Decision Trees Based Performance Analysis for Influence of Sensitizers Characteristics in Dye-Sensitized Solar Cells

Hisham A. Maddah
Department of Chemical Engineering, Faculty of Engineering—Rabigh Branch, King Abdulaziz University, Jeddah, Saudi Arabia
Email: hmaddah@kau.edu.sa

***Abstract*—The focus of the scientific community has shifted towards renewable and sustainable natural photosensitizers for Dye-Sensitized Solar Cells (DSSCs). Here, we statistically investigate the possibility to achieve relatively high PCEs in naturally-sensitized-photoanode-based DSSCs using decision trees (machine learning). We studied the chemical structure and bandgap of 27 sensitizers, which were then correlated to the literature reported PCEs. Tree training was 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 regarding the statistical possibility to achieve high PCEs (Yes/No). Trained datasets revealed the controlling parameters responsible for increasing PCEs. Testing (future) datasets were chosen to check for built models' accuracy in performance prediction for enhanced charge injection (current density). This work shows the potential of natural sensitizers used in DSSCs for renewable, cost-effective, and sustainable energy production.**

***Index Terms*—power conversion efficiency, natural sensitizers, machine learning, dye-sensitized solar cells**

I. Introduction

Photosensitization is the basis for designing efficient Dye-Sensitized Solar Cells (DSSCs) capable of initiating electron injection and charge transfer from dye molecules to the semiconductor [1]-[3]. There are many different kinds of natural and environmental-friendly photosensitizers [4], which can be extracted from light-harvesting complexes of anthocyanin, carotenoid, and chlorophyll biomolecules [5] emerged as an attempt to substitute the expensive and toxic [6] metal-based ruthenium polypyridyl dyes [7]. Natural dyes extracted from different natural sources (e.g. anthocyanin, carotenoid, flavonoid, etc.) have been previously proposed to be used as sensitizers in DSSCs due to their low cost and environmental friendliness [4]. However, the Power Conversion Efficiency (PCE) from naturally-sensitized DSSCs is typically in the average range (<0.05-1.7%) [8] 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 [9]-[11].

A photosensitizer is considered efficient when it fulfills these requirements [12]: (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. 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 known as "Decision trees" is well-known for its ability to categorize datasets. The method consists of a tree with internal nodes that are nothing but tests and with leaf nodes used as categories. This builds classification models from observations of datasets attributes or predictors (branches as terminal nodes) to reach conclusions based on categorized responses [15]. There are different types of decision tree algorithms including the common ones as Iterative Dichotomies 3 (ID3), the successor of ID3 (C4.5), classification and regression tree (CART), and Conditional Inference Trees (CTREE) [14]. 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].

Odabas *et al.* [16] applied decision trees to analyze the impact of materials selection on the stability of organo-lead halide perovskite solar cells from 404 cells stability profiles. Decision trees deduced rules and guidelines for fabricating long-term stable perovskite solar cells [16]. PCE prediction of DSSCs was earlier studied *via* multi-learner ensembles based on clustering and modeling approaches for achieving high accuracy. The L-SVM-KNN-WMA based achieved high accuracy >91% for PCE prediction [17]. Im *et al.* applied Gradient-Boosted Regression Trees (GBRT) to predict bandgap for lead-free perovskites [18]. Prediction of dye adsorption on titania and absorption capabilities was previously studied *via*

Manuscript received November 19, 2021; revised April 22, 2022.

doi: 10.12720/jait.13.3.271-276

classification methods which accurately indicated spectral shifts in 70–80% of inspected photosensitizers [19].

In this work, we statistically investigate the possibility to achieve high PCEs in naturally-sensitized-photoanode-based DSSCs using decision trees machine learning [17], [20], [21] of dye structural, electronic, and molecular properties. An earlier introduced concept in our work [21], 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). A "parameters importance" analysis was conducted to find the prime factors and controlling variables that would enhance dye abilities to absorb more of the visible-light energy (photons) and separate generated electron-hole pairs for maximum performance.

