Check for updates

Solar RRL
www.solar-rrl.com

Predictive Modeling and SHAP (SHapley Additive ExPlanations) Analysis for Enhancing Natural Dye-Sensitized Solar Cell Performance

*Burcu Oral, Hisham A. Maddah, and Ramazan Yildirim**

Achieving high power conversion efficiency (PCE) in natural dye-sensitized solar cells remains a challenge. To better understand such challenges and explore potential solutions, a dataset is created from 113 experimental articles published recently. The data are analyzed using random forest and gradient boosting algorithms, and predictive models for open-circuit voltage (V_{oc}), short-circuit current density (J_{sc}), fill factor (FF), and PCE are developed. The model predictions are quite successful for all four performance indicators, with root mean square errors of 0.1, 1.7, 0.09, and 0.5 for V_{oc}, J_{sc}, FF, and PCE, respectively. The SHAP (SHapley Additive exPlanations) analysis is also performed to determine the effects of the descriptors on output variables. It is found that the dye extraction (such as dye/solvent ratio and extraction time) and deposition methods are highly influential for all four performance variables. It is also observed that chlorophyll, anthocyanin, and carotenoid dyes can improve V_{oc}, whereas there is no major dye type that can be identified for improvement of J_{sc}. Flavonoids, curcumin, and tannins dyes are found to be capable of increasing the cell FF; only the anthocyanin and chlorophyll can have a direct positive impact on the PCE output.

1. Introduction

The rising energy demands and environmental problems caused by the use of fossil fuels have increased the interest in renewable and clean energy sources like solar, wind, and geothermal energy, with solar energy being a major focus due to its abundance. While silicon-based solar cells have dominated the field since the 1950s, their high costs have led to the search for new alternatives like dye-sensitized (DSSC), organic, and organolead halide perovskite solar cells. These emerging technologies promise cost-effectiveness, scalability, and efficient solar energy conversion, though challenges remain for their widespread commercialization.[1,2]

A DSSC consists of a photoactive semiconducting anode (usually TiO_2) sensitized by a light-sensitive dye, a cathode to close the electric circuit, and an iodine electrolyte for charge transfer.[3] A lot of effort has been spent to find or develop natural or synthetic dyes with high visible light absorption capacity. The use of natural dyes has several advantages in cost, fabrication, environmental effects, and availability even though their stability and efficiency ranges remain low[4]; they can be extracted from the roots, leaves, and barks of plants, and they are generally classified into six categories[5]: anthocyanins (blue-purple), betalains (red-brown), carotenoids (yellow-orange), chlorophyll (green), curcumin (orange-yellow), and flavonoids (red-purple). The compatibility of the dye and anode material (TiO_2) is crucial for DSSC; the carboxyl and hydroxyl groups are required to ensure good electronic binding with TiO_2.[4] Dye aggregation is another problem lowering the photocurrent; it can be solved by the modifications of the dye[6] or cosensitization to increase the electron injection efficiency.[7]

Nowadays, a large amount of data is generated in the research field for DSSCs, and the ability to use this data to extract knowledge is crucial for future work. Machine learning (ML), as a branch of artificial intelligence, offers effective tools for knowledge extraction from large datasets. The search for new materials, new dyes, and performance prediction of DSSCs using ML have been studied widely in recent years. For example, Venkatraman et al. created a database for the properties of DSSCs from 4,000 articles[8] while Venkatraman and Chellappan developed a database for dye aggregation and type of aggregation from 1,500 articles.[9] Al-Sabana et al.[10] in contrast, utilized the random forest (RF) algorithm to optimize DSSCs by focusing on the mesoporous TiO_2 layer's thickness and porosity. Their model, built from literature and numerical data, achieved 99.87% accuracy in predicting and optimizing

B. Oral, R. Yildirim
Department of Chemical Engineering
Bogazici University
Istanbul 34342, Turkey
E-mail: yildirra@bogazici.edu.tr

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

H. A. Maddah
Energy and Water Research Center (EWRC)
Al-Maddah Group
Jeddah 23613, Saudi Arabia

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/solr.202400432.

© 2024 The Author(s). Solar RRL published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/solr.202400432

DSSC performance, with experimental validation showing a 4.17% conversion efficiency. This study demonstrates the effectiveness and flexibility of RF algorithms for enhancing DSSC performance and optimizing design parameters. In addition to performance-related ML works, Wen et al.[11] developed a robust and interpretable quantitative structure–property relationship model by integrating ML with computational quantum chemistry. This model enabled virtual screening and assessment of synthetic accessibility to identify efficient organic dyes for DSSCs. Eight promising dyes with high power conversion efficiency (PCE) and good synthetic accessibility were selected. The model also provided insights into chemical rules, related to electron-donating capabilities, π bridge planarity, and so on, for high-performance dyes. Maddah also investigated the natural DSSCs using decision trees with simple dye descriptors.[12] The author used four dye properties (number of π bonds, anchoring groups, highest occupied molecular orbital–lowest unoccupied molecular orbital (HOMO–LUMO) levels, and bandgap) of 27 sensitizers to predict if the natural DSSC efficiency would exceed 1.82%. Further information on the application of ML on DSSCs can be found in the review article by Tomar et al.[13] The molecular information of a dye can be briefly associated with conjugated bonds and free electrons within the dye structure as well as the bandgap.[12,14]

ML has also been used in other new-generation solar cell designs. For example, Odabaşı and Yildirim developed a database of 1921 performance data points from 800 publications on organo-lead-halide perovskite solar cells, investigated their performance, and identified key factors for high efficiency, such as mixed cation perovskites and specific fabrication techniques.[15] Im et al. utilized gradient-boosted regression trees to analyze a dataset of electronic structures for halide double perovskites in search of lead-free alternatives for solar cells. Their approach identified key features and design guidelines for discovering new perovskites with desirable properties, such as heat of formation and bandgap.[16] Alwadai et al.[17] demonstrated the application of ML in optimizing organic solar cells, which is another new generation of solar cells, by predicting PCEs with high accuracy. Utilizing literature data and a molecular descriptor calculator, their study applied ML models like RF and bagging regressors, achieving R^2 values of 0.89.

Chen et al. reviewed the application of ML in perovskite solar cell research, highlighting recent advancements, and the types of questions being explored. They discuss methods for extracting meaningful features from perovskite data, accelerating synthesis and characterization, and theoretical screening for new materials.[18] Similarly, Mahmood and Wang[19] reviewed the application of ML in the development of organic solar cells, highlighting its potential to overcome the challenges of selecting suitable organic semiconductors.

To the best of our knowledge, there are no published works that focus on the prediction of natural DSSC performance under various cell fabrication and operation conditions (materials, variables, and procedures used to prepare the cells and irradiation conditions) as we have done in this work. We constructed a comprehensive dataset from the literature of natural DSSCs over the past five years and used it to predict the performance of the cell fabrication and operation variables using ML algorithms. We used RF and gradient boosting (GB) algorithms to analyze the literature data and predict the PCE, open circuit voltage (V_{oc}), short-circuit density (J_{sc}), and the fill factor (FF) of the solar cells as the performance measure; we also calculated the SHAP (SHapley Additive exPlanations) values of input variables (descriptors) for each outcome to be able to interpret the factor effects more effectively. We utilized fabrication-based descriptors, such as the type of natural dye used, the thickness of the photoelectrode, and the electrolyte composition, to relate to the performance of the natural DSSCs. Since our data were extracted from multiple sources (hence, we have both high and low-performance data), we could not directly compare our results with the literature for the performance levels. However, we compared our prediction to the experimental values reported in the literature to test the fitness of our ML models; we found that our models accurately predicted the performance of the solar cells while providing valuable information on the descriptor effects through SHAP analysis. Our study also highlights the potential use of ML to predict the performance of natural DSSCs and offers valuable insights for researchers and engineers in the field of photovoltaics.

2. Computational Details

A Web of Science (WOS) search was conducted using the keywords natural dye solar cell for the years 2018-2023 to create a dataset of recent experimental results in natural DSSCs. To minimize selection bias, the most relevant 150 articles were screened, resulting in 522 data points extracted from 113 selected articles (we used the relevance criterion of WOS for sorting). These articles were chosen for their inclusion of photoelectrochemical testing information (i.e., J–V measurements) and details about the source and type of dye used. Open circuit voltage (V_{oc}), short circuit current (J_{sc}), FF, and PCE were chosen as the output variables to assess the performance of the solar cells. Descriptors include details about the semiconductor, dye, their preparation procedures, as well as the electrolyte and operational conditions. Additional information about dyes, including free electrons and anchoring groups based on dye class from ref. [12] was also added. The final dataset comprises 26 descriptors and four output variables. This dataset is provided as an embedded Excel file in the Supporting Information file, while the list of variables and their ranges/levels are detailed in **Table 1**. Variables appearing with low frequency in the dataset were combined into a category labeled others.

