MACHINE LEARNING IN SOLAR ENERGY UTILIZATION

by

Burcu Oral

B.S., Chemical Engineering, Boğaziçi University, 2017

M.S., Chemical Engineering, Boğaziçi University, 2019

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Graduate Program in Chemical Engineering

Boğaziçi University

2023

ACKNOWLEDGEMENTS

First, I would like to express my gratitude to my supervisor Prof. Ramazan Yıldırım for his valuable guidance and support in the planning and development of my thesis. He encouraged me to be a better researcher.

Besides my advisor, I would like to thank the rest of my thesis committee: Assoc. Prof. Damla Eroğlu Pala, Prof. Mehmet Erdem Günay, Assoc. Prof. Burak Alakent and Assoc. Prof. Sarp Kaya for their valuable time and interest in my thesis.

I appreciate the friendship and support given by my dear friends, Elif Esvap, İrem Gülçin Zırhlıoğlu, İlke Kaykanat, Özlem Özbek, Destina Ekingen Genç, Ayşegül Kılıç and my laboratory mates Elif Can Özcan, Pınar Özdemir and Beyza Yılmaz.

I would like to thank Alara Timurtaş for her support and belief in every part of my life.

Finally, I would to like to thank my dear family, my mother Bahriye, my father İbrahim and my dear sister Ezgi, for their unconditional love and support. Their guidance and belief in me have a big contribution to my success.

ABSTRACT

MACHINE LEARNING IN SOLAR ENERGY UTILIZATION

In this dissertation, the solar energy utilization processes were studied via machine learning algorithms. First, a bibliometric analysis of the topic was performed; it was observed that photovoltaics and hydrogen-related applications have been studied more extensively while machine learning, mostly neural network and deep learning algorithms, has been mainly used for forecasting and optimization. Desalination, dye-sensitized solar cells and photo(electro)catalysis, which are the most commonly studied solar technologies, were selected for further analysis in this dissertation. The data were collected using Web of Science database search with related keywords while R language was used for the analysis. Tree-based ensemble methods gave higher predictive power, apparently due to the presence of a large number of categorical variables; the decision trees were used to deduce heuristic rules for high performance, especially for the cases in which the predictive models were not successful. Although the band gaps of semiconductors were generally predicted successfully, the accuracy in the prediction of gas production rate and photocurrent density was lower; hence, the classification with a decision tree was employed to identify the range of input variables for high performance. Results in desalination analysis suggested that the inlet water and air temperatures and flow rates have a higher influence on the performance. The dye type was found to be the most important variable for the performance of both synthetic and natural dye solar cells; the type of counter electrode also becomes important for synthetic dyes whereas electrolyte is more influential in natural dye counterparts. In photocatalysis, it was also found that the band gap mostly depended on the preparation method of the semiconductor while the photocurrent density was affected by bias and electrolyte. For CO_2 reduction, the method used for co-catalyst deposition and reaction temperature were important for the gas and liquid phase systems, respectively.

ÖZET

SOLAR ENERJİ KULLANIMINDA MAKİNE ÖĞRENMESİ

Bu tezde, makine öğrenimi algoritmaları aracılığıyla güneş enerjisi kullanım süreçleri incelenmiştir. İlk olarak, konunun bibliyometrik analizi yapılmış, fotovoltaik ve hidrojenle ilgili güneş enerjisi uygulamalarının daha yaygın şekilde incelendiği görülmüştür; makine öğreniminin ise öngörü ve optimizasyon amaçlı olarak özellikle yapay sinir ağları ve derin öğrenme algoritmaları şeklinde uygulandığı görülmüştür. Bu tez için, en yaygın çalışılan güneş enerjisi teknolojileri olan, deniz suyunun arıtımı, boyaya duyarlı güneş pilleri ve foto(elektro)kataliz seçilmiştir. Bu konular için veriler, ilgili anahtar kelimelerle Web of Science veritabanı araştırması kullanılarak toplanmış, analiz dili olarak da R kullanılmıştır. Ağaç tabanlı yöntemler, muhtemelen kategorik değişkenlerin çokluğundan dolayı, yüksek tahmin gücü vermiş, karar ağaçları ise özellikle tahminin zor olduğu durumlarda yüksek performansı sağlayabilecek kuralların çıkarılmasına yardımcı olmuştur. Kullanılan yarıiletkenlerin bant aralığı, oldukça başarılı biçimde tahmin edilebilirken, gaz üretim hızı ve fotoakım yoğunluğu tahmininde hata daha yüksek olmuş, bu durumlarda karar ağacı algoritmasıyla sınıflandırma yapılarak yüksek performansı veren değişkenlerin aralıkları belirlenmiştir. Deniz suyu arıtma sonuçları, giren su ve havanın sıcaklıkları ile akış hızlarının deniz suyu arıtma performansı üzerinde daha büyük bir etkiye sahip olduğunu göstermiştir. Sentetik ve doğal boya duyarlı güneş hücrelerinin analizide, boya tipinin performansı belirleyen en önemli değişken olduğunu ortaya koymuştur. Ayrıca sentetik boyalar için karşı elektrot tipi daha önemliyken, doğal boyalı güneş hücrelerinde elektrolit daha önemli çıkmıştır. Fotokataliz uygulamalarında, yarı iletkenlerin bant aralığının yarı iletkenin hazırlama yöntemine, fotoakım yoğunluğunun ise voltaj ve elektrolite bağlı olduğu bulunmuştur. Ayrıca eş-katalizör yüklemesi ve reaksiyon sıcaklığı sırasıyla gaz ve sıvı fazı CO_2 indirgeme süreçlerinde önemli çıkmıştır.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

b	Bias
i	i^{th} index
j	j^{th} index
J	Current density, mA/cm^2
N	Number of observation
oc	Open circuit
p	Probability
sc	Short circuit
V	Voltage, V
w	Weight
x	Independent variable
X	Precedent
Y	Ancedent
y	Dependent variable
β	Intercept
θ	Coefficient
η	Efficiency

LIST OF ACRONYMS/ABBREVIATIONS

ANN	Artificial Neural Network
ARM	Association Rule Mining
AUC	Area Under Curve
CACW	Closed Air Closed Water
CAOW	Closed Air Open Water
CV	Cross Validation
DFT	Density Functional Theory
DP	Deposition Precipitation
DSSCs	Dye Sensitized Solar Cells
DT	Decision Tree
DW	Distilled Water
FF	Fill Factor
FN	False Negative
FP	False Positive
GA	Genetic Algorithm
GB	Gradient Boosting
GOR	Gained Output Ratio
HDH	Humidification-Dehumidification
KNN	K-nearest Neighbor
LCA	Life Cycle Analysis
LHS	Left Hand Side
LR	Linear Regression
MAE	Mean Absolute Error
ML	Machine Learning
NHE	Normal Hydrogen Electrode
OACW	Open Air Closed Water
OAOW	Open Air Open Water
PCE	Power Conversion Efficiency

PD	Photocurrent Density
PECWS	Photoelectrochemical Water Splitting
PEC	Photoelectrochemical
QDs	Quantum Dots
ReLU	Rectified Linear Unit
RF	Random Forest
RHS	Right Hand Side
RMSE	Root Mean Squared Error
ROC	Receiver Operator Curve
RR	Recovery Ratio
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
TTIP	Titanium Tetra Isopropoxide
WOS	Web of Science

1. INTRODUCTION

Solar energy has emerged as a promising sustainable alternative to conventional energy sources in recent years. As the demand for clean and renewable energy continues to rise, researchers and engineers are constantly seeking ways to maximize the efficiency and effectiveness of solar energy systems. Solar energy technologies harness the energy of the sun to provide heat, light, hot water, electricity, and even cooling for houses, business centers, and industrial facilities. There are a variety of technologies that use one of two main ways of utilizing solar energy: physical and chemical. Physical methods involve converting solar radiation directly into electricity or heat using devices such as solar cells, solar thermal collectors, and solar cookers. Chemical methods, on the other hand, involve using solar energy to drive chemical reactions that produce useful substances such as fuels and valuable chemicals. Solar thermal technologies can be used for heating water or air in residential and commercial buildings, and concentrated solar power uses mirrors or lenses to concentrate the sunlight in a large area onto a much smaller area to increase temperature sufficiently to generate steam or vapours of low boiling point chemicals to generate electricity; photocatalysis mimics photosynthesis and uses energy from the sun to govern reactions. These technologies provide a sustainable way to produce energy, have significant potential to reduce greenhouse gas emissions and play a crucial role in the transition towards a more sustainable and greener chemical industry.

One of the potential uses of solar energy in future is desalination, which is a process of removing salt and other minerals from seawater or brackish water to produce freshwater. Solar desalination can be used to provide clean and sustainable water for drinking, irrigation, industry, and other purposes in areas where conventional desalination methods are not feasible or affordable. There are two basic methods of achieving desalination using solar energy: direct and indirect [1]. Direct solar desalination uses a solar collector to heat the seawater and then evaporate it into steam. The steam is then condensed into freshwater on another surface, such as a metal plate or a fabric

membrane. This method is also known as thermal evaporation or solar distillation [2]. Direct solar desalination has been used for thousands of years by ancient civilizations such as the Greeks and Persians [3]. However, it has some limitations, such as low efficiency, high cost, and dependence on weather conditions [4]. Indirect solar desalination, on the other hand, uses a device that converts sunlight directly into electricity or heat using semiconductors or other materials. The electricity or heat is then used to power a membrane process that separates the freshwater from the saltwater. This method is also known as photovoltaic or concentrated solar power desalination [5]. Indirect solar desalination has some advantages over direct solar desalination, such as higher efficiency, lower cost, and greater scalability, however, it also faces some challenges, such as land use, environmental impact, and grid integration [6].

Solar desalination is one of the most promising technologies for addressing the global water crisis. It can provide fresh water in remote areas where conventional sources are scarce or expensive. It can also reduce greenhouse gas emissions by replacing fossil fuels in the energy sector. However, it also requires further research and development to overcome its technical and economic barriers. Some of the current research topics include improving the performance and durability of solar collectors and membranes; reducing the cost and complexity of system design and operation; integrating solar desalination with other renewable energy sources; and evaluating the social and environmental impacts of solar desalination projects [1]. Dye-sensitized solar cells (DSSCs) can be shown as an example of electricity production from solar irradiation; they are a type of thin-film solar cells that use organic or inorganic dyes to absorb light and generate electricity. They are one of the oldest and most widely studied solar cell technologies, as they have many advantages such as low cost, simple fabrication, flexibility, transparency, and environmental friendliness. However, they also face some challenges such as low efficiency, stability, scalability, and durability. Recent research has focused on finding new materials or designs that can overcome these challenges and achieve higher performance under various conditions. The highest reported efficiency is 15% under direct sunlight with a co-sensitized TiO_2 photoanode [7]. Although metal halide perovskites, as more efficient solar sensitizers, have been studied more exten-

sively [8,9], the studies with synthetic and natural dyes also continue due to being more mature and suitable for different market applications [10] as well as having low-cost and simple fabrication processes compared to other solar cells [11].

Solar photocatalysis and solar-to-fuel applications are two promising fields in solar energy utilization through chemical reactions. Solar photocatalysis leverages solar energy to drive chemical reactions; the interaction of solar irradiation with a semiconductor creates photoexcited electrons that can be used in redox reactions. Solar photocatalysis has been extensively applied in environmental remediation, such as wastewater treatment, due to its ability to degrade various pollutants [12]. Solar-to-fuel conversion is another way to use photocatalysis. Solar-to-fuel technologies aim to convert solar energy into chemical fuels, providing a sustainable solution to energy storage and transportation [13]. These technologies harness solar energy to produce fuels from basic chemical feedstocks like water (H_2O) and carbon dioxide (CO_2). Examples of solar fuels include hydrogen, ammonia, and hydrazine. The production of these fuels involves various processes such as photochemical, photobiological, electrochemical, and thermochemical reactions. These fuels can be stored for later use, making them an attractive alternative to fossil fuels and batteries.

Machine learning (ML) is a branch of artificial intelligence that helps researchers analyze large datasets and extract trends and patterns hidden in the data. ML algorithms can be used in classification, regression, association, clustering, and dimensionality reduction problems. In classification, the goal is to predict discrete categories or labels based on input data. On the other hand, regression aims to predict a continuous value. Both classification and regression are types of supervised learning, where the model learns from a dataset that includes both the input data and the correct output. The model's performance is then evaluated based on its ability to accurately predict the unseen data. This ability to generalize from learned data to predict new data is what makes machine learning so powerful and widely applicable across numerous fields including solar energy technologies.

The most common algorithm for association detection is association rule mining (ARM) which helps to identify the relations between input and output variables by considering their restricted conditional probability distributions. Linear regression (LR) is the simplest regression algorithm that models the relationship between a dependent variable and one or more independent variables using a linear function. Support vector machines (SVM), artificial neural networks (ANN), Decision tree (DT) and tree-based ensemble algorithms such as random forest (RF) and gradient boosting (GB) algorithms can be used for both classification and regression problems. SVM algorithm aims to find the best hyperplane to divide the data. ANN uses a network of connections and simple transformations to learn the non-linear data. DT develops selection rules and heuristics for desired output variables by creating several if/then statements using input variables. In RF, on the other hand, a large number of trees are created, and the output variable is predicted by majority voting (if classification) or by taking the average (if regression). In gradient boosting trees, the learning is sequential, as opposed to parallel tree building in RF, and the weight of variables is changed according to the previous tree in every tree creation. Each of these algorithms has its strengths and weaknesses, and they all have numerous settings and configurations that can be tuned based on the specifics of the dataset and the problem. The choice of algorithm depends on the size, quality, and nature of the data set, and the insights that one wants to get from the data.

Machine learning can be used to analyze all solar technologies mentioned above. For example, by integrating machine learning algorithms into solar-powered desalination systems, researchers have made significant strides in optimizing efficiency [14] and reducing costs [15]. These algorithms allow for real-time monitoring and control, adjustment of operational parameters, and prediction of the energy output based on environmental factors. Similarly, machine learning algorithms have enabled researchers to predict the optical and electronic properties of various dyes, allowing for the discovery or design of more efficient and cost-effective materials. By training algorithms on a large dataset of known dyes and their properties, researchers can create models that can accurately predict the performance of new candidate dyes. This approach significantly

accelerates the discovery of highly efficient dyes for DSSCs, reducing the time and cost associated with extensive experimental synthesis and testing [16]. Another area where machine learning has shown promise is in optimizing the parameters of DSSC devices. By considering a multitude of factors, such as the thickness of the dye-sensitized layer, the composition of the electrolyte, and the configuration of the electrodes, machine learning algorithms can identify the optimal combination of parameters for maximum efficiency. These algorithms utilize advanced optimization techniques to explore the multidimensional parameter space, searching for the most favourable configurations. By iteratively adjusting the parameters and evaluating the performance of each configuration, the machine learning algorithms can converge on the most efficient design, leading to higher power conversion efficiencies in DSSCs [17].

Machine learning-integrated photocatalysis is also a promising field that aims to accelerate the discovery of efficient photocatalysts [18]. The input variables for machine learning models in photocatalysis depend on the specific application and the type of data available. In general, the input variables can include structural and electronic properties of the photocatalyst, such as bandgap energy, surface area, crystal structure, and doping concentration. Other variables such as reaction conditions, light intensity, and temperature can also be included in the model [19]. Machine learning algorithms can also be trained on large datasets of materials with known morphologies to predict the morphology of new materials. This approach has been shown to be effective in accelerating the discovery of new materials for photocatalysis [20].

In this dissertation, a text mining analysis of the topic was performed first to observe the general research trends in the area; then, five solar energy utilization subjects were analyzed using machine learning tools. The selected subjects summarized above were popular developing technologies for a sustainable and clean future; although there was a tremendous amount of experimental work conducted in those areas, the utilization of machine learning in understanding the process and performance was limited. Most machine learning algorithms were focused on forecasting solar power or new material research as explained in each chapter. Our work aimed to infer knowledge

from already published data and guide researchers based on statistical analysis. The use of literature data to analyze the performance of a solar energy system is relatively rare in literature; most of the published works contain the density functional theory (DFT) based analysis for new material search. In the analysis of each subject, general ML workflow was adopted; first, the data was collected, then it was cleaned and pre-processed into machine-readable format after that data analysis was performed to understand the database, this was followed by model building and model selection, in which the datasets were divided into training and testing subsets, and k-fold cross validation was performed in the model selection.

This dissertation consists of nine chapters. The second chapter gives brief information about ML workflow and algorithms. Text mining analysis of the solar energy utilization and ML in solar energy utilization was done to observe common themes and trend topics in the last years and the results were given in chapter three. The fourth chapter includes the analysis of the humidification-dehumidification desalination system by machine learning tools. Performance analysis of dye-sensitized solar cells and natural dye-sensitized solar cells were given in the fifth and sixth chapters, respectively. Chapter Seven includes machine learning analysis of photoelectrochemical water splitting. Analysis of photocatalytic CO_2 reduction over metal oxide semiconductors was given in chapter eight. Finally, the conclusion and recommendations for future work were given in chapter nine.

2. MACHINE LEARNING

We are in the data age; a large amount of data is generated every second. To understand and utilize such big data, there is a tremendous effort in the development of computational tools. In recent decades, machine learning has become a trending topic in the analysis of big data. By utilising machine learning, it is possible to detect patterns, make generalizations, deduct rules and predict the outcome of a new condition [21]. Machine learning concept was developed in 1950 but it became trending in recent decades due to increases in computational power [22].

The quality, quantity and volume of data are very important for machine learning applications. Data consists of features (descriptors, input variables), which can be discrete (categorical) or continuous and may have a target (output variable). When there is no target in the data, the approach is called unsupervised learning. The main aim of unsupervised learning is to detect patterns in data and find similar data points. Clustering algorithms are mostly applied in unsupervised learning. By examining the similarity in data, it is possible to group instances and understand trends. On the other hand, when there is a defined target, the approach is called supervised learning. The aim is to develop a model that predicts the defined target by learning from a given database. There are two tasks in supervised learning: classification and regression. In classification, the aim is to predict the labelled group/class of data. The classes can be binary such as yes-no outcome or multivariable such as the type of the crystal structure of a material. Classification is usually done for categorical target types, but it is also possible to discretize continuous variables and predict the range as a classification task. Class imbalance can be an important problem in this method [23]; sometimes having a large amount in one class compared to others might lead to poor prediction performance for minority classes because the imbalance creates a bias toward the most populated class. There are several ways to overcome class imbalance problems; the amount of data in classes is balanced so that the model may have high predictive power for all classes. One way of doing this is obtaining the same number of classes from the dataset

by over- or under-sampling. For instance, for a dataset that has 700 instances of A and 300 instances of B class, either random 300 data of class A can be selected (under-sampling) or by over-sampling 400 data can be added by randomly selecting class B data (with replacement). Another way of dealing with class imbalance is synthetic data generation, some algorithms can generate new data for a class based on other instances of that class [24]. Regression, on the other hand, requires a continuous output. The simplest and widely known regression algorithm is linear regression. The aim is to obtain coefficients for features that will be used in the calculation of the target value. Those coefficients give information about the importance of that feature on the output.

2.1. Model Development

After the raw data is collected, it should be processed into machine-readable format. Some algorithms cannot handle missing values; hence imputation of those values is required. There are several ways of imputing missing values: by mean, mode, median or even it is possible to develop a model for missing columns as the target and predict missing values. In addition to imputing missing values, input variables should also be pre-processed if necessary, this process is called feature engineering.

There are lots of variables and parameters that can be collected for the model development. Having reliable and powerful models requires excellent representation of data by features. According to the output variable, the effective variables should be identified correctly in order to obtain effective models. In the database, there might be variables that do not influence the output hence, complicating the model and resulting in low predictive powers. Therefore, it is important to select the right features. There are various ways to do feature selection algorithms such as backward (shrinking) or forward (growing) feature selection [25]. In forward selection, the variables are added, and the models are developed whereas in backward elimination variables are deleted and the optimum feature set that gives the minimum error is found.

In small datasets, the number of features should be limited since as the number of features increases the model will be prone to overfitting [26]. When the input variables are sparse, it is necessary to reduce the number of features without losing significant information. Two techniques that analyse the correlations of descriptors without model development are the Pearson correlation map [27] and maximal relevance minimal redundancy [28]. One example of the techniques that involve model development during feature selection is the genetic algorithm (GA), which is used to find a set of key features based on the root mean square error (RMSE) values calculated from the developed model [29]. Principle component analysis is another method to reduce dimensionality i.e. number of variables in the model. It creates projections of original variables and combines them linearly to generate new variables with minimal information loss [25].

2.2. Algorithms

2.2.1. Linear Regression

Linear regression is the simplest regression algorithm. It models the relationship between input and output variables by fitting a linear equation that fits best the data. The equation is

$$estimation(y) = \beta_0 + \beta_1 x + \epsilon, \tag{2.1}$$

where y is the dependent variable, x is the independent variable, β_0 is the y-intercept, β_1 is the slope, and ϵ is the error term. The aim is to find β_0 and β_1 values using least squares estimation.

There are some assumptions such as linearity, independence, homoscedasticity, and normality for a model to be valid and reliable. Linear regression can also be used when multiple independent variables are used to predict a dependent variable. The performance of the model can be improved by several methods such as transforming or standardizing the data, adding interaction or polynomial terms or even employing regularization. Lasso and Ridge Regressions approaches involve regularization terms

to penalize coefficients of the linear regression model. The main difference between them is Lasso penalizes proportional to the absolute value of the coefficients, whereas Ridge uses the square of coefficients [30]. Though they can improve the accuracy of the model, they may create problems like multicollinearity and overfitting [31].

2.2.2. Support Vector Machines

Support vector machines is another algorithm that is popular in ML applications. It can be used for both classification and regression approaches and the main goal is to find a hyperplane that separates the data. The hyperplane is chosen such that it maximizes the margin between the separated classes, which is the distance between the hyperplane and the nearest data points from each class [32]. SVMs can be used for both linear and nonlinear classification tasks. In the case of nonlinear classification, SVMs use a technique called the kernel trick to transform the data into a higher-dimensional space where it can be separated by a hyperplane [33].

One of the advantages of SVMs is that they are effective in high-dimensional spaces, which makes them suitable for tasks such as image classification and text classification [34]. However, SVMs have some limitations. For example, they can be sensitive to the choice of kernel function and the regularization parameter, and they can be computationally expensive to train on large datasets [35].

2.2.3. Tree Based Algorithms

Tree-based methods are simple and more interpretable compared to other algorithms. Decision Tree is based on learning simple decision rules inferred from the variables. The rules are simple if-else conditions that lead to a predicted outcome (either regression or classification prediction). The models developed by decision tree algorithms are easy to read and visualize, therefore they are white-box algorithms. In a DT model, there is a top node with all of the instances and then based on rules the data is divided into further nodes according to the decision criteria (Figure 2.1).

Some of the parameters that can change decision tree model are the depth of tree, complexity parameter and number of instances in final node.

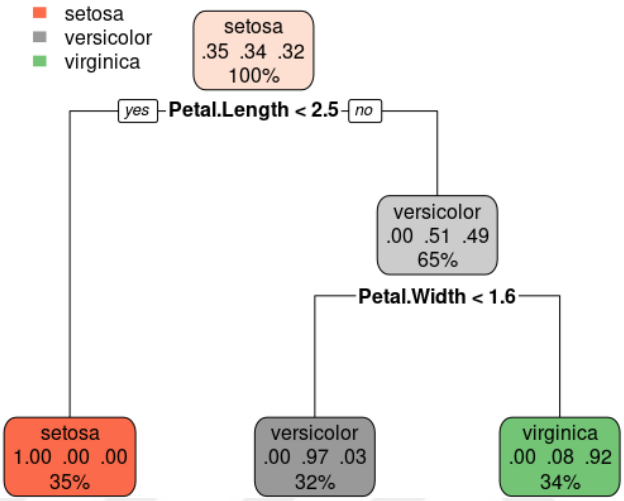

Figure 2.1. Example of a Decision Tree Model.

DT algorithm aims to decide the best set of rules that describes the data correctly. Decision trees are prone to overfitting; therefore, careful use of parameters is important. When the complexity of the tree increases, the model tends to overfit the noise and lose predictive power for unseen data. *Complexity parameter* is responsible for the complexity of the decision tree and it eliminates unnecessary branches that do not increase the accuracy; this process is called pruning [23]. In classification tasks, the optimum rules are decided according to the reduction of entropy or Gini index, whereas in regression tasks, the rule that gives the minimum residual sum of squares is selected as the decision criteria. Entropy measures the impurity or randomness of a model; it has a minimum value of 0 showing the total purity [36].Formula for Entropy is

$$Entropy = \sum -p \log p, \tag{2.2}$$

Gini index is an alternative metric to entropy which measures the probability of a variable being wrongly classified in the model. Gini index can have values between 0 and 1, Gini equal 0 denotes that all elements belong to a certain class, and 1 denotes that the elements are randomly distributed across various classes; hence, a low Gini

index is preferred in the selection of decision criteria [37]. Gini is calculated as

$$Gini = 1 - \sum p^2, \tag{2.3}$$

where p denotes the probability. Due to some disadvantages of decision trees such as being non-robust and prone to overfitting, ensemble methods have emerged [36]. The ensemble approaches aim to combine multiple trees and extract information from those trees. The construction of multiple trees can be in parallel or series. Random Forest and Gradient Boosted trees are the two most commonly used ensemble methods for parallel and series tree constructions respectively.

Random forest models are a robust and black box algorithm working based on the majority voting concept [38]. Some of the parameters are *number of trees*, the minimum number of samples required to split a node (*misnplit*) etc. The name random comes from the random selection of variables to be used in split criteria. In random forest model development, bagging is used. Bagging is a procedure where a portion of the dataset is drawn, with replacement, from the database. After bagging, a tree is grown on the selected dataset using random features. This method is reported to increase the accuracy of random forest models [38]. This way lots of copies of the dataset are formed and lots of trees are created to increase the prediction performance of a single tree [36].

Gradient boosted trees are using boosting approach instead of bagging. This time the production of trees is sequential, the information from one tree is passed to another. There are three main parameters in boosted trees; number of trees, shrinkage and number of splits in each tree [39]. *Number of trees* parameter decides how many trees will be generated in the algorithm, since a very high number of trees may result in overfitting in GB. Learning rate, *shrinkage*, is an important parameter that is used in regularization. Low shrinkage values often result in better models at the expense of computation time. Number of splits or *interaction depth* controls the complexity of the GB, setting this parameter to one makes trees which have a single split. Due to the nature of boosting, unlike random forests, smaller trees are enough in GB models [39].

2.2.4. Artificial Neural Networks and Deep Learning

ANN is a popular non-linear fitting algorithm with easy training, adaptive structure and tuneable parameters [40]. Inspired by neurons in human brains, artificial neural networks use a combination of nodes to predict an output variable. As seen from Figure 2.2, it consists of minimum of three parts: an input layer, a hidden layer and an output layer. Neurons are connected to the neurons in the following layer and these connections represent a weight and by activation functions a prediction can be generated. Below, the prediction formula for ANN algorithm is

$$NET = b + \sum_{i=1}^{n} x_i w_{ij}, \tag{2.4}$$

w_{ij} represents weight of a connection, x_i represents an input variable and b represents a bias.

In the training of an ANN model, the selection of a proper activation function and optimization of these weights are important. Rectified Linear Unit (ReLU) and Sigmoid activation functions are the most preferred activation functions [41]. The main principle of ReLU is to map negative outputs to zero, and this property creates a non-linearity; hence ANN could learn complex patterns. ReLU is often preferred due to its efficiency and performance in training deep neural networks. However, it can cause "dead neurons" that only output zero, which is a limitation to consider during the network design process. The optimization of the weights is mostly done by the backpropagation algorithm. It optimizes the weight of a connection by iteratively analysing errors generated on the following layer and regulating weight in the previous layer accordingly [42].

Aside from conventional ANN methods, new algorithms based on ANN concept have emerged. General regression neural networks [43], extreme machine learning [44] and deep neural networks [40] are some examples.

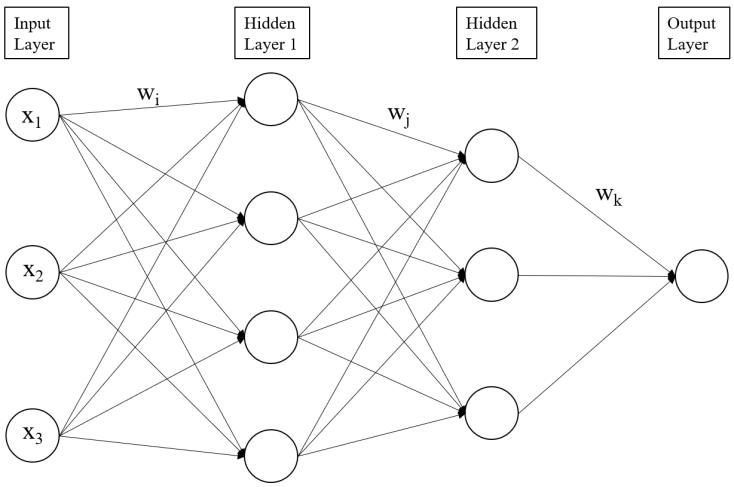

Figure 2.2. Example of an ANN network with two hidden layers.

Deep learning is composed of multiple processing layers to learn the representation of data with multiple levels of complexity [40]. It uses the backpropagation algorithm to determine the parameters that are used to compute the output in each layer by using the error information from the preceding layer [42]. These parameters, often called weights, are numbers that define the input-output function of the model. In a typical deep-learning system, there may be millions of these adjustable weights and millions of labelled examples with which to train the machine. To properly adjust the weights, the algorithm computes a gradient vector that calculates by what amount the error would change if the weight were changed slightly. The weight vector is then adjusted in the opposite direction to the gradient vector.

2.2.5. Association Rule Mining

Association rule mining is a statistical analysis tool for unsupervised learning applications. It reveals the frequently occurring if-then relations between two (for instance input and output) variables within a large database. This if-then relation is also called antecedent (X) and consequent (Y) relation or simply rule. All the variables should be in categorical form for the ARM algorithm. There are three metrics to define a rule in ARM algorithm; support, confidence and lift. *Support* $(freq(X,Y)/N)$ indicates how often that antecedent and consequent relation is found in the dataset, $confidence$

($freq(X,Y)/freq(X)$) represents how often the given relation is true and $lift$ value ($Confidence/freq(Y)$) indicates the dependence of antecedent and consequent; the values greater than one show a strong association.

2.3. Performance Evaluation

2.3.1. Train Test Split

The goal of machine learning is to accurately predict the unseen, new data by the rules obtained from available data. Since it is not possible to evaluate the performance of the models in the prediction of unseen data, a test set should be extracted from the dataset. This set will only be used in the final model evaluation and give insights about the possible prediction power of the developed model, hence, the performance in the prediction of the test set will be an objective measure of the strength of the model. As a rule of thumb, influenced by the Pareto principle, generally, data is split into 80% train, 20% test set [45], however, there is a lot of research trying to find the optimum value for train-test split [46, 47].

2.3.2. Model Verification

Complex models may result in overfitting, i.e. prediction of train set may be very good, but it may fail to predict unseen test data correctly because the model is fitted to the noise in the train set, hence, losing its generalization power. Therefore, in order to obtain reliable models, the verification of models is important and cross-validation is a tool that is commonly employed for model verification. As Cai et. al. summarized, there are 6 different cross-validations: Average cross-validation on multiple hold-out estimates, leave p-out cross-validation, repeated learning test cross-validation, Monte Carlo cross-validation, k-fold cross-validation and bootstrap cross-validation [22]. The simplest and most widely used cross-validation technique is k-fold cross-validation [48] since it is computationally less expensive compared to other methods. In k-fold cross validation train set is divided into k folds and the model is developed for k-1 folds and

tested with the remaining fold. After each model development, the folds are shifted so that every fold is utilized as test set. The average error can be found for the model after each fold is tested. Most employed k values are 4,5 and 10 according to database size.

2.3.3. Model Evaluation Metrics

Selection of model parameters and evaluation of the performance of the proposed model must be based on objective metrics. These metrics are different for classification and regression tasks. Confusion matrix gives insights about the performance for classification models. It is a table of the distribution of the predicted and real classes; a simple example is given in Table 2.1.

Table 2.1. Typical Confusion matrix for classification algorithms.

		Predicted class	
		A	B
True class	A	*True Positive* (TP)	*False Negative* (FN)
	B	*False Positive* (FP)	*True Negative* (TN)

In classification algorithms, total accuracy, precision, class accuracy (recall) and area under receiver operating characteristic (ROC) curve (AUC) are some of the metrics that can be used in model comparison. AUC is a popular measure of the accuracy of a diagnostic test, higher AUC values indicate better test performance. AUC changes from 0.5 (no performance) to 1.0 (perfect performance). Total accuracy is calculated using the correctly classifed samples

$$Total\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{2.5}$$

Recall of a class is calculated using the samples in that class

$$Recall_A = \frac{TP}{TP + FN}. \tag{2.6}$$

Precision of a class is calculated using the samples that were classified in that class

$$Precision_A = \frac{TP}{TP + FP}. \tag{2.7}$$

All of these three metrics shows some predictive power of the developed model.

In regression, the error associated with the model is reported by mean absolute error (MAE) and/or root mean squared error (RMSE). The equation for MAE is

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|x_i - \hat{x}|, \tag{2.8}$$

and for RMSE is calculated by

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x})^2}, \tag{2.9}$$

respetively. Coefficient of determination (R^2) may also be used in reporting model evaluation but it should be noted that it is not a reliable metric to be the only decision metric since it can be biased [49]

$$R^2 = \mathbf{1} - \frac{\sum(x_i - \hat{x})^2}{\sum(x_i - \bar{x})^2}, \tag{2.10}$$

where x_i is real value, $\hat{x}$ is predicted value, $\bar{x}$ is mean value of x.

Even though test error is the ultimate criteria in the determination of model success, validation and training errors should also be reported. Here, the validation error represents the error obtained from cross-validation results for the selected optimum model, the train error is the error calculated for the optimum model performance in the prediction of the train set and the test error is the performance of the optimum model on the unseen data.

3. BIBLIOMETRIC ANALYSIS OF SOLAR ENERGY UTILIZATION

3.1. Background Information on Solar Energy and Bibliometric Analysis

Increasing concerns for the depletion of fossil fuels, as well as the air pollution and global warming caused by the use of fossil fuels, have driven researchers to develop alternative energy technologies, aiming to reduce our reliance on fossil fuels for the high energy demands of modern life. Although various technologies that do not rely on fossil fuels and are sustainable in principle, they have their disadvantages and challenges. For example, hydroelectric power, a form of renewable energy, generates electricity from flowing water; however, the construction of dams may have negative impacts on environmental aspects. Similarly, geothermal energy harnesses heat from Earth's core through ground resources, but its use is limited to specific geological areas that can eventually be depleted while it may also harm the environment around the geothermal wells. Biomass energy utilizes organic materials as an energy source and is easily adaptable to the current infrastructure, but its efficiencies remain low and it may compete with food production. Wind energy converts wind movement into electricity, but inconsistent wind patterns and the need for regular maintenance pose challenges for widespread adoption. Nuclear energy provides substantial power, but it operates at high costs and has significant safety concerns. The direct use of solar energy, which is practically free, environmentally friendly and abundant, stands out as one of the most promising renewable sources among the new and renewable alternative energy forms like hydroelectric energy, biofuels and wind (which are also the other forms of solar energy anyway).

Various technologies utilize solar energy, in the form of light and heat, through chemical or physical processes. The heat from solar energy serves various applications including cooking food, heating water and houses, desalination of seawater, and generating steam for electricity using concentrated solar power. Solar irradiation (light)

is also used extensively in photovoltaics and photocatalytic/photoelectrocatalytic systems, which rely on the excitation of electrons from a suitable material by photons from the sun, and by mimicking the photosynthesis, utilize them in electricity generation or reactions; the photovoltaic technologies are newly developed with significant challenges while the photocatalytic/photoelectrocatalytic systems, especially recently popularized water splitting and CO_2 reduction processes are still at infancy stage, the these technologies are constantly evolving and improving to enhance efficiency, affordability, and accessibility.

Accumulation of data on solar energy, coupled with increased computational power, has led to the emergence of data-driven applications in recent decades. Machine learning, a branch of artificial intelligence, is a trending tool for extracting knowledge from literature and analyzing the impact of parameters on desired outcomes using both experimental and computational data; the knowledge is used to improve the effectiveness of future studies in that field. Indeed, ML has been extensively employed in forecasting, performance analysis, and the search for new materials in various solar energy technologies.

The literature can be also examined using bibliometric analysis and text mining, which intersect at the crossroads of information retrieval, natural language processing and scientific research, to get valuable insights into prevailing research trends and identify potential research gaps [50]. While bibliometric analysis quantitatively assesses scholarly publications, authors, journals, and research trends; text mining involves extracting valuable insights from unstructured textual data, such as research papers, news articles, and social media posts. By applying techniques like information extraction, sentiment analysis, and topic modelling, text mining transforms raw text into structured information.

Bibliometric analysis and text mining have been used in many research areas in recent years. There are also works involving solar energy technologies. In one of these works, Beard et.al. created a database for organic and perovskite solar cells using text

mining tools and recorded photovoltaic device parameters [51]. Life cycle assessment (LCA) of solar photovoltaics was studied in another work; by analyzing relevant literature, the authors identified emerging trends in LCA on solar panels, which primarily involve environmental impacts, resource utilization, and end-of-life management [52]. David et. al. conducted a bibliometric analysis for photovoltaic solar energy management between 2000 and 2018 using the Scopus database. They found ten future trends in the solar photovoltaic field; forecasting techniques, demand management, acceptance of the technology, module research for better performance, load status modeling, energy management, recycling and waste, battery integration, circular economy and cost reduction [53]. There are several other works on specific bibliometric analysis of solar energy technology such as concentrated solar panels [54], solar thermal energy storage [55], photovoltaics [56, 57] and perovskite solar cells [58, 59]. In another work, researchers aimed to analyze solar energy research trends from 1992 to 2011 with keywords like "solar cell," "solar energy," "solar power," "solar radiation," and "solar thermal," and found that thin-film solar cells, organic photovoltaics and solar thermal technologies were emerging topics and international collaboration was increasing [60]. Paulo and Porto analyzed solar energy technologies and open innovation literature between 2000 and 2014 [61].

Aside from these references, there hasn't been any attempt to investigate recent trends in solar energy and a study on bibliometric analysis of the use of ML in solar energy has not been found. Ajibade et. al. conducted a bibliometric study on ML applications in renewable energy in the last decade and emphasized that machine learning plays a crucial role in predicting, operating, and optimizing renewable energy technologies and assessment of the social and economic aspects related to these technologies are lacking in the literature [62]. In this work, a bibliometric analysis of solar energy utilization and machine learning in solar energy utilization was applied in order to gain insights about recent research trends and possible research gaps. This study includes solar-to-fuel approaches such as photocatalytic and photoelectrochemical systems as well as machine learning applications compared to previously published articles on bibliometric analysis of solar energy utilization.

3.2. Computational Details

Web of Science database was searched with *("solar") AND ("energy" OR "utilization" OR "systems" OR "to fuel" OR "fuel")* keyword structure and the most relevant 10 000 (out of 216,761) articles was used for the analysis. Keywords for the machine learning analysis in solar energy were in the structure of *("solar") AND ("energy" OR "utilization" OR "systems" OR "to fuel" OR "fuel") AND ("machine learning" OR "data-driven")* and 2759 articles were found and used for the analysis. Bibliometrix package [63] of the R language was used for the analysis of the trends in the solar energy utilization literature. The data was pre-processed to remove synonym words and plural forms. The list of removed words and synonyms is given in Appendix A. The minimum frequency of keywords is selected as five in trend analysis, the clustering method for co-occurrence analysis was the "walktrap algorithm" and "association" was used as normalization (default values of biblioshiny function).

3.3. Results and Discussion

Initially, all publications related to 'solar energy utilization' from 1957 to 2022 were considered; however, the data from only the last 20 years were used in most of the analysis, because the number of publications increased significantly after 2005 (as depicted in Figure 3.3). The time span was even shorter for machine learning applications in solar energy utilization as ML has gained great attention in recent years due to advancements in computational power and data management capabilities. As the ML in the solar energy domain noticeably increased after 2013, we used the data from the last 10 years for the ML-related trends. There are various publishers and journals that focus on articles related to solar energy utilization (Figure 3.3). Among them, *Renewable Energy*, *Energy Conversion Management*, *Solar Energy*, *Energy*, *International Journal of Hydrogen Energy* have the high number of solar energy related publications. Energies journal seems to be the most preferred source of machine learning applications in solar energy even though the journals that published a large number of solar energy papers (as some mentioned above) also have a considerable

number of ML-related papers as well. Next, the author productivity over time was analyzed for the most relevant authors in the field; it is evident that İbrahim Dinçer and JJ Wang have been active in this area for an extended period, and they continue to contribute (Figure 3.3a).

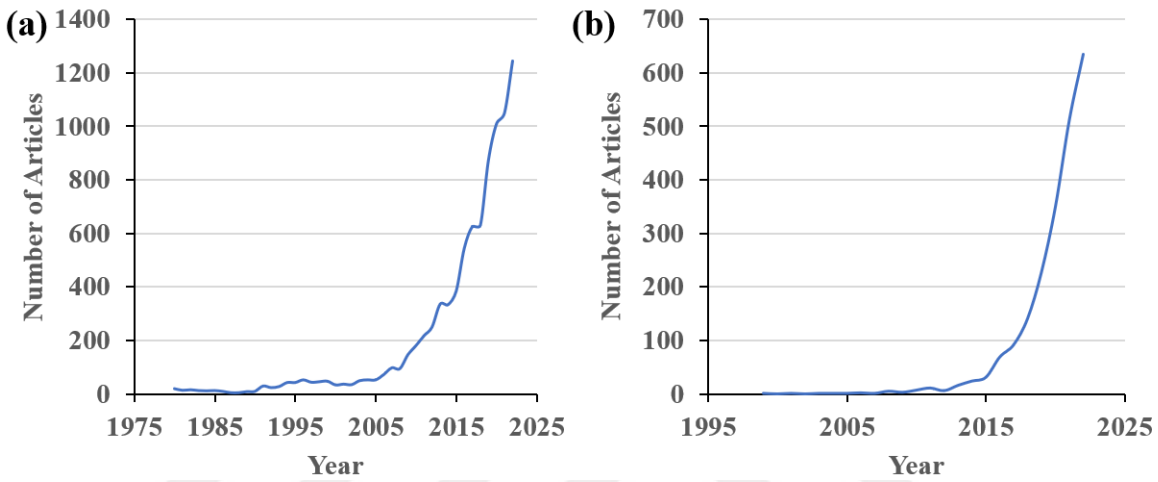

Figure 3.1. The annual publication numbers for (a) solar energy utilization and (b) machine learning in solar energy.

China is the primary contributor to the total number of published articles on solar energy utilization (Figure 3.3a and 3.3c); followed by USA, India, Iran, and Turkey. In addition, single-country publications have been more prevalent than international collaborations over the past 20 years. However, the landscape shifts when we delve into machine learning applications in solar energy utilization (Figure 3.3b and 3.3d). In the ML domain, the USA takes the lead, while China, followed by India, is in second place. Countries like Spain, Australia, and Korea also contribute to ML research in solar energy utilization. Figure 3.3a shows the change of publication frequencies for leading countries through the years; while the USA initiated research earlier than most other countries and published more solar energy-related papers, China has surpassed USA, significantly in the last 5 years with a remarkable increase in the number of publications. India, Iran and Turkey are the other countries that have a significant number of publications in the field. Similar trends are also observed in ML publications in the solar energy area with an about 10-year delay in time; the total number of

publications increased only in recent years and Chinese publications are about to exceed the number of USA publications in upcoming years. Spain and Australia also appeared in the list as significant contributors to the ML papers on solar energy.

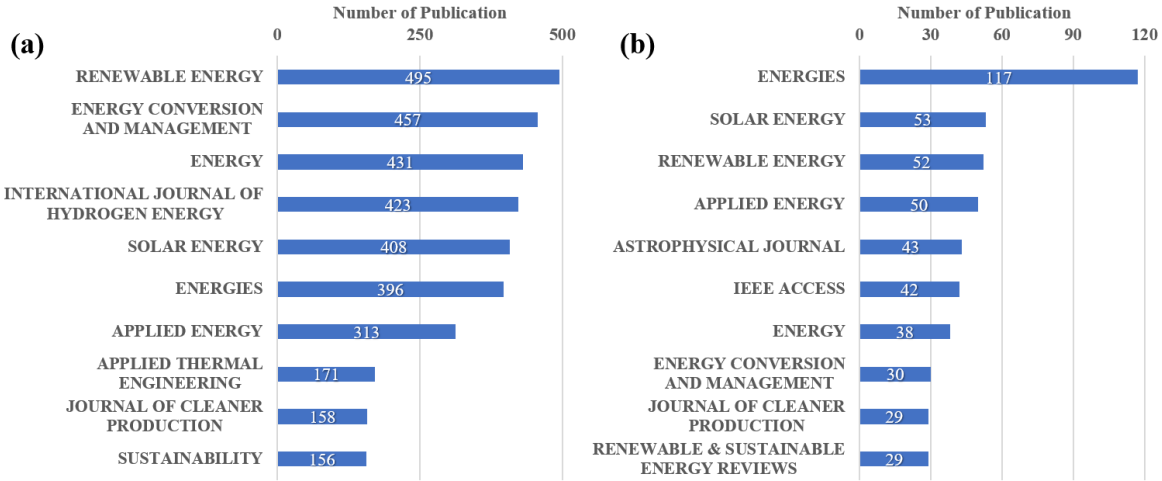

Figure 3.2. Most relevant journals in (a) solar energy and (b) machine learning in solar energy.

The most frequently appeared keywords were also presented using *word cloud*, which is an effective visualization method for textual data points (Figure 3.3). The keyword *photovoltaics* is the most reported keyword in the solar energy papers as expected considering the remarkable increase in the attention to photovoltaic electricity both in academic and industrial perspectives. This is followed by *hydrogen*, which is considered as an important energy carrier for the future and its production via electrolysis as well as photocatalytic and photoelectrochemical processes (water splitting also appeared as another keyword) have been studied extensively. CO_2 reduction, which is another popular solar application (coupled with hydrogen production) appeared less frequently; this may be attributed to the fact that this field gained its popularity more recently). Solar collectors, desalination and solar cells are some of the other solar energy technologies which are reported more frequently compared to others. Performance and optimization-related studies seem to be the majority type of research conducted followed by economic analysis and exergy.

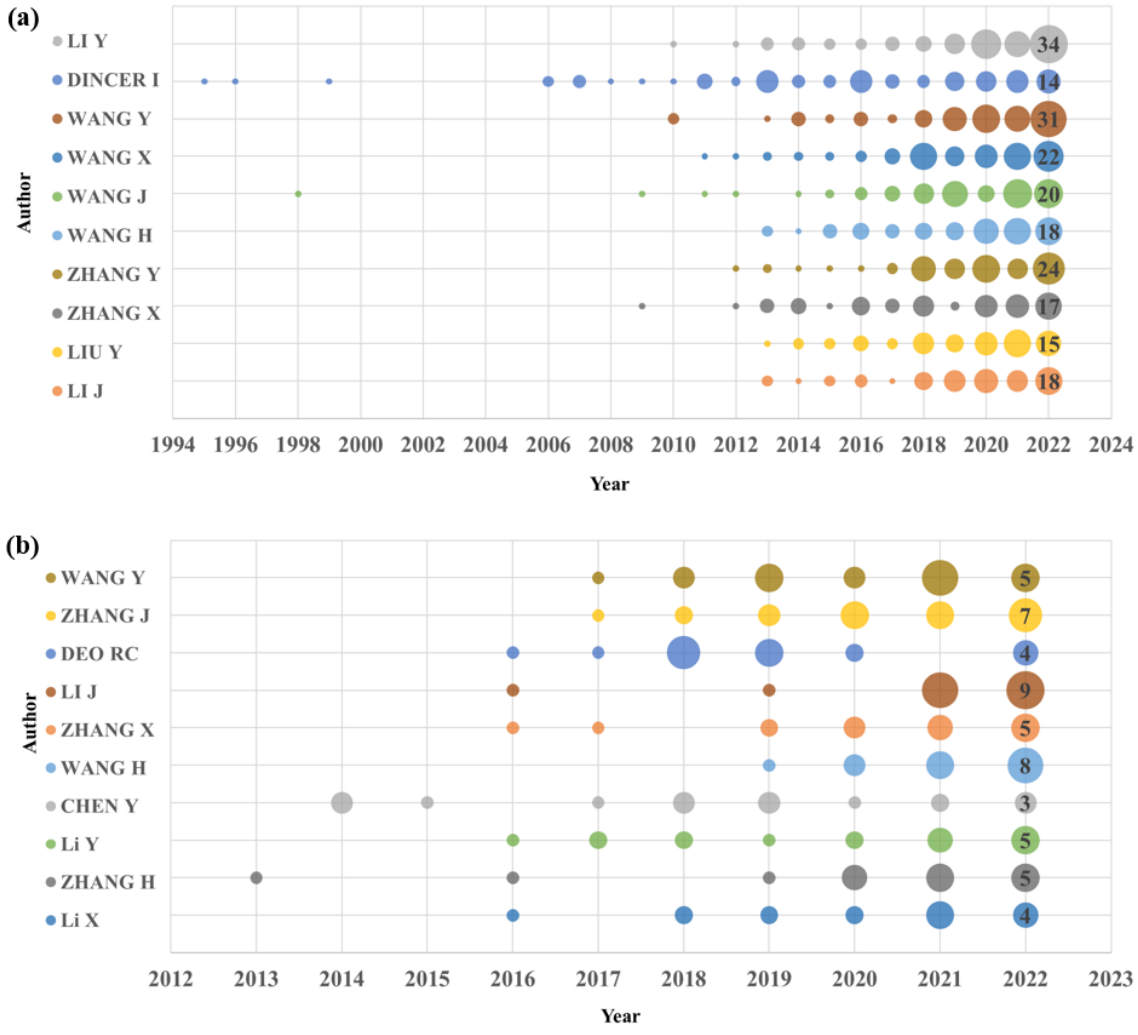

Figure 3.3. Most relevant authors for (a) solar energy and (b) machine learning in solar energy (bubble sizes show the total publication number).

Similarly, *photovoltaic* is also one of the most frequently reoccurring keywords in ML applications of solar energy utilization, while ML algorithms such as neural networks, deep learning, support vector machines and random forest are the other most popular keywords related to ML applications. It is interesting that forecasting, followed by optimization, seems to be the most common task performed by ML; apparently, demand/supply forecasting (especially for solar electricity) as well as forecasting for the future of solar technologies, most of which are in the development stage, are being studied.

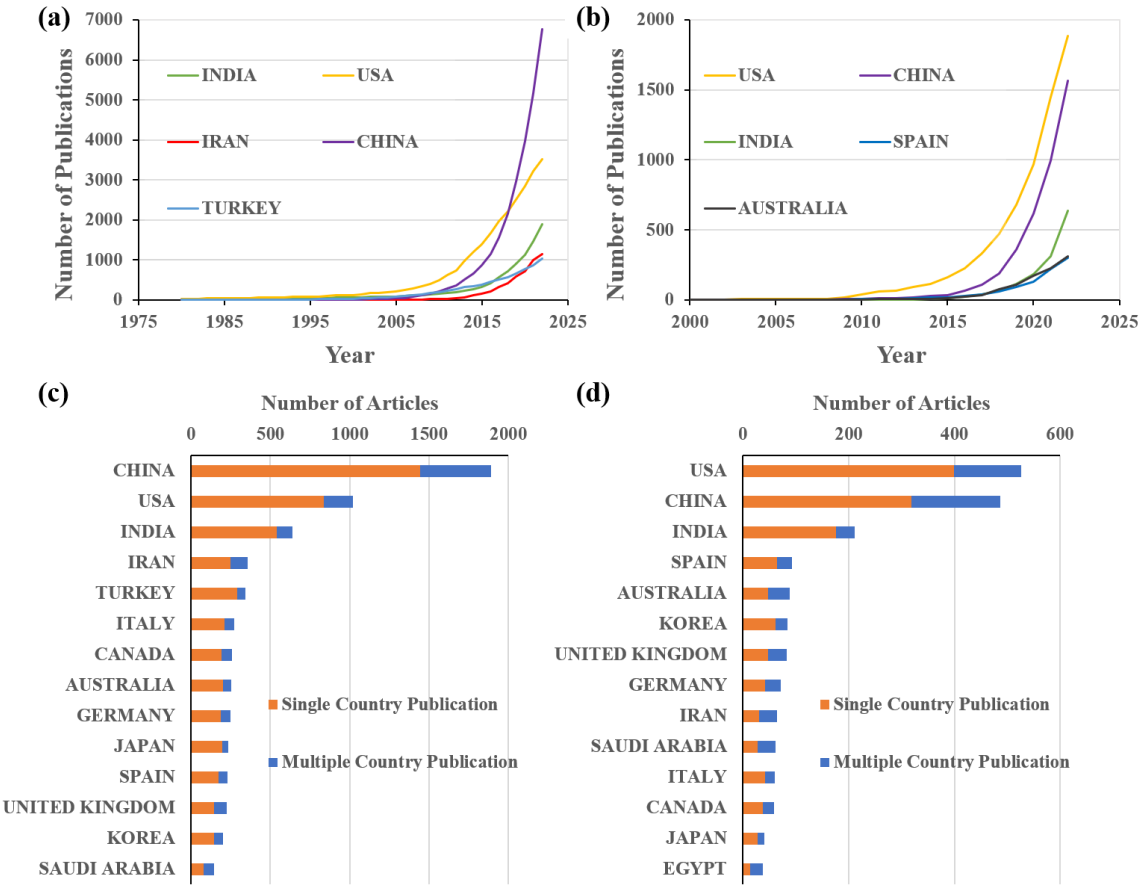

Figure 3.4. Countries publications over time for (a) solar energy (b) ML in solar energy and corresponding authors country for (c) solar energy (d) ML in solar energy

.

