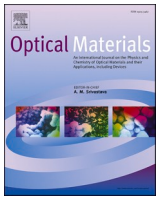
Optical Materials 128 (2022) 112343

Contents lists available at ScienceDirect

Optical Materials

journal homepage: www.elsevier.com/locate/optmat



Machine learning analysis on performance of naturally-sensitized solar cells

Hisham A. Maddah *

Department of Chemical Engineering, Faculty of Engineering—Rabigh Branch, King Abdulaziz University, Jeddah, 21589, Saudi Arabia

ARTICLE INFO

Keywords:
Supervised classification
Machine learning models
Dye solar cells
Natural sensitizers
Decision trees

ABSTRACT

Understanding natural photosensitizers characteristics in dye-sensitized solar cells (DSSCs) is necessary to achieve high power conversion efficiency (PCE). Here, we statistically investigate the possibility to achieve relatively high PCEs in naturally-sensitized-photoanode-based DSSCs using decision trees and support vector training of dye structural, electronic, and molecular properties. We studied the chemical structure and bandgap of 27 sensitizers, correlated to the literature reported PCEs while applying "in-between randomization" for datasets expansion. Training and testing algorithms were carried out *via* 4 (dye) predictors including the number of π-bonds (PI), anchoring groups (X), HOMO(H)-LUMO(L), and bandgap energy (BG), with 2 responses for the possibility to achieve PCEs >1.82% (Yes/No). Both HLBG-input and PIXBG-input models were found promising with the highest accuracies of $\sim$ 92% using trees classifiers and $\sim$ 96% with support vector classification, respectively. Testing datasets were chosen to check for built models accuracy and similarities were evident between the models' results (PIX, BG, HLBG, and PIXBG). Residual analysis showed trees models had the minimum statistical errors with narrow violons ($\pm$ 0.25 ranges). Despite that the existence of more anchoring groups allows firm molecules attachment to semiconductors for enhanced charge injection, the HLBG-input analysis confirmed that BG is the foremost controlling parameter ($\sim$ 3-fold > H), with BG/X importance ratio of 12 leading to the parameter's importance: BG (1) > H (0.32) > PI (0.08) > X (0.04). This work shows the potential of adopting trained classifiers for analyzing natural sensitizer's abilities to inject and separate generated electron-hole pairs for producing renewable, cost-effective, and sustainable energy.

1. Introduction

The current worldwide trends show the increasing population's demand for renewable and clean energies which can be met by securing utilization of highly efficient and environmental-friendly solar cells [1–6]. 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 [7–9]. There are many different kinds of natural and environmental-friendly photosensitizers [10], which can be extracted from light-harvesting complexes of anthocyanin, carotenoid, and chlorophyll biomolecules [4] emerged as an attempt to substitute the expensive and toxic [11] metal-based ruthenium polypyridyl dyes [12]. Natural dyes extracted from different natural sources (e.g. anthocyanin, carotenoid, flavonoid, aurone, chlorophyll, tannin, betalain obtained from fruits, flowers, leaves, seeds, barks, and various parts of plants or other biological sources) [13] have been previously proposed to be used as sensitizers in DSSCs due to their low cost and environmental friendliness [10]. However, the power conversion efficiency (PCE) from naturally-sensitized DSSCs is typically in the average range (<0.05–1.7%) [14–16] and requires 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 [17–19].

A photosensitizer is considered efficient for DSSCs when it fulfills these requirements [20]: (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 the semiconductor surface. The pigment's molecular structure, properties (i.e. hydrophilicity/hydrophobicity, solubility, surface chemistry, and stability of dye molecules), surface morphology, self-assembly, aggregation tendency, anchoring groups, and electrolyte interaction with photosensitizers are some of the basic parameters that need to be well understood to optimize DSSCs performance through using commercial and/or

* Corresponding author.
E-mail address: hmaddah@kau.edu.sa.

https://doi.org/10.1016/j.optmat.2022.112343
Received 6 February 2022; Received in revised form 1 April 2022; Accepted 6 April 2022
Available online 18 April 2022

natural photosensitizers [21]. Uniformly dispersed dyes in an optimal solvent prevent dye agglomeration and enhance dye/semiconductor surface interactions required for the attachment of dye acceptor segments, reducing series resistance and improving electron injection.

Enhanced photoanode sensitization is attainable by stacking semiconducting materials [22–24] with different bandgaps, or more viably by utilizing low-bandgap energy cosensitizers [25–28]. Hug et al. (2014) [21] collected available data for natural dyes utilized in DSSCs. Bixin, crocetin, crocin, betaxanthin, betalains, angostin, rutin, neoxanthin, violaxanthin, and lutein were among the investigated natural sensitizers extracted from plant-based sources. More importantly, anthocyanin and carotenoids (e.g. β-carotene) have been identified as one of the promising natural dyes with many studies on their application in naturally-sensitized DSSCs for photons-to-electrons conversion [21]. Carotenoids are highly light-sensitive pigments [21] due to their conjugated double π-bonds structure with an optimal chain length of seven [29] giving an approximated light-absorption range of 400–500 nm [30]. The highest observed performance with single carotenoids in DSSCs was 2.6% with optimal structure length consisting of double conjugated bonds [31].

The use of dye blends from the combination of carotenoids/chlorophylls derivatives can increase the efficiency up to 4.2% as found in earlier works from testing modified chlorophyll/ β-carotene, modified chlorophyll/lutein, modified chlorophyll/violaxanthin, and modified chlorophyll/neoxanthin [32]. Rutin [20] from mangosteen pericarp extract showed the highest efficiency of 1.17% while extract of rhoeo spathacea showed an efficiency of 1.49% [33]. Further, sicilian prickly pear extract (betaxanthin) was determined to be capable of achieving high efficiency of 2.06% [34]. Proteins pigment complexes (PPCs) might be good alternatives to carotenoids since they have a higher absorption coefficient, wider absorbance range (300–1100 nm), and higher conversion efficiency [35]. Under 1 sun radiation (100 mW/cm^2), the DSSCs performance have been estimated from reported photovoltaic properties from several natural photosensitizers extracted from plants and other biosources: black rice (anthocyanin) 3.27% [14], capsicum (carotenoid) 0.58% [14], erythrina variegata flower (carotenoid, chlorophyll) 2.06% [14], rosa xanthine (anthocyanin) 1.63% [14], kelp (chlorophyll) 1.18% [14], zinc phthalocyanines 4.6% [36] and 6.4% [37], cyanine 4.8% [38] and 7.62% [39], rose bengal (xanthenes) in ZnO-based DSSCs 1.56% [40], coumarin in TiO_2-based DSSCs 7.7% [41] and 9% [42].

Moreover, natural pigments have many advantages over commercial metal-synthetic dyes for DSSCs. Natural pigments from plant and/or bacteria sources are promising candidates to be integrated into DSSCs which can be simply installed as rolls in many daily used items such as handbags and clothing as well as building walls, windows, and integrated bio-photovoltaics [43,44]. Advantages of using natural pigments from natural sources include [30,35,43,44]:

(i) Plants, bacteria, and their proteins and carotenoids are abundant and cost-effective.
(ii) Dye extraction is easy, feasible, and can be also utilized on large scales (scalable).
(iii) Biodegradable, renewable, and sustainable which makes them very convenient.
(iv) Noncarcinogenic, environmental-friendly, and pose no health concerns to humans.
(v) Absorb most of the light energy due to wide absorption spectra (multi wavelengths).

In this work, we statistically investigate the possibility to achieve relatively high PCEs in naturally-sensitized-photoanode-based DSSCs using decision trees and support vector machines (SVMs) machine learning [45–47] of dye structural, electronic, and molecular properties. An earlier introduced concept in our work [46], called "in-between randomization", was then applied for an expansion of datasets including information on the structural, electronic, and bandgap of 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). Collected data were then divided into training and testing datasets to check for the classification accuracy of the four established input-parameters models: PIX-input, BG-input, HLBG-input, and PIXBG-input. Residual analysis, quartile range (QR) and inter-quartile range (IQR) methods, and confusion matrices were applied to find the best models with minimum statistical errors (outliers) based on the 1.5 IQR range-median decision rule. Lastly, we conducted a "parameters importance" analysis complemented with classification tree graphs 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 of the photoanode composite in DSSCs.

2. Machine learning for solar cells and renewable energies

Since the beginning of the technology and information age, machines (computers) in many trials have been used to learn specific patterns from provided data following certain algorithms to create what is called machine learning for data classification and decision making [48]. Grouping similar data together is also known as data mining (knowledge management) based on gathered data from literature [49]. Classification algorithm distributes variously mixed datasets into categories by constructing a model from learning the relation between input attributes and an output-dependent parameter (response) [50]. It is much easier to infer such relationships between inputs and outputs *via* supervised learning which correlates each input object to the desired output value to create unique vectors that can be used in new mapping predictions [48].

Decision Trees: A decision tree is a tree with internal nodes that are nothing but tests (based on input data patterns) and with leaf nodes used as categories (of these patterns). In short, tree builds classification models from observations of datasets attributes or predictors (represented in branches as a decision or terminal nodes) to reach conclusions about the target classification based on categorized responses (represented in leaf nodes) [51]. Classification trees are found to be predictable for outcomes and decisions and can resolve problems of data shortage or incompleteness [52]. A roots node of a tree is the parent of all existing sub-nodes, with nodes for attributes, each link (branch) shows a decision (rule) and each leaf shows an outcome [52–54]. A decision tree is a hierarchical representation of knowledge that works efficiently with discrete data for data separating sequence until a Boolean outcome at the leaf node is achieved [50]. It is an iterative process that splits the data into partitions with the continuous splitting to select the split that minimizes the sum of the squared deviations from the mean in the two separate partitions, applied to each of the new branches [51,55,56]. 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 [48–50]. 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) [50].

Support ***Vector Machines:*** Support Vector Machines (SVMs) analysis was first identified by Vladimir Vapnik and his colleagues in 1995 [57] as a nonparametric statistical regression technique relying on kernel function and parameters selection. SVM is also capable of building the nonlinear boundaries among the classes suitable in almost all classification tasks. SVMs work by searching for a particular line or decision boundary (hyperplane) for separating the datasets or classes while avoiding extra overfitting [48]. Cross-validation learning and gradient descent learning are some of the primary methods which are commonly used for kernel optimization and parameters selection. Considering a mixed kernel function strategy would result in models with decent

learning ability for generalization purposes [58]. Both predictor parameters and response values must be selected and analyzed carefully, respectively, from the training detests. Such selection would ensure having models with minimum errors and highest accuracy from finding a flat function $f(x)$ with ε as the maximum deviation from y_i for each training point x [59]. In other words, the function should have at most ε-deviation from the target from convex optimization based on three constraints and a tradeoff complexity. Typically, we need to find regression function: $f : R^D \rightarrow R$:

$$y = f(x) = \omega^T \varphi(x) + b \tag{1}$$

Knowing the following definitions, ω is a weight vector, $\varphi(x)$ is a selected function for data mapping of x from a low dimension to a high dimension space, and b is an up or down numeric value. SVMs regression adopts ε-insensitive function, where training data are assumed to follow a linear trendline with an accuracy associated with the ε value. Thus, function minimization can be optimized by converting the problem to an objective function as shown in the following [60]:

$$\min \ \frac{1}{2}\omega^2 + \frac{C}{2}\sum_{i=1}^{m}\left(\xi_i^2 + \xi_i^{*2}\right) \tag{2}$$

Under constraints:

$$\begin{aligned} &\omega^T\varphi(x_i) + b - y_i \leq \ \varepsilon + \xi_i \ , \ i = 1, 2, \ \ldots, \ m \\ &y_i - \omega^T\varphi(x_i) - b \leq \ \varepsilon + \xi_i^* \ , \ i = 1, 2, \ \ldots, \ m \\ &\xi_i \ , \xi_i^* \geq 0 \ , \ i = 1, 2, \ \ldots, \ m \end{aligned} \tag{3}$$

where $\xi_i \ , \xi_i^*$ is the relaxation factor, which should be equal to 0 when there is no error in the fitting. The performance of support vector regression is affected by the error penalty parameter C, which is the degree of punishment that is used to process the mistakenly divided sample. C is a tradeoff between the algorithm complexity and degree of mistakenly classified samples. In other words, C is the penalty factor, which is used as the weight between the error and the optimization objective. The first term (left term) of the function shown in **Eq. (2)**, for optimization purposes, allows generalizing the model from the improved fitting smoothness. The second term (right term) of the function shown in **Eq. (2)** reduces the error and that when $C > 0$, there will be errors in the estimated regression with penalty indicated by the error ε [58]. There should be an appropriate selection of the model that determines the most suitable kernel function for the data characteristics [60]. This would ensure accurate data training based on the constructed kernel function type and its relevant parameters [58].

Odabas et al. [61] applied machine learning tools and decision trees to analyze the impact of materials selection and deposition methods on the stability of organo-lead halide perovskite solar cells. Constructed datasets were gathered from 404 cells stability profiles over time under various testing conditions. Decision trees were built to deduce rules and guidelines that would serve in fabricating long-term stable perovskite solar cells [61]. Another work investigated the factors related to cell fabrication from 800 publication database; statistical tools including decision tree classification determined major trends or patterns and significance of factors for generalizing models for building efficient cells [62]. PCE prediction of DSSCs was earlier studied *via* multi-learner ensembles (GBDT, RF, SVM-KNN-WMA, L-SVM) based on clustering and then modeling approach for achieving high accuracy and generalization. The L-SVM-KNN-WMA based on the optimal subset of clustering was the optimal method for small datasets with high accuracy >91% for PCE prediction of all-organic DSSCs [47]. Machine learning allows manipulating datasets for predicting unknowing relationships as explained by Im et al. who applied gradient-boosted regression trees (GBRT) to predict structural heat of formation and bandgap from electronic structures for designing new lead-free perovskites [63]. 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 [64].