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) [8], [22]-[25]. 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 [21], called "in-between randomization", was applied for an expansion of datasets by 5-fold. Simply, we took leverage of inevitable errors from reported experimental and theoretical results by considering errors of ±1% and ±2% in PCEs of the cell and their associated dye bandgaps for generating further numbers in the datasets. This 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 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.

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

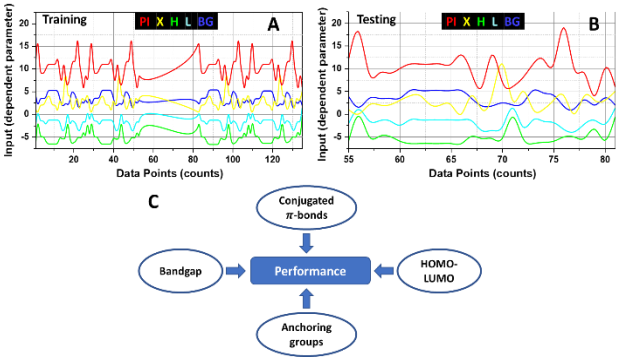

Figure 1. Raw datasets from literature used in the supervised machine learning: (A) Training, (B) Testing; (C) Factors affecting performance.

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).

1. Collecting datasets from various literature sources.
2. Find original datasets to correlate PI, X, H, L, and BG to PCE.
3. Expand datasets size by 5-fold from "In-Between Randomization"
4. Use 80% for Training (108) and the other 20% for Testing (27).
5. Apply 'Classification Learner' in MATLAB defining dependent variables.
6. Train the Learner to check if dye achieve high PCE>1.82% or not.
7. Compare statistical results of from the different trained models.
8. Select and test best models based on accuracy
9. Check if the model prediction accurate or not.
10. Analyze response patterns and residuals from 1.5IQR
11. Further study FT & FGSVM "best models".
12. Analyze best models from confusion matrices & classification.

Figure 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 & Predictor Importance

Controlling parameters or included inputs used in the various built decision tree models have been evaluated *via* predictor importance analysis. 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.

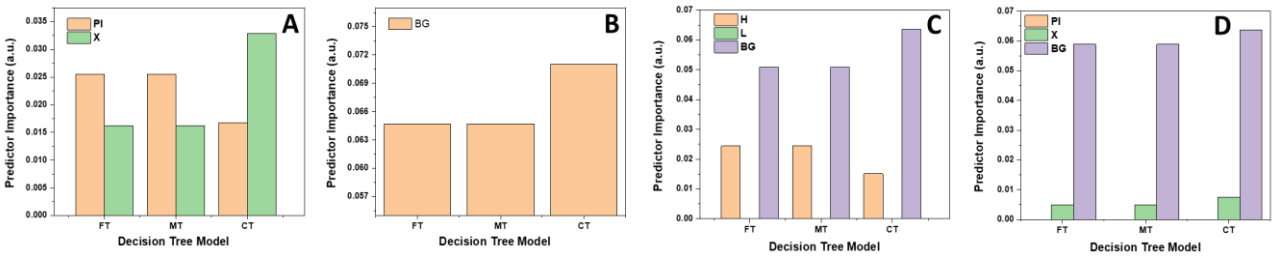

Figure 3. Predictor importance of independent input variables used in decision tree models based on selection of inputs: (A) PIX-input models, (B) BG-input models, (C) HLBG-input models, (D) PIXBG-input models.

