CONSTRUCTION AND ANALYSIS OF A DATABASE FOR PHOTOCATALYTIC WATER SPLITTING FROM THE PUBLISHED ARTICLES

by

Elif Can

B.S., Chemical Engineering, Boğaziçi University, 2013

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

Graduate Program in Chemical Engineering
Boğaziçi University

2015

ACKNOWLEDGEMENTS

This thesis was ensued with the help and support of many individuals and I would like to start by mentioning them.

First of all, I would like thank my thesis supervisor, Prof. Ramazan Yıldırım due to his endless patience, enthusiasm and guidance during my graduate study. He was always accessible whenewer I need his help and wisdom. Besides, he always encouraged me to do my best and he has never lost his faith in me and my studies.

I would like to acknowledge my thesis committee, Prof. Ahmet Erhan Aksoylu and Assist. Prof. Mehmet Erdem Günay for their contibution to my thesis by sparing their valuable time.

I devote this thesis to my sensitive father Yılmaz Can, my warmhearted mother Gülsüm Can, and my dear and irreplacable sisters İpek Can and Betül Can. During my thesis studies I was far away from them by 736.1 km, but distance doesn't matter I can always feel their love and support with me.

I want to especially thank my fellow and housemate Tuğçe Nur Eren, to see her after long and tiring study nights was always pleasure to me, also Elif Gençtürk and Meltem Baysal deserves my special thanks for their moral support and unique sense of humor. I wish to express my sincere thanks to Barış Burnak, he was always with me on my left side and without his endless help and support this thesis would not be possible.

This long and difficult way of thesis writing became one of the most enjoyable period of my life with Elif Erdinç, Manouchehr Nadjafi, Özgür Yaşar Çağlar, Coşar Doğa Demirhan, Salih Emre Demirel and Serhat Erşahin, in other words "yemekhane" group. I feel very lucky for having chance to meet them and I know I will always miss those peerless people.

I would like to thank also all CATREL members and ChE group who I could not mention their names.

ABSTRACT

CONSTRUCTION AND ANALYSIS OF A DATABASE FOR PHOTOCATALYTIC WATER SPLITTING FROM THE PUBLISHED ARTICLES

The aim of this thesis was to develop a comprehensive database from published articles about photocatalytic water splitting oin literature, then to extract knowledge and to examine whether the results of an unperformed experiment could be estimated or not. Total number of instances gathered was 6378 from 129 different articles; later this large database was divided into 4 major subsets to improve the effectiveness of the analysis performed, these subsets included data about UV over TiO_2, visible light over TiO_2, ABO_3 perovskites and ABS_3 perovskites The database was gone through a pre-processing step to eliminate noisy instances, filling missing values and reducing number of dimensions or size of data while keeping its content. The hydrogen production rate (μmol/g-cat/h) was selected as the output variable and tried to be estimated by using several input variables such as semiconducting material, preparation methods, catalytic properties, or operational variables. Linear regression, artificial neural network, random forest and principal component analysis were applied to each dataset by using libraries and functions of "R". The parameters of each algorithm were changed within a certain intervals to find optimum conditions for best modeling. The models developed were evaluated using their standard error, root mean squared error and r-squared values. To evaluate model performances, 10-fold cross validation and standardized residual error analysis were performed. The best results were achieved with random forest algorithm for all subsets; the absolute error, rmse, and r-squared values were in the range of, 0.08 - 0.42, 0.17 - 0.61, and 0.64 - 0.97 respectively. In general, the error were scattered randomly, but some outliers were detected in subsets. The input significance and sensitivity analysis were also applied to subsets to extract more information. Catalytic variables (band gap, surface area, and particle size) were found as deterministic in titanium based photocatalysts whereas operational variables were more significant for hydrogen evolution over perovskite based photocatalysts.

ÖZET

YAYINLANMIŞ MAKALELERDEN SUYUN FOTOKATALİTİK AYRIŞTIRILMASI İLE İLGİLİ VERİTABANI OLUŞTURULMASI VE ANALİZİ

Bu tezin amacı suyun fotokatalitik olarak ayrıştırılıp hidrojen üretilmesi ile ilgili literatürde yer alan makaleleri inceleyerek kapsamlı bir veritabanı oluşturmak ve bu veri tabanından bilgi çıkarımı yapıp, daha önce yapılmamış bir deneyin sonucunu tahmin etmeye çalışmaktır. 129 farklı makaleden toplam 6378 örnek toplanmış ve bu veritabanı 4 anlamlı alt kümeye ayrılmıştır (UV ışığıyla TiO_2, görünür ışıkla TiO_2, ABO_3 ve ABS_3 perovskitleri kullanılarak yapılan çalışmalar). Hatalı ve tutarsız verileri çıkarıp, eksik verileri tamamlamak, ayrıca anlamsız değişkenleri elemek amacıyla her alt küme modellenmeden önce ön işleme tabi tutulmuştur. Katalitik özellikler, hazırlama yöntemleri ve operasyon koşulları gibi giriş değişkenleri kullanılarak, sonuç değişkeni olarak belirlenen hidrojen üretim oranı (μmol/g-kat/s) tahmin edilmeye çalışılmıştır. R yazılım programının içerdiği farklı fonksiyonlar kullanılarak doğrusal regresyon, yapay sinir ağı, rastgele orman ve temel bileşen analizi metodları her alt kümeye uygulanmıştır. Her algoritma için çeşitli parametreler, en iyi sonucu veren modeli bulana kadar belirli aralıklar içinde değiştirilmiştir. Elde edilen modeller standard hata, karesel ortalama hata ve determinasyon katsayılarına göre değerlendirilmiştir. Bütün modeller arasında en iyi sonuca her zaman için rastgele orman algoritması ile ulaşılmıştır. Bu algoritma sonucunda hesaplanan standard hata, karesel ortalama hata ve determinasyon katsayı aralıkları sırasıyla şöyledir; 0.08 – 0.42, 0.17 – 0.61 ve 0.64 – 0.97. Yapay sinir ağı algoritması genel olarak daha başarısız sonuçlar vermiş olsa da titanyum bazlı fotokatalizörlerde "rastgele orman" modeline yakın souçlar elde edilmiştir. Modelin başarısını ölçmek amacıyla kalıntı değer analizi ve 10 katlı çapraz geçerlilik sınaması da uygulanmıştır. Elde edilen veritabanından mümkün olduğu kadar fazla bilgi çıkarabilmek amacıyla girdi önem analizi ve duyarlılık analizi her bir alt kümeye uygulanmıştır. Titanyum bazlı fotokatalizörlerle hazırlanmış veritabanıyla yapılan modellerde katalitik özellikler (enerji bant aralığı, yüzey alanı ve parçacık boyutu) daha önemli, perovskit tipi fotokatalizörler kullanılarak oluşturulan modellerde operasyon koşullarının daha önemli olduğu anlaşılmıştır.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

R^2	Coefficient of determination
Δ	Delta
α	Regression coefficients
β	Regression coefficients
μ	Mean
σ	Standard deviation

LIST OF ACRONYMS/ABBREVIATIONS

ACS	American chemical society
ANN	Artificial neural network
CART	Classification and regression trees
DEA	Diethylamine
e	Electron
eV	Electro volt
g-cat	One gram of catalyst
ID3	Iterative Dichotomiser
kg	Kilogram
kJ	Kilo joule
k-NN	K Nearest Neighbor
MAE	Mean absolute error
MJ	Mega joule
MLP	Multilayer perceptron
PCA	Principal component analysis
PWS	Photocatalytic water splitting
RE	Rare earth
Rh B	Rhodamine B dye
RMSE	Root mean squared error
QE	Quantum efficiency

SEM	Scanning electron microscopy
SSE	Sum of squared error
TEM	Transmission electron microscopy
TEOA	Triethanolamine
TGR	Temperature gradient reactor
TiNT	Titanium nanotube
UV	Ultraviolet
W	Watt
XRD	X-Ray diffraction

1. INTRODUCTION

Global energy demand is increasing and it is expected to be doubled by 2050. Solar energy is predicted to be an ideal alternative because it is renewable, sustainable, and free. Hydrogen has also been considered as a promising energy source due to its high energy density (140 MJ kg^{-1}), which is superior to those of gasoline and coal. Besides energy production from hydrogen combustion does not cause carbon emission, it results in a clean and useful by-product of water (Xie *et al.*, 2013). For these reasons, photocatalytic water splitting into H_2 and O_2 using semiconductor materials has received much attention.

In a simple PWS system, a semiconducting material absorbs light from sunlight or an illuminated light source in experiments. When an electron in the valence band of the semiconductor absorbs a photon from light it gains the energy of the photon and if this energy exceeds the band gap of semiconductor, it leads excitation of electron to the conduction band and a hole created in the valence band. That electron and hole act as reducing and oxidizing agent to produce hydrogen from H^+ and oxygen from O^{2-}. Electrons and holes may also recombine with each other without any chemical reaction (Ismail and Bahnemann, 2014). Since the Gibbs energy of PWS is +237 kJ/mol (1.23 eV/e) the band gap of the photocatalyst should be greater than 1.23 eV to produce hydrogen and oxygen by PWS (Chen *et al.*, 2010).