The co-appearance of the keywords was also analyzed using *the co-occurrence network*, which shows the terms used together in keywords or abstracts to see the connections among them (Figure 3.6). The size of the nodes shows the frequency of the keyword, and the thickness of the lines shows the strength of the connection of two keywords (i.e. it represents how many times they appeared together). For instance, performance, optimization and design are the most frequently used keywords together. This is an expected result considering that, as we also mentioned above most of the solar energy technologies are still under the development phase. Consequently, most of the research focused on improving the design and performance of these technologies; for example, even in the relatively well-established field of photovoltaics, various new technologies (like halide perovskite and organic solar cells) have been investigated

extensively to replace silicon-based cells, which is the current dominant design in the field. There is another, but smaller, cluster in the network that apparently relates the solar hydrogen-related keywords. *Hydrogen* appears to be the dominant keyword in the second cluster, which is expected since it is a widely studied product of photocatalytic and photoelectrochemical systems. *Carbon dioxide* and *water* are strongly connected to hydrogen since the splitting of water and reduction of CO_2 are investigated for hydrogen generation [64,65]. The use of nanoparticles for photoconversion reactions are also widely studied since they offer enhanced performance; for instance, utilization of copper and gold nanoparticles increased the CO_2 photoconversion rate [66] and optimized nanoparticle size was reported to increase hydrogen yield in CO_2 photoreduction [67].

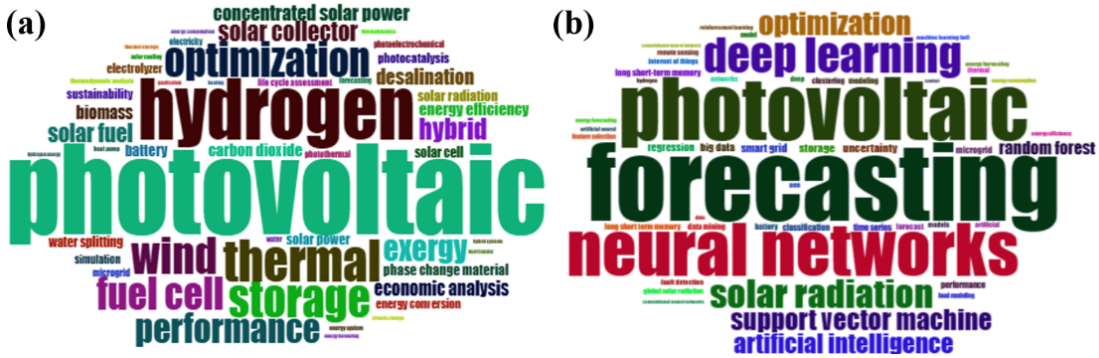

Figure 3.5. Word cloud for keywords of (a) solar energy utilization (b) ML in solar energy utilization.

Similar to the solar energy research presented in Figure 3.6, ML-related publications also lead to two groups; one being about forecasting and the other about optimization (Figure 3.7). The dominating cluster relating forecasting, photovoltaics, ANN and deep learning seems to be mostly related to forecasting works on photovoltaics-related fields (including solar irradiation) while the most commonly employed ML technique is ANN and deep learning, which is a special application of ANN. Optimization cluster, on the other hand, mainly consists of ML in solar energy storage applications such as batteries or load modelling in grid applications and reinforcement learning seems to be more relevant to this cluster compared to other keywords. Load modelling and control

of the integration of solar energy to the grid via ML seems to be the popular research area and reinforcement learning was used for microgrid energy management [68–70].

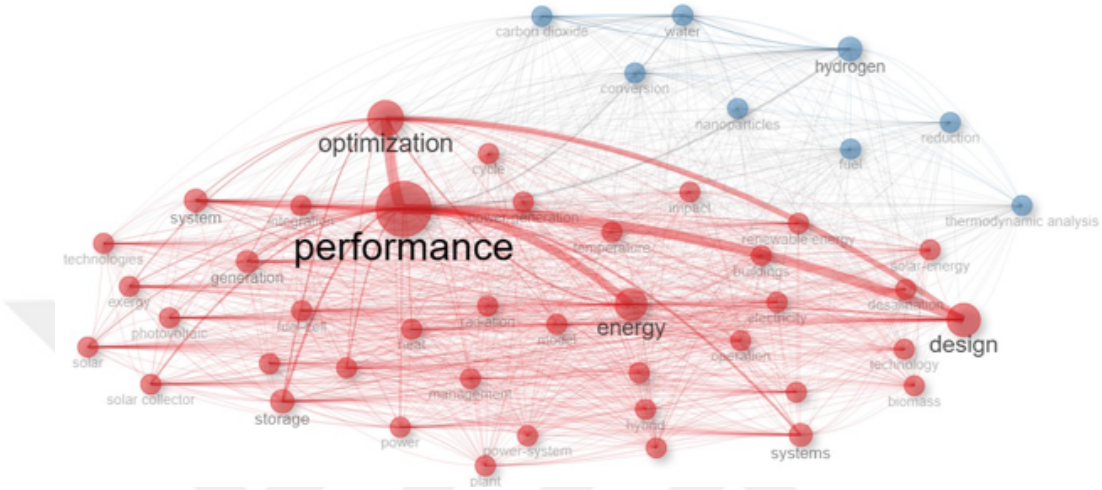

Figure 3.6. Co-occurrence network for keywords of solar energy.

Finally, a trend analysis was performed to see the change in the popularity of major keywords. The bubble sizes show the frequency of the keyword at its peak year and the lines start at the first appearance of that keyword. Technologies such as water-splitting, desalination, photovoltaic and hydrogen were in peak use around 2019 but their frequency decreased even though they are still being researched significantly; apparently, the research has shifted to system efficiency, machine learning and photothermal systems in recent years. *Green hydrogen* keyword has gained popularity in recent years [71,72], it represents that the energy for the production of hydrogen is from clean resources and solar is the major clean energy source for this application. *Multi-objective* optimization is another word that seems to be trending around 2021 possibly representing the works focusing on the integration of solar energy applications to electricity/smart grid [73, 74]. Although *zinc* is not one of the most frequently appeared keywords, it appeared over many years, starting early 2000s. The possible reason is that zinc is widely used for both solar panel and photocatalytic applications including protective coatings for preventing rust and in the protection of solar systems, while ZnO is a well-known and widely studied photocatalyst in many applications [75–77].

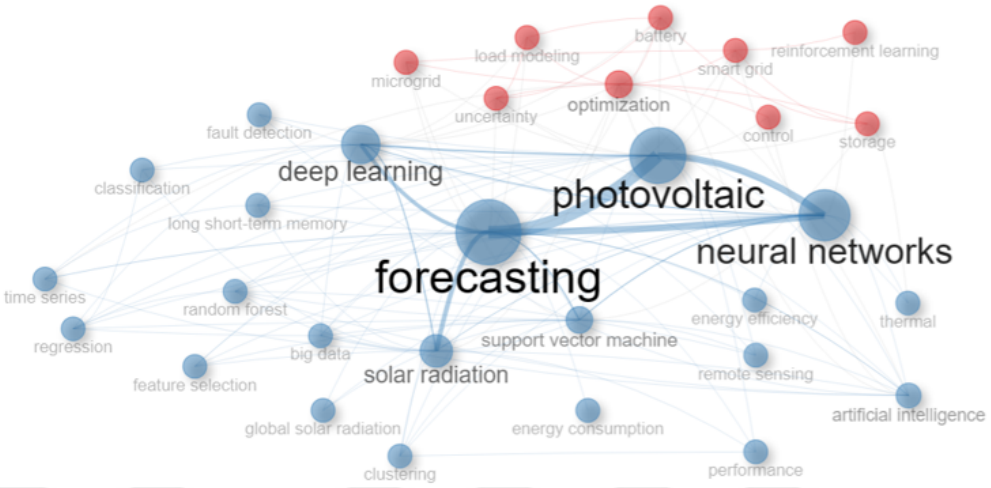

Figure 3.7. Co-occurrence network for keywords of ML in solar energy.

Various algorithms appear in the trend analysis for machine learning in solar energy utilization (Figure 3.9). *Monitoring* and *long-short term memory* seem to be the latest trends in ML in solar energy utilization. Monitoring of the condition of solar energy systems such as photovoltaics [78] or monitoring of energy production [79] was done by machine learning. Long-short-term memory is a deep learning method that was used in solar power prediction of photovoltaics [80, 81]. Hydrogen, photovoltaics and solar radiation were the main solar energy technologies that were analyzed by machine learning. Multi-objective optimization and techno-economic analysis are some of the areas that machine learning was used. As stated above, machine learning applications were widely used for forecasting and optimization. Since decarbonization and photothermal methods are some trending keywords for solar energy utilization, it can be expected for ML to be widely used in those areas of solar energy utiilization in the near future. Early applications of ML for solar energy utilization focused on the prediction of solar radiation, and performance prediction of photovoltaics and hydrogen generation became trending around 2021. The appearance of decision tree in recent years is interesting since it is a relatively old model that is mainly used for classification purposes and ensemble models such as random forest and gradient boosting usually have higher predictive power. Possibly, it is used in studies that compare various machine learning algorithms in solar radiation prediction [82] or photocatalytic

performance prediction [39], however, some recent research is focusing on utilizing decision trees for solar radiance prediction [83] or identification of optimal location for solar power plants [84].

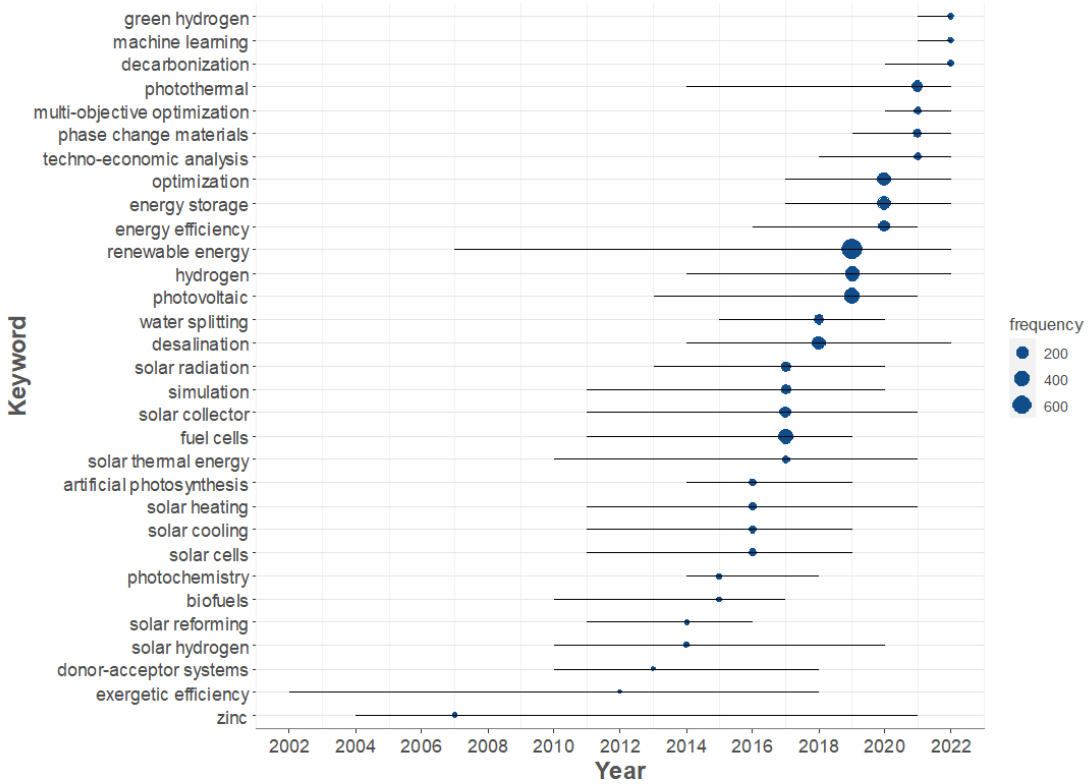

Figure 3.8. Trend analysis for the keywords of solar energy utilization.

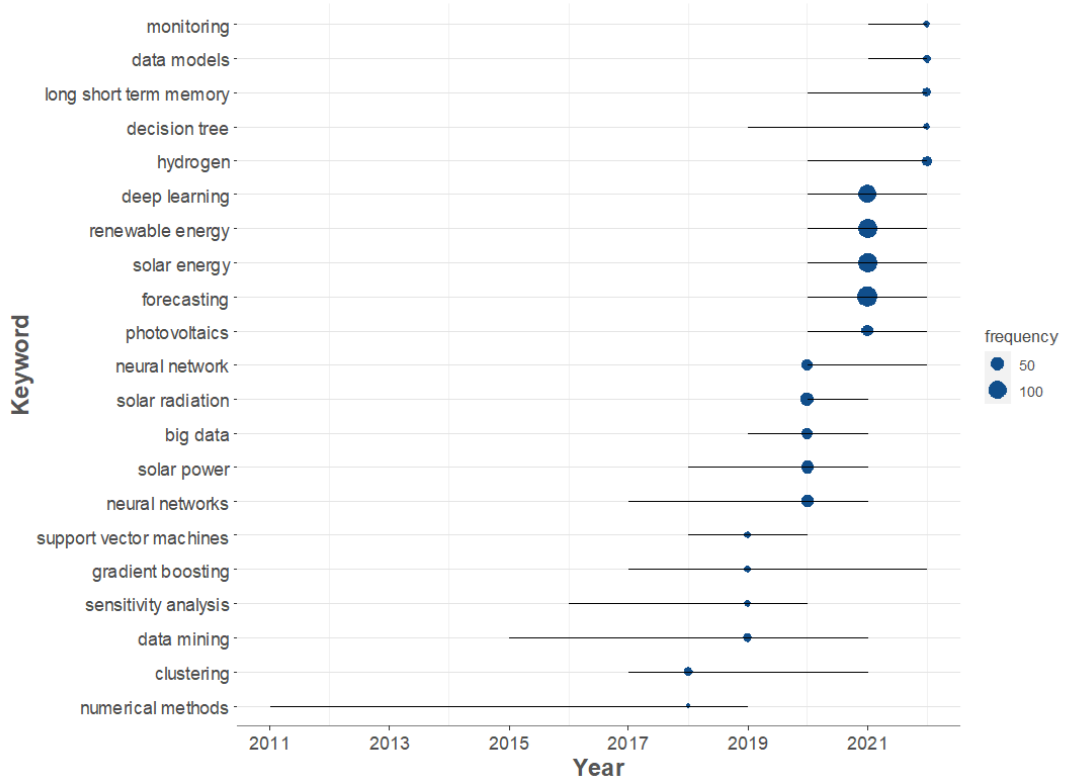

Figure 3.9. Trend analysis for keywords of ML in solar energy utilization.

4. ANALYSIS OF HUMIDIFICATION DEHUMIDIFICATION DESALINATION VIA MACHINE LEARNING

This work was performed in collaboration with the group of Prof. Majid Amidpour of University of Tehran, and the data for this work was collected by Dr. Mohsen Salimi; the manuscript was submitted to Desalination journal on September 2023.

4.1. Background Information on Solar Desalination

Desalination has emerged as a critical treatment in the fight against the growing scarcity of freshwater supplies [85]. The pressing worries about human population growth, pollution of the environment by industry, and poor farming techniques have resulted in the rapid depletion of available potable water. As, part of the solution, thousands of desalination units are responsible for producing over 95 million cubic meters of fresh water on a daily basis [86]. However, despite this significant progress, there remains a pressing demand for improving freshwater production techniques. Desalination through sustainable energy sources and cutting-edge techniques is a potential alternative for tackling the global water scarcity dilemma and relieving the load on freshwater reserves. Desalination methods are categorized as electrochemical, thermal, and membrane-based processes as shown in Figure 4.1 [87, 88].

Renewable and low-grade energy sources may be used to power the humidification-dehumidification (HDH) desalination system making them ideal for supplying freshwater needs around the world. To operate HDH desalination systems, solar energy is the most frequently used renewable energy source; it is used to boost the working fluid's temperature. The transporting medium and the water flow are usually heated simultaneously. In HDH desalination systems, air is the most often used carrier gas; the solar air heaters are used to increase the air temperature in the HDH system as the air is better able to absorb water vapour by elevating its thermal energy. Several HDH

desalination cycles for utilizing solar air collectors have been developed [89]. Figure 4.2 displays the heaters that are used to run the HDH system [90].

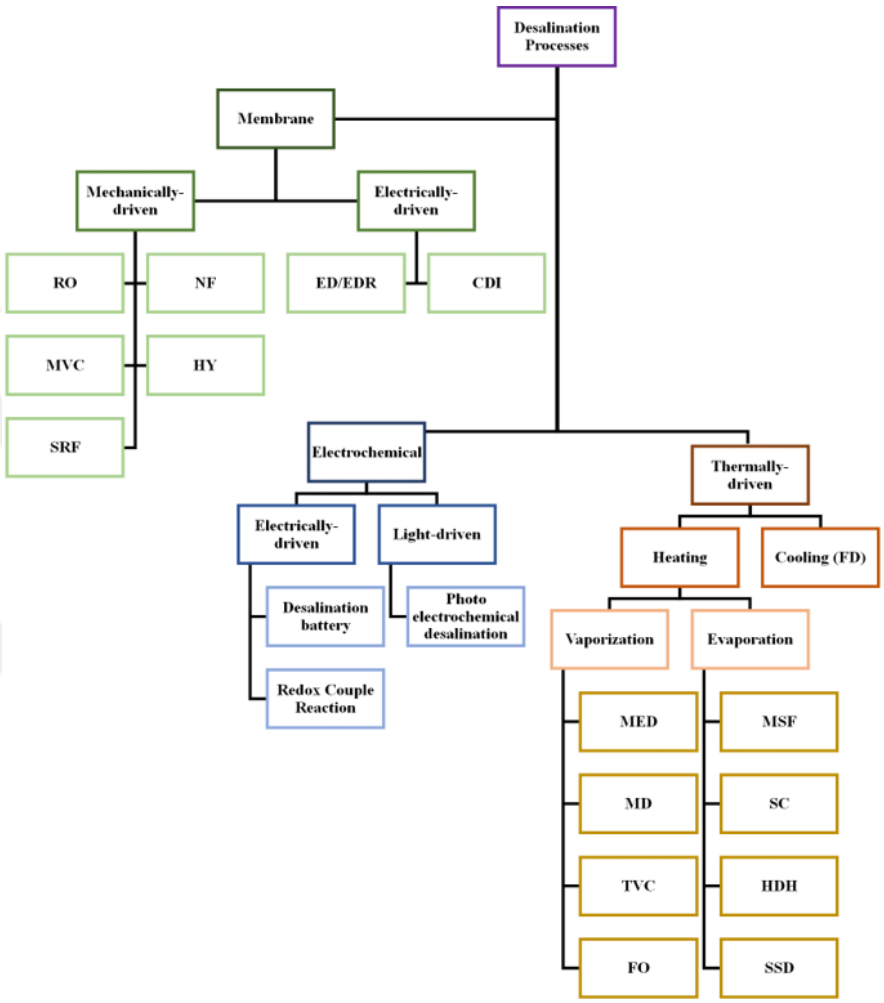

Figure 4.1. Types of desalination technologies. Adopted from [39].

The capacity of phase change technologies, notably humidification-dehumidification systems, to support desalination operations while utilizing renewable and low-grade heat sources has sparked considerable attention. This characteristic displays a positive attitude towards environmental issues. Figure 4.3 depicts a variety of HDH desalination systems. The performance of HDH desalination is related to a variety of factors, such as feed water temperature, relative humidity, airflow velocity, and the surface area and structure of the heat exchanger [91]. For instance, higher temperatures cause increased evaporation rates, which improves system efficiency. Similarly, it is critical

to precisely regulate relative humidity levels in order to avoid impediments to the condensation process [91]. Optimal air flow rates aid in producing effective condensation and maintaining suitable humidity levels [91].

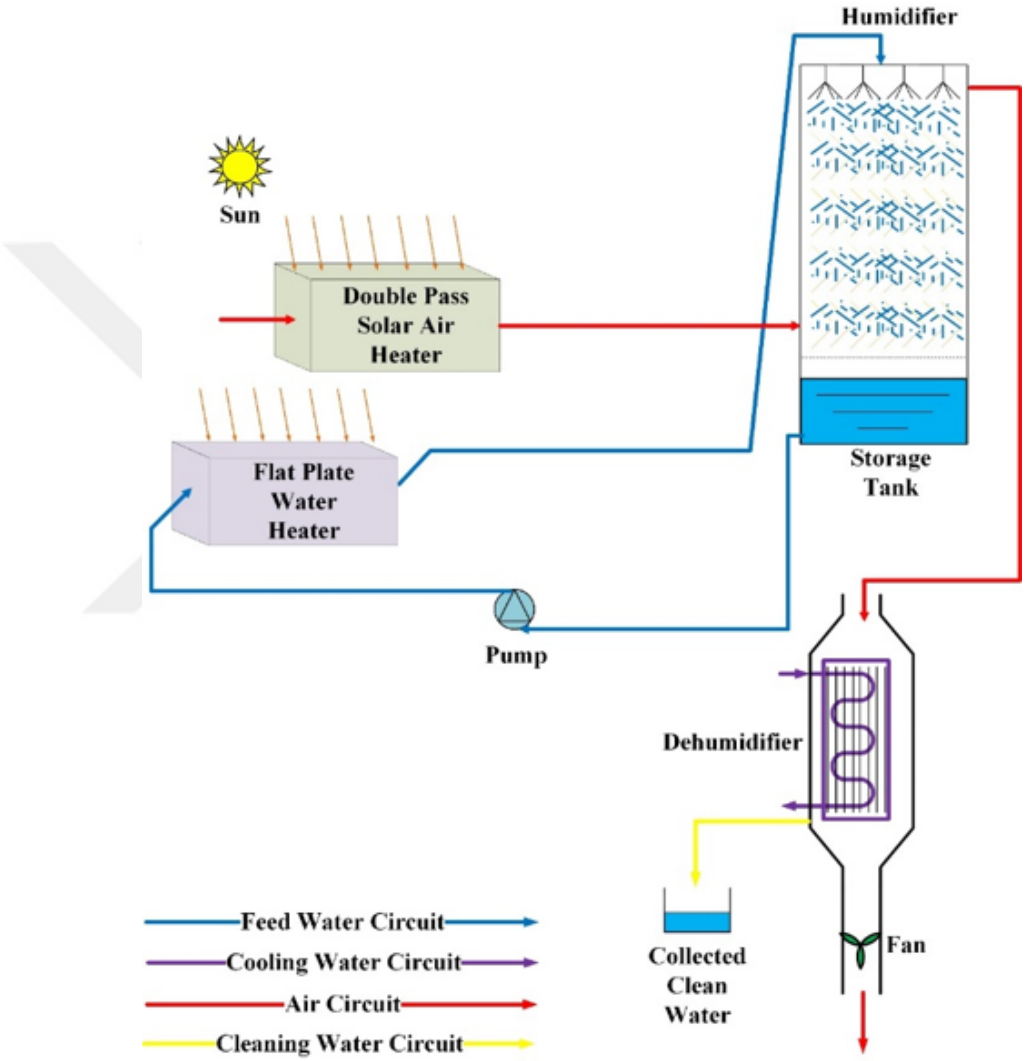

Figure 4.2. Solar-driven HDH desalination system. Adopted from [85].

The dimensions and construction of heat exchangers have a significant impact on heat transfer efficiency as well. Larger surface areas and optimal configurations have been reported to have the ability to improve system performance [92]. Consequently, it is critical to thoroughly evaluate and control these variables to maximize the effectiveness of HDH desalination systems.

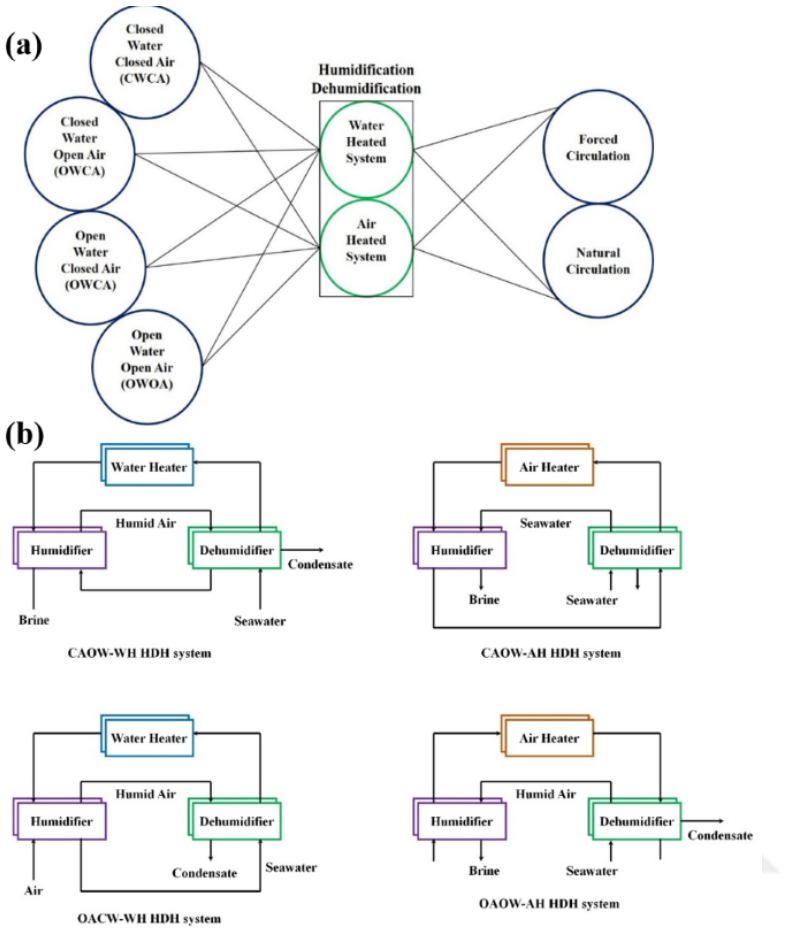

Figure 4.3. Classifications of the HDH desalination systems (a) and schematic of their operation (b).

ML has been used widely in various fields including renewable energy systems [93] including those utilizing solar energy such as photovoltaics [9], photoelectrochemical water splitting [94], photocatalytic CO_2 reduction [95] and solar desalination [96] with increasing availability of data through scientific publications and databases as well as the recent progress in data storage, retrieval and analysis technologies. As the supply of freshwater seems to be an important issue in future for a sustainable world, new desalination technologies including solar humidification dehumidification desalination systems to convert seawater to fresh water have been investigated extensively in recent years. ML has been also utilized in some of these works; for example, Priya et. al. [97] performed a material search for desalination systems using machine learning while Plasencia et. al. [98] developed support vector machines and decision tree models for reverse osmosis desalination. Similarly, Khayet and Cojocaru [99] developed

an artificial neural network model for sweeping gas membrane desalination using 53 experimental data points. Acevedo et. al. [100] employed ANN model to develop an innovative methodology for forecasting the permeate output in a gap membrane distillation module across diverse operational parameters, such as the temperatures at the condenser and evaporator inlets, and the flow rate of the feed seawater. K et. al. [101] used machine learning methodologies, including Response Surface Methodology, Artificial Neural Network, and Adaptive Neuro-Fuzzy Inference System to perform modelling and optimization of the Forward Osmosis process. Kandeal et. al. [102] employed experimental data that was gathered under the specific climatic conditions of Egypt; four distinct machine learning models, namely artificial neural network , random forest, support vector regression, and linear SVM, were utilized to forecast the performance of a double slope solar still. Wang et. al. [103] investigated the utilization of experimental data to construct and evaluate multiple models. This study examined various models for predicting the hourly performance of tubular solar stills. These models included classical artificial neural networks as well as random forest with and without Bayesian optimization, and traditional multilinear regression. Sohani et. al. [104] aimed to improve the design of a solar still desalination system by taking into account two primary performance indicators, namely the hourly distillate production and the temperature of water in the basin. To achieve this goal, the study employs different types of artificial neural networks, such as Feedforward, backpropagation, and radial basis function. Kiran Naik et. al. [105] employed a thermal model and a k nearest neighbor-ML tool to predict and assess the efficacy of a liquid desiccant regenerator that utilizes a membrane, powered by either waste heat or solar heat. Furthermore, an analysis was conducted on the rate of water extraction from the process of desalination. In another work, Faegh et. al. [106] compared various artificial neural network models for heat pump-assisted humidification-dehumidification desalination system; they used 180 experimental data and predicted gain output ratio, heat transfer rates of the evaporator, and evaporative condenser with saturation temperature of evaporator and the condenser, inlet flow rates and temperatures of air, water and refrigerant as input variables. In all these examples, the ML models on desalination used a limited number of data points generated in the same laboratory. As a bibliometric review study, the

work of Nia et. al. [107] was based on publications listed in Scopus through years but it involves a bibliometric analysis of desalination with artificial neural networks to assess the status of desalination research through the years.

As far as we know, there is no work reporting ML models for desalination performance based on the data collected from past publications combining the experience gained by various laboratories from the entire literature. In this part of the thesis, we constructed an extensive dataset from the experimental HDH solar desalination papers published in the literature and developed predictive ML models for distilled water flow rate (DW), recovery ratio (RR) and gained output ratio (GOR). We tested various ML regression algorithms such as random forest, gradient boosting, support vector machine and artificial neural networks and reported the models representing the data best for all three output variables; RF and GB models were chosen because they worked better due to the presence of categorical variables in the dataset.

4.2. Computational Details

4.2.1. Construction of the Dataset

The dataset was built from 38 experimental papers, which were published in 2003-2022 and available in online sources. The dataset contains 838 data points. Distilled water flow rate, recovery ratio and gained output ratio were used as performance (output) variables; the dataset contained 694 data points for DW, 571 data points for RR and 267 data points for GOR datasets. The input variables (descriptors), which describe the characteristics of the humidifier and dehumidifier as well as the operational conditions at which the data is taken. The ranges and levels of the descriptors and outputs are given in Table 4.1. Equivalent diameter was calculated using the same formula given by Koch [108]. The missing values in the dataset was filled with column means/modes of the training set.

Table 4.1. Levels and Ranges of all variables in the dataset.

Variable	**Level/Range**
Height of dehumidifier	13 - 240 cm
Equivalent diameter of dehumidifier	13 - 714.75 cm
Geometry of dehumidifier	Cylindrical, Rectangular
Type of dehumidifier	copper, cross flow, fin and tube, plexiglass, polypropylene, shell and tube
Surface Area of humidifier	0.12 - 30 m^2
Packing type in humidifier	aluminium, cellulose, gunny, honey-comb paper, none, paddy grass, pall ring, plastic, textile, wooden
Height of humidifier	10 - 2100 cm
Equivalent diameter of humidifier	5.8 - 714.7 cm
Geometry of dehumidifier	Cylindrical, Rectangular
Specific area of packing materials	5 - 2800 m^2/m^3
Type of HDH cycle	CACW, CAOW, OACW, OAOW
Water Heated	Yes, No
Variable	**Level/Range**
Air Heated	Yes, No
Inlet Water Temperature	22 - 90 °C
Inlet Air Temperature	20 - 60 °C
Inlet Water Flow Rate	0.01 - 60 kg/min
Inlet Air Flow Rate	0*- 49.7 kg/min (0*: natural flow)
Distilled Water Flow Rate	0.0 - 61.2 kg/h
Recovery Ratio	0.000006 - 0.18
Gained Output Ratio	0.05 - 2.4

4.2.2. Model Development

Regression models for DW, RR and GOR were developed in R environment using RandomForest [109] and gbm [110] packages for random forest (RF) and gradient boosting (GB) regression respectively. Boruta analysis of the datasets for feature selection was done using Boruta [111] package. Swarm plots were generated using ggplot2 [112] and ggbeeswarm [113] packages.

Number of trees (ntree) and number of variables (mtry) are used as hyperparameters for the optimization of the random forest model. The number of estimators (ntree), interaction depth and shrinkage (learning rate) were optimized as hyperparameters for gradient boosting. 5-fold cross-validation (CV) was used to tune the model hyperparameters [23] with the train-test split ratio of 80-20%. The test set was randomly separated first; then the training set was further divided into five subsets (5-fold) randomly. The best models for DW and RR prediction was achieved by RF (with ntree of 125 and mtry of 10 for DW, and ntree of 15 and mtry of 10 for RR). GB regression gave the highest predictive power for GOR set with ntree of 20, interaction depth of 4 and shrinkage of 0.41.

4.3. Results and Discussion

4.3.1. Pre-analysis of Database

Before developing predictive ML models, the structure of data was analyzed first using some simple descriptive statistics. Figure 4.4 shows the distribution of data based on three output variables (DW, RR and GOR); the range of the x-axis was chosen as the range of corresponding variables in the data set to show where the majority of data were accumulated. Most of the data are located in the lower end of ranges for all three variables even though some of the data were also gathered closely in higher values indicating some bimodal distribution; this pattern is more apparent in GOR data. This type of distribution usually indicates the presence of influential categorical variables with few alternatives (like two designs of materials alternatives with different performance levels); sometimes these variables are within the input variables in the dataset (as fortunate) or some differences in the experimental systems or measurement that are not reported in relevant papers (as unfortunate). Hence, we checked the influence of major input variables to search for the signs of the variables causing this distribution as well as other patterns that may be important for further analysis.

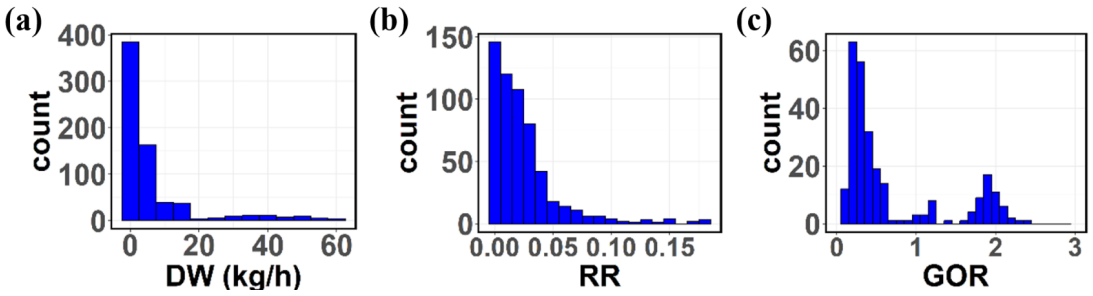

Figure 4.4. Distribution of the data for (a) DW (b) RR and (c) GOR.

First, we divided the data based on the type of desalination systems that mainly differ in characteristics of air and water flow; they may have closed (C) or open (O) loops for air (A) and/or water (W), while the solar energy can be used to heat inlet air, water or both. In Figure 4.5, the distribution of DW, RR and GOR data among different HDH types are presented. Most of the DW and RR data were from closed-air-open water (CAOW) loops operated with heated water systems and open-air close-water loops (OACW) with water-heated systems; DW data in CAOW and OACW are also more localized at lower regions while CACW and OAOW data scattered to higher values supporting the distribution of DW data in Figure 4.5a. Patterns are also similar in RR data in Figure 4.5b. Most of the GOR data, on the other hand, are from OACW loop systems with heated air and have lower values. There are also a significant number of data points obtained from CACW with heated water while CAOW data also on the higher GOR values explaining the more apparent bimodal character of distribution presented in Figure 4.5c for GOR data.

We should mention here that swarm plots presented can be normally used to see whether the comparison of average performance is meaningful or not. It can be done if the distributions are close to normal in which averages represent the entire dataset well. For example, Figure 4.5c indicates that the average performance of CACW system operating under heated water is much higher than the average performance of OACW system with heated air. Generally, the OACW configuration has the lowest GOR among different HDH cycle configurations [116]. It can be also said that RR

performance of water-heated CAOW data is slightly higher than the performance of air-heated OACW data. On the other hand, the average value of a distribution like DW data from water-heated CACW (Figure 4.5a) has no meaning because the distribution indicates that the next data may have any value in the range of the plot. Even though their distribution is not entirely normal, the averages of heated water CAOW and OACW DW data can be used as the indicator of similar performance considering that a large number of data points are located in a quite narrow range. This point of view can be extended be also extended for Figure 4.6 if necessary.

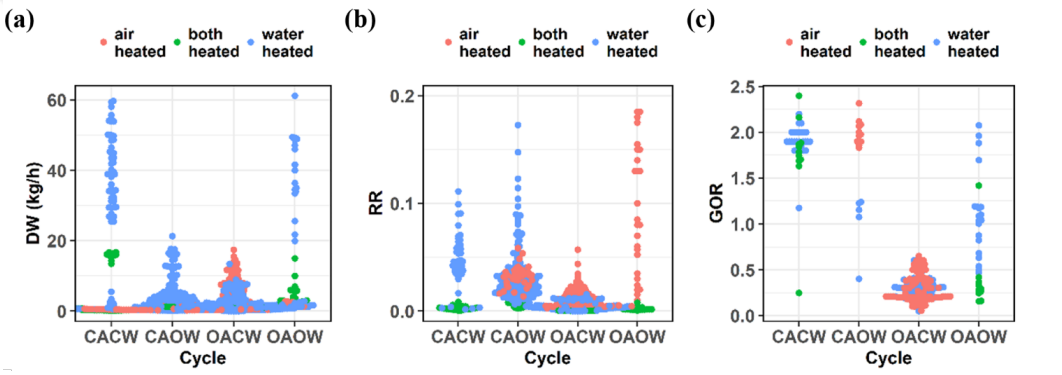

Figure 4.5. Effect of cycle and heating on (a)DW, (b)RR and (c)GOR.

Figure 4.6 indicates the effects of packing material in the humidifier. Cellulose is the most used packing material followed by plastic for all datasets. Cellulose is a good packing material for the HDH desalination system for several reasons: i) High surface area: Cellulose has a high surface area per unit volume, which allows for efficient heat and mass transfer. This is important for the HDH system, which relies on the transfer of heat and water vapour to produce fresh water. ii) Hydrophilic nature: Cellulose is hydrophilic, meaning it has a strong attraction to water molecules. iii) Low cost and abundance: Cellulose is a widely available, low-cost material that can be easily sourced from plant matter. This makes it an attractive option for use in the HDH desalination system, which requires large quantities of packing material. iv) Environmentally friendly: Cellulose is biodegradable and non-toxic, making it an environmentally friendly choice for use in the HDH system. Overall, the combination of cellulose's high surface area, hydrophilic nature, low cost and abundance, and environ-

mental friendliness make it a great choice for packing material in the HDH desalination system. As Figure 4.6a indicates, the use of plastic occasionally produces higher DW than the others while the majority of the data points are located at lower values than the others. Interestingly, the oscillation in RR performance seems to be lower with plastics while slightly increasing with cellulose. The fluctuation to higher GOR values is more apparent in Figure 4.6c as expected considering the clear bimodal distributions of GOR data in Figure 4.4c. The textile as a packing material has a low surface area and does not have structural stability. Therefore, its performance is poor.

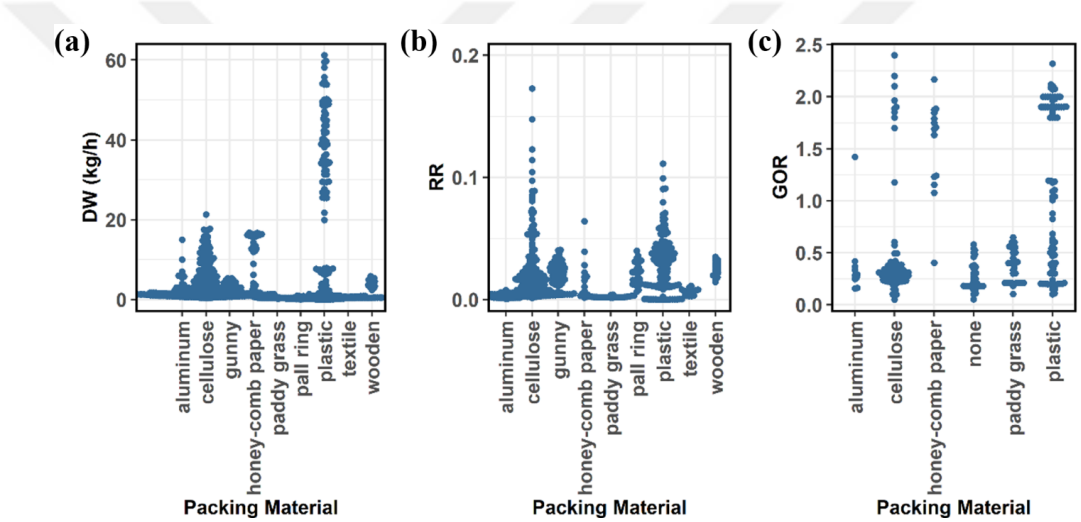

Figure 4.6. Effect of packing on (a)DW, (b) RR and (c)GOR.

Finally, we investigated the effects of water temperature on the DW, RR and GOR performance in Figure 4.7. Since temperature is a continuous variable, we should not expect the data located at certain values; instead, we should look at how the range of distribution changes with increasing temperature; as expected more data shift to higher values of output variables with increasing temperature. However, there seem to be two groups of data in DW and GOR data indicating that the temperature may be affecting the outcome under different values of the other variables.

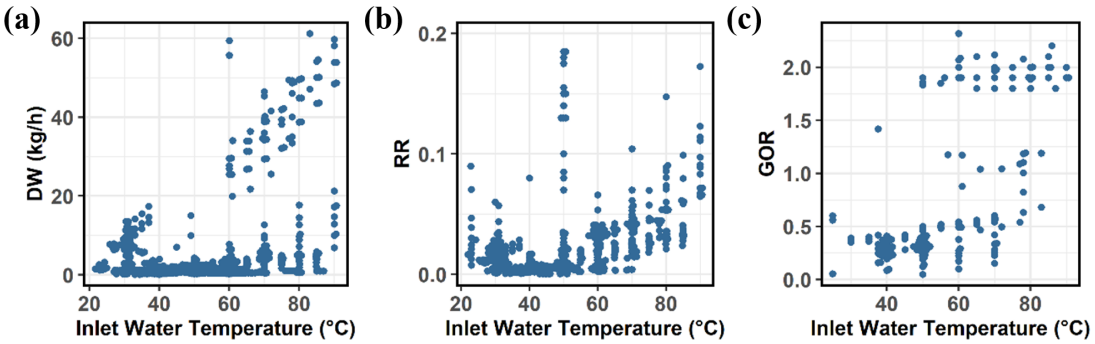

Figure 4.7. Effect of inlet temperature on (a)DW, (b)RR and (c)GOR.

4.3.2. Predictive Models for DW, RR and GOR

Since the ML model with a smaller number of descriptors are more robust and reliable (for a given number of data), only descriptors that have a significant effect on the output variables should be used. Boruta analysis, which is a feature selection technique relying on wrapper-based algorithm and tree-based modelling, was utilized for this purpose. This technique creates an extended version of the dataset by shuffling and fitting a random forest model iteratively to estimate the importance of the input variables [111]. Boruta analysis of DW and RR sets is given in Figure 4.8; GOR set could not be analyzed this way because the number of missing values was too high. Hence, all input variables were used in the predictive models of GOR. The most significant descriptor for the DW dataset was found to be inlet water temperature followed by inlet air flow rate (Figure 4.8a).

Inlet water flow rate and the equivalent diameter of the humidifier were descriptors of high importance for RR dataset while the surface area of the dehumidifier was significant for both of the datasets [114, 115]. On the other hand, Boruta analysis showed that the geometries of the humidifier and dehumidifier (cylindrical or rectangular) are not significant; probably, the size and surface area were sufficient to describe the system regardless of the geometrical structure.

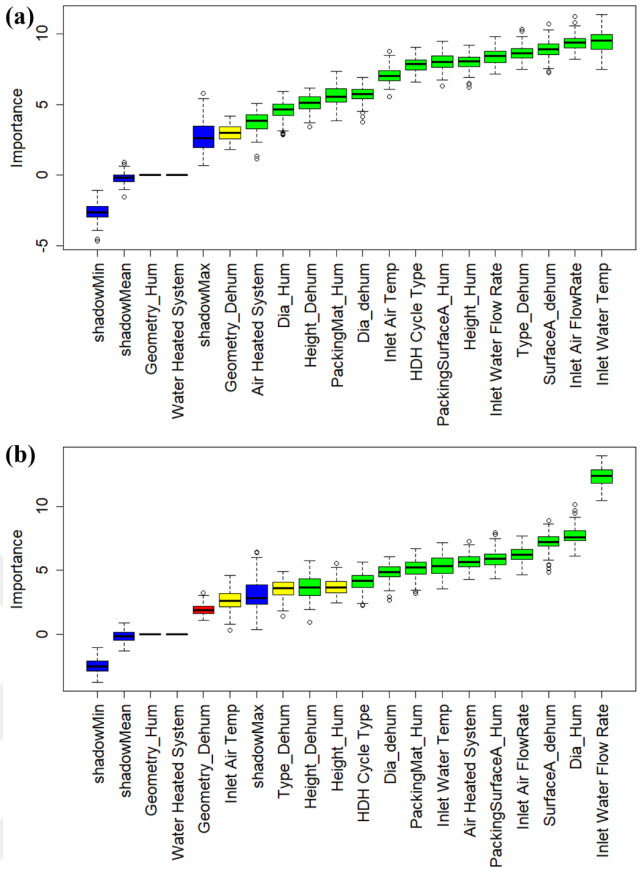

Figure 4.8. Boruta Analysis of (a) DW and (b) RR dataset (the colors represent the significance of the variables: green for significant variables, yellow for undetermined significance and red for insignificant variables).

Similarly, the descriptor indicating whether the system is "water heated or not" was found to be insignificant probably due to the redundancy of information considering that the "zero" value of the descriptors representing "air heating" is described already. It should be noted that these results only give a preliminary assessment of variables, the influential variables on the prediction of DW and RR were discussed later in this section.

Considering the results of Boruta analysis, the geometrical descriptors for humidifier and dehumidifier (being rectangular or cylindrical) and descriptor indicating whether the system is water heated or not were excluded for DW and RR models; for the GOR models, on the other hand, all variables were used as we have no results

suggesting otherwise. RF was found to be the best choice for DW and RR while gradient boosting performed better with GOR data. Only the best models are presented (Figure 4.9-Figure 4.11) and discussed in this section. Random forest model with ntree of 125 and mtry of 10 gave the lowest average validation error (4.9a); then this model was used to predict the entire dataset and test set that has not seen by the model before resulting in the predictions presented in Figure 4.9b and Figure 4.9c respectively. Although the model predictions are generally good, the low DW values were overestimated while the data at the high end of the spectrum were underestimated. The absence of a significant but uncounted descriptor may be the reason for this descriptor as the validation and testing plots in Figure 4.9b and Figure 4.9c have the same pattern even though it is less obvious. The most important variable for the model was the type of HDH cycle followed by water and air flow rates, which are important for obtaining suitable humidity levels which directly affect DW [91].

The best model for RR prediction was also achieved by RF with ntree of 15 and mtry of 10 (Figure 4.10). This time, although the accuracy of predictions is generally good, the model does not seem to represent some data points well (especially at high RR values). This may be caused by the presence of some uncounted descriptors (like differences in the systems in different papers even though they are labelled with the same names); another plausible explanation is that the number of data points may not be sufficient to develop a model that will describe the system in entire operational range in literature. Inlet water temperature, type of HDH cycle and packing surface area are important variables for the model. Temperature and cycle type are related to the amount of energy transferred between saline and circulating water.

For instance, lower RR can be observed during OACW operation because of lower air dehumidification when the inlet saline water temperature is high [1]. Surface area in the humidification part of the desalination setup is also important since the mass transfer increases with increased contact area and therefore, increases the RR [92].

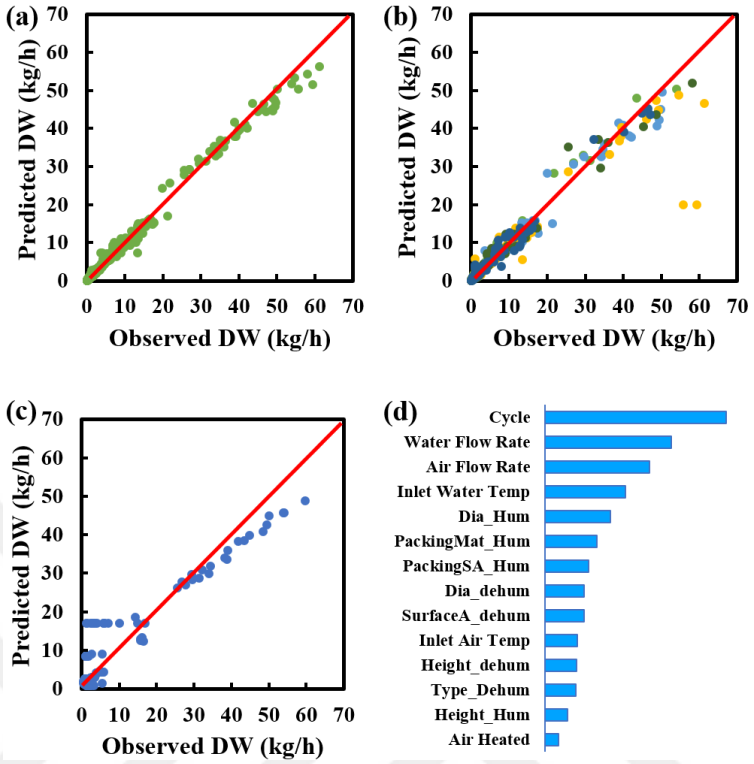

Figure 4.9. Predicted vs observed values for random forest model for DW (a) validation (b) train (c) test set results and (d)variable importance of model.

GB showed better predictive power compared to RF, for GOR dataset (Figure 4.11). The optimum hyperparameters were 20, 4 and 0.46 for ntree, interaction depth and shrinkage respectively. It seems that the data were clustered in three GOR ranges (around 0-1, around 1-1.5 and around 2). These trends were captured in the validation and training sets while the test set happened to be on the lower side of GOR; as clearly indicated in Computational details, the number of data points is relatively small for GOR (data from only twelve papers), and we made sure that all data taken from the same experimental systems (i.e. with same size of packing material) are either in training or testing set to prevent information leak. Apparently, this caused the testing set, which is only 20% of the entire dataset to be formed from the papers with low GOR values. Nevertheless, the fitness of the model is quite good for all validation, training and testing sets. Since we could not perform Boruta analysis for GOR data, we presented the significance analysis for the descriptors from GB analysis in Figure 4.11b. Similar to Boruta analysis, the descriptor stating the presence or

absence of water heating and geometries of humidifier and dehumidifier did not affect GOR; additionally, air inlet temperature, surface area and height of humidifier are also found to be insignificant for GOR [116].

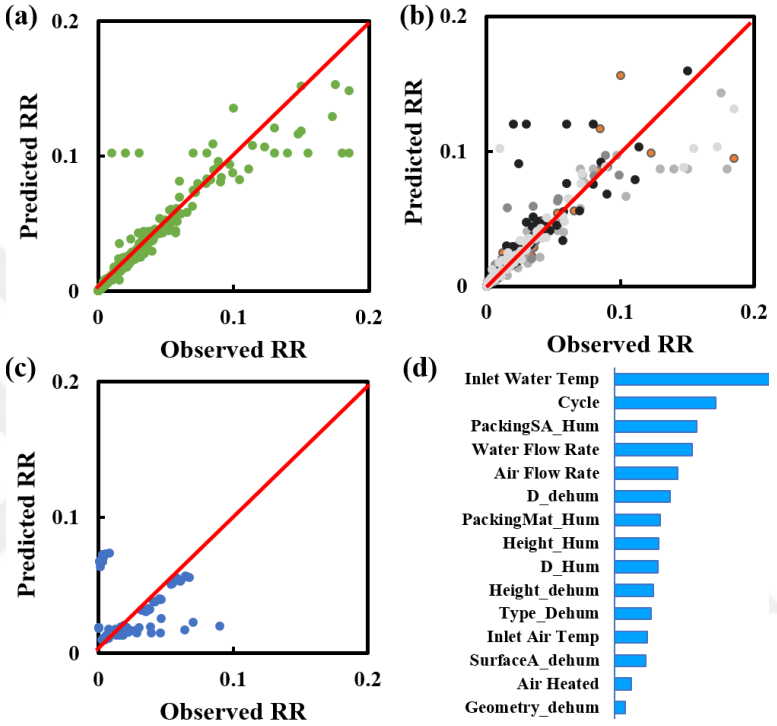

Figure 4.10. Predicted vs observed values of random forest model for RR (a) validation (b) train (c) test set results and (d) variable importance of model.

The air inlet temperature affects the amount of water that can be evaporated from the humidifier [117]. However, the effect is usually small because the air temperature can only be slightly lowered by the cooling system. Therefore, the impact of air inlet temperature on GOR may be negligible. The surface area of the humidifier affects the amount of water that can be evaporated, but increasing the surface area beyond a certain point may not have a significant effect on GOR [118]. The height of the humidifier refers to the vertical distance between the inlet and outlet of the humidifier, which determines the residence time of the air in the humidifier. It affects the residence time of the air inside the humidifier and thus the amount of water that can be evaporated. However, increasing the height of the humidifier beyond a certain point may not have a significant effect on GOR because the rate of heat and mass transfer is already high

due to the design of the humidifier. Additionally, increasing the humidifier height can increase the overall size and cost of the system, which can be a disadvantage.