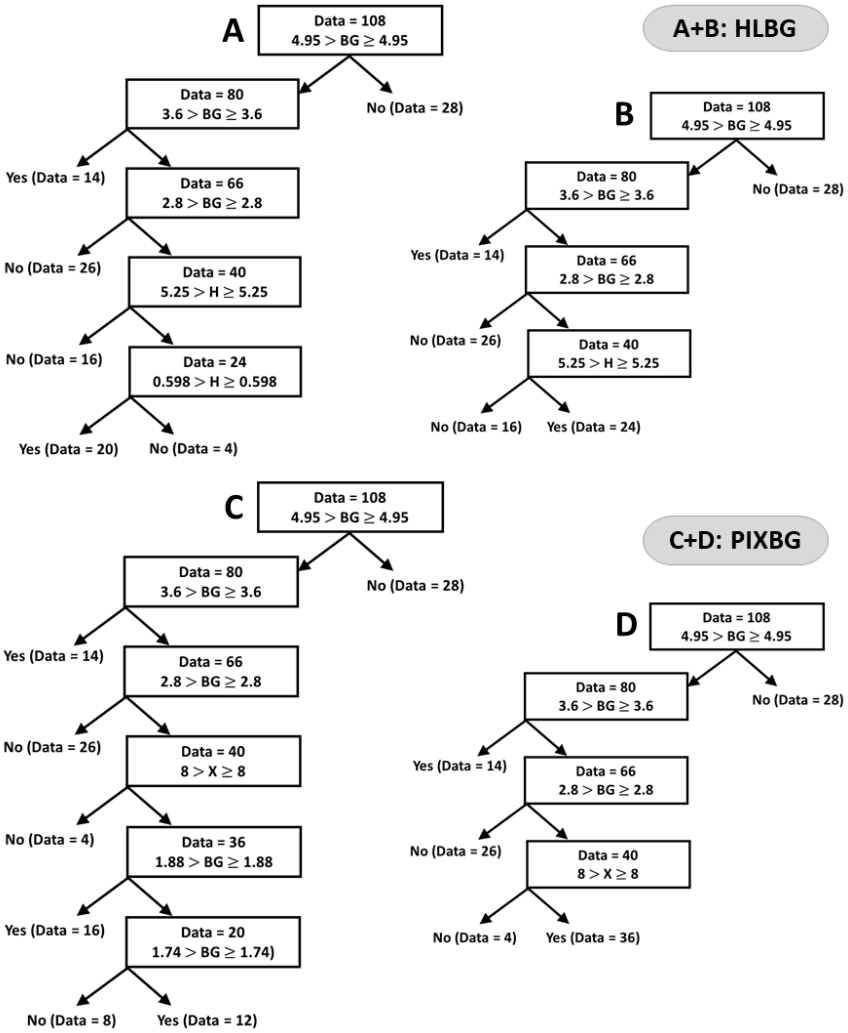

Figure 4. Classification tree graphs from fine tree (FT), medium tree (MT), and coarse tree (CT) trained classifiers: **(A)** FT and MT, **(B)** CT, **(C)** FT and MT, **(D)** CT. Note number in the parenthesis corresponds to probability to achieve PCE >1.82%.

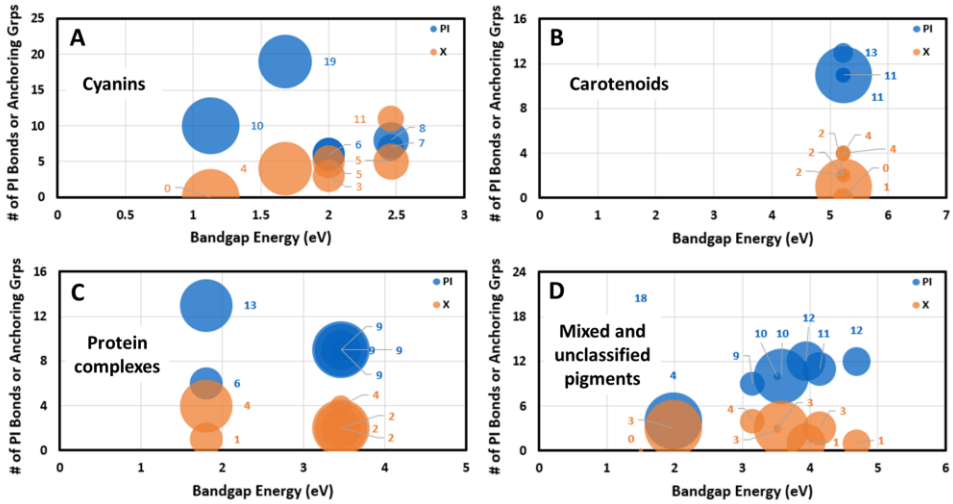

Figure 5. Bubble charts for visualizing changes in PCEs in TiO_2-based/iodide-triiodide-liquid-redox-based naturally-sensitized DSSCs showing the impact of independent inputs: (A) Cyanin dyes, (B) Carotenoids, (C) Protein complexes, (D) Mixed and unclassified pigments.