The anchoring group is generally either ═O, ─COOH, or ─OH, radicals; the differences between their radicals and their specific impacts are unclear. However, they all share the same function, which is to allow the dye to attach its molecules firmly to the semiconductor substrate.[20] The numbering of these radicals is done by counting the number of existing groups in each dye category, therefore allowing an understanding of the correlation between the number of existing radicals and the adsorption strength translated in the smooth transport of electrons from the dye molecules to the TiO_2 and thereby enhancing the generated photocurrent. In short, more radicals would result in more current density due to the firm attachment of dye molecules and reduced interfacial resistance and/or electron recombination.

© 2024 The Author(s). Solar RRL published by Wiley-VCH GmbH

2367198x, 2024, 22, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/solr.202400432 by Ukraine - Cochrane, Wiley Online Library on [11/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 1. Range and levels of all descriptors in the dataset.

Type	Variable	Range/category level
Anode related	Material	TiO_2, TiO_2-composite, ZnO, others
	Deposition method	commercial,[a] doctor blade, screen printing, spin coating, others[b]
	Deposition thickness [μm]	0.17–100
	Treatment ($TiCl_4$, TTIP)	yes (1), no (0)
	Calcination temperature [°C]	25–550
	Calcination time [h]	0.08–5
	Active cell area [cm^2]	0.02–10
	Bandgap [eV]	0.89–3.98
	Dye absorption time [h]	0.03–72
	Counter electrode	Carbon, Pt, others
Dye related	Dye	Anthocyanin, betalain, carotenoids, chlorophyll, curcumin, mix, others
	Free electrons	4–12
	Anchoring Groups	1–11
	Extraction solvent	Ethanol, methanol, none, organic solvent, water
	Additional solvent	Acid, inorganic, none, organic, water
	Extraction pH	1–11
	Extraction temperature [°C]	4–300
	Dye/solvent ratio [$g\,mL^{-1}$]	0.003–25
	Extraction time [h]	0.08–168
	Cosensitized dye (second dye)	Anthocyanin, betalain, chlorophyll, curcumin, flavonoids, non-natural, none
	Ratio of dye/ Cosensitized dye	0.5–4
	Peak absorption wavelength	242–800
	Light intensity [$mW\,cm^{-2}$]	3.5–1000
Electrolyte related	Salt	Iodine, KI, Li, other, t-butyl-pyridine
	Solvent	3-methoxypropionitrile, acetonitrile, acetonitrile mix, aqueous, ethylene glycol, organic
	IL	Yes (1), no (0)
Outputs	V_{oc} [V]	0.02–1.3
	J_{sc} [$mA\,cm^{-2}$]	0.003–19.92
	FF	0.06–0.93
	PCE [%]	0.0001–5.28

[a]Commercial: the electrode is used as is without any deposition technique; [b]Others: the cases in the deposition procedure are none of the common methods or not clearly described.

Preanalysis of the dataset was performed using box and whisker, and swarm plots, which were generated using *ggbeeswarm*[21] package in *R* to see the structure of data. Then, the regression analysis was employed for the prediction of solar cell performance defined in terms of V_{oc}, J_{sc}, FF, and PCE. Due to the presence of categorical variables, RF and GB algorithms were used, and models were developed using *randomForest*[22] and *gbm*[23] packages in *R*, respectively. Other algorithms, such as linear regression and support vector regression, were also tested in the prediction of target variables but their predictive power was not satisfactory; this is probably because most of the descriptors in the dataset are categorical, which are more suitable for tree-based algorithms. SHAP values of descriptors for each output variable were also computed using *fastshap* package of *R*.[24]

Each model was developed using an 80–20% train–test split ratio and missing values were imputed by using the mean values of the training set. The random seed number was set to 124 to ensure reproducible results. The dataset was randomly divided into training and testing sets, model selection and validation were done using the training set; the performance of selected models was tested using the testing set. Hyperparameters, which were chosen as number of trees (*ntree*), and maximum number of features (*mtry*) for RF, and number of trees (*ntree*), learning rate (*shrinkage*), and *interaction depth* (*int_depth*) for GB analysis, were tuned (optimized) using fivefold cross-validation. In fivefold cross-validation, the training set is further divided into five randomly similar-sized subsets, and the model with the same hyperparameters is trained five times using onefold as validation data and the remaining fourfold as the training data. After the search for parameters is over, the root mean square error (RMSE) of the validation sets for each hyperparameter is calculated and the model hyperparameters with the lowest error are selected as the optimum model; mean absolute error and R^2 were also computed and provided with the results while only RMSE was used in discussions. Two different feature structures were created for the dataset; in the first structure (categoric), the "dye" and "salt" variables were used as they are (categorical variables having the names of dye and salt used in experiments as the levels); in second structure (encoded), these variables are converted into numeric variables via one hot encoding. The categoric structure has a total of 26 descriptors, whereas the encoded structure has 37 descriptors; both structures are presented in the Supporting Information. The best models of each output are summarized in **Table 2**.

3. Results and Discussion

The distribution of the data was presented for each output variable in **Figure 1** as box-and-whisker plots, which are used

Table 2. Best models developed for each output variable.

Output variable[a]	Feature structure	Model and hyperparameters
J_{sc}	Categoric	RF – *ntree*: 250, *mtry*: 14
V_{oc}	Encoded	GB – *ntree*: 290, *shrinkage*: 0.15, *int_depth*: 8
FF	Encoded	RF – *ntree*: 150, *mtry*: 6
PCE	Encoded	GB – *ntree*: 410, *shrinkage*: 0.15, *int_depth*: 7

[a]*ntree*: number of trees; *mtry*: maximum number of features; *int_depth*: *interaction depth*.

© 2024 The Author(s). Solar RRL published by Wiley-VCH GmbH

2367198x, 2024, 22, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/solr.202400432 by Ukraine - Cochrane, Wiley Online Library on [11/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

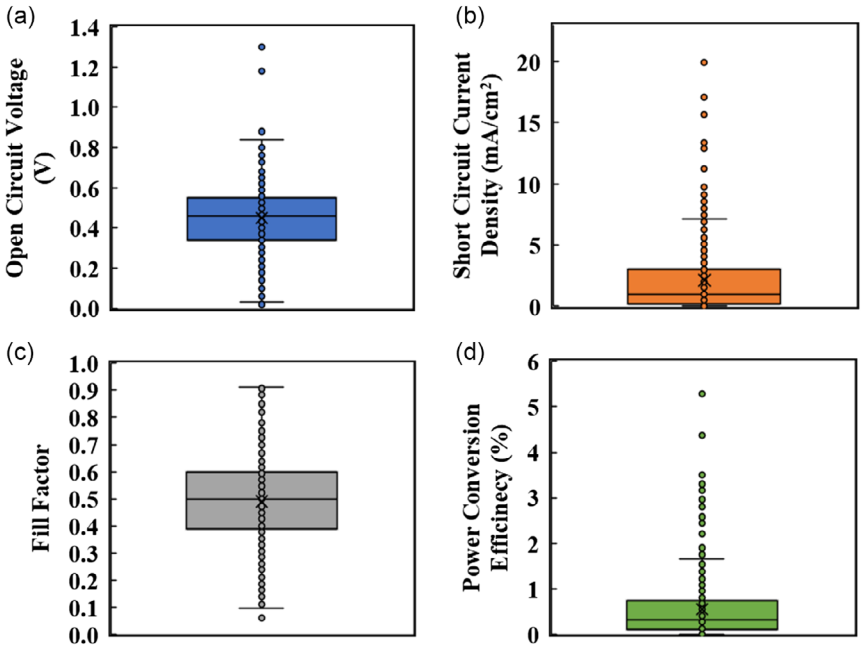

Figure 1. Box-and-whisker plots for outputs: a) open circuit voltage, b) short circuit current density, c) FF, and d) PCE.

to visually summarize the structure of the data. The box in the plot represents the middle 50% of the data, with the median line inside, while the tails show the upper and lower quartiles. Although some data points in plots can be excluded as outliers in some analyses, we used all data points in our models because our data is quite heterogeneous and some of the very low and very high values are not necessarily the results of experimental errors; on the contrary, they may be the outcome of new materials and methods (with better or worse performance than the more common alternatives) that should also be part of the analysis; thereby, outliers were also considered in the model training.