There are several examined photocatalysts in the literature such as metal oxide (oxides of Ti, Zr, Nb, Ta, W, Ga, Sn) and non-oxides. TiO_2 was the first examined material as the photocatalyst in the pioneering work of Fujishima and Honda (Fujishima and Honda, 1972). Since then, titanium based photocatalysts have been the most preferred semiconductors in PWS reaction due to its reliable activity and high applicability. Perovskites have gains attention for PWS reaction in the recent years because of easily modifiable catalyst properties; especially the band gap of a perovskite can be changed by changing their ingredients and compositions. However, most of these semiconductors requires UV light; hence, another focus in PWS research is to make semiconductor for visible light harvesting since the motivation behind the water splitting is to produce hydrogen by using sunlight as free and renewable energy source. Only the small portion (about four percent) of sunlight while the visible light made up about its 43%. At this point

dye-sensitization is one of the most common and promising way to make semiconductors visible light active. Another approach to have visible light driven systems is to change the bandgap of semiconductor by metal doping.

After significant amount of research, it was clearly shown that depositing promoter (co-catalyst) on semiconductor increases the efficiency (Fukuzumi *et al.*, 2013). Since the absorption edges of oxidation and reduction states are different, to use particular promoter provide separate oxidation and reduction reagents to the PWS. Also they prevent reverse reaction (formation of water). Pt, Ru, Ir and C, or N are the most preferred promoters used for this purpose.

A wide range of catalyst preparation methods have been also developed. The most common methods are sol-gel method and solid-state reaction; the photo-deposition is another preferred deposition method for co-catalysts.

The catalytic performance of a photocatalyst is controlled by its structure and morphology such as crystallinity, defects, structure distortion, particle size, surface area, and surface structure (Yan *et al.*, 2013). Defects on the surface of photocatalyst act as traps for gaseous products. Smaller band gap, smaller particle size and larger surface area are the preferred properties for a PWS catalyst.

External heating or any pressure adjustment is not necessary for PWS; most of the experiments are conducted under normal conditions. Batch type reactor is the most common one. Reaction solution is generally aqueous alcohol solution at different concentrations; methanol is the most used alcohol as sacrificial agent to inhibit electron hole recombination. Acidity of solution affects the hydrogen yield positively (Zhang *et al.*, 2013).

Significant progress has been made in recent years about PWS efficiency. Over 140 metal oxides, nano-metal-oxides, perovskites and other semiconductor materials are known to catalyze the photochemical water-splitting reaction and to develop an efficient photocatalyst, several parameters are needed to be balanced. Semiconductor material, promoter type, crystal structure, catalyst weight, reaction solution, type and intensity of light source are the major factors that affect catalytic performance. Great time, labor and money are required to adjust these parameters in the way that the best activity is achieved.

At that point, data mining is a method to extract non-trivial, previously unknown, and potentially useful knowledge from the published articles and performed experiments, then to decide the optimum preparation and operational variables of a new process (Baydoğan, 2014). There are various types of data mining tools; the common ones are linear regression, logistic regression, k-nearest neighbor algorithm, decision trees, support vector machines, and artificial neural networks.

Up to now, the most preferred pattern recognition models in chemical engineering are artificial neural network (ANN) and decision tree. ANN is inspired by the human neural processing; it can be considered as a simplified model of the brain. In that model, inter-connected nodes and weighted links, and a perceptron model are used the output node. Then it compares the output node with a defined threshold value, as long as the output node is bigger than that threshold value, iteration continues. In decision tree, set of parameters organized hierarchically in such a way that the output can be determined following these parameters (Sorzano, 2014). That model is easier to understand visually, and more appropriate to extract rules and trends from the dataset.

Yamada analyzed different operational and catalytic variables for methanol synthesis and made all-encompassing calculations to develop optimum catalysts (Yamada *et al.*, 2004). In 2007, Günay and Yıldırım modeled the design of Pt-Co-Ce/Al_2O_3 catalyst for the low temperature CO oxidation in hydrogen stream by using ANN with 30 points dataset (Günay and Yıldırım, 2007). Omata was used also ANN to optimize of temperature profile of temperature gradient reactor (TGR) for dimethyl ether synthesis from syngas (Omata *et al.*, 2009). Günay and Yıldırım published another article about ANN analysis of selective CO oxidation with different preparation and operational variables than that in 2007 (Günay *et al.*, 2010) Manfred Baerns and *et al.*, made a statistical analysis of past 1870 datasets on oxidative methane coupling to have a new insight into the composition of high-performance catalysts, in 2011 (Zavyalova *et al.*, 2011). As a novel approach, decision tree and modular neural network was used consecutively as data mining tools to model preferential CO oxidation over promoted Au/Al_2O_3 by Günay and Yıldırım, in 2013 (Günay *et al.*, 2013). The recent article about data mining in the area of chemical engineering was published by Odabaşı, *et al.*, in 2014; in the article, the water gas shift data collected from the literature were modeled using decision tree, support vector machine and artificial neural network (Odabaşı *et al*, 2014).

In this study, a database for PWS system was constructed from the published work in the literature and analyzed using linear regression, artificial neural networks, decision tree and principal component analysis. The model that describes the data best was found by systematic tuning its parameters and evaluating the performance after each adjustment using three criteria (standard error, root mean squared error and r-squared value); model validation was also made by cross validation and residual analysis. Once the model is well constructed some analyses (input significance, sensitivity analysis) were also applied to observe the correlation between variables and to extract knowledge.

In the “Literature Survey” (Chapter 2) published articles about photocatalytic water splitting and data mining in catalyst studies are summarized; the details of the photoreaction system and the data mining tools used in this work are also explained in detail. Process of constructing database as well as the computational details about linear regression, neural network and decision tree algorithms are presented in Chapter 3. In Chapter 4, (Results and Discussion), results obtained in this work were shown and discussed; then the major conclusions and recommendations for possible future works are summarized in the last Chapter 5.

2. THESIS BACKGROUND

2.1. Photocatalytic Water Splitting

Photocatalytic water splitting (PWS) can be investigated in 3 steps; (i) absorption of photons with energies exceeding the semiconductor band gap and leading to the generation of electron and hole pairs in the semiconductor; (ii) migration of these photo-generated particles (e^- and h^+) which resulted in charge separation; (iii) surface chemical reactions between these carriers and present compound (e.g., water, water-alcohol solution) (Ismail *et al.*, 2014). During the third step, these carriers may recombine with each other and form water instead of participating in chemical reactions. PWS gives higher efficiencies under UV light, however PWS is tried to achieve efficiently also under visible light.

The reaction mechanism for photocatalytic water splitting is described in the following half-cell reactions;

$$2H^+ + 2e^- \rightarrow H_2 \tag{2.1}$$

$$H_2O \rightarrow \frac{1}{2}O_2 + 2H^+ + 2e^- \tag{2.2}$$

$$\Delta G = +237\ kJ/mol \tag{2.3}$$

Water splitting reaction needs the standard Gibbs free energy change ΔG of 237 kJ/mol or 1.23 eV. (Chen *et al.*, 2010).

Quantum efficiency (QE) term is used to describe the effectiveness of semiconductor materials, it is described with the following equation:

$$\begin{aligned} &overall\ quantum\ efficiency\ (\%) \\ &= \left(\frac{(2xNumber\ of\ evolved\ H_2 molecules)}{Number\ of\ incident\ photons}\ x100(for\ H_2\ evolution) \right) \end{aligned} \tag{2.4}$$

Over 140 metal oxides, nano-oxides, perovskites and other semiconductor materials are known to catalyze the photochemical water-splitting reaction. Significant progress has

been made in recent years about PWS efficiency. To develop an efficient photocatalyst several variables are needed to be balanced; semiconductor material, promoter type, crystallic structure, catalyst weight, reaction solution, type and intensity of light are the major variables that affect catalytic performance.

2.2. Effective Parameters on Photocatalytic Hydrogen Production

2.2.1. UV-Active Photocatalysts for Water Splitting

In PWS, the type of semiconducting material is the key variable that affects efficiency directly. There are three main group of photocatalysts in the literature; metal oxide (oxides of Ti, Zr, Nb, Ta, W, Ga, Sn), non-oxide photocatalysts, and perovskites.

It was verified that TiO_2 can produce hydrogen and/or oxygen from water under UV irradiation in several articles. In 1972, Fujishima and Honda were the first ones who worked with TiO_2 as semiconductor to split water into hydrogen and oxygen under UV-light. After their work, in 1981 the group of Duonghong studied on PWS with TiO_2 and possible promoters. It was found that when Pt and RuO_2 particles are used with TiO_2 the quantum yield could reach up to 30% (Duonghong *et al.*, 1981). During the past 40 years, it was found that the photo catalytic materials which contained either transition-metal cations (e.g.,Ta^{5+}, Ti^{4+}, Zr^{4+}, Nb^{5+}, Ta^{5+}, W^{6+}, and Mo^{6+}) or typical metal cations (e.g.,In^{3+} and Sn^{4+}, Ga^{3+}, Ge^{4+}, Sb^{5+}) were more efficient than others (Ismail *et al.*, 2014).

The layered titanates, $Na_2Ti_3O_7$, $K_2Ti_2O_5$, and $K_2Ti_4O_9$ were also found effective for PWS even without any co-catalyst (Shibata *et al.*, 1987). The reason for this is the fact that, titanates with layers act as trapping agents. They provide appropriate regions for absorption of gaseous products so they gain time to proceed forward reaction. As it is mentioned before, there are numerous publications about TiO_2 catalyzed PWS, as they are tabulated in Table 2.1 Semiconductor material used in the experiments involved in Table 2.1 is TiO_2.

P25 is a commercial mixture of titanium phases, which was developed by Degussa and it includes 20% rutile and 80% anatase phases of titanium. , Pt promoted titanium dioxide prepared via sol-gel method is the most common and relatively more effective titanium based semiconductor for PWS. However, to project the variety of catalyst

preparation methods and promoter types, which have been used with titanium, the table includes some studied less efficient photocatalysts as well.