From the originally constructed datasets from literature, the visualized changes in PCEs can be seen in the created bubble charts illustrated in Fig. 5 for every studied pigment class. The relative size of the bubbles translates variations in PCEs of various studied TiO_2-based/iodide-triiodide-liquid-redox-based naturally-sensitized DSSCs as a function of PI, X, and BG of the selected dyes. A larger bubbles size indicates the ability of a dye to achieve a

relatively high PCE when compared to the other investigated dyes within the same pigment class. For example, cyanin dyes category [including rutin (RU), betaxanthin (BE), anthocyanin (AN), zinc phthalocyanines (ZP), cyanine (CYA), betalains (BET)] shown in Fig. 5(A) confirms that efficient.

DSSCs that are based on cyanins with the following characteristics PI=10–19, X=0–4, and BG=1.1–1.68 eV would achieve PCEs>5.5%. However, recommended dye characteristics in carotenoids class [from the following: xanthophylls carotenoids: yellow (XC-Y), xanthophylls carotenoids: red (XC-R), xanthophylls carotenoids: pure orange (XC-PO), xanthophylls carotenoids: raw orange (XC-RO), xanthophylls carotenoids: cocktail (XC-C), lycopene carotenoids (LC), carotenoid (CAR)] were inferred from Fig. 5(B) suggesting that carotenoids with approximately 11 free electrons and only one anchoring groups [PI=11, X=1] would yield in the highest PCEs>0.475% subjected to BG=5.23 eV. Alternatively, dyes from protein complexes [e.g. light-harvesting complex II (LH2-1), reaction centers (RC), light-harvesting complex II (LH2-2), RC photosystem I trimer (PSI), bacteriorhodopsin protein (BR-P), bacteriorhodopsin protein - Solid (BR-PS)] have shown that highest cell performance (PCEs>0.49%) was evident when the PPCs structural and electronic characteristics were in the following ranges PI=9–13, X=2–4, and BG=1.8–3.46 eV, as shown in Fig. 5(C). Mixed dyes [e.g. chlorophyll a + carotenoids (CC-1), bacteriorhodopsin proteins and bacterioruberin carotenoids (BRs), carotenoid + chlorophyll (CC-2), A. amentacea + P. pterocarpum (AP) from anthocyanin, carotenoid, and chlorophyll] and unclassified pigments [e.g. chromatophores (CHR), chlorophyll (CHL), xanthenes (XAN), coumarin (COU)] from Fig. 5(D) have shown that they can theoretically achieve the highest efficiency of PCEs>7.8% probably with the following constraints PI=4–10, X=3, and BG=1.98–3.57 eV. Such high-efficiency observations found for dyes with low free electrons from mixed and unclassified dyes might be explained by the fact that high numbers of free electrons in association with π-bonds could increase excitation competitions. Visible-light incident allows excited electrons to transport through anchoring groups (e.g. carboxyl).

V. Conclusion

We developed high-accuracy predictive models to study the impact of dye structural, electronic, and molecular properties on the PCE of DSSCs. Tree training algorithms 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) and altogether would ensure smooth charge injection and forward electron transport.

Conflict of Interest

The authors declare no conflict of interest.