3. Methods and study framework

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) [16,21,30,35,65–92], (see Table S1 in the Supplementary, noting that similar studies with the same kind of natural dye and/or results were dropped yielding in 27 raw datasets). The PCEs were confirmed to have resulted from experiments only (not from theoretical calculations) for reliable analysis. Whenever previous authors reported both voltages and currents, we have double-checked PCEs. Collected PCEs from the different DSSCs with various natural sanitizers were then correlated to the dye type, structural, electronic, and molecular properties. The chemical structures of the studied dyes were then carefully drawn and manually evaluated to

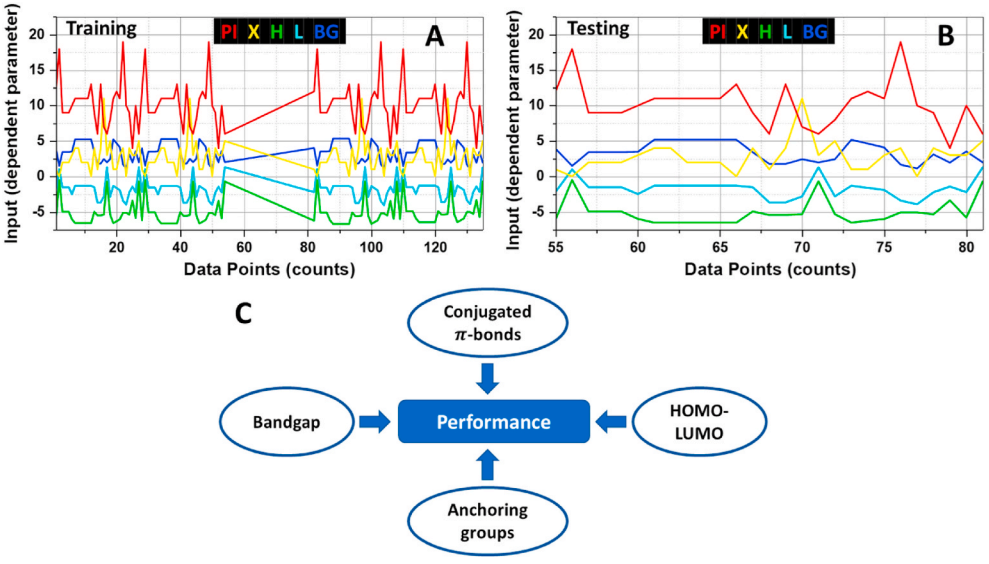


Fig. 1. Raw datasets (expanded) information taken from previous literature for the number of dye free π-electrons (PI), functional groups (X), HOMO (H), LUMO (L), and bandgap (BG) energy levels used in the supervised machine learning analysis for building accurate prediction models: **(A)** Training datasets, **(B)** Testing datasets, **(C)** Factors affecting DSSCs performance based on dye characteristics.

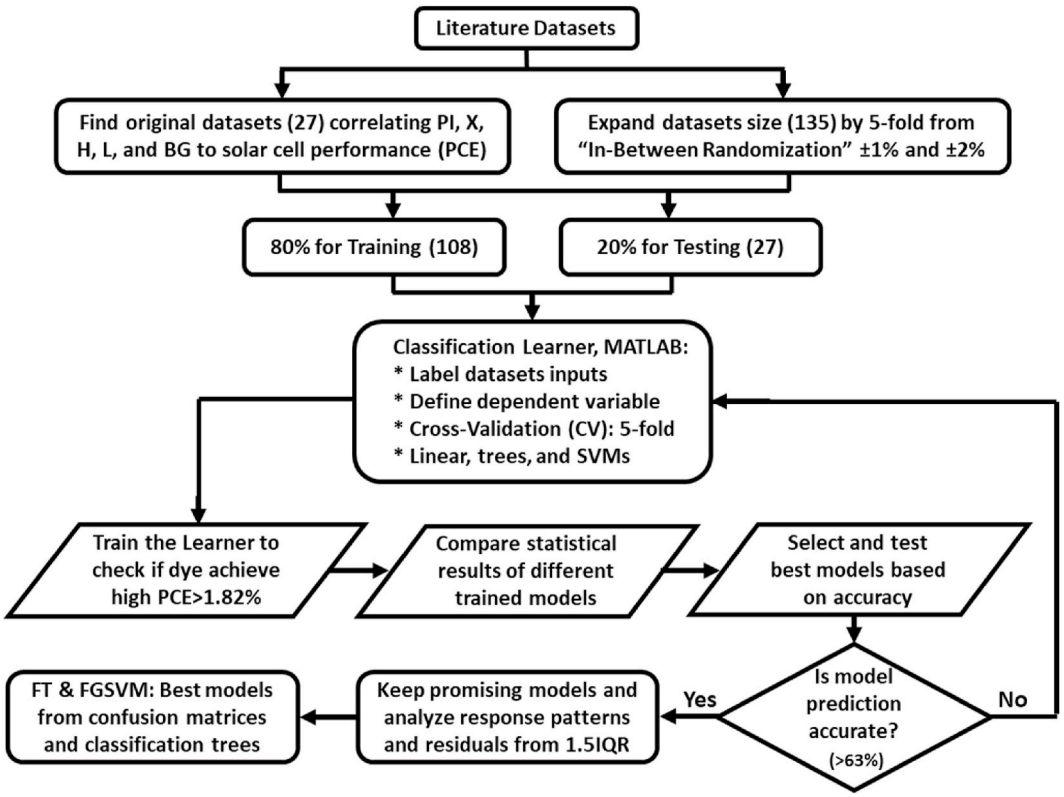


Fig. 2. Study framework showing the starting step of data collection, training, testing, followed by selection and analysis of the most accurate machine learning predictive models.

check for the existing number of double conjugated π-bonds and the existing number of anchoring or functional groups attached to the prime pigment structure. 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 literature [76–86] (bandgaps were not necessarily obtained from the same article that reported the DSSCs performance; however, in fact, most of the bandgap energies were taken from other theoretical related works on HOMO-LUMO levels) [4].

After gathering dyes structural and electronic information and estimating the number of free electrons (π-bonds), anchoring groups, and bandgaps, we have constructed original datasets which contain 27 different sensitizers mainly selected from dye classes such as carotenoids, protein complexes (LH2 and RC), flavonoids as cyanins (anthocyanins, betacyanin), chlorophyll, and chromatophores. Moreover, HOMO and LUMO energy levels were collected and added to the constructed datasets as a one-more-combined input parameter with the HOMO-LUMO data brought from the same theoretical works that reported the determined theoretical bandgaps of variously selected natural sensitizers. Thus, machine learning analysis made on the selected dyes in DSSCs had 4 predictors (independent parameters) 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). We used both HOMO/LUMO and bandgap as two separate predictors because we wanted to investigate more about absolute energy levels of both HOMO and LUMO and their association with the TiO_2 valence/conduction bands and their impact on electron excitation, injection, and forward transport. Such analysis would allow us to compare obtained results with the energy levels of the semiconductor in future studies so that one would come up with further discussion and maybe specific predictive models that would relate HOMO/LUMO to PCEs in TiO_2-based DSSCs. The 1.82% is 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 (except for the chromatophores-sensitized, PSI-sensitized, and BR-solids-sensitized solar cells that had different cell structures).

Then, an earlier introduced concept from our previous related and recently published work [46], called "in-between randomization", was applied for an expansion of datasets by 5-fold. Simply, we took leverage of inevitable errors from reported experimental [16,30,35,66–70,88,92–99] and theoretical results [76–86,100–108] by considering errors of $\pm$ 1% and $\pm$ 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 both HOMO and LUMO levels were also expanded with the taken errors since (BG = LUMO–HOMO), as in Fig. 1 (along with factors affecting performance). It is worth mentioning that PI and X values should have no errors since they solely depend on the well-known and fixed dye molecular structures and previously identified bonds and attached functional or anchoring groups. The expanded datasets were divided into two sets (80% training and 20% testing) to accurately establish classification models and at the same time to be able to test the established model's validity and prediction accuracy. The ultimate goal was to statistically investigate the possibility to achieve relatively high PCEs in naturally-sensitized-photoanode-based DSSCs using decision trees and SVMs machine learning. 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.

We conducted training analysis steps with trials and errors of the other existing machine learning classification models from the MATLAB toolbox which had shown low predictions accuracies of <63%. Accordingly, we selected the best-identified classification methods (decision trees and SVMs) for the training of the constructed and

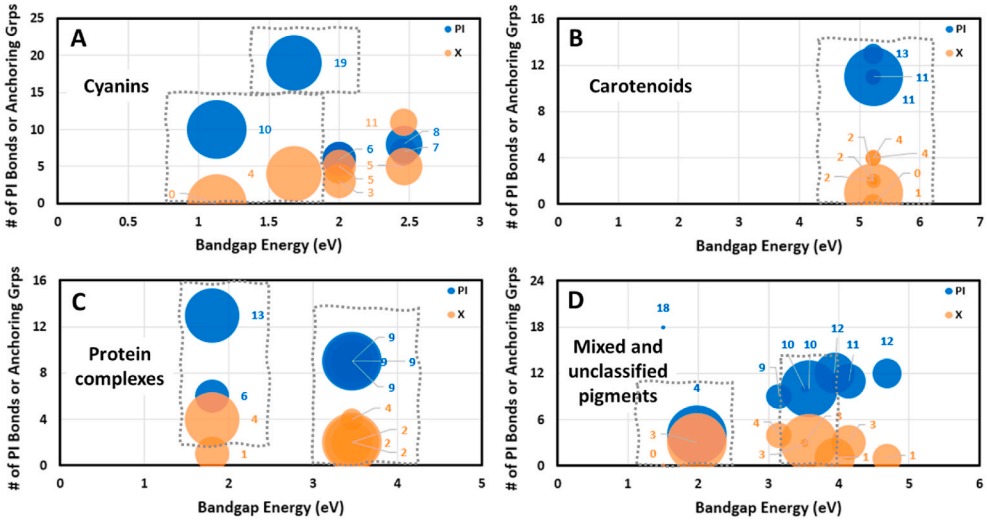


Fig. 3. Bubble charts for visualizing changes in PCEs of various studied TiO_2-based/iodide-triiodide-liquid-redox-based naturally-sensitized DSSCs showing the impact of independent inputs selected for the machine learning analysis of dye structural and electronic properties including number of π-bonds (PI), number of anchoring groups (X), and bandgap energy (BG): **(A)** Cyanin dyes (PCEs = 1.33% ~ 6.21%), **(B)** Carotenoids (PCEs = 0.008% ~ 0.475%), **(C)** Protein complexes (PCEs = 0.08% ~ 0.57%), **(D)** Mixed and unclassified pigments (PCEs = 0.04% ~ 8.35%). Gray dotted-boxes refer to the optimal ranges that would yield in the highest PCEs.

expanded datasets (Table S2). Testing datasets consisted of 27 sensitizers information (PI, X, HOMO-LUMO, and BG), and their PCEs translated into a relatively high or low performance based on data normalization from dividing by the found averaged-performance. 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 this numeric-to-character conversion decision analysis, we have been able to correlate every studied PCE to the various naturally-sensitized photoanodes and their pigments. Various trained classifiers were then tested statistically to check for selected models' accuracies, which in turn showed that only decision trees and SVMs had high prediction accuracies.

Residual, QR, and IQR methods were then used for outlier investigations of recorded responses from the trained models for training and testing datasets. Such analysis guided us towards models with minimum statistical errors (minimal deviations from actual observations) based on 1.5 IQR range-median decision rule and confusion matrices. The adopted study framework is shown in Fig. 2. Note that the various 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).

The interpretations of models errors *via* confusion matrices and 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 equations of identified statistical errors from the coefficient of determination (R^2) and residual are shown in **Eq. (4)** and **Eq. (5)**, respectively. 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 $\overline{x}_o$ is the experimentally obtained or observed values from averaging; $\overline{x}_p$ is the theoretically estimated or predicted values from averaging; and n refers to the datasets size or the number of experimental observations.

$$R^2 = \frac{\left[\sum_{i=1}^{n}\left(x_{o,i}-\overline{x}_o\right)\left(x_{p,i}-\overline{x}_p\right)\right]^2}{\sum_{i=1}^{n}\left(x_{o,i}-\overline{x}_o\right)^2 \times \sum_{i=1}^{n}\left(x_{p,i}-\overline{x}_p\right)^2} \quad (4)$$

$$\text{Residual} = x_o - x_p \quad (5)$$

From the originally constructed datasets from literature, the visualized changes in PCEs can be seen in the created bubble charts illustrated in Fig. 3 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. Keeping both PI and X within the same figure and sharing the same y-axis shows how both predictors are correlated to BG based on the different studied natural dyes associated numbers of free electrons and anchoring groups (i.e. the goal is to identify the optimal combination of BG, PI, X that would result in the highest PCEs or largest bubble sizes corresponding to each dye category. For example, tiny bubble size (18) in Fig. 3(D) indicates a very low efficiency at BG = 1.5 eV, PI = 18, X = 0 for the "Mixed + Unclassified" pigments.