Table 2.1. TiO_2 based Photocatalysts for Water Splitting.

Catalyst Preparation Method	**Promoter**	**Crystal Phase**	**Reaction Solution**	**Light Source/ Power/ Wavelength**	**H_2 Production Rate (µmol/h/gcat)**	**Reference Article**
RF Magnetron Sputtering	N-based	Anatase	10% Aqueous Methanol Solution	Xenon Lamp 300W 400 nm	598	Wang *et al.*, 2013 (283)
Sol-gel Method	-	Rutile:Anatase (1:1)	15% Aqueous Methanol Solution	Xenon Lamp 300W 400 nm	24.7	Kokporka *et al.*, 2013.
Solvo-Thermal Route	-	Anatase	50% Aqueous Methanol Solution	UV-light 300W 365 nm	2110	Lee *et al.*, 2013
Sol-gel Method	Ag-based	Anatase	Pure Water	Hg Lamp 300 W 254 nm	185	Onsuratoom *et al.*, 2011
Sol-gel Method	Pt	Anatase	10% Aqueous Methanol Solution	Metal Halide Lamp 150W 420 nm	817	D'Elia *et al.*, 2011.
Deposition-Precipitation	Ni	Rutile:Anatase (3:7)	50% Aqueous Methanol Solution	UV-light 300W 365 nm	674	Oros-Ruiz *et al.*, 2014
Sol-gel Method	Ir	Anatase	Pure Water	Xenon Lamp 300W 400 nm	3.2	Khan *et al.*, 2009
Hydrothermal Method	Fe	Anatase	Pure Water	Xenon Lamp 300W 400 nm	25	Khan *et al.*, 2008
Sol-gel Method	Pt	P25	9% Aqueous Methanol Solution	Hg Lamp 300 W 254 nm	6000	Sreethawong *et al.*, 2006
Sol-gel Method	Ni	Anatase	9% Aqueous Methanol Solution	Hg Lamp 300 W 254 nm	590	Sreethawong *et al.*, 2005

A wide range of metal oxide materials are known to work as efficient photocatalysts for water splitting under UV irradiation. For example, Lin and co-workers examined hydrogen production by PWS over niobium oxide photocatalysts. They also loaded a series of nanoparticles such as Pt, Au, Cu and NiO as co-catalyst to observe its effect on photocatalytic activity. The highest efficiency was exhibited by Pt loaded niobium oxide catalyst with a rate of H_2 production of 4647 μmol/h/g-cat (Lin *et al.*, 2011). Generally, the metal oxide semiconducting materials have been studied as combinations of two of them, and this topic is explained in one of the following sections in detail.

CdS-based semiconductors have been also studied very often for the photocatalytic hydrogen evolution from water splitting under visible light. Satnish and Viswanath examined the activity of the pure, noble metal loaded and supported mesoporous CdS in 2007. The Pt loaded mesoporous CdS showed the highest hydrogen production rate of 14150 μmol/h/g-cat among the catalysts tested. Al_2O_3 and MgO were also studied as support in the same study, and it was observed that MgO is a better option in terms of photocatalytic activity for both bulk and nanosize CdS (Sathish *et al.*, 2007).

SiC with different morphologies such as whiskery, worm-like and particulate were also employed as the catalyst for photocatalytic hydrogen production from water under visible light by Hao and his co-workers in 2013. The results showed that the particulate and worm-like SiC samples had hydrogen evolution rates of 3.75 μmol/h/g-cat and 3.69 μmol/h/g-cat, respectively; both were higher than that of SiC whiskers (2.04 μmol/h/g-cat). As a result, it was verified that physical properties of SiC samples had an impact on average hydrogen evolution rates and SiC was not a promising semiconducting material compared to the other oxide and non-oxide photocatalysts (Hao, *et al.*, 2013).

Perovskites are of significant interest in the field of many technological applications. Up to now, several perovskites materials have been studied as catalyst for PWS. The general formula presenting perovskites is ABX_3, the A represents the larger cation when B represents the smaller cation. X site is generally occupied by oxygen, however in recent studies S and F also have come out and sulfides (ABS_3) or fluorides (ABF_3) type perovskites have been examined (Kanhere, *et al.*, 2014).

The main requirements of a photocatalyst material to produce hydrogen by PWS are being in visible light-response range and ensuring minimum recombination rate. At that point, with extensive space for the combination of elemental compositions, perovskites provide a control of electronic band structure of ABX_3 (Lee *et al.*, 2004).

Among perovskites type semiconductors, alkali tantalates have drawn attention for efficient PWS under UV light. $NaTaO_3$ photocatalyst is the most common alkali tantalate catalysts studied for this purpose. Li *et al.* synthesized $NaTaO_3$ photocatalyst successfully by a two-step synthesis approach. They compared their results with previous articles in which $NaTaO_3$ was synthesized by solid state method. The results exhibited that the novel catalyst preparation method they used gave more than 4 times higher photoactivity than established methods. They also showed that promoting NiO as co-catalyst onto the surface of $NaTaO_3$ improved its photocatalytic efficiency (Li *et al.*, 2014).

Layered type perovskites act as trapping agents which prolong the life of electrons and holes, and reduces the recombination rate, so enhances the photocatalytic activity of related catalyst. In 2011, Huang *et al.* examined a series of layered perovskites semiconducting materials, $ASr_2Ta_xNb_{3-x}O_{10}$ (A=K, H; x=0, 1, 1.5, 2 and 3) synthesized by conventional solid-state reaction which is the most common preparation method for layered perovskites. They reached the maximum hydrogen evolution rate of 9280 μmol/h/g-cat when x was 1 and with Pt over $HSr_2TaNb_2O_{10}$ (Huang *et al.*, 2011).

Chen *et al.* claimed that it was probable to obtain novel photocatalysts by protonating layered perovskites. They prepared $H_{1.9}K_{0.3}La_{0.5}Bi_{0.1}Ta_2O_7$ and $H_{1.6}K_{0.2}La_{0.3}Bi_{0.1}Nb_2O_{6.5}$ by ion-exchange method in 2012. As a result of their research it was concluded that the interlayer modification for layered perovskites could be a promising way to construct a novel photocatalysts which had also high photocatalytic activity for water splitting (Chen *et al.*, 2012).

There are several semiconductor combinations have been studied in the literature. To provide better understanding, general characteristics of semiconductor combinations are given in the tabulated form of related articles as Table 2.2.

Table 2.2. Semiconductor Combinations for Photocatalytic Water Splitting.

Materials	**Catalyst Preparation Method**	**Promoter**	**Light Type (wavelength)**	**Highest H_2 Production Rate (µmol/h/gcat)**	**Reference Article**
Cadmium sulfide-Titanium dioxide	Ion exchange followed by sulfurization	-	Xe Lamp 420 nm	394	Li *et al.*, 2010
Zirconium phosphate-Titanium phosphate	Two-step sulfidation	CdS-ZnS	Hg Lamp 400 nm	2142	Biswal *et al.*, 2011
Zirconium dioxide-Titanium dioxide	Sol-gel process	Ag	Hg Lamp 254 nm	185	Onsuratoom *et al.*, 2011
Titanium dioxide-Zinc oxide	Sol-gel process	Zn	Hg Lamp 254 nm	1300	Pérez-Larios *et al.*, 2012
Titanium dioxide-Tin dioxide	Electrospinning an innovated precursor soln.	Sn	Hg Lamp 365 nm	200	Lee *et al.*, 2012
Silicon dioxide-Titanium dioxide	Sol-gel process	-	Xe Lamp 400 nm	12	Rungjaroentawon *et al.*, 2012
Zinc sulfide-Copper sulfide-Cadmium sulfide	Collodial precipitation	-	Xe Lamp 420 nm	837	Hong *et al.*, 2014
Copper sulfide-Titanium dioxide	Hydrothermal method	-	Xe Lamp 420 nm	570	Wang *et al.*, 2013

2.2.2. Catalyst Preparation Methods

Photocatalyst materials are obtained with different preparation methods and the most common one is sol-gel method for titanium based catalysts while the solid-state reaction

method is often used perovskite type catalysts. Some novel methods have been also tried to increase the photocatalytic activity.

In a novel study prepared by He and Guo in 2014, CdS nanorod particles were hydrothermally synthesized through a dissolution-recrystallization approach in concentrated ammonia solvent. They observed stacking fault structures within CdS nanorods and that type of structure promoted the separation of photo induced electrons and holes. As a result the best quantum yield they reached was 23.0% with the average hydrogen production rate of 5357 μmol/h/g-cat (He *et al.*, 2014)

It was already known that titanium based catalysts have good photocatalytic properties to be used as catalyst in water splitting reaction. There have been also some attempts to get titania with better crystalline structure and surface are. In 2014, Serrano *et al.*, have successfully applied a hard-templating pathway to obtain ordered mesoporous titania and they obtained larger surface are (241 m^2g^{-1}) and relatively smaller nanoparticles with 16nm mean crystal size. The highest hydrogen production rate of 13140 μmol/h/g-cat was achieved with 0.9% Pt promoted titanium dioxide (Serrano *et al.*, 2014).

The calcination, reduction, oxidation and drying steps of catalyst preparation affect the performance of the catalyst, and the influences of these steps have not been studied much; however there are still several articles in the literature about the impact of calcination temperature or calcination time on the activity of photocatalyst for PWS.