Further details for the distribution of the data are provided in the Supporting Information file. For instance, Figure S1, Supporting Information, shows the distribution of categorical data for each output variable. In our dataset, TiO_2 was the most utilized electrode material at 89%, while 60% of the experiments used a Pt counter electrode. Anthocyanin was the most frequently used dye at 42%; most of the experiments (87%) did not use any cosensitization but anthocyanin was the most preferred second dye in 62% of the cases that used cosensitization. The doctor blade was the most preferred coating method at 55%; in 84% of the experiments, pretreatment was not used for the photoanode. Ethanol was the most commonly used solvent for the dye extraction (53%) and additional organic solvent was added to the extraction medium in 51% of the cases. LiI was the most preferred electrolyte salt at 38% whereas the most commonly used electrolyte solvent was acetonitrile at 54%; in 15% of the experiments, ionic liquids (ILs) were used as the electrolyte solvent.

Swarm plots also provide a clear visualization of the distribution of data points, particularly useful for categorical data or when comparing multiple groups. Unlike scatter plots, swarm plots prevent overlapping points by arranging them horizontally based on their corresponding values. Each data point is represented by a marker, revealing the density and distribution of observations within the dataset. When the number of data points is low, and they are scattered along the y-axis, the reliability of swarm plots may be limited. In such cases, the swarm plot should be interpreted with caution. A small number of data points may not provide a representative sample of the underlying distribution, leading to potential misinterpretation of the data. Additionally, when points are scattered along the y-axis, the density and distribution of the data may not be effectively conveyed by the swarm plot, making it difficult to discern patterns or trends. Swarm plots for the descriptors were also generated to better understand the data structure and the relationship between input and output variables. The plots for the variables showing significant trends are presented in **Figure 2**, while the plots for the remaining variables and the distributions of the data points for each variable are given in (Figure S2–S18, Supporting Information).

As mentioned above, TiO_2 is the most commonly used semiconductor for the anode in our dataset; the other semiconductors, such as ZnO and TiO_2 composites are also utilized. As shown in Figure 2, however, none of these alternatives demonstrated better performance than TiO_2 across any of the four target variables (V_{oc}, J_{sc}, FF, and PCE). In this context, the term TiO_2 composites represents a collection of TiO_2 composites that are formed with different materials. Normally, different composites are expected to have different impacts on the outcome; hence, they should be treated separately. However, their numbers were not sufficient to do that; instead, we collected them together so that, at least, we can compare their performance with that of TiO_2 and see if they provide any additional benefit as all these materials were used to enhance the performance of TiO_2. Even though the majority of data points (and therefore their average) for TiO_2

© 2024 The Author(s). Solar RRL published by Wiley-VCH GmbH

2367198x, 2024, 22, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/solr.202400432 by Ukraine - Cochrane, Wiley Online Library on [11/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

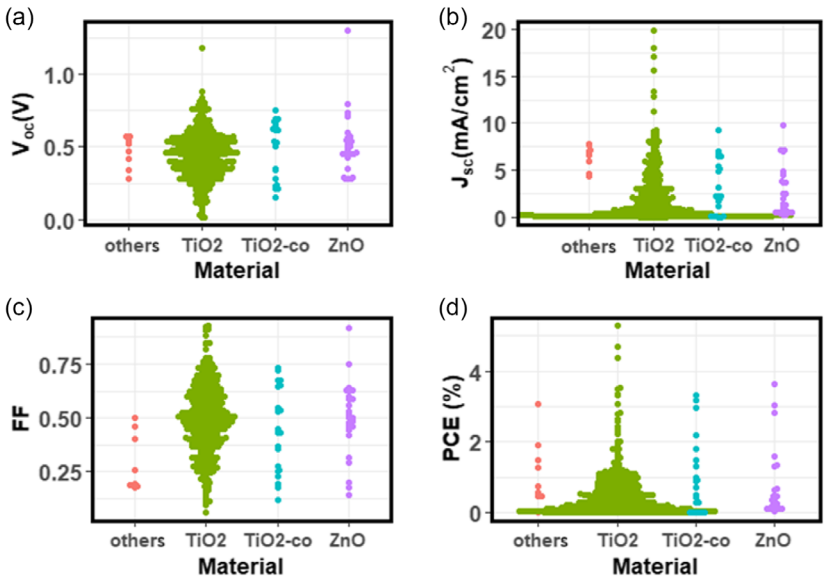

Figure 2. The distribution of anode materials for a) V_{oc}, b) J_{sc}, c) FF, and d) PCE.

are on the lower side of J_{sc} and PCE, the cases that have higher J_{sc} and PCE also outnumber the other alternatives shadowing the positive effects of other alternatives including TiO_2 composites if there are any (Figure 2b,d). Hence, we can state that the data here are not conclusive on the effects of semiconductor type on Jsc and PEC (however, SHAP analysis presented below shows that the semiconductors other than TiO_2 are more favorable for PCE). In contrast, TiO_2 seems to perform better in terms of V_{oc} and FF (Figure 2a,c) as the data were mostly located in relatively higher V_{oc} and FF levels.

To enhance the performance of TiO_2 photoanode, some researchers employed surface treatment using $TiCl_4$ or titanium isopropoxide. This seems to have some positive effects on the performance, especially on FF (**Figure 3**); the data accumulated in higher performance ranges (providing higher average performance) while this trend is more obvious for FF.

The observed improvements in the performance of photoanodes upon $TiCl_4$ treatment can be attributed to several factors. First, the filling of cracks and pinholes in the TiO_2 layer enhances the structural integrity and surface smoothness, which are

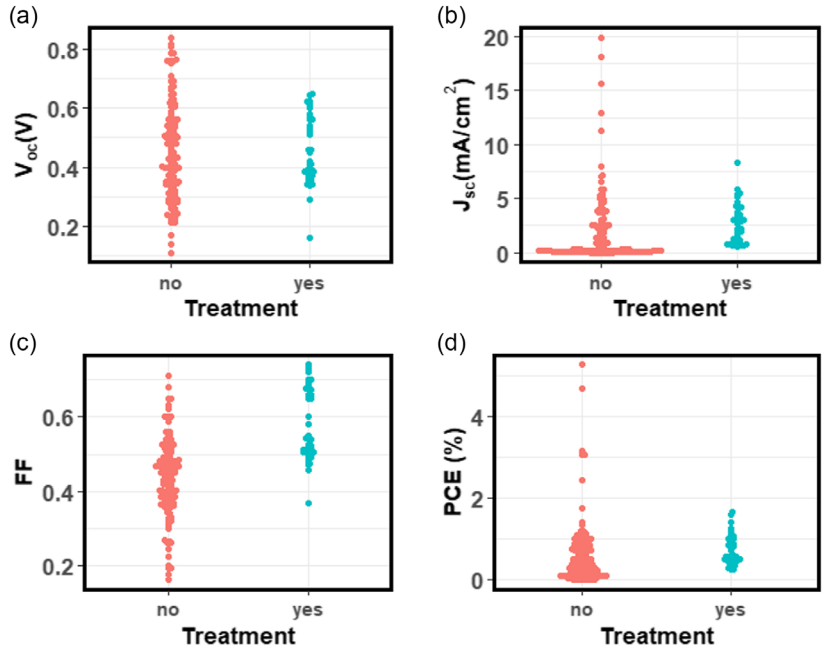

Figure 3. The distribution of using treatment or not for a) V_{oc}, b) J_{sc}, c) FF, and d) PCE.

© 2024 The Author(s). Solar RRL published by Wiley-VCH GmbH

2367198x, 2024, 22, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/solr.202400432 by Ukraine - Cochrane, Wiley Online Library on [11/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

crucial for efficient charge transport and reduced recombination losses. Additionally, the reducing the surface work function of the TiO_2 layer promotes better alignment of energy levels at the interfaces, facilitating more efficient charge extraction and transfer. Better and/or close energy-level alignments allow more forward-electron transport and more generation of current density, due to the low work function that results in less electron-excitement energy required for electron hopping and tunneling at the dye/TiO_2 interfacial contacts.[25] Both of these advantages helped perovskite solar cells utilizing $TiCl_4$-treated TiO_2 surpass the 20% efficiency threshold[26]; apparently, they have similar effects in natural dye-sensitized fuel cells as well.