The optimal ranges that would yield the highest PCEs are defined by gray dotted boxes as a function of BG, PI, and X for every dye class. 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. 3(A) confirms that efficient DSSCs results from preferably having cyanins with the following characteristics PI = 10–19, X = 0–4 and BG = 1.1–1.68 eV to 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. 3(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. 3(C). Mixed

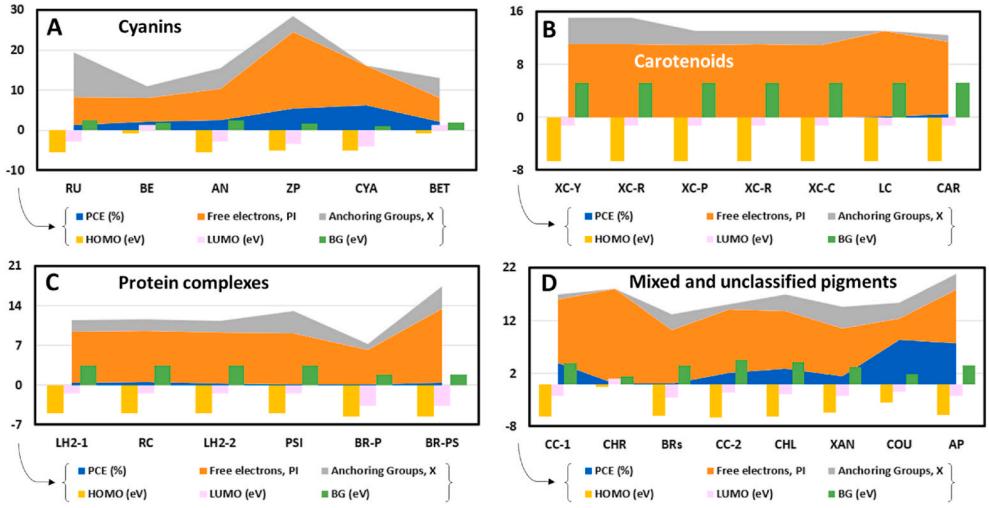


Fig. 4. Variations in observed PCEs (y-axis) of various studied TiO_2-based/iodide-triiodide-liquid-redox-based naturally-sensitized DSSCs correlated to the four studied independent inputs (PI, X, HOMO(H)-LUMO(L), BG) selected for the machine learning analysis and from using different dyes according to the studied pigment classes (x-axis) which were taken from literature: **(A)** Cyanin dyes (PCEs = 1.33% ~ 6.21%), **(B)** Carotenoids (PCEs = 0.008% ~ 0.475%), **(C)** Protein complexes (PCEs = 0.08% ~ 0.57%), **(D)** Mixed and unclassified pigments (PCEs = 0.04% ~ 8.35%). Stacked areas interpret information of PCE, PI, X; and clustered columns are for HOMO, LUMO, BG.

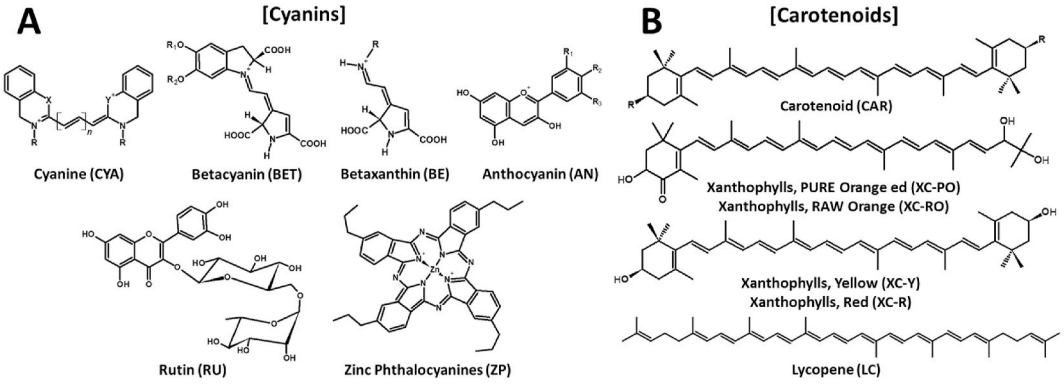


Fig. 5. Studied natural sensitizers (dyes) and their theoretically determined characteristics from dye molecular structures showing existing number of free electrons (PI) and anchoring groups (X) which were used in training machine learning models: **(A)** Cyanin dyes category, **(B)** Carotenoids.

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. 3(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 between free electrons. Visible-light incident allows excited electrons to transport through anchoring groups (e.g. carboxyl) and pigment components *via* donor-π-acceptor (D-π-A) segments or donor-acceptor-substituted π-conjugated bridge to the semiconductor [109]. Carboxylic groups attached to acceptor segments of dye molecules provide firm chemical attachment to semiconductor surface and rapid electron injection for reduced recombination; however, xanthene-based derivatives (e.g. rhodamine) might have shown less PCEs with more X because of their difficult quenching once covalently bonded to the TiO_2 surface which increases the unfavorable fluorescence that reduces conjugation donation power and decreases light absorption to the near-IR region [110–112]. Fig. 4 shows PCEs changes according to the studied predictors for the different dye types. Fig. 5 and Fig. 6 shows selected and studied natural dyes (sensitizers) molecular structures and their existing number of free electrons and anchoring groups used in model building (refer to Table S1 in the Supplementary for the constructed raw datasets of the studied dye categories and specific dye names with their determined structural and electronic characteristics).

4. Results and discussion

As discussed, tree training or SVMs algorithms were carried out *via* 4 predictors with 2 responses regarding the statistical possibility to achieve relatively high PCEs (Yes/No). The different four input models were built from training datasets including one or more of the following predictors: the number of dye structure π-bonds, number of dye anchoring groups, HOMO-LUMO, and bandgap energies. Experimental and/or previous theoretical observations were taken as a benchmark or a baseline to compare our built model's accuracy in predicting the impact of dye structural, electronic, and molecular properties on the power conversion efficiency of naturally-sensitized DSSCs, or specifically on the performance of the photoanode composite and its ability to absorb visible-light energy (photons) and separate generated electron-hole pairs.

The trained datasets of earlier observations from the literature were taken into consideration for checking the classification accuracy of the built models including decision trees and SVMs. The various 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). Comparisons of various models' predictions of the different trained data points with the earlier experimental or theoretical results are shown in Fig. 7. It is quite clear that FT,

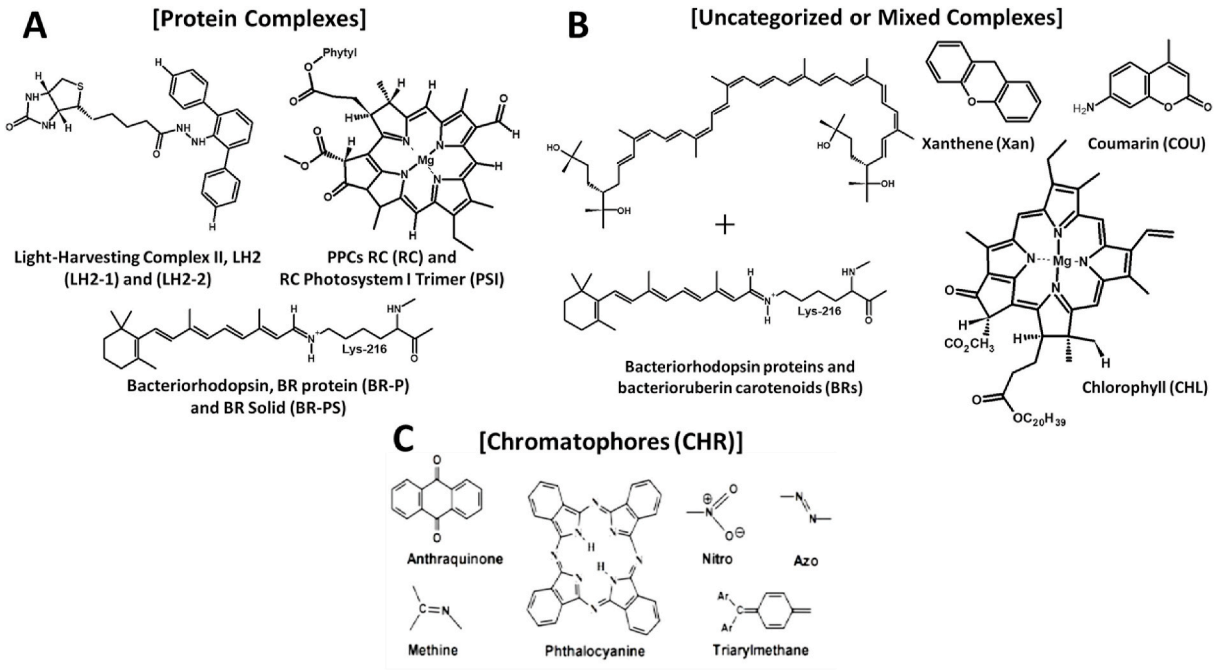


Fig. 6. Studied natural sensitizers (dyes) and their theoretically determined characteristics from dye molecular structures showing existing number of free electrons (PI) and anchoring groups (X) which were used in training machine learning models: **(A)** Protein complexes category, **(B)** Uncategorized and mixed dye complexes, **(C)** Chromatophores.

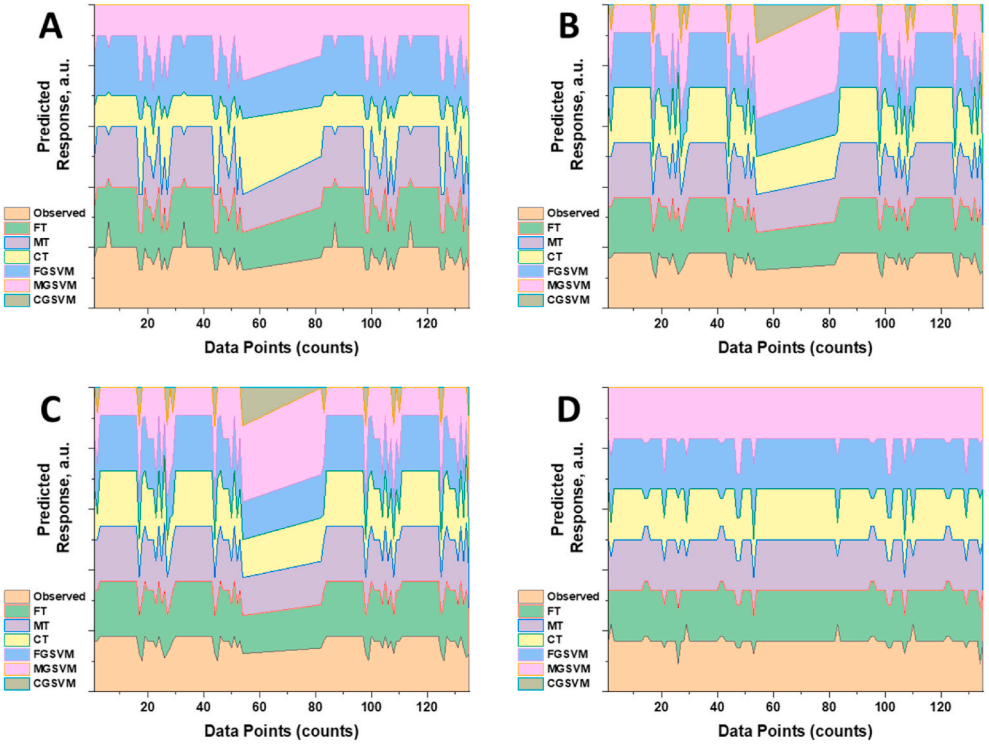


Fig. 7. Earlier experimental observations of training datasets compared with theoretical predictions identified from the trained classification models including decision trees and support vector machines (SVMs) using various selection of inputs as independent parameters: **(A)** PIX-input models, **(B)** BG-input models, **(C)** HLBG-input models, **(D)** PIXBG-input models.

MT, and FGSVM from the four unique input models had shown almost identical patterns and predictions to those original observations found previously according to the utilized raw data from the literature [4]. The CT models' results were found to have decent predictions placing it in the second rank from the top or among the best models that would predict actual observations. However, both MGSVM and CGSVM showed the worst classification accuracy to determine DSSCs with relatively high PCEs from the four unique input models.

There were similarities between the four models regarding the prediction responses (Yes/No), where the adjusted Yes = 2 and No = 1 are

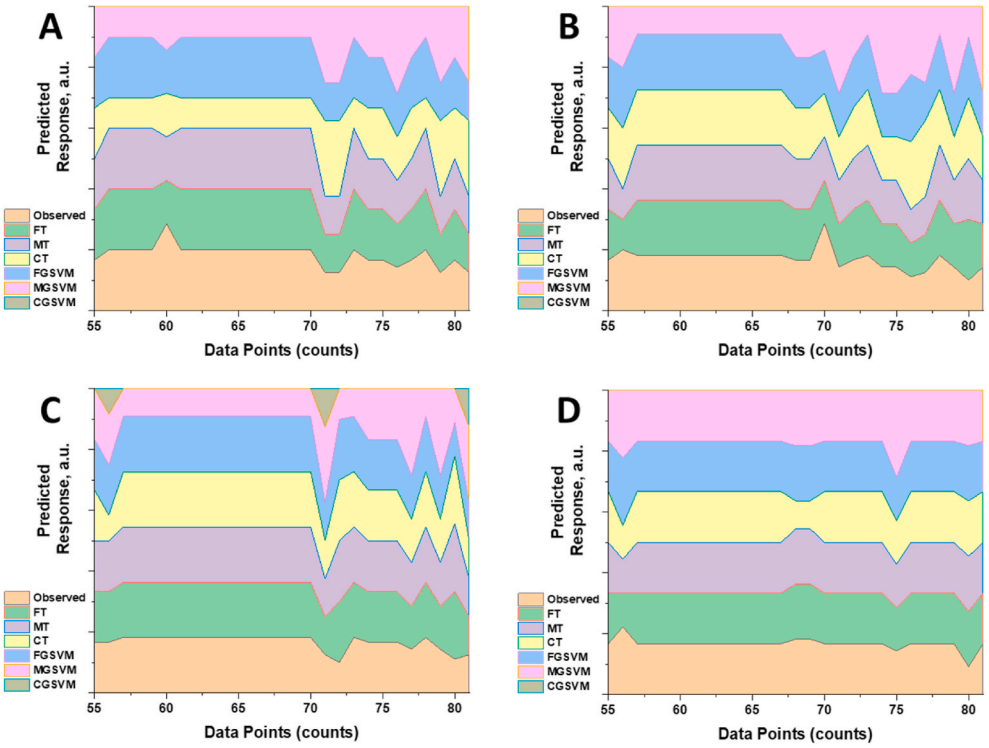


Fig. 8. Earlier experimental observations of testing datasets compared with theoretical predictions identified from the trained classification models including decision trees and support vector machines (SVMs) using various selection of inputs as independent parameters: **(A)** PIX-input models, **(B)** BG-input models, **(C)** HLBG-input models, **(D)** PIXBG-input models.

