certain dye and its correlated PCE (already known from the literature) would achieve below or above the average PCE. Thus, this allowed us to classify the problem according to the selected parameters to know the dyes that would produce high PCEs and those that would produce low PCE so that they can be correlated to their BGs and the other parameters indicated in the study methodology.

II. Methods and Equations

We collected raw information regarding the performance of various redox-liquid and TiO_2-based naturally-sensitized DSSCs from more than 30 recently published articles (2015–2020) [2], [21]–[24]. Collected PCEs were then correlated to the dye type, structural, electronic, and molecular properties. The chemical structures of the studied dyes were then carefully gathered and manually evaluated to check for the existing number of double conjugated π-bonds and the existing number of anchoring or functional groups. Then, we looked for the approximated values of bandgap energies of every single and different studied dye, where we have taken the average value of the reported theoretical bandgaps of pigments from the literature.

The constructed original datasets which contain 27 different sensitizers were mainly selected from dye classes such as carotenoids, protein complexes, flavonoids, cyanins, chlorophyll, and chromatophores. The 1.82% was the determined averaged performance of the naturally-sensitized DSSCs according to the selected dye types, based on TiO_2 photoanode and iodide-triiodide liquid redox. An earlier introduced concept [20], called "in-between randomization", was applied for an expansion of datasets by 5-fold.

The "in-between randomization" allowed us to expand the originally constructed datasets to 135 numbers whereas that both HOMO and LUMO levels were also expanded with the taken errors since (BG=HOMO–LUMO), as in **Fig. 1** (along with factors affecting performance). The original data points were plotted after raw sets were expanded *via* utilizing BG values. Notice that we assumed that there were only 4 dye-associated factors that would impact the performance of DSSCs. The expanded datasets were divided into two sets (80% training and 20% testing) to accurately establish classification models and be able to test their validity and prediction accuracy.

The interpretations of models errors *via* classification tree graphs were then considered to study the parameter's importance and select the best models among the different established input-parameters trained models. Prime factors or primary and secondary controlling variables in each of the best models were obtained from tree pruning based on the root node and internal nodes from tree branching. This would allow measuring the degree of impact of studied predictors on PCEs and dye absorption ability for visible-light energy and capability to separate generated electron-hole pairs. We then estimated the order of magnitude of parameters importance, while correlating the importance of existing anchoring groups to both PI and BG and respective dye impact on the solar cell PCEs. The equation of identified statistical error from the coefficient of determination (R^2) is shown in **Eq. (1)**, knowing that the observed value is symbolized as $x_{o,i}$ and/or x_o; $x_{p,i}$ and/or x_p refers to the values predicted by the model; predicted value $\bar{x}_o$ is the experimentally obtained or observed values from averaging; $\bar{x}_p$ is the theoretically estimated or predicted values from averaging; and n refers to the datasets size or the number of experimental observations.

$$\mathbf{R^2} = \frac{\left[\sum_{i=1}^{n}(x_{o,i}-\bar{x}_o)(x_{p,i}-\bar{x}_p)\right]^2}{\sum_{i=1}^{n}(x_{o,i}-\bar{x}_o)^2 \times \sum_{i=1}^{n}(x_{p,i}-\bar{x}_p)^2} \quad (1)$$

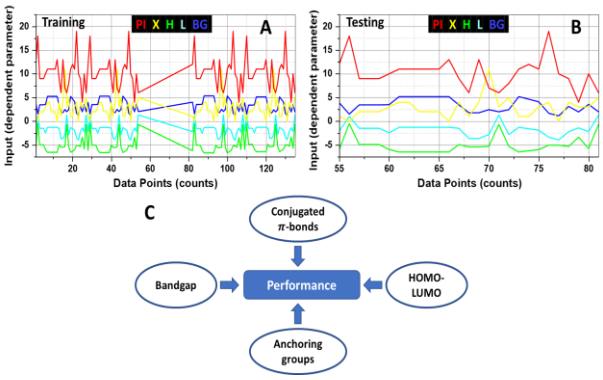

Fig. 1. Raw datasets from literature used in the supervised machine learning: (A) Training, (B) Testing; (C) Factors affecting performance.