Ding *et al.* prepared a visible light driven $CaIn_2S_4$ photocatalyst using hydrothermal method in 2013. During catalyst preparation, they calcined the catalyst at different temperatures to observe its effect on surface area, particle size, band gap and eventually hydrogen production rate. Increasing calcination temperature decreased the surface area, while it increased band gap and particle size values of the related catalyst. The hydrogen production rate increased with increasing calcination temperature, up to 873 K, but further increment in the temperature caused activity loss of the photocatalyst. Same research group also investigated effect calcination time; morphological and optical properties of $CaIn_2S_4$ stayed almost the same. However the average hydrogen evolution firstly increased with increasing calcination period, it was optimum when the photocatalyst calcined 3h and then production rate decreased with increasing operation time (Ding *et al.*, 2013).

Shen and co-workers carried out a similar research, they synthesized with Rh doped $SrTiO_3$ photocatalyst for PWS, again in 2013. They obtained supportive results with the article mentioned above. The optimum calcination temperature was found as 973K for that catalyst and under that condition the amount of evolved hydrogen was 960 μmol/h/g-cat. The activity decreased as the calcination temperature was increased or decreased (Shen *et al.*, 2013).

2.2.3. Approaches to Modifying for Visible-Light Harvesting

Up to now, the highest quantum efficiencies for PWS are 90% and 56% by using ZnS and $NiO/La/KTaO_3$ respectively. However, both of the experiments were conducted under UV irradiation (Ismail *et al.*, 2014). Since the motivation behind the production of hydrogen by PWS is using sun as light source and sunlight is only 3% ultra-violet at ground level, these high efficiencies lose their values. At this point, approaches to modifying the electronic band structure, dye sensitization or novel methods for visible light harvesting become important topics in this field of research.

The most common method is modifying the electronic band structure by doping metal and non-metal to the semiconductor material. They create an impurity energy level and there are two types of impurities which are donor and acceptor impurities. For example when acceptor impurities are added to a semiconductor, their energy level E_a is just slightly below the conduction band energy E_c. Therefore new band gap value become smaller and visible light absorption of related semiconductor is more possible. This process is described in Figure 2.1. E_c, E_v, and E_g stands for conduction band energy level, valence band energy level and bad gap energy, respectively.

In the pioneering work of Borgarello *et al.* it is found that Cr doped TiO_2 could produce hydrogen by water splitting under visible light irradiation (Borgarello *et al.*, 1982).

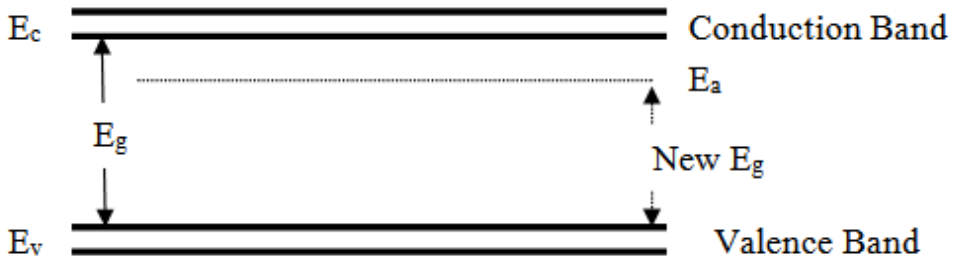

Figure 2.1. Changing Band Gap of Semiconductor by Ion Doping.

It is mentioned before that titanium-based semiconductor is the proven photocatalyst for water splitting reaction. Thus various types of metal and non-metal materials have been doped on TiO_2. In 2009, Khan *et al.* tried to change the band gap values of titanium nano-tube (TiNT) by introducing iridium and cobalt metal particles via ion-exchange method in 2009. They observed that band gap energies were successfully decreased to 2.5 eV and 2.6 eV from 3.1 eV of pure TiNT, and then TiNT became a visible light response material (Khan *et al.*, 2009).

Fe and Ni are widespread doping materials and Sun *et al.* examined pure TiO_2, single-doped and (Fe, Ni) co-doped TiO_2 nano particles for PWS. It was deduced from results of TEM and SEM, particle size decreased with metal doping. Also the photoluminescence spectroscopy indicated that the recombination rate suppressed relatively. The average hydrogen production rate of 361.64 μmol/h/gcat was measured when 5.0% Fe–4.0% Ni/TiO_2 catalyst was used (Sun *et al.*, 2012).

Nitrogen and carbon doping on titanium is another method to develop a visible light harvesting semiconductor. Wang *et al.* deposited N-doped titanium film by RF reactive magnetron sputtering method in 2013. They reached 601 μmol/h/gcat with 4.91% nitrogen doped TiO_2 as maximum hydrogen production rate. The band gap value was decreased down to 2.65 eV which was in required range for PWS (Wang *et al.*, 2013).

Relatedly, in 2013 Liu *et al.* developed a novel method to prepare carbon and nitrogen co-doping mesoporous TiO_2. Among all prepared catalyst, the photocatalytic generation of hydrogen with the rate of 81.8 μmol/h/gcat was the most successful one. It was concluded that the result might derived from the relatively high surface area of

fabricated photocatalysts and C,N-co-doping of titanium dioxide which was favorable for water splitting reaction under visible light (Liu *et al.*, 2013).

Jana *et al.* studied with rare earth (RE) elements (Y, La, Ce, and Yb) to fabricate doped $NaTaO_3$ in 2014. They used solid state synthesis method and different amounts of RE elements were doped on perovskite material successfully. Pt nano-particles were also promoted as co-catalyst by using situ-deposition method. The previously reported best doping element was La on $NaTaO_3$, and they reported that Y doping gave better results than that former catalyst (Jana *et al.*, 2014).

Dye sensitization is another promising method to harvest visible light for photocatalytic water splitting. The schematic representation of dye sensitized systems is given in Figure 2.2. (Chen *et al.*, 2010).

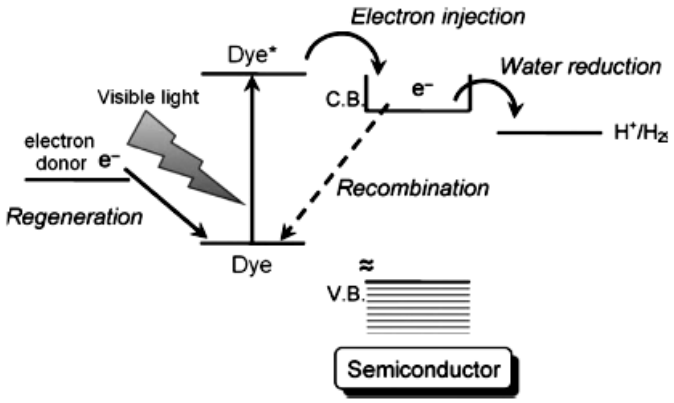

Figure 2.2. Working Principle of Dye Sensitization (Chen *et al.*, 2010).

Particularly sensitization of semiconductors with ruthenium complex dyes have been aroused interest in recent years.

Zheng *et al.* synthesized a new organic-inorganic visible light sensitive compound $[RuL(bpy)_2](PF_6)_2$ by polystep reaction in 2013. Pt was also loaded onto TiO_2 to improve the activity of photocatalysts by reducing recombination rate. In that study, pH was also tried to be optimized by adding sacrificial agent TEOA. As a result, the maximum hydrogen production rate was recorded as 2578 μmol/h/gcat when the optimal conditions were provided as pH 5 and 5% sacrificial agent (v/v) (Zheng *et al.*, 2013).

Metal-free dyes are the established material among compounds which sensitizes band gap of semiconductor to harvest visible light. In 2010, Fan *et al.* focused on the fabrication of $ZnIn_2S_4$/fluoropolymer fiber composites and they performed water splitting reactions by using that compound under Xe lamp irradiation which simulated visible light. Also zinc and indium ions are promoted onto fiber surface. They reached absorption edge value of 510 nm for those fiber composites and maximum hydrogen production rate of 406.25 μmol/h/gcat (Fan *et al.*, 2010).

Another effective way is rhodamine B (Rh B) dye sensitization and it was preferred in the research of Le and co-workers in 2013. They prepared Co doped Rh B sensitized titanium oxide catalyst and the moderate band gap of 2.58 eV which was lower than Co doped titanium oxide was achieved. 227.3 μmol/h/gcat was the highest hydrogen production rate reached during their research (Le *et al.*, 2012).

Wu *et al.* prepared La:$NaTaO_3$ photocatalysts by solid-state reaction method under hydrogen/argon and air. They also used Rh B dye to sensitize the band gap of semiconductor. The compound prepared under hydrogen and air atmosphere showed higher production rates with respect to air atmosphere (Wu *et al.*, 2014).

There are also some novel methods to enhances the photocatalyst activity of a compound by making it visible light responsive.

Khan *et al.* reported a new material which was platinum ion-exchanged TiO_2 nano-tubes to be used in photocatalytic water splitting reaction. They performed the experiments with both methanol solution and pure water. However, they reached relatively smaller hydrogen production rates which were 14.6 μmol/h/gcat and 2.3 μmol/h/gcat, respectively (Khan *et al.*, 2008).

In 2013, Zhang *et al.* used a novel method to synthesize CdS/bentonite nano-materials. They observed that the applied approach led to the expansion of layer spacing of nano powders and according to the results of analyses it was indicated that the absorption edge was blue-shifted. The 1% CdS promoted bentonite exhibited the highest activity of hydrogen production 114.7 μmol/h/gcat (Zhang *et al.*, 2008/7-11).

2.2.4. Approaches for Efficient Photogenerated Charge Separation

The charge recombination process, which means the electron in conduction band losses its energy and re-occupies the energy state of a hole in the valence band, competes with the photocatalytic process and reduces its efficiency. Two main approaches to suppress the surface back reaction are the addition of sacrificial agent (electron donor or acceptor) and creating photoactive sites, which are generally formed by promoting with noble metal or metal oxides co-catalyst on the surface of semiconductor.