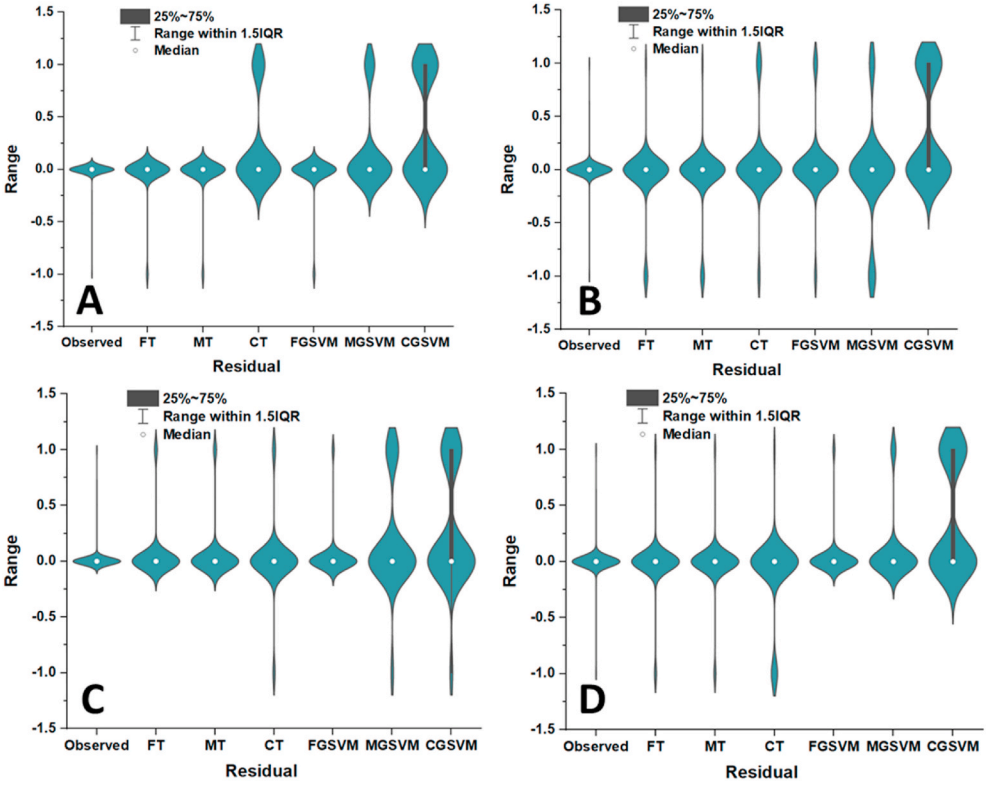


Fig. 9. Residual against datasets range within 1.5 IQR for outliers detection of the studied earlier experimental observations of training datasets compared with theoretical trees and SVMs trained-model predictions using various selection of inputs as independent parameters: **(A)** PIX-input models, **(B)** BG-input models, **(C)** HLBG-input models, **(D)** PIXBG-input models.

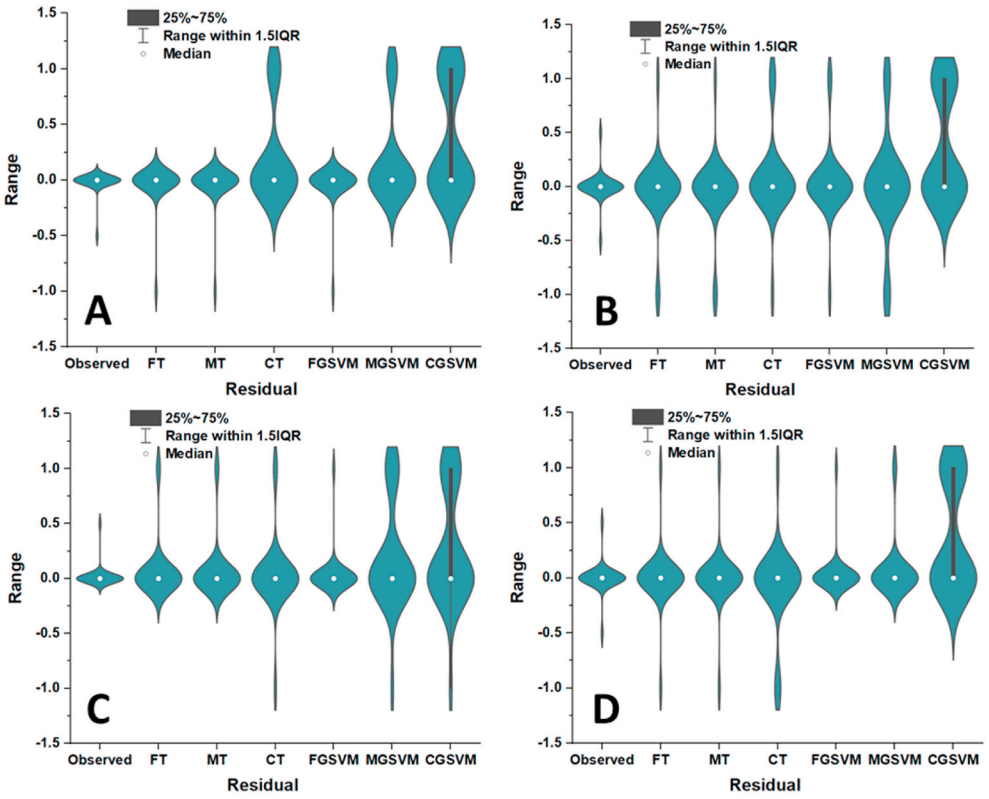


Fig. 10. Residual against datasets range within 1.5 IQR for outliers detection of the studied earlier experimental observations of testing datasets compared with theoretical trees and SVMs trained-model predictions using various selection of inputs as independent parameters: **(A)** PIX-input models, **(B)** BG-input models, **(C)** HLBG-input models, **(D)** PIXBG-input models.

from original datasets normalization (i.e. adding one to the following numeric-to-character conditions: Yes = 1, and No = 0). The PIX-input and BG-input models showed average prediction accuracy ~ 85% with FT and/or MT trained classifiers and maximum accuracy ~ 93% with FGSVM classification as illustrated in Fig. 7(A and B). Further, much more accurate predictions were possible from using the promising HLBG-input and PIXBG-input models, whereas that prediction accuracy can reach up to ~ 92% using FT and/or MT trained classifiers and ~ 96% with FGSVM classification as shown in Fig. 7(C,D) and Fig. S1 in the Supplementary**.** Such high prediction accuracies of the latter two discussed models were possible from introducing +3 predictor variables or controlling parameters into the classifiers MATLAB code (toolbox). This would result in models with better judgment abilities (the case for HLBG-input and PIXBG-input models) to show the promising dyes or pigments structures and electronic energies that would allow achieving relatively high PCEs in DSSCs.

In short, SVMs trained models were preferred over decision tree classifiers only if fine gaussian classification (FGSVM) is the applied classifier; otherwise, FT and MT tree models were found to be good enough to outperform the other SVMs classifications. To further prove our conceptual analysis, identified predictions were tested using trained models applied to selected testing datasets as shown in Fig. 8. The testing datasets were taken from the original (raw) data obtained from the literature and after carrying out the "in-between randomization" step for datasets expansion for training/testing steps. The only 27-point testing datasets showed results with very close predictions to those responses obtained from the trained models' training datasets analysis. Similarities between the same-model (PIX, BG, HLBG, and PIXBG) results for trained and tested datasets were evident as shown in Fig. 7(A-D), and Fig. 8(A-D), respectively. This confirms the earlier conclusions that FGSVM classification, as well as FT and MT tree classifiers, are effective in models training when using +3 predictors as controlling parameters for the estimation of whether a dye solar cell will be efficient enough or not using a dye (pigment) with similar characteristics of one of the studied dye specific characteristics.

Conducted residual versus range analysis for the four different input models showed that the least range of residual (close to the zeroth line) was found for FT, MT, and FGSVM results from PIX, HLBG, and PIXBG models with an exception for the BG-input model as shown in Fig. 9 and Fig. 10 training and testing datasets, respectively. These identified small ranges within 1.5 IQR indicate and confirm the three highly accurate statistical models which have the minimum detected outliers on a selected dataset. The distribution of the dataset's residual based on the normalized results (Yes = 1 and No = 0) shows the model deviations whereas those minimal deviations from actual observations are noticed when ranges are small (i.e. maximum value minus minimum value in the dataset is less which would mean much less spread of residuals indicating fewer prediction errors).

The previously well-known quartile range "QR Method" can be understood and applied from visualizing a box plot where the median is the center point, Q_1 and Q_3 are the lower and upper borders, respectively, of the inter-quartile range (IQR). The lower/first quartile is then from the minimum point up to the Q_1 border that is for the 25th percentile (i.e. 25% of the data lies between minimum and Q_1), and the upper/third quartile is then from the minimum point up to the Q_3 border that is for 75th percentile (i.e. 75% of the data lies between minimum and Q_3). Accordingly, the IQR can be determined from the difference between Q_3 and Q_1 ($IQR = Q_3 - Q_1$) which can produce decision ranges of various used datasets to detect outliers according to obtained ranges for residual from every model type. Any data point which was found to be set outside the identified ranges is considered as an outlier, knowing that both lower bound and upper bound are calculated with a scale of 1.5 as ($Q_1 - 1.5 \times IQR$) and ($Q_3 - 1.5 \times IQR$), respectively [113]. This is equivalent to considering outliers only for any data which lies beyond 2.7 of standard

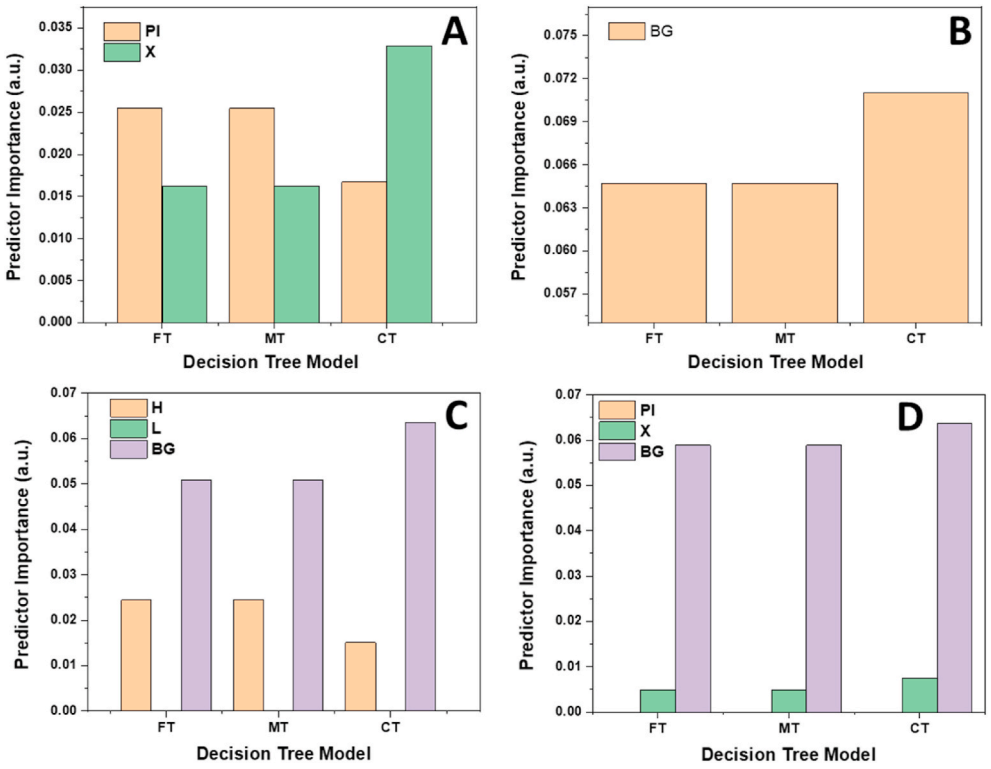


Fig. 11. Predictor importance of various involved independent input variables (or controlling parameters) that were used in building the three different decision tree models based on the unique selection of inputs as independent parameters: **(A)** PIX-input models, **(B)** BG-input models, **(C)** HLBG-input models, **(D)** PIXBG-input models.