On the other hand, the diameter of the humidifier, the HDH cycle and the type of the dehumidifier have a strong influence on the prediction of GOR. The diameter of the humidifier is important because it determines the cross-sectional area available for the airflow and, consequently, affects the air velocity through the system. The air velocity, in turn, affects the heat and mass transfer rates in the humidifier, which are critical for the performance of the system. In general, a larger diameter humidifier will allow for a greater air flow rate, which can lead to higher heat and mass transfer rates and, therefore, a higher GOR. However, there is a trade-off between the humidifier diameter and the system cost [117]. A larger diameter humidifier will require a larger system and, therefore, may be more expensive to build and operate [115]. On the other hand, a smaller diameter humidifier may lead to a more compact system but may have lower heat and mass transfer rates and, consequently, a lower GOR [115]. As a result, the diameter of the humidifier is an important design parameter to consider when designing an HDH desalination system; its optimum value will depend on various factors, including the required system capacity, the available space for installation, and the available budget for construction and operation [115]. The cycle type affects the GOR of the HDH desalination system because of the differences in the energy requirements and efficiencies. Additionally, the cycle type can also affect the overall cost and complexity of the HDH desalination system.

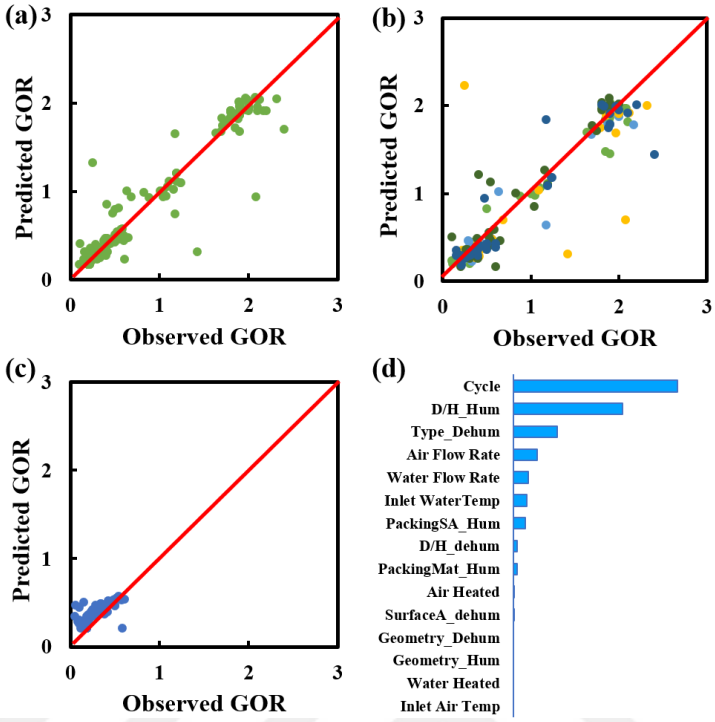

Figure 4.11. Predicted vs observed values by gradient boosting for GOR (a) validation (b) train (c) test set results and (d) variable importance of model.

5. MACHINE LEARNING ANALYSIS OF DYE SENSITIZED SOLAR CELL

The data for this part was collected by Neslihan Şener, Fadimenur Uslu and İlknur Boyan.

5.1. Background Information on Dye-Sensitized Solar Cells

The world's energy demand is increasing at an unprecedented rate, and fear of the depletion of fossil fuels, as well as continuously increasing concerns for the environment, has led researchers to focus on renewable energy sources such as solar, wind, biofuels and geothermal. Among these alternatives, solar energy is the most promising source since sunlight is abundant as the origin of most of the energy sources, free and well distributed around the world. Converting solar irradiation into electricity through solar cells is one of the most attractive and therefore extensively studied ways to harvest solar energy. The history of solar cells dates back to the discovery of the photovoltaic effect in 1839 [119]. The first practical silicon solar cell with an efficiency of 6% was developed in Bell laboratory in 1954 [120]. Since then, the efficiency of solar cells has improved significantly, reaching over 40% (in tandem cells) in recent years [121]. This increase in efficiency, together with the decrease in the cost of solar cells, has made them a popular choice for renewable electricity generation. Solar cells are usually classified chronically in three classes (called generation); the first generation usually refers to silicon solar cells while the second-generation solar cells contain thin films solar cell technologies; the dye-sensitized, organic and perovskite solar cells, which are still in research stage, are called as third generation solar cells.

Dye-sensitized solar cells are a type of low-cost solar cell that was invented at UC Berkeley in 1991 [122]. DSSCs are a photoelectrochemical system that consists of a photo-sensitized anode, a counter electrode and an electrolyte. The anode is typically made of semiconductor material, such as silicon or titanium dioxide coated

with a broad-spectrum dye as a sensitizer; it absorbs light and generates electrons. The counter electrode, on the other hand, is usually made of a conductive material, such as graphite or platinum, that collects the electrons from the external circuit and facilitates the redox reduction in the electrolyte. The electrolyte is a solution that contains ions that can move between the anode and the cathode, allowing the flow of current. The most commonly used electrolyte in dye-sensitized solar cells is an iodine/tri-iodine solution. The electron generated in the photoanode is transferred to the front electrode, passed through a circuit to produce electricity and returns back to the anode through the electrolyte solution. The corrosive nature of the conventional electrolyte formulations limits the industrialization of DSSCs [123]. Acetonitrile and valeronitrile are low boiling point solvents that cause leakage in long-term stability tests [124] and are mostly used in DSSCs. Our dataset involves many organic solvents such as ethanol, dichloromethane etc. Various solvents are used for solving the iodine salts; organic solvents, ionic liquids and solid-state electrolytes [124].

Dyes used in DSSC applications can be natural or synthetic dyes. They should have good stability in redox conditions, great visual absorption and efficient charge injection and regeneration. Most commonly used dye is ruthenium-based metal complex dyes; N719 (di-tetrabutylammonium cis-bis(isothiocyanato) bis(2,2'-bipyridyl-4,4'- dicarboxylato) ruthenium(II)), as one of the Ru-containing dye, is considered as a benchmark dye due to its excellent absorption efficiency and high stability [125]. However, since ruthenium is an expensive material, cheaper dye alternatives such as porphyrins, phthalocyanines, and metal-free organic dyes have been also researched extensively [126].

The performance of a DSSC is mainly measured by photoconversion efficiency (PCE) but other electrochemical characteristics such as short circuit current density (J_{sc}), open circuit voltage (V_{OC}) and fill factor (FF) also used as the other performance indicators as each provides different and valuable insights for the performance characteristics and the problems of DSSC under investigations. J_{sc} represents the current density when there is no voltage applied across the system and it is determined by

minimizing the resistance encountered during charge transfer at the interface between the anode, dye, and electrolyte. On the other hand, V_{OC} represents the maximum voltage that can be obtained from the system, i.e. refers to the electrical potential measured across a circuit when there is no flow of current, and the resistance within the circuit is infinitely high. FF is a parameter that determines the maximum power from a solar cell by comparing the maximum power produced to the product of the external open circuit voltage and short circuit current. It is a reliable indicator of the device's quality and is influenced by losses in both series and shunt connections. PCE can easily be calculated using these properties with the following formula [127]

$$\eta = FF \ \frac{J_{sc}V_{oc}}{P_{in}}, \tag{5.1}$$

where P_{in} is incoming solar power.

Significant amount of research has been conducted to increase the PCE of solar cells; examples of such works are the increase the porosity of the anode material, i.e. TiO_2 [128] use of co-sensitized dyes [129], developments of molecularly engineered [130] or novel dyes [71] and testing different configurations for the structure of DSSC [131]. As the results of all these efforts and investigation of optimum parameters, a large amount of data were accumulated on the subject through years; machine learning tools can be used to analyze this accusation to extract information that is not observable with the human eye, and the results can be used to improve the performance of DSSC further. Previously Wang et al used ML to find suitable organic sensitizers using pi-spacer details [132]. In another work, the peak absorption wavelength of the azo-based dyes was predicted using xgboost regression [20]. Our group has analyzed the performance of perovskite cells [8], [133] and various photocatalytic reactions such as water splitting and CO_2 reduction via machine learning [95], [134, 135]. However, to the best of our knowledge, there is no work involving the machine learning analysis of the effects of cell materials (dye, semiconductor, electrolyte etc.) and manufacturing variables on the performance of the dye-sensitized solar cell. In this work, an extensive database correlating the cell materials and cell manufacturing variables was constructed and analyzed by using ML tools. Predictive models for all performance measures (PCE, J_{sc}, V_{OC} and FF) were developed using random forest and gradient boosting

algorithms; the most significant input variables (descriptors) affecting all performance measures were also identified.

5.2. Computational Details

5.2.1. Dataset Construction

The Web of Science database was searched with the keywords "dye-sensitized solar cells" for publications between 2009 and 2022. Only the synthetic dyes were considered; articles related to natural dyes were not recorded because they have a different level of performance requiring separate analysis. A total of 764 data from 102 articles, which report the results of experimental works, were used in the analysis. The 16 descriptors were used in ML models, and 4 output variables characterized the performance of a DSSC: PCE, J_{sc} V_{OC} and FF as presented in Table 5.1 with their ranges.

Table 5.1. The variables used in the dataset and their ranges/levels.

Variable	**Range/Category**
TiO_2 **addition**	Ag, Au, GeO_2, Gr, Ni, None, other, P-Z, SiO_2, Zn, ZnO
transparent layer method	anodization, deposition, doctor blade, hydrothermal, rf sputtering, screen printing, spin coating
blocking layer (nm)	0-871
blocking layer (mM)	0-500
transparent thickness (µm)	0-57.8
treatment mM	0-50
active area (cm^2)	0.09-5.6
thickness (µm)	2.5-55.57

Table 5.1. The variables used in the dataset and their ranges/levels. (cont.)

Variable	**Range/Category**
counter electrode	C, Co-Se, CoS, MoS_2, $NiCO_2S_4$, NiS, other, PEDOT, Pt, rGO
counter electrode method	deposition, dip coating, doctor blade, drop casting, electrodeposition, hydrothermal, in-situ polymerization, ion sputtering, phosphorization, rf sputtering, screen printing, spin coating, spray coating, sputtering
dye	C218, D-131, D102, N3, N719, N719-other, NO, other, phthalocyanine, ruthenium, SQ-140, W2, XW, Y-123, YD2-o-C8
dye solvent	acetonitrile-tert butyl alcohol, anhydrous ethanol, chenodeoxycholic acid, dichloromethane, ethanol, ethanol-acetonitrile, ethanol-dimethylsulfoxide, other, tert butyl alcohol-anhydrous ethanol, tetrahydrofuran-ethanol, toluene- ethanol
dye concentration (mM)	0.1-500
dipping time (h)	0-48
intensity (mW/cm^2)	10-100

Table 5.1. The variables used in the dataset and their ranges/levels. (cont.)

Variable	**Range/Category**
electrolyte_solvent	3-methoxypropionitrile, acetonitrile-ethylene carbonate, acetonitrile-valeronitrile, acetonitrile, acetonitrile-methoxypropionitrile, liquid iodine electrolyte
Photo Conversion Efficiency, %	0-12
Fill Factor	0.1-0.86
Open Circuit Voltage (V)	0.25-1.1
Short Circuit Current Density (mA/cm^2)	0.02-25

5.2.2. Model Development

The descriptive analysis of the dataset was performed using violin plots with ggplot2 package [112] of R. The regression analysis was conducted using various machine learning algorithms such as linear regression (stats package) [136], support vector machine (e1071 package) [137], random forest (randomForest package) [109] and gradient boosting (gbm package) [110] algorithms. Variable importance of the optimum models was obtained using a caret package [138]. However, the linear regression and support vector machine algorithms did not perform well; apparently, because most of the descriptors were categorical and transforming them to numerical descriptors via one hot encoding weakened the models. ntree and mtry for RF and ntree and learning rate for GBM were optimized as model hyperparameters using 5-fold cross-validation. The models were developed using the test split of 20% and random seed number of 124. Root mean square error was used as the performance measure in determining the optimum model for each output. The optimum hyperparameters and best algorithms for each output are summarized in given in Table 5.2.

Table 5.2. The algorithms used in prediction of performance.

Output	**Best Algorithm**	**Best hyperparameters**
PCE	GB	nree: 50 & learning rate: 0.25
J_{sc}	GB	ntree: 150 & learning rate: 0.1
V_{oc}	RF	ntree: 300 & mtry: 10
FF	RF	ntree: 200 & mtry:12

5.3. Results and Discussion

5.3.1. Pre-analysis of the Dataset

The data is first analyzed by simple statistical tools to detect outliers and major patterns. Anode as the working electrode, is an important part of the DSSC, and TiO_2 is the most commonly used semiconductor for this; however, in some works, some additional materials, mostly metals and metal oxides were added to TiO_2 to increase the electron transfer in the DSSC. Figure 5.1 shows the effects of these additional materials on efficiency, fill factor, open circuit voltage and short circuit current. For instance, SiO_2 addition to TiO_2 seems to enhance the efficiency of the DSSC (Figure 5.1a) since it has a higher average efficiency compared to bare TiO_2 (represented as "None"). The efficiency increase of SiO_2 incorporated TiO_2 photoanode could be related to decreased charge recombination due to SiO_2 being an energy barrier [139], this is also evident from increased FF. FF depends on the resistances within the system [140], therefore, decreased recombination would result in decreased resistances, hence improving FF. On the other hand, graphene-TiO_2 pair has the lowest efficiency due to a reduction in fill factor and short circuit current density even though open circuit voltage improved slightly; Sadikin et. al. suggests that this decrease could be due to insufficient dipping time of the electrodes in the dye solution since the structure and dye adsorption properties of the graphene-coated TiO_2 differs from bare TiO_2 the optimization of dipping time might result in higher PCE [141]. GeO_2 addition gave the highest fill factor followed by Au-TiO_2 and Ag-TiO_2(Figure 5.1b). The addition of nanoparticles like Au, Ag and Zn is reported to cause surface plasmonic resonance effect

which enhances the light absorption of the photoanode [142, 143]. In addition, Duan et.al. showed that TiO_2/GeO_2 and GeO_2/electrolyte interface had light scattering properties that enhanced the PCE of the DSSC [144]. Except for ZnO addition, which has negative impacts, all other materials had similar open circuit voltage (Figure 5.1c). Yeoh et. al. investigated the effect of the thickness ZnO blocking layer addition to TiO_2 photoanode and confirmed that the increased thickness of the ZnO layer reduced the performance due to rapid decrease in J_{sc} [145]. Short circuit current densities vary a lot according to the anode type, some works utilizing Ag-TiO_2 seem to achieve higher J_{sc} values but overall only SiO_2 and Zn addition seems to have a positive impact on the short circuit current density of the anode. The SiO_2 incorporation to TiO_2 resulted in Si-O-Ti covalent bond which increased the generation of photoelectrons i.e. the J_{sc} [146].

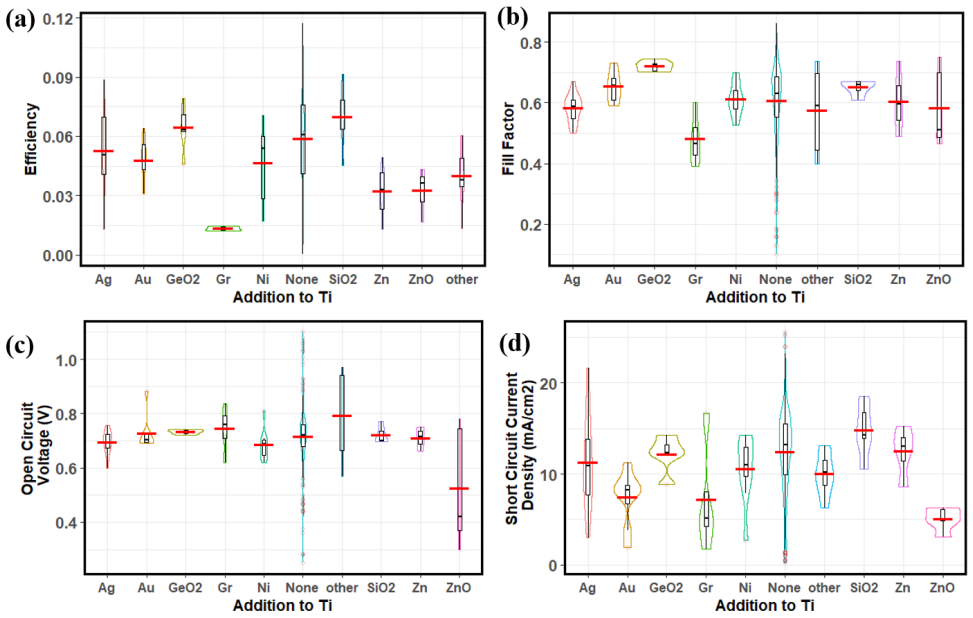

Figure 5.1. The effect on anode type on (a) PCE, (b) FF, (c) V_{OC} and (d) J_{sc} (red line represents the average values).

Counter electrode type is also important in the overall DSSC performance. Researchers have tried many counter electrodes in their DSSC and their comparison is given in Figure 5.2. Due to Pt being an expensive material and showing corrosion under iodine redox cycle [127], researchers tried different alternatives that showed good activity in tri-iodide reduction; carbon-based materials, polymers, metal oxides, metal sulfides and alloys are some of the materials that were utilized as counter electrode in DSSC applications [147].In our dataset, carbon-based materials, metal sulfides and polymers were mostly used as counter electrodes but the number of studies using Pt is still high (almost 50%).

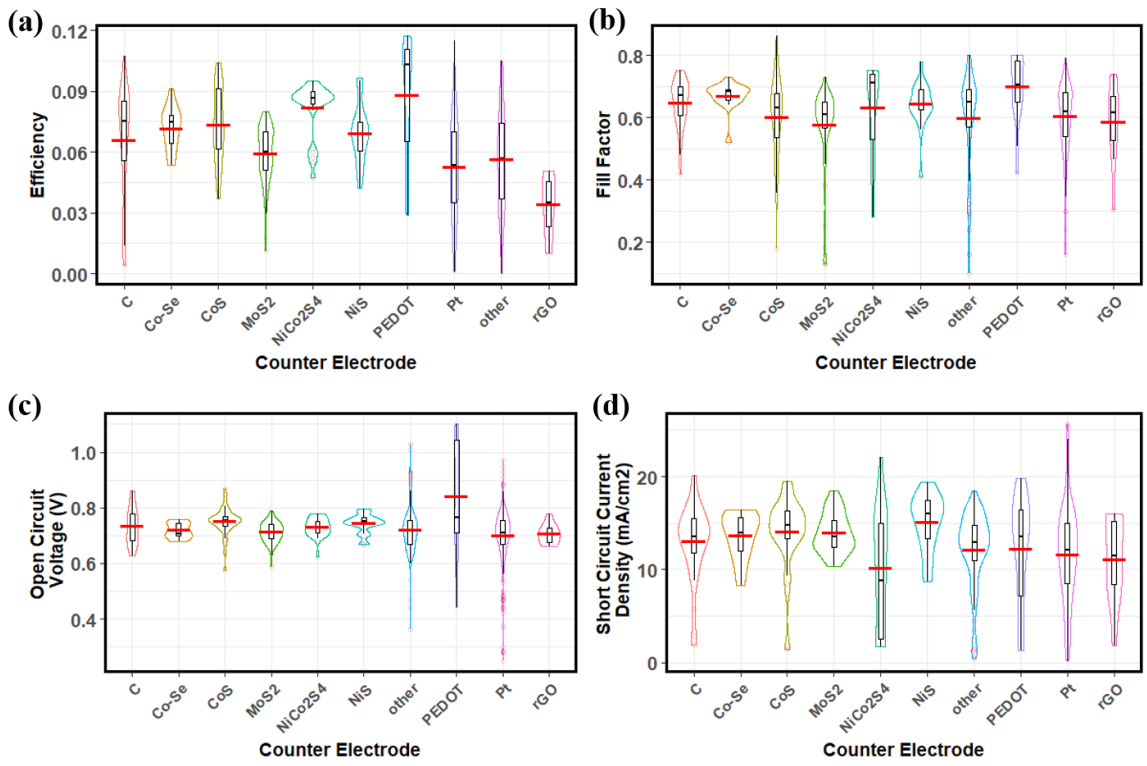

Figure 5.2. The effect on cathode type on a) PCE, (b) FF, (c) V_{OC} and (d) J_{sc} (red line represents the average values).

PEDOT and $NiCO_2S_4$ had higher efficiency compared to Pt, which is a common counter electrode in many electrochemical processes. Fill factor doesn't seem to be affected by the counter electrode type, and only PEDOT slightly increased open circuit voltage and $NiCO_2S_4$ decreased the short circuit current density compared to other types. Multicomponent sulfides are attractive counter-electrode materials due to their tunable properties due to having multiple cations. $NiCO_2S_4$ showed a high electrocatalytic reduction in triiodide reduction and high transmittance in the visible wavelength region, though its electron conductivity is low [148]. On the other hand, PEDOT provides multiple interfacial electron transfer pathways which enhances electron transfer when utilized as a counter electrode [149] but it lacks sufficient electrocatalytic active sites hence, the J_{sc} was not higher compared to other alternatives [150].

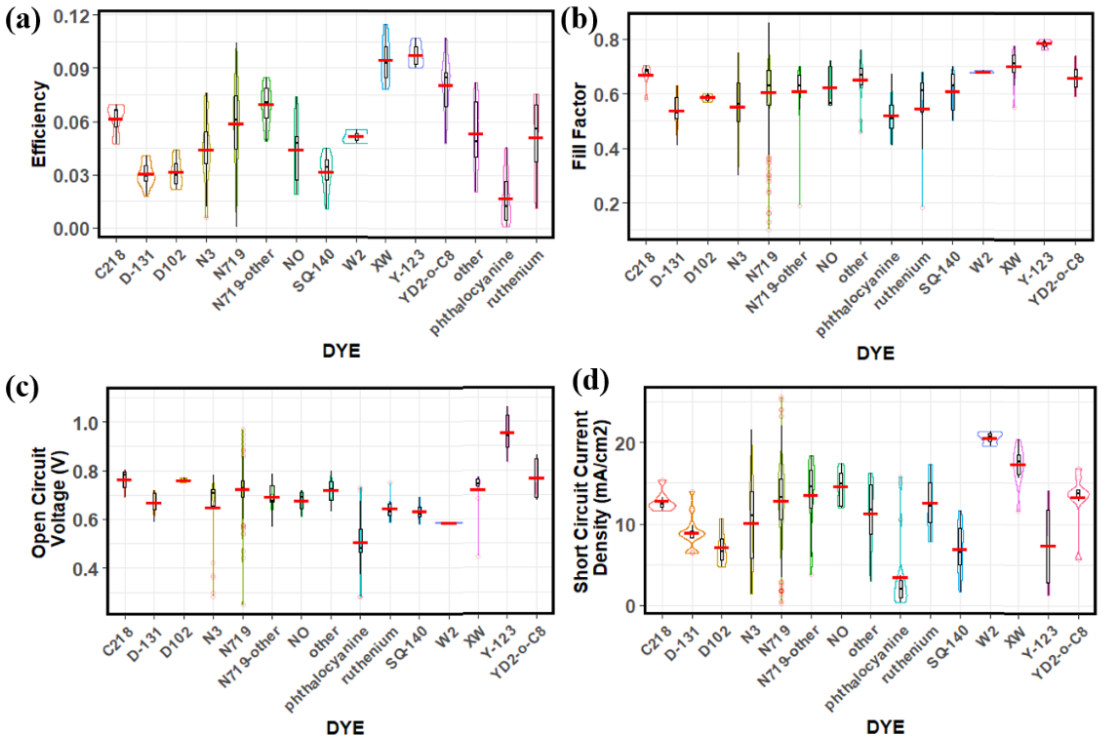

Figure 5.3. The effect on dye type on a) PCE, (b) FF, (c) V_{OC} and (d) J_{sc} (red line represents the average values).

Dye is the most important part of the DSSC since it is responsible for the absorption of the light and charge transfer. Since ruthenium is an expensive material, researchers have worked on developing alternative dyes containing porphyrin, carbazole or metal-free dyes [151].

The importance is also evident from the fluctuations in all the measurements about the DSSC performance (Figure 5.3), this kind of variation was not observed with other parameters. Dyes such as XW (porphyrin-based), W2 (triarylamine-based), Y-123 (metal-free) and Yd2-o-C8 (porphyrin-based) dyes showed remarkable performance compared to benchmark dye N719.

5.3.2. Performance Prediction of DSSCs

Two different ensemble algorithms, random forest and gradient boosting, performed better in the prediction of DSSC performance, which was expressed in terms of four measures (photoconversion efficiency, fill factor, open circuit voltage and short circuit current); the algorithm tested (linear regression, SVM, ...) did not perform well probably due to the fact that most of the descriptors were categorical. All models for performance prediction were powerful and variable importance was different for each performance metric (Figure 5.4-5.7).

PCE, the main performance metric of a DSSC, was best predicted by gradient boosting algorithm with RMSE values of 1.07, 0.81 and 1.55 for training, validation and testing respectively. The model predictions were quite successful as it can be also seen in Figure 5.4; the distribution of predicted vs observed data is quite close to the x=y line even though there is a tendency to overpredict low PCE values as apparent from all training, validation and testing plots sets (Figure 5.4). Variable importance plot (Figure 5.4d) suggests that dye type is the most important component of the DSSC followed by electrolyte solvent type and the method used for counter electrode fabrication. The main part of a DSSC is the dye and the PCE mostly depends on the dye's light absorption ability and its compatibility with TiO_2 surface. For instance,

Yan et.al. designed a triarylamine-based sensitizer and showed that the introduction of thiazole group improved the absorption ability and decreased the recombination of excited charges [152].The intermolecular π-stacking and aggregation of dyes on the photoanode surface causes lower efficiencies [151]. Optimization of the electrolyte is also required to have high efficiencies, for instance, solvents containing 4-tert-butylpyridine and ionic liquids are investigated for improving PCE [131].

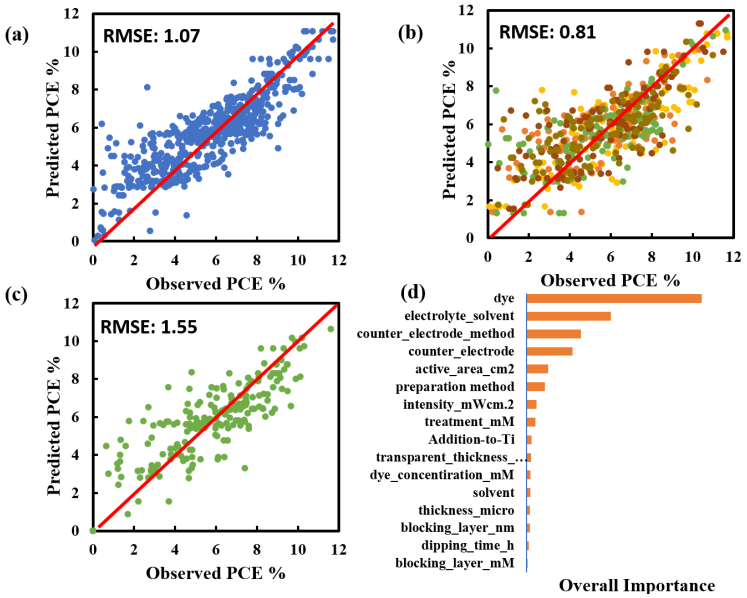

Figure 5.4. Gradient boosting results for PCE prediction (a) training, (b) validation, (c) testing results and (d) variable importance of model.

Fill factor was also predicted with good performance using random forest model and the overprediction of low values was observed here as in PCE prediction. The number of studies having fill factors below 0.5 are small therefore it is possible that the model could not effectively determine the conditions resulting in low fill factor values. Most probably, the overprediction in PCE arises from the same reason for FF as they are interdependent (Equation 4.1). There could be also an important descriptor, like electrolyte concentration, that would help the model to distinguish low FF but could not be used in the model due to insufficient data. It was reported that the concentration of electrolytes affects the ability of dye regeneration, causes dye aggregation and increases the recombination, hence the resistance [153]. The most important variable

was found to be the electrolyte solvent followed by the counter electrode and preparation method but all variables except light intensity had similar importance (Figure 5.5d). Since fill factor is the measure of the ideality of the constructed DSSC, it is reasonable that the parameters related to the fabrication of DSSC are all important while the experiment conditions i.e. light intensity are not as important in the prediction. The FF of the DSSC depends on the resistances within the DSSC and fabrication of the DSSC, choice of materials etc. plays an important role in the FF of the system. For instance, the recombination of photogenerated electrons and holes increases recombination resistances [154], and the distribution of the energy levels in the electrolyte has a direct influence on the recombination of charges at the electrolyte-anode interface [155]. Additionally, the counter electrode can also affect resistances; changing the counter electrode to PEDOT was reported to decrease charge transfer resistance, hence, increasing FF [156].

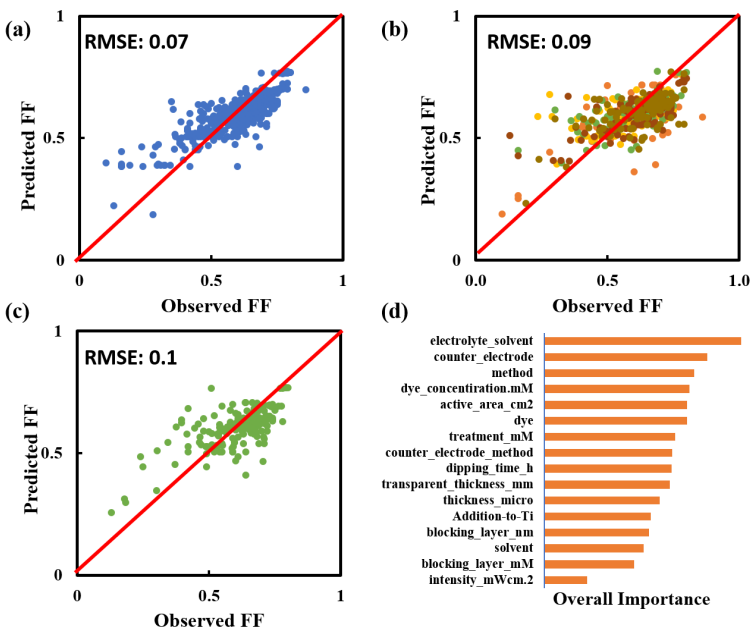

Figure 5.5. Random forest results for FF prediction (a) training, (b) validation, (c) testing results and (d) variable importance of model.

Open circuit voltage predictions were also successful as the previous two performance metrics with RMSE of 0.04, 0.07 and 0.05 for training, validation and testing respectively (Figure 5.6). The distribution of the predicted vs observed is very close to the x=y line for training and testing sets. There are some low V_{OC} values that were overpredicted by the model, this is probably due to the amount of low V_{OC} values being small as in fill factor prediction. Dye type and its concentration were the two most important variables, it was expected since the dye is the main contributor to the generation of photoexcited electrons upon illumination. The increased density of the excited electrons coming from dye changes the quasi-Fermi level of TiO_2 structure, and consequently, increases the difference between Fermi level of TiO_2 and the redox potential of the electrolyte [139].

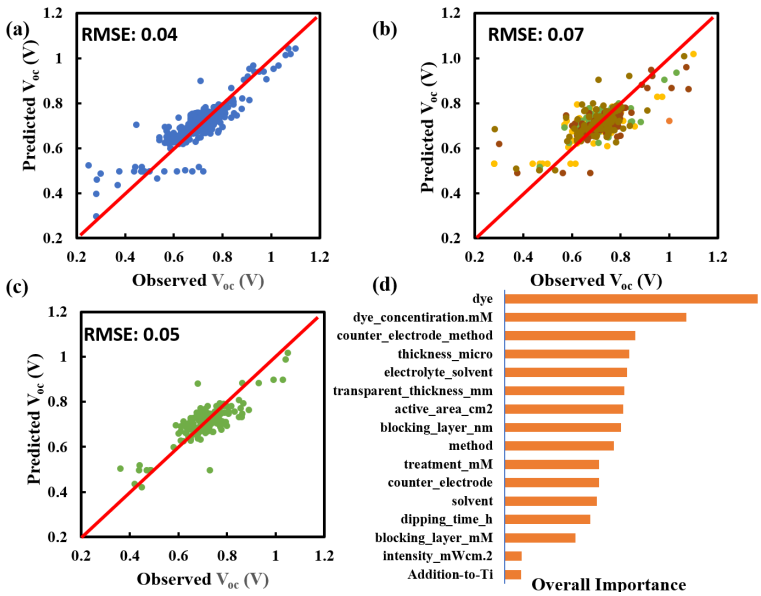

Figure 5.6. Random forest results for V_{OC} prediction (a) training, (b) validation, (c) testing results and (d) variable importance of model.

The prediction power of the gradient boosting model for short circuit current density was lower than the other three performance metrics but still sufficiently good. The RMSE values were 2.87, 3.65 and 3.85 for training, validation and testing respectively. The values are higher than the errors in other performance measures since the range

of J_{sc} values changes between 0 and 25 mA/cm^2 whereas, for example, PCE changes between 0 and 12 %; hence, the higher RMSE values for J_{sc} are reasonable. In addition, the non-standard reporting conditions for indoor DSSC testing might affect the reliability of the J_{sc} data [157]. J_{sc} is the measure of generated current per illuminated area; hence any discrepancy between the reporting of active area or the light intensity could directly influence the resulting J_{sc} value [158]. It was reported that the distance from the light source and utilization of AM 1.5G filter or any cutoff filter would change the light flux on the anode surface [159, 160].

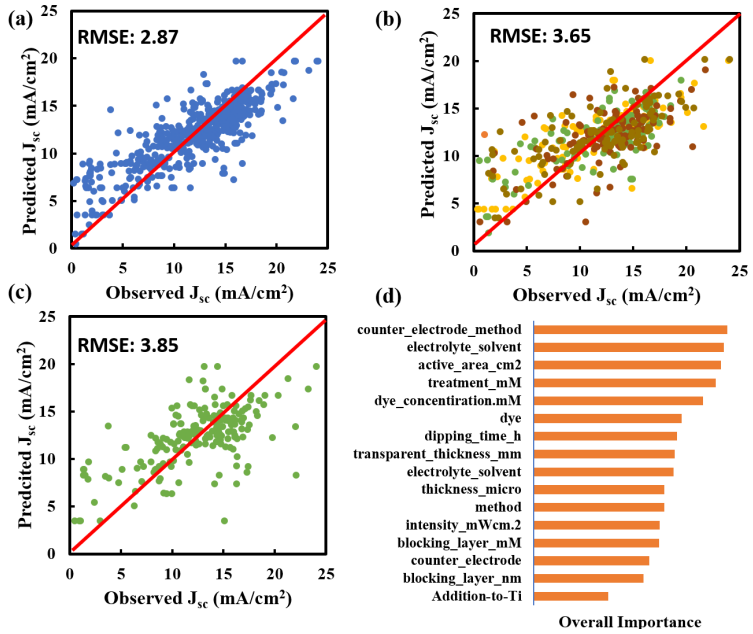

Figure 5.7. Gradient boosting results for J_{sc} prediction training, (b) validation, (c) testing results and (d) variable importance of model.

Variable importance plot (Figure 5.7d) suggests that the method used for counter electrode fabrication, electrolyte solvent type, active area of the DSSC and amount of $TiCl_4$ or TiPP treatment applied to the working electrode had similar importance while the remaining variables showed relatively lower (but similar to each other) importance except type of the material added to TiO_2 anode (represented as Addition-to-Ti), which seems to be insignificant. This could be related to TiO_2 being used without any addition in most of the studies (Figure 5.1), which hinders its effect of material addition

to TiO_2 anode on the overall model. The photocurrent generated in the DSSC depends on the redox reactions happening in the electrolyte, therefore the electrolyte properties influence the J_{sc} more; for instance, utilization of 4-tert-butylpyridine as an electrolyte additive decreased the J_{sc} of the DSSC due to low electron injection rate [161].

6. PREDICTION OF PERFORMANCE OF NATURAL DYE SENSITIZED SOLAR CELLS VIA MACHINE LEARNING

This work was performed in collaboration with the group of Assist. Prof. Hisham A. Maddah of King Abdulaziz University.

6.1. Background Information on Natural Dyes Sensitized Solar Cells

As briefly introduced in the previous section, a dye-sensitized solar cell consists of a photoactive anode (usually TiO_2) sensitized by a light-sensitive dye, which is responsible for light absorption and generation of electron-hole pairs, a cathode to close the circuit and an iodine electrolyte for charge transfer [162]. The sensitizer dye is an important part of the DSSCs and 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 compared to synthetic routes even though the stability and current efficiency ranges are challenges to overcome [163]. The natural dyes can be extracted from the roots, leaves and barks of the plants, and they can be generally classified into six categories: 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 in dye structure to ensure good electronic binding with TiO_2 [166]. Dye aggregation is another problem in DSSC lowering the photocurrent; in addition to modifications to dye chemistry [167], co-sensitization, which is also employed to increase the electron injection efficiency, may also ease the aggregation problem [164].

Nowadays, a large amount of data is generated in all fields of research including DSSCs, and the ability to use this data to extract knowledge is crucial for future

works. The search for new materials, new dyes or performance prediction of DSSCs have been studied widely in recent years and some databases in the field emerged. For example, Venkatraman et. al created a database for the properties of dye-sensitized solar cells from 4000 articles [165]. Venkatraman and Chellappan developed a database for dye aggregation and type of aggregation from 1500 articles [166]. Consequently, the prediction of the performance of DSSCs is possible with machine learning algorithms. Previously, Maddah investigated the natural dye-sensitized solar cells using decision trees with simple dye descriptors [167]. He used four dye properties (number of pi bonds, anchoring groups, molecular orbit levels and band gap) of 27 sensitizers to predict if the natural DSSC efficiency would exceed 1.82%. To best of our knowledge, there are no works focused on the prediction of natural DSSC performance using details about solar cell fabrication. In this work, we gathered literature data from the last 5 years for natural dye-sensitized solar cells and predicted the performance of the solar cells using random forest and gradient boosting algorithms.

6.2. Computational Details

6.2.1. Construction of the Dataset

A Web of Science (WOS) search was performed using the keywords "natural dye solar cell" between the years 2018-2023 to create a database for the most recent natural dye-synthesized solar cell applications. The most relevant 150 papers were screened and 522 data points from 113 of the papers, were selected as they provide information for photoelectrochemical testing (i.e. J-V measurements) and the details about the source and type of dye were used. Short Circuit Current density, Open Circuit Voltage, Fill Factor, and power conversion efficiency of the solar cells were collected as output variables whereas details about semiconductor, dye, and their preparation procedure, electrolyte and operation conditions were used as input variables. Additional details about dyes such as free electron and anchoring groups of the dye classes were added from the work of Maddah [167]; in the final dataset, there were 26 input variables (descriptors) and 4 output variables. The list of variables, with their ranges/levels, is

given in Table 6.1. The variables that appeared in low frequency in the dataset were combined in a group as others.

Table 6.1. Range and levels of all input variables in the dataset.

Variable	**Range/Category Level**
Material	other, TiO_2, TiO_2-composite, ZnO
Deposition Method	commercial, deposition, doctor blade, screen printing, spin coating
Deposition Thickness (µm)	0.17 - 500
Treatment ($TiCl_4$ etc.)	yes (1), no (0)
Calcination Temperature (°C)	25-550
Calcination Time (h)	0.08-5
Active cell area (cm^2)	0.02-10
Bang Gap (eV)	0.89-3.98
Dye absorption time (h)	0.03-72
Counter Electrode	Carbon, Pt, other
DYE	anthocyanin, betalain, carotenoids, chlorophyll, curcumin, mix, other
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)	0.003-25
Extraction Time (h)	0.08-168

Table 6.1. Range and levels of all input variables in the dataset. (cont.)

Variable	**Range/Category Level**
Co-sensitized Dye	anthocyanin, betalain, chlorophyll, curcumin, flavonoids, non-natural, none
Ratio of Dye/Co-sensitized Dye	0.5-4
Peak absorption Wavelength	242-800
Light Intensity (mW/cm^2)	3.5-1000
Salt	iodine, KI, Li, other, t_butyl_pyridine
Solvent	3-methoxypropionitrile, acetonitrile, acetonitrile mix, aqueous, ethylene glycol, organic
IL	yes (1), no (0)
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

6.2.2. Model Development

Pre-analysis of the dataset was performed first to see the structure of data using box and whisker, and swarmplots were generated using ggbeeswarm [168] package in R. Then, the regression analysis was employed for prediction of solar cell performance defined in terms of V_{OC}, J_{sc}, FF and PCE. Random forest and gradient boosting algorithms were used, and models were developed using randomForest [38] and gbm [110] packages in R respectively. Variable importance graphs were constructed using varImp function of caret [138] package in R. Each model was developed using 80-20% train-test split ratio and missing values were imputed by using mean values of the training set. Hyperparameters, which were chosen as the number of trees (ntree), and 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 5-fold cross validation. 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 the second structure (encoded), these variables are converted into numeric variables via one hot encoding. The best models of each output are given in Table 6.2.

Table 6.2. Best models developed for each output variable.

Output	**Feature Structure**	**Model and hyperparameters**
J_{sc}	Categoric	RF - ntree: 230 , 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

6.3. Results and Discussion

6.3.1. Pre-analysis of the Dataset

The data distribution was presented for each output variable using box and whisker plots in Figure 6.1. Since the outliers should be discarded to prevent their negative influence on the model performance [169]; the thresholds values of 1 V, 10 mA/cm^2 and 4% for open circuit V_{OC}, J_{sc} and PCE; the points above these limits were removed from the dataset. No threshold value was selected for fill factor because no outlier (data far from the main group) was identified.

Swarm plots for the input variables for each output were also plotted to see the general trends in the variables and their relationship with output variables and to understand the data structure better for the model development.

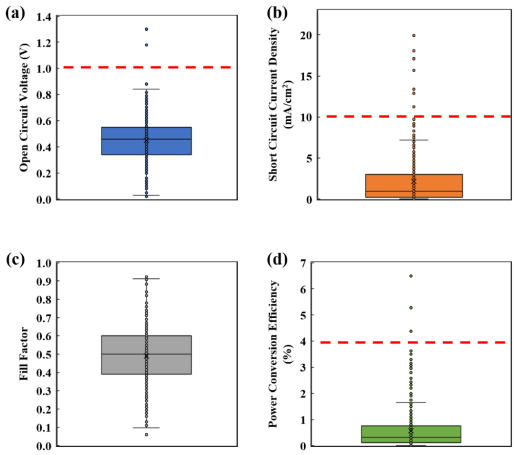

Figure 6.1. Box plot for outputs (a) V_{OC}, (b) J_{sc}, (c) FF and (d) PCE.

TiO_2 is the most commonly used semiconductor for anode in our dataset as expected; other semiconductors like ZnO, PbS and composites of TiO_2 are also utilized but they have not shown better performance than TiO_2 in any of four target variables (V_{OC}, J_{sc}, FF and PCE) used as the performance measure (Figure 6.2a).

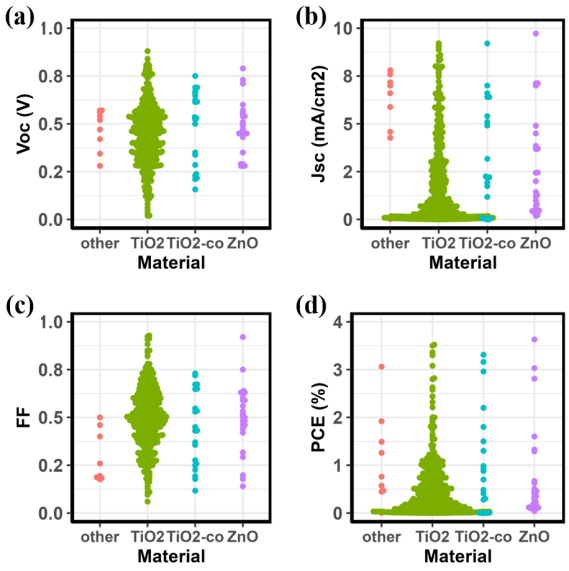

Figure 6.2. The distribution of anode materials for (a) V_{OC}, (b) J_{sc}, (c) FF and (d) PCE.

In order to increase the performance of TiO_2 photoanode, some researchers employed surface treatment using $TiCl_4$ or titanium isopropoxide. These seem to have

some positive effects on the performance, especially on FF (Figure 6.3); the data were accumulated in higher-performance regions (providing higher average performance) while this trend is more obvious for FF. This could arise from the fact that the resistances in the photoanode decreased upon treatment (i.e. the recombination of excited charge carriers decreased [170]).

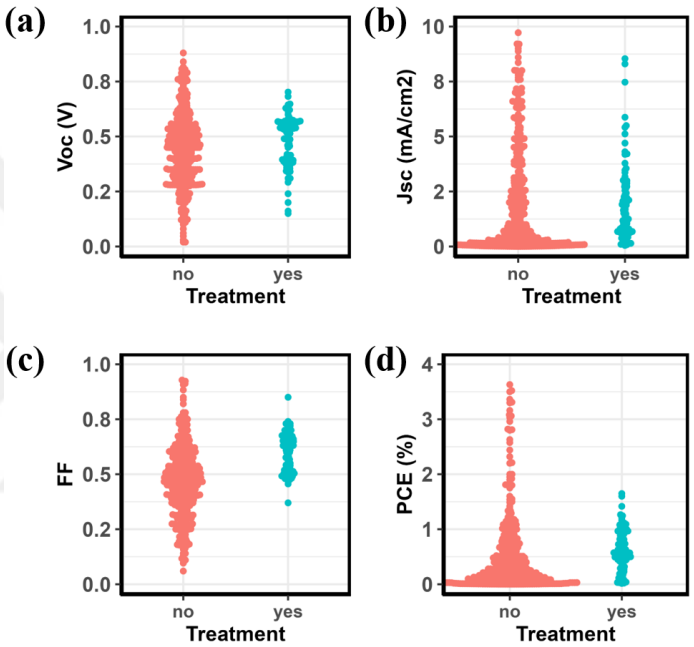

Figure 6.3. The distribution of using treatment for (a) V_{OC}, (b) J_{sc}, (c) FF and (d) PCE.

The choice of dye affects the performance of the natural dye-sensitized solar cells since they are the main component responsible for light absorption. The effect of natural dye type on the DSSC performance metrics is given in Figure 6.4.

The V_{OC} normally depends on the photoanode and dye [171]. However, no remarkable difference was observed between different dye types; this could be due to the fact that the majority of the studies used TiO_2 based photoanodes. On the other hand, J_{sc} can be affected by the light absorption ability of the dye [151]; however, the accumulation of the data was close for all dyes, hence the effects of dye type on the J_{sc} was not observed. As explained in the previous chapter, there is no standard reporting

for DSSC therefore the effect of dye type on the J_{sc} might be hindered by fluctuations in other variables such as light intensity and active area.

Curcumin dye shows slightly improved FF (Figure 6.4c). As discussed in the previous chapter, FF mostly depends on the resistances in the system [154]; therefore, this could mean that the charge transfer or shunt resistances are lower when curcumin dye is used compared to other dye types, however, no literature suggests this argument was found hence the difference is probably related to other variables which would affect FF more, such as improved photoanode or electrolyte coupled with curcumin dye.

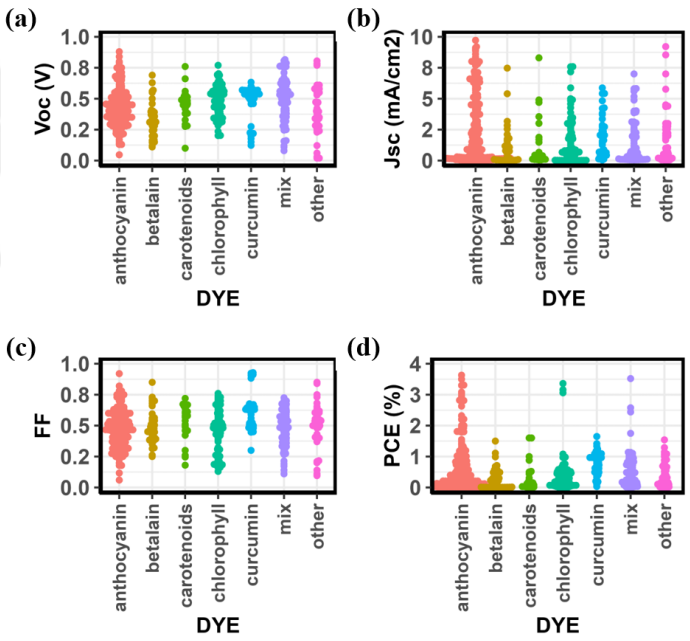

Figure 6.4. The distribution of dyes for (a) V_{OC}, (b) J_{sc}, (c) FF and (d) PCE.

Researchers have also tested the cocktails of dyes to increase the overall performance of the natural dye-sensitized solar cells. In our dataset, anthocyanin is most frequently employed co-sensitization dye even though using it as a co-sensitizer does not give better results compared to not co-sensitization (Figure 6.5); although some other dyes such as flavonoids, seem to have higher average J_{sc} and PCE, the number of cases containing these dyes are quite low for reliable results.

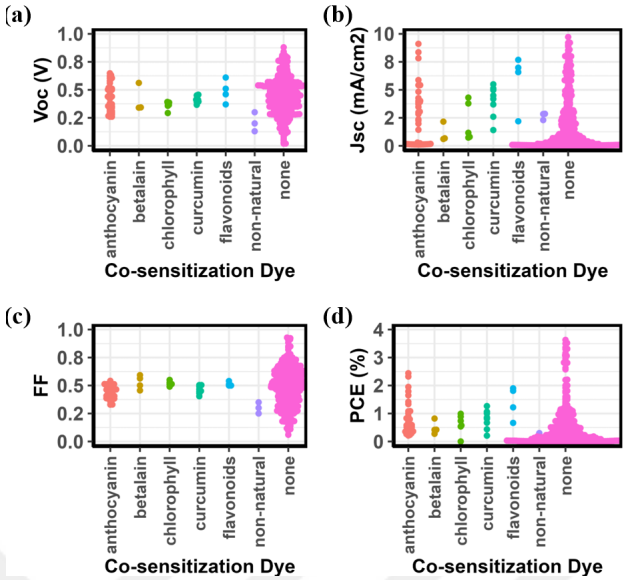

Figure 6.5. The distribution of Co-sensitization dye data for (a) V_{OC}, (b) J_{sc}, (c) FF and (d) PCE.

The reason behind the low improvements could be related to the ratios of those cocktail dyes; the ratio of two dyes affects the output variables and the dye ratio around 2.5 has improved all outputs, including efficiency. This shows the co-sensitization dye should not be used in very small or large amounts compared to the main dye, the effect of the co-sensitization dye might not be apparent in low amounts (ratio above 3) and when large amounts are used (ratio below 2), the two dyes might not be compatible or competitive light absorption might decrease the performance [172, 173].

The presence of ionic liquid in electrolytes promotes charge transfer within the solar cell as seen in Figure 6.7 which indicates the distribution of data with and without the use of ionic liquid. Even though the efficiency is lower when the ionic liquid is present, V_{OC} and FF are higher; it seems that ionic liquid contributes to the improvement of voltage generated by the cell. As V_{OC} is related to the difference between the Fermi level of TiO_2 and the redox potential of the electrolyte [139], the addition of ionic liquid in the electrolyte could have increased that difference and hence increased V_{OC}.

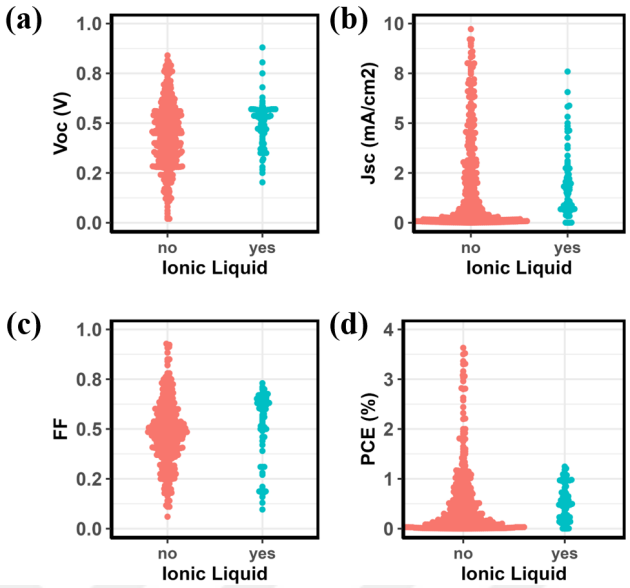

Figure 6.6. The distribution of data using or not using ionic liquid for (a) V_{OC}, (b) J_{sc}, (c) FF and (d) PCE.

6.3.2. Prediction of Open Circuit Voltage

Upon elimination of outliers and data with missing target variables, the dataset for V_{OC} has decreased to 512 data points. The GB model with encoded data, and ntree of 290, shrinkage of 0.15 and int_depth of 8 gave the lowest root mean square error 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 0.002, 0.09 and 0.07 respectively) are given in Figure 6.7; the model predictions, especially for testing data, which are 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.

The variable importance plot (Figure 6.7d) suggests that dye extraction variables are the most influential variables in the prediction of V_{OC}, followed by the deposition method of the anode. Indeed, the voltage of the solar cell depends on the amount of active dye extracted from the solvent during the cell preparation. The RMSE of GB model for the categorically structured data and RF models is given in Appendix B.