In water splitting systems the methanol and ethanol are the most common sacrificial agents that preferred to favor photogenerated charge separation. Almost all of the researchers who study in this field conduct experiments with aqueous methanol or ethanol solutions. The concentration of alcohol in reaction solutions can reach up to 50% percent. Besides, there are some other sacrificial agents and acid solutions which are used to enhance the photocatalytic activity and they are tabulated in Table 2.3.

Table 2.3. Some Sacrificial Agent for Photocatalytic Water Splitting.

Materials & Promoter	**Catalyst Preparation Method**	**Sacrificial Agent**	**Light Type (wavelength)**	**Highest H_2 Production Rate (µmol/h/gcat)**	**Reference Article**
$Ta_2O_5/LiTaO_3$	Impregnation Method	Formic Acid	Hg Lamp 400	909	Zielinska *et al.*, 2012
$TiO_2/HTiNbO_5$	Exfoliation Restacking route	Diethanolamine (DEA)	Xe Lamp 420 nm	4735	Fan *et al.*, 2013
$(CuAg)_{0.15}In_{0.3}Zn_{1.4}S_2$	Novel Method	KI solution	Xe Lamp 420 nm	525	Zhang *et al.*, 2013

Pt promoted titanium based catalyst can be regarded as the most common method to separate photogenerated charges during water splitting reaction. In 2006, Sreethawond *et al.* fabricated 0.1-0.9 wt% Pt promoted mesoporous TiO_2 by sol gel process, incipient to wetness impregnation and photochemical deposition. Pt loading for these methods were 0.6, 0.4, and 0.5 wt%, respectively. The highest hydrogen production rate was 1385

μmol/h/gcat over the catalyst was prepared by sol gel process; the other methods also gave significant results (Sreethawong *et al.*, 2006).

TiO_2 microspheres were prepared by hydrothermal method by Wei *et al.* in 2013 and Pt was loaded by the impregnation-reduction method. Under UV light irradiation 1.2 wt% Pt/TiO_2 microspheres exhibited about 125 times higher hydrogen production rates than the bare titania microspheres (Wei *et al.*, 2013).

In 2012, Fang *et al.* tried to promote Au with different concentrations on TiO_2 nano-composites via a co-polymer assisted sol-gel method and they produced up to 16000 μmol/h/gcat hydrogen by using the prepared catalysts water splitting reaction. Similarly in other two articles it was mentioned that Au promoting onto TiO_2 was a promising way to suppress the recombination rate and slows down the electron-hole pair reactions and favors the photocatalytic water splitting reaction. One of those articles Au was promoted on TiO_2 with oxides of various metals such as Ag, Cu and Ni. The combinations of $Au\text{-}Cu_2O/TiO_2$ and $Au\text{-}NiO/TiO_2$ catalysts increased the hydrogen production rate significantly (2064 and 1636 μmol/h/gcat, respectively) (Fang *et al.*, 2012, Rayalu *et al.*, 2013, Oros-Ruiz *et al.*, 2014).

Some other noble metals such as Ag, Cu, Ni, Pd, and Rh were also promoted on different types of semiconducting materials. In 2013, Wang *et al.* studied Pd/CdS catalyst for water splitting and they concluded that 3.0 wt% was the optimal value for promoter amount and it gave average 1555 μmol/h/gcat hydrogen production rate (Wang *et al.*, 2013).

Gomathisankar and his group were also examined different noble metals to promote on ZnO and they found that deposition of Cu was approximately 130 times better than the other metal co-catalysts (Gomathisankar *et al.*, 2013).

Promoting noble co-catalysts on perovskite type semiconductor is also a promising way to separate photo-generated electrons and holes. In 2006, Jeong *et al.* synthesized NiO loaded $Sr_3Ti_2O_7$ which was a layered perovskite type oxide by solid state reaction method and polymerized complex method. The loading amount of NiO was changed from 0 to 8 percent and its effect on rate of hydrogen was observed. The layered structure of $Sr_3Ti_2O_7$

and presence of promoter suppressed the recombination of electrons and holes. As a result, the optimum catalyst was chosen as 3 wt% NiO/ $Sr_3Ti_2O_7$ in which hydrogen evolution rate reached up to 164 μmol/h/gcat (Jeong *et al.*, 2006).

Li and co-workers fabricated a new layered compound $H_{2.23}Sr_{0.67}Nb_5O_{14.335}$ to be used in PWS, in 2009. Pt was also promoted and its enhancement on activity was observed clearly. When bare $H_{2.23}Sr_{0.67}Nb_5O_{14.335}$ was used the total amount of hydrogen evolution could reach only 1700 μmol/gcat after 6 hours. However, loading 1 %wt Pt on that perovskite increased the cumulative hydrogen evolution amount up to 40000 μmol/gcat (Li *et al*, 2009).

In 2014, Yoshida *et al.* prepared Potassium hexatitanate ($K_2Ti_6O_{13}$) photocatalysts by a flux method. By using oxidative photo-deposition method rhodium co-catalyst was loaded. Different loading amounts were also examined to find the optimum value and it was concluded that 0.01wt% Rh loading gave the best activity values. The results of experiments showed that the less recombination of electron-hole pairs was achieved (Yoshida *et al.*, 2014).

Martha *et al.* tried a relatively unknown materials to catalyze water splitting reaction in 2013. They fabricated Rh and Cr_2O_3 promoted N-doped GaZn photocatalyst and the results were sufficient. The 3 wt% Rh – 1.5 wt % Cr_2O_3 loaded N-doped GaZn was considered as the optimum catalyst produced and they collected 2232.1 mmol hydrogen in 3 h (Martha *et al.*, 2012). Same year, Lee *et al.*, developed anatase TiO2 loaded onto pyrite FeS_2 (FeS_2/TiO_2) to enhance the production of hydrogen. They also produced considerable amount of hydrogen which was 11200 μmol in 10 h with 10 wt % FeS_2 on TiO_2 (Lee *et al.*, 2013).

Parayil *et al.* prepared carbon-modified TiO_2 composite materials with a green and facile approach which was called as hydrothermal synthesis followed by pyrolytic treatment, in 2012. They reached 210 μmol/h/gcat hydrogen production rate and it is believed the carbon promoting enhanced the photocatalytic activity by minimizing recombination rate (Parayil *et al.*, 2012).

In 2014, Chen *et al.* proved that the reaction sites of protonated layered perovskite in the presence of Pt at edge of layers had an obvious advantage to separate electrons and holes. The highest photocatalytic activity for hydrogen evolution was recorded as 491 μmol/h/gcat (Chen *et al.*, 2014).

2.2.5. Modification of Crystal Structure and Morphology

As it is mentioned before the separation of photo-induced electrons and holes is a fundamental requirement for photocatalytic water splitting reaction, and the crystal structure and morphology are key parameters which strongly affects the charge separation.

In 2007, Jitputti *et al.* synthesized mesoporous TiO_2 via hydrothermal method and according to the XRD and SEM results the crystal size of nanoparticles were 8nm and specific surface area of those particles were 215 m^2/g. The photocatalytic activity of titanium oxide was considerably higher than commercial titanium oxide and it was concluded that high surface area and small particle size favored the photocatalytic activity (Jitputti *et al.*, 2007).

In 2011, D'Elia *et al.* performed a photocatalytic water splitting with titanium oxide in different morphologies such as nano-particles, nano-tubes and aerogels. The nano-particles were more active because of their high surface areas. Nevertheless, the nano-tubes and aerogels also revealed high photocatalytic activity because of the possible role of their specific morphology which creates traps for photogenerated electrons and holes (D'Elia *et al.*, 2011).

There is also some novel morphologies experimentally created such as corn-like structure which was developed by Liu *et al.*, in 2011. That type of structure enhanced the light utilization, enlarged the surface area, and promoted the photogenerated electrons transfer (Liu *et al.*, 2011).

Similarly, in 2014 *Guo et al.* constructed a novel hierarchical 3D flower-like nano-micro titanium phosphate photocatalyst and used it in water splitting reaction. That catalyst exhibited high photocatalytic activity under visible light irradiation due to its suitable structure for charge separation (Guo *et al.*, 2014).

Chan *et al.* synthesized a series of cubic zinc-blend phase of $Zn_xCd_{1-x}S$ photocatalysts using Na_2S as the S source, in 2014. The XRD patterns showed new structural peaks different than a simple compound of ZnS and CdS and that catalyst exhibited very stable characteristics at least 50 h during photocatalytic water splitting under sunlight irradiation (Chan *et al.*, 2014).

2.2.6. Light Type

Wavelength, power, intensity and the type of light are the most important operational parameters that affect directly the photocatalytic activity of a PWS mechanism.

In 2013, Baniasadi *et al.* performed an experimental study with cadmium sulfide and zinc sulfide photocatalysts to investigate the effects of radiation intensity on hydrogen evolution rates. Hybrid lamps with three different light intensities such as 1000, 900, and 800 W/m^2 were used, the other qualifications of those lamps kept constant. It was clearly observed that produced total amount of hydrogen was directly proportional with light intensity (Baniasadi *et al.*, 2013).

D'Elia *et al.* synthesized titanium dioxide with different morphologies and conducted PWS in 2011. They also investigated the effect of type of lamp on PWS activity. 150 W metal halide lamp was used to simulate visible light and 450 W mercury lamp was chosen as UV light source. As it was expected, all titanium based catalysts with different morphologies gave better catalytic results under UV irradiation (D'Elia *et al.*, 2011).