The output of the classifier is set to be correlated to the average PCE of 1.82% that the average efficiency found from many studied naturally-sensitized DSSCs with the same studied dyes. The idea here is to check whether a certain dye and its correlated PCE (already known from the literature) would achieve below or above the average PCE. Thus, this allowed us to classify the problem according to the selected parameters to know the dyes that would produce high PCEs and those that would produce low PCE so that they can be correlated to their BGs and the other parameters indicated in the study methodology. The outcome from the classification analysis would help scientists to conduct their studies according to the found parameters which have a high impact on the PCE such as BG and H of the dye molecules. The author suggests that dye blending or hybrid natural-synthetic dyes should be synthesized based on each dye BG and H to better promote the PCE in DSSCs.

III. Study Framework

Training steps were carried out using four different input-parameters models: PIX-input, BG-input, HLBG-input, and PIXBG-input (i.e. mix/match of selected independent parameters). The selection of various input parameters is important to define controlling factors that would chiefly result in changing PCEs based on attributes analysis.

The classification output (response) was linked to the normalized scores which were in the range [0.004 – 4.54], and PCEs were identified to be relatively high or low if score > 1 (Yes) and score < 1 (No), respectively. In other words, the "COUNT IFS" statement was applied in "EXCEL" to translate literature PCEs numbers to (Yes=1) and (No=0) whereas that average PCE = 1.82 is the boundary limits [i.e.

If PCEs >1.82%, return 1=Yes, else 0=No). By doing a numeric-to-character conversion decision analysis, we correlated studied PCEs to the various naturally-sensitized photoanodes and their pigments. Various trained classifiers were then tested statistically to check for their accuracies, which in turn showed that only decision trees and SVMs had high prediction accuracies. . The adopted study framework is shown in **Fig. 2**. The selection of inputs as independent parameters resulted in the possibility of establishing a minimum of four unique models: PIX-input models, BG-input models, HLBG-input models, and PIXBG-input models (i.e. mix/match the studied independent parameters).

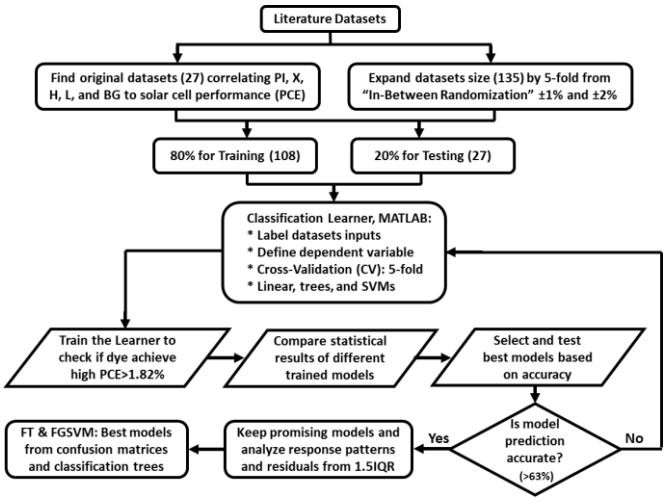

Fig. 2. Study framework for data collection, training, testing, followed by the analysis of the most accurate machine learning predictive models.

IV. Results and Discussion

The classification accuracy results of the different input models found from MATLAB analysis are shown in **Table I**, which have approximate 81%, 85%, and 90% accuracies for PIX-input, BG-input, and [HLBG, and PIXBG]-input models, respectively. Also, the built models' accuracies were determined by taking the average accuracy obtained from the various decision trees trained models shown in **Table I** (which shows predictions from classification for each of the studied input models).

TABLE I. The prediction accuracies of different input models for evaluation of natural dyes predictors impact on PCEs.