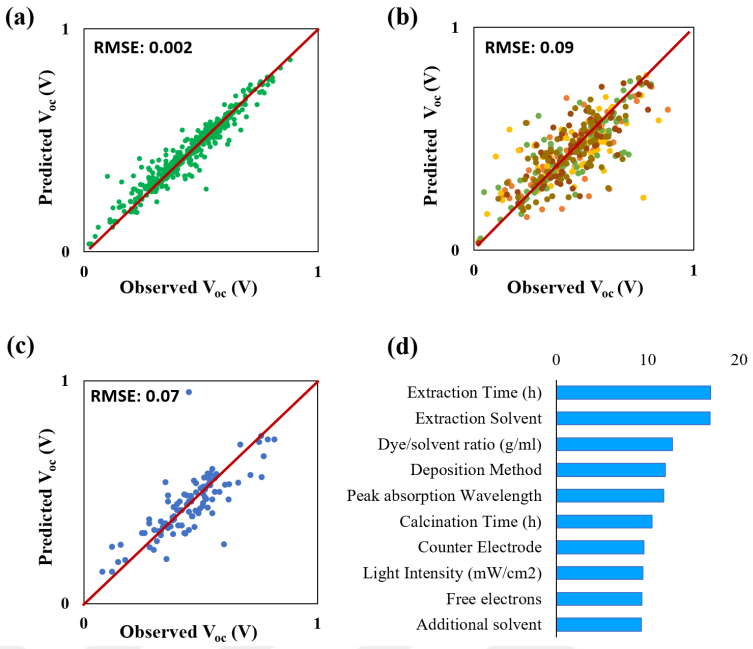

Figure 6.7. The GB model for V_{OC} prediction (a) training (b) validation (c) testing and (d) variable importance of top ten variables.

6.3.3. Prediction of Short Circuit Current Density

The predictions for J_{sc} were better with categorically structured data (500 data points) while the RF algorithms, which are highly suitable for categorical data, performed better for J_{sc}; the optimum values of model hyperparameters, ntree and mtry, were found to be 230 and 14 respectively. The actual versus predicted J_{sc} for training, validation and testing are presented in Figure 6.9 with the RMSE of 0.5, 1.3, and 1.5 respectively. Although the model predictions are not as accurate as the V_{OC} predictions, they are still reasonably well; the dye/solvent ratio, band gap of the anode and active cell area were found to be the most important variables for J_{sc} prediction (Figure 6.8d). This is an expected result considering that the current generated depends heavily on the light absorption properties of the dye and the properties of anode material.

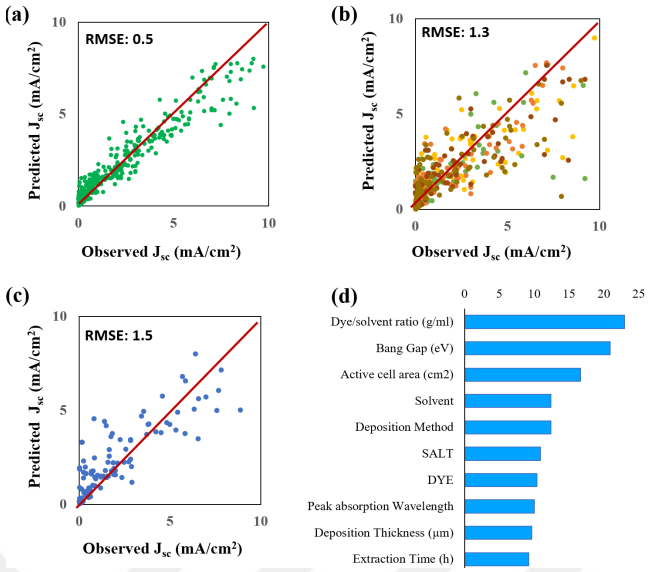

Figure 6.8. The RF model for J_{sc} prediction (a) training (b) validation (c) testing and (d) variable importance of top ten variables.

6.3.4. Prediction of Fill Factor

Fill Factor was predicted with high accuracy if the encoded structure of data, containing 513 cases, and random forest as the ML algorithm was used; the hyperparameters of the optimum model were 150 and 6 for ntree and mtry respectively. As can be seen from Figure 6.9, the model predictions are quite successful with the RMSE of 0.04, 0.1, and 0.08 for training, validation and testing sets respectively. The most important input variables were dye/solvent ratio, dye extraction time, and dye absorption wavelength. As different from other output predictions, this time anode treatment, as well as electrolyte solvent, also appeared to be important.

The mobility-lifetime of the material was found to affect FF [140], which is a term used to evaluate the performance of semiconductors. It is calculated by multiplying the quantum efficiency with the mobility and lifetime of charge carriers. High mobility lifetime shows that semiconductors can transfer charges efficiently with a long lifetime.

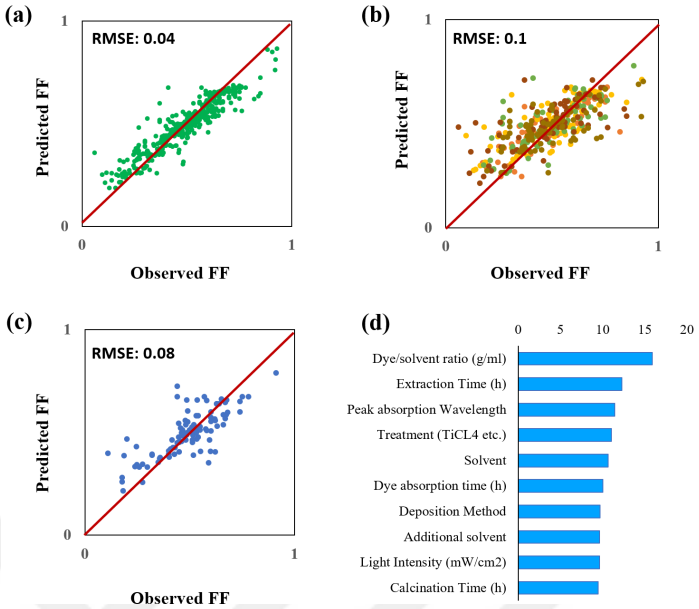

Figure 6.9. The RF model for FF prediction (a) training (b) validation (c) testing and (d) variable importance of top ten variables.

The ability to extract charges and their transfer rates are important in the determination of FF of the DSSC and treatment of the photoanode and optimizing the electrolyte solvent could result in lower charge recombination. As explained in the previous chapter, decreased recombination of charges would also decrease the resistances due to recombination and hence improve FF of a DSSC [154].

6.3.5. Prediction of Efficiency

Finally, the efficiency, which is the most important performance measure of a solar cell as it directly measures the fraction of light energy converted to electricity, is modelled. Encoded data structure and GB model with ntree of 410, shrinkage of 0.15 and int_depth of 7 was found to be the best for the efficiency predictions. The optimum model resulted in the RMSE of 0.03, 0.4 and 0.2 for training, validation and testing respectively; again, the model predictions are reasonably good as presented in Figure 6.10.

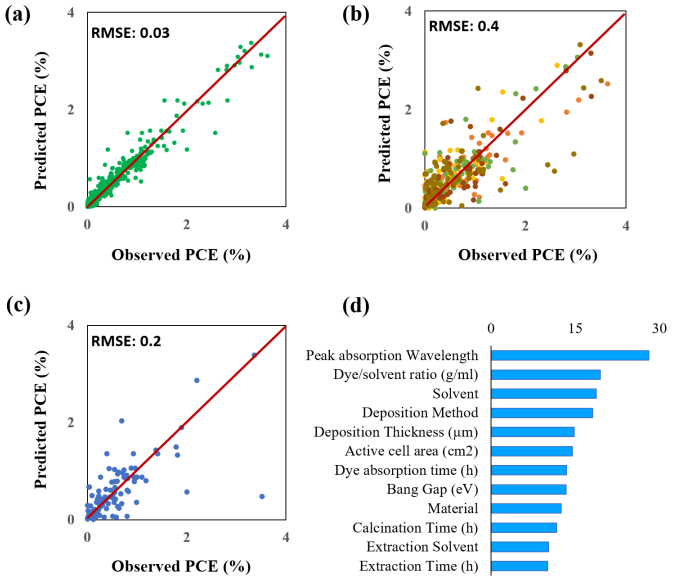

Figure 6.10. The GB model for PCE prediction (a) training (b) validation (c) testing and (d) variable importance of top ten variables.

The most important variables were the peak absorption wavelength of the dye, (equivalent to the dye band gap) dye/solvent ratio followed by electrolyte solvent, deposition method and thickness. It is reasonable that these variables affect the prediction of efficiency more compared to other variables because they contribute to light absorption and electron transport in the cell.

7. MACHINE LEARNING ANALYSIS OF PHOTOELECTROCHEMICAL WATER SPLITTING

The data collection for this work was conducted with the help of Dr. Elif Can, who is a former PhD student in our group. This work was published in International Journal of Hydrogen Energy on January 2022 [134].

7.1. Background Information on Photoelectrochemical Water Splitting

Due to the increase in energy demand and environmental pollution, clean and sustainable energy sources have been researched extensively in recent years [174–177]. Hydrogen is considered as a potential energy carrier for the future since it is environmentally friendly and can be used to store and transport solar energy as easy-to-use chemical energy [178, 179] while it is also an important raw material for industry. Today, the major production method of hydrogen is hydrocarbon reforming [180] resulting in carbon dioxide emissions, which is the main cause of global warming. Hence, alternative and cleaner ways of hydrogen production are explored, and one of these methods investigated extensively is hydrogen production from water via photoelectrochemical (PEC) processes [181, 182]. TiO_2 is the most commonly employed semiconductor for water splitting while the other semiconductors such as WO_3, CdS, Ta_2O_5 and $BiVO_4$ have been also extensively used due to their high activity and stability under solar irradiation [183, 184].

An optimum semiconductor should have suitable band edge positions to harvest visible light and convert it into chemical energy [185]. However, most of the major semiconductors are active under UV light, which is only a small portion (4%) of the solar irradiation. For example, TiO_2 has a band gap of 3.2 eV which limits its activity under visible light, hence, it should be modified by element doping into its crystal structure [186] or the use of co-catalyst loading [12] to enhance its activity under solar light illumination via better charge separation. Mostly noble metals are used

as co-catalysts but it is also possible to load less expensive materials such as CdS, Cu_2O, NiO and even perovskites [187]. The use of bilayer or composite materials may also increase the charge transfer rates by providing two-step photo-excitation [188]. Another limitation of photoelectrochemical reactions is the high possibility of the recombination of photogenerated electron-hole pairs; the materials should have fast charge transfer ability with low recombination rate. TiO_2, ZnO and WO_3 are widely studied because they have good carrier mobility but their band gap sizes and alignments require modification [187]. Fe_2O_3 and Cu_2O have suitable band gap but they have fast recombination which limits the charge transfer [185]. In recent years, carbon materials such as graphene oxide, carbon nanotubes and graphitic carbon nitride(g-C_3N_4) have been also used as semiconductors, co-catalysts and sensitizers to increase light absorption and charge transfer rate [189, 190]. Another recent development in the photoelectrochemical processes is the use of easily tunable nanomaterials [191, 192]; such as quantum dots (QDs) [193], nanowire or rod formation and nanolayer heterostructures [185].

Investigation of new photocathode materials is also important for the commercialization of PEC water-splitting systems. Silicon is an attractive material for high charge transfer rates and suitable band edge positions [194]; Cu-based materials are commonly employed as photocathodes but the challenges in protection of these materials hinder the advantages such as low cost and high photoactivity [195]. Sulphides and selenides are other groups of investigated materials due to their tunability of band gap and earth abundance [196]. Besides those, nitride-, phosphide-, arsenide- and iron-based compounds are also being researched as photocathode materials [197].

Photoelectrochemical water splitting is an attractive way of producing large-scale pure hydrogen with a theoretical efficiency of 30.7% [198, 199], but current efficiencies are well below 10% in practical applications [200]. As mentioned above, the semiconductor materials should have good chemical stability, low cost, rapid charge transfer without recombination and suitable band gap for an effective overall water-splitting process, without the need for an external bias. The band edge positions are also criti-

cal to executing photoelectrochemical water splitting; the conduction band should be more negative than the reduction potential (0 V vs NHE) whereas the valence band should be more positive than the oxidation potential (1.23 V vs NHE) of water [201]. If the band edges are not appropriate, an external bias should be applied to split water [202]. As mentioned above, since visible light is a large portion of the solar spectrum, the photoelectrodes should be also able to utilize visible light (i.e. have a low band gap [203, 204]. The technology is still at its infancy period; hence, a large number of studies have been performed to find a way to optimize large number of factors related to photoelectrodes and electrolytes; this creates a massive amount data in literature that can be used to extract knowledge by machine learning and utilize this knowledge for the future experimental studies.

ML have been practised in various fields of science including catalytic and photocatalytic reactions [205]. We applied ML on various systems such as biofuels [206], [207], solar cells [133, 208] and photocatalytic systems [135]. It was also used by various researchers to analyze photocatalytic [209, 210], electrochemical [206,207], [211, 212] and photoelectrochemical [213] reactions including CO_2 reduction [214, 215]. However, as far as we know, there is no ML work analyzing the experimental data reported in literature for photoelectrochemical water splitting.

In this work, we constructed a comprehensive dataset for photoelectrochemical water splitting from 180 articles (10560 data points from 584 experiments) published in the literature and analyzed using machine learning tools. We started with simple descriptive statistics to understand the data structure and review the subject in detail. Then we used association rule mining to see the relations between the system features and performance measures (band gap or photocurrent density), and developed predictive models for both band gap or photocurrent density using various regression algorithms such as linear regression, lasso regression, SVM, RF, etc. Finally, decision tree analysis was used to deduce rules leading desired level of band gap and high photocurrent density.

7.2. Computational Details

7.2.1. Dataset Construction

The dataset was constructed by scanning hundreds of scientific papers on photoelectrochemical water splitting published between 2007 and 2020 at various online sources including Elsevier, American Chemical Society, Royal Society of Chemistry, Wiley Online Library and Springer. After a detailed review (starting from the top hits in relevance sorting), 10560 data points (584 experiments) from 180 experimental articles were included in the dataset.

The materials and methods used in the manufacturing of electrodes, the details about the electrolyte solution and the properties of the light source were used as input variables (descriptors) while the band gap (eV) of the working electrode and photocurrent density (mA/cm^2) obtained were defined as the output variables. The band gap was also used as input variable in the prediction of photocurrent density. The complete list of input and output variables with their ranges (for numerical variables) or sub-categories (for categorical variables) is represented in Table 7.1.

Table 7.1. The details for input and output variables.

Variables	**Ranges / Levels of variables**
First Layer	$BiVO_4$, Bi_2WO_6, CdS, Cd_2SnO_4, composite, $CuWO_4$, Fe_2O_3, g-C_3N_4, GaN, Ga_2O_3, InO_2, $LaFeO_3$, $LaTaON_2$, MoS_2, $NaNbO_3$, Nb_2O_5, other, $PbTiO_3$, Sb_2O_3, Si, SiO_2, SnO_2, $SrTiO_3$, Ta_3N_5, Ta_2O_5, TiO_2, VO_2, WO_3, $YFeO_3$, ZnCuInS, ZnO
First Layer Doping	B, C, Cl, Co, Cr, Cs, double dop, Fe, K, Li, Mn, Mo, N, Nb, other, P, Pt, Sb, Si, Sn, Ta, Ti, undoped, W
Co-catalyst	Au, CdS, CePi, Co(OH), CoPi, Ni/NiO, NiO, other, Pt, RuO, not provided

Table 7.1. The details for input and output variables. (cont.)

Variables	**Ranges / Levels of variables**
Second Layer	BiOCl, BiOI, BiS, $BiVO_4$, CdSe, CuInS, CuO, Er_2O_3, Fe_2O_3, FeOOH, g-C_3N_4, InGaN, Co-$LaFeO_3$, $LaFeO_3$, NiOOH, other, rGO, SnO_2, TiO_2, ZnO, ZnSe, not provided
Second Layer Doping	Co, Ga, Mo, undoped, Ti
Third layer	Au, $BiVO_4$, Carbon Quantum Dots, FeOOH/NiOOH, other, not provided
First Layer Synthesis Method	Al reduction, anodization, chemical deposition, chemical growth, commercial, electrodeposition, etching, flame Oxidation, hydrothermal, hydrolysis, molecular beam epitaxy, metal organic deposition, other, polymerization, pyrolysis, sol-gel, solvothermal, spin coating, spray coating, sputtering, solid state reaction, vapor deposition, not provided
Calcination Temperature (First Layer)	25-1300 Celcius
Calcination Time (First Layer)	0-30 hours
Calcination Atmosphere (First Layer)	air, Ar/H_2, H_2, N_2, NH_3, O_2, other, not provided
Second Layer Synthesis Method	Atomic layer deposition, chemical deposition, commercial, dip coating, electrodeposition, hummers method, hydrothermal, impregnation, other, photo deposition, photo reduction, sol-gel, SILAR, solvothermal, spin coating, sputtering, not provided
Calcination Temperature (Second Layer)	25-623 °C
Calcination Time (Second Layer)	0-4 hours
Calcination Atmosphere (Second Layer)	air, N_2, not provided
Third Layer Synthesis Method	chemDep, chemRed, electrodep, other, photoRed, solvoT, not provided
Counter Electrode	Pt, Pt-Ti, Pt coil, Pt foil, Pt gauze, Pt plate, Pt wire, not provided

Table 7.1. The details for input and output variables. (cont.)

Variables	**Ranges / Levels of variables**
Reference Electrode	Ag/AgCl, Hg/HgO, RHE, SCE, not provided
WE area	0.014-14 cm^2
CE area	0.005-15 cm^2
Substrate	AZO, FTO, ITO, Nb, other, Pt/Ti, Si, Stainless steel, Ta, Ti, W, Zn
Coating Method (First Layer)	deposition, dip coating, doctor blade, drop casting, electrodep in-situ, spin coating, spray coating, sputtering
Coating Method (Second Layer)	dep, dip coating, drop casting, electrodeposition, imp, in-situ, other, spin coating, not provided
Coating Method (Third Layer)	Drop casting, in-situ, spin coating, not provided
Last Calcination Temperature	25-900 C
Last Calcination Time	0-24 hours
Last Calcination Atmosphere	Air, Ar, N_2, NH_3, NP, O_2
Light Type	Hg, Laser, Led, W, Xe, not provided
Wavelength (nm)	299-520
Light Intensity (mW/cm^2)	1-1000
AM 1.5G Filter	Yes or No
Electrolyte Type & Molarity (mol/L)	$B_4K_2O_7$, borate Buffer, H_2SO_4, H_3PO_4, HBr, $HClO_4$, K_2HPO_4, K_3PO_4, KBr, KH_2PO_4, $KHCO_3$, KOH, Na_2S, Na_2SO_3, Na_2SO_4, NaCl, $NaClO_4$, $NaNO_3$, NaOH, NaS, other, phosphate Buffer & 0.01 - 5
Additive & Molarity (mol/L)	C_2H_5OH, $C_2H_6O_2$, CH_3OH, H_2O_2, H_3PO_4, Na_2SO_3, Na_2SO_4, NaOH, other, PBS, phosphate Buffer, none & 0 - 5
pH	0 - 14
Crystal. Structure first layer	anatase, cubic, hexagonal, monoclinic, orthorhombic, rhombohedral, rutile, tetragonal, triclinic, not provided

Table 7.1. The details for input and output variables. (cont.)

Variables	**Ranges / Levels of variables**
Crystal. Structure second layer	Amorphous, anatase, cubic, hexagonal, monoclinic, rhombohedral, tetragonal, not provided
Crystal. Structure third layer	Hexagonal, monoclinic, rutile, not provided
Bias (V vs RHE)	-0.7 - +3.1
Band Gap (eV)	1.16 - 4.37
Photocurrent Density (mA/cm^2)	0 - 47.9

7.2.2. Model Development

Association rule mining was used to identify the influential input variables on the band gap of the working electrode and high photocurrent density. To apply ARM algorithm, the numerical variables were discretized in ten levels (see Appendix C for details). For the band gap analysis, the data was divided into three classes containing approximately equal numbers of instances, as low, medium and high band gap as shown in Table 7.2. The ARM analysis of photocurrent density was performed for three bias levels separately because the most suitable conditions for the high photocurrent density are highly dependent on the bias; the conditions suitable for (desirable) low bias operations would be overshadowed by high photocurrent densities obtained at high bias if the entire dataset was used in the analysis. The low bias subset contained the data obtained at voltage values lower than and equal to 0.5 V vs NHE while the medium bias corresponded to the voltage values higher than 0.5 V and lower than and equal to 1.23 V vs NHE; the high bias data contains the data obtained at higher than 1.23 V vs NHE. Then, the data was further divided into three classes as low, medium, and high photocurrent density classes to associate the descriptor values for high photocurrent densities. The details of class structure are presented in Table 7.3. The arules [216] package of R was utilized for ARM application in the R environment.

Table 7.2. The class structure used in the classification of band gap.

Name of Class	**Lower Bound** ($\leq$)	**Upper Bound** ($>$)	**Number of Instances**
High	3.00 eV	4.37 eV	135
Medium	2.40 eV	3.00 eV	138
Low	1.16 eV	2.40 eV	128

Various regression algorithms were used for the prediction of band gap and photocurrent density; tree-based ensemble algorithms resulted in good predictive power for band gap, however a regression model with sufficient predictive power could not be achieved for photocurrent density. The e1071 [137], randomForest [38] and gbm [110] packages of R were used for regression algorithms. The data collected from the literature have some missing band gap values (78 out of 479 non-duplicated instances), which were also predicted by using the RF model developed to complete the dataset for the analysis of photocurrent density. To validate the model, k-fold cross-validation (k-fold CV) was applied using the 401 instances with known band gap values by changing n as 3, 5 and 10; the dataset was divided into three parts as testing, training, and validation data; 20%, 25% and 40% of data were randomly separated first as testing subset. The remaining data were divided into k (3, 5, and 10) subsets. A wide range of ntree and mtry were scanned, and the results were evaluated to find the model with minimum complexity. Then the optimum model, which has the lowest root mean square error for validation, was selected, and tested with testing data; the RMSE for testing was used to evaluate the prediction ability of the model. The optimum model was found to be with ntree of 220 and mtry of 15 in 25% test split and 5-fold cross-validation.

The decision tree analysis was applied to extract heuristic rules indicating the input variables to be adjusted to obtain the desired band gap and high photocurrent density. The data for photocurrent density was not divided into subsets based on bias as we did in ARM analysis because they were not sufficient to construct reliable decision tree models; instead, entire dataset was used in DT classification. The output

variables (band gap and photocurrent density) were divided into three classes in a way that all classes have approximately equal number of instances to avoid class imbalance problem [217]. The class structures for band gap and photocurrent density are given in Table 7.2 and Table 7.3. The limits of the classes were determined using the distribution of the training set output variables. DT analysis was performed using rpart [218] package of R for modelling and visualization, respectively. The complexity parameter was chosen as the hyperparameter to be optimized; the decision trees with the highest accuracy and minimum complexity were selected as the model representing the data best. This is done by changing the complexity parameter under various test/train set ratio (20/80 to 30/70) using k-fold cross-validation (k is changed as 3, 5, 7 and 10). The best model was obtained when 25% of the data was taken as the test set and 5-fold CV was applied.

Table 7.3. The class structure used in classification of photocurrent density.

Name of Class	**Lower Bound** ($\leq$)	**Upper Bound** ($>$)	**Number of Instances**
A	1.23 mA/cm^2	47.9 mA/cm^2	3521
B	0.80 mA/cm^2	1.23 mA/cm^2	3382
C	0.00	0.80 mA/cm^2	3656

In the DT analysis of photocurrent density, all photocurrent density values obtained in the same experiment using different biases were placed in the same subset (into the testing or one of the five folds in training) to prevent data leakage (because they are the same materials).

7.3. Results and Discussion

7.3.1. Pre-analysis of the Dataset

The steps in the photoelectrochemical water splitting (PECWS) can be summarized as i) absorption of the energy from the light source by the semiconductor, ii)

generation and separation of photoexcited electron and holes, iii) reduction of water to H_2 by electrons and oxidation of water to O_2 by holes [201]. Theoretically 1.23 eV is needed to split water, but energy losses in practical applications increase the required energy up to 1.8 eV [219]. There are many efforts for increasing the efficiency (photocurrent) of photoelectrochemical water splitting by focusing on the materials, electrode fabrication procedures, light properties, and reaction solution, we will briefly review those efforts made in the literature in terms of major variables before the detailed ML analysis.

Clearly, the light-sensitive materials (i.e. semiconductors) should be used as electrode materials in photoelectrochemical water splitting; the position of the valence and the conduction bands of the semiconductor determines whether it can be used in reduction or oxidization reactions [220]. Semiconductors can be classified as p-type or n-type; p-type semiconductors are hole-abundant and generally suitable for the reduction of water (hydrogen generation) whereas electron-abundant n-type semiconductors are suitable for oxidation (oxygen generation) [221]. In our dataset, the studies on n-type materials (photoanodes) were collected and used because the reaction taking place on the electrode surface is different in n and p-type materials and the former one is more common in the literature since oxygen generation is the rate-limiting step and it is more open to development To enhance the water splitting performance of photoelectrodes, ion-doping, the use of co-catalyst, nanoparticle/quantum dot sensitization, heterojunction structures and thermal treatments are some methods which enable a change in the band gap, electron transport properties, crystal structure or band edge structure of the semiconductors [222]. Distribution of the semiconductors is given in Figure 7.1. There are total of 30 semiconductor types that were used as first layer of the photoanode in the dataset (most of the times this is also single layer). At 1 V vs NHE bias, which is the most frequent bias value in our dataset, the highest photocurrent density was observed for GaN photoanode followed by SnO_2 photoanode. Most frequently studied semiconductors are TiO_2, Fe_2O_3, ZnO, $BiVO_4$ and ZnO respectively. Among these semiconductors, TiO_2 and ZnO based photoanodes exhibited highest average photocurrent density at 1 V vs NHE bias.

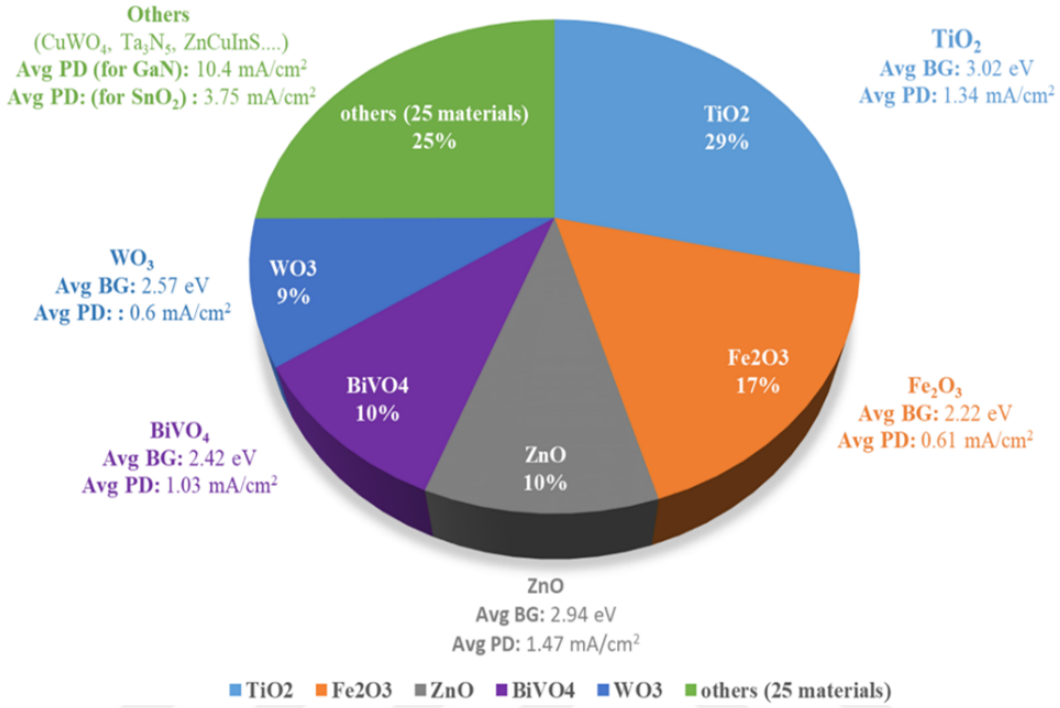

Figure 7.1. Distribution of semiconductors in the dataset with their average band gap (Avg BG) and photocurrent density (Avg PD) values.

Metal oxides are the most commonly studied photoanode materials due to their high chemical stability and low cost [223–225]. Aside being the first semiconductor to split water photoelectrochemically, TiO_2 has been the most frequently studied material for the photoanode throughout history. In our dataset 47 different articles, around 29% of data (167 instances out of 584) studied TiO_2. It is environmentally friendly, abundant, and chemically stable but its large band gap (around 3.2 eV) makes it UV active and the band edge structure of TiO_2 is not suitable for both hydrogen and oxygen evolution; therefore, it requires an external bias to perform overall water splitting [226]. In order to increase the activity of TiO_2 for photoelectrochemical water splitting; doping, use of co-catalyst, sensitization, thermal treatment, composite structures and layer structures have been studied. In our dataset, TiO_2 is mostly doped with carbon that lowered the band gap to 2.7 eV and a mid-gap around 1.6 eV [227–229] while the other ions have been also used for doping. For example, Rani

et. al. doped TiO_2 by Fe, Ce and La and showed that the band gap of 2.65 eV could be achieved by Fe doping whereas the highest improvement in photocurrent density was obtained after La doping [230]. Other doped ions such as Cr [231, 232], Li [233], Ta [234], Si [235], W [236], and Nb [201] resulted in band gap reduction and performance enhancement. Hydrogen doping [236] also caused the reduction of TiO_2 and created oxygen vacancies and therefore improvement in the photocurrent density. Figure 7.2 shows the effect of the doping on average band gap and current density under 1 V vs NHE bias conditions for TiO_2; although various successful examples of doping have been reported, we observed no significant improvement in the average in our dataset. The only exception to this is that the carbon doping indeed increased the photocurrent density significantly, as clearly seen in Figure 7.2.

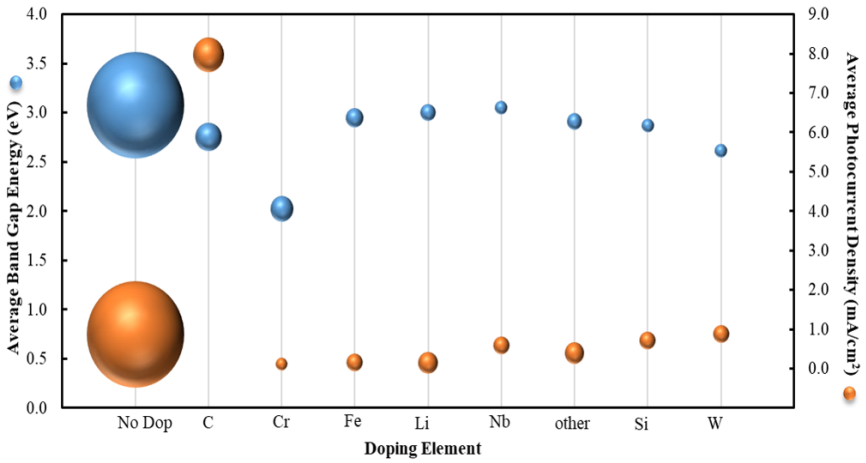

Figure 7.2. Effect of doping elements for TiO_2 (under 1 V vs NHE bias) on average band gap (left) and average photocurrent density (right).

Au [198], Pt [237] and Ru [238] were often used with TiO_2 as the co-catalyst to promote charge separation. Similarly, graphene oxide (GO) [239], CdS [240], $CdSe$ [241] and g-C_3N_4 [242] nanoparticle decoration of TiO_2 was proved to be effective for band gap reduction and better charge separation. It was also reported that, when TiO_2 was coupled with $BiVO_4$ to form a composite structure, the performance was improved by 5 folds in the visible region due to the effective charge separation [243]. Hybrid nanostructures were also shown to improve the performance [244]. For example,

ZnO layer over TiO_2 resulted in better charge separation and lowering the band gap, thus increasing the efficiency [245]. W-doped TiO_2 coupled with $BiVO_4$ [246] showed visible light activity and high photoelectrochemical performance while the use of BiOCl improved both stability and the performance of TiO_2 by forming a p-n junction [247].

In addition to doping and heterostructures, different structural modifications have been also used to improve the performance. For example, the phases of TiO_2, anatase and rutile, were coupled in a core-shell structure, and almost 20 times increase in the photocurrent, compared to the anatase phase, was observed [240]. Thermal treatment also plays important role in the performance of TiO_2 photoelectrodes. For instance, annealing at temperatures above 400°C increased the conductivity and lowered the band gap due to rutile formation in the structure [248]. Thermal treatment under argon and hydrogen [248], or ammonia [249] flow increased light absorption and resulted in better performance due to reduction of TiO_2. Other treatments such as reduction with Al [250], H_2O_2 treatment [251] and NaBH4 treatment [252] caused improvements in the light absorption and charge carrier properties of TiO_2 photoanodes.

Fe_2O_3 is the second most studied semiconductor in our dataset by 17% of the total data (98 instances out of 584 from 33 different articles). Its band gap is suitable for visible light absorption (2.1 eV) but it has slow charge transfer and so high recombination rate [253,254]. In order to overcome the charge transfer problems, Al [255], P [249], Pt [250], Sn [251], Ti [255–260] and Zn [261] ions were doped to Fe_2O_3. Heterostructure formation is another approach to increase the charge transfer properties of Fe_2O_3. SiO_2 under-layer in Fe_2O_3 photoanode suppressed the electron-hole recombination and improved the performance [262]. Coating of Fe_2O_3 by Au [263], FeOOH [264], $LaFeO_3$ [265], NiO [266], NiOOH [267] and TiO_2 [268,267] layers favoured charge separation and photoelectrochemical reaction. As in TiO_2, studies are showing the temperature and atmosphere for thermal treatment can be effective on the performance of Fe_2O_3 photoanode [223], [252].

ZnO have been also studied widely for PEC water splitting. It has large band gap like TiO_2 but it has suitable band edge positions for overall water splitting [268–273]. ZnO has been studied in 22 different articles corresponding to 10% of the total data (60 instances out of 584). It is mostly used in heterojunction structures rather than being doped with an element or promoted by a co-catalyst. However, doping by C [273], Co [274] and Li [275] ions shifted its band gap to visible light region as well, and decreased rate of combination due to separation of charges enabling oxidation and reduction. Sensitization by carbon quantum dots [276], Au nanoparticles [277] and AgSbS2 [278] increased the performance due to the change in band gap and charge separation properties. Thermal treatment is also effective on the performance of ZnO [279]; it was shown that annealing at 450°C could increase the efficiency up to 7.5 times compared to unannealed ZnO [280]. In addition to thermal treatment, treatment environment is also important to obtainhigh performance ZnO photoelectrodes; it was reported that the structure of ZnO changed with the addition of nitrogen gas to the argon environment during synthesis [281]. Heterojunction by CuO, Ag, and Au metals [282], core-shell structure with TiO_2 [283], $ZnFe_2O_4$ [284], CdSe [285] or Er_2O_3 [286] and multilayer composite forming by ZnSe [203], CdS [287,288] or Ga-ZnO [289] were employed to enhance charge separation and visible light absorption of ZnO photoanodes.

Our dataset contains 21 articles and 58 instances (10% of total data) focused on $BiVO_4$ photoelectrode. Like ZnO, $BiVO_4$ is mostly used in the heterostructure forms. It was used with TiO_2 [236], WO_3 [300] and reduced graphene oxide [301-303] to overcome charge transfer problems and increase the performance. It was mostly doped with higher valence metal ions like Mo [304–308] and W [309] to improve the charge separation. Doping can be done to the different (Bi^{3+} and V^{5+}) sites of $BiVO_4$, and it was reported that doping to V^{5+} site improved the performance better than doping to Bi^{3+} site [310]. For co-catalyst, the cobalt phosphate [311,312], cobalt silicate [313], $LaFeO_3$ [313], Bi nanoparticles [306], AgS [307] and Au [308] were employed to increase the photocurrent density.

WO_3 also attracted a lot of attention in the field of photoelectrochemical water splitting due to its high stability and low band gap (2.6 eV) but it suffers from fast recombination of charges [314,315]. In our dataset 20 different articles with 55 instances (9% of the data) reported WO_3 as photoanode. It was reported that the performance of WO_3 increased with nitrogen doping if doping is in high concentrations [311]. It was also shown that co-catalysts, like PtO_x and RuO_2, offered more reduction and oxidation sites for ions and increased performance [312]. Heterojunction composite formation was used in order to enhance the performance of WO_3 since it lowers the need for external bias and improves the charge transfer properties. Besides doping, it was published that the use of co-catalyst [237], treatment of WO_3 with hydrogen [313] or alcohol vapour [314] increased photocurrent density due to the formation of oxygen vacancies and favoured charge transfer rate. WO_3 was mostly coupled with $BiVO_4$ since $BiVO_4$ can absorb more portion of the solar spectrum also photogenerated charge transfer from $BiVO_4$ to WO_3 increased the efficiency [307],[315,316]; WO_3 was also coupled with reduced graphene oxide [317], g-C_3N_4 [319], TiO_2 [318,319], and BiOI [320,321]. Optimization of thermal treatment temperature was proved to be also effective in increasing the photocurrent density [320-323].

Besides these photoelectrode materials 25% of data is composed of various other semiconductors such as oxides [253–255], nitride and oxynitride compounds [256–258], perovskites [237], [259, 260] and chalcogenides [199], [240], [252], [261, 262]. The preparation method affects the semiconductor properties such as surface area, band gap, and crystal structure. In our dataset, the hydrothermal method is the most commonly used semiconductor preparation route with 29% data (168 instances out of 584) followed by anodization (13%, 77 instances), electrodeposition (13%, 76 instances), solvothermal method (7%, instances), sol-gel synthesis (6%, instances) and the other methods with less than 5% frequencies. It was reported that TiO_2 nanostructure [263,264] and porous WO_3 were obtained by hydrothermal synthesis [265]. Qin et, al., also synthesized WO_3 by using hydrothermal method, and they obtained different lengths and thicknesses of WO_3-flakes by varying reaction time; they observed that thicker flakes had stronger visible light absorption [266]. The hydrothermal time can be tuned to synthesize the

best performing semiconductor; even novel materials such as $Cu_2In_2ZnSn_3$ [267], $CaBi_2O_4$ [268] and $NaInSn_2$ [269] were synthesized by this way, and they showed a remarkable performance in photoelectrochemical water splitting.

Anodization is used to obtain metal oxide structures from metal foils; it was used to prepare Fe_2O_3, TiO_2 [270], Ta_2O_5 [271] and Nb_2O_5 [272, 273] semiconductors in our dataset. For example, self-assembled titania nanotube arrays [234], [274] and nanotube photonic crystal TiO_2 [275] synthesized by anodization showed enhanced photoelectrochemical water splitting performance and good stability. It was shown that the anodization time and rotation speed of the electrode during anodization were important parameters to obtain Fe_2O_3 with optimum thickness and structure [276].

Electrodeposition/electrochemical deposition is similar to anodization, but this time precursor solution is deposited on a substrate by applying a potential difference. Synthesis of Fe_2O_3 by electrodeposition and Fe-oxidation were compared, and it was observed that electrodeposition resulted in a hematite structure with good light absorption but low photocurrent density due to slow charge transfer; however, the Fe-oxidation method resulted in better photocurrent density while its light absorption capacity was low because of the magnetite structure [277]. In order to increase the performance, a modified electrodeposition was employed; Fe_2O_3 precursor was electrodeposited on polystyrene colloids and by this way, it was possible to tune the thickness by electrodeposition time and the geometry by polystyrene sphere size [278]. Zhang et, al. reported that it was possible to control the amount of nanoparticle decoration or per cent of doping by adjusting the electrodeposition time [220]. In addition to the method or procedure employed, the ratio of precursors and additional agents is important in the synthesis of semiconductors as expected. For example, it was possible to tune the structure of bi-phasic Cd_2SnO_4 by the addition of urea in different amounts; the junction of crystal structures improved the performance by reducing surface recombination of photogenerated electrons and holes [279]. Similarly, changing the acid type, optimizing the precursor concentration and the reaction time improved the performance of $BiVO_4$ photoanode [226].

In Figure 7.3, the effect of the synthesis method on the crystal structure and the band gap energies are given for the most frequent bottom materials used in the dataset. It was seen that $BiVO_4$ was used mostly in the monoclinic form and only the sol-gel method resulted in a different crystal structure, which has the lowest band gap of 2.19 eV compared to the average band gap value of 2.45 eV obtained with other methods. Fe_2O_3 was used in hexagonal and rhombohedral crystal structures in the dataset and the lowest band gap value (1.85 eV) was achieved with hexagonal Fe_2O_3 prepared by hydrothermal method while the overall band gap for Fe_2O_3 was 2.19 eV. For TiO_2, mostly the anatase phase was utilized (around 60%), and it is seen that the rutile phase had generally a lower band gap than the anatase phase. For WO_3, on the other hand, the lowest band gaps were observed in monoclinic crystal structure prepared by sol-gel synthesis and chemical solution route (the average band gap for all WO_3 in the dataset was 2.57 eV). ZnO was mostly in hexagonal structure in the dataset with the average band gap of 3 eV (excluding the hydrothermal method) while the hydrothermal method gave the lowest average band gap (2.59 eV) with cubic crystal structure.

The steps in the photoelectrochemical water splitting can be summarized as: i) absorption of the energy from light source by the semiconductor, ii) generation and separation of photoexcited electron and holes, iii) reduction of water to H_2 by electrons and oxidation of water to O_2 by holes [201]. Theoretically, 1.23 eV is needed to split water, but energy losses in practical applications increase the required energy up to 1.8 eV [219]. There are many efforts for increasing the efficiency (photocurrent) of photoelectrochemical water splitting by focusing on the materials, electrode fabrication procedures, light properties, and reaction solution, we will briefly review those efforts made in the literature in terms of major variables before the detailed ML analysis.

In addition to the synthesis of semiconductors, the methods used in electrode fabrication are also important. It is possible to fabricate electrodes in situ; the conductive substrate can be coated while the semiconductor synthesis. In our dataset 70% of the experiments (412 instances) used in-situ fabrication using methods like hydrothermal, sol-gel, chemical deposition, vapor deposition, and electrodeposition. When in-situ is

not preferred, spin coating (50 instances), dip coating (15 instances), doctor blade (14 instances) and electrodeposition (14 instances) were used to prepare electrodes from previously synthesized semiconductors. In spin-coating and dip coating, the thickness of the coating layer can be controlled via cycle number whereas the deposition time plays an important role in methods like electrodeposition as explained above.

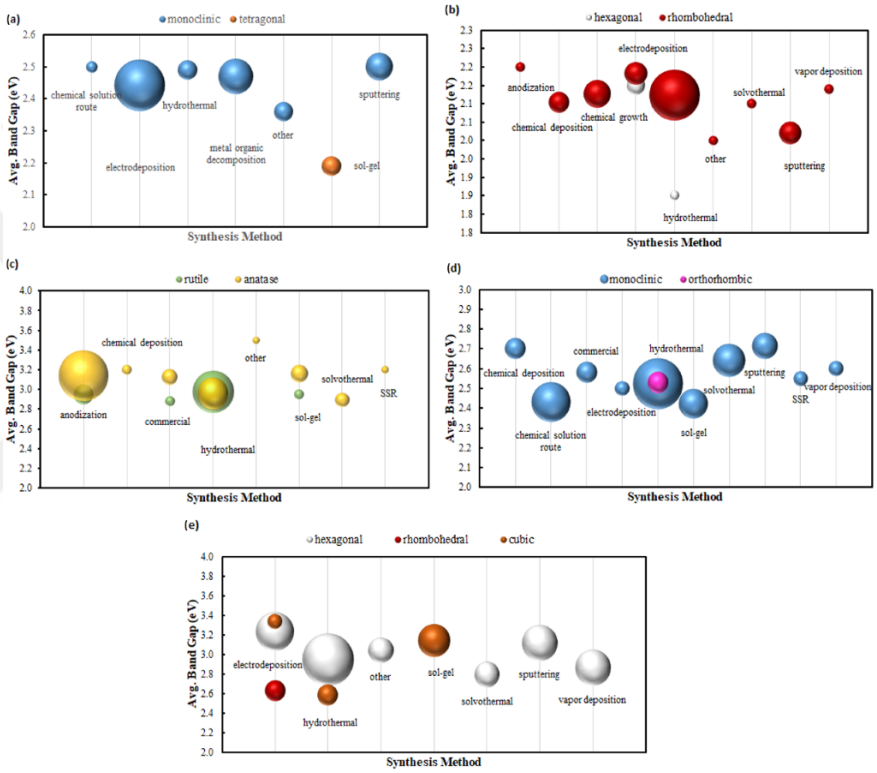

Figure 7.3. The effect of synthesis method on the crystal structure and the band gap energies for (a) $BiVO_4$ (b) Fe_2O_3 (c) TiO_2 (d) WO_3 (e) ZnO photoanodes.

In practical applications, the photoelectrodes are supposed to absorb sunlight, and the spectrum of sunlight mainly consists of visible light [203]. Hence, it is important to utilize the visible light portion of the spectrum and develop semiconductors that can do this efficiently. However, different wavelengths (in both visible and UV region) and intensities of light were used in our dataset. Increasing the light intensity resulted in photocurrent density increase as expected since more photons were involved in the excitement of electrons [250], [280, 281]. In some works, cut-off filters were used to investigate the visible light performance of photoelectrodes [248], [282–286]. The widely

used light type is Xe lamp with 487 instances (82%) in the dataset. The intensity was mostly 100mW/cm^2 is used (404 instances, 69%) while some other values were also used. The use of AM 1.5G filter, which is a standard used to simulate solar light on the earth's surface [287], was reported in 376 instances (64%) [250].

Electrolytes are the media for ion transfer that can also prevent electron-hole recombination [268], [288]; the choice of electrolyte is important for the photoelectrochemical performance since the reaction on the catalyst surface is affected by the concentration, pH, and type of the ions [289]. Oxidation/reduction potentials of ions are important to achieve effective water splitting. For example, anions like iodine and bromide are oxidized easier than water, therefore they should not be used in electrolytes to have water oxidation reaction; however, the chlorine ions can be used because they have higher oxidation potential than water [290]. The choice of electrolyte is also depending on the semiconductor used. For instance, it was found that the photoelectrochemical performance over TiO_2 was highest for Li+ containing alkaline electrolytes (compared to those containing Na+ and K+) [291]; however, when Li+ was used (instead of Na+), the injection of Li+ ions to mesostructure was easier and resulted in electrochromism, which decreases the photocurrent density, on WO_3 photoelectrodes [292]. For ZnO, on the other hand, the use of Na_2SO_3 electrolyte instead of $NaSO_4$ was reported as having six times higher water splitting performance due to the consumption of holes by SO_3^{2-} ions, which promoted electron-hole separation [293]. Similarly, WO_3 gave better results in acidic conditions whereas the best results over TiO_2 were obtained in basic solutions. The pH of the electrolyte is also important since the gas evolution reactions and the rate-determining steps depend on pH [294]. For example, electrolytes with different pH values (1,3 and 7) were used for $CuWO_4$ based photoelectrode, and the decrease of pH of buffered solution also decreased the current density [295]. The electrolytes and average photocurrent density they obtained are presented in Figure 7.4 for TiO_2 and WO_3 as examples.

Additives can also be used in electrolytes to improve the photoelectrochemical performance; these chemicals work as electron donors (hole scavenger) or acceptors

and they prevent electron-hole recombination. For example, methanol is widely used as electron donor in both photocatalytic and photoelectrochemical water splitting [259]. For $CuWO_4$ photoelectrode, the use of methanol as additive doubled the photocurrent since it is a strong electron donor [295]. Similarly, ethanol was employed as additive for photoelectrochemical system with WO_3 photoelectrode and photocurrent obtained was increased [292]. H_2O_2 is another additive that can be used to increase the photocurrent [198], [204], [249], [296] but not preferred due to its corrosive nature [297]. Other than those, the additives such as ethylene glycol [236], [298] and EDTA [251] were also used in the works covered in the dataset. Na_2S and Na_2SO_3 additives were mostly used as hole scavenger to promote sulphide oxidation on the working electrode [203], [299], [300–305].In summary, NaOH is the most dominant electrolyte with 169 instances (29%) in our dataset, followed by Na_2SO_4 with 147 instances (25%) and then KOH with 111 instances (19%); there are also other electrolytes (such as HBr, NaCl, phosphate buffer) with the frequency of less than 5% each. The large majority of instances (481 experiments) did not use any additive in the electrolyte while Na_2SO_3 was used as hole scavenger in 38 instances out of 102 (37%); the materials such as methanol, ethanol and ethylene glycol were also used in lower frequencies.

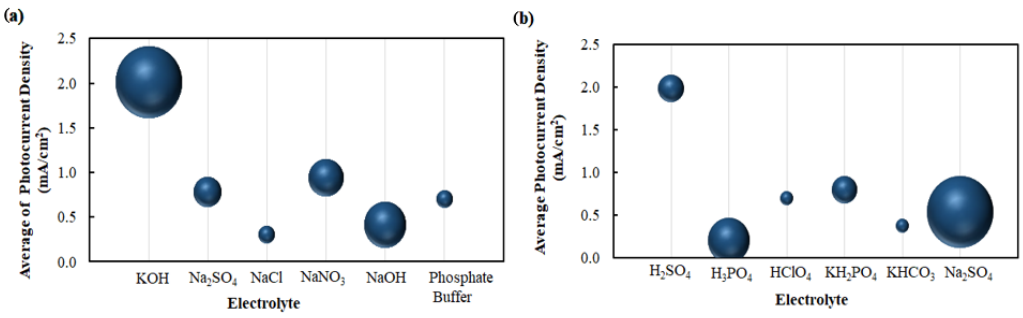

Figure 7.4. The effect of electrolytes on photocurrent density for different semiconductors (a) TiO_2 and (b) WO_3.

7.3.2. Association Rule Mining Analysis

Association rule mining was used to identify the previously unknown association (if-then relations) between the (discretized) input as antecedent or left-hand side

(LHS) variables and the band gap as consequent or right hand-side (RHS) variable. Antecedent and consequent can be considered as "if" and "then" part of the statement respectively and three parameters (support, confidence and lift) are used to evaluate the relation. The determination and interpretation of these parameters are explained for the rule below as an example

$$Method = anodization \Rightarrow BandGap\ Class = high. \tag{7.1}$$

Suppose the rule above has support:0.095, confidence:0.72 and lift:2.25. Itassociates the anodization (as the preparation method for the first layer of the semiconductor) with the probability of obtaining a high band gap (higher than 3 eV) semiconductor. The support is the probability of instances that meet both conditions in RHS and LHS (i.e. the number of high band gap cases prepared with anodization, which is equal to 38, divided by the total dataset). Since the total number of instances is 401, the support is 38/401=of 0.095. The confidence is the ratio of instances that meet conditions in both RHS and LHS to the number of instances that meet the condition in only LHS (i.e. total number of electrodes prepared by anodization). Then, the confidence is 38/53= 0.72. Finally, lift is equal to the division of confidence with the fraction of instances that meet the restriction in RHS (i.e. total high band gap cases). Since the total number of cases obeying the RHS condition is 128, the lift can be calculated as (38/53)/128/401) = 0.72/0.32=2.25. In general, lift value of one indicates that there is no relation between the conditions defined in LHS and RHS while lift larger (smaller) than one implies a positive (negative) relation. The lift value of 2.25 means that the fraction of high band gap photoelectrodes among those prepared by the anodization method is 2.25 times higher than the fraction of high band gap cases in the total dataset. This clearly indicates that the use of anodization for the first semiconductor layer of electrode results in high band gap electrodes. The reliability of this generalization can be checked with the support and confidence. It should be noted again that these band gap values are for the electrodes, but they also correspond to the materials because a single semiconductor was utilized for most of the times (84%). We treated rare cases that may have a second semiconductor layer under the category of first layer semiconductor to prevent confusion.

The lifts for the low, medium, and high band gap class electrodes based on the preparation methods for the first layer semiconductors are compared in Appendix C. The semiconductor prepared by the chemical growth tends to have a band gap value smaller than 2.40 eV (low class) since the radius of the red bubble is bigger than the radius of the yellow and purple bubbles. When the chemical deposition is used in the synthesis of the first layer, however, the probability of having an electrode with medium class band gap (between 2.4 -3 eV) is higher even though there are also some low and high band gap cases. Some synthesis methods seem to give only one class of band gap; however, this usually happens because those methods were probably used for specific materials (for example, etching produces only low-band materials, but it was used only for Si electrodes).

The change of the band gap energy of a semiconductor with respect to the doping material, synthesis method and thermal treatment for frequently used materials are given in Appendix C. When conditions for low band gap energy are investigated (considering that the low band gap is required for visible light absorption), it is clearly seen that for $BiVO_4$, Li, and Mo doping has more potential for this compared to other doping materials. The use of the sol-gel method for $BiVO_4$ and calcining between 500 °C and 600 °C for 2 h provide a high probability of obtaining a low band gap (Table 6.4). Fe_2O_3 also has low band gap when P or Pt is doped, anodization or sol-gel is used with no calcination. For TiO_2, on the other hand, the chromium doping, and electrodeposition method are found to be more suitable to have low band gaps. WO_3 seems to have low band gap only with sol-gel method while no rule can be applied to obtain low band gap ZnO; in our dataset, all band gap values of ZnO are either in medium or high class. ARM analysis of photocurrent density is performed for each of the commonly used bottom semiconductors ($BiVO_4$, TiO_2, Fe_2O_3, ZnO, and WO_3) separately because the preliminary analysis showed that the different descriptors (input variables) were important for different semiconductor.

