# Machine learning analysis of photoelectrochemical water splitting

Burcu Oral , Elif Can and , Ramazan Yıldırım
Bogazici University, Department of Chemical Engineering, Istanbul, 34342, Turkey

Hydrogen is a promising energy carrier and important raw material for the chemical industry[1]. Large-scale hydrogen can be produced by photoelectrochemical water splitting; hence, significnant amount of research has been focused on ways to improve the efficiencies of these systems. This study aims to analyze photoelectrochemical water splitting (PECWS) literature between 2007 and 2020 via machine learning tools. The band gap of the semiconductors and photocurrent density of the PECWS cell were analyzed and predicted using association rule mining, decision trees, and random forest algorithms.

The dataset for PECWS consists of 584 experiments and 10560 data points extracted from 180 articles published between 2007- and 2020. Material properties, PECWS cell details, and operating conditions were used as descriptors, where Band gap (eV) and photocurrent density (mA/cm$^2$) were target variables.

Models were developed using R. Association rule mining (*apriori*) was used for identifying the influential descriptors on the desired target values (low band gap and high photocurrent density). Decision tree algorithm (*rpart*) was used for the classification of band gap and photocurrent density ranges and the random forest algorithm (*randomForest*) was used for regression. Hyperparameters were optimized using grid search and the best model was selected using cross-validation. Various train-test split ratios (10-40%) and k values for cross-validation (3, 5, and 10) were tried. The complexity parameter (*cp*) was optimized for the decision tree while the number of trees (*ntree*) and minimum split size (*nodesize*) were selected as hyperparameters for random forest algorithm. In all models, it was made sure that the Voltage (bias) - Photocurrent data from one experiment is kept in the same set (validation, train, or test) to prevent information leak from training to testing.

Association rule mining (ARM) analysis for band gap revealed that the anodization method for $TiO_2$ photoanode resulted in a high band gap. Li and Mo doping for $BiVO_4$ photoanode have a higher probability to give a low band gap compared to other dopants. Decision tree classification of band gap had the cp of 0.1 and accuracy of 0.78 and 0.75 for training and testing respectively. The class accuracy values show that high band gap values can be accurately predicted by the model, hence the rules to avoid high band gap could be trusted.

Figure 1. gives the predictive model developed for band gap using random forest. The best model was found using 40% test splitting and 10-fold cross-validation. The optimum model had *ntree* and *nodesize* of 13 and 35 respectively. The root mean square error of 0.24 and 0.27 were obtained for validation and testing sets respectively.

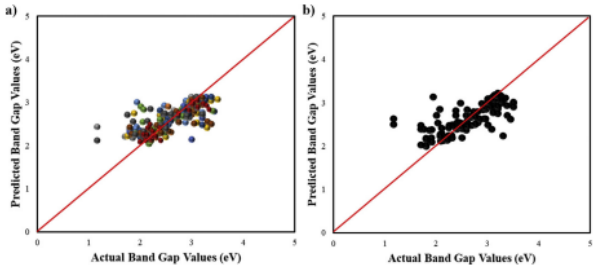
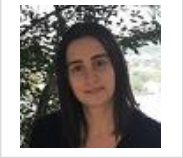



Figure 1. Predicted versus real band gap for a)validation and b) testing sets

Association rule mining analysis of photocurrent density was performed for each of the commonly used photoanodes (i.e. $BiVO_4$, $TiO_2$, $Fe_2O_3$, ZnO, and $WO_3$) separately because the important descriptors for the photoanodes differed. The bias and photocurrent density values were discretized into three categories based on water-splitting domain knowledge and intuition. Influences of each descriptor and their combined effect were obtained for high photocurrent density in low-bias conditions. $BiVO_4$ photoanode tends to have high photocurrent density at low bias when it is doped with Mo, or calcined around 600-700ºC, etc. $TiO_2$ photoanode doped with C and without any co-catalyst performs well under low bias. Decision tree for the photocurrent density (cp=0.01) model had training and testing accuracy of 0.61 and 0.54 respectively. The overall accuracy was lower than that for band gap classification since this dataset included reaction conditions; the non-standard testing procedure for PECWS may cause noise and decrease performance. Due to similar reasons, a strong predictive regression model could not be developed.

**References**

[1] D. Kumar, S. Singh, N. Khare Int. J Hydrogen Energy, 43 (2018) 507–512.



Burcu Oral is a Ph.D. candidate in the Department of Chemical Engineering at Boğaziçi University. She obtained her Bachelor's and M.Sc. degrees from Boğaziçi University and started her Ph.D. in 2019. Her work includes machine learning analysis of energy systems; solar energy applications, catalytic applications, batteries, etc.

Burcu Oral, e-mail: burcu.oral@boun.edu.tr tel: +90(212) 359-6873