2.2.7. Operating Pressure and Temperature

Photocatalytic water splitting reactions are usually carried out under atmospheric pressure and temperature, and there are barely articles in the literature about temperature and pressure dependency of PWS.

2.3. Data Mining Methods

The amount of data in the world has been increasing day by day and there is no end due to the ubiquitous computers' enhanced ability of data recording (Witten *et al.*, 2011).

At this point data mining and knowledge extraction make sense to convert large amounts of unsupervised or supervised data into meaningful sentences. It is a new branch of technology that helps especially companies focus on the most important information in the data they have collected about the behavior of their customers and potential customers. Not only industrial companies use, but also medical, biology and genetics researchers use these data extraction methods to predict the situation of a new comer (patient, subject eg.,).

Data mining draws ideas from machine learning, pattern recognition, statistics and database systems by using various tools. The common techniques under data mining are linear regression, logistic regression, k-nearest neighbor algorithm, decision trees, support vector machine, and artificial neural networks. Other traditional techniques may remain incapable due to high dimensionality, heterogeneity or enormity of data.

In the widest sense, data mining can be classified as descriptive and predictive. The descriptive one is useful in characterization of the data in a database. In the latter data mining category, a model is formed by using available data to make predictions about an unseen data. The subset groups of descriptive and predictive data mining tasks are categorized and showed in Figure 2.3. Some classification and regression methods which are for the predictive data mining tasks are explained in the following sections.

A simple data set comprises of rows and columns and depending on the application domain rows may be referred to instances, objects, points or feature vectors, and columns may be considered as attributes, properties or variables. The number of instances may be called as the “size of the data” whereas the number of features is referred to the “dimensionality” of the data.

There are some quality measures of a data set such as completeness, uniqueness, consistency; the presence of errors during data entry is also an important characteristic that weaken its reliability. There are several methods to preprocess data set before building model to overcome data quality problems (Baydoğan, 2014).

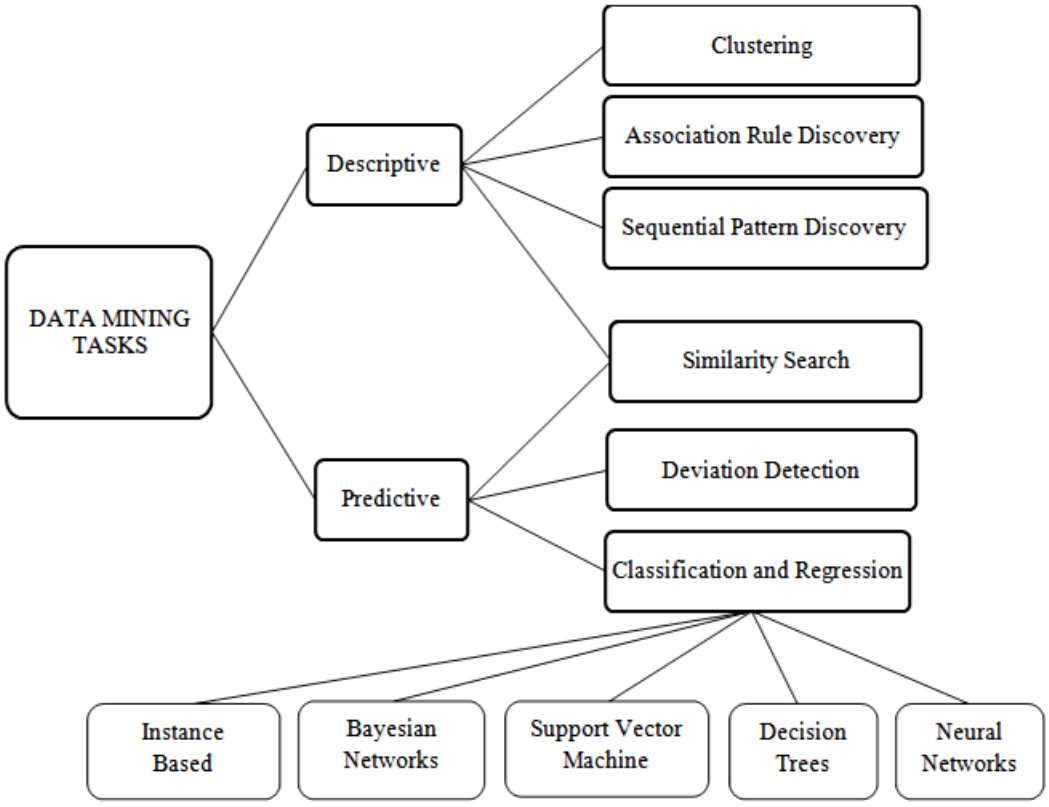

Figure 2.3. Data Mining Tasks.

Data mining differs from other model driven methods due to its data driven characteristic. In statistics, the motivation of researchers is finding the smallest data group that gives enough true estimates. As opposed to that, data mining is interested in developing models which describes the larger data sets well (Cios *et al.*, 2007). Building a good and easy to understand model is the key point in that sense.

Model evaluation is the principal part of data mining; it helps to decide which one is the best model that characterizes related data set and also gives hint about the future applicability of that model. Hold-out, k-fold cross validation, leave-one-out, and bootstrapping are the most common ways to evaluate model performance (Baydoğan, 2014).

Up to now, the most preferred pattern recognition models in chemical engineering are artificial neural network (ANN) and decision tree. Examples of studies in that field are mentioned in following sections.

2.3.1. Linear Regression

Linear regression model is the most frequently used and simplest statistical technique to describe the relationship between the trend of output variable and input variable. Output can be also called as dependent variable and inputs can be referred as independent variables. The general formula for regression is given in Equation 2.5.

$$Y = \alpha + \beta X \tag{2.5}$$

Where α and β are the regression coefficients. These coefficients can be found by least square analysis in which the error between actual and predicted data is minimized. In the basis of this analysis, the variance of Y is assumed as constant. The reliability of prediction can be evaluated by sum of square error (SSE) (Kantardzic, 2003).

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - y'_i)^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \tag{2.6}$$

Where y_i is the real output value and y'_i is the predicted value via model.

Multiple regression is an extended version of simple linear regression, in which there are more than one predictor variable. If the predictor variables are x_1, x_2, x_3 the general formula for multiple linear regression can be expressed as in Equation 2.7.

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \tag{2.7}$$

Where α and β_1, β_2, β_3 are the regression coefficients, again (Kantardzic, 2003). To find the coefficients, the same path can be followed with simple linear regression. The only difference is that may be useful to analyze the unknown parameters through a matrix calculation.

Outliers in a data set lower performance of least square analysis. As it is explained before, that analysis works by minimizing the sum of the squared error, so any data point which has a dependent value differs a lot from the rest of the data deviates the resulting value of coefficients so much. Moreover, linear regression methods are unable to define

non-linear relationships, so it is the major deficiency of that method due to the non-linearity of most systems in reality.

2.3.2. K-Nearest Neighbor Algorithm

K-nearest neighbor algorithm (k-NN) has been especially used in the field of pattern recognition. In this method each training sample is represented with a data point in an n-dimensional space (Han *et al.*, 2012). When a new and unknown sample is given to the space, the algorithm searches for the k training samples which are closest to the new comer. This closeness can be defined with different distance metrics such as Euclidean, Manhattan and Chebyshev distance. The most common one is Euclidean distance and can be evaluated with the following formula (Equation 2.8).

$$Euclidean\ Distance(x_1, x_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2} \tag{2.8}$$

Where x_1 and x_2 are two data points in the n-dimensional space. Application of this algorithm requires finding all distances between the test sample and all previous training samples. Both regression and classification are possible with k-NN algorithm. In k-NN classification, the class of test sample is defined by a majority vote of its neighbors, and for k-NN regression the output value of the test sample is the average of the values of its k nearest neighbors.

The key point in this algorithm is choosing optimum k value. When k value is too small than it should be, model becomes sensitive to outliers, if k-value is selected too large, data points from different classes may be included in neighborhood (Baydoğan, 2014).

One of the drawbacks of k-NN algorithm is that there is no model can be used for interpretation. Also, since it requires similarity measure for each data point, this algorithm may be problematic when the size of data set is too large.

2.3.3. Decision Tree

Decision trees are supervised learning methods and they can be categorized into two groups such as regression and classification trees. In the machine learning and applied statistics literature, there are a large number of algorithms based on decision tree induction (Kantardzic, 2003). In 1980s, J.Ross Quinlan developed a tree based algorithm known as Iterative Dichotomiser (ID3). Then he improved his research and presented a successor of ID3 which is called as C4.5, that algorithm became conducive to newer supervised learning algorithms. In 1984, a group of researchers in statistics published a book which is called as "*Classification and Regression Trees"*(CART) and binary decision trees were explained in detail (Han *et al.*, 2012).

A typical decision tree system searches for a solution in a part of the space by following a path from the root to the leaf nodes. In a decision tree, *internal node* is represented by a test on an attribute, each *branch* refers to an outcome of the test, and each *leaf node* stands for a class label. The top node in a decision tree can be considered as the *root* node. A sample listed data and a decision tree which is formed by using the given data set are shown in Table 2.4 and Figure 2.4 (Baydoğan, 2014).

Table 2.4. Sample Data Set for Decision Tree.