Table 7.4. ARM analysis of Band gap for $BiVO_4$.

Variable	**Level**	**LOW BG**	**MEDIUM BG**	**HIGH BG**
Doping	Li	3.69		
	Mo	2.39	0.91	
	Sn		3.51	
	W		3.51	
	No doping	0.69	1.97	
True class	electrodeposition	1.32	1.65	
	hydrothermal		0.82	
	metal org. decomp	0.7	2.14	
	sol-gel	2.46	1.17	
Calcination Temperature	400-500	0.81	1.83	
	500-600	2.88	0.41	
	No calcination	0.53	2.05	
Calcination time	$<$ 1h	1.85		
	$<$ 2h	3.69		
	$<$ 3h	1.00	1.7	
	$<$ 6h		3.51	

Each dataset is divided into three subsets by considering the level of bias since not only the photocurrent density itself but also the effects of the materials and methods on the current density may depend on the bias level. The lower limit of high bias subset is selected as 1.23 V, since it is the theoretical energy needed for water splitting and when further bias values are employed the feasibility of the process decreases. The value of 0.5 V is selected to separate medium bias and low bias subsets by considering the distribution of data. After dividing data into three as low, medium, and high bias, photocurrent density values are also divided into three classes for each subset as given in Table C.2. The upper or lower limits of photocurrent density classes for each subset were determined again in a way that the number of instances in each class became approximately equal. Consequently, the upper and lower limits of photocurrent density classes are different for each bias level otherwise the results obtained with high/medium

bias would dominate the high photocurrent density class and overshadows the potential semiconductors that may be effective under no or very low bias conditions. Indeed, as it is apparent from Table 7.4 and Table C.4-C.7, the conditions leading to high photocurrent density under high bias conditions are quite rare for commonly used semiconductors indicating that some special materials listed under "other" are more effective. It should be noted that the common semiconductors (except F_2O_3) may still perform better under high-bias conditions than they do under low bias, but they cannot compete with some other materials at high-bias conditions.

The ARM results for $BiVO_4$ are presented in Table 7.5 under low bias conditions, Table 7.6 under medium bias conditions and Table 7.7 under high bias conditions. It seems that the use of $BiVO_4$ photoanode without any co-catalyst is better in low bias values; this is the only option with high lift (1.84) in low bias operations, and it is supported by a significant number of data (25 cases represented as count) indicating its high reliability (see Table 7.5). However, as the bias increases to a medium level, $BiVO_4$ performs much better with $CoPi$ co-catalyst while some other co-catalysts (like AgS, Co-Ni or Pd) also became useful to improve the photocurrent density even though the use of no co-catalyst is still a good option. Mo seems to be the best doping choice for $BiVO_4$, to obtain high photocurrent values at low bias (with the lift and count of 3.39 and 13 respectively) as well as the medium bias conditions, but it loses its significance with further increase of bias; no doping is also a good option for low and medium bias conditions. Among various choices for the second layer, the reduced graphene oxide, FeOOH, and $LaFeO_3$ are favourable for high photocurrent density in both low and medium bias while $Co - LaFeO_3$ also become effective at medium bias levels. Electrodeposition (ED) appears to be the best choice as the synthesis method to obtain high photocurrent densities in both low and medium bias; calcination at higher temperatures and shorter times seems to be better. Additionally, the use of drop casting as the coating method provides better results than in-situ and spin coating. Finally, $K_2B_4O_7$ or Na_2SO_3 electrolytes, with molarities between 0.25-0.5 M or 1-1.25 M respectively, and pH of 11 give high photocurrent densities with increasing strength as the bias increases. As Table 7.5 indicates, $BiVO_4$ (as the other commonly used

semiconductors except for Fe_2O_3), there seem to be no conditions that will provide high photocurrent density at higher biases relative to some rarely used semiconductors (such as SiO_2 and Cd_2SnO_4).

For TiO_2, the use of no co-catalyst gives high photocurrent densities at low and medium bias while C doping seems to be also effective. Anodization and flame oxidation are the most important methods for high photocurrent densities at low bias, but anodization loses its strength with increasing bias. Calcination between 500-900 °C for a short time (2 h or less) is also important for obtaining high photocurrent densities. The advantage of in-situ coating at low bias values decreases with increasing bias. The use of KOH electrolyte with a molarity higher than 3 M, and a pH of 14 seems to be beneficial for achieving high photocurrent density in low and medium bias.

We could not deduce any rule for Fe_2O_3 that gives high photocurrent in low bias, which is expected since it has slow charge transfer and high recombination rate [268]. Hence, as the external bias increases, it overcomes charge transfer limitations and produces higher photocurrent densities. The use of Au co-catalyst (even though its effect decreases in high bias), TiO_2 as the second layer, or vapour deposition method for synthesis results in high photocurrent density values with high lift supported with a reasonably large amount of data if a certain level of bias was applied. Similarly, no rule could be deduced for WO_3 to obtain high photocurrent density in low bias values. Although there are data involving the use of WO_3 synthesized by chemical deposition and used with $NaClO_4$ electrolyte producing high photocurrent density with low bias, we did not generalize because they are from the same source [272].The empty lift and count values indicate that the rule for that conditions were not found; it can be due to those data not existing or existing in low frequency.

For ZnO photoanode, although CdS as a co-catalyst does not seem to be influential in obtaining high photocurrent in low bias, it increases the performance with increasing bias. We may say that the doping of this semiconductor does not result in high photocurrent regardless of the level of bias. However, the use of CdSe as a

Table 7.5. Association rule mining results for BiVO4 (Low Bias (<0.5 V)).

Variable	**Category**	**High PD**		**Medium PD**		**Low PD**	
		lift	count	lift	count	lift	count
Doping	Li	-	-	3.9	2	-	-
	Mo	3.4	13	1.5	17	0.5	13
	Sn	-	-	-	-	-	-
	W	-	-	-	-	1.5	2
	No doping	1.2	13	1.5	50	0.8	64
Method	ED	2.2	22	1.5	45	0.6	47
	Hydrothermal	-	-	2.3	7	0.6	5
	MOD	2.1	4	2.2	11	0.4	6
	Sol-gel	-	-	0.9	2	1.2	7
Calcination Temperature	400-500 (°C)	1.6	8	1.5	23	0.7	27
	600-700 (°C)	3.6	11	2.1	29	0.2	7
	No calcination	-	-	1.4	17	1.1	45
Calcination Time	<1h	4.5	4	2.3	6	-	-
	<2h	3.7	13	1.9	19	0.3	7
	<3h	1.5	8	1.7	27	0.7	27
Coating Method	Drop Casting	4.5	4	2.3	6	-	-
	Spin Coating	-	-	1.5	5	-	-
	in-situ	1.7	22	1.5	56	0.7	64
Co-catalyst	NiO	-	-	-	-	1.5	2
	$CoPi$	-	-	-	-	-	-
	other	-	-	1.5	7	0.9	10
	No co-catalyst	1.8	25	1.6	61	0.7	66
Second Layer	$FeOOH$	4.0	5	2.5	9	-	-
	$Co-LaFeO_3$	-	-	1.9	5	0.8	5
	$LaFeO_3$	2.0	4	2.2	13	0.4	6
	rGO	6.7	3	1.6	2	-	-
	No layer	1.3	13	1.2	36	0.9	64
Electrolyte	$K_2B_4O_7$	3.2	9	1.8	15	0.4	8
	K_2HPO_4	-	-	3.9	4	1.5	6
	Na_2SO_3	5.6	10	1.7	9	-	-
	Na_2SO_4	0.9	4	1.3	18	0.9	30
Electrolyte Molarity (M)	0.1-0.25	1.3	7	1.3	20	0.8	33
	0.25-0.5	3.6	8	1.9	12	0.3	5
	0.5-1	-	-	1.4	24	0.9	40
	1 - 1.25	4.7	10	2.9	13	-	-
pH	6	-	-	3.9	4	-	-
	7	0.8	7	1.4	37	0.9	61
	10	2.7	9	1.5	15	0.6	14
	11	4.7	10	2.9	13	-	-

Table 7.6. Association rule mining results for $BiVO_4$ (Medium Bias (0.5 <V< 1.23)).

Variable	**Category**	**High PD**		**Medium PD**		**Low PD**	
		lift	count	lift	count	lift	count
Doping	Li	-	-	2.6	4	0.5	2
	Mo	1.9	18	1.4	39	0.7	52
	Sn	-	-	2.2	9	0.7	7
	W	-	-	1.7	13	0.9	17
	No doping	1.6	51	1.2	109	0.8	191
Method	ED	2.4	53	1.5	94	0.6	104
	Hydrothermal	0.7	2	1.9	15	0.7	14
	MOD	1.2	14	1.5	48	0.8	65
	Sol-gel	-	-	1.3	2	1.2	4
Calcination Temperature	400-500 (°C)	2.8	30	1.4	58	0.7	74
	600-700 (°C)	4.2	32	1.9	43	0.2	11
	No calcination	0.3	7	1.9	78	1.6	192
Calcination Time	<1h	6.5	14	1.6	10	-	-
	<2h	4.3	18	2.2	29	0.9	3
	<3h	2.0	30	1.4	62	0.7	77
Coating Method	Drop Casting	6.5	14	1.6	10	-	-
	Spin Coating	-	-	1.4	39	1.0	70
	in-situ	1.6	55	1.3	128	0.8	198
Co-catalyst	NiO	-	-	-	-	1.5	8
	$CoPi$	-	-	3.2	5	-	-
	other	1.4	7	1.3	18	0.8	30
	No co-catalyst	1.6	62	1.4	156	0.8	218
Second Layer	$FeOOH$	3.7	6	2.6	12	-	-
	$Co-LaFeO_3$	1.3	3	3.0	20	0.2	3
	$LaFeO_3$	4.3	18	2.3	28	-	-
	rGO	11.2	12	-	-	-	-
	No layer	0.8	28	1.6	113	1.1	271
Electrolyte	$K_2B_4O_7$	5.9	32	1.7	26	0.8	3
	K_2HPO_4	-	-	1.3	4	1.2	8
	Na_2SO_3	6.4	12	1.7	9	-	-
	Na_2SO_4	0.8	16	1.2	70	1.0	150
Electrolyte	0.1-0.25	1.3	20	0.9	43	1.0	114
	0.25-0.5	6.6	27	1.6	19	0.4	10
	0.5-1	-	-	1.5	107	0.9	158
	1 - 1.25	5.0	12	1.4	10	0.3	5
pH	6	-	-	1.9	33	0.8	33
	7	0.9	25	1.2	100	0.9	189
	10	3.4	32	1.2	33	0.6	40
	11	4.8	12	1.5	13	0.4	8

Table 7.7. Association rule mining results for $BiVO_4$ (High Bias (>1.23 V)).

Variable	**Category**	**High PD**		**Medium PD**		**Low PD**	
		lift	count	lift	count	lift	count
Doping	Li	-	-	3.2	9	0.3	2
	Mo	-	-	0.6	8	1.3	41
	Sn	-	-	-	-	1.5	8
	W	-	-	-	-	1.5	20
	No doping	-	-	0.3	12	1.4	156
Method	ED	-	-	0.7	13	1.2	55
	Hydrothermal	-	-	-	-	-	-
	MOD	-	-	0.3	6	1.4	62
	Sol-gel	-	-	-	-	1.5	6
Calcination Temperature	400-500 (°C)	-	-	0.5	9	1.3	60
	600-700 (°C)	-	-	1.9	12	0.8	13
	No calcination	-	-	0.3	18	1.4	189
Calcination Time	<1h	-	-	-	-	-	-
	<2h	-	-	2.6	2	-	-
	<3h	-	-	1.1	19	1.1	54
Coating Method	Drop Casting	-	-	-	-	-	-
	Spin Coating	-	-	0.3	6	1.4	68
	in-situ	-	-	0.6	33	1.3	170
Co-catalyst	NiO	-	-	-	-	1.5	5
	$CoPi$	-	-	-	-	1.5	4
	other	-	-	1.1	5	1.8	12
	No co-catalyst	-	-	0.5	34	1.3	227
Second Layer	$FeOOH$	-	-	-	-	-	-
	$Co-LaFeO_3$	-	-	2.6	2	-	-
	$LaFeO_3$	-	-	2.9	3	-	-
	rGO	-	-	-	-	-	-
	No layer	-	-	0.5	34	1.4	256
Electrolyte	$K_2B_4O_7$	-	-	3.0	7	0.3	2
	K_2HPO_4	-	-	1.8	10	0.8	12
	Na_2SO_3	-	-	-	-	-	-
	Na_2SO_4	-	-	0.3	9	1.4	115
Electrolyte	0.1-0.25	-	-	0.4	12	1.4	103
	0.25-0.5	-	-	3.9	5	-	-
	0.5-1	-	-	0.5	22	1.4	159
	1 - 1.25	-	-	-	-	-	-
pH	6	-	-	0.6	5	1.3	29
	7	-	-	0.2	7	1.5	139
	10	-	-	0.7	17	1.2	72
	11	-	-	1.8	10	0.8	12

second layer is effective for high photocurrent values at both low and medium bias conditions. Electrodeposition seems to be the best choice for synthesis, but it loses its advantages with increasing bias to medium level where the hydrothermal method becomes a better option. Drop casting, as the coating method for semiconductors over the electrode substrate, is effective at low biases but it loses its significance to dip coating with increasing bias level (to medium). The use of Na_2S electrolyte with a molarity of 0.1-0.25 M gives high photocurrent values in both low and medium biases.

7.3.3. Prediction of Band Gap

The optimum decision tree, with the cp value of 0.1, for the band gap classification is presented in Figure 7.5. The training and testing accuracies of the model were found to be 78% and 72%, respectively while the balanced testing accuracy, which is used when the data is unbalanced, is computed as 75%; the small difference between those values implies that the distribution of the instances through classes are well balanced. In Table 7.8, the recall (the fraction of class X data correctly classified as X) and precision (the fraction of true class X data in those classified as class X) can be used as a criterion to evaluate the fitness of the model. In applications like ours, precision is more important because missing some suitable conditions (due to the low recall) may not be important if there are some other conditions for selection, but the conditions identified as good should be really good (implying high precision) to avoid misleading. Both recall and precisions are reasonably well high for high-class and low-class band gap; such models usually predict medium class less successfully due to the leaks from both sides. As can be seen from the confusion matrix in Table 7.8, the recall is higher for high band gap classes while the precision is much higher for low band gap classes indicating that the conditions suggested by the decision tree for low band gap, which is desirable, is quite reliable (with testing precision of 84 %). On the other hand, the model may omit some other suitable conditions for band gap considering that the testing recall for this class is only 60 %.

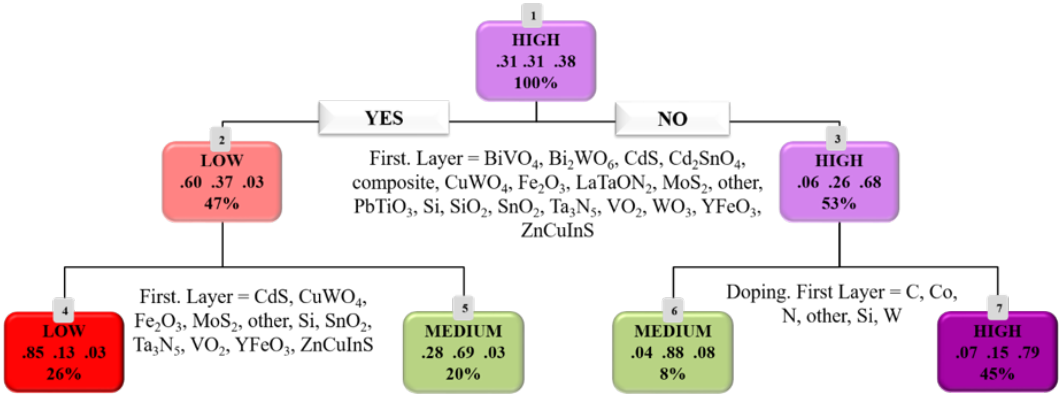

Figure 7.5. The decision tree for band gap classification.

In Figure 7.5 the nodes containing instances with mostly low band gap values (class low) are coloured as red while those having mostly high band gap values (class high) are purple; the nodes containing instances with medium band gap values are indicated as green. The colors of the nodes become darker from the root to the leaf nodes with increasing purity in the nodes. The first line in a node indicates the band gap class of that node determined by majority voting. The second line gives the real fractions of each band gap class in that node (the number on the leftmost is for the low band gap class while the number on the rightmost is for the high band gap), and the last line is the per cent of the total instances placed in that node. The band gap class of the first node is labeled as high since it includes 38% instances belonging high class. Then the branching starts with the first rule of whether the first layer includes particular semiconductors or not; and if those specified semiconductors are in the first layer, then the tree continues to grow leftward, if not then it goes rightward. The decision tree should be followed from the top (root) to the bottom (leaves) until terminal nodes. For example, in node 4, the tree indicates that if the previous statements are answered as "yes", which is solely answered by the type of semiconductor material, the band gap will be in low class. The accuracy of this decision equals the real fraction of the instances in low class in node 4, which is quite high (85%). The decision or rule stated in this branch (i.e., semiconductor material selection) is highly reliable because 26% of the total data corresponding 78 instances obey this rule. One can also consider the branch leading node 7, which is also influenced by doping, as the path that should not be followed if the high band gap of the electrode is going to be avoided.

Table 7.8. Confusion matrix for decision tree model for band gap.

	Overall Accuracy	**Real Class**	**Total Number**	**Predicted class**			**Class Accuracy (Recall)**
				A	**B**	**C**	
Train	0.78	**A**	94	67	18	9	0.71
		B	94	10	64	20	0.68
		C	113	2	4	107	0.95
			Precision	0.85	0.74	0.79	
Test	0.75	**Low**	35	21	9	5	0.60
		Medium	44	3	31	10	0.70
		High	21	1	0	20	0.95
			Precision	0.84	0.78	0.57	

As one of the key factors in photoactive reactions, the band gap should be among the key descriptors in the photocurrent density modelling of photoelectrochemical systems. However, the band gap of working electrodes was not reported in some articles in our dataset (77 out of 476 experiments in the dataset). Hence, we develop a predictive model for band gap using several regression algorithms such as SVM, GB and RF; the best model will also allow us to calculate the missing band gap for the analysis of photocurrent density while this model will be also a valuable tool for the band gap predictions by itself. The descriptors were selected as the electrode materials, preparation methods as well as thermal treatment and calcination conditions, which are known to be effective for the band gap of a semiconductor. Best predictive power was achieved with RF model. The model with the lowest RMSE for validation, which had the *ntree* and *mtry* of 220 and 15 respectively, was found by using 25% of data for testing (101 instances out of 399) and 5-fold cross-validation. The average RMSE validation (average of 5-folds) and testing were calculated as 0.26 and 0.19 respectively. The statistical fitness of the model can be considered as sufficiently high as it is also apparent from the predicted versus real band gap plots in Figure 7.6 The most influential variables were related to the first layer material, the type of material at the bottom layer by far had the most impact on determination of band gap.

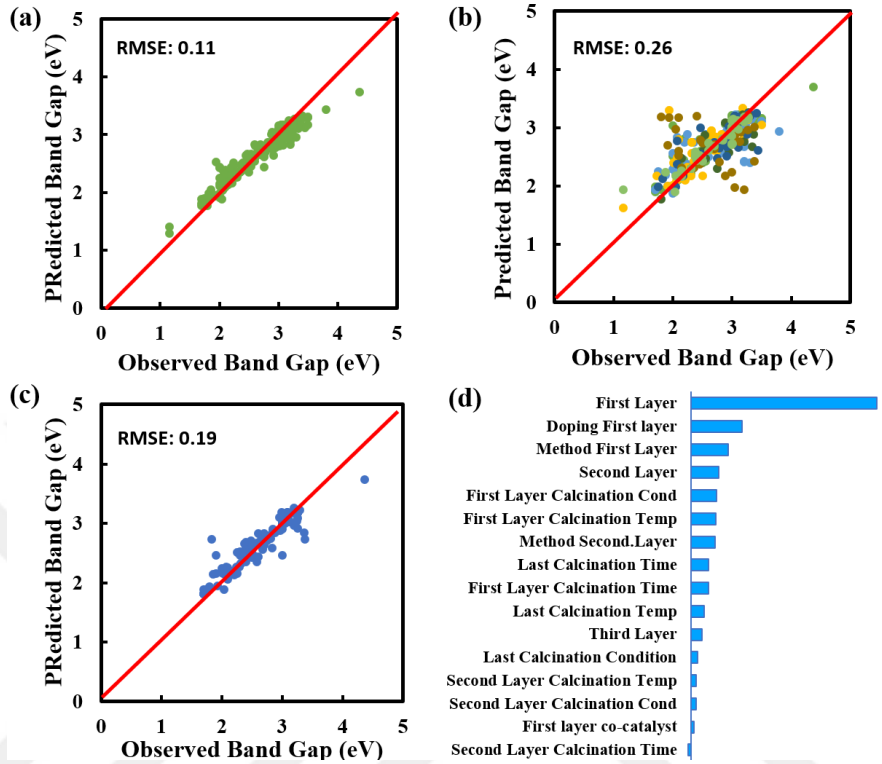

Figure 7.6. Predicted versus actual band gap values by using RF model for (a) training, (b) validation, (c) testing set and d) variable importance plot of the model.

7.3.4. Prediction of Photocurrent Density

Predictive models for photocurrent density, including those developed using random forest were not as good as that for band gap, probably due to the experimental noise arising differences in the experimental conditions (especially non-standard irradiation conditions) in different laboratories. Hence, we employed decision tree classification, which, is less precise than a good predictive model, but they are more suitable for categorical data and less sensitive to the noise in the data. The same procedure described above for the band gap analysis was used here as well; the optimum tree (with the *complexity parameter* of 0.01) is given in Figure 7.7.

In Table 7.9, the confusion matrix for the optimum decision tree model for the photocurrent density is given. The recall and precision have the same meaning and

significance as the decision tree model that was discussed above for the band gap. The same is also true for numbers on the nodes and the interpretation of the tree; hence, we will not repeat the similar discussion that we made above here. Even though the fitness of the model is low compared to that given for the band gap above, it still provides some information. Considering that the major difference in the two models is the presence of reaction (and irradiation), it is reasonable to conclude that the data is much nosier due to the nonstandard conditions mentioned above even though the model seems to be improvable with the availability of more accurate experimental data in the future.

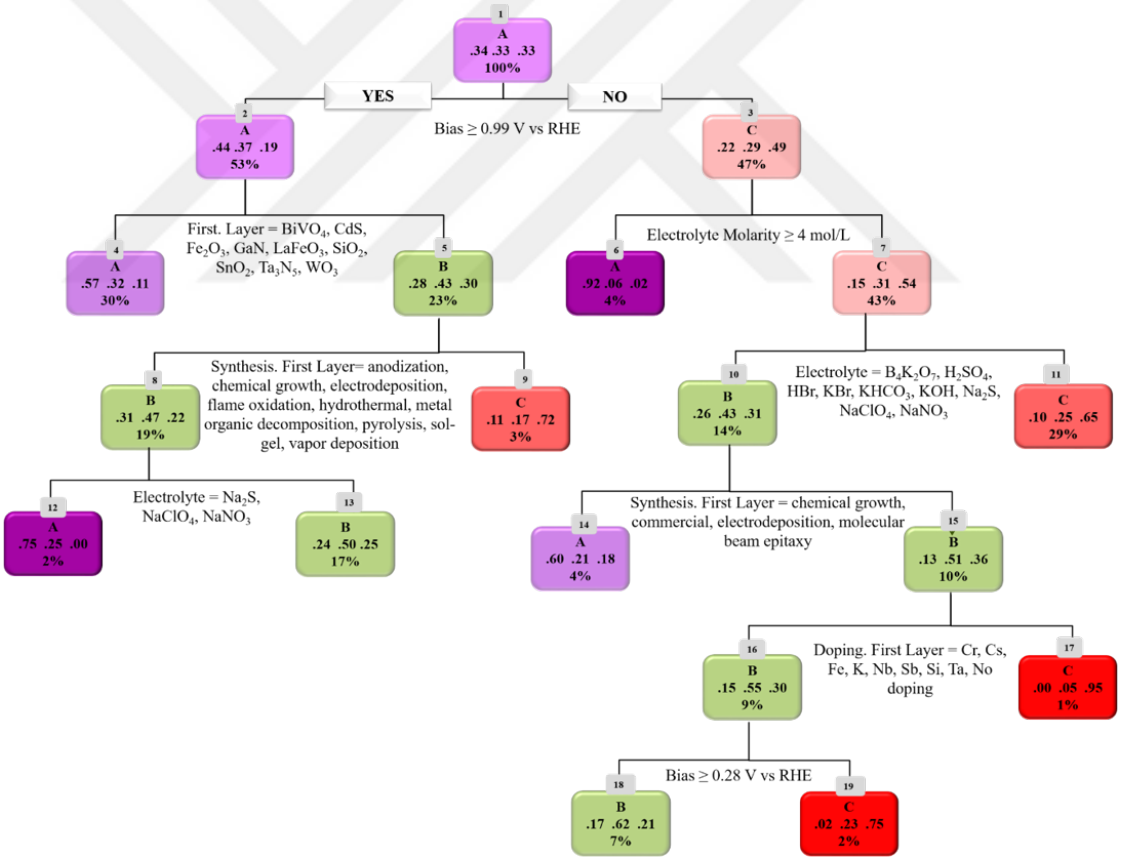

Figure 7.7. The decision tree for current density classification.

Table 7.9. Confusion matrix for decision tree model for photocurrent density.

	Overall Accuracy	**Real Class**	**Total Number**	**Predicted class**			**Class Accuracy (Recall)**
				A	**B**	**C**	
Train	0.61	**A**	2671	1977	438	256	0.74
		B	2623	908	1063	652	0.40
		C	2622	326	469	1827	0.70
			Precision	0.62	0.54	0.67	
Test	0.54	**A**	865	543	187	135	0.63
		B	909	276	348	285	0.38
		C	869	129	222	518	0.60
			Precision	0.57	0.46	0.55	

8. MACHINE LEARNING ANALYSIS OF PHOTOCATALYTIC CO_2 REDUCTION FOR HYDROGEN PRODUCTION

The data collection for this work was conducted by Dilara Saadetnejad, who is a PhD student in Boğaziçi University, Chemical Engineering Department. This work was published in International Journal of Hydrogen Energy [95].

8.1. Background Information on Photocatalytic CO_2 Reduction

The world's energy consumption has been continuously increasing due to the rapid growth of the world population and the energy consumption per capita; the rate of global energy demand is estimated to increase to 25-27 TW by 2050 [306]. Today, most of the world's energy supply comes from non-renewable fossil fuels, which are the main reasons for global warming; about 30.4 Gt of carbon dioxide is annually emitted to the atmosphere [307]. This is expected to cause serious problems including sea level rise due to the melting of polar ice and droughts in future [308]. According to the prediction by the Intergovernmental Panel on Climate Change, the amount of CO_2 will reach to 590 ppm by 2100 leading to a global temperature rise by 1.9 °C [309].

On the other hand, CO_2 can be utilized as a raw material and converted into more valuable chemicals such as hydrogen, methane and methanol [310]. This may help to reduce the net carbon emission to the atmosphere; however, due to its high stability, the chemical reactions of CO_2 are thermodynamically unfavourable requiring a large amount of energy. Consequently, although it is highly desirable, CO_2 reduction is a highly challenging process [147].

The history of CO_2 reduction goes back to the 19^{th} century. In 1870, Royer first reported formic acid production on Zn electrodes via reduction of CO_2 [311]. Between the 1970s and 1980s, a new phase was opened in electrocatalytic CO_2 reduction

research; Ito and Murata studied the electrocatalytic performances of several metals such as In, Cd, Sn, Zn, and Pb to reduce CO_2 to formic acid [312]. The studies on photocatalytic CO_2 reduction also started in 1970s; for instance, in 1978, Halmann reported that CO_2 was reduced to CH_3OH and CO on a p-type GaP electrode under light illumination [313]. Following these works, the electrocatalytic and photocatalytic CO_2 reduction have been attracted increasing attention with some significant progress that was made in recent years.

Solar radiation provides an almost infinite amount of energy as the main source behind all the renewable energy forms; the solar energy reaching Earth in one hour approximately corresponds to the annual energy consumption in the entire world [314]. Hence, if it can be commercialized, the CO_2 reduction using solar energy seems to be an excellent solution to control the CO_2 emission. Two main approaches have been studied extensively to reduce CO_2 via solar energy. First, the photovoltaic cells are utilized to convert the solar radiation into electricity; then CO_2 is reduced electrochemically on metal electrodes utilizing the solar electricity. In the second approach, CO_2 is photocatalytically reduced in the presence of some reductants such as H_2O; this direct approach has been improved significantly with the recent developments in the photocatalytic systems [307], [315], and up to now, many heterogeneous and homogeneous photocatalysts have been extensively studied for this purpose [316, 317]. The heterogeneous photocatalysts, which are based on solid-state semiconductors, have more advantages over homogeneous systems.

Photocatalytic CO_2 reduction can be carried out in both liquid and gas phases [318, 319]. Although the liquid phase processes seem to be more efficient in promoting the formation of oxygenated products like methanol, the gas phase processes have the potential to produce hydrogen or syn-gas under visible light conditions, promoting hydrogen economy while eliminating CO_2. In both cases, various photocatalyst preparation and operational variables affect the performance. The selection of the semiconductor is highly critical; its bandgap has to be suitable for visible light because solar energy is expected to be the source for irradiation in practical applications while

the band edge positions should be appropriate for the reaction. Indeed, the common semiconductors are usually modified to have suitable bandgaps and band edge positions by using various methods such as metal/nonmetal doping [320], dye sensitization [321] and structural modifications as in the case of black TiO_2 [322] during the synthesis and pretreatment stages. The use of a co-catalyst to ease the charge separation and prevent electron-hole recombination, which is another major bottleneck in photochemical reactions, may also improve the bandgap and band edge alignment [323]. Mixed semiconductors and co-catalysts as well as multilayered structures are also used to improve the light-harvesting capabilities and photocatalytic reaction rates while the operational variables like temperature, pressure, CO_2 concentration, the characteristic of the light source and additives in reaction solution are also important to improve photocatalytic performance. All these variables interact with each other and make it hard to understand and determine the most suitable conditions; however, the presence of a large number of works accumulated in literature also provides an opportunity to build large data sets and analyze them using machine learning tools to identify the hidden relations and patterns among the variables as it has been implemented in many field or research including catalysis and photocatalysis [205].

In this work, the gas phase photocatalytic CO_2 reduction over solid photocatalysts was investigated by using machine learning techniques, and the results were compared with those obtained in liquid processes. The datasets containing 268 cases for gas phase and 281 cases for liquid phase CO_2 reduction (total 549 cases) were extracted from 80 published papers in literature and investigated using random forest and decision tree techniques to identify the major patterns and characteristics in the literature data, determine the critical factors and their effects on the bandgap and CO_2 photoreduction rates; a predictive model for the bandgap and heuristic rules for high CO_2 reduction rates were also developed.

8.2. Computational Details

8.2.1. Dataset Preparation

The dataset was built from 80 experimental papers, which were published from 1993 to 2019 and available in online sources including Elsevier, Wiley Online Library, American Chemical Society, and Royal Society of Chemistry. The dataset contains 268 cases for the gas phase and 281 cases for liquid phase processes (549 data points in total). The gas and liquid phase data were analyzed separately due to the differences in input variables (descriptors); this also allowed us to compare the basic natures of the two processes, including the product distribution. Descriptors in the dataset involve the characteristics of semiconductors and co-catalysts (including preparation methods), and the reaction parameters which are given in Table 8.1 and Table 8.2 for the gas and liquid phase data, respectively. Some papers did not report the bandgap for semiconductors; hence, we developed a predictive random forest model for bandgap using the reported values in literature and to calculate the missing data to be able to use a complete bandgap dataset for the gas production; the bandgap was not reported in 180 works out of 549 data.

Although the details of individual gas products were recorded in the dataset, the total gas production rate, which was the sum of CH_3OH, CH_4, H_2, CO, CH_2O and CH_2O_2 production rates, was used as the single output variable in the analysis after the main differences in the product distributions of gas and liquid phase processes were presented and discussed. All descriptors in Table 8.1 and Table 8.2 were used for the total gas production rate while only the descriptors involving the photocatalyst (excluding reaction medium and operation conditions) were utilized for the bandgap predictions.

Table 8.1. List of variables with their ranges for gas phase dataset.

Variable	**Range**
Light Source	UV, VIS
Temperature (K)	278-353
Pressure (bar)	0.7-1.75
Feed molar ratio (CO_2:H_2O)	0.2-97.7
Additive	H_2SO_4, K_2CO_3, $KHCO_3$, Na_2CO_3, $NaHCO_3$, None
Amount of additive (Mol/L)	0-0.5
Semiconductor prepation method	Chemical reduction, Commercial, Hydrothermal, Ion exchange, NP*, Sol gel, Solid state, Thermal method
Co-catalyst Prepation Method	Atomic layer deposition, Deposition precipitation (DP), Hydrothermal, Incipient-to-wetness impregnation, NP*, Photodeposition, Sol gel, Solid state, Thermal method
Calcination temperature (K)	298-1223
Loading of co-catalyst-1, wt %	0-100
Co-catalyst-1	Ag, Au, Co, Cr, Cu, CuO, ln_2O_3, MgO, None, Pd, Pt, Ru,V
Loading of co-catalyst-2, wt %	0-2
Co-catalyst- 2	Cu,Fe,None
Dopant-1/Semiconductor-1, wt %	0-90
Dopant-1	Cu, Fe, Mg, None, V, W
Dopant-2/Semiconductor-1, wt%	0-5
Dopant-2	None, W
Ratio of Semiconductor-1, %	4.2-100
Semiconductor-1	C_3N_4, CeO_2, GO, $SrTiO_3$, TiO_2, ZnS, $Zn_2Ti_3O_8$, ZrO_2
Semiconductor-2	None, SiO_2, TCPP, Zeolite**, $ZnTiO_3$
Band Gap Energy, eV	1.5 - 3.8
Total gas production rate (μmol/gcat/h)	0.01-310

*NP: not provided; ** Modified

Table 8.2. List of variables with their ranges for liquid phase dataset.

Variable	**Range**
Light Source	UV, VIS
Temperature (K)	278-973
Pressure (bar)	0.25-7.5
Additive	C_3H_8O, CH_3CN, K_2CO_3, $KHCO_3$, $NaHCO_3$, None, NaOH, TEOA
Amount of additive (Mol/L)	0-6
Semiconductor preparation method	Chemical reduction, Commercial, Hydrothermal, NP*, Polymerizable complex method, Sol gel, Solid state, Thermal method
Co-catalyst preparation method	Chemical reduction, Deposition-precipitation, Freeze-drying, Hydrothermal, Incipient-to-wetness impregnation, None, Photodeposition, Sol gel, Solid state, Thermal method
Calcination temperature (K)	298-1473
Loading of co-catalyst-1, wt %	0-175
Co-catalyst-1	Ag, $AgBr$, Au, Bi_2S_3, Co, Cu, CPDs, Eu, No, NiO, Pd, Pt, Ru, Ru (II) dinuclear complex
Loading of co-catalyst-2, wt %	0-1
Co-catalyst- 2	Co, Cu, Fe, No, Ni, Ru(bpy3)+2, Ru (II) dinuclear complex
Ratio of semiconductor_1, %	0-100
Semiconductor-1	$BaLa_4Ti_4O_15$, $BiVO_4$, BP, C_3N_4, $CaFe_2O_4$, CdS, CeO_2, Fe_2O_3, $lnTaO_4$, No, NiO, $PbBiO_2Br$, rG, rGO, SiC, $SrNb_2O_6$, Ta_2O_5, $TaON$, TiO_2, $ZnGa_2O_4$, ZnO, ZnS, $ZnTe$
Semiconductor-2	ACF, CdS, CeO_2, CuI, Fe_2O_3, Molecular Sieve**, No, NiO, rGO, SiO_2, Ta_2O_5, WO_3, $ZnTe$
Band Gap Energy, eV	1.6 - 4.7
Total gas production rate (μmol/gcat/h)	0-840

*NP: not provided; ** Modified

8.2.2. Model Development

The random forest regression algorithm was applied to predict the total gas production rates and the missing bandgap values using the reported values and the other semiconductor properties; the randomForest package [38] in RStudio was used for this purpose. The variable importance of the models was obtained using caret package [138]. The *node size* and number of trees (*ntree*), as the model hyperparameters, were optimized by k-fold cross-validation. The dataset was randomly divided into two subsets (training and testing) using various percentages (as the test set from 10% to 40%); Then the training set was further divided into n subsets (k was changed from 3 to 10); k-1 subsets were used for model building while the remaining one subset was used for validation (k-fold cross validation). This procedure was repeated k times (by changing the validation set) with the varying model hyperparameters (node size from five to 50 and ntree from five to 300) for all test set ratios; band gap prediction model with node size of 5 and ntree of 75 (obtained with 10% of data as the test data and 10-fold cross-validation) resulted in minimum root mean square error for validation. The model was finally tested with 10% testing data to assess their prediction capacity for unseen data. The random forest regression was also applied to the total gas production rates using the similar procedure described above; two separate models were developed for gas and liquid data because the product distributions are different in the two processes indicating that the descriptors output relations may be different in gas and liquid datasets. However, the band gap (together with operational variables) was also added to the descriptor list. The preliminary results showed that the predictions were not satisfactory for the total gas production rate; therefore, log (10 base) transformation was applied to the output so that, at least an order of magnitude prediction could be performed Both gas and liquid phase models were developed using 5-fold cross-validation and using 25% data as test set since lowest validation RMSE was achieved using those values. The best model for the gas phase had node size of 5 and ntree of 480 whereas the model for the liquid phase had node size and ntree of 5 and 475, respectively.

Decision tree algorithm was used for the classification of data based on total gas production rates; the rpart package [218] in RStudio was used for this purpose. The dataset was divided into three equal-sized classes as high (class A), intermediate (class B) and low (class C) production rates for gas and liquid phase systems separately. The class structures for both phases are presented in Table 8.3. Similar to RF regression described above, the DT analysis was performed using 10-40% of data for testing and 3-10 for the n in n-fold cross-validation to optimize the minsplit and complexity parameter as the model hyperparameters within the range of 5-45 and 0.01-0.1 respectively. The optimum tree was selected by the overall validation accuracy while the accuracy and purity in the individual terminal nodes and the simplicity of the tree structure were also considered among the alternatives with similar overall accuracy. The optimum minsplit and complexity parameter were found to be seven and 0.01 respectively for the gas phase dataset (with 20% testing data and 10-fold cross-validation). The best model for the liquid phase dataset, obtained by using 10% of data for testing and 10-fold cross-validation, had the minsplit and complexity parameters of 11 and 0.015 respectively.

Table 8.3. Class structures for gas and liquid phase data.

Class	**Gas dataset**		**Liquid dataset**	
	Total gas production ranges (μmol/gcat/h)	**# of data points**	**Total gas production ranges (μmol/gcat/h)**	**# of data points**
A (high)	(6.97- 310]	67	(27.9-840]	70
B (medium)	(0.75- 6.97]	67	(4.05-27.9]	70
C (low)	[0.01- 0.75]	67	[0-4.05]	70

8.3. Results and Discussion

8.3.1. Pre-analysis of the Dataset

In the photoreduction of CO_2, methanol (CH_3OH), methane (CH_4), carbon monoxide (CO) and hydrogen are the main products whereas formic acid ($HCOOH$), formaldehyde , ethanol , acetic acid, oxalic acid, acetaldehyde, ethylene and ethane have been also seen in some cases. The free energy, $\triangle$G0 and the standard redox potential, $\triangle$E0 of the multi-electron CO_2 reduction reactions are listed below [317]. CO production from CO_2 reduction reaction is

$$CO_2(g) \rightarrow CO(g) + 1/2O_2(g), \tag{8.1}$$

where the the energy needed is $\triangle$ G = 257 kJ/mol ($\triangle$ E = 1.33 V) and CH_4 production from CO_2 reduction reaction is

$$CO_2(g) + 2H_2O(l) \rightarrow CH_4(g) + 2O_2(g), \tag{8.2}$$

where the the energy needed is $\triangle$G = 818 kJ/mol ($\triangle$E = 1.06 V). HCOOH is produced by the following reaction

$$CO_2(g) + H_2O(l) \rightarrow HCOOH(l) + 1/2O_2(g), \tag{8.3}$$

where the the energy needed is $\triangle$G = 286 kJ/mol ($\triangle$E = 1.48 V). Methanol production reaction is as follows

$$CO_2(g) + 2H_2O(l) \rightarrow CH_3OH(l) + 3/2O_2(g), \tag{8.4}$$

where the the energy needed is $\triangle$G = 703 kJ/mol ($\triangle$E = 1.21 V).

All $\triangle$G0 values above are positive and large implying that CO_2 reduction is a highly endothermic process, which might make it difficult to proceed with the reaction at ambient temperatures. The distribution of average production rates of major products is different in the gas and liquid phase processes while the light source (UV versus visible) also affects the distribution as presented in Figure 8.1. The total gas production rate is much higher in liquid phases under UV irradiation while the difference significantly diminishes if the visible light sources is employed. Furthermore,

hydrogen production is much higher in the gas phase, and much higher under visible light illumination, which also favours the CO production as the second product. Considering that visible light (i.e., solar energy) must be the main energy source in commercial applications in the future, and the product stream seems to be mainly $H_2 + CO$ (syn-gas), which is a valuable mixture to produce synthetic fuels, the gas phase processes can be said to have some potential to contribute the hydrogen economy.

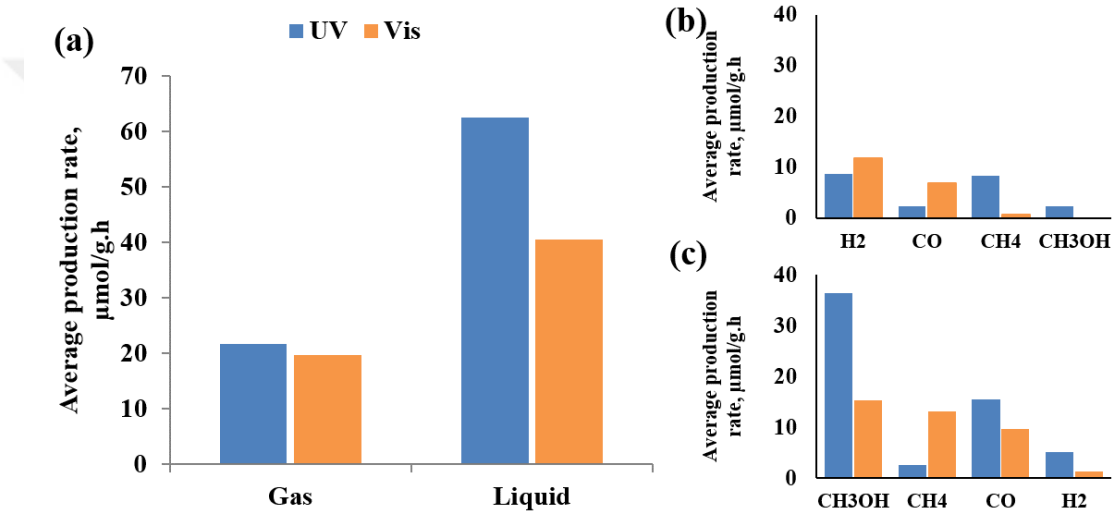

Figure 8.1. Average products in CO_2 reduction (a) total gas production rate (b) product distribution for gas phase processes (c) product distribution for liquid phase processes.

To develop an effective photocatalyst, variables such as material type (both semiconductor and co-catalyst), co-catalyst loading, preparation methods and thermal conditions should be optimized; these variables affect the performance through various photocatalytic properties such as band gap energy, surface area, pore structure, crystal structure and so on [138]. The reaction conditions like the light source and the temperature (especially in the liquid phase) can also influence the performance. Hence, we reviewed the publications included in the dataset and summarized the basic trends below as a pre-analysis to provide some general idea before the machine learning analysis of the dataset; we discussed the semiconductors and co-catalysts in more detail because those variables investigated more frequently due their significance for the process.

Figure 8.2a and Figure 8.2b show the average total gas production rate over various semiconductors used in our dataset for the gas and liquid phase processes, respectively; the data obtained under visible and UV light were presented in copper and blue colours respectively while the bubble sizes indicate the number of data points. In the gas phase, 58 % of data (155 data points) were obtained under UV light while the remaining 113 cases were under visible light; this number is 56% (156 data points) for UV and 125 for visible light in the liquid phase reactions. TiO_2 is the most commonly used semiconductor in both phases, as in the other photocatalytic processes, due to its stability, low cost, abundance in nature, and low impact on human health and environment. The main drawback of this semiconductor is the poor visible light utilization potential due to its relatively larger band gap [324]. This is also apparent from Figure 8.2a and Figure 8.2b that both the number of cases and the average performances of TiO_2 are higher under UV. The difference appears to be small in Figure 8.1a due to the larger size of UV symbol compared to the scale of y-axis but the average performance is 53% higher under UV (35% for the liquid phase). Consequently, a significant number of studies focus on the modification of TiO_2 to improve its photocatalytic activity; elemental doping, the use of noble metal or metal oxide co-catalysts, and adding quantum dots or dyes are the most common methods employed for this purpose. The facet engineering is another important way to enhance the activity of TiO_2; for instance, the highly active facet 001 of TiO_2 crystals exhibited significantly better photocatalytic performance than other common facets [325]. The other semiconductors have been also studied for this purpose even though their frequency of appearances. are low. It can be clearly seen from Figure 8.2a that ZrO_2 is the best semiconductor in the gas phase even though it was studied under UV irradiation only while $SrTiO_3$ and ZnS also performed better than TiO_2. However, the number of cases involving these semiconductors are quite small for reliable generalizations. In the liquid phase, on the other hand, the visible light performance of ZnO is much better than the other semiconductor regardless of lighting; however, the number of cases containing ZnO are also relatively small for a reliable conclusion. Although the number of data points for C_3N_4, CdS, and $ZnGa_2O_4$ are higher than the other alternatives (also C_3N_4 and CdS are more active in visible light), their average performances do not seem to be better

than TiO_2 even though they are usually treated otherwise. For example, g-C_3N_4, with a band gap of 2.7 eV, has been one of the hot research topics in the field in recent years. It is believed to have suitable properties for the catalysis application due to the presence of basic surface sites while its visible light absorption capacity can be further improved by modifying its band gap via various methods such as elemental doping or copolymerization [326].

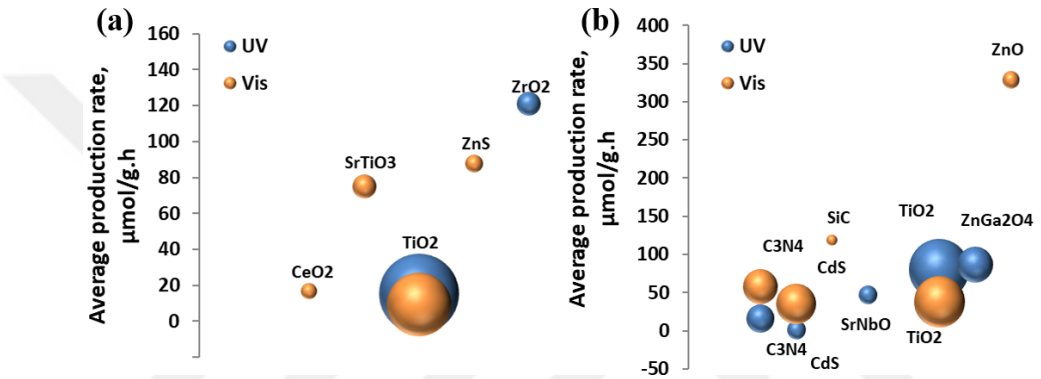

Figure 8.2. Semiconductor used in photocatalytic CO_2 reduction (a) gas phase (b) liquid phase.

Figure 8.3a and Figure 8.3b show the distribution of co-catalysts reported in the dataset for the gas and liquid phases respectively. Similar to the semiconductors, the bubble sizes indicate the number of instances in the dataset while the y-axis represents the average gas production rate obtained with the use of those materials; the data taken under UV (blue symbols) and visible light (copper colour symbols) are also separated. The most commonly used co-catalysts in the dataset are Cu, CuO, Ag, $AgBr$, Pt and In_2O_3 while the cases with no co-catalyst are still the highest group in both phases. Au seems to be the best in the gas phase (under UV) while NiO has the highest average performance in the liquid phase (under UV); however, both co-catalysts were employed less frequently than those mentioned above.

The application of Cu-based co-catalysts for CO_2 conversion, especially with TiO_2, has been widely reported for several decades for mostly liquid phases while there are also cases reported for the gas phase. For instance, Bellardita and coworkers

studied Cu/TiO_2 for CO_2 reduction in the gas phase and observed that CO_2 converted to several hydrocarbons such as methane, methanol, formaldehyde and formic acid [327]; the use of mesoporous silica as the support for Cu/TiO_2 photocatalyst was also reported to improve the selectivity towards the production of methane in gas phase [325]. On the other hand, Pu et. al. found that the CO production was enhanced when Cu was used with CeO_2 (liquid phase) [328]; similarly, Jiang et. al. investigated Cu-based ($CuCO_2O_4$) CdS catalyst on photocatalytic CO_2 reduction and observed CO production as well [329]. Shown and co-workers also reported the efficient photocatalytic conversion of CO_2 to methanol and acetaldehyde over graphene oxide (GO) loaded with Cu nanoparticles under visible-light irradiation in the gas phase [330]. Recently, Zhang and co-workers reported that the use Cu/TiO_2 resulted in CO_2 reduction in the presence of water vapour and CO was found to be the major product with some CH_4 [331]. As reported in the references mentioned above, the use of copper clearly changes the product distribution. However, it does not seem to affect the overall gas production rate (or CO_2 conversion) significantly in the gas phase because the average performance does not seem to be better than no co-catalyst. It (as Cu) seems to improve the total production rate in the liquid phase under UV irradiation while it (as CuO) gives nearly the same performance with the catalyst having no co-catalyst.

The Ag-based co-catalysts have been also used for CO_2 reduction due to their excellent light harvesting and photocatalytic properties; the Ag nanoparticles enhance the visible light adsorption thanks to the localized surface plasmon resonance (LSPR) effects [332]. For instance, Liu et. al. worked on Ag/TiO_2 with varying weight ratios of Ag in the liquid phase and found that 2.5 wt. % of Ag gave the methanol yield of 405.2 μmol/g.cat, which was nine times higher than the TiO_2 alone [333]. Pang and coworkers investigated Ag loading on $SrNb_2O_6$ by three different methods as chemical reduction, incipient-to-wetness impregnation and photodeposition. They found that chemical reduction resulted in higher photocatalytic efficiency and CO production [334]. $AgBr$ has been also used as an effective co-catalyst in various applications. For example, Liu et. al. showed that AgBr/TiO_2 was a promising photocatalyst for

dye degradation application under visible light [176] while Zhang et. al. reported that $AgBr$ on Bi_2WO_6 nanosheets showed good performances in the removal of pollutants [335]. Indeed, Ag has the highest average visible light performance in gas phase Figure 8.3a while its effect, which is mostly tested under UV, is not that apparent in the liquid phase Figure 8.3b; the same is also true for $AgBr$.

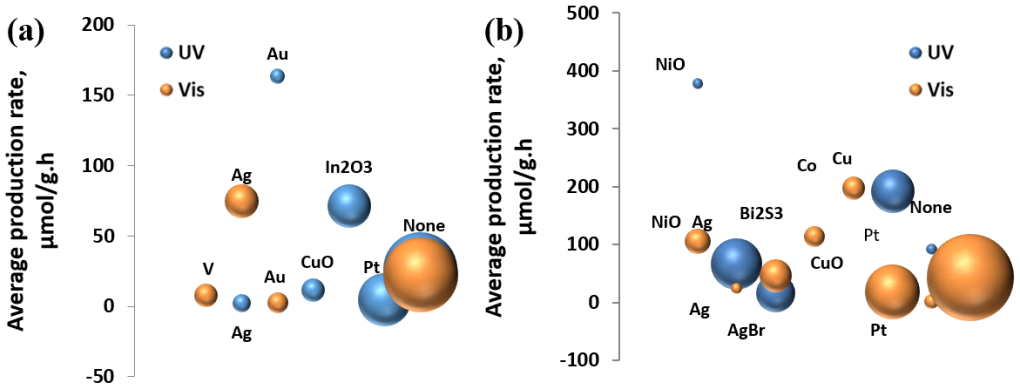

Figure 8.3. Co-catalyst used in photocatalytic CO_2 reduction (a) gas and (b) liquid phase.

Pt is also among the preferred noble metal co-catalysts in photocatalytic applications as it also for the gas phase CO_2 photoreduction processes; it improves charge carrier separation and surface adsorption of various compounds. Pt is mostly used with TiO_2. For instance, Zhang et. al. found that Pt/TiO_2 improved the activity for CH_4 production in the liquid phase; they reported that Pt weight ratio on TiO_2 was important, and 0.12 weight per cent gave more promising results [336]. Pan and coworkers, on the other hand, investigated the Pt weight ratio over various semiconductors in the gas phase, and reported that CeO_2 gave higher yields than ZrO_2 and TiO_2; they also found that 0.12 wt% Pt was optimum for CH_4 production [337]. However, Figure 8.3 indicates that the average gas production rate does not seem to be enhanced by the use of this metal.