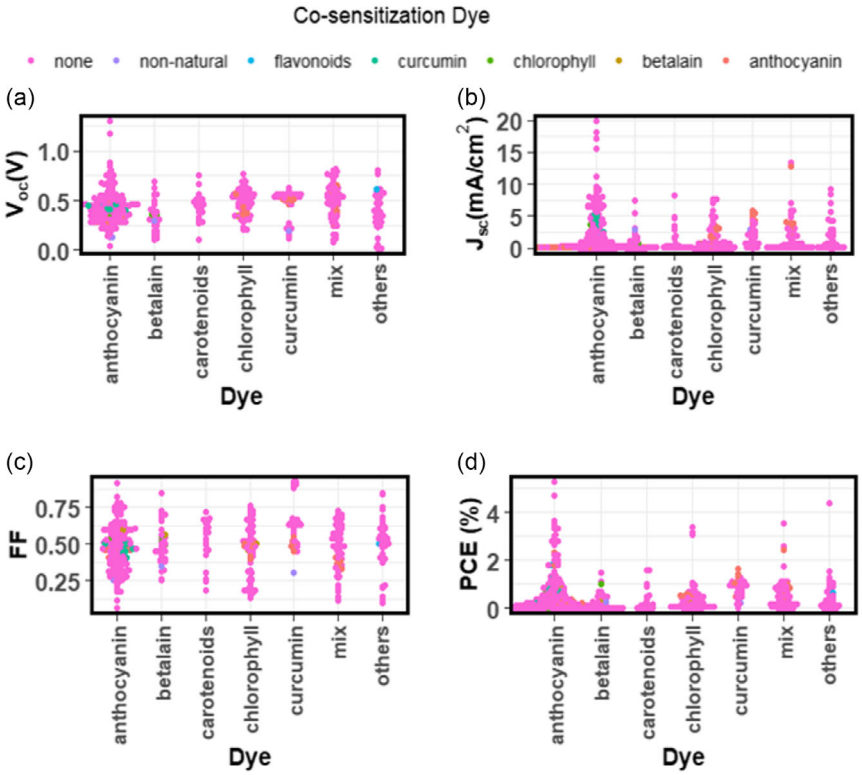

Figure 4. The distribution of dyes for a) V_{oc}, b) J_{sc}, c) FF, and d) PCE (colors represent the cosensitization dye applied on top of the dyes).

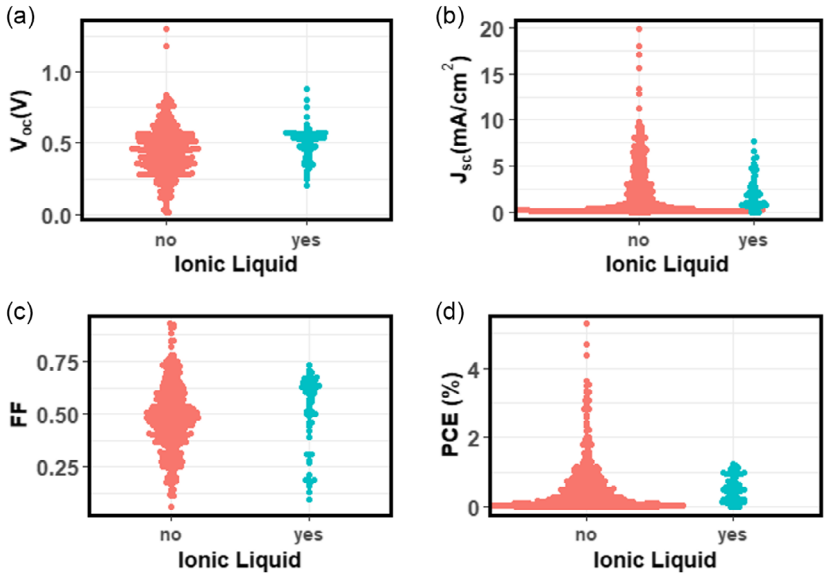

Figure 5. The distribution of data with or without ionic liquids in electrolyte. a) V_{oc}, b) J_{sc}, c) FF, and d) PCE.

© 2024 The Author(s). Solar RRL published by Wiley-VCH GmbH

2367198x, 2024, 22, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/solr.202400432 by Ukraine - Cochrane, Wiley Online Library on [11/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The choice of dye affects the performance of the natural DSSCs as they are the main component that is responsible for light absorption. As shown in **Figure 4**, anthocyanin dyes are the most preferred and best-performing group in our dataset. Anthocyanins, a type of natural pigment found in plants, are known for their ability to absorb visible light due to their conjugated structures. This property makes them promising natural dyes for various applications, including DSSCs (their higher average performance is more apparent in V_{oc} and FF (Figure 4a,c)). Recent advances in anthocyanin dyes extracted from plants have shown promising efficiency in water-based DSSCs, achieving a PCE of 1.49%.[27] In contrast, chlorophyll showed slightly higher open circuit voltage compared to anthocyanin (Figure 4a); the main reason for this is that V_{oc} is influenced by the energy levels of dye, which was determined by the conjugated bonds present in the dye molecule, and anthocyanin-based natural dyes have a lower number of conjugated bonds compared to chlorophyll-based dyes.[28] This means that the energy gap between the valence band and the conduction band of the semiconductor material sensitized with anthocyanin dyes is larger than that of chlorophyll-based alternatives. As a result, the V_{oc} of anthocyanin-based solar cells is slightly lower on average than that of chlorophyll-based solar cells. Various researchers have been also testing the cocktails of dyes to increase the overall performance of natural DSSCs. In our dataset, anthocyanin dye is the most frequently employed cosensitization dye even though Figure 4 shows that the cosensitization does not give better results compared to single dye sensitization (none refer to the cases without cosensitization). The majority of articles used single dye sensitization and the ones that used additional dyes for cosensitization did not seem to have significant improvement on the target variables except for J_{sc} where using the mixture of curcumin and betalain showed increased J_{sc}. In few cases, it was reported the presence of more than one dye in the extract from the plant (without details about the ratios); we separated those from intentional cosensitization and labeled them as mixed dyes.

The use of IL in electrolytes promotes charge transfer within the solar cell, as shown in **Figure 5** showing the distribution of data with and without IL. Even though the effects of ILs in electrolytes on efficiency and current are not clear, V_{oc} and FF are higher in their presence. ILs have been shown to improve the performance of natural DSSCs (DSSCs) by enhancing the adsorption of dye molecules on the surface of the photoelectrode and increasing the stability of the dye-sensitized electrode; they

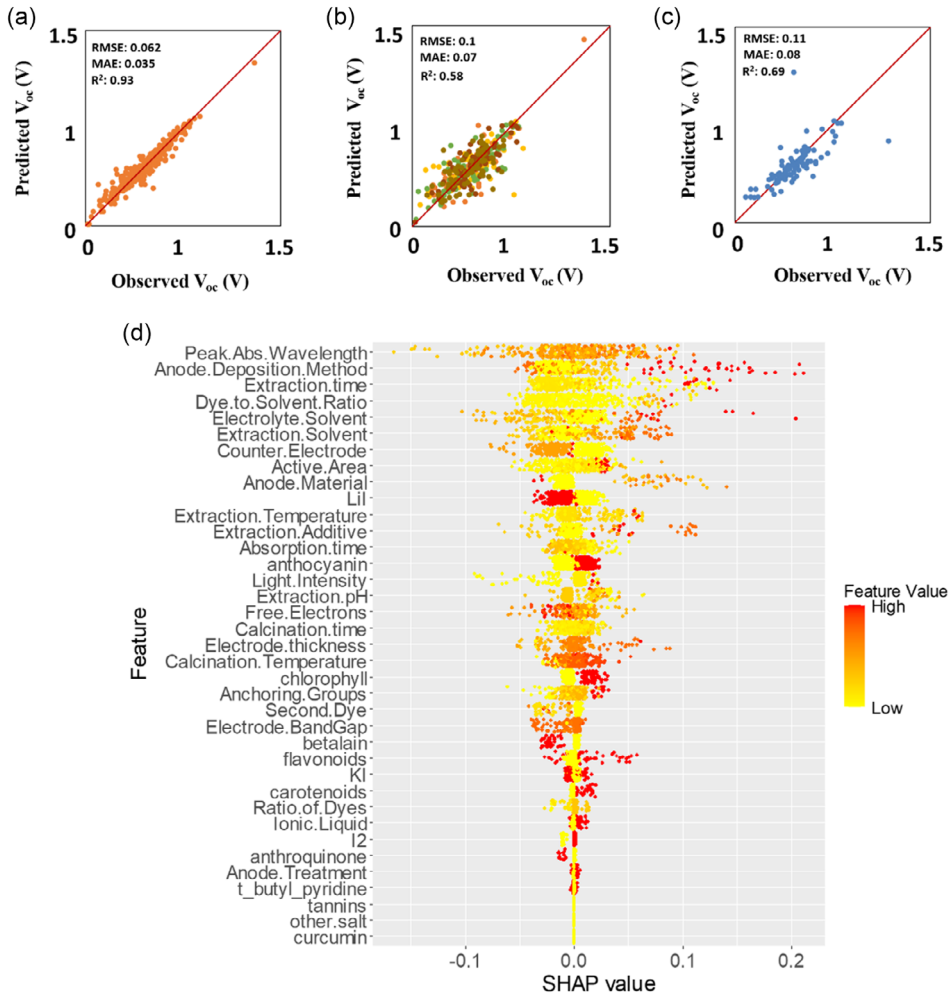

Figure 6. The GB model for V_{oc} prediction: a) training, b) validation, c) testing, and d) SHAP plot for the model.