#No	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

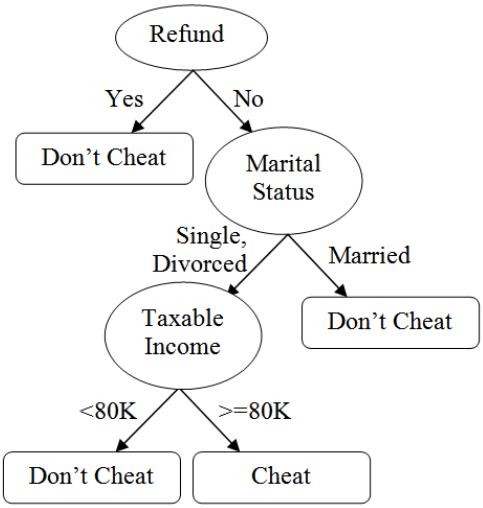

Figure 2.4. Sample Decision Tree.

In the Figure above square boxes in which the outcome is written represent *leaf nodes* and *internal nodes* are circle boxes that refers to the variables of the data set. Meaningful sentences can be deduced from that decision tree. For example, refund is a key parameter to decide whether cheating is probable or not. Also marital status is another deterministic parameter on a person bias to cheating. Although the given data set above is not so complicated, it is still not possible to extract these arguments easily with the naked eyes.

Partitioning the feature space depends on the feature types which can be nominal, ordinal and continuous, also the way of splits may be binary or multi-way split. For example, car types can be considered as nominal features and they can be categorized as family, sports and luxury by multi-way splits or as family and sports-luxury by binary splits. Determining best split way and when to stop splitting are very important in regression decision trees to inference logical knowledge. Splitting continues until certain level of impurity, certain depth level or certain number of terminal nodes achieved. The remarkable point during splitting is preventing model from under-fitting and over-fitting problems. By analyzing change in error rate of training set and test set together while increasing the number of nodes is the best way to solve those problems. In Figure 2.5, that situation is well explained visually, as a side note model complexity refers to number of nodes (Yuret, 2014). Prediction error of test sample decreases until certain number of nodes, before that point model performance can be considered as under-fitting. After that

point, prediction error of test sample starts to increase while error rate of training sample continues to decrease. It means model memorizes the training sample trend, that's why error rate of it decreases. However if an unknown test set is introduced, the model, fails to predict outcome of test set correctly. Exactly that point which is colored green in Figure 2.5, is the point where splitting should stop, in other words that number of nodes is sufficient enough for the model (Seni, 2013).

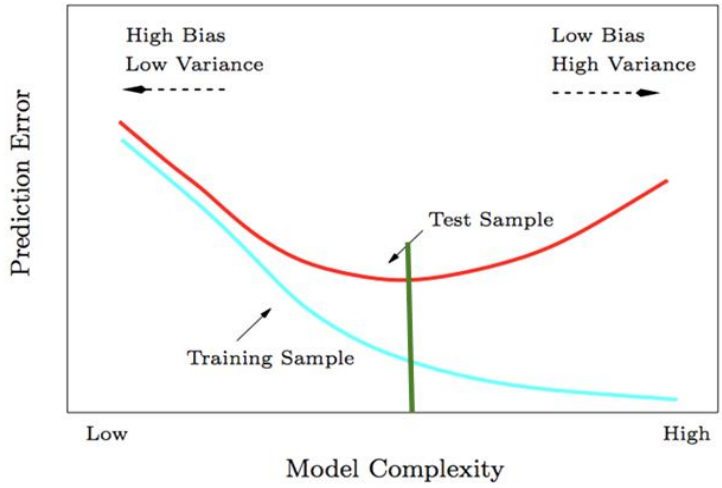

Figure 2.5. Model Complexity vs. Prediction Error Rate (Seni, 2013).

Other machine learning algorithms can perform around the same level with decision tree algorithm, but this is simpler in many ways. It consists lower computational complexity, more interpretable than other algorithms, easier to train because less sensitive to noisy data and generalizes a basic data set better.

In random forest, the individual trees are generated using a random set of variables and samples from present data set. Trevor Stephens, graduated from Analytics department of USF developed an analogy which relates random forest with a group of musicians. In an ensemble of talented instrumentalist, an off-note done by one musician can be tolerated by the others in the group. In analogy to this, random forest creates a large collection of individually imperfect decision tree models, and the ultimate result is evaluated by major voting in classification or taking average in regression problems. Thus a satisfactory output can be obtained even some poor decision trees in random forest model.

Random forest use bagging in which randomized samples are taken with replacement, so that some samples may occur in model more than once. The number of selected attributes to form individual trees is much smaller than the real number of attributes in data set (Han *et al.*, 2012). The trees are grown as much as possible without pruning, in that point over-fitting can come to mind as a problem. To avoid this issue tuning parameters that governs the number of features gains importance. K-fold cross validation can be used to minimize the test sample prediction error while searching optimal intervals for parameters.

2.3.4. Artificial Neural Network

Artificial neural networks (ANN) are inspired by human neural processing, briefly human brain. It is a simplified version of human brain and works as a function; it converts coming inputs into outputs by its ability of modeling (Baydoğan, 2014). The human brain has an approximately 10^{11} neurons and they are interconnected with each other with much more links (Kantardzic, 2003). ANN has also artificial neurons that cooperate to perform the desired function. These artificial neurons can be named as nodes in ANN and each node refers to a processing unit, and these processing units are connected via casual links. Figure 2.6 - 2.7 shows the similar structures of human brain and ANN. In a neuron, received stimuli from environment are transmitted via several dendrites to another neuron; the contact point is called synapse. Analogous to that process, in a neural network model, the inputs given to the input layer are transferred to next forward layer after applying a chosen activation function. That operation proceeds through hidden layer, if any, until ultimate output layer is reached.

Figure 2.6. Schematic of A Human Brain Cell (DeBellis, 2012).

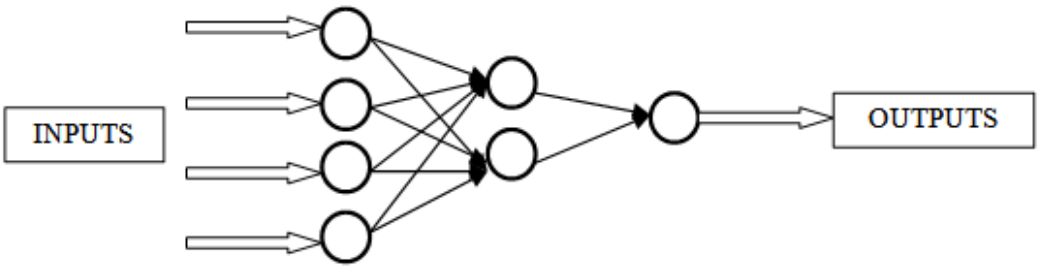

Figure 2.7. Neural Network Algorithm.

ANN algorithm has several properties and capabilities superior to other machine learning algorithms. First of all, ANN is a highly nonlinear algorithm and it is especially important if it is considered that most of systems in real world are non-linear. The weights of interconnection are arranged by tuning the parameters and applying a set of training examples. That's why ANN has a high adaptability to changes in the surrounding environment. ANN also offers a robust computation; its performance barely lowers with missing or noisy data, or disconnections of neurons. The same principles and same steps in methodology are used in all domains; it provides uniformity of analysis and design (Kantardzic, 2003).

Inter-connected nodes and weighted links form the model together, and output node simply sums up its entered inputs according to the weights of their links. There is a threshold value for the difference between predicted and real value, the weights are changed to find convincing output value until threshold value is reached. This process is illustrated in Figure 2.8 (Sadhu, 2012).

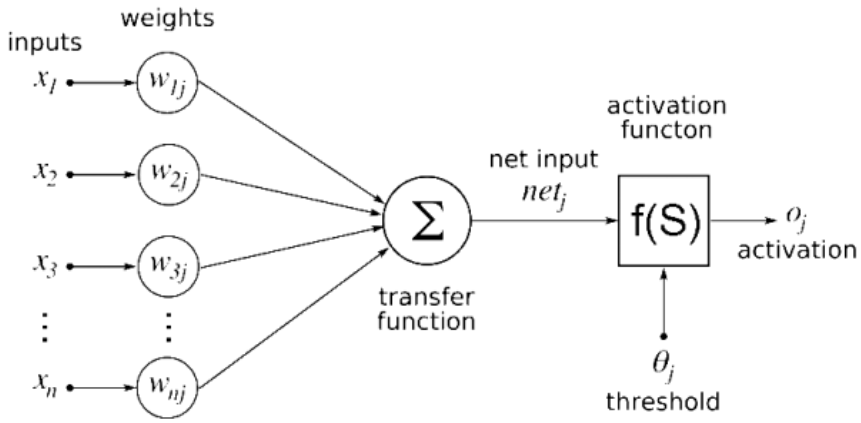

Figure 2.8. The Structure of Neural Network.

Where x_n's and w_{nj}'s represent inputs and weights of link, respectively, and models also includes an externally applied bias. Transfer function can be written as,

$$net_j = w_0 x_{j0} + w_1 x_{j1} + w_2 x_{j2} + \cdots + w_m x_{jm} = \sum_{i=1}^{m} x_i w_{ji} \tag{2.9}$$

There are several forms of activation function and they are listed in Table 2.5., activation function f(S) is one of the key parameters of network.

Table 2.5. Some Common Activation Functions.

Activation Function	Step	Linear	Log-sigmoid	Hyperbolic Tangent Sigmoid
Mathematical Relation	$y_j = \begin{cases} A \; if \; net \geq 0 \\ B \; if \; net < 0 \end{cases}$	$y_j = net$	$y_j = \frac{1}{(1 + e^{-net})}$	$y_j = \frac{(e^{net} - e^{-net})}{(e^{net} + e^{-net})}$

Perceptrons, in other words nodes, learn with back propagation method in conjunction with gradient descent method. In this method generally non-linear differentiable activation functions are used to model, and the aim is minimizing error. Let's assume that a network is trained to teach the first four letters (A, B, C and D) of the alphabet. Figure 2.9 shows how the back propagation algorithm works. By starting to train from the first letter "A" and change all the weights in the network once, next apply the second, third and last letter until the error becomes small enough which means network has recognized all the letters.