In 1989, Haruta et. al. discovered that Au could have an important activity in CO oxidation if it was applied as nanoparticles [338]. Since then, this metal has been

used in many catalytic and photocatalytic applications. In addition to its catalytic role, the interband transitions and surface plasmon resonance effect of Au nanoparticles can also have a significant role in the photocatalytic process [339]. Indeed, Cui et. al. investigated Au nanoclusters together with some metal cations such as Fe2+, CO2+, Ni2+ and Cu2+ on photocatalytic CO_2 reduction in the liquid phase and found that the metal cations can accept photogenerated electrons from Au nanoclusters and serve as catalytic sites for CO_2 reduction [340]. Au-modified TiO_2 microspheres were also investigated by Liu and coworkers (in the gas phase), who reported synergetic effects between light harvesting, charge separation, CO_2 adsorption and activation. They also investigated the effects of Au loading and found that 1 wt.% Au gave the best results for CH_4 production; at the same time, CO production was enhanced along with Au utilization [242].

Indium oxide has also gained attention from researchers in recent years as it is also appearing in our dataset; it is commonly used with SnO_2, which is used a stable conductive transparent layer and thin-film transistors. Additionally, the indium is an efficient co-catalyst that enhances the photocatalytic activity; for instance, Tahir et. al. investigated In_2O_3 for CO_2 photoreduction and reported that it improved photoactivity of TiO_2 towards CH_4 production [341]. Additionally, although it has very small number of cases in our dataset, Au has been also used in many catalytic and photocatalytic applications since Haruta et. al. discovered that Au could have an important activity in CO oxidation if it was applied as nanoparticles in 1989 [342]. In addition to its catalytic role, the interband transitions and surface plasmon resonance effect of Au nanoparticles can also have a significant role in the photocatalytic process [343].

The methods used for semiconductor preparation or co-catalyst loading may be also influential on the CO_2 photoreduction performance of the photocatalysts. Hydrothermal synthesis is the most commonly employed method (159 data points; 80 for gas, 79 for liquid phase) to prepare semiconductors probably because it produces highly homogeneous nanocomposites in mild conditions [342]. Commercially available materials were used without any treatment in 107 data points (51 for gas, 56 for liquid)

while the sol-gel was also employed quite often (75 data points; 51 for gas, 24 for liquid) as a method to produce materials with high surface area and stable surfaces [343]. The other methods applied for this purpose were the thermal methods, chemical reduction, ion exchange, and solid-state reaction. The average production rates for the gas products obtained over the photocatalysts synthesized via various methods are presented in Figure 8.4. Interestingly, commercial materials have the highest production rate under UV light conditions in both phases while hydrothermal synthesis provides a clear advantage in gas phase processes under visible light; the thermal synthesis and the sol-gel seem to be better under visible light irradiation in liquid phase process.

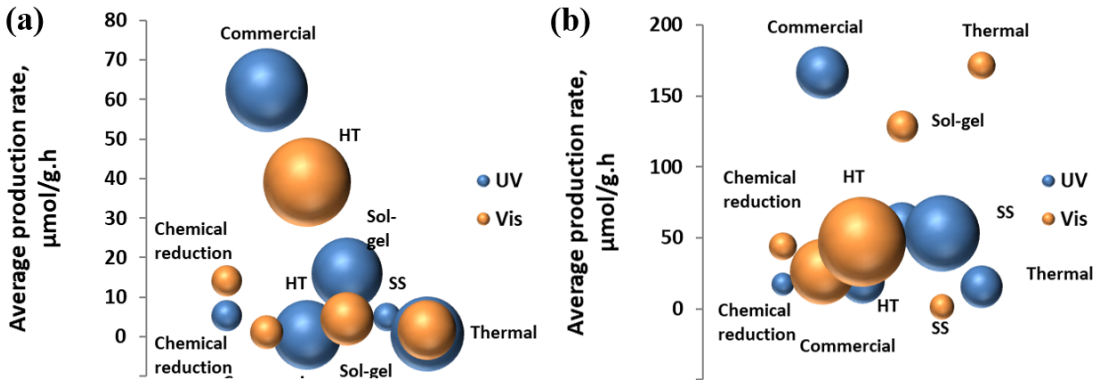

Figure 8.4. Semiconductor preparation methods in photocatalytic CO_2 reduction (a) gas and (b) liquid phase.

The deposition method of the co-catalyst on the semiconductor may be also important; as mentioned above, Pang and coworkers compared the performance of $Ag/SrNb_2O_6$ photocatalysts using three different methods (chemical reduction, incipient-to-wetness impregnation, and photodeposition) for Ag loading and found that the chemical reduction was more efficient for photocatalytic efficiency and CO_2 conversion than others [317]. We also checked the frequency of the deposition methods in our dataset and found that the sol-gel is the most commonly employed method with 70 data points (23% of the entire dataset). Incipient to wetness (62 data points), thermal methods (53 data points) and hydrothermal synthesis method (28 data points) were also used frequently. Although the choice of deposition methods also depends on the type of co-catalyst (some materials are more effective when they are loaded with a

specific method), the data are not sufficient to extract detailed correlations between the individual co-catalysts and deposition method; however, some major trends are still observable as shown in Figure 8.5. The photocatalysts prepared with deposition precipitation (DP) and used under UV conditions clearly resulted in a much higher average gas production rate compared to others in the gas phase while thermal and sol-gel processes are also distinguished from others; it is also interesting that the thermal method results in better performance under visible light (Figure 8.5a). The other methods resulted in a similar level of performance regardless of the irradiation type used in the tests. In the liquid phase, on the other hand, the influences of co-catalyst loading methods are more apparent (Figure 8.5b); the difference between the average visible light and UV performances is also observable. Incipient to wetness impregnation seems to be the best choice for the UV tests while the photocatalysts prepared by sol-gel gave much better performance under visible light.

Although photocatalytic CO_2 reduction should eventually utilize sunlight, most of photocatalysts work better under UV light (Figure 8.1a). While some investigators modified the photocatalysts for visible light harvesting, others continue to study using UV light to understand the mechanism or catalytic phenomena better. Indeed, 51% of the data (280 data points; 155 for gas, 125 for liquid) in our data set were obtained under visible light illumination whereas 49% of the data points (269 data points; 113 for gas, 156 for liquid) were investigated under UV light. Clearly, the liquid phase processes perform better in both UV and visible light while the difference between the average performances under UV and visible irradiation is also significant. In the gas phase, on the other hand, the level of gas production rates obtained under UV appears to be also possible under visible light conditions; however, the gas phase performances in both light conditions are still much lower than those in the liquid phase processes. This seems to influence the choice of the light source as well; the fraction of gas phase data taken under visible light is 58% (155/ 268) while this ratio is only 44% (125/ 281) in the liquid phase.

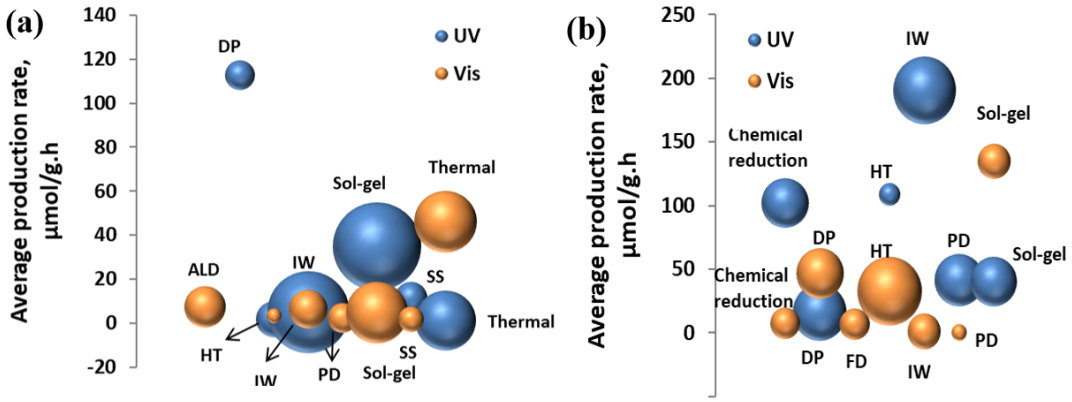

Figure 8.5. Co-catalyst deposition methods in photocatalytic CO_2 reduction (a) gas phase (b) liquid phase.

The data in our dataset indicate that the temperature may be important for the CO_2 reduction rates in the liquid phase reactions probably due to the temperature dependency of CO_2 solubility in water. Indeed, Slamet et. al. investigated the effect of temperature over Cu/TiO_2 photocatalyst in the liquid phase and reported that 348 K resulted in 402 μmol/gcat/h CH_3OH, which was the highest production rate (higher than that at the temperature of 316, and 333 K) [344]. Saladdin et. al. also argued that the kinetics of CO_2 reduction could be modelled with the adsorption /desorption equilibrium. At low temperatures, the products cannot be easily desorbed due to high surface coverage, and become rate-limiting; at high temperatures, on the other hand, the products desorption is much easier [345]. Consequently, the temperature should be at a certain level so that the solubility of CO_2 would be sufficient while the products could be also desorbed easily. No direct relation between temperature and CO_2 reduction was observed in the gas phase data set; the main reason for this result appears to be the fact that the gas phase data were obtained in the narrower temperature range so weaker temperature dependency compared to the liquid phase can be also expected. The gas phase works were usually performed at 298 K whereas the temperature was changed within the range of 278-353 K in the liquid phase experiments.

8.3.2. Model Development for Band Gap Energy

Random forest regression was used to predict the band gap energy for a given set of descriptors (photocatalyst properties); this would help to determine the missing band gap values in the dataset while it may also provide a model to be used with similar studies. As explained in Section 8.2.2, the model with the hyperparameters of node size=5 and ntree=75 (obtained with 10% data as test set and 10-fold cross-validation) was found to be optimum; the plot of actual versus predicted band gap is given for training, validation and testing in Figure 8.6a, Figure 8.6b and Figure 8.6c, respectively. The fitness of the model was measured using the root mean square error (RMSE), which is one of the most common indicators used for this purpose, and found to be 0.19, 0.12 and 0.15 for validation, training and testing, respectively.

As it is clearly seen in Figure 8.6, the predictive power of the model is remarkably high. The variable importance of the descriptors is also given in Figure 8.6d where the effect of each variable on the band gap prediction is seen. Mean square error (MSE) was used for assessing the relative importance of the variables and it shows the contribution of a variable to the model accuracy. As expected, the type of materials and their preparation method play important roles in the band gap of the photocatalyst; as discussed above, the co-catalyst seems to be also influential even though its primary purpose is to help the charge separation. Normally the dopants and their percent are expected to be also important considering that they are used exactly for this purpose; however, the low number of data points involved in doped semiconductors must have hindered its effect doping on the model.

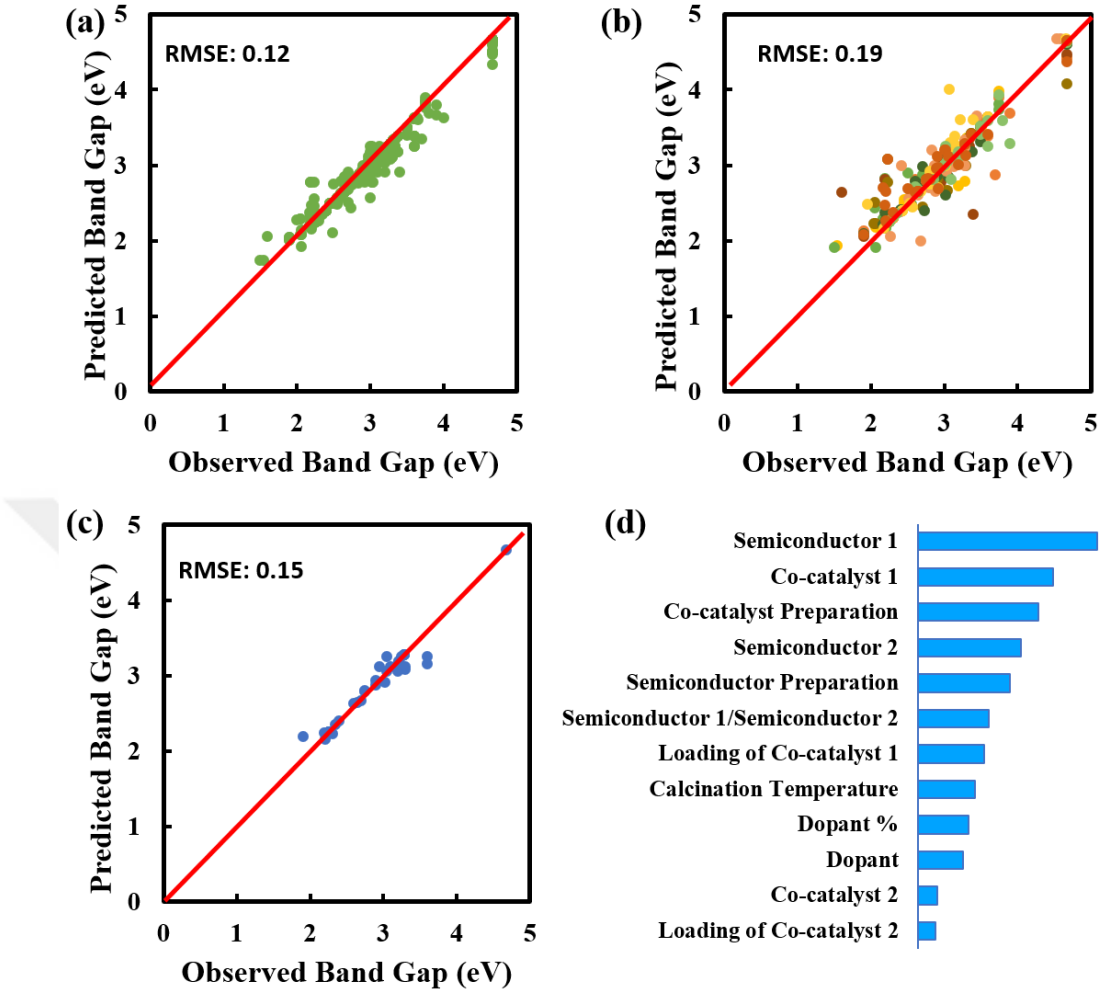

Figure 8.6. Predicted versus real bandgap for (a) validation (b) training (c) testing and (d) importance of input variables in RF model.

8.3.3. Model Development for Total Gas Production Rate

The predictive models for the total gas production rate were also developed for the gas and liquid phases separately (given in Figure 8.7 and Figure 8.8). The RMSE values for the gas phase model were 0.34,0.48 and 0.56 for training, validation and testing sets respectively; for the liquid phase, the RMSE errors were 0.37, 0.52 and 0.57 for training, validation and testing sets respectively. The distribution of the data suggests that there is a tendency to underpredict the total gas production rate for both gas and liquid phase systems. Variable importance plots show that co-catalyst and semiconductor preparation methods have a high impact on the prediction of gas

phase total gas production rate whereas, band gap and reaction temperature as well as the semiconductor preparation method become important for liquid phase total gas production rate.

Although both models have some predictive power, they are not sufficiently accurate for practical purposes; the predictions for unseen data (testing) may not be used to plan future experiments in confidence. The weakness of the models may be arising from the noises in the data because the standard testing and reporting conditions (including the properties of light sources used) are not well established in the field. Other possible explanations are the insufficiency of the dataset due to its small size or the failure to identify and use some important descriptors that may be unknown at this stage. Consequently, we developed these models for order of magnitude estimation and analyzed the total gas production rate using decision tree classification in detail, which is more robust against the noisy data [135], [346, 347]. This technique may provide reliable predictions for the performance range (class) while it can also be used to develop heuristic rules for high performance by inspecting the decision criteria in the individual branches as discussed below.

Decision tree algorithm was used to classify the total gas production rates for the gas phase and liquid phase data separately. The optimum model developed for the gas phase (with the minsplit and complexity parameters of seven and 0.01 respectively) is given in Figure 8.9. The overall accuracy of the model was 80% for training and 79% for testing, which are sufficiently high. The confusion matrix, which shows the classification accuracies for the individual classes, is given in Table 8.4 with the class recall (fraction of data in class X that was correctly classified as class X) and class precision (fraction of actual class X cases in the data that were classified as class X).

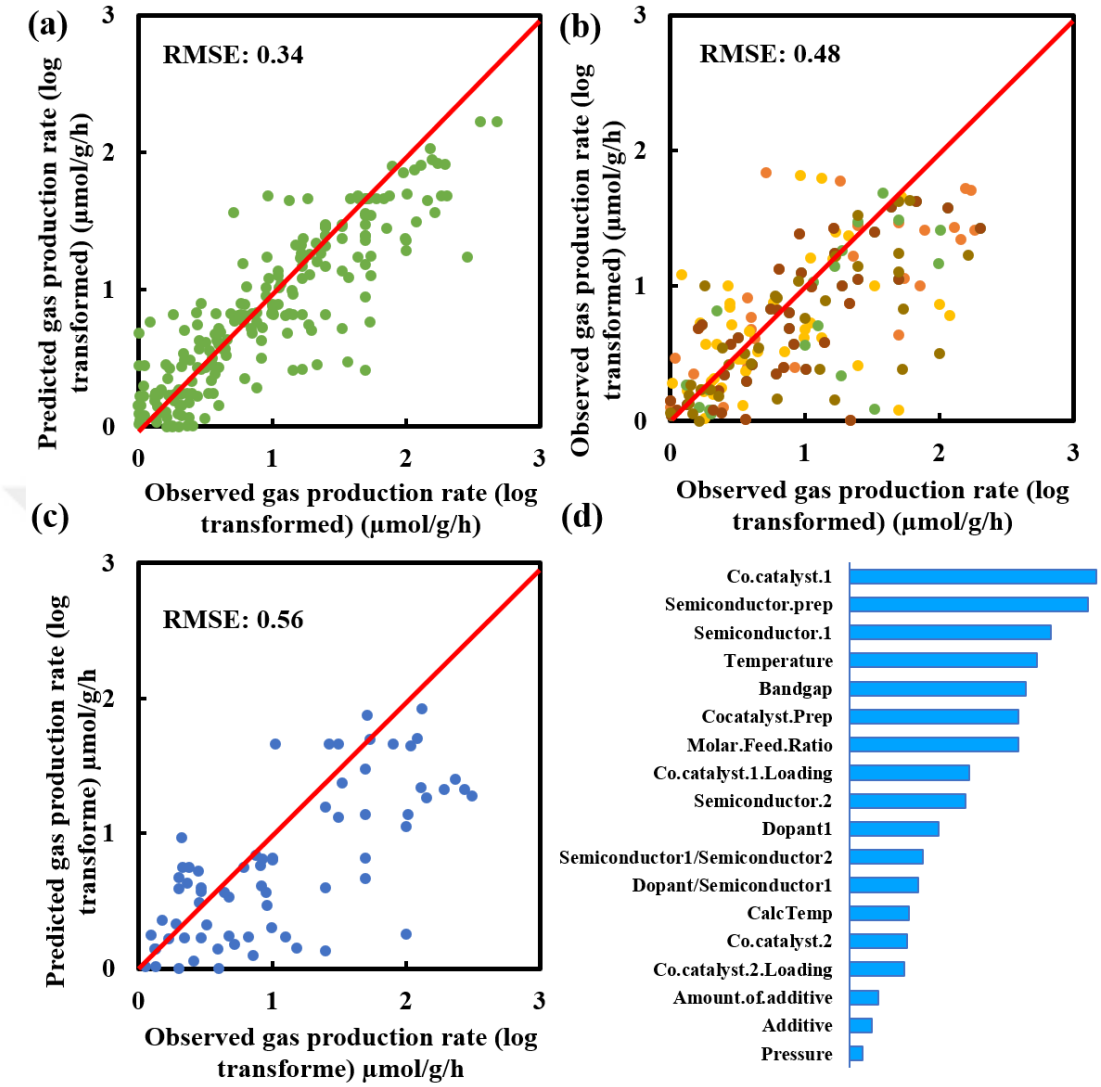

Figure 8.7. Predicted versus real gas phase total gas production rate (log transformed) for (a) validation, (b) training, (c) testing and d) importance of input variables in RF model.

The high precision of class A (81% for training and 71% for testing) is especially desirable because we want to deduce heuristic rules that can be used to reach class A performance. In such aim, it is important to know whether an apparently favourable set of variables is indeed favourable for high performance (expressed by precision); failure to identify one good set of variables (expressed by recall) may not harm if we find another one to work on but labelling an unfavourable set as favourable will lead wrong action. The same may also be true for class C performance if one wants to identify

what should not be done; the precision of class B is usually lower due to the leaks from both sides.

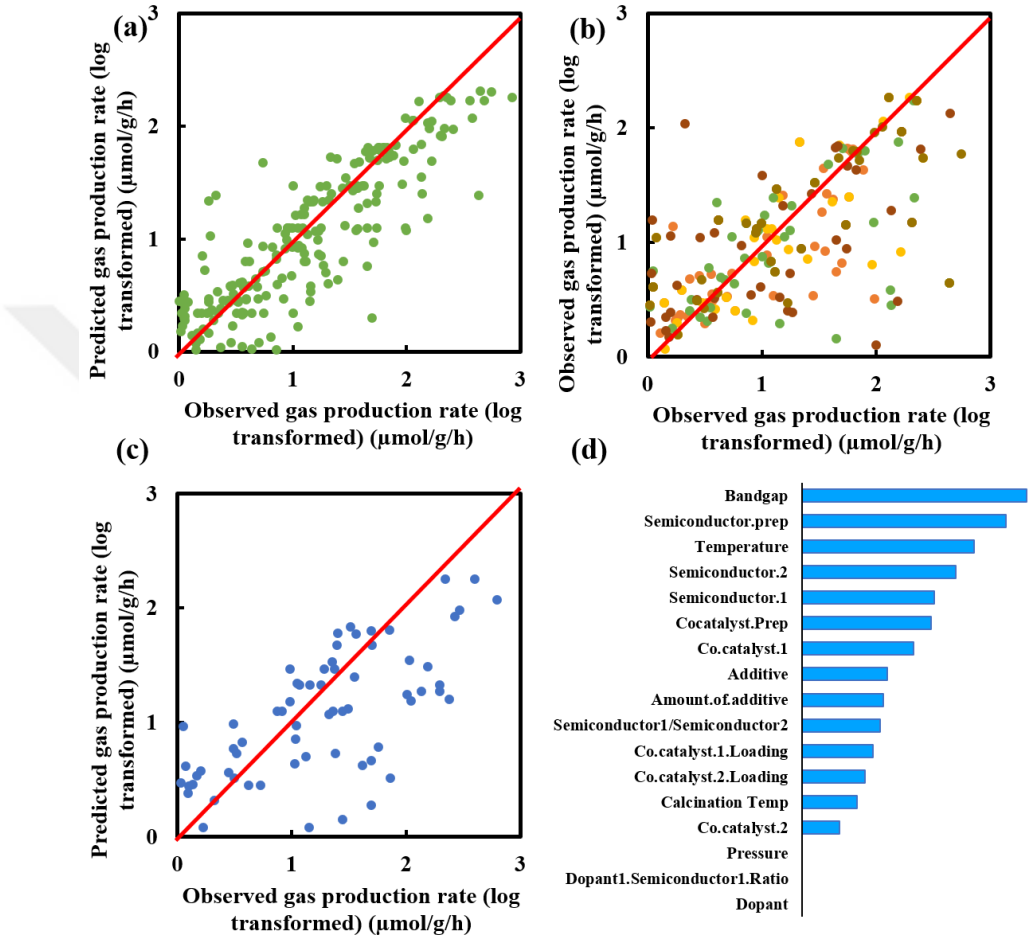

Figure 8.8. Predicted versus real liquid phase total gas production rate (log transformed) for (a) validation, (b) training, (c) testing and d) importance of input variables in RF model.

To deduce heuristic rules, the decision tree model in Figure 8.9 should be read from the top node to the bottom. When the condition (criterion) given for a node is satisfied, the left branch is followed; otherwise, the data are sent to the right. Then, the next criterion is used further down purifying the data in each split until the terminal nodes. The capital letters in the nodes indicate the majority classes while the numbers in medium rows show the fraction of data in each class (from class A on the left to

class C on the right) while the values at the bottom rows indicate the fraction of total data classified in that node. If the fraction of data and purity of one class (especially class A) are significantly high in one terminal node, the splitting criteria used on the branch leading to that terminal node can be used as a heuristic rule to obtain that class of performance.

Table 8.4. Confusion matrix and accuracy for gas phase model.

	Overall Accuracy	**Real Classes**	**# of data Points**	**Predicted class**			**Class Accuracy (Recall)**
				A	**B**	**C**	
Train	0.80	**A**	67	60	78	0	0.89
		B	67	9	49	9	0.73
		C	67	5	11	51	0.76
			Precision	0.81	0.73	0.85	
Test	0.79	**A**	18	15	2	0	0.83
		B	12	3	8	1	0.66
		C	37	3	4	30	0.81
			Precision	0.71	0.53	0.96	

As shown in Figure 8.9, the data is first divided based on the type of semiconductor used; the data was divided whether CeO_2, $SrTiO_3$, ZnS or ZrO_2 as semiconductor resulting in the left-most terminal node which contains 13% of total data (35 out of 268 points representing about 39% of A) with 100% purity in class A (i.e. 35 data points are meeting the conditions described on this branch and all of them represent class A performance). The remaining data was further divided depending on the type of first co-catalyst; the use of CuO or In2S3 resulted in class A level performance in node 6 (24 data points representing 9% of total and 27% of class A data) with 89% purity. The band gap smaller than 3.2 eV and the use of Ag, Modified Zeolite or TCPP as a second semiconductor also lead to class A performance (node 12) with reasonably high purity in A (88%). Since the precision of the class C in our model is also high, the criteria followed through the branch leading to class C nodes (for example node 9) can be also considered as a heuristic rule showing what should be avoided. The terminal

nodes involving class B or low purity in class A or class C cannot be utilized to develop heuristic rules.

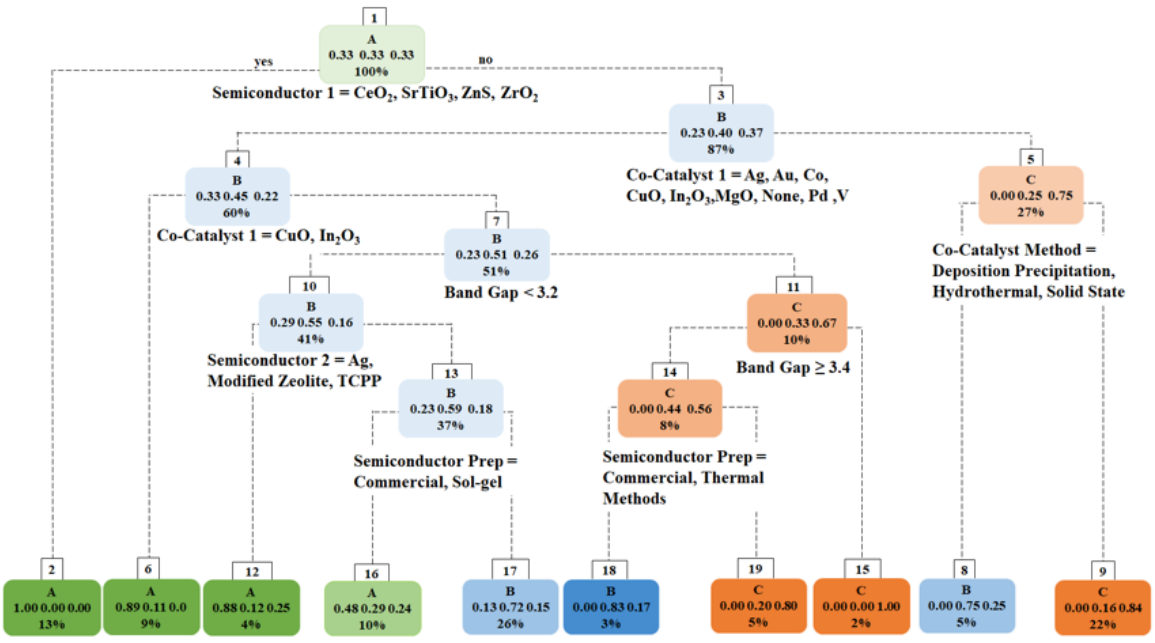

Figure 8.9. Decision tree model for gas phase CO_2 reduction.

The decision tree model for the liquid phase is also given in Figure 8.10. The overall accuracy for training is 77% while it is 65% for testing. Although these are lower than the gas phase model, they are still significant; besides, as we mentioned above, we are interested in class A performance, and the precisions for class A is (95% and 75% for training and testing respectively) are quite high. (Table 8.5) allowing to deduce some useful heuristic rules. Indeed, as it is clearly seen from Figure 8.10, there are terminal nodes with a high number of cases with relatively high purity in class A that can be safely used as heuristics for high CO_2 reduction rate.

Similar to the gas phase data, the first splitting was also done by the semiconductor in the liquid phase data (Figure 8.10); the use of a particular set of semiconductors $BaLa_4Ti_4O_15$, SiC, $SrNb_2O_6$, $ZnGa_2O_4$ and ZnO) resulted in class A performance with high reliability considering that 13% of total data (36 data points out of 281) conform these conditions with 96% purity (meaning that 34 of 36 data points in this

branch is indeed class A data). This seems to be a valuable heuristic rule. On the second terminal node on the left, the tree divides the data by considering the co-catalyst 1 and then by temperature resulting in again a high purity (92%) classification in node 6 indicating that 16 of 17 data points obtained at temperatures higher than 303 K gave high performance probably due to the higher solution of CO_2 and easy desorption of product from the surface as discussed above [346].

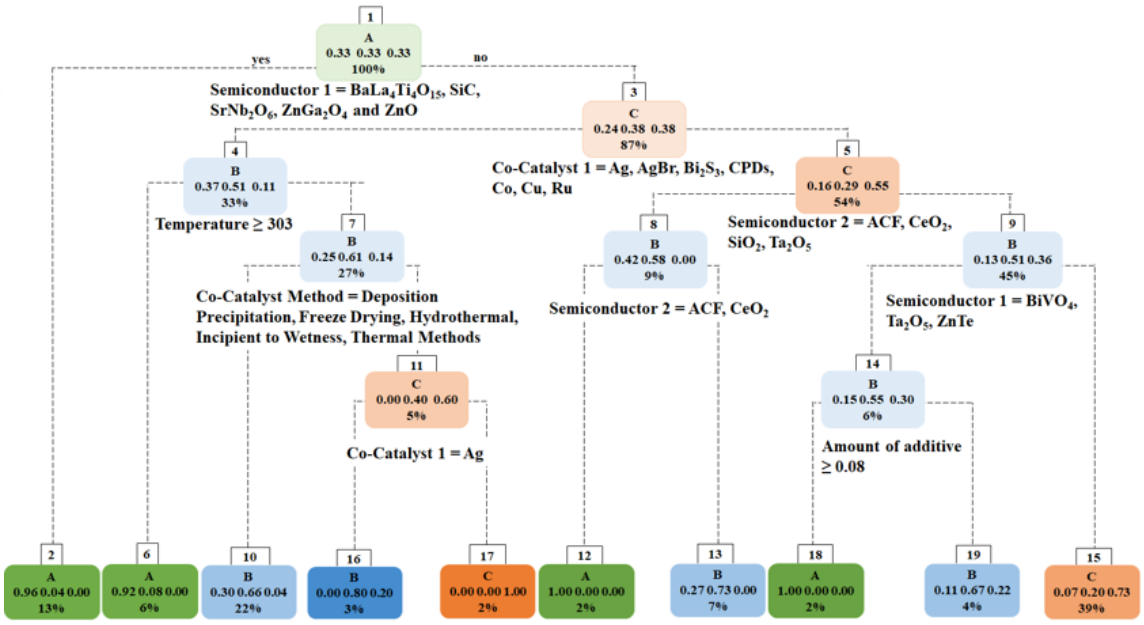

Figure 8.10. Decision tree model for liquid phase CO_2 reduction.

It is important that the tree was quite successful in determining class A performance considering that all nodes rich in A had a purity of more than 92%. The tree also collected the majority of the class C production data at the rightmost node. Although 73% is not as high as those for class A in liquid phase data, it is still acceptable considering that it covers 39% of data; 110 out of 281 total cases are in this node and 80 of them are indeed class C level.

Table 8.5. Confusion matrix and accuracies for liquid phase model.

	Overall Accuracy	**Real Classes**	**# of data Points**	**Predicted class**			**Class Accuracy (Recall)**
				A	**B**	**C**	
Train	0.77	**A**	70	45	19	6	0.64
		B	70	2	52	16	0.74
		C	70	06	5	65	0.92
			Precision	0.95	0.68	0.74	
Test	0.65	**A**	26	12	8	6	0.46
		B	17	4	11	2	0.65
		C	28	0	5	23	0.82
			Precision	0.75	0.46	0.74	

9. CONCLUSION AND RECOMMENDATION

9.1. Conclusion

9.1.1. Bibliometric Analysis of Solar Energy Utilization

The bibliometric analysis of "solar energy utilization" and "machine learning in solar energy utilization" through Web of Science database was conducted and research trends were identified. It was seen that China has taken over the research in solar energy utilization and is about to become a leader also in ML in solar energy. Photovoltaic and hydrogen-related applications of solar energy have been widely studied in the last 20 years and ML was used for forecasting and optimization mainly with neural network and deep learning algorithms in the last 10 years. Co-occurrence networks suggest that the solar energy utilization domain can be divided roughly into two clusters; performance, design and optimization are important concepts in one of the clusters and photocatalysis such as hydrogen production, water splitting, and carbon dioxide reduction are important concepts in the second cluster. ML in solar energy also splits into two clusters; forecasting and optimization. *Forecasting* cluster consists of ML use cases and algorithms whereas *optimization* cluster includes grid applications of solar energy technologies. Trend analysis shows that system green hydrogen, machine learning and decarbonization are being investigated more extensively in recent years and in ML applications hydrogen, monitoring and deep learning research are on the rise.

9.1.2. Analysis of Solar HDH Desalination

A dataset containing 38 articles and 838 data points was constructed from the publications on humidification dehumidification solar desalination. There were a total of 17 descriptors and 3 target variables. Simple statistical analysis showed that water-heated CACW and OAOW systems may produce higher DW under some conditions

even though this is not always the case. RR data obtained with CACW, CAOW and OACW are scattered allowing the comparison of average performance, which decreases with the same order. This much obvious for GOR data indicating that CACW with heated water stream has much better performance than OACW with heated air. The most common packaging materials are cellulose and plastics; while plastic occasionally provides better DW and GOR results, the trend for RR is less conclusive. As expected, all three of the output variables increase with inlet water temperature.

Boruta analysis showed that descriptors indicate whether water heating is used or not, and geometries of humidifier and dehumidifier (rectangular or cylindrical) are not important for DW and RR models, hence they were disregarded. Best model for DW prediction was achieved with random forest with ntree of 125 and mtry of 10, and had the validation, training, and testing RMSE of 2.7, 0.99 and 7.8 respectively. Random forest also gave higher predictive power for RR with ntree of 15 and mtry of 10, and resulted in the validation, training, and testing RMSE of 0.014, 0.01 and 0.035 respectively. Gradient boosting regression had the highest predictive power for GOR prediction with ntree of 20, interaction depth of 4 and shrinkage of 0.46; the validation, training, and testing RMSE were 0.25, 0.19 and 0.13 respectively. The predictive power of all models was reasonably well.

9.1.3. Analysis of DSSC Performance

A comprehensive dataset about synthetic dyes used in the DSSC between 2009 and 2022 has been constructed and analyzed using machine learning tools. PCE, FF, J_{sc} and V_{OC} were predicted with low RMSE values with gradient boosting and random forest algorithms. RMSE for gradient boosting for PCE % prediction in the test set was 1.55%; that of J_{sc} was 3.85 mA/cm^2; similarly, for random forest model for FF and V_{OC} prediction resulted in test RMSEs of 0.1 and 0.05 V respectively. The variable importance plots show that details related to the type of dye and type of counter electrode are main determining variables for DSSC performance for our dataset.

9.1.4. Analysis of Natural DSSC Performance

A dataset of 522 data coming from 113 published articles was created for the prediction of performance measures of natural DSSC. Pre-analysis revealed that TiO_2 is the most commonly used anode material and other materials do not generally result in better performance. Treatment of TiO_2 layer with $TiCl_4$ or TTIP seems to have a positive influence on the FF but does not significantly change efficiency and other measures. Anthocyanin is the most commonly used dye (by itself or in a cocktail dye with others), and its performance is generally higher than the dye types. In co-sensitization, the dye ratio is important and dye ratio between two and three seems to have a positive effect on the natural DSSC performance. Ionic liquid presence in the electrolyte affects V_{OC} and FF positively but results in no improvement in efficiency. V_{OC} prediction model had RMSE of 0.002, 0.09 and 0.07 for training, validation and testing set respectively; J_{sc} was model had RMSE of 0.5, 1.3 and 1.5 for training, validation and testing set respectively; FF has RMSE of 0.04, 0.1, and 0.08 for training, validation and testing set respectively and finally PCE model had RMSE of 0.03, 0.4 and 0.2 for training, validation and testing set respectively. The variable importance shows that dye-related input variables were important for all performance measures, however, deposition method, band gap, treatment and electrolyte solvent were also important variables specific to open circuit voltage, short circuit current density, fill factor and efficiency respectively.

9.1.5. Analysis of Photoelectrochemical Water Splitting

The simple descriptive statistics provided some overall evaluation of the dataset while the association rule mining was successfully employed for both band gap and photocurrent density as the output. Decision tree model for the band gap was quite successful (with the overall training and testing accuracies of 78% and 72% respectively) while the fitness of the model decreased significantly for the photocurrent density (training accuracies= 61%; testing accuracy = 54%) probably due to the non-standard testing (especially irradiation) conditions. The predictive model developed for the band

gap using random forest was also quite remarkable (with the testing root mean square error 0.27, respectively); however, no such model could be developed for photocurrent density probably due to the uncertainties mentioned above. The most commonly used semiconductor in the dataset for the first layer of electrodes is TiO_2 (29%) followed by Fe_2O_3 (17%), ZnO (10%), $BiVO_4$ (10%), and WO_3 (9); the remaining 25% of data involved various less frequently used semiconductors. In most of the experiments, the semiconductors were used without doping, and the doping rarely improved the average photocurrent densities when it was employed. The preparation method used for the semiconductor at the first layer of the electrode influenced the band gap significantly. Except Fe_2O_3, none of the most commonly used semiconductors have a significant fraction of high photocurrent density cases at high bias. It was also found that $BiVO_4$ photoanode performed better without co-catalyst at low bias conditions while some co-catalyst like $CoPi$, AgS, Co-Ni, and Pd helped to have better results at higher bias conditions. For TiO_2, the use no co-catalyst gives high photocurrent in low and medium bias while C doping is effective in low and medium bias for high photocurrent density. No rule could be deduced for Fe_2O_3 for high photocurrent density at low bias. For ZnO photoanode, on the other hand, CdS has positive effects at low and medium bias conditions while the use of CdSe as a second layer is also beneficial.

9.1.6. Analysis of CO_2 Photoreduction over Metal Oxides

In summary, an extensive dataset was constructed from the published papers on the photoreduction of CO_2 and analyzed using machine learning techniques The conclusions obtained from simple descriptive analysis can be summarized as follows: H_2, CO, and CH_4 are the main products in gas phase while CH_3OH production is more dominant in liquid phase. Hydrogen production is much higher in gas phase, and much higher under visible light, which also favors the CO production as the secondary product. TiO_2 (324 instances) is the most common semiconductor as expected while C_3N_4 (42 instances) and CdS (34 instances) are also utilized often. The most commonly used co-catalysts are Cu (96 instances), Ag (40 instances) and Pt (30 instances); Au gives the most efficient result in gas phase (under UV) whereas NiO is better in liquid

phase (also under UV). Preparation methods of catalysts were also significant for the performance.

The random forest prediction of bandgap was quite successful; the model developed was used to compute the missing bandgap values for the analysis of the total gas production rate indicating that the bandgap of the photocatalysts can be accurately predicted for a given photocatalyst constituents and preparation conditions. Decision tree is suited much better for the total gas production rates for both gas and liquid phase data; some useful heuristic rules. For example, it was found that, in the gas phase, CeO_2, $SrTiO_3$, ZnS and ZrO_2 usually give a high total gas production rate; otherwise, the co-catalyst (like CuO and In_2O_3 became important. The co-catalyst and its deposition methods also seem to influence the low total gas production rate. In the liquid phase, on the other hand, $BaLa_4Ti_4O_15$, SiC, $SrNb_2O_6$, $ZnGa_2O_4$ and ZnO) resulted in class A performance regardless of the other conditions within the range of dataset; higher temperatures seem to also promote high total gas production rate for the other semiconductors with the conditions that some co-catalyst has to be used.

9.2. Recommendations

While machine learning offers tremendous potential for enhancing the efficiency of solar energy technologies, several challenges need to be addressed. The availability of high-quality data for training machine learning models is crucial. Additionally, the interpretability and transparency of machine learning algorithms need to be ensured, as they play a major role in such critical applications. Furthermore, scalability and computational complexity are important considerations for implementing machine learning algorithms in real-world solar energy systems. In upcoming studies, solar energy systems with high commercialization potential could be investigated for cost reduction, demand management and fault detection [348] , besides further insights from already commercialized systems such as photovoltaics could be analyzed using deep learning tools since the size of the real-world performance data is growing each day.

Researchers could develop hybrid models that combine both machine learning algorithms and physical models to achieve the best-performing models. Machine learning integrated with physic-based models could improve the predictive power of the models since the algorithms would learn the underlying physics behind the process. There are several examples of this approach in materials science and catalyst research [349, 350]. In addition, the analysis of high-throughput experimental data could be valuable for designing better experiments and optimizing the conditions. There are various databases based on DFT studies which can be utilized in new material research or the expansion of the feature set; a study focusing on the prediction of performance using experimental data like the ones in this dissertation could be enhanced using materials properties from databases like Materials Project [351] and ICSD [352].

The improved explainability of the machine learning models would be an important development for a better understanding of the underlying effects and patterns in the data. There are various explainable machine learning applications in the computer science domain, and transferring those applications into experimental data analysis could guide researchers into developing better-performing solar energy technologies. Computer vision algorithms could also improve the model power and generate new insights; for instance, through examination of SEM images of perovskite solar cells, researchers successfully predicted the distribution of the perovskite crystal sizes [353].

REFERENCES

1. Abu El-Maaty, A. E., M. M. Awad, G. I. Sultan and A. M. Hamed, "Innovative Approaches to Solar Desalination: a Comprehensive Review of Recent Research", *Energies*, Vol. 16, No. 9, pp. 30–46, 2023.

2. Shakerian, M., M. Karrabi, M. Gheibi, A. M. Fathollahi-Fard and M. Hajiaghaei-Keshteli, "Evaluating the Performance of a Solar Distillation Technology in the Desalination of Brackish Waters", *Processes*, Vol. 10, No. 8, pp. 2–11, 2022.

3. Tiwari, G. N. and L. Sahota, "History of Passive Solar-Distillation Systems", *Green Energy and Technology*, Vol. 10, No. 1, pp. 10–18, 2017.

4. Bamasag, A., E. Almatrafi, T. Alqahtani, P. Phelan, M. Ullah, M. Mustakeem, M. Obaid and N. Ghaffour, "Recent Advances and Future Prospects in Direct Solar Desalination Systems Using Membrane Distillation Technology", *Journal of Cleaner Production*, Vol. 385, No. 211, pp. 230–245, 2023.

5. Ahmed, F., M. Sharizal Abdul Aziz, P. Palaniandy and F. Shaik, "A Review on Application of Renewable Energy for Desalination Technologies with Emphasis on Concentrated Solar Power", *Sustainable Energy Technologies and Assessments*, Vol. 53, No. 21, pp. 1230–1245, 2022.

6. Ali, M. T., H. E. Fath and P. R. Armstrong, "A Comprehensive Techno-Economical Review of Indirect Solar Desalination", *Renewable and Sustainable Energy Reviews*, Vol. 15, No. 8, pp. 110–120, 2011.

7. Burschka, J., N. Pellet, S. J. Moon, R. Humphry-Baker, P. Gao, M. K. Nazeeruddin and M. GräTzel, "Sequential Deposition As a Route to High-Performance Perovskite-Sensitized Solar Cells", *Nature*, Vol. 499, No. 7458, pp. 413–426, 2013.

8. Odabasi, C. and R. Yıldırım, "Performance Analysis of Perovskite Solar Cells

in 2013–2018 Using Machine-Learning Tools", *Nano Energy*, Vol. 56, No. 4, pp. 770–791, 2019.

9. Yılmaz, B., C. Odabası and R. Yıldırım, "Efficiency and Stability Analysis of 2D/3D Perovskite Solar Cells Using Machine Learning", *Energy Technology*, Vol. 10, No. 3, pp. 11–23, 2022.

10. Vidal, R., J. A. Alberola-Borràs, N. SáNchez-Pantoja and I. Mora-Seró, "Comparison of Perovskite Solar Cells with Other Photovoltaics Technologies from the Point of View of Life Cycle Assessment", *Advanced Energy and Sustainability Research*, Vol. 2, No. 5, pp. 550–560, 2021.

11. Alizadeh, A., M. Roudgar-Amoli, S. M. Bonyad-Shekalgourabi, Z. Shariatinia, M. Mahmoudi and F. Saadat, "Dye Sensitized Solar Cells Go Beyond Using Perovskite and Spinel Inorganic Materials: a Review", *Renewable and Sustainable Energy Reviews*, Vol. 157, No. 19, pp. 150–160, 2022.

12. Etacheri, V., C. Di Valentin, J. Schneider, D. Bahnemann and S. C. Pillai, "Visible-Light Activation of TiO2 Photocatalysts: Advances in Theory and Experiments", *Journal of Photochemistry and Photobiology C: Photochemistry Reviews*, Vol. 25, No. 14, pp. 12–29, 2015.

13. Shih, C. F., T. Zhang, J. Li and C. Bai, "Powering the Future with Liquid Sunshine", *Joule*, Vol. 2, No. 10, pp. 145–157, 2018.

14. He, Q., H. Zheng, X. Ma, L. Wang, H. Kong and Z. Zhu, "Artificial Intelligence Application in a Renewable Energy-Driven Desalination System: a Critical Review", *Energy and Artificial Intelligence*, Vol. 7, No. 4, pp. 340–354, 2022.

15. Abid, A., M. A. Jamil, N. Us Sabah, M. U. Farooq, H. Yaqoob, L. A. Khan and M. W. Shahzad, "Exergoeconomic Optimization of a Forward Feed Multi-Effect Desalination System with and Without Energy Recovery", *Desalination*, Vol. 499,

No. 10, pp. 110–118, 2021.

16. Parikh, N., M. Karamta, N. Yadav, M. Mahdi Tavakoli, D. Prochowicz, S. Akin, A. Kalam, S. Satapathi and P. Yadav, "Is Machine Learning Redefining the Perovskite Solar Cells?", *Journal of Energy Chemistry*, Vol. 66, No. 53, pp. 234–246, 2022.

17. Li, F., X. Peng, Z. Wang, Y. Zhou, Y. Wu, M. Jiang and M. Xu, "Machine Learning (Ml)-Assisted Design and Fabrication for Solar Cells", *Energy and Environmental Materials*, Vol. 2, No. 4, pp. 253–266, 2019.

18. Singh, A. K., R. Gorelik and T. Biswas, "Data-Driven Discovery of Robust Materials for Photocatalytic Energy Conversion", *Annual Review of Condensed Matter Physics*, Vol. 14, No. 9, pp. 237–259, 2023.

19. Ge, L., Y. Ke and X. Li, "Machine Learning Integrated Photocatalysis: Progress and Challenges", *Chemical Communications*, Vol. 59, No. 39, pp. 136–148, 2023.

20. Mai, H., T. C. Le, D. Chen, D. A. Winkler and R. A. Caruso, "Machine Learning for Electrocatalyst and Photocatalyst Design and Discovery", *Chemical Reviews*, Vol. 122, No. 16, pp. 152–168, 2022.

21. Ahmad, T., D. Zhang, C. Huang, H. Zhang, N. Dai, Y. Song and H. Chen, "Artificial Intelligence in Sustainable Energy Industry: Status Quo, Challenges and Opportunities", *Journal of Cleaner Production*, Vol. 289, No. 3, p. 125834, 2021.

22. Cai, J., X. Chu, K. Xu, H. Li and J. Wei, "Machine Learning-Driven New Material Discovery", *Nanoscale Advances*, Vol. 2, No. 8, pp. 3115–3130, 2020.

23. Alpaydin, E., *Introduction to Machine Learning*, The MIT Press, London, 3rd edn., 2014.

24. Lim, P., C. K. Goh and K. C. Tan, "Evolutionary Cluster-Based Synthetic Oversampling Ensemble (Eco-Ensemble) for Imbalance Learning", *Transactions on Cybernetics*, Vol. 47, No. 9, pp. 15–21, 2017.

25. Khalid, S., T. Khalil and S. Nasreen, "A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning", *Proceedings of 2014 Science and Information Conference*, Vol. 1, pp. 372–378, New York, 2014.

26. Ying, X., "An Overview of Overfitting and Its Solutions", *Journal of Physics*, Vol. 1168, No. 2, pp. 150–158, 2019.

27. Dufera, A. G., T. Liu and J. Xu, "Regression Models of Pearson Correlation Coefficient", *Statistical Theory and Related Fields*, Vol. 7, No. 2, pp. 143–156, 2023.

28. Xie, S., Y. Zhang, D. Lv, X. Chen, J. Lu and J. Liu, "A New Improved Maximal Relevance and Minimal Redundancy Method based on Feature Subset", *Journal of Supercomputing*, Vol. 79, No. 3, pp. 43–56, 2023.

29. Almomani, O., "A Feature Selection Model for Network Intrusion Detection System based on Pso, Gwo, Ffa and Ga Algorithms", *Symmetry*, Vol. 12, No. 6, pp. 1210–1230, 2020.

30. Mcneish, D. M., "Using Lasso for Predictor Selection and to Assuage Overfitting: a Method Long Overlooked in Behavioral Sciences", *Multivariate Behavioral Research*, Vol. 50, No. 5, pp. 10–20, 2015.

31. Trivedi, U. B., M. Bhatt and P. Srivastava, "Prevent Overfitting Problem in Machine Learning: a Case Focus on Linear Regression and Logistics Regression", *Advances in Science, Technology and Innovation*, Vol. 5, No. 15, pp. 110–120, 2021.

32. Guenther, N. and M. Schonlau, "Support Vector Machines", *Stata journal*,

Vol. 16, No. 4, pp. 1510–1560, 2016.

33. Suykens, J. A., "Nonlinear Modelling and Support Vector Machines", *Instrumentation and Measurement Technology*, Vol. 1, No. 50, pp. 510–560, 2001.

34. Burges, C. J., "A Tutorial on Support Vector Machines for Pattern Recognition", *Data Mining and Knowledge Discovery*, Vol. 2, No. 2, pp. 5210–5260, 1998.

35. Shawe-Taylor, J. and S. Sun, "A Review of Optimization Methodologies in Support Vector Machines", *Neurocomputing*, Vol. 74, No. 17, pp. 50–60, 2011.

36. James, G., D. Witten, T. Hastie and R. Tibshirani, *An Introduction to Statistical Learning*, Springer Texts in Statistics, Springer New York, New York, 2013.

37. Breiman, L., J. H. Friedman, R. A. Olshen and C. J. Stone, "Classification and Regression Trees", *Classification and Regression Trees*, 2017.

38. Breiman, L., "Random Forests", *Machine Learning*, Vol. 45, No. 10, pp. 5–32, 2001.

39. Jaffari, Z. H., A. Abbas, S. M. Lam, S. Park, K. Chon, E. S. Kim and K. H. Cho, "Machine Learning Approaches to Predict the Photocatalytic Performance of Bismuth Ferrite-Based Materials in the Removal of Malachite Green", *Journal of Hazardous Materials*, Vol. 442, No. 15, pp. 1336–1355, 2023.

40. Schmidhuber, J., "Deep Learning in Neural Networks: An Overview", *Neural Networks*, Vol. 61, No. 1, pp. 85–117, 2015.

41. Florek, D. and M. Milosz, "Comparison of An Effectiveness of Artificial Neural Networks for Various Activation Functions", *Journal of Computer Sciences Institute*, Vol. 26, No. 5, pp. 346–352, 2023.

42. Lecun, Y., Y. Bengio and G. Hinton, "Deep Learning", *Nature*, Vol. 521, No.

7553, pp. 436–444, 2015.

43. Specht, D., "A General Regression Neural Network", *IEE Transactions on Neural Networks*, Vol. 2, No. 6, pp. 1550–1560, 1991.

44. Wang, J., S. Lu, S. H. Wang and Y. D. Zhang, "A Review on Extreme Learning Machine", *Multimedia Tools and Applications*, Vol. 81, No. 29, pp. 550–560, 2022.

45. Joseph, V. R., "Optimal Ratio for Data Splitting", *Statistical Analysis and Data Mining*, Vol. 15, No. 4, 2022.

46. Pawluszek-Filipiak, K. and A. Borkowski, "On the Importance of Train-Test Split Ratio of Datasets in Automatic Landslide Detection by Supervised Classification", *Remote Sensing*, Vol. 12, No. 18, pp. 1210–1215, 2020.