© 2024 The Author(s). Solar RRL published by Wiley-VCH GmbH

2367198x, 2024, 22, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/solr.202400432 by Ukraine - Cochrane, Wiley Online Library on [11/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

can improve the long-term stability of the cells by reducing the leaching of dye molecules from the photoelectrode and preventing the degradation of the electrolyte.[27]

3.1. Prediction of Open Circuit Voltage

The GB model with encoded structure and the hyperparameters *ntree* of 290, *shrinkage* of 0.15, and *int_depth* of 8 gave the lowest RMSE for validation; therefore, it is selected as the best model for prediction of V_{oc}. The actual versus predicted plots obtained with this model for training, validation, and testing (with the RMSE of 0.062, 0.1, and 0.11 respectively) are given in **Figure 6**; the model predictions, especially for testing data, which was not seen by the model during construction, are reasonably well indicating that the model can be used to predict the V_{oc} values of a new set of design variables.

RF and GB models are often considered black box models due to their complexity, making them less interpretable compared to simpler algorithms like linear regression. Therefore, we used SHAP analysis to improve the explainability of our models and to understand the contribution of individual descriptors to the outcome better. The SHAP plot for V_{oc} prediction is given in Figure 6d. The x-axis represents the SHAP values, which indicate the impact of a feature on the model's prediction. Positive SHAP values indicate that the descriptor has a positive impact on the target variable while negative SHAP values have the opposite. For the continuous variables, the color of the points indicates the value of the feature from low (yellow) to high (red). We used different tones of color for different categories (materials or methods) for categorical variables in the same color scale; we assigned the colors in a way that most commonly used categories in the yellow color so that we could see whether they lead to high or low level of performance and compare them with alternatives. The color code for the categorical variables is given in **Figure 7**, which is also valid for the other SHAP figures.

The results for the peak absolute wavelength at the top of the graph seem to be inconclusive as all the tones of data were distributed in both negative and positive directions (only very low wavelengths appear to be on the negative side). However, the effect of the electrode deposition method is more observable; the articles that did not specify the deposition method, as one of the common deposition techniques used in the field, generally had higher V_{oc}. We labeled the method in these cases as other deposition methods (we shortened as other in figures); considering that the performances in these data are better in general; the deposition procedures likely involve some unique steps or modifications. Spin coating also seems to have a more positive effect compared to screen printing and doctor blades; the use of commercially ready electrodes leads to lower performance.

Dye extraction time, which is the third most important descriptor, may influence the amount of dye deposited, and, in return, higher loading promotes electron-hole separation and may increase V_{oc}; however, excessive dye loading may also lead to aggregation or overlapping of dye molecules, which can hinder effective charge injection into the semiconductor and increase charge recombination rates, thereby diminishing the overall voltage output.[29] In our case, the higher dye/solvent ratio, which indicates higher dye loading, seems to promote higher V_{oc} indicating that the amount of dye on the surface is not too high to show its inverse effect described above. Higher extraction times and dye-to-solvent rations positively influence the V_{oc} while the use of organic solvents (followed by methanol) improves the performance further. V_{oc} is mainly the difference between the Fermi level of the anode and the redox

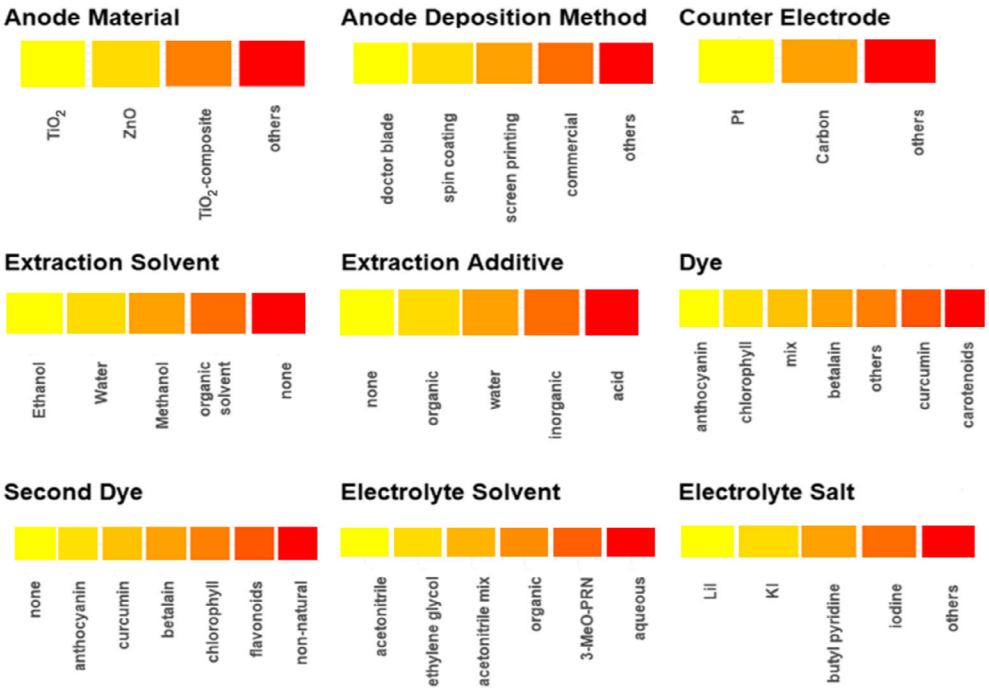

Figure 7. Color code used to label the data for categorical variables in Figure 6d, 8d, and 9d.

© 2024 The Author(s). Solar RRL published by Wiley-VCH GmbH

2367198x, 2024, 22, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/solr.202400432 by Ukraine - Cochrane, Wiley Online Library on [11/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

potential of the electrolyte[30]; the dye sensitization of the anode may affect the HOMO and LUMO level of the dye through the changes of functional groups on the dye with changing extraction conditions (like solvent used and extraction time). The solvent and the duration of dye extraction are also shown to affect the type of functional groups present in the dyes.[31,32]

The choice of solvents used for the electrolyte also appears to be important, and the aqueous electrolytes exhibited better performance. Although they are in the lower ranks in the SHAP values, it is also worth the mention the contribution of some descriptors as their effects are very clear. For example, the use of LiI as the salt in electrolytes seems to have negative effects on V_{oc}; the same is also true for betalain dye while the effects of anthocyanin and chlorophyll are positive. The SHAP values of continuous other descriptors in Figure 6d can be interpreted using scale of continuous variables in Table 1, and color code shown at the right side of SHAP plot; we used a similar color code for categorical variables through **Figure** 6d, **8**d, and **9**d in accordance with the specific color assigned for each category in Figure 7. Categories with the highest number of data points are shaded yellow, indicating greater representation, while categories with fewer data points shift toward red. This gradient visually emphasizes the change in different levels of categories, allowing the effect of each level on the SHAP values to be clearly identified.

3.2. Prediction of Short Circuit Current Density

The predictions for J_{sc} were better with categorically structured data (500 data points) while the RF algorithm, which is well-suited for categorical data, also performed better for J_{sc}; the optimum values of model hyperparameters of *ntree*, and *mtry*, were found to be 250 and 14 respectively. The actual versus predicted J_{sc} for training, validation, and testing are presented in Figure 8 with the RMSE of 0.8, 1.7, and 1.7, respectively. Although the model predictions are not as accurate as for V_{oc}, they are still reasonably well.