Predictors / Model	Decision Trees		
	FT	***MT***	***CT***
PIX	85.2%	85.2%	75%
BG	85.2%	85.2%	83.3%
HLBG	91.7%	91.7%	86.1%
PIXBG	90.7%	90.7%	79.6%

A. *Controlling Parameters and Predictor Importance*

Controlling parameters or included inputs used in the various built decision tree models have been evaluated *via* predictor importance analysis. Predictor importance is nothing but weight fraction corresponding to each parameter that would show the impact of a parameter on categorizing DSSCs according to the selected dyes and based on the system PCE. For instance, PIX-input analysis showed that the PI (free dye electrons) is almost as twice important as the X (anchoring groups) in indicating whether a dye type would effectively increase PCEs or not based on FT and MT that were found to be much more accurate than that CT as shown in **Fig. 3(A)**. The BG-input model only has BG as an independent variable which was found to be of high importance in defining dye capabilities, **Fig. 3(B)**. The HLBG-input model's analysis confirmed that BG is among the top controlling parameters that is ~ 3-fold more important than H (HOMO) energy level of the dye, **Fig. 3(C)**. Yet, HOMO levels must be taken into consideration since this is the lowest dye molecular energy level from where electrons should be excited to reach L (LUMO) and overcome the BG energies to produce excitons (free e-h pairs). Moreover, the PIXBG-input model's analysis determined that BG/X importance ratio was about 12 as shown in **Fig. 3(D),** which concludes that the order of magnitude of parameters importance as BG (1) > H (0.32) > PI (0.08) > X (0.04) that should be adopted when analyzing natural dye abilities for charge generation/injection to achieve high PCEs.

B. *Decision Trees Classification*

According to the classification tree graphs plotted in **Fig. 4** from FT, MT, and CT trained classifiers, it was evident that only BG and H are the controlling factors when it comes to the HLBG model with only two pruning levels. The first controlling parameter or feature (BG) has classified >63% of the datasets from HLBG based on the root node and internal nodes from tree branches and sub-branches as shown in **Fig. 4(A,B)**. Conversely, the HOMO level, which is important for the dye absorption abilities, is not as critical as the overall required energy needed to be expressed in BG. From analyzing the generated trees from PIXBG trained models, BG was also the prime classifier among the three input factors including free electrons and anchoring groups from illustrations in **Fig. 4(C,D)**. Both BG and X were the controlling factors in the case of using PIXBG, which emphasizes that PI is not as important as X in finding dye impact on PCEs in DSSCs. Moreover, The BG was found to control >85% of datasets for FT/MT (PIXBG) acting as a prime parameter.

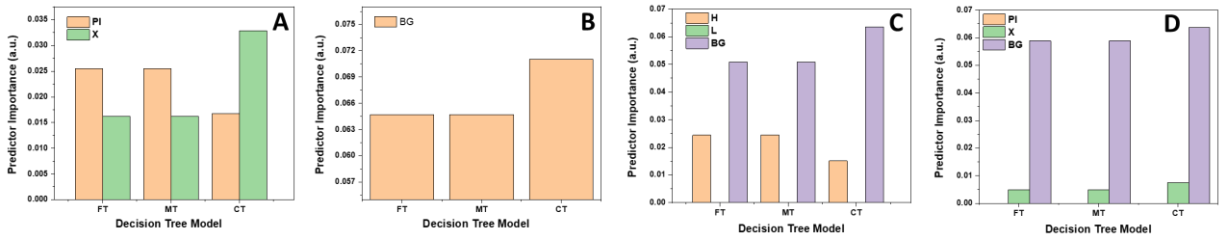

Fig. 3. Predictor importance of independent input variables used in decision tree models based on selection of inputs: (A) PIX-input models, (B) BG-input models, (C) HLBG-input models, (D) PIXBG-input models.