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] A. Hagfeldt, "Brief overview of dye-sensitized solar cells," *Ambio*, vol. 41, no. 2, 2012.
[2] V. Sugathan, E. John, and K. Sudhakar, "Recent improvements in dye sensitized solar cells: A review," *Renewable and Sustainable Energy Reviews*, vol. 52, pp. 54-64, 2015.
[3] B. Parida, S. Iniyan, and R. Goic, "A review of solar photovoltaic technologies," *Renewable and Sustainable Energy Reviews*, vol. 15, no. 3, 2011.
[4] N. Sawhney, A. Raghav, and S. Satapathi, "Utilization of naturally occurring dyes as sensitizers in dye sensitized solar cells," *IEEE J. Photovoltaics*, vol. 7, no. 2, 2017.
[5] 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*, vol. 121, p. 109678, 2020.
[6] 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)," *AIP Conference Proceedings*, vol. 2014, no. 1, 2018.
[7] S. Mathew, *et al.*, "Dye-sensitized solar cells with 13% efficiency achieved through the molecular engineering of porphyrin sensitizers," *Nature Chemistry*, vol. 6, no. 3, 2014.
[8] S. K. Srivastava, P. Piwek, S. R. Ayakar, A. Bonakdarpour, D. P. Wilkinson, and V. G. Yadav, "A biogenic photovoltaic material," *Small*, vol. 14, no. 26, 2018.
[9] H. Hug, M. Bader, P. Mair, and T. Glatzel, "Biophotovoltaics: Natural pigments in dye-sensitized solar cells," *Applied Energy*, vol. 115, 2014.
[10] K. Nagai, and A. Takagi, *Conjugated Objects: Developments, Synthesis, and Applications*, Pan Stanford, 2017.
[11] 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.
[12] M. R. Narayan, "Review: Dye sensitized solar cells based on natural photosensitizers," *Renewable and Sustainable Energy Reviews*, vol. 16, no. 1, 2012.
[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 *Proc. 2nd International Conference on Computing, Communication, Control and Automation*, 2016.
[14] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 1, 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] Ç. Odabaşı and R. Yıldırım, "Machine learning analysis on stability of perovskite solar cells," *Solar Energy Materials and Solar Cells*, vol. 205, p. 110284, 2020.
[17] H. Li, *et al.*, "Ensemble learning for overall power conversion efficiency of the all-organic dye-sensitized solar cells," *IEEE Access*, vol. 6, 2018.

[18] J. Im, S. Lee, T. W. Ko, H. W. Kim, Y. Hyon, and H. Chang, "Identifying Pb-free perovskites for solar cells by machine learning," *Npj Computational Materials*, vol. 5, no. 1, p. 37, 2019.
[19] V. Venkatraman, A. E. Yemene, and J. D. Mello, "Prediction of absorption spectrum shifts in dyes adsorbed on Titania," *Scientific Reports*, vol. 9, no. 1, 2019.
[20] H. A. Maddah, V. Berry, and S. K. Behura, "Cuboctahedral stability in Titanium halide perovskites via machine learning," *Computational Materials Science*, vol. 173, p. 109415, 2020.
[21] 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," *Renewable Energy*, vol. 162, 2020.
[22] 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," *Journal of Photochemistry and Photobiology B: Biology*, vol. 162, 2016.
[23] 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," *Journal of Photochemistry and Photobiology B: Biology*, vol. 162, 2016.
[24] K. Liu, *et al.*, "Spiro[fluorene-9,9′-xanthene]-based hole transporting materials for efficient perovskite solar cells with enhanced stability," *Materials Chemistry Frontiers*, vol. 1, 2017.
[25] Y. Shao, *et al.*, "Stable graphene-two-dimensional multiphase perovskite heterostructure phototransistors with high gain," *Nano letters*, vo. 17, no. 12, 2017.

Dr. Hisham A. Maddah is an assistant professor in the Chemical Engineering Department at King Abdulaziz University (KAU) in Rabigh. He earned his PhD in Chemical Engineering by the summer of 2020 from the University of Illinois at Chicago in "Naturally-Sensitized Photoanodes for Molecular Photovoltaics". Dr. Maddah completed his MS and BS from the University of Southern California in 2017 and KAU in 2012, respectively. He is an expert in dye-sensitized solar cells and the utilization of natural sensitizers for solar energy harvesting. His research interests include renewable energies, machine learning, and statistical analysis for solar-desalination systems, and materials stability.