47. Salazar, J. J., L. Garland, J. Ochoa and M. J. Pyrcz, "Fair Train-Test Split in Machine Learning: Mitigating Spatial Autocorrelation for Improved Prediction Accuracy", *Journal of Petroleum Science and Engineering*, Vol. 209, 2022.

48. Rodriguez, J. D., A. Perez and J. A. Lozano, "Sensitivity Analysis of K-Fold Cross Validation in Prediction Error Estimation", *IEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32, No. 3, pp. 162–178, 2010.

49. Li, J., "Assessing the Accuracy of Predictive Models for Numerical Data: not R nor R2, Why Not? Then What?", *Plos One*, Vol. 12, No. 8, p. 142–158, 2017.

50. Nicolaisen, J., "Bibliometrics and Citation Analysis: from the Science Citation Index to Cybermetrics", *Journal of the American Society for Information Science and Technology*, Vol. 61, No. 1, pp. 1532–2882, 2010.

51. Beard, E. J. and J. M. Cole, "Perovskite- and Dye-Sensitized Solar-Cell Device Databases Auto-Generated Using Chemdataextractor", *Scientific Data*, Vol. 9, No. 1, pp. 205–213, 2022.

52. Naves, A. X., C. Barreneche, A. I. FernáNdez, L. F. Cabeza, A. N. Haddad and D. Boer, "Life Cycle Costing As a Bottom Line for the Life Cycle Sustainability Assessment in the Solar Energy Sector: a Review", *Solar Energy*, Vol. 192, No. 15, pp. 138–149, 2019.

53. David, T. M., P. M. Silva Rocha Rizol, M. A. Guerreiro Machado and G. P. Buccieri, "Future Research Tendencies for Solar Energy Management Using a Bibliometric Analysis, 2000–2019", *Heliyon*, Vol. 6, No. 7, pp. 2405–2440, 2020.

54. Ferruzzi, G., C. Delcea, A. Barberi, V. D. Dio, M. D. Somma, P. Catrini, S. Guarino, F. Rossi, M. L. Parisi, A. Sinicropi and S. Longo, "Concentrating Solar Power : the State of the Art , Research Gaps and Future Perspectives", *Energies*, Vol. 16, No. 8082, pp. 1–41, 2023.

55. CalderóN, A., C. Barreneche, K. HernáNdez-Valle, E. Galindo, M. Segarra and A. I. FernáNdez, "Where Is Thermal Energy Storage (Tes) Research Going? – a Bibliometric Analysis", *Solar Energy*, Vol. 200, No. 69, pp. 380–392, 2020.

56. Michas, S., V. Stavrakas, N. A. Spyridaki and A. Flamos, "Identifying Research Priorities for the Further Development and Deployment of Solar Photovoltaics", *International Journal of Sustainable Energy*, Vol. 38, No. 3, pp. 276–296, 2019.

57. Obaideen, K., A. G. Olabi, Y. Al Swailmeen, N. Shehata, M. A. Abdelkareem, A. H. Alami, C. Rodriguez and E. T. Sayed, "Solar Energy: Applications, Trends Analysis, Bibliometric Analysis and Research Contribution to Sustainable Development Goals (Sdgs)", *Sustainability*, Vol. 15, No. 2, pp. 657–665, 2023.

58. Yeo, J. S. and Y. Jeong, "Pathway Toward Market Entry of Perovskite Solar Cells: a Detailed Study on the Research Trends and Collaboration Networks Through Bibliometrics", *Energy Reports*, Vol. 6, No. 2, pp. 2075–2085, 2020.

59. Fonteyn, P., S. Lizin and W. Maes, "The Evolution of the Most Important Re-

search Topics in Organic and Perovskite Solar Cell Research from 2008 to 2017: a Bibliometric Literature Review Using Bibliographic Coupling Analysis", *Solar Energy Materials and Solar Cells*, Vol. 207, No. 51, pp. 927–932, 2020.

60. Du, H., N. Li, M. A. Brown, Y. Peng and Y. Shuai, "A Bibliographic Analysis of Recent Solar Energy Literatures: the Expansion and Evolution of a Research Field", *Renewable Energy*, Vol. 66, No. 3, p. 09181, 2014.

61. De Paulo, A. F. and G. S. Porto, "Solar Energy Technologies and Open Innovation: a Study based on Bibliometric and Social Network Analysis", *Energy Policy*, Vol. 108, No. 13, pp. 91–101, 2017.

62. Ajibade, S. S. M., D. D. C. Flores, M. Ayaz, Y. A. Dodo, F. O. Areche, A. O. Adediran, O. J. Oyebode and J. P. Dayupay, "Application of Machine Learning in Renewable Energy: a Bibliometric Analysis of a Decade", *2023 International Conference on Automatic Control and Intelligent Systems*, Vol. 1, pp. 65–70, San Francisco, 2023.

63. Aria, M. and C. Cuccurullo, "Bibliometrix: An R-Tool for Comprehensive Science Mapping Analysis", *Journal of Informetrics*, Vol. 11, No. 4, pp. 1875–1889, 2017.

64. Kumar, B., J. M. Smieja, A. F. Sasayam and C. P. Kubiak, "Tunable, Light-Assisted Co-Generation of Co and H2 from CO2 and H2O by Re(Bipy-Tbu)(Co)3cl and P-Si in Non-Aqueous Medium", *Chemical Communications*, Vol. 48, No. 2, p. 13645, 2012.

65. Alotaibi, B., H. P. Nguyen, S. Zhao, M. G. Kibria, S. Fan and Z. Mi, "Highly Stable Photoelectrochemical Water Splitting and Hydrogen Generation Using a Double-Band Ingan/Gan Core/Shell Nanowire Photoanode", *Nano Letters*, Vol. 13, No. 9, pp. 4356–4361, 2013.

66. Natividad, R., "CO2 Photoconversion Catalyzed by Nanoparticles Supported on

TiO2", *Nanoparticles in Green Organic Synthesis: Strategy Towards Sustainability*, Vol. 124, No. 19, pp. 99–110, 2023.

67. Ni, D., H. Shen, H. Li, Y. Ma and T. Zhai, "Synthesis of High Efficient Cu/TiO2 Photocatalysts by Grinding and Their Size-Dependent Photocatalytic Hydrogen Production", *Applied Surface Science*, Vol. 409, No. 19, p. 16332, 2017.

68. Nakabi, T. A. and P. Toivanen, "Deep Reinforcement Learning for Energy Management in a Microgrid with Flexible Demand", *Sustainable Energy, Grids and Networks*, Vol. 25, pp. 2352–2367, 2021.

69. Muriithi, G. and S. Chowdhury, "Optimal Energy Management of a Grid-Tied Solar Pv-Battery Microgrid: a Reinforcement Learning Approach", *Energies*, Vol. 14, No. 9, pp. 199–210, 2021.

70. Leo, R., R. S. Milton and S. Sibi, "Reinforcement Learning for Optimal Energy Management of a Solar Microgrid", *Global Humanitarian Technology*, Vol. 1, No. 1, pp. 10–18, 2014.

71. Khan, M. A., I. Al-Shankiti, A. Ziani and H. Idriss, "Demonstration of Green Hydrogen Production Using Solar Energy At 28% Efficiency and Evaluation of Its Economic Viability", *Sustainable Energy and Fuels*, Vol. 5, No. 4, pp. 239–248, 2021.

72. Karayel, G. K., N. Javani and I. Dincer, "Green Hydrogen Production Potential for Turkey with Solar Energy", *International Journal of Hydrogen Energy*, Vol. 47, No. 45, pp. 360–379, 2022.

73. Zhu, X., P. Gui, X. Zhang, Z. Han and Y. Li, "Multi-Objective Optimization of a Hybrid Energy System Integrated with Solar-Wind-Pemfc and Energy Storage", *Journal of Energy Storage*, Vol. 72, No. 5, pp. 139–148, 2023.

74. Allouhi, A., M. Benzakour Amine and C. Reisch, "Multi-Objective Optimization

of Solar Energy Systems for Electricity and Hot Water Generation in Collective Residential Buildings Considering the Power-to-Heat Concept", *Applied Thermal Engineering*, Vol. 230, No. 15, pp. 1359–1364, 2023.

75. Wibowo, A., M. A. Marsudi, M. I. Amal, M. B. Ananda, R. Stephanie, H. Ardy and L. J. Diguna, "ZnO Nanostructured Materials for Emerging Solar Cell Applications", *Royal Science of Chemistry Advances*, Vol. 10, No. 70, pp. 2046–2069, 2020.

76. Khalafi, T., F. Buazar and K. Ghanemi, "Phycosynthesis and Enhanced Photocatalytic Activity of Zinc Oxide Nanoparticles Toward Organosulfur Pollutants", *Scientific Reports*, Vol. 9, No. 1, pp. 1045–1057, 2019.

77. LimóN-Rocha, I., C. A. GuzmáN-GonzáLez, L. M. Anaya-Esparza, R. Romero-Toledo, J. L. Rico, O. A. GonzáLez-Vargas and A. PeRez-Larios, "Effect of the Precursor on the Synthesis of ZnO and Its Photocatalytic Activity", *Inorganics*, Vol. 10, No. 2, pp. 2304–2314, 2022.

78. Berghout, T., M. Benbouzid, T. Bentrcia, X. Ma, S. DjuroviĆ and L. H. Mouss, "Machine Learning-Based Condition Monitoring for Pv Systems: State of the Art and Future Prospects", *Energies*, Vol. 14, No. 19, 2021.

79. Ledmaoui, Y., A. E. Maghraoui and A. Chebak, "Solar Charging Station for Electric Vehicles with Iot Solution for Monitoring Energy Production", *Communications in Computer and Information Science*, Vol. 1677 Ccis, 2022.

80. Huang, C. and M. Yang, "Memory Long and Short Term Time Series Network for Ultra-Short-Term Photovoltaic Power Forecasting", *Energy*, Vol. 279, pp. 2057–2062, 2023.

81. Sabri, M. and M. El Hassouni, "Photovoltaic Power Forecasting with a Long Short-Term Memory Autoencoder Networks", *Soft Computing*, Vol. 27, No. 15,

pp. 387–396, 2023.

82. Essam, Y., A. N. Ahmed, R. Ramli, K. W. Chau, M. S. Idris Ibrahim, M. Sherif, A. Sefelnasr and A. El-Shafie, "Investigating Photovoltaic Solar Power Output Forecasting Using Machine Learning Algorithms", *Engineering Applications of Computational Fluid Mechanics*, Vol. 16, No. 1, pp. 100–115, 2022.

83. Gupta, S., A. Kumar and K. K. Tejavath, "A Pharmacognostic Approach for Mitigating Pancreatic Cancer: Emphasis on Herbal Extracts and Phytoconstituents", *Future Journal of Pharmaceutical Sciences*, Vol. 7, No. 1, p. 96, 2021.

84. Shorabeh, S. N., N. N. Samany, F. Minaei, H. K. Firozjaei, M. Homaee and A. D. Boloorani, "A Decision Model based on Decision Tree and Particle Swarm Optimization Algorithms to Identify Optimal Locations for Solar Power Plants Construction in Iran", *Renewable Energy*, Vol. 187, No. 10, pp. 1879–1788, 2022.

85. Amidpour, M., M. Salimi and W. He, *Advances in Sustainable Humidification-Dehumidification Thermal Desalination System Integrated with Renewable and Hybrid Energy Resources 4e Analysis, Process Integration, and Material*, Elsevier B.V, Amsterdam, 2024.

86. Advisory, A., *Water Desalination Market - Global Outlook & Forecast 2021-2026*, Technavio, London, 2021.

87. Curcio, E., G. D. Profio, E. Fontananova and E. Drioli, "Membrane Technologies for Seawater Desalination and Brackish Water Treatment", *Advances in Membrane Technologies for Water Treatment: Materials, Processes and Applications*, Vol. 15, No. 9, pp. 411–441, 2015.

88. Goh, P. S., W. J. Lau, M. H. Othman and A. F. Ismail, "Membrane Fouling in Desalination and Its Mitigation Strategies", *Desalination*, Vol. 425, pp. 130–155, 2018.

89. Nematollahi, F., A. Rahimi and T. T. Gheinani, "Experimental and Theoretical Energy and Exergy Analysis for a Solar Desalination System", *Desalination*, Vol. 317, No. 23, pp. 23–31, 2013.

90. Ho, C. D., H. M. Yeh and R. C. Wang, "Heat-Transfer Enhancement in Double-Pass Flat-Plate Solar Air Heaters with Recycle", *Energy*, Vol. 30, No. 15, pp. 2796–2817, 2005.

91. Giwa, A., N. Akther, A. A. Housani, S. Haris and S. W. Hasan, "Recent Advances in Humidification Dehumidification (Hdh) Desalination Processes: Improved Designs and Productivity", *Renewable and Sustainable Energy Reviews*, Vol. 57, No. 5, pp. 929–944, 2016.

92. Zarei, T. and M. R. Miroliaei, "Performance Evaluation of An Hdh Desalination System Using Direct Contact Packed Towers: Experimental and Mathematical Modeling Study", *Journal of Water Reuse and Desalination*, Vol. 12, No. 1, pp. 92–110, 2022.

93. Chen, C., Y. Zuo, W. Ye, X. Li, Z. Deng and S. P. Ong, "A Critical Review of Machine Learning of Energy Materials", *Advanced Energy Materials*, Vol. 10, No. 8, p. 1903242, 2020.

94. Oral, B., E. Can and R. Yildirim, "Analysis of Photoelectrochemical Water Splitting Using Machine Learning", *International Journal of Hydrogen Energy*, Vol. 47, No. 45, pp. 19633–19654, 2022.

95. Saadetnejad, D., B. Oral, E. Can and R. Yıldırım, "Machine Learning Analysis of Gas Phase Photocatalytic CO2 Reduction for Hydrogen Production", *International Journal of Hydrogen Energy*, Vol. 47, No. 45, pp. 19655–19668, 2022.

96. Ray, S. S., R. K. Verma, A. Singh, M. Ganesapillai and Y. N. Kwon, "A Holistic Review on How Artificial Intelligence Has Redefined Water Treatment and Sea-

water Desalination Processes", *Desalination*, Vol. 546, No. 19, p. 116221, 2023.

97. Priya, P., T. C. Nguyen, A. Saxena and N. R. Aluru, "Machine Learning Assisted Screening of Two-Dimensional Materials for Water Desalination", *Nano*, Vol. 16, No. 2, pp. 1929–1939, 2022.

98. Marichal Plasencia, G. N., J. Camacho-Espino, D. ávila Prats and B. PeÑate Suárez, "Machine Learning Models Applied to Manage the Operation of a Simple Swro Desalination Plant and Its Application in Marine Vessels", *Water*, Vol. 13, No. 18, pp. 856–869, 2021.

99. Khayet, M. and C. Cojocaru, "Artificial Neural Network Model for Desalination by Sweeping Gas Membrane Distillation", *Desalination*, Vol. 308, No. 19, pp. 102–110, 2013.

100. Acevedo, L., J. Uche and A. Del-Amo, "Improving the Distillate Prediction of a Membrane Distillation Unit in a Trigeneration Scheme by Using Artificial Neural Networks", *Water*, Vol. 10, No. 3, p. 310, 2018.

101. K, A., A. Mungray, S. Agarwal, J. Ali and M. Chandra Garg, "Performance Optimisation of Forward-Osmosis Membrane System Using Machine Learning for the Treatment of Textile Industry Wastewater", *Journal of Cleaner Production*, Vol. 289, No. 95, p. 125690, 2021.

102. Kandeal, A. W., M. An, X. Chen, A. M. Algazzar, A. Kumar Thakur, X. Guan, J. Wang, M. R. Elkadeem, W. Ma and S. W. Sharshir, "Productivity Modeling Enhancement of a Solar Desalination Unit with Nanofluids Using Machine Learning Algorithms Integrated with Bayesian Optimization", *Energy Technology*, Vol. 9, No. 9, p. 2100189, 2021.

103. Wang, Y., A. Kandeal, A. Swidan, S. W. Sharshir, G. B. Abdelaziz, M. Halim, A. Kabeel and N. Yang, "Prediction of Tubular Solar Still Performance by Ma-

chine Learning Integrated with Bayesian Optimization Algorithm", *Applied Thermal Engineering*, Vol. 184, No. 2, p. 116233, 2021.

104. Sohani, A., S. Hoseinzadeh, S. Samiezadeh and I. Verhaert, "Machine Learning Prediction Approach for Dynamic Performance Modeling of An Enhanced Solar Still Desalination System", *Journal of Thermal Analysis and Calorimetry*, Vol. 147, No. 5, pp. 3919–3930, 2022.

105. Kiran Naik, B., M. Chinthala, S. Patel and P. Ramesh, "Performance Assessment of Waste Heat/Solar Driven Membrane-Based Simultaneous Desalination and Liquid Desiccant Regeneration System Using a Thermal Model and Knn Machine Learning Tool", *Desalination*, Vol. 505, No. 9, pp. 100–119, 2021.

106. Faegh, M., P. Behnam, M. B. Shafii and M. Khiadani, "Development of Artificial Neural Networks for Performance Prediction of a Heat Pump Assisted Humidification-Dehumidification Desalination System", *Desalination*, Vol. 508, No. 2, p. 115052, 2021.

107. Nia, B. B., M. Khorrami, I. Farahbakhsh, M. Amin and F. Chekab, "Application of Artificial Intelligence and Machine Learning in Desalination : a Bibliometric and Review Study", *Desalination and Water Purification*, Vol. 10, No. 1, pp. 136–145, 2021.

108. Koch, P., "Equivalent Diameters of Rectangular and Oval Ducts", *Building Services Engineering Research and Technology*, Vol. 29, No. 4, pp. 341–347, 2008.

109. Liaw, A. and M. Wiener, "Classification and Regression by Randomforest", *R News*, Vol. 2, No. 3, pp. 18–22, 2002.

110. Greenwell, B., B. Boehmke and J. Cunningham, "Generalized Boosted Regression Models", *Cran- R packages*, Vol. 1, No. 1, pp. 1–39, 2022.

111. Kursa, M. B. and W. R. Rudnicki, "Boruta: Wrapper Algorithm for All Relevant

Feature Selection", *Cran- R packages*, Vol. 1, No. 1, pp. 1–17, 2020.

112. Wickham, H., *Ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York, New York, 2nd edn., 2016.

113. Sherrill-Mix, S. and C. Dawson, "Plot Categorical Data Using Quasirandom Noise and Density Estimates", *Cran- R packages*, Vol. 1, No. 1, pp. 1–37, 2017.

114. Zhao, Y., H. Zheng, S. Liang, N. Zhang and X. Long Ma, "Experimental Research on Four-Stage Cross Flow Humidification Dehumidification (Hdh) Solar Desalination System with Direct Contact Dehumidifiers", *Desalination*, Vol. 467, No. 7, pp. 147–157, 2019.

115. Rahimi-Ahar, Z., M. S. Hatamipour and L. R. Ahar, "Air Humidification-Dehumidification Process for Desalination: a Review", *Progress in Energy and Combustion Science*, Vol. 80, No. 9, p. 100850, 2020.

116. Ahmed, H. A., I. M. Ismail, W. F. Saleh and M. Ahmed, "Experimental Investigation of Humidification-Dehumidification Desalination System with Corrugated Packing in the Humidifier", *Desalination*, Vol. 410, No. 191, pp. 19–29, 2017.

117. Garg, K., S. K. Das and H. Tyagi, "Thermal Design of a Humidification-Dehumidification Desalination Cycle Consisting of Packed-Bed Humidifier and Finned-Tube Dehumidifier", *International Journal of Heat and Mass Transfer*, Vol. 183, No. 2, p. 122153, 2022.

118. Chiranjeevi, C. and T. Srinivas, "Design and Development of An Air Humidifier Using Finite Difference Method for a Solar Desalination Plant", *Materials Science and Engineering*, Vol. 263, No. 11, p. 062041, 2017.

119. Hanania, J., K. Stenhouse, Yyelland. Brodie and J. Donev, "Photovoltaic Effect - Energy Education", *Energy Education*, Vol. 20, No. 6, pp. 1062–1079, 2017.

120. Blandford, R. and M. Watkins, "This Month in Physics History: April 25, 1954: Bell Labs Demonstrates the First Practical Silicon Solar Cell", *Aps News*, Vol. 18, No. 4, pp. 1220–1239, 2009.

121. King, R. R., D. Bhusari, D. Larrabee, X. Q. Liu, E. Rehder, K. Edmondson, H. Cotal, R. K. Jones, J. H. Ermer, C. M. Fetzer, D. C. Law and N. H. Karam, "Solar Cell Generations over 40% Efficiency", *Progress in Photovoltaics: Research and Applications*, Vol. 20, No. 6, pp. 1062–1079, 2012.

122. O'regan, B. and M. GräTzel, "A Low-Cost, High-Efficiency Solar Cell based on Dye-Sensitized Colloidal TiO2 Films", *Nature*, Vol. 353, No. 6346, pp. 836–852, 1991.

123. Iftikhar, H., G. G. Sonai, S. G. Hashmi, A. F. Nogueira and P. D. Lund, "Progress on Electrolytes Development in Dye-Sensitized Solar Cells", *Materials*, Vol. 12, No. 12, pp. 1944–1956, 2019.

124. Pettersson, H. and T. Gruszecki, "Long-Term Stability of Low-Power Dye-Sensitised Solar Cells Prepared by Industrial Methods", *Solar Energy Materials and Solar Cells*, Vol. 70, No. 2, pp. 997–1008, 2001.

125. Niedzwiedzki, D. M., "Photophysical Properties Ofn719andz907dyes, Benchmark Sensitizers for Dye-Sensitized Solar Cells, At Room and Low Temperature", *Physical Chemistry Chemical Physics*, Vol. 23, No. 10, pp. 1463–1476, 2021.

126. Aboulouard, A., B. Gultekin, M. Can, M. Erol, A. Jouaiti, B. Elhadadi, C. Zafer and S. Demic, "Dye Sensitized Solar Cells based on Titanium Dioxide Nanoparticles Synthesized by Flame Spray Pyrolysis and Hydrothermal Sol-Gel Methods: a Comparative Study on Photovoltaic Performances", *Journal of Materials Research and Technology*, Vol. 9, No. 2, pp. 1569–1577, 2020.

127. Kumar, R., V. Sahajwalla and P. Bhargava, "Fabrication of a Counter Electrode

for Dye-Sensitized Solar Cells (DSSCs) Using a Carbon Material Produced with the Organic Ligand 2-Methyl-8-Hydroxyquinolinol (Mq)", *Nanoscale Advances*, Vol. 1, No. 8, pp. 251–262, 2019.

128. GonzáLez-Verjan, V. A., B. Trujillo-Navarrete, R. M. FeLix-Navarro, J. N. De LeóN, J. M. Romo-Herrera, J. C. Calva-YáÑEz, J. M. HernáNdez-Lizalde and E. A. Reynoso-Soto, "Effect of TiO2 Particle and Pore Size on DSSC Efficiency", *Materials for Renewable and Sustainable Energy*, Vol. 9, No. 2, pp. 219–224, 2020.

129. Wang, X., A. Bolag, W. Yun, Y. Du, C. Eerdun, X. Zhang, T. Bao, J. Ning, H. Alata and T. Ojiyed, "Enhanced Performance of Dye-Sensitized Solar Cells based on a Dual Anchored Diphenylpyranylidene Dye and N719 Co-Sensitization", *Journal of Molecular Structure*, Vol. 1206, No. 15, p. 22860, 2020.

130. Ramasamy, S., M. Bhagavathiachari, S. A. Suthanthiraraj and M. Pichai, "Mini Review on the Molecular Engineering of Photosensitizer: Current Status and Prospects of Metal-Free/Porphyrin Frameworks At the Interface of Dye-Sensitized Solar Cells", *Dyes and Pigments*, Vol. 203, No. 36, pp. 187–193, 2022.

131. Sharma, K., V. Sharma and S. S. Sharma, "Dye-Sensitized Solar Cells: Fundamentals and Current Status", *Nanoscale Research Letters*, Vol. 13, No. 8, pp. 125–139, 2018.

132. Arooj, Q. and F. Wang, "Switching on Optical Properties of D-Π-A DSSC Sensitizers from Π-Spacers Towards Machine Learning", *Solar Energy*, Vol. 188, No. 2, pp. 380–392, 2019.

133. Yılmaz, B. and R. Yıldırım, "Critical Review of Machine Learning Applications in Perovskite Solar Research", *Nano Energy*, Vol. 80, No. 12, p. 105546, 2021.

134. Oral, B., B. Tekin, D. Eroglu and R. Yildirim, "Performance Analysis of Na-Ion

Batteries by Machine Learning", *Journal of Power Sources*, Vol. 549, No. 5, p. 232126, 2022.

135. Can, E. and R. Yildirim, "Data Mining in Photocatalytic Water Splitting over Perovskites Literature for Higher Hydrogen Production", *Applied Catalysis B: Environmental*, Vol. 242, No. 3, pp. 267–283, 2019.

136. R Core Team, "R Core Team 2021 R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.R-Project.Org/", *R Foundation for Statistical Computing*, Vol. 2, No. 1, pp. 1–14, 2022.

137. Hornik, K., A. Weingessel, F. Leisch and M. D. M. Davidmeyerr-Projectorg, "Package 'E1071'", *Cran- R packages*, Vol. 1, No. 1, pp. 1–66, 2022.

138. Kuhn, M., "Building Predictive Models in R Using the Caret Package", *Journal of Statistical Software*, Vol. 28, No. 8, pp. 1–26, 2008.

139. Nguyen, t. V., H. C. Lee, M. Alam Khan and O. B. Yang, "Electrodeposition of TiO2/SiO2 Nanocomposite for Dye-Sensitized Solar Cell", *Solar Energy*, Vol. 81, No. 4, pp. 529–534, 2007.

140. Gupta, D., S. Mukhopadhyay and K. S. Narayan, "Fill Factor in Organic Solar Cells", *Solar Energy Materials and Solar Cells*, Vol. 94, No. 8, pp. 1309–1313, 2010.

141. Sadikin, S. N., M. Y. Rahman, A. A. Umar and T. H. Aziz, "Improvement of Dye-Sensitized Solar Cell Performance by Utilizing Graphene-Coated TiO2 Films Photoanode", *Superlattices and Microstructures*, Vol. 128, No. 7, pp. 1096–1103, 2019.

142. Bhojanaa, K. B., J. J. Mohammed, M. Manishvarun and A. Pandikumar, "Dye-Sensitized Solar Cells with Efficiency Enhancement Surpassing 65% Through

Layer-By-Layer Assembled Plasmonic Photoanodes", *Journal of Power Sources*, Vol. 558, No. 7, p. 3783, 2023.

143. Fallah, M., I. Maleki, M. R. Zamani-Meymian and Y. Abdi, "Enhancing the Efficiency of Dye-Sensitized Solar Cell by Increasing the Light Trapping and Decreasing the Electron-Hole Recombination Rate Due to Ag@TiO2 Core-Shell Photoanode Structure", *Materials Research Express*, Vol. 7, No. 1, 2019.

144. Duan, Y., Q. Tang, Z. Chen, B. He and H. Chen, "Enhanced Dye Illumination in Dye-Sensitized Solar Cells Using TiO2/Geo2 Photo-Anodes", *Journal of Materials Chemistry A*, Vol. 2, No. 31, pp. 12459–12465, 2014.

145. Yeoh, M. E. and K. Y. Chan, "Efficiency Enhancement in Dye-Sensitized Solar Cells with ZnO and TiO2 Blocking Layers", *Journal of Electronic Materials*, Vol. 48, No. 7, pp. 361–372, 2019.

146. Yuan, S., Q. Tang, B. He, L. Men and H. Chen, "Transmission Enhanced Photoanodes for Efficient Dye-Sensitized Solar Cells", *Electrochimica Acta*, Vol. 125, No. 16, pp. 106–112, 2014.

147. Wu, J., Z. Lan, J. Lin, M. Huang, Y. Huang, L. Fan, G. Luo, Y. Lin, Y. Xie and Y. Wei, "Counter Electrodes in Dye-Sensitized Solar Cells", *Chemical Society Reviews*, Vol. 46, No. 19, pp. 146–157, 2017.

148. Lin, J. Y. and S. W. Chou, "Highly Transparent NiCO2S4 Thin Film As An Effective Catalyst Toward Triiodide Reduction in Dye-Sensitized Solar Cells", *Electrochemistry Communications*, Vol. 37, No. 10, pp. 1775–1789, 2013.

149. Huang, Y. J., M. S. Fan, C. T. Li, C. P. Lee, T. Y. Chen, R. Vittal and K. C. Ho, "Mose2 Nanosheet/Poly(3,4-Ethylenedioxythiophene): Poly(Styrenesulfonate) Composite Film As a Pt-Free Counter Electrode for Dye-Sensitized Solar Cells", *Electrochimica Acta*, Vol. 211, No. 12, pp. 1775–1789, 2016.

150. Seo, H., M. K. Son, N. Itagaki, K. Koga and M. Shiratani, "Polymer Counter Electrode of Poly(3,4-Ethylenedioxythiophene):Poly(4-Styrenesulfonate) Containing TiO2 Nano-Particles for Dye-Sensitized Solar Cells", *Journal of Power Sources*, Vol. 307, No. 2, pp. 775–789, 2016.

151. Kaliramna, S., S. S. Dhayal, R. Chaudhary, S. Khaturia, K. L. Ameta and N. Kumar, "A Review and Comparative Analysis of Different Types of Dyes for Applications in Dye-Sensitized Solar Cells", *Brazilian Journal of Physics*, Vol. 52, No. 4, pp. 167–184, 2022.

152. Yan, W., K. Chaitanya, Z. D. Sun and X. H. Ju, "Theoretical Study on p-Type D-π-A Sensitizers with Modified π-Spacers for Dye-Sensitized Solar Cells", *Journal of Molecular Modeling*, Vol. 24, No. 3, pp. 140–157, 2018.

153. Hoshikawa, T., T. Ikebe, R. Kikuchi and K. Eguchi, "Effects of Electrolyte in Dye-Sensitized Solar Cells and Evaluation by Impedance Spectroscopy", *Electrochimica Acta*, Vol. 51, No. 25, pp. 101–126, 2006.

154. Pulfrey, D. L., "On the Fill Factor of Solar Cells", *Solid State Electronics*, Vol. 21, No. 3, pp. 519–520, 1978.

155. Filipic, M., M. Berginc, F. Smole and M. Topic, "Analysis of Electron Recombination in Dye-Sensitized Solar Cell", *Current Applied Physics*, Vol. 12, No. 1, pp. 156–167, 2012.

156. Kouhnavard, M., D. Yifan, J. M. D' Arcy, R. Mishra and P. Biswas, "Highly Conductive PEDOT Films with Enhanced Catalytic Activity for Dye-Sensitized Solar Cells", *Solar Energy*, Vol. 211, No. 11, pp. 258–264, 2020.

157. Devadiga, D., M. Selvakumar, P. Shetty and M. S. Santosh, "Dye-Sensitized Solar Cell for Indoor Applications: a Mini-Review", *Journal of Electronic Materials*, Vol. 50, No. 6, 2021.

158. O'reilly, T., C. Andrei, J. O'reilly and D. Zerulla, "The Influence of Light Intensity, Active Area, and Excitation Wavelength on the Temporal Response of a Dye Sensitized Solar Cell", *Optical Modeling and Measurements for Solar Energy Systems*, Vol. 7410, No. 5, pp. 110–118, 2009.

159. Andari, R., "Distance Variation of Light Source Effects Toward Dye Sensitized Solar Cell (DSSC) Performance Using Anthocyanin Extract from Rosella Flower", *Journal of Physical Science and Engineering*, Vol. 5, No. 1, pp. 255–265, 2020.

160. Bera, S., D. Sengupta, S. Roy and K. Mukherjee, "Research into Dye-Sensitized Solar Cells: a Review Highlighting Progress in India", *Journal of Physical Energy*, Vol. 3, No. 3, pp. 2515–2525, 2021.

161. Saidi, N. M., N. K. Farhana, S. Ramesh and K. Ramesh, "Influence of Different Concentrations of 4-Tert-Butyl-Pyridine in a Gel Polymer Electrolyte Towards Improved Performance of Dye-Sensitized Solar Cells (DSSC)", *Solar Energy*, Vol. 216, No. 54, pp. 380–392, 2021.

162. Al-Alwani, M. A., A. B. Mohamad, N. A. Ludin, A. A. H. Kadhum and K. Sopian, "Dye-Sensitised Solar Cells: Development, Structure, Operation Principles, Electron Kinetics, Characterisation, Synthesis Materials and Natural Photosensitisers", *Renewable and Sustainable Energy Reviews*, Vol. 65, No. 18, pp. 183–213, 2016.

163. Richhariya, G., A. Kumar, P. Tekasakul and B. Gupta, "Natural Dyes for Dye Sensitized Solar Cell: a Review", *Renewable and Sustainable Energy Reviews*, Vol. 69, No. 5, pp. 705–718, 2017.

164. Chayal, G., K. R. Patel, M. S. Roy, M. Kumar, N. Prasad and K. Shitiz, "Effect of Surface Treatment on Photovoltaic Properties of Dye-Sensitized Solar Cell based on Natural Dye Quercetin", *J. Indian Chem. Soc*, Vol. 96, No. 8, pp. 1059–1065, 2019.

165. Venkatraman, V., R. Raju, S. P. Oikonomopoulos and B. K. Alsberg, "The Dye-Sensitized Solar Cell Database", *Journal of Cheminformatics*, Vol. 10, No. 1, pp. 1–9, 2018.

166. Venkatraman, V. and L. K. Chellappan, "An Open Access Data Set Highlighting Aggregation of Dyes on Metal Oxides", *Data*, Vol. 5, No. 2, pp. 157–167, 2020.

167. Maddah, H. A., "Decision Trees based Performance Analysis for Influence of Sensitizers Characteristics in Dye-Sensitized Solar Cells", *Journal of Advances in Information Technology*, Vol. 13, No. 3, pp. 271–276, 2022.

168. Clarke, E., S. Sherrill-Mix and C. Dawson, "Package 'Ggbeeswarm'", *Cran- R packages*, Vol. 1, No. 1, pp. 1–40, 2017.

169. Wilcox, R. R., "Robust and Exploratory Regression", *Applying Contemporary Statistical Techniques*, Vol. 2, No. 1, pp. 457–515, 2003.

170. Bartesaghi, D., I. D. C. PeRez, J. Kniepert, S. Roland, M. Turbiez, D. Neher and L. J. A. Koster, "Competition Between Recombination and Extraction of Free Charges Determines the Fill Factor of Organic Solar Cells", *Nature Communications*, Vol. 6, No. 5, pp. 259–267, 2015.

171. Zhou, W., X. Jia, L. Chen, Z. Yin, Z. Zhang and G. Gao, "Low Cost Nis As An Efficient Counter Electrode for Dye-Sensitized Solar Cells", *Materials Letters*, Vol. 163, No. 19, pp. 1146–1157, 2016.

172. Pepe, G., J. M. Cole, P. G. Waddell and J. R. Griffiths, "Molecular Engineering of Fluorescein Dyes As Complementary Absorbers in Dye Co-Sensitized Solar Cells", *Molecular Systems Design and Engineering*, Vol. 1, No. 4, 2016.

173. Pepe, G., J. M. Cole, P. G. Waddell and S. Mckechnie, "Molecular Engineering of Cyanine Dyes to Design a Panchromatic Response in Co-Sensitized Dye-Sensitized Solar Cells", *Molecular Systems Design and Engineering*, Vol. 1, No. 1, 2016.

174. Zhou, S., K. Chen, J. Huang, L. Wang, M. Zhang, B. Bai, H. Liu and Q. Wang, "Preparation of Heterometallic Coni-Mofs-Modified Bivo4: a Steady Photoanode for Improved Performance in Photoelectrochemical Water Splitting", *Applied Catalysis B: Environmental*, Vol. 266, No. 11, p. 118513, 2020.

175. Wang, P., D. Wang, J. Lin, X. Li, C. Peng, X. Gao, Q. Huang, J. Wang, H. Xu and C. Fan, "Lattice Defect-Enhanced Hydrogen Production in Nanostructured Hematite-Based Photoelectrochemical Device", *Applied Materials and Interfaces*, Vol. 4, No. 4, pp. 2295–2301, 2012.

176. Liu, J., Y. Yu, Z. Liu, S. Zuo and B. Li, "Agbr-Coupled TiO2: a Visible Heterostructured Photocatalyst for Degrading Dye Pollutants", *International Journal of Photoenergy*, Vol. 2012, No. 4, pp. 1991–2003, 2012.

177. Momeni, M. M. and Y. Ghayeb, "Visible Light-Driven Photoelectrochemical Water Splitting on ZnO–TiO2 Heterogeneous Nanotube Photoanodes", *Journal of Applied Electrochemistry*, Vol. 45, No. 6, pp. 557–566, 2015.

178. Kumar, D., S. Singh and N. Khare, "Plasmonic Ag Nanoparticles Decorated Nanbo3 Nanorods for Efficient Photoelectrochemical Water Splitting", *International Journal of Hydrogen Energy*, Vol. 43, No. 17, pp. 8198–8205, 2018.

179. Dom, R., G. S. Kumar, N. Y. Hebalkar, S. V. Joshi and P. H. Borse, "Eco-Friendly Ferrite Nanocomposite Photoelectrode for Improved Solar Hydrogen Generation", *Royal Science of Chemistry Advances*, Vol. 3, No. 35, pp. 15217–15224, 2013.

180. Park, Y., D. Kang and K. S. Choi, "Marked Enhancement in Electron-Hole Separation Achieved in the Low Bias Region Using Electrochemically Prepared Mo-Doped Bivo4 Photoanodes", *Physical Chemistry Chemical Physics*, Vol. 16, No. 3, pp. 1238–1246, 2014.

181. Grigorescu, S., B. BäRhausen, L. Wang, A. Mazare, J. E. Yoo, R. Hahn and

P. Schmuki, "Tungsten Doping of Ta3N5-Nanotubes for Band Gap Narrowing and Enhanced Photoelectrochemical Water Splitting Efficiency", *Electrochemistry Communications*, Vol. 51, No. 35, pp. 85–88, 2015.

182. Kim, K., I. H. Kim, K. Y. Yoon, J. Lee and J. H. Jang, "Fe2O3 on Patterned Fluorine Doped Tin Oxide for Efficient Photoelectrochemical Water Splitting", *Journal of Materials Chemistry A*, Vol. 3, No. 15, pp. 7706–7709, 2015.

183. Saraswat, S. K., D. D. Rodene and R. B. Gupta, "Recent Advancements in Semiconductor Materials for Photoelectrochemical Water Splitting for Hydrogen Production Using Visible Light", *Renewable and Sustainable Energy Reviews*, Vol. 89, No. 7, pp. 228–248, 2018.

184. Tournet, J., Y. Lee, S. K. Karuturi, H. H. Tan and C. Jagadish, "Iii–V Semiconductor Materials for Solar Hydrogen Production: Status and Prospects", *Energy Letters*, Vol. 5, No. 2, pp. 345–356, 2020.

185. Daulbayev, C., F. Sultanov, B. Bakbolat and O. Daulbayev, "0D, 1D and 2D Nanomaterials for Visible Photoelectrochemical Water Splitting. a Review", *International Journal of Hydrogen Energy*, Vol. 45, No. 58, pp. 33325–33342, 2020.

186. Dholam, R., N. Patel, M. Adami and A. Miotello, "Hydrogen Production by Photocatalytic Water-Splitting Using Cr- Or Fe-Doped TiO2 Composite Thin Films Photocatalyst", *International Journal of Hydrogen Energy*, Vol. 34, No. 13, pp. 965–974, 2009.

187. Bin Adnan, M. A., K. Arifin, L. J. Minggu and M. B. Kassim, "Titanate-Based Perovskites for Photochemical and Photoelectrochemical Water Splitting Applications: a Review", *International Journal of Hydrogen Energy*, Vol. 43, No. 52, pp. 23209–23220, 2018.

188. Marlinda, A. R., N. Yusoff, S. Sagadevan and M. R. Johan, "Recent Devel-

opments in Reduced Graphene Oxide Nanocomposites for Photoelectrochemical Water-Splitting Applications", *International Journal of Hydrogen Energy*, Vol. 45, No. 21, pp. 11976–11994, 2020.

189. Singla, S., S. Sharma, S. Basu, N. P. Shetti and T. M. Aminabhavi, "Photocatalytic Water Splitting Hydrogen Production Via Environmental Benign Carbon based Nanomaterials", *International Journal of Hydrogen Energy*, Vol. 46, No. 68, pp. 33696–33717, 2021.

190. Bellani, S., M. R. Antognazza and F. Bonaccorso, "Carbon-Based Photocathode Materials for Solar Hydrogen Production", *Advanced Materials*, Vol. 31, No. 9, pp. 12330–12340, 2019.

191. Joy, J., J. Mathew and S. C. George, "Nanomaterials for Photoelectrochemical Water Splitting – Review", *International Journal of Hydrogen Energy*, Vol. 43, No. 10, pp. 4804–4817, 2018.

192. Aguilera GonzáLez, E. N., S. Estrada-Flores and A. MartiNez-LueVanos, "Nanomaterials: Recent Advances for Hydrogen Production", *Handbook of Nanomaterials and Nanocomposites for Energy and Environmental Applications*, Vol. 4, No. 2, pp. 2767–2792, 2021.

193. Jin, L., H. Zhao, Z. M. Wang and F. Rosei, "Quantum Dots-Based Photoelectrochemical Hydrogen Evolution from Water Splitting", *Advanced Energy Materials*, Vol. 11, No. 12, pp. 234–250, 2021.

194. Liu, C., N. P. Dasgupta and P. Yang, "Semiconductor Nanowires for Artificial Photosynthesis", *Chemistry of Materials*, Vol. 26, No. 1, pp. 897–997, 2014.

195. Liu, C., T. Liu, Y. Li, Z. Zhao, D. Zhou, W. Li, Y. Zhao, H. Yang, L. Sun, F. Li and Z. Li, "A Dendritic Sb2Se3/In2S3 Heterojunction Nanorod Array Photocathode Decorated with a MOSx catalyst for Efficient Solar Hydrogen Evolution",

Journal of Materials Chemistry A, Vol. 8, No. 44, pp. 205–214, 2020.

196. Kim, B., G. S. Park, Y. J. Hwang, D. H. Won, W. Kim, D. K. Lee and B. K. Min, "Cu(In,Ga)(S,Se)2 Photocathodes with a Cuxs Catalyst for Solar Water Splitting", *Energy Letters*, Vol. 4, No. 12, pp. 238–245, 2019.

197. Zhao, Y., Z. Niu, J. Zhao, L. Xue, X. Fu and J. Long, "Recent Advancements in Photoelectrochemical Water Splitting for Hydrogen Production", *Electrochemical Energy Reviews*, Vol. 6, No. 1, pp. 7–18, 2023.

198. Zhang, L., Y. Song, J. Feng, T. Fang, Y. Zhong, Z. Li and Z. Zou, "Photoelectrochemical Water Oxidation of Lataon2 Under Visible-Light Irradiation", *International Journal of Hydrogen Energy*, Vol. 39, No. 15, pp. 7697–7704, 2014.

199. Ullah, R., M. Pei, J. Wu, Y. Tian, Z. Gu, Q. Zhang, C. Song, Y. Yang, M. Ahmad, J. Zeb, F. Qayyum, Y. Liu, X. An, L. Gu, X. Wang and J. Zhang, "Bifunctional Photoelectrode Driven by Charged Domain Walls in Ferroelectric Bi2WO6", *Applied Energy Materials*, Vol. 3, No. 5, pp. 4149–4154, 2020.

200. Kado, Y., C. Y. Lee, K. Lee, J. Müller, M. Moll, E. Spiecker and P. Schmuki, "Enhanced Water Splitting Activity of M-Doped Ta3N5 (M = Na, K, Rb, Cs)", *Chemical Communications*, Vol. 48, No. 69, pp. 8685–8687, 2012.

201. Das, C., P. Roy, M. Yang, H. Jha and P. Schmuki, "Nb Doped TiO2 Nanotubes for Enhanced Photoelectrochemical Water-Splitting", *Nanoscale*, Vol. 3, No. 8, pp. 3094–3096, 2011.

202. Kim, J. K., K. Shin, S. M. Cho, T. W. Lee and J. H. Park, "Synthesis of Transparent Mesoporous Tungsten Trioxide Films with Enhanced Photoelectrochemical Response: Application to Unassisted Solar Water Splitting", *Energy and Environmental Science*, Vol. 4, No. 4, pp. 1465–1470, 2011.

203. Cho, S., J. W. Jang, S. H. Lim, H. J. Kang, S. W. Rhee, J. S. Lee and K. H. Lee,

"Solution-Based Fabrication of ZnO/ZnSe Heterostructure Nanowire Arrays for Solar Energy Conversion", *Journal of Materials Chemistry*, Vol. 21, No. 44, pp. 17816–17822, 2011.

204. Hill, J. C. and K. S. Choi, "Synthesis and Characterization of High Surface Area Cuwo4 and Bi2wo6 Electrodes for Use As Photoanodes for Solar Water Oxidation", *Journal of Materials Chemistry A*, Vol. 1, No. 16, pp. 5006–5014, 2013.

205. Günay, M. E., R. Yıldırım, M. Erdem Günay and R. Yıldırım, "Recent Advances in Knowledge Discovery for Heterogeneous Catalysis Using Machine Learning", *Catalysis Reviews - Science and Engineering*, Vol. 63, No. 1, pp. 120–164, 2021.

206. Tapan, N. A., R. Yıldırım, M. E. Günay, N. Alper Tapan, R. Yıldırım and M. E. Günay, "Analysis of Past Experimental Data in Literature to Determine Conditions for High Performance in Biodiesel Production", *Biofuels, Bioproducts and Biorefining*, Vol. 10, No. 4, pp. 422–434, 2016.

207. Coşgun, A., M. E. Günay and R. Yıldırım, "Exploring the Critical Factors of Algal Biomass and Lipid Production for Renewable Fuel Production by Machine Learning", *Renewable Energy*, Vol. 163, No. 1, pp. 1299–1317, 2021.

208. Odabasi, C. and R. Yıldırım, "Machine Learning Analysis on Stability of Perovskite Solar Cells", *Solar Energy Materials and Solar Cells*, Vol. 205, No. 2, p. 110284, 2020.

209. Tao, Q., P. Xu, M. Li and W. Lu, "Machine Learning for Perovskite Materials Design and Discovery", *Nature Computational Materials*, Vol. 7, No. 1, pp. 1–18, 2021.

210. Masood, H., C. Y. Toe, W. Y. Teoh, V. Sethu and R. Amal, "Machine Learning for Accelerated Discovery of Solar Photocatalysts", *Catalysis*, Vol. 9, No. 12, pp. 11774–11787, 2019.

211. Min, K., B. Choi, K. Park and E. Cho, "Machine Learning Assisted Optimization of Electrochemical Properties for Ni-Rich Cathode Materials", *Scientific Reports*, Vol. 8, No. 1, pp. 1230–1240, 2018.

212. Yoon, Y., M. J. Kim and J. J. Kim, "Machine Learning to Electrochemistry: Analysis of Polymers and Halide Ions in a Copper Electrolyte", *Electrochimica Acta*, Vol. 399, No. 18, pp. 84–99, 2021.

213. Hogerwaard, J., I. Dincer and G. F. Naterer, "Experimental Investigation and Optimization of Integrated Photovoltaic and Photoelectrochemical Hydrogen Generation", *Energy Conversion and Management*, Vol. 207, No. 81, pp. 1004–1019, 2020.

214. Chen, J., C. Dong, H. Idriss, O. F. Mohammed and O. M. Bakr, "Metal Halide Perovskites for Solar-to-Chemical Fuel Conversion", *Advanced Energy Materials*, Vol. 10, No. 13, pp. 1614–1624, 2020.

215. Zhang, H., Z. Yu, R. Jiang, Y. Hou, J. Huang, H. Zhu, F. Yang, M. Li, F. Li and Q. Ran, "Metal Organic Frameworks Constructed Heterojunction with NiS/CdS: the Effect of Organic-Ligand in Uio-66 for Charge Transfer of Photocatalytic Hydrogen Evolution", *Renewable Energy*, Vol. 168, No. 11, pp. 187–196, 2021.

216. Saxena, A. and V. Rajpoot, "A Comparative Analysis of Association Rule Mining Algorithms", *Materials Science and Engineering*, Vol. 1099, No. 1, p. 12032, 2021.

217. Prati, R. C., G. E. A. P. A. Batista and M. C. Monard, "A Study with Class Imbalance and Random Sampling for a Decision Tree Learning System", *International Federation for Information Processing*, Vol. 276, No. 68, pp. 131–140, 2008.

218. Therneau, T. and E. Atkinson, "Recursive Partitioning and Regression Trees (Rpart)", *Comprehensive R Archive Network*, Vol. 1, No. 4, pp. 1–60, 2019.

219. Choudhary, S., S. Upadhyay, P. Kumar, N. Singh, V. R. Satsangi, R. Shrivastav and S. Dass, "Nanostructured Bilayered Thin Films in Photoelectrochemical Water Splitting - a Review", *International Journal of Hydrogen Energy*, Vol. 37, No. 24, pp. 18713–18730, 2012.

220. Zhang, W., W. Wang, H. Shi, Y. Liang, J. Fu and M. Zhu, "Surface Plasmon-Driven Photoelectrochemical Water Splitting of Aligned ZnO Nanorod Arrays Decorated with Loading-Controllable Au Nanoparticles", *Solar Energy Materials and Solar Cells*, Vol. 180, No. February, pp. 25–33, 2018.

221. Kment, S., F. Riboni, S. Pausova, L. Wang, L. Wang, H. Han, Z. Hubicka, J. Krysa, P. Schmuki and R. Zboril, "Photoanodes based on TiO2 and Fe2O3 for Solar Water Splitting-Superior Role of 1d Nanoarchitectures and of Combined Heterostructures", *Chemical Society Reviews*, Vol. 46, No. 12, pp. 3716–3769, 2017.

222. Li, R., "Latest Progress in Hydrogen Production from Solar Water Splitting Via Photocatalysis, Photoelectrochemical, and Photovoltaic-Photoelectrochemical Solutions", *Cuihua Xuebao/Chinese Journal of Catalysis*, Vol. 38, No. 1, pp. 5–12, 2017.

223. Wang, L., C. Y. Lee and P. Schmuki, "Solar Water Splitting: Preserving the Beneficial Small Feature Size in Porous Fe2O3 Photoelectrodes During Annealing", *Journal of Materials Chemistry A*, Vol. 1, No. 2, pp. 212–215, 2013.

224. Hu, Y., X. Yan, Y. Gu, X. Chen, Z. Bai, Z. Kang, F. Long and Y. Zhang, "Large-Scale Patterned ZnO Nanorod Arrays for Efficient Photoelectrochemical Water Splitting", *Applied Surface Science*, Vol. 339, No. 1, pp. 122–127, 2015.

225. Hassan, N. K., M. R. Hashim and N. K. Allam, "ZnO Nano-Tetrapod Photoanodes for Enhanced Solar-Driven Water Splitting", *Chemical Physics Letters*, Vol. 549, No. 92, pp. 62–66, 2012.

226. Iwase, A. and A. Kudo, "Photoelectrochemical Water Splitting Using Visible-Light-Responsive Bivo4 Fine Particles Prepared in An Aqueous Acetic Acid Solution", *Journal of Materials Chemistry*, Vol. 20, No. 35, pp. 7536–7542, 2010.

227. Shaban, Y. A. and S. U. Khan, "Surface Grooved Visible Light Active Carbon Modified (Cm)-N-TiO2 Thin Films for Efficient Photoelectrochemical Splitting of Water", *Chemical Physics*, Vol. 339, No. 1-3, pp. 73–85, 2007.

228. Xu, C., Y. A. Shaban, W. B. Ingler and S. U. Khan, "Nanotube Enhanced Photoresponse of Carbon Modified (Cm)-N-TiO2 for Efficient Water Splitting", *Solar Energy Materials and Solar Cells*, Vol. 91, No. 10, pp. 938–943, 2007.

229. Shaban, Y. A. and S. U. Khan, "Visible Light Active Carbon Modified N-TiO2 for Efficient Hydrogen Production by Photoelectrochemical Splitting of Water", *International Journal of Hydrogen Energy*, Vol. 33, No. 4, pp. 1118–1126, 2008.

230. Rani, B. J., M. Praveenkumar, S. Ravichandran, V. Ganesh, R. K. Guduru, G. Ravi and R. Yuvakkumar, "Ultrafine M-Doped TiO2 (M=Fe, Ce, La)Nanosphere Photoanodes for Photoelectrochemical Water-Splitting Applications", *Materials Characterization*, Vol. 152, No. 4, pp. 188–203, 2019.

231. Tsai, C. C. and H. Teng, "Chromium-Doped Titanium Dioxide Thin-Film Photoanodes in Visible-Light-Induced Water Cleavage", *Applied Surface Science*, Vol. 254, No. 15, pp. 4912–4918, 2008.

232. Momeni, M. M. and Y. Ghayeb, "Fabrication, Characterization and Photoelectrochemical Performance of Chromium-Sensitized Titania Nanotubes As Efficient Photoanodes for Solar Water Splitting", *Journal of Solid State Electrochemistry*, Vol. 20, No. 3, pp. 683–689, 2016.