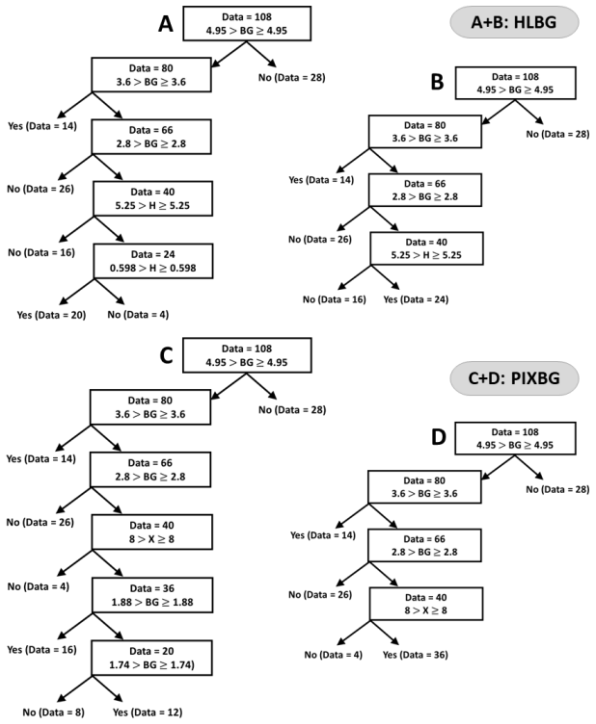

Fig. 4. Classification tree graphs from fine tree (FT), medium tree (MT), and coarse tree (CT) trained classifiers: **(A)** FT and MT, **(B)** CT, **(C)** FT and MT, **(D)** CT. Note number in the parenthesis corresponds to probability to achieve PCE >1.82%.

A parameter that would be able to control the classification of DSSCs for high or low PCE would practically also result in making DSSCs performance highly susceptible to that parameter. For instance, the identified parameters importance BG (1) > H (0.32) > PI (0.08) > X (0.04) can be also thought of as the PCE would be very susceptible to DSSCs operated with dyes that have very high or very low BGs, whereas that the very low BG would increase efficiency and the very high BG would result in low PCE due to the high energy requirements for electron excitation.

V. Conclusion

We developed high-accuracy predictive models to study the impact of dye structural, electronic, and molecular properties on the PCE of DSSCs. Statistical analyses were carried out *via* 4 predictors [the number of dye structure π-bonds (PI), number of dye anchoring groups (X), HOMO(H)-LUMO(L), and bandgap energy (BG)] with 2 responses for the high PCEs (Yes/No). The HLBG-input and PIXBG-input models were found promising with the highest accuracies of 91% using FT/MT. The results confirmed that BG is among the top controlling parameters with the order of magnitude of parameters importance as BG (1) > H (0.32) > PI (0.08) > X (0.04). Both BG and X were controlling factors when applying PIXBG, which emphasizes that PI is not as important as X in impacting PCEs whereas the BG parameter was found to control >85% of the datasets (FT/MT).

Acknowledgment

The author would like to acknowledge the Deanship of Scientific Research (DSR) at King Abdulaziz University (KAU) for their support and motivation to complete this work.

References

[1] N. Sawhney, A. Raghav, and S. Satapathi, "Utilization of Naturally Occurring Dyes as Sensitizers in Dye Sensitized Solar Cells," *IEEE J. Photovoltaics*, 2017.

[2] S. K. Srivastava, P. Piwek, S. R. Ayakar, A. Bonakdarpour, D. P. Wilkinson, and V. G. Yadav, "A Biogenic Photovoltaic Material," *Small*, 2018.

[3] H. Hug, M. Bader, P. Mair, and T. Glatzel, "Biophotovoltaics: Natural pigments in dye-sensitized solar cells," *Appl. Energy*, 2014.

[4] K. Nagai, A., & Takagi, *Conjugated Objects: Developments, Synthesis, and Applications*. Pan Stanford, 2017.

[5] H. Maddah, A. Jhally, V. Berry, and S. Behura, "Highly Efficient Dye-sensitized Solar Cells with Integrated 3D Graphene-based Materials," in *Graphene-based 3D Macrostructures for Clean Energy and Environmental Applications*, Royal Society of Chemistry, 2021, pp. 205–236.