SHAP values for the Jsc prediction (Figure 8d) show that the dye/solvent ratio, the active cell area, and calcination temperature are the most influential variables. The bandgap of the anode and the electrolyte solvent were also found to be important for J_{sc} prediction. The strong influence of the dye/solvent ratio is an expected result considering that the generated current heavily depends on the light absorption properties of the dye, and the characteristics of anode material including the bandgap, which influences the recombination rate of charge carriers in a DSSC. In contrast, having a higher active area seems to have negative effects on the photocurrent density. This may be because the distribution of trap states and the recombination rate of charge carriers can be affected by the active area of the cell; a larger active area can deepen the

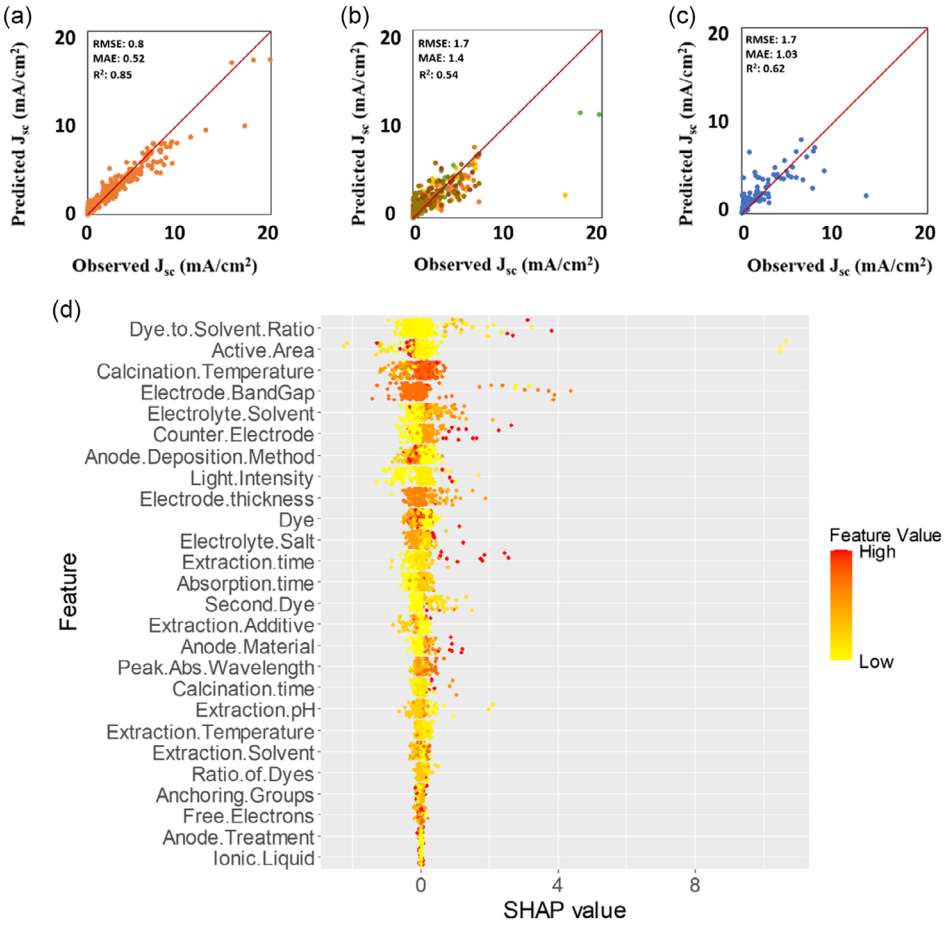

Figure 8. The RF model for J_{sc} prediction: a) training, b) validation, c) testing, and d) SHAP plot for the model.

© 2024 The Author(s). Solar RRL published by Wiley-VCH GmbH

2367198x, 2024, 22, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/solr.202400432 by Ukraine - Cochrane, Wiley Online Library on [11/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

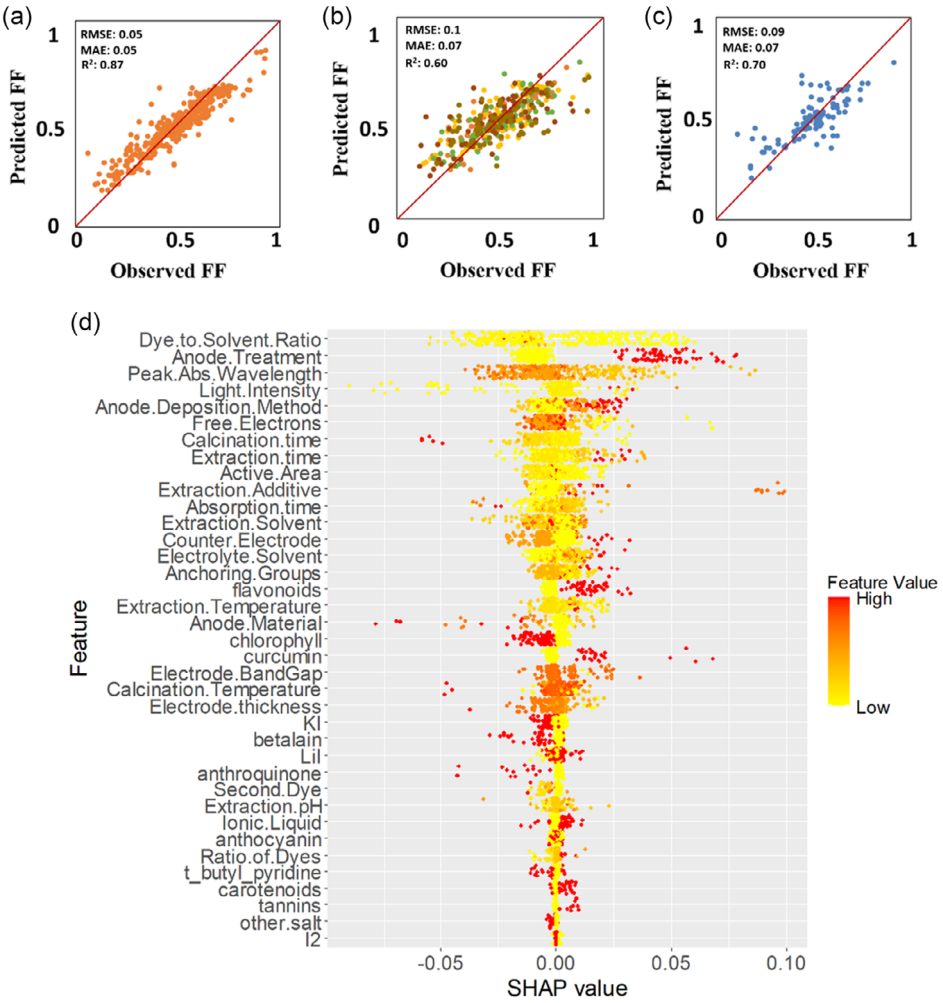

Figure 9. The RF model for FF prediction: a) training, b) validation, c) testing, and d) SHAP plot for the model.

distribution of trap states and increase the recombination rate of electrons, reducing the electron collection efficiency and the photocurrent density.[33] The higher calcination temperatures also slightly lead to the higher photocurrent densities.

Figure 8d also indicates that the photocurrent density is positively affected if the bandgap of the semiconductor is narrow (even though relatively few cases are supported). Normally, a wider bandgap in the semiconductor leads to a larger energy difference between the conduction band and the redox potential of the electrolyte, reducing the likelihood of charge recombination and increasing the open-circuit voltage of the cell.[34] However, a wider bandgap also reduces the absorption of visible light, leading to lower photocurrent generation. Therefore, instead of a single general conclusion for the bandgap effect, optimizing the bandgap of the photoanode and the dye molecule seems to be more important to have high efficiency in DSSCs.

3.3. Prediction of FF

FF was predicted with high accuracy using the encoded data, containing 513 cases, which was used with the RF algorithm; the hyperparameters of the optimum model were 150 and 6 for *ntree* and *mtry*, respectively. As can be seen from Figure 9, the model predictions are quite successful with the RMSE of 0.05, 0.1, and 0.09 for training, validation, and testing sets, respectively.