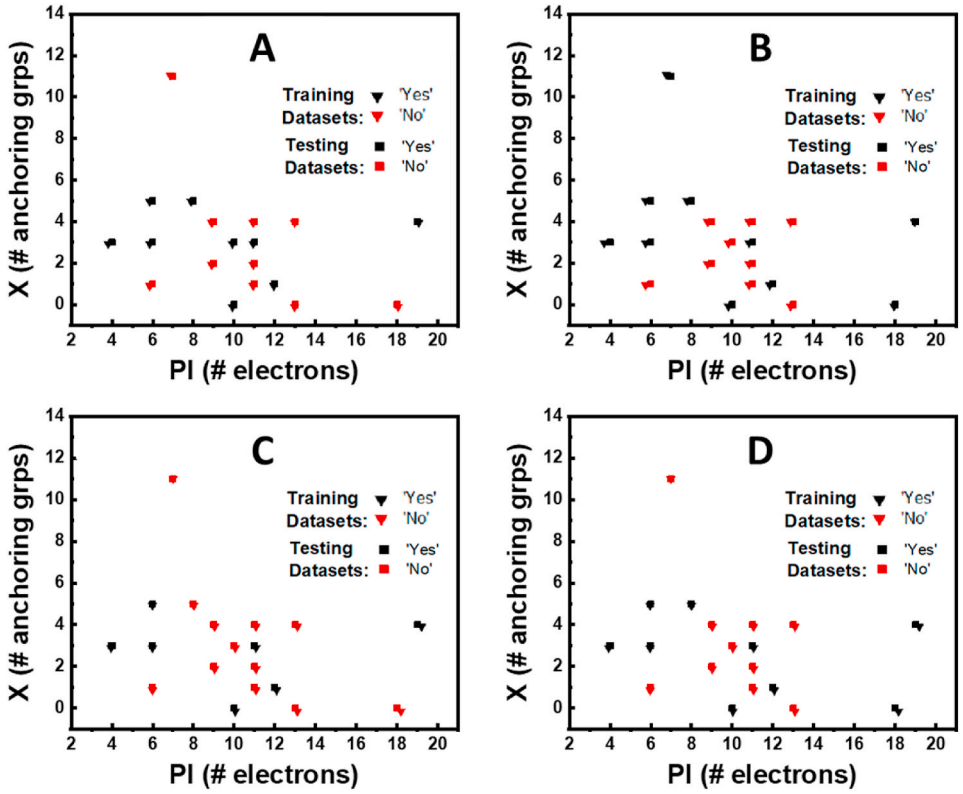


Fig. 12. Observed and predicted number of dye free π-electrons (PI) correlated with the existing functional groups (X), according to both the training datasets and their corresponding testing datasets obtained from the various trained FT tree models (input/response) and based on selection of different inputs as independent parameters: **(A)** PIX-input models, **(B)** BG-input models, **(C)** HLBG-input models, **(D)** PIXBG-input models.

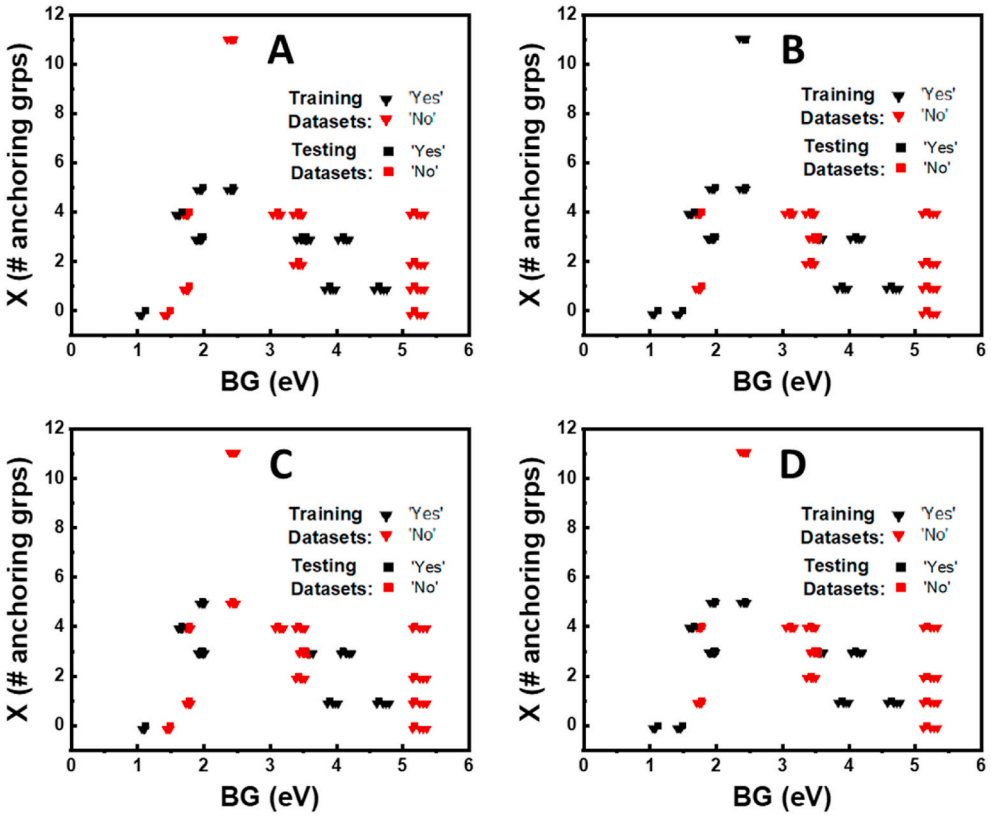


Fig. 13. Observed and predicted dye bandgap (BG) energy levels correlated with the existing functional groups (X), according to both the training datasets and their corresponding testing datasets obtained from the various trained FT tree models (input/response)and based on selection of different inputs as independent parameters: **(A)** PIX-input models, **(B)** BG-input models, **(C)** HLBG-input models, **(D)** PIXBG-input models.

deviation (σ) from the mean (μ) on either side of a normal distribution "bell curve" [113–115]. Hence, Gaussian distribution for outlier detection is relevant when the 1.5 IQR range-median decision rule is applied, which would allow us to check for the data outlier and draw conclusions about the accuracy of the built trained models.

Note that the first x-axis values (observed) in both Figs. 9 and 10 reserved the minimum residual since observed residual from experiments should be shown close to the zeroth line for the experimental results which were taken as our baseline. That is if we consider observations from earlier literature correct with minimal errors, residuals of the observed must align with the zero-line indicating minimum errors. We plotted the observed results concerning Yes/No responses minus itself showing the range from 0 to ± 0.1 (considering error ± 2%) where most of the studied dyes >50% were found to not have the ability to achieve a relatively high PCE >1.82%. The reason behind selecting FT, MT, and FGSVM as the best models appears from the narrow violons with small ± 0.25 ranges of residual from the four models. On the contrary, CT and MGSVM were found with larger errors numerically defined by violins with ± 0.5 ranges of residual indicating undesired doubled prediction errors for the dye abilities in DSSCs. CGSVM showed the worst accuracy among the built models with $-0.75 <$ residual $< +1$. Similarities between trained and tested datasets violins are evident from Figs. 9 and 10, respectively.

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. 11(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. 11(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. 11(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. 11 (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.

The correspondent number of free π-electrons estimated in the different studied dyes have been plotted against the number of existing anchoring groups in the dye structure from both the training and testing datasets as shown in Fig. 12. It was observed that the four input models from FT classification have been able to predict the number of either electrons or functional groups existing in various pigments very close to the experimental or theoretical observations reported in the literature, where (Yes/No) responses refer to dyes ability to efficiently absorb visible-light energy and produce excitons to be separated and transported towards the photoanode semiconductor for the achievement of maximum PCEs>1.82% from naturally-sensitized DSSCs. The good replication of the patterns and the relationships between anchoring groups and PI electrons is proof of the built FT model's reliability for the prediction of dye performance and its role in DSSCs. It is more probable to achieve high PCEs from using dye with the following characteristics [X = 2–6 & PI = 4–8], [X = 0–6 & PI = 10–12 or PI > 18].

To further analyze the impact of the dye bandgap on the DSSCs performance, the plotted charts of observation versus response patterns shown in Fig. 13 were used to correlate bandgaps to the available anchoring groups in the different studied dye structures. The generated plots were for the optimum trained FT model which has shown

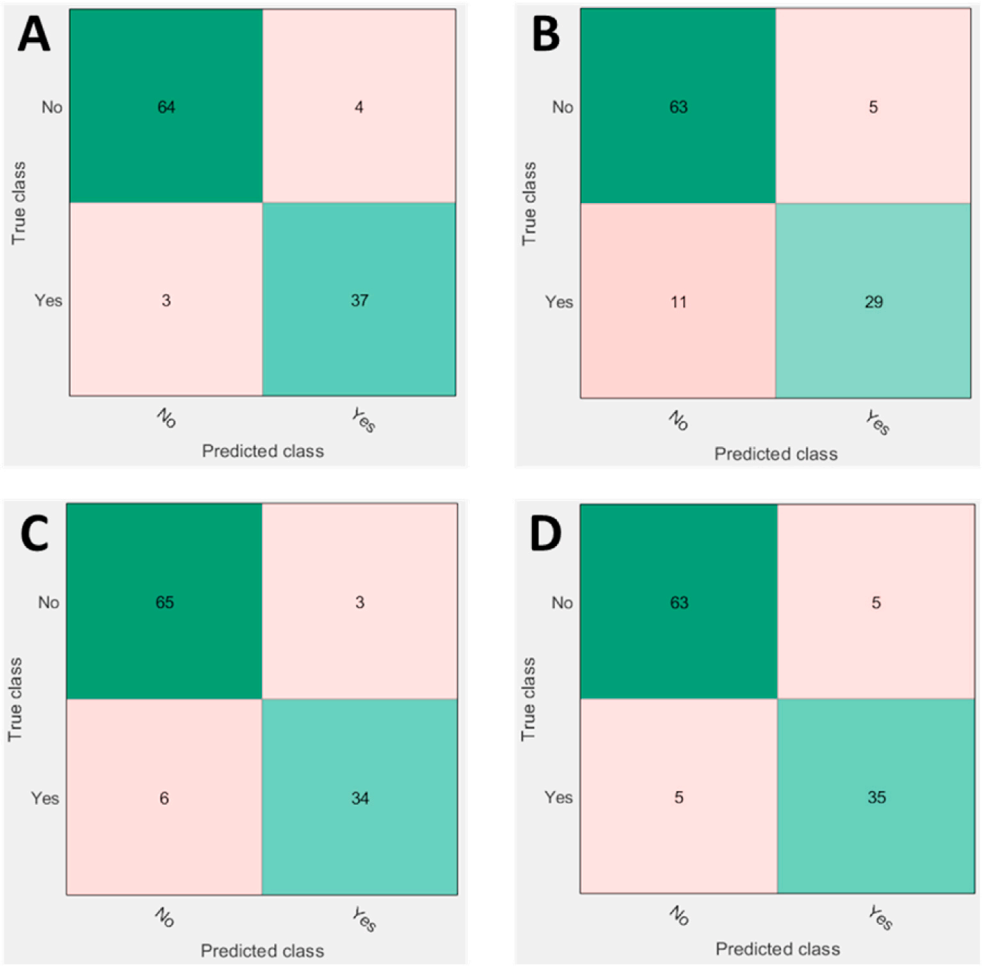


Fig. 14. Confusion matrices of the trained FT decision trees supervised machine learning models with highest accuracy showing prediction versus true responses (Yes/No) for a specific studied dye class or pigment type and whether it can achieve a relatively high PCE (>1.82%) in naturally-sensitized DSSCs: **(A)** PIX-input models, **(B)** BG-input models, **(C)** HLBG-input models, **(D)** PIXBG-input models.

minimum residuals according to the violin-box plots from Figs. 9 and 10. Again, trained models were able to produce similar-to-the-observations patterns predicted from the testing datasets confirming good prediction accuracy. No specific relationship can be obtained for the optimum bandgap and X [mix/match]; however, it seems that dye molecules would maximize the cell performance with [X = 3–4.5 and BG = 1.5–2.5], or with [X = 0–3 and BG = 1–2 or 4–5] as illustrated in Fig. 13 from the four input FT trained models. Further, confusion matrices of the trained FT decision trees and FGSVM supervised machine learning classifiers are shown in Fig. 14 and Fig. 15, respectively from the four input models.

The three FT decision trees from PIX, HLBG, and PIXBG input models showed the minimum statistical errors as per the results obtained from their confusion matrices shown in Fig. 14, with false responses less than 10 out of 108 training data points (i.e. >91% accuracy). However, the remaining created BG-input model from FT classification had the highest deviation from true responses with ± 15% errors indicating that bandgap of dyes cannot be taken as a sole independent parameter in defining the dye capabilities to improve PCEs in naturally-sensitized DSSCs. These results are comparable to the determined accuracies from MATLAB classification analysis as shown in Table 1, which calculated approximate 85% and 91% accuracies for BG-input and [PIX, HLBG, and PIXBG]-input models, respectively.

The FGSVM from PIX, HLBG, and PIXBG input models showed the highest accuracies with >94% among the various classifications utilized from the MATLAB machine learning toolbox. The reason behind this high accuracy is due to the models' abilities to predict true class responses as shown in Fig. 15. Nevertheless, the BG-input model achieved less than 85% as previously discussed.

Both HLBG and PIXBG input models have been proven to be the optimal decision tree classification for determining the dye impact on the overall performance of DSSCs based on the sensitizer's bandgap and HOMO energy levels as well as existing free electrons ready for excitation. According to the classification tree graphs plotted in Fig. 16 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 (see Figs. S2–S5 in the Supplementary). 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. 16(A and 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. 16(C and 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.