[6] A. Hagfeldt, "Brief overview of dye-sensitized solar cells," in *Ambio*, 2012.

[7] V. Sugathan, E. John, and K. Sudhakar, "Recent improvements in dye sensitized solar cells: A review," *Renewable and Sustainable Energy Reviews*. 2015.

[8] B. Parida, S. Iniyan, and R. Goic, "A review of solar photovoltaic technologies," *Renewable and Sustainable Energy Reviews*. 2011.

[9] M. R. Narayan, "Review: Dye sensitized solar cells based on natural photosensitizers," *Renewable and Sustainable Energy Reviews*. 2012.

[10] H. A. Maddah, V. Berry, and S. K. Behura, "Biomolecular photosensitizers for dye-sensitized solar cells: Recent developments and critical insights," *Renewable and Sustainable Energy Reviews*. 2020.

[11] C. Cari, K. Khairuddin, T. Y. Septiawan, P. M. Suciatmoko, D. Kurniawan, and A. Supriyanto, "The preparation of natural dye for dye-sensitized solar cell (DSSC)," in *AIP Conference Proceedings*, 2018.

[12] S. Mathew *et al.*, "Dye-sensitized solar cells with 13% efficiency achieved through the molecular engineering of porphyrin sensitizers," *Nat. Chem.*, 2014.

[13] M. Somvanshi, P. Chavan, S. Tambade, and S. V. Shinde, "A review of machine learning techniques using decision tree and support vector machine," in *Proceedings - 2nd International Conference on Computing, Communication, Control and Automation, ICCUBEA 2016*, 2017.

[14] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, 2021.

[15] S. Sayad, "Decision Tree - Regression," *Data Science: Predicting the Future, Modeling & Regression*. [Online]. Available: https://www.saedsayad.com/decision_tree_reg.htm.

[16] H. Li *et al.*, "Ensemble learning for overall power conversion efficiency of the all-organic dye-sensitized solar cells," *IEEE Access*, vol. 6, 2018.

[17] H. Im, J., Lee, S., Ko, T. W., Kim, H. W., Hyon, Y., & Chang, "Identifying Pb-free perovskites for solar cells by machine learning," *npj Comput. Mater.*, vol. 5, no. 1, p. 37, 2019.

[18] V. Venkatraman, A. E. Yemene, and J. de Mello, "Prediction of Absorption Spectrum Shifts in Dyes Adsorbed on Titania," *Sci. Rep.*, vol. 9, no. 1, 2019.

[19] H. A. Maddah, V. Berry, and S. K. Behura, "Cuboctahedral stability in Titanium halide perovskites via machine learning," *Comput. Mater. Sci.*, vol. 173, 2020.

[20] H. A. Maddah, M. Bassyouni, M. H. Abdel-Aziz, M. S. Zoromba, and A. F. Al-Hossainy, "Performance estimation of a mini-passive solar still via machine learning," *Renew. Energy*,

2020.

[21] N. Órdenes-Aenishanslins, G. Anziani-Ostuni, M. Vargas-Reyes, J. Alarcón, A. Tello, and J. M. Pérez-Donoso, "Pigments from UV-resistant Antarctic bacteria as photosensitizers in Dye Sensitized Solar Cells," *J. Photochem. Photobiol. B Biol.*, 2016.

[22] J. Chellamuthu, P. Nagaraj, S. G. Chidambaram, A. Sambandam, and A. Muthupandian, "Enhanced photocurrent generation in bacteriorhodopsin based bio-sensitized solar cells using gel electrolyte," *J. Photochem. Photobiol. B Biol.*, 2016.

[23] K. Liu *et al.*, "Spiro[fluorene-9,9′-xanthene]-based hole transporting materials for efficient perovskite solar cells with enhanced stability," *Mater. Chem. Front.*, 2017.

[24] Y. Shao *et al.*, "Stable Graphene-Two-Dimensional Multiphase Perovskite Heterostructure Phototransistors with High Gain," *Nano Lett.*, 2017.