The results of the SHAP analysis are given in Figure 9d. The FF is directly influenced by the voltage-dependent losses in the solar cell, which include series resistance (R_s) and shunt resistance (R_{sh}).[35] Unlike other output variables, the anode treatment (with $TiCl_4$ or TTIP) also appeared to be important for FF with a positive impact; this treatment has a direct effect on the surface properties, and it may lower the resistance and improve the current generated. The FF is further influenced by the current-dependent losses, particularly related to charge transport and recombination, which also contribute to the total resistances in the DSSC.[36]

It is also noteworthy that the use of tannins, curcumin, and flavonoid dyes improves the FF while chlorophyll, carotenoids, and anthocyanin have negative impacts. Notably, higher calcination times and temperatures also reduce FF. The adverse effects of increased particle size and decreased porosity might hinder the benefits (such as improved crystallinity and reduction in

© 2024 The Author(s). Solar RRL published by Wiley-VCH GmbH

2367198x, 2024, 22, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/solr.202400432 by Ukraine - Cochrane, Wiley Online Library on [11/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

grain boundaries) of high calcination temperatures. Larger grains would decrease the surface area for dye absorption (and therefore light absorption), which would lower the performance of the DSSC.

3.4. Prediction of Efficiency

Finally, efficiency, the key performance measure of a solar cell as it directly quantifies the amount of light converted into electricity, was modeled using 522 data points. Encoded data structure and GB model with *ntree* of 410, *shrinkage* of 0.15, and *interaction depth* of 7 were found to be the best case for the efficiency predictions. The optimum model resulted in the RMSE of 0.2, 0.4, and 0.5 for training, validation, and testing sets, respectively; again, the model predictions are reasonably good as presented in **Figure 10**.

Figure 10d presents the SHAP values for descriptors influencing PCE; the electrolyte solvent seems to be the most influential descriptor because it affects the charge transfer kinetics and voltage loss mechanisms within the DSSC; it is reported that electrolytes with organic solvents often lead to better performance compared to inorganic solvents.[37] In our model, the organic solvents and 3-methoxypropionitrile were found to be the most positively affecting electrolyte solvent type for PCE whereas acetonitrile and ethylene glycol had more negative impact. The following important descriptor was calcination time, of which higher values have a slightly positive impact on PCE while the lower calcination temperatures seem to be more favorable suggesting that this step should be implemented for a longer duration and lower temperatures. Although the dye-to-solvent ratio, which was at the top for J_{sc} and FF and last for V_{oc}, is in lower order here, it is still important, and higher values are more favorable for the PEC as in the case of V_{oc} and J_{sc}.

It is also noticeable that the choice of electrode material and deposition method has a great influence on PCE. Generally, the materials used other than TiO_2 resulted in better PCE (as they are usually investigated to replace TiO_2). The deposition method is also critical since it affects the thickness and surface area of the photoanode layer and the charge transfer kinetics at the photoanode/electrolyte interface. Therefore, optimizing the

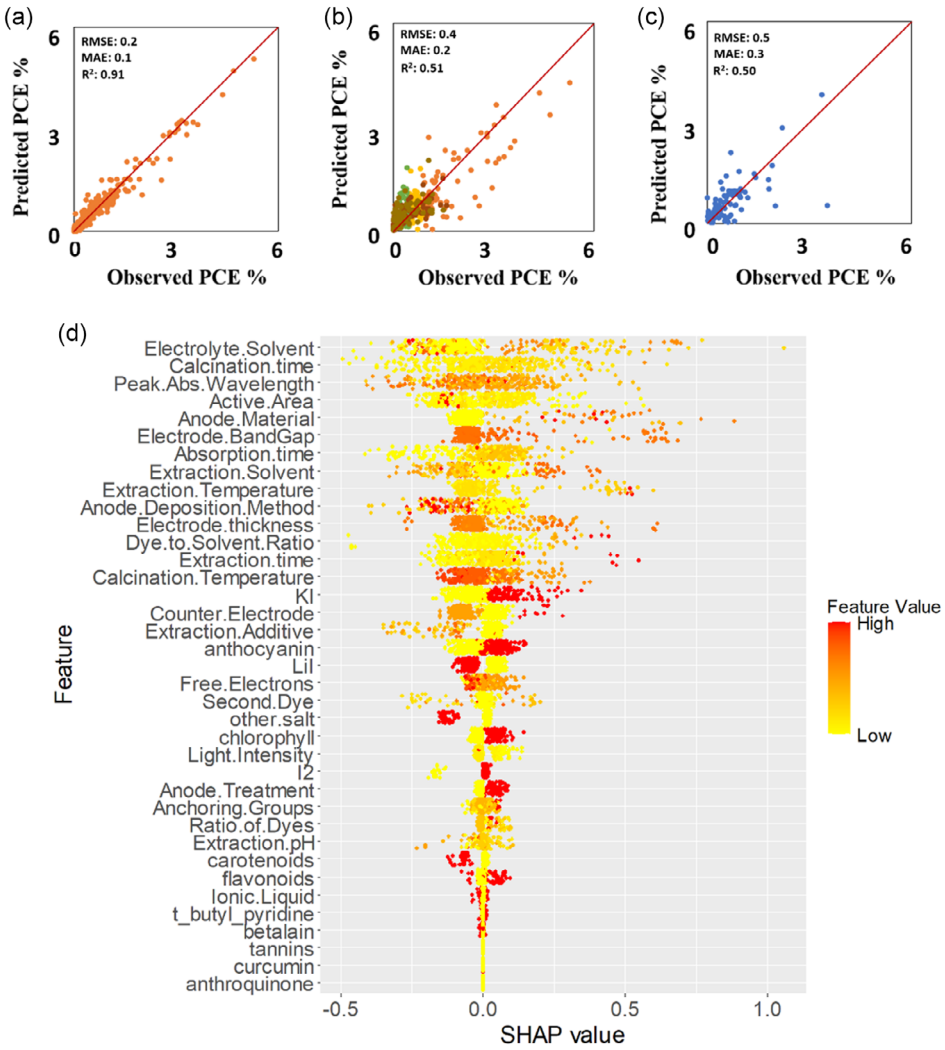

Figure 10. The RF model for PCE prediction: a) training, b) validation, c) testing, and d) SHAP plot for the model.

© 2024 The Author(s). Solar RRL published by Wiley-VCH GmbH

2367198x, 2024, 22, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/solr.202400432 by Ukraine - Cochrane, Wiley Online Library on [11/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

deposition method is crucial for achieving high efficiency in DSSCs.[38] Doctor blade and screen printing seem to have the most positive impact on the PCE prediction, whereas the remaining methods have mostly a negative impact.

As shown in Figure 10d, a higher dye/solvent ratio can lead to a greater amount of dye adsorbed onto the photoanode surface, which can increase the photocurrent density and efficiency. However, excessive dye loading can lead to aggregation or overlapping of dye molecules, which can hinder effective charge injection into the semiconductor and increase charge recombination rates, thereby diminishing the overall voltage output.[39] Finally, the thickness of the photoanode also plays a critical role in determining the efficiency of the DSSC. A thicker photoanode can provide a larger surface area for dye adsorption and increased light absorption, leading to higher photocurrent density and efficiency. However, excessive thickness can lead to increased recombination rates and decreased charge transfer efficiency, resulting in decreased efficiency.[40] In our dataset, there seem to be a few high-thickness cases that have higher PCE.

4. Conclusion

A dataset of 522 data points coming from 113 published articles was created for the prediction of performance measures of natural DSSC. Preanalysis indicated that TiO_2 is the most commonly used anode material, with other materials generally not showing improved performance. Additionally, treating the TiO_2 layer with $TiCl_4$ or TTIP appears to positively influence FF, though it does not significantly affect efficiency or other performance measures. Anthocyanin, used either alone or in combination with other dyes, emerged as the most frequently used dye and generally exhibits higher performance.

Predictive models for performance variables were developed using RF and GB algorithms with categorical or encoded data. This is followed by a SHAP analysis of descriptors for each output variable. 1) The best model for predicting V_{oc} was the GB model with 290 trees, a *shrinkage* parameter of 0.15, and an *interaction depth* of 8. This model performed optimally with encoded data, achieving RMSE values of 0.062, 0.1, and 0.11 for training, validation, and testing datasets, respectively. The most influential descriptor was the peak absorption wavelength of the dye, followed by the deposition method for the anode. The other important variables were related to the dye extraction process (including extraction time, solvent, and dye/solvent ratio). It is noteworthy that these factors significantly influence the amount of dye loaded. 2) The best model for predicting J_{sc} was the RF model with 250 trees and a *mtry* of 14, using categorical data. This model resulted in RMSE values of 0.8, 1.7, and 1.7 for training, validation, and testing datasets, respectively. The most influential descriptor was the dye/solvent ratio used during extraction, as it affects the concentration and purity of the dye. This was followed by the active area and calcination temperature. 3) For predicting FF, the best model was the RF with 150 trees and a *mtry* of 6, using encoded data. It achieved RMSE values of 0.05, 0.1, and 0.09 for training, validation, and testing datasets, respectively. The main influential factors were dye/solvent ratio and anode treatment, where the treatment on the photoanode had a clear positive impact on the prediction of the FF. 4) The GB model for predicting PCE, with 410 trees, 0.15 as *shrinkage*, and an *interaction depth* of 7, performed best with encoded data, yielding RMSE values of 0.2, 0.4, and 0.5 for training, validation, and testing datasets, respectively. The most important variables for PCE prediction were the electrolyte solvent and the calcination time followed by the peak absorption wavelength of the dye.