The relationships between the various studied features (predictors) and their impact on PCEs of DSSCs have been plotted in Fig. 17. There were 27 natural sensitizers and their data was gathered from literature

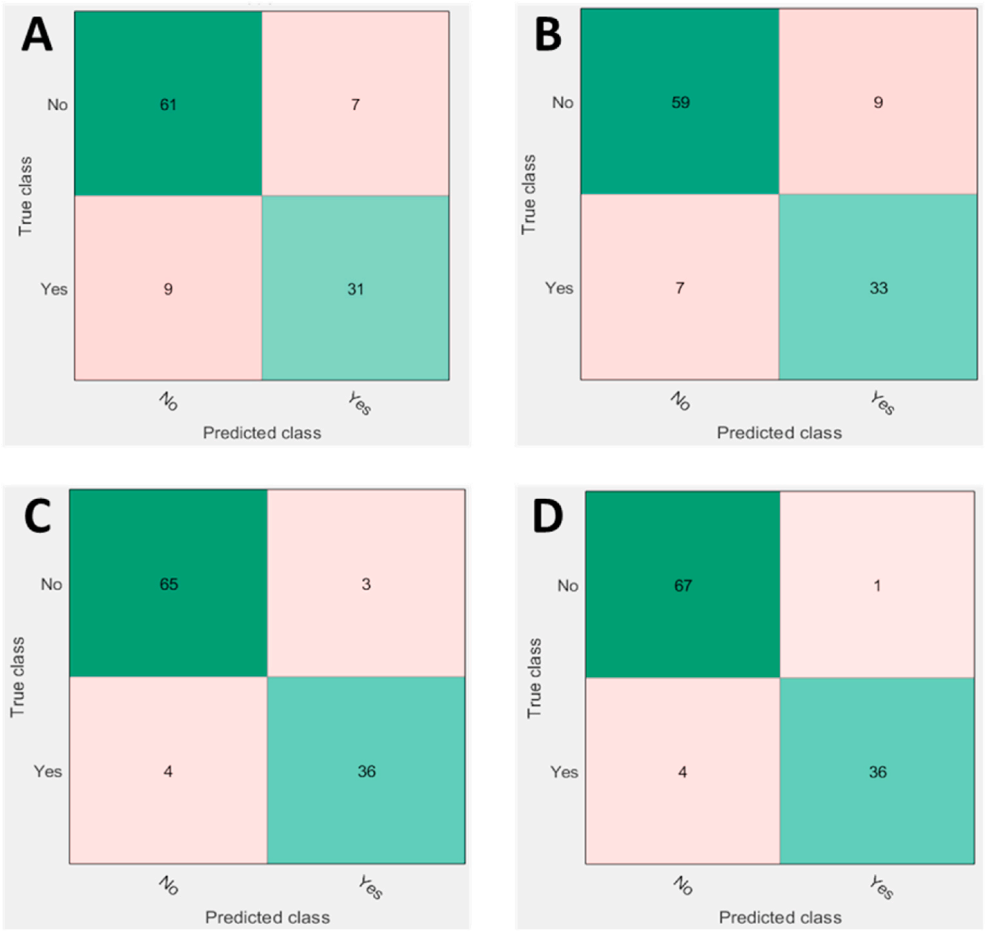


Fig. 15. Confusion matrices of the trained FGSVM support vector machines supervised machine learning models with highest accuracy showing prediction versus true responses (Yes/No) for a specific studied dye class or pigment type and whether it can achieve a relatively high PCE (>1.82%) in naturally-sensitized DSSCs: **(A)** PIX-input models, **(B)** BG-input models, **(C)** HLBG-input models, **(D)** PIXBG-input models.

Table 1
The input models for evaluation of natural dyes impact on PCEs with their predictors and prediction accuracies from classification training (expanded datasets)[a].

Predictors Model	Decision Trees			Support Vector Machines (SVMs)		
	FT	MT	CT	FGSVM	MGSVM	CGSVM
PIX	85.2%	85.2%	75%	93.5%	80.6%	63%
BG	85.2%	85.2%	83.3%	85.2%	71.3%	63%
HLBG	91.7%	91.7%	86.1%	93.5%	68.5%	66.7%
PIXBG	90.7%	90.7%	79.6%	95.4%	87%	63%

[a] The HLBG-input model's analysis confirmed that BG is among the top controlling parameters that is ~ 3-fold more important than H (HOMO) of the dye. The PIXBG-input model's analysis determined that BG/X importance ratio was about 12, which concluded the order of magnitude of parameters importance as BG (1) > H (0.32) > PI (0.08) > X (0.04).

(corresponding dye reference numbers can be found in Table S1 and Table S2 in the Supplementary). As expected, it is preferable to utilize dyes with the following characteristics: (i) can be sensitized easily from their low BG energy <3.57 eV, (ii) contains a reasonable number of π-conjugated bonds or free electrons (PI = 4–12) for less competition between free electrons for excitation, and (iii) with optimally a minimum number of anchoring groups of X = 3–5 for perfect dye attachment onto the semiconductor surface and provided charge pathways which would ensure smooth charge injection and forward electron transport.

The more the number of included features in classification training for model building, the more prediction accuracy we get from the model as shown in Fig. 18(A). The numbers shown in x-axis [1, 2, 3*, 3**] of Fig. 18(A) refers to # of features according to the different studied input models [BG, PIX, HLBG, PIXBG], respectively, whereas that accuracies were calculated from taking the average accuracy obtained from both decision trees and SVMs trained models shown in Tables 1 and 2 (shows determined prediction accuracies from classification training for each of studied input models). On average, it has been estimated that decision trees would outperform SVMs models by approximately a maximum +20% better accuracy based on average data analysis, Fig. 18(B). Nonetheless, FGSVM is the best overall model that can be adopted for the highest accuracy >95.4% for relating PCEs of DSSCs to sensitizers characteristics.

5. Conclusion

We developed high-accuracy predictive models to study the impact of dye structural, electronic, and molecular properties on the PCE of naturally-sensitized DSSCs. Models training was 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 relatively high PCEs (Yes/No). Most of the studied dyes >50% were found to not have the ability to achieve PCE >1.82%. It was estimated that decision trees would outperform SVMs by a maximum +20% better accuracy based on average data analysis. Conducted residual and confusion matrices analysis showed that

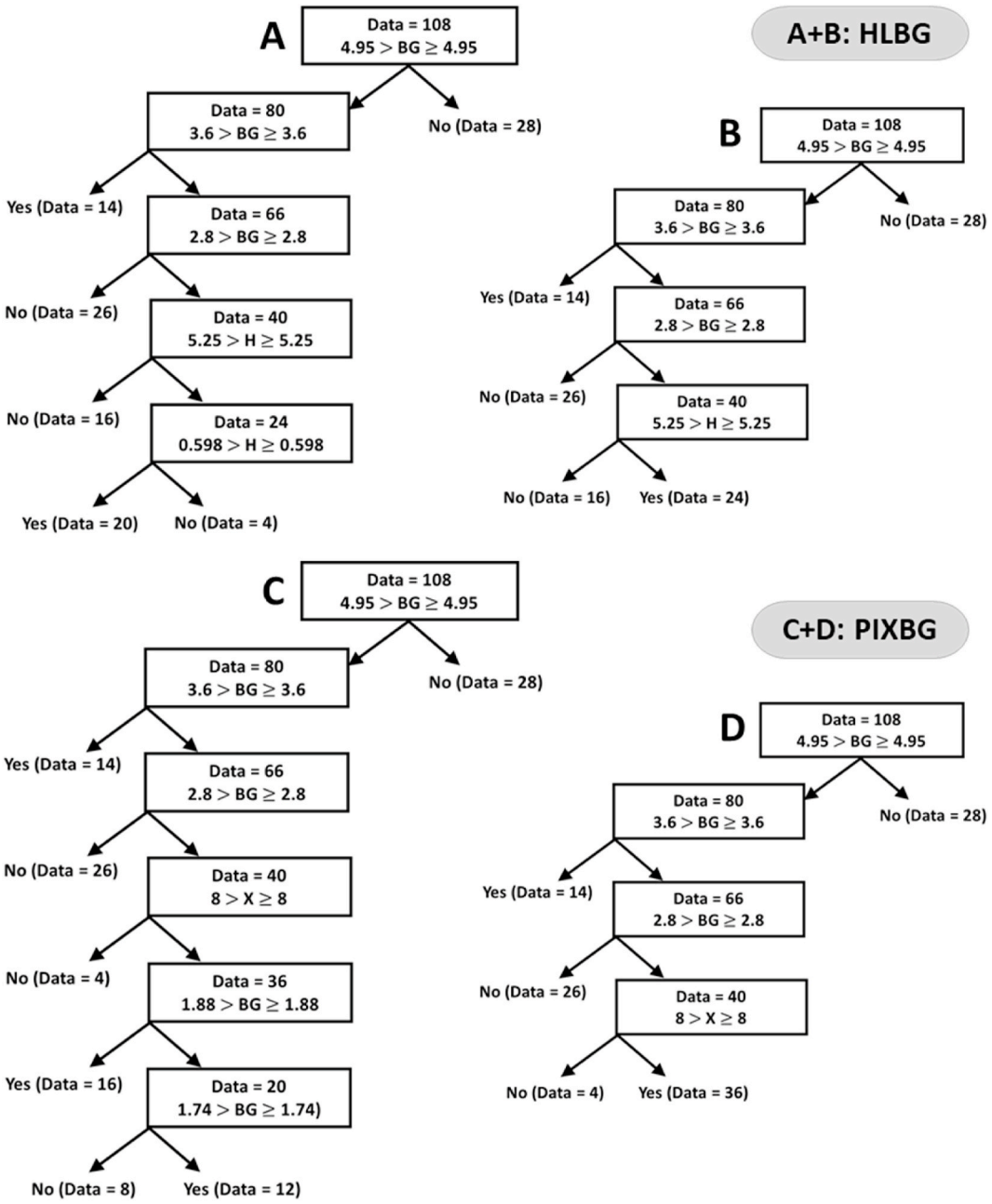


Fig. 16. Classification tree graphs from fine tree (FT), medium tree (MT), and coarse tree (CT) trained classifier models using HLBG and PIXBG parameters selection in input models: **(A)** FT and MT decision trees, **(B)** CT decision tree, **(C)** FT and MT decision trees, **(D)** CT decision tree. Note that BG, H, and X refer to bandgap, HOMO, and number of functional groups, respectively, with two responses (Yes/No) and number in the parenthesis corresponds to probability to achieve high DSSCs performance of >1.82% using the studied dyes.

minimum statistical errors (± 0.25 ranges) were found from FT, MT, and FGSVM models with almost identical patterns to original observations and with the highest accuracies of ~ 92% and ~ 96% for FT/MT and FGSVM, respectively. The HLBG-input model's analysis confirmed that BG is among the top controlling parameters that is ~ 3-fold more important than H (HOMO) of the dye, indicating the absolute energy level (HOMO) is not as critical as BG for the dye absorption abilities in TiO_2-based DSSCs. Moreover, the PIXBG-input model's analysis determined that BG/X importance ratio was about 12, which concluded the order of magnitude of parameters importance as BG (1) > H (0.32) > PI (0.08) > X (0.04). In summary, it is preferable to utilize dyes with the following characteristics: (i) can be sensitized easily from their low BG energy <3.57 eV, (ii) contains a reasonable number of π-conjugated bonds or free electrons (PI = 4–12) for less competition between free electrons for excitation, and (iii) with optimally a minimum number of anchoring groups of X = 3–5 for perfect dye attachment onto the semiconductor surface and provided charge pathways to ensure smooth charge injection and forward electron transport. The built supervised classification models would allow the scientific community to further study the impact of dye molecules and their absorption characteristics/capabilities on the performance of the photoanode composites for efficient and long-lasting naturally-sensitized solar cells.

CRediT authorship contribution statement

Hisham A. Maddah: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation,

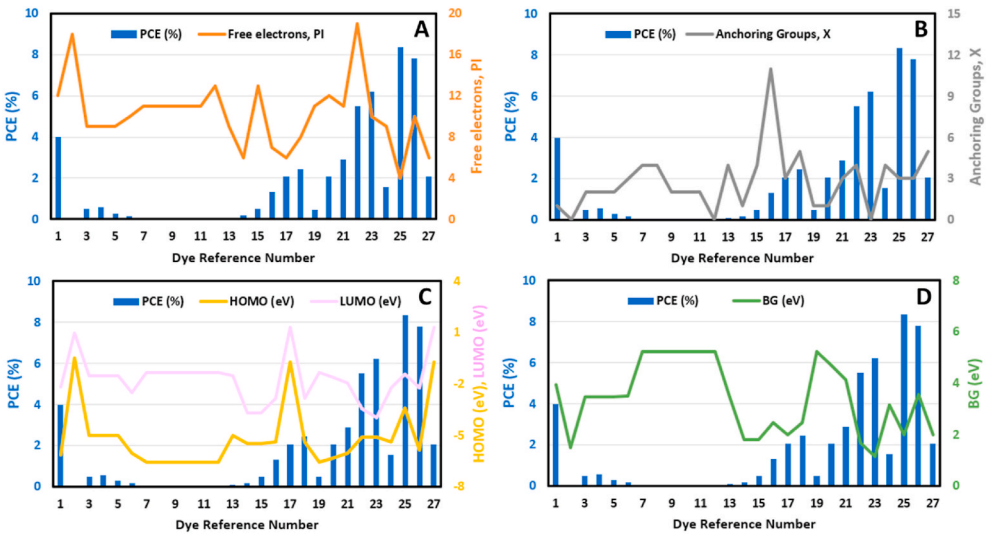


Fig. 17. Determined relationships between selected and studied dye features (predictors) and their impact on the performance of DSSCs according to different natural photosensitizers: **(A)** Free electrons (PI) vs. PCE, **(B)** Anchoring groups (X) vs. PCE, **(C)** HOMO and LUMO energy levels vs. PCE, **(D)** Bandgap (BG) energy vs. PCE.