233. SáNchez-Tovar, R., E. Blasco-Tamarit, R. M. FernáNdez-Domene, B. Lucas-Granados and J. GarciA-AntóN, "Should TiO2 Nanostructures Doped with Li+

Be Used As Photoanodes for Photoelectrochemical Water Splitting Applications?", *Journal of Catalysis*, Vol. 349, No. 48, pp. 41–52, 2017.

234. Altomare, M., K. Lee, M. S. Killian, E. Selli and P. Schmuki, "Ta-Doped TiO2 Nanotubes for Enhanced Solar-Light Photoelectrochemical Water Splitting", *Chemistry - a European journal*, Vol. 19, No. 19, pp. 5841–5844, 2013.

235. Dong, Z., D. Ding, T. Li and C. Ning, "Ni-Doped TiO2 Nanotubes Photoanode for Enhanced Photoelectrochemical Water Splitting", *Applied Surface Science*, Vol. 443, No. 234, pp. 321–328, 2018.

236. Xu, C., Y. Song, L. Lu, C. Cheng, D. Liu, X. Fang, X. Chen, X. Zhu and D. Li, "Electrochemically Hydrogenated TiO2 Nanotubes with Improved Photoelectrochemical Water Splitting Performance", *Nanoscale Research Letters*, Vol. 8, No. 1, pp. 1–7, 2013.

237. Ahn, C. W., P. H. Borse, J. Y. J. H. J. Y. Kim, J. Y. J. H. J. Y. Kim, J. S. Jang, C. R. Cho, J. H. Yoon, B. Seob Lee, J. S. Bae, H. G. Kim and J. S. Lee, "Effective Charge Separation in Site-Isolated Pt-Nanodot Deposited PbtiO3 Nanotube Arrays for Enhanced Photoelectrochemical Water Splitting", *Applied Catalysis B: Environmental*, Vol. 224, No. 8, pp. 804–809, 2018.

238. Roy, P., C. Das, K. Lee, R. Hahn, T. Ruff, M. Moll and P. Schmuki, "Oxide Nanotubes on Ti-Ru Alloys: Strongly Enhanced and Stable Photoelectrochemical Activity for Water Splitting", *Journal of the American Chemical Society*, Vol. 133, No. 15, pp. 5629–5631, 2011.

239. Subramanian, A., Z. Pan, H. Li, L. Zhou, W. Li, Y. Qiu, Y. Xu, Y. Hou, C. Muzi and Y. Zhang, "Synergistic Promotion of Photoelectrochemical Water Splitting Efficiency of TiO2 Nanorods Using Metal-Semiconducting Nanoparticles", *Applied Surface Science*, Vol. 420, No. 111, pp. 631–637, 2017.

240. Su, F., J. Lu, Y. Tian, X. Ma and J. Gong, "Branched TiO2 Nanoarrays Sensitized with Cds Quantum Dots for Highly Efficient Photoelectrochemical Water Splitting", *Physical Chemistry Chemical Physics*, Vol. 15, No. 29, pp. 12026–12032, 2013.

241. Ji, I. A., M. J. Park, J. Y. Jung, M. J. Choi, Y. W. Lee, J. H. Lee and J. H. Bang, "One-Dimensional Core/Shell Structured TiO2/ZnO Heterojunction for Improved Photoelectrochemical Performance", *Bulletin of the Korean Chemical Society*, Vol. 33, No. 7, pp. 2200–2206, 2012.

242. Liu, X., M. Ye, S. Zhang, G. Huang, C. Li, J. Yu, P. K. Wong and S. Liu, "Enhanced Photocatalytic CO2 Valorization over TiO2 Hollow Microspheres by Synergetic Surface Tailoring and Au Decoration", *Journal of Materials Chemistry A*, Vol. 6, No. 47, pp. 1360–1371, 2018.

243. Ho-Kimura, S., S. J. Moniz, A. D. Handoko and J. Tang, "Enhanced Photoelectrochemical Water Splitting by Nanostructured BiVO4-TiO2 Composite Electrodes", *Journal of Materials Chemistry A*, Vol. 2, No. 11, pp. 3948–3953, 2014.

244. Lee, J. G., D. Y. Kim, J. J. Park, Y. H. Cha, J. Y. Yoon, H. S. Jeon, B. K. Min, M. T. Swihart, S. Jin, S. S. Al-Deyab and S. S. Yoon, "Graphene-Titania Hybrid Photoanodes by Supersonic Kinetic Spraying for Solar Water Splitting", *Journal of the American Ceramic Society*, Vol. 97, No. 11, pp. 3660–3668, 2014.

245. SáNchez-Tovar, R., E. Blasco-Tamarit, R. M. FernáNdez-Domene, M. Villanueva-Pascual and J. GarciA-AntóN, "Electrochemical Formation of Novel TiO2-ZnO Hybrid Nanostructures for Photoelectrochemical Water Splitting Applications", *Surface and Coatings Technology*, Vol. 388, No. 3, p. 125605, 2020.

246. Jia, Y., Z. Wang, Y. Ma, J. Liu, W. Shi, Y. Lin, X. Hu and K. Zhang, "Boosting Interfacial Charge Migration of TiO2Bivo 4 Photoanode by W Doping for Photoelectrochemical Water Splitting", *Electrochimica Acta*, Vol. 300, No. 12, pp.

138–144, 2019.

247. Fan, W. Q., X. Q. Yu, S. Y. Song, H. Y. Bai, C. Zhang, D. Yan, C. B. Liu, Q. Wang and W. D. Shi, "Fabrication of TiO2-BiOCl Double-Layer Nanostructure Arrays for Photoelectrochemical Water Splitting", *Crystengcomm*, Vol. 16, No. 5, pp. 820–825, 2014.

248. Mohajernia, S., S. Hejazi, A. Mazare, N. T. Nguyen and P. Schmuki, "Photoelectrochemical H2 Generation from Suboxide TiO2 Nanotubes: Visible-Light Absorption Versus Conductivity", *Chemistry - a European journal*, Vol. 23, No. 50, pp. 12406–12411, 2017.

249. Bu, X., Y. Gao, S. Zhang and Y. Tian, "Amorphous Cerium Phosphate on P-Doped Fe2O3 Nanosheets for Efficient Photoelectrochemical Water Oxidation", *Chemical Engineering journal*, Vol. 355, No. 8, pp. 910–919, 2019.

250. Hu, Y.-S., A. Kleiman-Shwarsctein, A. J. Forman, D. Hazen, J.-N. Park, E. W. Mcfarland, S. Barbara and S. Barbara, "Pt-Doped R -Fe 2 O 3 Thin Films Active for Photoelectrochemical Water Splitting ", *Chemistry of Materials*, Vol. 20, No. 12, pp. 3803–3805, 2008.

251. Cai, J., Y. Liu, S. Li, M. Gao and G. Qin, "Photoelectrochemical Behavior of Sn-Doped Fe2O3 Photoanode with Different Reducer", *Chinese Journal of Chemistry*, Vol. 34, No. 8, pp. 778–782, 2016.

252. Bosso, P., A. Milella, G. Barucca, P. Mengucci, V. Armenise, F. Fanelli, R. Giannuzzi, V. Maiorano and F. Fracassi, "Plasma-Assisted Deposition of Iron Oxide Thin Films for Photoelectrochemical Water Splitting", *Plasma Processes and Polymers*, Vol. 18, No. 1, pp. 1–12, 2021.

253. Han, H., S. Kment, F. Karlicky, L. Wang, A. Naldoni, P. Schmuki and R. Zboril, "Sb-Doped SnO2 Nanorods Underlayer Effect to the Fe2O3 Nanorods Sheathed

with TiO2 for Enhanced Photoelectrochemical Water Splitting", *Small*, Vol. 14, No. 19, pp. 1–8, 2018.

254. Weng, B., F. Xu and J. Xu, "Core-Shell Photoanode Developed by Atomic Layer Deposition of Bi2o3 on Si Nanowires for Enhanced Photoelectrochemical Water Splitting", *Nanotechnology*, Vol. 25, No. 45, pp. 436–452, 2014.

255. Qamar, M., M. Abdalwadoud, M. I. Ahmed, A. M. Azad, B. Merzougui, S. Bukola, Z. H. Yamani and M. N. Siddiqui, "Single-Pot Synthesis of ⟨001⟩-Faceted N-Doped Nb2o5/Reduced Graphene Oxide Nanocomposite for Efficient Photoelectrochemical Water Splitting", *Applied Materials and Interfaces*, Vol. 7, No. 32, pp. 17954–17962, 2015.

256. Alotaibi, B., S. Fan, H. P. Nguyen, S. Zhao, M. G. Kibria and Z. Mi, "Photoelectrochemical Water Splitting and Hydrogen Generation Using Ingan/Gan Nanowire Arrays", *2014 Summer Topicals Meeting Series*, Vol. 3, pp. 206–207, Sicilia, 2014.

257. Alotaibi, B., M. Harati, S. Fan, S. Zhao, H. P. Nguyen, M. G. Kibria and Z. Mi, "High Efficiency Photoelectrochemical Water Splitting and Hydrogen Generation Using Gan Nanowire Photoelectrode", *Nanotechnology*, Vol. 24, No. 17, pp. 1596–1623, 2013.

258. Son, H., J. H. Park, P. Uthirakumar, A. Y. Kuznetsov and I. H. Lee, "Impact of Chloride Surface Treatment on Nano-Porous Gan Structure for Enhanced Water-Splitting Efficiency", *Applied Surface Science*, Vol. 532, No. 7, p. 147465, 2020.

259. Pinheiro, A. N., E. G. Firmiano, A. C. Rabelo, C. J. Dalmaschio and E. R. Leite, "Revisiting SrTiO3 As a Photoanode for Water Splitting: Development of Thin Films with Enhanced Charge Separation Under Standard Solar Irradiation", *Royal Science of Chemistry Advances*, Vol. 4, No. 4, pp. 2029–2036, 2014.

260. Guo, Y., N. Zhang, H. Huang, Z. Li and Z. Zou, "A Novel Wide-Spectrum Response Hexagonal Yfeo3 Photoanode for Solar Water Splitting", *Royal Science of Chemistry Advances*, Vol. 7, No. 30, pp. 18418–18420, 2017.

261. Li, M., J. Su and L. Guo, "Preparation and Characterization of ZnIn2S4 Thin Films Deposited by Spray Pyrolysis for Hydrogen Production", *International Journal of Hydrogen Energy*, Vol. 33, No. 12, pp. 2891–2896, 2008.

262. Sitara, E., H. Nasir, A. Mumtaz, M. F. Ehsan, M. Sohail, S. Iram and S. A. B. Bukhari, "Efficient Photoelectrochemical Water Splitting by Tailoring Mos2/Cote Heterojunction in a Photoelectrochemical Cell", *Nanomaterials*, Vol. 10, No. 12, pp. 1–14, 2020.

263. Huang, X., Y. Li, X. Gao, Q. Xue, R. Zhang, Y. Gao, Z. Han and M. Shao, "The Effect of the Photochemical Environment on Photoanodes for Photoelectrochemical Water Splitting", *Dalton Transactions*, Vol. 49, No. 35, pp. 12338–12344, 2020.

264. Fei, H., Y. Yang, D. L. Rogow, X. Fan and S. R. Oliver, "Polymer-Templated Nanospider TiO2 Thin Films for Efficient Photoelectrochemical Water Splitting", *Applied Materials and Interfaces*, Vol. 2, No. 4, pp. 974–979, 2010.

265. Hou, Y., C. Zheng, Z. Zhu and X. Wang, "Microwave-Assisted Fabrication of Porous Hematite Photoanodes for Efficient Solar Water Splitting", *Chemical Communications*, Vol. 52, No. 42, pp. 6888–6891, 2016.

266. Qin, D. D., C. L. Tao, S. A. Friesen, T. H. Wang, O. K. Varghese, N. Z. Bao, Z. Y. Yang, T. E. Mallouk and C. A. Grimes, "Dense Layers of Vertically Oriented Wo3 Crystals As Anodes for Photoelectrochemical Water Oxidation", *Chemical Communications*, Vol. 48, No. 5, pp. 729–731, 2012.

267. Hong, T., Z. Liu, J. Zhang, G. Li, J. Liu, X. Zhang and S. Lin, "Flower-Like

Cu2in2zns5 Nanosheets: a Novel Promising Photoelectrode for Water Splitting", *Chemcatchem*, Vol. 8, No. 7, pp. 1288–1292, 2016.

268. Wang, X., Z. Liu and Z. Liu, "A Dumbbell Cabi2o4 Photoelectrode for Photoelectrochemical Water Splitting", *Chemcatchem*, Vol. 9, No. 21, pp. 4029–4034, 2017.

269. Chen, E. X., M. Qiu, Y. F. Zhang, Y. S. Zhu, L. Y. Liu, Y. Y. Sun, X. Bu, J. Zhang and Q. Lin, "Acid and Base Resistant Zirconium Polyphenolate-Metalloporphyrin Scaffolds for Efficient CO2 Photoreduction", *Advanced Materials*, Vol. 30, No. 2, pp. 152–168, 2018.

270. Chahrour, K. M., F. K. Yam and A. M. Eid, "Water-Splitting Properties of Bi-Phased TiO2 Nanotube Arrays Subjected to High-Temperature Annealing", *Ceramics International*, Vol. 46, No. 13, pp. 21471–21481, 2020.

271. Zhang, S., M. Feng, Y. Liu and D. Wang, "Ta2O5 Nts-TiO2 Nanodots Heterostructure Photocatalyst Material for Enhanced Photodegradation and Photoelectrochemical Performance Under Simulated Solar Light", *Journal of Nanoparticle Research*, Vol. 22, No. 12, pp. 568–575, 2020.

272. Shaheen, B. S., A. M. Hafez, B. Murali, A. R. Kirmani, A. Amassian, O. F. Mohammed and N. K. Allam, "10-Fold Enhancement in Light-Driven Water Splitting Using Niobium Oxynitride Microcone Array Films", *Solar Energy Materials and Solar Cells*, Vol. 151, No. 92, pp. 149–153, 2016.

273. Hu, C., W. Y. Teoh, S. Ji, C. Ye and A. Iwase, "in Situ Metal Doping During Modified Anodization Synthesis of Nb2o5 with Enhanced Photoelectrochemical Water Splitting", *AICHE journal*, Vol. 62, No. 2, pp. 352–358, 2016.

274. Mahajan, V. K., S. K. Mohapatra and M. Misra, "Stability of TiO2 Nanotube Arrays in Photoelectrochemical Studies", *International Journal of Hydrogen Energy*,

Vol. 33, No. 20, pp. 5369–5374, 2008.

275. Liu, K., G. Wang, M. Meng, S. Chen, J. Li, X. Sun, H. Yuan, L. Sun and N. Qin, "TiO2 Nanotube Photonic Crystal Fabricated by Two-Step Anodization Method for Enhanced Photoelectrochemical Water Splitting", *Materials Letters*, Vol. 207, No. 72, pp. 96–99, 2017.

276. Lucas-Granados, B., R. SáNchez-Tovar, R. M. FernáNdez-Domene and J. GarciA-AntóN, "Controlled Hydrodynamic Conditions on the Formation of Iron Oxide Nanostructures Synthesized by Electrochemical Anodization: Effect of the Electrode Rotation Speed", *Applied Surface Science*, Vol. 392, No. 52, pp. 503–513, 2017.

277. Chou, J. C., S. A. Lin, C. Y. Lee and J. Y. Gan, "Effect of Bulk Doping and Surface-Trapped States on Water Splitting with Hematite Photoanodes", *Journal of Materials Chemistry A*, Vol. 1, No. 19, pp. 5908–5914, 2013.

278. Shi, X., K. Zhang, K. Shin, J. H. Moon, T. W. Lee and J. H. Park, "Constructing Inverse Opal Structured Hematite Photoanodes Via Electrochemical Process and Their Application to Photoelectrochemical Water Splitting", *Physical Chemistry Chemical Physics*, Vol. 15, No. 28, pp. 11717–11722, 2013.

279. Deshpande, A., S. Kelkar, S. Rayalu and S. Ogale, "Orthorhombic/Cubic Cd2sno4 Nanojunctions: Enhancing Solar Water Splitting Efficiency by the Suppression of Charge Recombination", *Journal of Materials Chemistry A*, Vol. 2, No. 2, pp. 492–499, 2014.

280. Zhang, H., X. Liu, Y. Li, Q. Sun, Y. Wang, B. J. Wood, P. Liu, D. Yang and H. Zhao, "Vertically Aligned Nanorod-Like Rutile TiO2 Single Crystal Nanowire Bundles with Superior Electron Transport and Photoelectrocatalytic Properties", *Journal of Materials Chemistry*, Vol. 22, No. 6, pp. 2465–2472, 2012.

281. Shen, S., J. Jiang, P. Guo and L. Guo, "Facile Growth of Porous Hematite Films for Photoelectrochemical Water Splitting", *International Journal of Photoenergy*, Vol. 201, No. 10, pp. 1–9, 2013.

282. Reddy, I. N., A. Sreedhar, J. Shim and J. S. Gwag, "Multifunctional Monoclinic Vo 2 Nanorod Thin Films for Enhanced Energy Applications: Photoelectrochemical Water Splitting and Supercapacitor", *Journal of Electroanalytical Chemistry*, Vol. 835, No. 10, pp. 40–47, 2019.

283. Zhao, X., Y. Wu, W. Yao and Y. Zhu, "Photoelectrochemical Properties of Thin Bi2wo6 Films", *Thin Solid Films*, Vol. 515, No. 11, pp. 4753–4757, 2007.

284. Kitano, M., M. Takeuchi, M. Matsuoka, J. M. Thomas and M. Anpo, "Photocatalytic Water Splitting Using Pt-loaded Visible Light-Responsive TiO2 Thin Film Photocatalysts", *Catalysis Today*, Vol. 120, No. 2, pp. 133–138, 2007.

285. Sun, Y., C. J. Murphy, K. R. Reyes-Gil, E. A. Reyes-Garcia, J. P. Lilly and D. Raftery, "Carbon-Doped In2o3 Films for Photoelectrochemical Hydrogen Production", *International Journal of Hydrogen Energy*, Vol. 33, No. 21, pp. 5967–5974, 2008.

286. Yang, W., Y. Xiong, L. Zou, Z. Zou, D. Li, Q. Mi, Y. Wang and H. Yang, "Plasmonic Pd Nanoparticle- and Plasmonic Pd Nanorod-Decorated Bivo4 Electrodes with Enhanced Photoelectrochemical Water Splitting Efficiency Across Visible-Nir Region", *Nanoscale Research Letters*, Vol. 11, No. 1, pp. 155–162, 2016.

287. Gueymard, C. A., D. Myers and K. Emery, "Proposed Reference Irradiance Spectra for Solar Energy Systems Testing", *Solar Energy*, Vol. 73, No. 6, pp. 443–467, 2002.

288. Abe, R., "Recent Progress on Photocatalytic and Photoelectrochemical Water Splitting Under Visible Light Irradiation", *Journal of Photochemistry and Pho-*

tobiology C: Photochemistry Reviews, Vol. 11, No. 4, pp. 179–209, 2010.

289. Ding, C., J. Shi, Z. Wang and C. Li, "Photoelectrocatalytic Water Splitting: Significance of Cocatalysts, Electrolyte, and Interfaces", *Catalysis*, Vol. 7, No. 1, pp. 675–688, 2017.

290. Crawford, S., E. Thimsen and P. Biswas, "Impact of Different Electrolytes on Photocatalytic Water Splitting", *Journal of the Electrochemical Society*, Vol. 156, No. 5, pp. 1236–1250, 2009.

291. Ding, C., X. Zhou, J. Shi, P. Yan, Z. Wang, G. Liu and C. Li, "Abnormal Effects of Cations (Li+, Na+, and K+) on Photoelectrochemical and Electrocatalytic Water Splitting", *Journal of Physical Chemistry B*, Vol. 119, No. 8, pp. 3560–3566, 2015.

292. Sfaelou, S., L. C. Pop, O. Monfort, V. Dracopoulos and P. Lianos, "Mesoporous Wo3 Photoanodes for Hydrogen Production by Water Splitting and Photofuelcell Operation", *International Journal of Hydrogen Energy*, Vol. 41, No. 14, pp. 5902–5907, 2016.

293. Guo, S., X. Zhao, W. Zhang and W. Wang, "Optimization of Electrolyte to Significantly Improve Photoelectrochemical Water Splitting Performance of ZnO Nanowire Arrays", *Materials Science and Engineering B: Solid-State Materials for Advanced Technology*, Vol. 227, No. June 2017, pp. 129–135, 2018.

294. Jozwiak, L., J. Balcerzak and J. Tyczkowski, "Plasma-Deposited Ru-Based Thin Films for Photoelectrochemical Water Splitting", *Catalysts*, Vol. 10, No. 3, pp. 2083–2104, 2020.

295. Yourey, J. E. and B. M. Bartlett, "Electrochemical Deposition and Photoelectrochemistry of Cuwo4, a Promising Photoanode for Water Oxidation", *Journal of Materials Chemistry*, Vol. 21, No. 21, pp. 7651–7660, 2011.

296. Bohra, D. and W. A. Smith, "Improved Charge Separation Via Fe-Doping of Copper Tungstate Photoanodes", *Physical Chemistry Chemical Physics*, Vol. 17, No. 15, pp. 9857–9866, 2015.

297. Kuo, Y. and K. J. Klabunde, "Hydrogen Generation from Water/Methanol Under Visible Light Using Aerogel Prepared Strontium Titanate (Srtio 3) Nanomaterials Doped with Ruthenium and Rhodium Metals", *Nanotechnology*, Vol. 23, No. 29, pp. 117–132, 2012.

298. Momeni, M. M., Y. Ghayeb and F. Ezati, "Investigation of the Morphology, Structural, Optical, and Photoelectrochemical Properties of Wo3–Fe2O3/CrtiO2 Thin-Film Photoanodes for Water Splitting", *Applied Physics A: Materials Science and Processing*, Vol. 126, No. 4, pp. 1–9, 2020.

299. He, W., W. Wu, Q. Li, K. Chen and X. Lu, "Facile Fabrication of Ga2o3 Nanorods for Photoelectrochemical Water Splitting", *Chemisty Nanomaterials*, Vol. 6, No. 2, pp. 208–211, 2020.

300. Zhang, H., J. Wei, J. Dong, G. Liu, L. Shi, P. An, G. Zhao, J. Kong, X. Wang, X. Meng, J. Zhang and J. Ye, "Efficient Visible-Light-Driven Carbon Dioxide Reduction by a Single-Atom Implanted Metal–Organic Framework", *Angewandte Chemie - International Edition*, Vol. 55, No. 46, pp. 1521–1537, 2016.

301. Prakash, J., U. Prasad, X. Shi, X. Peng, B. Azeredo and A. M. Kannan, "Photoelectrochemical Water Splitting Using Lithium Doped Bismuth Vanadate Photoanode with Near-Complete Bulk Charge Separation", *Journal of Power Sources*, Vol. 448, No. 11, p. 227418, 2020.

302. Song, J. P., P. F. Yin, J. Mao, S. Z. Qiao and X. W. Du, "Catalytically Active and Chemically Inert Cdin2s4 Coating on a Cds Photoanode for Efficient and Stable Water Splitting", *Nanoscale*, Vol. 9, No. 19, pp. 6296–6301, 2017.

303. Lee, S. A., T. H. Lee, C. Kim, M. G. Lee, M. J. Choi, H. Park, S. Choi, J. Oh and H. W. Jang, "Tailored Niox/Ni Cocatalysts on Silicon for Highly Efficient Water Splitting Photoanodes Via Pulsed Electrodeposition", *Catalysis*, Vol. 8, No. 8, pp. 7261–7269, 2018.

304. Zhang, B., C. Chen, J. Liu, W. Qiao, J. Zhao, J. Yang, Y. Yu, S. Chen and Y. Qin, "Simultaneous Ni Nanoparticles Decoration and Ni Doping of Cds Nanorods for Synergistically Promoting Photocatalytic H2 Evolution", *Applied Surface Science*, Vol. 508, No. 10, pp. 1645–1657, 2020.

305. Kim, K., J. Yang and J. H. Moon, "Unveiling the Effects of Nanostructures and Core Materials on Charge-Transport Dynamics in Heterojunction Electrodes for Photoelectrochemical Water Splitting", *Applied Materials and Interfaces*, Vol. 12, No. 19, pp. 21894–21902, 2020.

306. Habisreutinger, S. N., L. Schmidt-Mende and J. K. Stolarczyk, "Photocatalytic Reduction of CO2 on TiO2 and Other Semiconductors", *Angewandte Chemie - International Edition*, Vol. 52, No. 29, pp. 1433–1441, 2013.

307. Mori, K., H. Yamashita and M. Anpo, "Photocatalytic Reduction of CO2 with H2O on Various Titanium Oxide Photocatalysts", *Royal Science of Chemistry Advances*, Vol. 2, No. 8, pp. 134–145, 2012.

308. Li, K., X. An, K. H. Park, M. Khraisheh and J. Tang, "A Critical Review of CO2 Photoconversion: Catalysts and Reactors", *Catalysis Today*, Vol. 224, No. 118, pp. 1134–1145, 2014.

309. Zhani, K. and H. Ben Bacha, "Experimental Investigation of a New Solar Desalination Prototype Using the Humidification Dehumidification Principle", *Renewable Energy*, Vol. 35, No. 11, pp. 2610–2617, 2010.

310. Cokoja, M., C. Bruckmeier, B. Rieger, W. A. Herrmann and F. E. Kühn, "Trans-

formation of Carbon Dioxide with Homogeneous Transition-Metal Catalysts: a Molecular Solution to a Global Challenge?", *Angewandte Chemie - International Edition*, Vol. 50, No. 37, pp. 265–278, 2011.

311. Christophe, J., T. Doneux and C. Buess-Herman, "Electroreduction of Carbon Dioxide on Copper-Based Electrodes: Activity of Copper Single Crystals and Copper-Gold Alloys", *Electrocatalysis*, Vol. 3, No. 2, pp. 186–193, 2012.

312. Vassiliev, Y. B., V. S. Bagotzky, N. V. Osetrova and A. A. Mikhailova, "Electroreduction of Carbon Dioxide: Adsorption and Reduction of CO2 on Platinum", *Journal of Electroanalytical Chemistry*, Vol. 189, No. 2, pp. 368–374, 1985.

313. Halmann, M., "Photoelectrochemical Reduction of Aqueous Carbon Dioxide on P-Type Gallium Phosphide in Liquid Junction Solar Cells", *Nature*, Vol. 275, No. 5676, p. 28086, 1978.

314. Porosoff, M. D. and J. G. Chen, "Trends in the Catalytic Reduction of CO2 by Hydrogen over Supported Monometallic and Bimetallic Catalysts", *Journal of Catalysis*, Vol. 301, No. 5, pp. 30–37, 2013.

315. Peterson, A. A. and J. K. NØRskov, "Activity Descriptors for Co 2 Electroreduction to Methane on Transition-Metal Catalysts", *Journal of Physical Chemistry Letters*, Vol. 3, No. 2, pp. 1948–1957, 2012.

316. Morris, A. J., G. J. Meyer and E. Fujita, "Molecular Approaches to the Photocatalytic Reduction of Carbon Dioxide for Solar Fuels", *Accounts of Chemical Research*, Vol. 42, No. 12, pp. 148–156, 2009.

317. Li, X., J. Wen, J. Low, Y. Fang and J. Yu, "Design and Fabrication of Semiconductor Photocatalyst for Photocatalytic Reduction of CO2 to Solar Fuel", *Science China Materials*, Vol. 57, No. 1, pp. 2095–2126, 2014.

318. Ali, S., M. C. Flores, A. Razzaq, S. Sorcar, C. B. Hiragond, H. R. Kim, Y. H.

Park, Y. Hwang, H. S. Kim, H. Kim, E. H. Gong, J. Lee, D. Kim and S. I. In, "Gas Phase Photocatalytic CO2 Reduction, "A Brief Overview for Benchmarking"", *Catalysts*, Vol. 9, No. 9, pp. 2073–2104, 2019.

319. Olivo, A., E. Ghedini, M. Signoretto, M. Compagnoni and I. Rossetti, "Liquid vs. Gas Phase CO2 Photoreduction Process: Which Is the Effect of the Reaction Medium?", *Energies*, Vol. 10, No. 9, pp. 199–210, 2017.

320. Pan, L., H. Mei, G. Zhu, S. Li, X. Xie, S. Gong, H. Liu, Z. Jin, J. Gao, L. Cheng and L. Zhang, "Bi Selectively Doped SrtiO3-X Nanosheets Enhance Photocatalytic CO2 Reduction Under Visible Light", *Journal of Colloid and Interface Science*, Vol. 611, No. 156, pp. 1095–2013, 2022.

321. Mahmoud Idris, A., X. Jiang, J. Tan, Z. Cai, X. Lou, J. Wang and Z. Li, "Dye-Sensitized Fe-Mof Nanosheets As Visible-Light Driven Photocatalyst for High Efficient Photocatalytic CO2 Reduction", *Journal of Colloid and Interface Science*, Vol. 607, No. 16, pp. 195–213, 2022.

322. Zhang, W., J. Xue, Q. Shen, S. Jia, J. Gao, X. Liu and H. Jia, "Black Single-Crystal TiO2 Nanosheet Array Films with Oxygen Vacancy on 001 Facets for Boosting Photocatalytic CO2 Reduction", *Journal of Alloys and Compounds*, Vol. 870, No. 56, pp. 201–213, 2021.

323. Anandan, S., N. Ohashi and M. Miyauchi, "ZnO-Based Visible-Light Photocatalyst: Band-Gap Engineering and Multi-Electron Reduction by Co-Catalyst", *Applied Catalysis B: Environmental*, Vol. 100, No. 3, pp. 495–517, 2010.

324. Liu, G., J. C. Yu, G. Q. Lu and H. M. Cheng, "Crystal Facet Engineering of Semiconductor Photocatalysts: Motivations, Advances and Unique Properties", *Chemical Communications*, Vol. 47, No. 24, pp. 136–145, 2011.

325. Xiong, Z., Z. Lei, Y. Li, L. Dong, Y. Zhao and J. Zhang, "A Review on Modifi-

cation of Facet-Engineered TiO2 for Photocatalytic CO2 Reduction", *Journal of Photochemistry and Photobiology C: Photochemistry Reviews*, Vol. 36, No. 4, pp. 1036–1050, 2018.

326. Fu, J., J. Yu, C. Jiang and B. Cheng, "G-C3n4-Based Heterostructured Photocatalysts", *Advanced Energy Materials*, Vol. 8, No. 3, pp. 161–176, 2018.

327. Bellardita, M., A. Di Paola, E. GarciA-LoPez, V. Loddo, G. Marci and L. Palmisano, "Photocatalytic CO2 Reduction in Gas-Solid Regime in the Presence of Bare, SiO2 Supported Or Cu-loaded TiO2 Samples", *Current Organic Chemistry*, Vol. 17, No. 21, p. 1385, 2013.

328. Pu, Y., Y. Luo, X. Wei, J. Sun, L. Li, W. Zou and L. Dong, "Synergistic Effects of Cu2O-Decorated Ceo2 on Photocatalytic CO2 Reduction: Surface Lewis Acid/Base and Oxygen Defect", *Applied Catalysis B: Environmental*, Vol. 254, No. 121, p. 09263, 2019.

329. Jiang, M., Y. Gao, Z. Wang and Z. Ding, "Photocatalytic CO2 Reduction Promoted by a Cuco2o4 Cocatalyst with Homogeneous and Heterogeneous Light Harvesters", *Applied Catalysis B: Environmental*, Vol. 198, No. 11, pp. 923–937, 2016.

330. Shown, I., H. C. Hsu, Y. C. Chang, C. H. Lin, P. K. Roy, A. Ganguly, C. H. Wang, J. K. Chang, C. I. Wu, L. C. Chen and K. H. Chen, "Highly Efficient Visible Light Photocatalytic Reduction of CO2 to Hydrocarbon Fuels by Cu-Nanoparticle Decorated Graphene Oxide", *Nano Letters*, Vol. 14, No. 11, pp. 153–162, 2014.

331. Zhang, Q., T. Gao, J. M. Andino and Y. Li, "Copper and Iodine Co-Modified Tio2 Nanoparticles for Improved Activity of Co 2 Photoreduction with Water Vapor", *Applied Catalysis B: Environmental*, Vol. 124, No. 10, pp. 253–263, 2012.

332. Ingram, D. B., P. Christopher, J. L. Bauer and S. Linic, "Predictive Model for the Design of Plasmonic Metal/Semiconductor Composite Photocatalysts", *Catalysis*, Vol. 1, No. 10, pp. 653–668, 2011.

333. Liu, E., L. Kang, F. Wu, T. Sun, X. Hu, Y. Yang, H. Liu and J. Fan, "Photocatalytic Reduction of CO2 Into Methanol over Ag/TiO2 Nanocomposites Enhanced by Surface Plasmon Resonance", *Plasmonics*, Vol. 9, No. 1, pp. 215–230, 2014.

334. Pang, R., K. Teramura, H. Asakura, S. Hosokawa and T. Tanaka, "Highly Selective Photocatalytic Conversion of CO2 by Water over Ag-loaded Srnb2o6 Nanorods", *Applied Catalysis B: Environmental*, Vol. 218, No. 24, pp. 996–1003, 2017.

335. Zhang, P., Z. Dong, Y. Ran, H. Xie, Y. Lu and S. Ding, "Preparation and Photocatalytic Application of Agbr Modified Bi2wo6 Nanosheets with High Adsorption Capacity", *Journal of Materials Research*, Vol. 33, No. 23, pp. 201–213, 2018.

336. Zhang, Q. H., W. D. Han, Y. J. Hong and J. G. Yu, "Photocatalytic Reduction of CO2 with H2O on Pt-loaded TiO2 Catalyst", *Catalysis Today*, Vol. 148, No. 3, pp. 191–210, 2009.

337. Pan, F., X. Xiang, W. Deng, H. Zhao, X. Feng and Y. Li, "A Novel Photo-Thermochemical Approach for Enhanced Carbon Dioxide Reforming of Methane", *Chemcatchem*, Vol. 10, No. 5, pp. 1867–1899, 2018.

338. Haruta, M., N. Yamada, T. Kobayashi and S. Iijima, "Gold Catalysts Prepared by Coprecipitation for Low-Temperature Oxidation of Hydrogen and of Carbon Monoxide", *Journal of Catalysis*, Vol. 115, No. 2, pp. 367–393, 1989.

339. Saadetnejad, D. and R. Yıldırım, "Photocatalytic Hydrogen Production by Water Splitting over Au/Al-SrtiO3", *International Journal of Hydrogen Energy*, Vol. 43, No. 2, pp. 360–379, 2018.

340. Cui, X., J. Wang, B. Liu, S. Ling, R. Long and Y. Xiong, "Turning Au Nanoclusters Catalytically Active for Visible-Light-Driven CO2 Reduction through Bridging Ligands", *Journal of the American Chemical Society*, Vol. 140, No. 48, pp. 361–375, 2018.

341. Tahir, M., B. Tahir, N. Saidina Amin and H. Alias, "Selective Photocatalytic Reduction of CO2 by H2O/H2 to Ch4 and Ch3oh over Cu-Promoted In2o3/TiO2 Nanocatalyst", *Applied Surface Science*, Vol. 389, No. 125, pp. 125–134, 2016.

342. Wang, W., L. Li, K. Wu, K. Zhang, J. Jie and Y. Yang, "Preparation of Ni-Mo-S Catalysts by Hydrothermal Method and Their Hydrodeoxygenation Properties", *Applied Catalysis A: General*, Vol. 495, No. 16, pp. 1305–1314, 2015.

343. Yilmaz, E. and M. Soylak, "Functionalized Nanomaterials for Sample Preparation Methods", *Handbook of Nanomaterials in Analytical Chemistry: Modern Trends in Analysis*, Vol. 2, No. 6, pp. 135–144, 2019.

344. Slamet, H. W. Nasution, E. Purnama, S. Kosela and J. Gunlazuardi, "Photocatalytic Reduction of CO2 on Copper-Doped Titania Catalysts Prepared by Improved-Impregnation Method", *Catalysis Communications*, Vol. 6, No. 5, pp. 156–167, 2005.

345. Saladin, F. and I. Alxneit, "Temperature Dependence of the Photochemical Reduction of CO2 in the Presence of H2O At the Solid/Gas Interface of TiO2", *Journal of the Chemical Society*, Vol. 93, No. 23, pp. 967–975, 1997.

346. Baysal, M., M. E. Günay and R. Yıldırım, "Decision Tree Analysis of Past Publications on Catalytic Steam Reforming to Develop Heuristics for High Performance: a Statistical Review", *International Journal of Hydrogen Energy*, Vol. 42, No. 1, pp. 243–254, 2017.

347. Yildiz, M. G., T. Davran-Candan, M. E. Günay and R. Yildirim, "CO2 Cap-

ture over Amine-Functionalized MCM-41 and SBA-15: Exploratory Analysis and Decision Tree Classification of Past Data", *Journal of CO2 Utilization*, Vol. 31, No. 5, pp. 27–42, 2019.

348. Gaviria, J. F., G. NarváEz, C. Guillen, L. F. Giraldo and M. Bressan, "Machine Learning in Photovoltaic Systems: a Review", *Renewable Energy*, Vol. 196, No. 10, p. 54–57, 2022.

349. Esterhuizen, J. A., B. R. Goldsmith and S. Linic, "Theory-Guided Machine Learning Finds Geometric Structure-Property Relationships for Chemisorption on Subsurface Alloys", *Chem*, Vol. 6, No. 11, p. 142–158, 2020.

350. Liu, C.-Y. and T. P. Senftle, "Finding Physical Insights in Catalysis with Machine Learning", *Current Opinion in Chemical Engineering*, Vol. 37, No. 14, p. 14–18, 2022.

351. Jain, A., S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, "Commentary: the Materials Project: a Materials Genome Approach to Accelerating Materials Innovation", *Advanced Physical Letters Materials*, Vol. 1, No. 10, p. 011002, 2013.

352. Ye, Y., S. Hou, L. Chen, X. Li, L. Zhao, S. Xu, J. Wang and Q. Xiong, "ICSD,An Automatic System for Insecure Code Snippet Detection in Stack Overflow over Heterogeneous Information Network", *Computer Security Applications*, Vol. 19, No. 10, p. 542–552, 2018.

353. Chen, T., Y. Zhou and M. Rafailovich, "Application of Machine Learning in Perovskite Solar Cell Crystal Size Distribution Analysis", *Materials Research Congress Advances*, Vol. 4, No. 14, p. 142–152, 2019.

APPENDIX A: DETAILS ABOUT BIBLIOMETRIC ANALYSIS OF SOLAR ENERGY UTILIZATION

Here the list of synonyms used to replace the words are given.

Table A.1. List of Synonyms used in Bibliometric analysis.

Replace to	**Pattern to Replace**
algorithm	algorithms
ANFIS	ADAPTIVE NEURO FUZZY INFERENCE SYSTEM
ANN	ARTIFICIAL NEURAL NETWORK
ANN	ARTIFICIAL NEURAL NETWORKS
ARTIFICIAL INTELLIGENCE	ARTIFICIAL INTELLIGENCE TECHNIQUES
battery	batteries
BISMUTH VANADATE	$BIVO_4$
BISMUTH VANADATE	$BIVO_4$ NANOSHEET
Biohydrogen	BIO-HYDROGEN
Biomass	BIOMASS-TO-LIQUID
Biomass	BIOMASS AND BIOFUELS
Biomass	BIOMASS FUEL
Biomass	BIOMASS CONVERSION
carbon dioxide	CO_2
carbon dioxide	CO_2 reduction
carbon dioxide	CO_2 conversion
carbon dioxide	carbon-dioxide
carbon dioxide	CO_2 splitting
carbon dioxide	CARBON DIOXIDE UTILIZATION
concentrated solar power	solar concentrator
concentrated solar power	solar concentration
concentrated solar power	SOLAR CONCENTRATING TECHNOLOGIES
concentrated solar power	concentrating solar power
desalination	WATER DESALINATION
desalination	HDH DESALINATION
desalination	SEAWATER DESALINATION

Table A.1. List of Synonyms used in Bibliometric analysis. (cont.)

Replace to	**Pattern to Replace**
desalination	solar desalination
design	optimal-design
economic analysis	economic
economic analysis	techno-economic analysis
electrolyzer	electrolyser
energy	energy analysis
energy	energy management
exergy	exergy analysis
forecasting	prediction algorithms
forecasting	forecasting model
forecasting	forecasting models
forecasting	WEATHER FORECASTING
forecasting	WEATHER FORECAST
forecasting	SOLAR POWER FORECASTING
forecasting	predictive models
forecasting	energy prediction
forecasting	predictive
forecasting	prediction
fuel cell	fuel cells
hybrid	hybrid system
hybrid	hybrid energy system
hybrid	HYBRID ENERGY SYSTEMS
hydrogen	hydrogen-production
hydrogen	hydrogen evolution
hydrogen	hydrogen energy storage
hydrogen	hydrogen fuel
hydrogen	hydrogen production
hydrogen	hydrogen generation
hydrogen	solar hydrogen
hydrogen	hydrogen storage
optimization	multi-objective optimization
optimization	optimisation
performance	efficiency
performance	performance analysis

Table A.1. List of Synonyms used in Bibliometric analysis. (cont.)

Replace to	**Pattern to Replace**
performance	feasibility
perovskite solar cell	ALL-INORGANIC PEROVSKITE SOLAR CELL
phase change material	phase change materials
photocatalyst	ADVANCED PHOTOCATALYST
photoelectrochemical	ALL-VANADIUM PHOTOELECTROCHEMICAL CELL
photoelectrochemical	photoelectrochemistry
photoelectrochemical	PHOTOELECTROCHEMICAL PROCESS
photoelectrochemical	PHOTOELECTROCHEMICAL CELL
photoelectrochemical	PHOTOELECTROCHEMICAL CELLS
photoelectrochemical	PHOTOELECTROCHEMICAL ACTIVITY
photoelectrochemical	PHOTOELECTROCHEMICAL REACTION
photoelectrochemical	PHOTO-ELECTROCHEMICAL SOLAR DEVICES
photovoltaic	pv
photovoltaic	solar photovoltaics
photovoltaic	solar photovoltaic system
photovoltaic	photovoltaic power systems
photovoltaic	solar photovoltaic
photovoltaic	photovoltaic systems
photovoltaic	solar pv
photovoltaic	photovoltaics
photovoltaic	photovoltaic system
photovoltaic	BIFACIAL SOLAR PHOTOVOLTAIC
photovoltaic	BIFACIAL PV
renewable energy	renewable energy sources
renewable energy	renewable
solar cell	solar cells
solar cell	ALL-POLYMER SOLAR CELLS
solar cell	BIFACIAL SOLAR CELLS
solar collector	collector
solar collector	collectors
solar fuel	solar fuels
storage	energy-storage
storage	energy storage
thermal	thermal energy storage

Table A.1. List of Synonyms used in Bibliometric analysis. (cont.)

Replace to	**Pattern to Replace**
thermal	solar thermal
thermal	solar thermal energy
thermal	thermal energy
titanium dioxide	TiO_2
water	H_2O
water splitting	H_2O SPLITTING
water splitting	PHOTOELECTROCHEMICAL WATER SPLITTING
wind	wind energy
wind	wind power
wind	wind turbine
energy system	energy systems
optimization	multiobjective optimization
optimization	multi-objective optimization
solar collector	solar collectors
solar collector	solar absorber
photothermal	photothermal conversion
energy conversion	energy transition
conservation	energy conservation
solar radiation	solar irradiance
neural networks	artificial neural networks
neural networks	artificial neural network
neural networks	artificial neural network ann
neural networks	neural network ann
neural networks	neural network
support vector machine	support vector machines
support vector machine	support vector regression
support vector machine	SVR
long short-term memory	lstm
machine learning	machine learning algorithm
machine learning	machine learning algorithms
machine learning	machine learning model
machine learning	machine learning models
machine learning	machine learning techniques
machine learning	machine learning method

Table A.1. List of Synonyms used in Bibliometric analysis. (cont.)

Replace to	**Pattern to Replace**
machine learning	machine learning methods
machine learning	machine learning ml
support vector machine	vector machine svm
random forest	random forest rf
photovoltaic	solar photovoltaic pv
photovoltaic	solar photovoltaic
support vector machines	machines svm
data driven	driven models
long short-term memory	term memory lstm

APPENDIX B: DETAILS ABOUT NATURAL DYE SENSITIZED SOLAR CELLS

Here the results of usingdifferent algortihms and data structures are given. Lowest rmse giving pair was selected for each output.

Table B.1. Comparison of ML algorithms.

Output	Structure	Algorithm	Validation rmse	Train rmse	Test rmse
Efficiency	Encoded	GB	0.4	0.03	0.3
		RF	0.5	0.03	0.5
	Categoric	GB	0.5	0.02	0.4
		RF	0.5	0.3	0.5
FF	Encoded	GB	0.11	0.04	0.1
		RF	0.10	0.04	0.08
	Categoric	GB	0.12	0.06	0.1
		RF	0.11	0.06	0.09
J_{sc}	Encoded	GB	1.8	0.4	1.7
		RF	1.7	0.8	1.7
	Categoric	GB	1.8	0.6	1.6
		RF	1.2	0.5	1.5
V_{oc}	Encoded	GB	0.11	0.004	0.10
		RF	0.09	0.002	0.07
	Categoric	GB	0.11	0.005	0.12
		RF	0.11	0.005	0.11

APPENDIX C: DETAILS ABOUT PHOTOELECTROCHEMICAL WATER SPLITTING

Here details about ARM analysis and the results for the analysis of remaining semiconductors are given.

Table C.1. Details about ARM algorithm.

	intervals	given name
Calcination Temperature (First Layer)	0-100	0
	100-400	1
	400-500	2
	500-600	3
	600-700	4
	700-800	5
	800-900	6
	900-1000	7
	1000+	8
Calcination Time (First Layer)	0-1	0
	1-2	1
	2-3	2
	3-4	3
	4-5	4
	5-6	5
	6-7	6
	7-8	7
	8+	8
Calcination Temperature (Second Layer)	0-100	0
	100-400	1
	400-500	2
	500-600	3
	600-700	4
	700-800	5
	800-900	6
	900-1000	7

Table C.1. Details about ARM algorithm. (cont.)

	intervals	given name
Calcination Time (Second Layer)	0-1	0
	1-2	1
	2-3	2
	3-4	3
	4-5	4
	5-6	5
	6-7	6
	7-8	7
	8+	8
Last Calcination Temperature	0-100	0
	100-400	1
	400-500	2
	500-600	3
	600-700	4
	700-800	5
	800-900	6
	900-1000	7
	1000+	8
Last Calcination Time	0-1	0
	1-2	1
	2-3	2
	3-4	3
	4-5	4
	5-6	5
	6-7	6
	7-8	7
	8+	8
Power	0-100	0
	100-200	1
	200-300	2
	300-400	3
	400-500	4
	500+	5

Table C.1. Details about ARM algorithm. (cont.)

	intervals	given name
Wavelength	0-300	0
	300-400	1
	400-410	2
	410-420	3
	420-450	4
	450-500	5
	500+	6
Intensity	0-300	0
	300-400	1
	400-410	2
	410-420	3
	420-450	4
	450-500	5
	500+	6
Electrolyte Molarity	0-0.1	0
	0.1-0.25	1
	0.25-0.5	2
	0.5-1	3
	1-1.25	4
	1.25-1.5	5
	1.5-2	6
	2-3	7
	3+	8
Additive Molarity	0-0.1	0
	0.1-0.25	1
	0.25-0.5	2
	0.5-1	3
	1-1.25	4
	1.25-1.5	5
	1.5-2	6
	2-3	7
	3+	8

Table C.1. Details about ARM algorithm. (cont.)

	intervals	given name
pH	0-1	1
	1-2	2
	2-3	3
	3-4	4
	4-5	5
	5-6	6
	6-7	7
	7-8	8
	8-9	9
	9-10	10
	10-11	11
	11-12	12
	12-13	13
	13-14	14

Table C.2. Photocurrent limits for ARM analysis with three subsets.

Data subset	**Class**	**Lower Bound ($\leq$ mA/cm2)**	**Upper Bound ($>$ mA/cm2)**
Low Bias (V vs NHE $\leq$ 0.5 V)	A	1	30
	B	0.05	1
	C	0	0.05
Medium Bias (0.5 V $<$ V vs NHE $\leq$1.23 V)	A	3	48
	B	0.5	3
	C	0	0.5
High Bias (V vs NHE $>$ 1.23 V)	A	10	33
	B	3	10
	C	0	3

Table C.3. Prediction of Band gap with different model parameters.

	Percentage of Testing Set (100-%(Train+Validation Sets))					
	10%		25%		40%	
k-folds	Validation RMSE	Testing RMSE	Validation RMSE	Testing RMSE	Validation RMSE	Testing RMSE
3-folds	0.33	0.25	0.30	0.35	0.30	0.33
	ntree=25, nodesize=13		ntree=19, nodesize=27		ntree=16, nodesize=15	
5-folds	0.27	0.22	0.29	0.28	0.30	0.25
	ntree=19, nodesize=11		ntree=13, nodesize=17		ntree=19, nodesize=13	
10-folds	0.27	0.20	0.29	0.32	0.24	0.27
	ntree=16, nodesize=21		ntree=13, nodesize=21		ntree=13, nodesize=35	

Table C.4. ARM analysis of Band gap for Fe_2O_3.

Variable	**Level**	**LOW BG**	**MEDIUM BG**	**HIGH BG**
Doping	P	3.69		
	Pt	3.69		
	Sn	0.64		
	Ti	1.44	0.48	
	No doping	1.87	0.16	
Synthesis Method	anodization	3.69		
	chemical deposition	1.09	1.06	
	chemical growth	1.71		
	electrodeposition	2.77		
	hydrothermal	1.65	0.52	
	sol-gel	3.69		
	vapor deposition	0.92		
Calcination time	<1h	0.76	0.72	
	<2h	1.25		
	<6h	1.55		

Table C.4. ARM analysis of Band gap for Fe_2O_3. (cont.)

Variable	Level	LOW BG	MEDIUM BG	HIGH BG
Calcination Temperature	100-400	2.01		
	500-600	1.08		
	700-800	0.93		2.67
	800-900		2.22	
	No calcination	2.6	0.37	

Table C.5. ARM analysis of Band gap for ZnO.

Variable	Level	LOW BG	MEDIUM BG	HIGH BG
Doping	C		3.51	
	Co		3.51	
	Li			3.57
	Mn			3.57
	No doping	0.33	0.92	1.95
Synthesis Method	electrodeposition		0.71	2.85
	hydrothermal	0.42	1.22	1.51
	sol-gel			3.57
	sputtering			3.57
	vapor deposition		2.46	1.07
Calcination Temperature	100-40			1.75
	500-600			2.41
	No calcination	0.26	0.89	2.27
Calcination time	<1 h			2.07

Table C.6. ARM analysis of Band gap for TiO_2.

Variable	**Level**	**LOW BG**	**MEDIUM BG**	**HIGH BG**
Doping	C		3.51	
	Cr	2.34		
	Fe		1.15	
	Li		1.17	
	Nb			
	Ta			
	W	1.85	1.75	
	Si		3.51	
	No doping	0.05	0.58	
Synthesis Method	anodization		0.9	2.1
	commercial		2.45	1.07
	electrodeposition	3.69		
	flame oxidation		3.51	
	hydrothermal	0.43	0.45	2.6
	sol-gel		0.78	2.78
	chemical deposition			3.57
Calcination Temperature	400-500		0.66	2.91
	500-600		2.28	1.25
	600-700	0.19	1.89	1.47
	800-900	0.26	3.03	0.23
	No calcination	0.21	0.45	2.44
Calcination time	<1h	0.19	2.75	0.59
	<2h	0.07	1.9	1.57
	<3h		0.82	2.74
	<4h		0.48	3.09
	<5h		3.51	
	<6h			3.57
	<7h		3.51	

Table C.7. ARM analysis of Band gap for WO_3.

Variable	**Level**	**LOW BG**	**MEDIUM BG**	**HIGH BG**
Doping	N		3.51	
	No doping	1	2.4	0.11
Synthesis Method	chemical deposition		3.51	
	commercial		3.51	
	electrodeposition		3.51	
	hydrothermal	1.37	2.21	
	sol-gel	2.8	0.84	
	solvothermal	0.95	2.6	
	sputtering		3.15	0.37
	vapor deposition		3.51	
Calcination Temperature	100-400		3.51	
	400-500		3.51	
	500-600	1.15	1.92	
	800-900		3.51	
	No Calcination	0.98	2.46	0.12
Calcination time	<1 h		3.51	
	<3 h	1.22	2.35	
	<6 h		3.51	

APPENDIX D: DETAILS ABOUT CO_2 PHOTOREDUCTION

The results for different split and cross validation values in band gap prediction is given below.

Table D.1. Average validation RMSE for different train split percentage and fold number pairs for band gap prediction.

		Testing/training split% (p)			
		10	**20**	**30**	**40**
Fold number (k)	**3**	0.28	0.30	0.30	0.34
	5	0.26	0.27	0.28	0.31
	7	0.26	0.27	0.27	0.30
	10	0.25	0.26	0.27	0.28

APPENDIX E: COPYRIGHT LICENSES FOR FIGURES

The licenses of all adopted figures are given below. The rest of the figures and tables which are not mentioned are the original.

Table E.1. Copyright licenses for figures.

Number of figures	**Licensed content publisher**	**License Number**
Figure 3.1	Elsevier	5700871388739
Figure 3.2	Elsevier	5700880046437