Finally, we should briefly mention the limitations of the models. First, the data from the literature often brings noise to the model and these can significantly affect the model's performance. Cross-validation and random splitting of training, validation, and test sets help ensure that the models generalize well, but their performance may vary when applied to entirely new experimental conditions that were not previously investigated. Another limitation is that the models developed from data are often valid within the limits of the dataset; hence their predictive power may be limited on the outside of these constraints (i.e., completely different semiconductors, dyes, or conditions). Finally, ML relies on statistical learning; hence, rare occurrences, including newly tested materials and methods that are promising in the long run, may not contribute to the models as the common alternatives that are repeated more frequently.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Burcu Oral: Data curation (lead); Formal analysis (lead); Investigation (lead); Writing—original draft (lead). **Hisham A. Maddah**: Data curation (supporting); Investigation (supporting); Writing—review and editing (supporting). **Ramazan Yildirim**: Conceptualization (equal); Investigation (supporting); Supervision (lead); Writing—review and editing (lead).

Data Availability Statement

The data that support the findings of this study are available in the supplementary material of this article.

Keywords

gradient boosting, machine learning, natural dyes, random forest, solar cells

Received: June 12, 2024
Revised: September 18, 2024
Published online: October 4, 2024

[1] A. H. Alami, S. Alasad, H. Aljaghoub, M. Ayoub, A. Alashkar, A. Mdallal, R. Hasan, *Advances in Science, Technology and Innovation*, Vol. *F663*, Springer, Cham **2023**, p. 115.
[2] B. Yılmaz, Ç. Odabaşı, R. Yıldırım, *Energy Technol.* **2022**, *10*, 2100948.

© 2024 The Author(s). Solar RRL published by Wiley-VCH GmbH

2367198x, 2024, 22, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/solr.202400432 by Ukraine - Cochrane, Wiley Online Library on [11/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

[3] M. A. M. Al-Alwani, A. B. Mohamad, N. A. Ludin, A. A. H. Kadhum, K. Sopian, *Renewable Sustainable Energy Rev.* **2016**, *65*, 183.
[4] G. Richhariya, A. Kumar, P. Tekasakul, B. Gupta, *Renewable Sustainable Energy Rev.* **2017**, *69*, 705.
[5] M. Hosseinnezhad, S. Safapour, *Renewable Dyes and Pigments*, Elsevier, Amsterdam **2023**.
[6] L. Zhang, J. M. Cole, *J. Mater. Chem. A* **2017**, *5*, 19541.
[7] G. Chayal, K. R. Patel, M. S. Roy, M. Kumar, N. Prasad, K. Shitiz, *J. Indian Chem. Soc.* **2019**, *96*, 1059.
[8] V. Venkatraman, R. Raju, S. P. Oikonomopoulos, B. K. Alsberg, *J. Cheminf.* **2018**, *10*, 18.
[9] V. Venkatraman, L. K. Chellappan, *Data* **2020**, *5*, 45.
[10] O. Al-Sabana, S. O. Abdellatif, *Optoelectron. Lett.* **2022**, *18*, 148.
[11] Y. Wen, L. Fu, G. Li, J. Ma, H. Ma, *Sol. RRL* **2020**, *4*, 2000110.
[12] H. A. Maddah, *Opt. Mater.* **2022**, *128*, 112343.
[13] N. Tomar, G. Rani, V. S. Dhaka, P. K. Surolia, *Int. J. Energy Res.* **2022**, *46*, https://doi.org/10.1002/er.7959.
[14] H. A. Maddah, *Mater. Sci. Eng.: B* **2024**, *302*, 117197.
[15] Ç. Odabaşı Özer, R. Yıldırım, *Nano Energy* **2019**, *56*, 770.
[16] J. Im, S. Lee, T. W. Ko, H. W. Kim, Y. K. Hyon, H. Chang, *npj Comput. Mater.* **2019**, *5*, 37.
[17] N. Alwadai, S. U. D. Khan, Z. M. Elqahtani, S. Ud-Din Khan, *Molecules* **2022**, *27*, 5905.
[18] C. Chen, A. Maqsood, T. J. Jacobsson, *J. Alloys Compd.* **2023**, *960*, 170824.
[19] A. Mahmood, J. L. Wang, *Energy Environ. Sci.* **2021**, *14*, 90.
[20] H. A. Maddah, V. Berry, S. K. Behura, *Renewable Sustainable Energy Rev.* **2020**, *121*, 109678.
[21] E. Clarke, S. Sherrill-mix, C. Dawson **2017**, https://doi.org/10.32614/CRAN.package.ggbeeswarm.
[22] L. Breiman, *Mach. Learning* **2001**, *45*, 5.
[23] B. Greenwell, B. Boehmke, J. Cunningham **2022**, https://doi.org/10.32614/CRAN.package.gbm.
[24] E. Štrumbelj, I. Kononenko, *Knowl. Inf. Syst.* **2014**, *41*, 647.
[25] H. A. Maddah, L. Aryadwita, V. Berry, S. K. Behura, *Renewable Sustainable Energy Rev.* **2021**, *151*, 111606.
[26] Y. Xu, C. Gao, S. Tang, J. Zhang, Y. Chen, Y. Zhu, Z. Hu, *J. Alloys Compd.* **2019**, *787*, 1082.
[27] N. Y. Amogne, D. W. Ayele, Y. A. Tsigie, *Mater. Renewable Sustainable Energy* **2020**, *9*, 23.
[28] W. A. Dhafina, M. Z. Daud, H. Salleh, *Optik* **2020**, *207*, 163808.
[29] K. Sharma, V. Sharma, S. S. Sharma, *Nanoscale Res. Lett.* **2018**, *13*, 381.
[30] K. B. Bhojanaa, J. J. Mohammed, M. Manishvarun, A. Pandikumar, *J. Power Sources* **2023**, *558*, 232593.
[31] O. Adedokun, Y. K. Sanusi, A. O. Awodugba, *Optik* **2018**, *174*, 497.
[32] H. Aliah, R. N. Iman, S. Sriwidati, A. Sawitri, A. Setiawan, A. P. D. Putri, F. Kurniawati, *J. Ecol. Eng.* **2023**, *24*, 312.
[33] W. Yan, M. M. Huo, R. Hu, Y. Wang, *RSC Adv.* **2019**, *9*, 1734.
[34] H. Mohammadian-Sarcheshmeh, R. Arazi, M. Mazloum-Ardakani, *Renewable Sustainable Energy Rev.* **2020**, *134*, 110249.
[35] J. Greulich, M. Glatthaar, S. Rein, *Prog. Photovoltaics* **2010**, *18*, 511.
[36] Y. Lv, H. Tong, W. Cai, Z. Zhang, H. Chen, X. Zhou, *J. Alloys Compd.* **2021**, *851*, 156785.
[37] P. Gu, D. Yang, X. Zhu, H. Sun, P. Wangyang, J. Li, H. Tian, *AIP Adv.* **2017**, *7*, 105219.
[38] A. Agrawal, A. Choudhary, *APL Mater.* **2016**, *4*, 053208.
[39] M. Kokkonen, P. Talebi, J. Zhou, S. Asgari, S. A. Soomro, F. Elsehrawy, J. Halme, S. Ahmad, A. Hagfeldt, S. G. Hashmi, *J. Mater. Chem. A* **2021**, *9*, 10527.
[40] K. Magiswaran, M. N. Norizan, N. Mahmed, I. S. Mohamad, S. N. Idris, M. F. M. Sabri, N. Amin, A. V. Sandu, P. Vizureanu, M. Nabiałek, M. A. A. M. Salleh, *Coatings* **2023**, *13*, 20.

© 2024 The Author(s). Solar RRL published by Wiley-VCH GmbH

2367198x, 2024, 22, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/solr.202400432 by Ukraine - Cochrane, Wiley Online Library on [11/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License