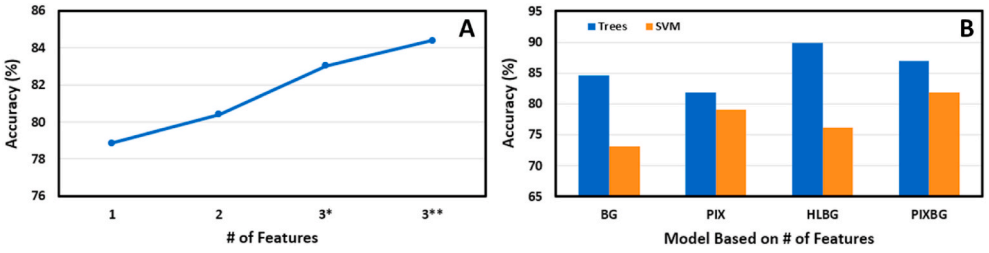


Fig. 18. Impact of selected number of features (based on input models) on the classification training accuracy: **(A)** Total average accuracy, **(B)** Specific average accuracy from decision trees and SVM. Note that [1, 2, 3*, 3**] refers to # of features from [BG, PIX, HLBG, PIXBG], respectively, and accuracies were calculated from taking the average accuracy; *considering H and L to be two features instead of a one combine predictor, **here we have 3 distinct features as PI, X, and BG.

Table 2
The input models for evaluation of natural dyes impact on PCEs with their predictors and prediction accuracies from classification training (unexpanded datasets)[a].

Predictors Model	Decision Trees			Support Vector Machines (SVMs)		
	FT	MT	CT	FGSVM	MGSVM	CGSVM
PIX	51.9%	51.9%	51.9%	63%	63%	63%
BG	74.1%	74.1%	74.1%	74.1%	59.3%	63%
HLBG	66.7%	66.7%	66.7%	70.4%	55.6%	63%
PIXBG	63%	63%	63%	63%	59.3%	63%

[a] The HLBG-input model's analysis confirmed that BG is among the top controlling parameters, but is equivalent to the importance of H (HOMO) of the dye. The PIXBG-input model's analysis determined that PI and X have no contributions (not logical), concluding the order of magnitude of parameters importance as BG (0.5) > H (0.5) > PI (0) > X (0).

Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research work was funded by General Research Fund Projects under grant no. (G: 556-829-1443). Therefore, the authors gratefully acknowledge technical and financial support from the Ministry of Education and King Abdulaziz University, DSR, Jeddah, Saudi Arabia.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.optmat.2022.112343.

References

[1] B. Gupta, T. Kumar Mandraha, P. Edla, M. Pandya, Thermal modeling and efficiency of solar water distillation: a review, Am. J. Eng. Res. 2 (12) (2013) 203–213.
[2] H.A. Maddah, Modeling the relation between carbon dioxide emissions and sea level rise for the determination of future (2100) sea level, Am. J. Environ. Eng. 6 (2) (2016) 52–61.
[3] G.N. Tiwari, H.N. Singh, R. Tripathi, Present Status of Solar Distillation, Sol. Energy, 2003.
[4] H.A. Maddah, V. Berry, S.K. Behura, Biomolecular Photosensitizers for Dye-Sensitized Solar Cells: Recent Developments and Critical Insights, Renewable and Sustainable Energy Reviews, 2020.
[5] H.A. Maddah, Modeling the feasibility of employing solar energy for water distillation, in: C. Hussain (Ed.), Handbook of Environmental Materials Management, Springer, Cham, 2018.
[6] H.A. Maddah, M.A. Shihon, Activated carbon cloth for desalination of brackish water using capacitive deionization, in: Desalination and Water Treatment, 2018.
[7] A. Hagfeldt, Brief Overview of Dye-Sensitized Solar Cells, Ambio, 2012.
[8] V. Sugathan, E. John, K. Sudhakar, Recent Improvements in Dye Sensitized Solar Cells: A Review, Renewable and Sustainable Energy Reviews, 2015.
[9] B. Parida, S. Iniyan, R. Goic, A Review of Solar Photovoltaic Technologies, Renewable and Sustainable Energy Reviews, 2011.

[10] N. Sawhney, A. Raghav, S. Satapathi, Utilization of Naturally Occurring Dyes as Sensitizers in Dye Sensitized Solar Cells, IEEE J. Photovoltaics, 2017.
[11] C. Cari, K. Khairuddin, T.Y. Septiawan, P.M. Suciatmoko, D. Kurniawan, A. Supriyanto, The Preparation of Natural Dye for Dye-Sensitized Solar Cell (DSSC), 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. 6 (3) (2014) 242–247.
[13] G. Richhariya, A. Kumar, P. Tekasakul, B. Gupta, Natural Dyes for Dye Sensitized Solar Cell: A Review, Renewable and Sustainable Energy Reviews, 2017.
[14] S. Hao, J. Wu, Y. Huang, J. Lin, Natural Dyes as Photosensitizers for Dye-Sensitized Solar Cell, Solar Energy, 2006.
[15] W.O. Nirwana Sari Halidun, E. Cahya Prima, B. Yuliarto, Suyatman, Fabrication dye sensitized solar cells (DSSCs) using β-carotene pigment based natural dye, in: MATEC Web of Conferences, 2018.
[16] S.K. Srivastava, P. Piwek, S.R. Ayakar, A. Bonakdarpour, D.P. Wilkinson, V. G. Yadav, A Biogenic Photovoltaic Material, Small, 2018.
[17] K. Nagai, A. Takagi, Conjugated Objects: Developments, Synthesis, and Applications, Pan Stanford, 2017.
[18] Samal, A Brief Discussion on Color: Why Does Such Conjugation Allow Absorption Of Visible Light?, 2014 [Online]. Available: https://people.chem.umass.edu/samal/269/color.pdf.
[19] H. Maddah, A. Jhally, V. Berry, 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.
[20] M.R. Narayan, Review: Dye Sensitized Solar Cells Based on Natural Photosensitizers, Renewable and Sustainable Energy Reviews, 2012.
[21] H. Hug, M. Bader, P. Mair, T. Glatzel, Biophotovoltaics: Natural Pigments in Dye-Sensitized Solar Cells, Appl. Energy, 2014.
[22] Musyaro'Ah, I. Huda, W. Indayani, B. Gunawan, G. Yudhoyono, Endarko, Fabrication and Characterization Dye Sensitized Solar Cell (DSSC) Based on TiO2/SnO2 Composite, AIP Conference Proceedings, 2017.
[23] M. Rani, S.K. Tripathi, A comparative study of nanostructured TiO2, ZnO and bilayer TiO2/ZnO dye-sensitized solar cells, J. Electron. Mater. 44 (4) (2015) 1151–1159.
[24] Y. Zhou, et al., Dye-sensitized solar cells based on nanoparticle-decorated ZnO/SnO 2 core/shell nanoneedle arrays, Appl. Surf. Sci. 292 (2014) 111–116.
[25] A. Kojima, K. Teshima, Y. Shirai, T. Miyasaka, Organometal halide perovskites as visible-light sensitizers for photovoltaic cells, J. Am. Chem. Soc. 131 (17) (2009) 6050–6051.
[26] J.H. Im, C.R. Lee, J.W. Lee, S.W. Park, N.G. Park, 6.5% efficient perovskite quantum-dot-sensitized solar cell, Nanoscale 3 (10) (2011) 4088–4093.
[27] K. Kakiage, T. Kyomen, M. Hanaya, Improvement in photovoltaic performance of dye-sensitized solar cells by cosensitization with an organometal halide perovskite, Chem. Lett. 42 (12) (2013) 1520–1521.
[28] H.S. Kim, et al., High efficiency solid-state sensitized solar cell-based on submicrometer rutile TiO2 nanorod and CH3NH3PbI3 perovskite sensitizer, Nano Lett. 13 (6) (2013) 2412–2417.
[29] X.F. Wang, et al., Dye-sensitized solar cells using retinoic acid and carotenoic acids: dependence of performance on the conjugation length and the dye concentration, Chem. Phys. Lett. 416 (1–3) (2005) 1–6.
[30] N. Órdenes-Aenishanslins, G. Anziani-Ostuni, M. Vargas-Reyes, J. Alarcón, A. Tello, J.M. Pérez-Donoso, Pigments from UV-resistant Antarctic bacteria as photosensitizers in dye sensitized solar cells, J. Photochem. Photobiol. B Biol. 162 (2016) 707–714.
[31] X.F. Wang, Y. Koyama, H. Nagae, Y. Yamano, M. Ito, Y. Wada, Photocurrents of solar cells sensitized by aggregate-forming polyenes: enhancement due to suppression of singlet-triplet annihilation by lowering of dye concentration or light intensity, Chem. Phys. Lett. 420 (4–6) (2006) 309–315.
[32] X.F. Wang, et al., Effects of plant carotenoid spacers on the performance of a dye-sensitized solar cell using a chlorophyll derivative: enhancement of photocurrent determined by one electron-oxidation potential of each carotenoid, Chem. Phys. Lett. 423 (4–6) (2006) 470–475.
[33] H. Zhou, L. Wu, Y. Gao, T. Ma, Dye-sensitized solar cells using 20 natural dyes as sensitizers, J. Photochem. Photobiol. Chem. 219 (2–3) (2011) 188–194.
[34] G. Calogero, J.H. Yum, A. Sinopoli, G. Di Marco, M. Grätzel, M.K. Nazeeruddin, Anthocyanins and Betalains as Light-Harvesting Pigments for Dye-Sensitized Solar Cells, Sol. Energy, 2012.
[35] A. Molaeirad, S. Janfaza, A. Karimi-Fard, B. Mahyad, Photocurrent generation by adsorption of two main pigments of Halobacterium salinarum on TiO2nanostructured electrode, Biotechnol. Appl. Biochem. 62 (1) (2015) 121–125.
[36] S. Mori, et al., Enhancement of incident photon-to-current conversion efficiency for phthalocyanine-sensitized solar cells by 3D molecular structuralization, J. Am. Chem. Soc. 132 (12) (2010) 4054–4055.
[37] T. Ikeuchi, H. Nomoto, N. Masaki, M.J. Griffith, S. Mori, M. Kimura, Molecular engineering of zinc phthalocyanine sensitizers for efficient dye-sensitized solar cells, Chem. Commun. 50 (16) (2014) 1941–1943.
[38] W. Wu, J. Hua, Y. Jin, W. Zhan, H. Tian, Photovoltaic properties of three new cyanine dyes for dye-sensitized solar cells, Photochem. Photobiol. Sci. 7 (1) (2007) 63–68.
[39] X. Ma, et al., A high-efficiency cyanine dye for dye-sensitized solar cells, Tetrahedron 64 (2) (2008) 345–350.
[40] I.B. Karki, J.J. Nakarmi, P.K. Mandal, S. Chatterjee, Effect of organic dyes on the performance of ZnO based dye-sensitized solar cells, Appl. Sol. Energy 49 (1) (2013) 40–45.
[41] K. Hara, et al., Design of new coumarin dyes having thiophene moieties for highly efficient organic-dye-sensitized solar cells, New J. Chem. 27 (5) (2003) 783–785.
[42] Z.S. Wang, Y. Cui, K. Hara, Y. Dan-Oh, C. Kasada, A. Shinpo, A high-light-harvesting-efficiency coumarin dye for stable dye-sensitized solar cells, Adv. Mater. 19 (8) (2007) 1138–1141.
[43] B.P. Jelle, C. Breivik, H. Drolsum Røkenes, Building integrated photovoltaic products: a state-of-the-art review and future research opportunities, Sol. Energy Mater. Sol. Cells 100 (2012) 69–96.
[44] L.P. Heiniger, et al., See-through dye-sensitized solar cells: photonic reflectors for tandem and building integrated photovoltaics, Adv. Mater. 25 (40) (2013) 5734–5741.
[45] H.A. Maddah, V. Berry, S.K. Behura, Cuboctahedral stability in Titanium halide perovskites via machine learning, Comput. Mater. Sci. 173 (2020).
[46] H.A. Maddah, M. Bassyouni, M.H. Abdel-Aziz, M.S. Zoromba, A.F. Al-Hossainy, Performance Estimation of a Mini-Passive Solar Still via Machine Learning, Renew. Energy, 2020.
[47] H. Li, et al., Ensemble learning for overall power conversion efficiency of the all-organic dye-sensitized solar cells, IEEE Access 6 (2018).
[48] M. Somvanshi, P. Chavan, S. Tambade, S.V. Shinde, A review of machine learning techniques using decision tree and support vector machine 2016, Proceedings - 2nd International Conference on Computing, Communication, Control and Automation, ICCUBEA, 2017.
[49] M.N. Anyanwu, S.G. Shiva, Comparative analysis of serial decision tree classification algorithms, Int. J. Comput. Sci. Secur. 3 (3) (2009).
[50] B. Charbuty, A. Abdulazeez, Classification based on decision tree algorithm for machine learning, J. Appl. Sci. Technol. Trends 2 (2021), 01.
[51] S. Sayad, "Decision Tree - Regression," Data Science: Predicting the Future, Modeling & Regression. [Online]. Available: https://www.saedsayad.com/decision_tree_reg.htm.
[52] A. Navada, A.N. Ansari, S. Patil, B.A. Sonkamble, Overview of use of decision tree algorithms in machine learning 2011, Proceedings - 2011 IEEE Control and System Graduate Research Colloquium, ICSGRC 2011, 2011, pp. 37–42.
[53] S.D. Jadhav, H.P. Channe, Efficient recommendation system using decision tree classifier and collaborative filtering, Int. Res. J. Eng. Technol. 3 (8) (2016).
[54] A. Gershman, A. Meisels, K. Lüke, L. Rokach, A Decision Tree Based Recommender System, Iics, 2010.
[55] Frontline Systems, Regression Trees," *Frontline Solvers*, 2020 [Online]. Available: https://www.solver.com/regression-trees.
[56] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, 2017.
[57] V.N. Vapnik, The Nature of Statistical Learning Theory, 1995.
[58] H. Wang, D. Xu, Parameter selection method for support vector regression based on adaptive fusion of the mixed kernel function, J. Control Sci. Eng. 2017 (2017) 1–12.
[59] MathWorks, Understanding Support Vector Machine Regression, 2020 [Online]. Available: https://www.mathworks.com/help/stats/understanding-support-vector-machine-regression.html.
[60] S. Ghassem Pour, F. Girosi, Joint prediction of chronic conditions onset: comparing multivariate probits with multiclass support vector machines, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016.
[61] Ç. Odabaşı, R. Yıldırım, Machine learning analysis on stability of perovskite solar cells, Sol. Energy Mater. Sol. Cells 205 (2020).
[62] Ç. Odabaşı Özer, R. Yıldırım, Performance analysis of perovskite solar cells in 2013–2018 using machine-learning tools, Nano Energy 56 (2019).
[63] J. Im, S. Lee, T.W. Ko, H.W. Kim, Y. Hyon, H. Chang, Identifying Pb-free perovskites for solar cells by machine learning, NPJ Comput. Mater. 5 (1) (2019) 37.
[64] V. Venkatraman, A.E. Yemene, J. de Mello, Prediction of absorption spectrum shifts in dyes adsorbed on titania, Sci. Rep. 9 (1) (2019).
[65] D.J. Kim, N. Takasuka, H. Nishino, H. Tsuda, Chemoprevention of Lung Cancer by Lycopene, BioFactors, 2000.
[66] K. Woronowicz, et al., Near-IR Absorbing Solar Cell Sensitized with Bacterial Photosynthetic Membranes, Photochem. Photobiol., 2012.
[67] Q. Fu, C. Zhao, S. Yang, J. Wu, The photoelectric performance of dye-sensitized solar cells fabricated by assembling pigment-protein complexes of purple bacteria on nanocrystalline photoelectrode, Mater. Lett. 129 (2014) 195–197.
[68] D. Yu, G. Zhu, S. Liu, B. Ge, F. Huang, Photocurrent activity of light-harvesting complex II isolated from spinach and its pigments in dye-sensitized TiO2 solar cell, Int. J. Hydrogen Energy 38 (36) (2013) 16740–16748.
[69] A. Mershin, et al., Self-assembled photosystem-I biophotovoltaics on nanostructured TiO2 and ZnO, Sci. Rep. 2 (1) (2012) 1–7.
[70] J. Chellamuthu, P. Nagaraj, S.G. Chidambaram, A. Sambandam, A. Muthupandian, Enhanced photocurrent generation in bacteriorhodopsin based bio-sensitized solar cells using gel electrolyte, J. Photochem. Photobiol. B Biol. 162 (2016) 208–212.
[71] K. Furuichi, T. Sashima, Y. Koyama, The first detection of the 3Ag- state in carotenoids using resonance-Raman excitation profiles, Chem. Phys. Lett. 356 (5–6) (2002) 547–555.
[72] H.R. Shahmohammadi, et al., Protective roles of bacterioruberin and intracellular KCl in the resistance of Halobacterium salinarium against DNA-damaging agents, J. Radiat. Res. 39 (4) (1998) 251–262.

[73] C. Gallardo, et al., Low-temperature biosynthesis of fluorescent semiconductor nanoparticles (CdS) by oxidative stress resistant Antarctic bacteria, J. Biotechnol. 187 (2014) 108–115.
[74] D.O. Plaza, C. Gallardo, Y.D. Straub, D. Bravo, J.M. Pérez-Donoso, Biological Synthesis of Fluorescent Nanoparticles by Cadmium and Tellurite Resistant Antarctic Bacteria: Exploring Novel Natural Nanofactories, Microb. Cell Fact., 2016.
[75] I.J. Iwuchukwu, M. Vaughn, N. Myers, H. O'Neill, P. Frymier, B.D. Bruce, Self-organized photosynthetic nanoparticle for cell-free hydrogen production, Nat. Nanotechnol. 5 (1) (2010) 73–79.
[76] G. Calogero, et al., Synthetic Analogues of Anthocyanins as Sensitizers for Dye-Sensitized Solar Cells," *Photochem*, Photobiol. Sci., 2013.
[77] T. Ruiz-Anchondo, N. Flores-Holguín, D. Glossman-Mitnik, Natural Carotenoids as Nanomaterial Precursors for Molecular Photovoltaics: A Computational DFT Study, Molecules, 2010.
[78] J.B.L. Martins, J.A. Durães, M.J.A. Sales, A.S.F.A. Vilela, G.M.E. Silva, R. Gargano, Theoretical investigation of carotenoid ultraviolet spectra, Int. J. Quant. Chem. 109 (4) (2009) 739–745.
[79] X.F. Wang, H. Tamiaki, O. Kitao, T. Ikeuchi, S.I. Sasaki, Molecular engineering on a chlorophyll derivative, chlorin e6, for significantly improved power conversion efficiency in dye-sensitized solar cells, J. Power Sources 242 (2013) 860–864.
[80] F.A. Castro, et al., On the use of cyanine dyes as low-bandgap materials in bulk heterojunction photovoltaic devices, Synth. Met. 156 (14–15) (2006) 973–978.
[81] Y.-S. Kim, J.-I. Shin, S.-Y. Park, K. Jun, Y.-A. Son, Electrochemical Studies on Heptamethine Cyanine Dyes, Text. Color. Finish., 2013.
[82] K. Liu, et al., Spiro[fluorene-9,9′-xanthene]-based hole transporting materials for efficient perovskite solar cells with enhanced stability, Mater. Chem. Front. 1 (1) (2017) 100–110.
[83] Y. Qian, et al., Spiro[fluorene-9,9′-xanthene]-based Universal Hosts for Understanding Structure-Property Relationships in RGB and White PhOLEDs, RSC Adv., 2015.
[84] B.B. Carbas, A.M. Önal, New fluorene-xanthene-based hybrid electrochromic and fluorescent polymers via donor-acceptor approach, Electrochim. Acta 66 (2012) 38–44.
[85] R. Sánchez-De-Armas, M.Á. San Miguel, J. Oviedo, J.F. Sanz, Coumarin derivatives for dye sensitized solar cells: a TD-DFT study, Phys. Chem. Chem. Phys. 14 (1) (2012) 225–233.
[86] S. Agrawal, P. Dev, N.J. English, K.R. Thampi, J.M.D. MacElroy, First-principles study of the excited-state properties of coumarin-derived dyes in dye-sensitized solar cells, J. Mater. Chem. 21 (30) (2011) 11101–11108.
[87] E. Daviso, et al., The electronic structure of the primary electron donor of reaction centers of purple bacteria at atomic resolution as observed by photo-CIDNP 13C NMR, Proc. Natl. Acad. Sci. U.S.A. 106 (52) (2009) 22281–22286.
[88] X.F. Wang, et al., Dye-sensitized solar cells using a chlorophyll a derivative as the sensitizer and carotenoids having different conjugation lengths as redox spacers, Chem. Phys. Lett. 408 (4–6) (2005) 409–414.
[89] M. Fairhead, D. Shen, L.K.M. Chan, E.D. Lowe, T.J. Donohoe, M. Howarth, Love-Hate ligands for high resolution analysis of strain in ultra-stable protein/small molecule interaction, Bioorg. Med. Chem. 22 (19) (2014) 5476–5486.
[90] F. Gai, K.C. Hasson, J.C. McDonald, P.A. Anfinrud, Chemical dynamics in proteins: the photoisomerization of retinal in bacteriorhodopsin, Science 80 (1998).
[91] A.M. Dummer, J.C. Bonsall, J.B. Cihla, S.M. Lawry, G.C. Johnson, R.F. Peck, Bacterioopsin-Mediated regulation of bacterioruberin biosynthesis in Halobacterium salinarum, J. Bacteriol. 193 (20) (2011) 5658–5667.
[92] T. Montagni, P. Enciso, J.J. Marizcurrena, S. Castro-Sowinski, C. Fontana, D. Davyt, M.F. Cerdá, Dye sensitized solar cells based on Antarctic Hymenobacter sp. UV11 dyes, Environ. Sustain. (1–9) (2018).
[93] H.A. Maddah, A.M. Chogle, Applicability of low pressure membranes for wastewater treatment with cost study analyses, Membr. Water Treat. 6 (6) (2015).
[94] H.A. Maddah, Optimal operating conditions in designing photocatalytic reactor for removal of phenol from wastewater, ARPN J. Eng. Appl. Sci. 11 (3) (2016).
[95] H.A. Maddah, et al., Determination of the treatment efficiency of different commercial membrane modules for the treatment of groundwater, J. Mater. Environ. Sci. 8 (6) (2017) 2006–2012.
[96] H.A. Maddah, A.S. Alzhrani, Quality monitoring of various local and imported brands of bottled drinking water in Saudi Arabia, World J. Eng. Technol. (2017) 551–563, 05, no. 04.
[97] H.A. Maddah, A.S. Alzhrani, M. Bassyouni, M.H. Abdel-Aziz, M. Zoromba, A. M. Almalki, Evaluation of various membrane filtration modules for the treatment of seawater, Appl. Water Sci. 8 (6) (2018) 150.
[98] H.A. Maddah, Modeling and designing of a novel lab-scale passive solar still, J. Eng. Technol. Sci. 51 (2019) 303.
[99] H.A. Maddah, Highly efficient solar still based on polystyrene, Int. J. Innovative Technol. Explor. Eng. 8 (2019) 3423–3425.
[100] H.A. Maddah, Polypropylene as a promising plastic: a review, Am. J. Polym. Sci. 6 (1) (2016) 1–11.
[101] H.A. Maddah, Modeling the relation between carbon dioxide emissions and sea level rise for the determination of future (2100) sea level, Am. J. Environ. Eng. 6 (2) (2016) 52–61.
[102] H.A. Maddah, Application of finite Fourier transform and similarity approach in a binary system of the diffusion of water in a polymer, J. Mater. Sci. Chem. Eng. 4 (4) (2016) 20.
[103] H. Maddah, A. Chogle, Biofouling in reverse osmosis: phenomena, monitoring, controlling and remediation, Appl. Water Sci. 7 (6) (2016) 2637–2651.
[104] H. Maddah, Analytical derivation of diffusio-osmosis electric potential and velocity distribution of an electrolyte in a fine capillary slit, Int. J. Eng. Technol. 18 (3) (2018) 1–9.
[105] H.A. Maddah, Numerical analysis for the oxidation of phenol with TiO2 in wastewater photocatalytic reactors, Eng. Technol. Appl. Sci. Res. 8 (5) (2018) 3463–3469.
[106] H.A. Maddah, Transport of electrolyte solutions along a plane by diffusion-osmosis, ARPN J. Eng. Appl. Sci. 15 (1) (2020) 46–51.
[107] H.A. Maddah, Predicting flux rates against pressure via solution-diffusion in reverse osmosis membranes, Eng. Technol. Appl. Sci. Res. 11 (2) (2021) 6902–6906.
[108] H.A. Maddah, Simulating fouling impact on the permeate flux in high-pressure membranes, Int. J. Adv. Appl. Sci. 8 (8) (2021) 1–8.
[109] A. Mishra, M.K. Fischer, P. Bauerle, Metal-free organic dyes for dye-sensitized solar cells: from structure: property relationships to design rules, Angew Chem. Int. Ed. Engl. 48 (14) (2009) 2474–2499.
[110] H. Klfout, A. Stewart, M. Elkhalifa, H. He, BODIPYs for dye-sensitized solar cells, ACS Appl. Mater. Interfaces 9 (46) (2017).
[111] K. Razmkhah, H. Little, S. Sandhu, T.R. Dafforn, A. Rodger, Optical properties of xanthene based fluorescent dyes studied by stretched-film linear dichroism, RSC Adv. 4 (71) (2014).
[112] E. Torres, CF® Dyes. What Started it All? Part 1. A History of Fluorescence, Biotium Company, 2019 [Online]. Available: https://biotium.com/blog/cf-dyes-what-started-it-all-part-1-a-history-of-fluorescence/.
[113] S. Chaudhary, Why '1.5' in IQR Method of Outlier Detection? Towards Data Science, 2019 [Online]. Available: https://towardsdatascience.com/why-1-5-in-iqr-method-of-outlier-detection-5d07fdc82097.
[114] B. Moska, D. Kostrzewa, R. Brzeski, Influence of the applied outlier detection methods on the quality of classification, in: Advances in Intelligent Systems and Computing, vol. 1061, 2020.
[115] G. Barbato, E.M. Barini, G. Genta, R. Levi, Features and performance of some outlier detection methods, J. Appl. Stat. 38 (10) (2011).