Error can be calculated with the given formula in Equation 2.10.

$$e(n) = t_j - y_j \tag{2.10}$$

$$E(n) = \frac{1}{2}(t_j - y_j)^2 = e^2 \tag{2.11}$$

$E(n)$ is called as cost function, where t_j is the target value and y_j represents output value. In order to fulfill the object, the derivative of cost function with respect to weights is taken

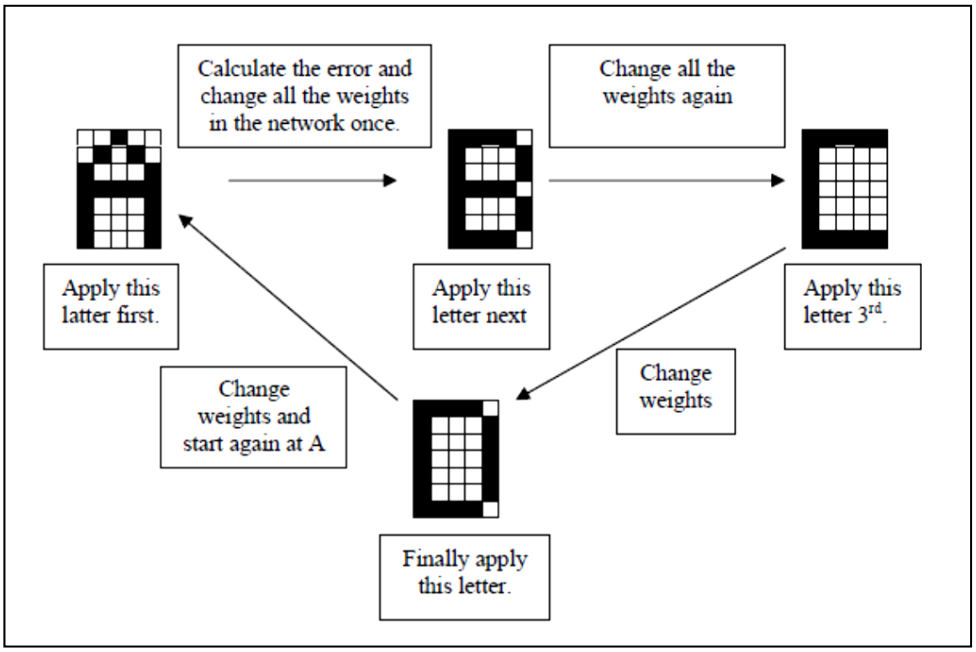

Figure 2.9. Schematic of A Sample Back Propagation Algorithm (Mallick *et al.*, 2013).

as shown in Equation 2.12 and calculation of derivatives flows backwards through the network. These derivatives give the direction of the maximum increase of the cost function. Then the weights are updated with a learning rate of α until small enough error is reached as shown in Equation 2.13. The magnitude of learning rate is so important because it defines whether model will converge or not and how fast will it converge (Baydoğan, 2014).

$$\frac{\partial E}{\partial w_i} = \frac{\partial E}{\partial y}\frac{\partial y}{\partial net}\frac{\partial net}{\partial w_i} \tag{2.12}$$

$$\frac{\partial E}{\partial w_i} \rightarrow How\ the\ error\ changes\ with\ respect\ to\ weights$$

$$\frac{\partial E}{\partial y} \rightarrow How\ the\ error\ changes\ with\ respect\ to\ output$$

$$\frac{\partial y}{\partial net} \rightarrow How\ the\ output\ changes\ with\ respect\ to\ weighted\ sum$$

$$\frac{\partial net}{\partial w_i} \rightarrow How\ the\ weighted\ sum\ changes\ with\ respect\ to\ weights$$

$$w_{new} = w_{old} + \alpha\left(t_j - y_j\right) * x_j \tag{2.13}$$

2.3.5. Principle Component Analysis (PCA)

Size and dimensionality of data set are some key parameters that affect performance of a model and predictability of outputs. There is a problem called as curse of dimensionality and this phenomenon was first expressed by Bellman in 1961. It means as the dimensionality of data set increases like hundreds or thousands of dimensions, it becomes sparse in the space. To avoid curse of dimensionality also accompanies such good features. First of all reducing dimension decreases the required time and memory by data mining algorithms. There may be some noise, irrelevant features or closely related features with each other; lowering dimension may help to eliminate them. It also provides a good visualization of data set (Baydoğan, 2014).

The most popular solution of that problem is the principal component analysis. It identifies important features which can explain the total variance of data, and then reduces the original map of observations to a lower dimensional space. Let's assume we have a p-dimensional data set, PCA sorts the principal axes of related data set in decreasing order by considering their variance, and it is important that all principal axes are uncorrelated with each other. After that, data set can be reduced to a lower n dimension by using outputs of PCA.

A geometric interpretation is shown in Figure 2.10. PCA is applied to a two dimensional data and the one dimensional projection of with principal component 1 and 2 is given. 2^{nd} PC has a lower variance, thus the projection of data on it may cause information loss about data set. In this specific case, to choose 1^{st} PC may be more logical to reduce dimension of current data (Liu, 2008).

2.4. Data Mining Studies in the Field of Catalysis

There is several catalyst studies combined with a data mining tool, and artificial neural network (ANN) is the most common one. Muneyoshi Yamada and his co-workers applied a radial basis functional ANN and all-encompassing calculations to develop the catalysts for methanol synthesis. The different compositions of defined elements (Cu, Zn, Al, Sc, B, Zr), preparation variables reaction conditions, calcination temperature (573-633K) and were analyzed as variables that affect the catalytic performance. 234 catalysts at

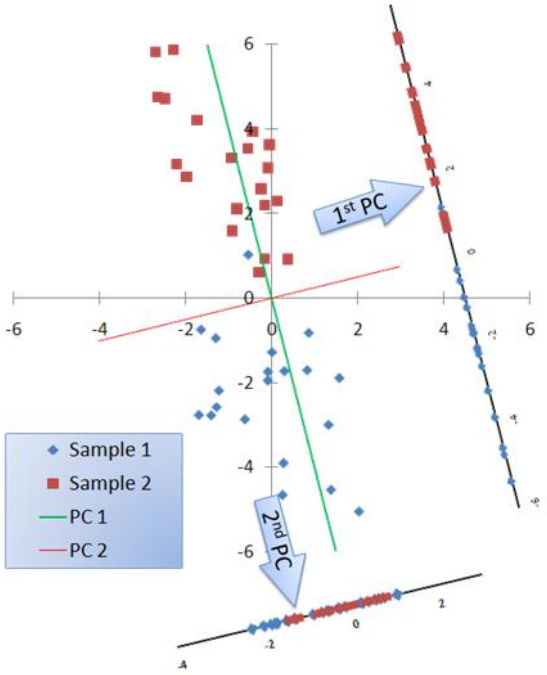

Figure 2.10. Geometric Interpretation of Principal Component Analysis.

different conditions were analyzed. At total, all-encompassing calculation is consisted of 2.6 million activities of all combinations. $Cu_{0.43}$ $Zn_{0.17}Al_{0.23}Sc_{0.11}B_{0.00}Zr_{0.06}O_{1.22}$ precipitated by 2.2 equiv of oxalic acid and calcined at 607 K was the global optimum one. That method also provided them good visualization (Yamada *et al.*, 2004).

In 2007, Günay and Yıldırım modeled the design of Pt-Co-Ce/Al_2O_3 catalyst for the low temperature CO oxidation in hydrogen stream by using ANN with 30 points dataset. The investigated parameters were percentage of Pt-Co-Ce metals, calcination temperature and time, etc. Then by adding time parameter as another input dataset points were increased to 120. It was concluded that the neural network modelling can be very helpful the design and efficiency in catalytic systems (Günay and Yıldırım, 2007).

A slightly different work was performed by Kohji Omata and his co-workers in 2009. They used artificial neural network and grid search to optimize of temperature profile of temperature gradient reactor (TGR) for dimethyl ether synthesis from syngas. In the TGR, the catalyst bed was divided into 5 parts and the data obtained from these distinct parts were used for training the ANN. Then to find the optimum temperature, grid search was applied on to the trained ANN. The results were analyzed and optimum condition was successfully found (Omata *et al.*, 2009).

Same group, Günay and Yıldırım, published another article about analysis of selective CO oxidation over promoted Pt/Al_2O_3 catalysts using modular neural networks, in 2010. The difference between that article and the one in 2007 was different process methods of preparation and operational variables in the network. In that sense, it was aimed that the accuracy of the model could be improved. The results were consistent with the expectations, root mean squared error values were decreased and the number of correctly classified instances was increased (Günay *et al.*, 2010)

Another research group, Manfred Baerns and *et al.* made a statistical analysis of past 1870 datasets on oxidative methane coupling to have a new insight into the composition of high-performance catalysts, in 2011. From last 30 years, about 1000 full-text references were analyzed, and 18 catalytic key elements were selected from originally 68 elements. For the identification of elements piecewise-constant function was applied in regression tree analyses. ANOVA test was also performed to identify the highly significant elements of the catalysts. Regression tree analysis could also be considered as an effective method to data extraction due to the satisfactory and logical test results (Zavyalova *et al.*, 2011).

After two years, in 2012, Yıldırım and his co-workers investigated the effects of second promoter on the water gas shift activity of $Pt\text{-}CeO_2/Al_2O_3$ catalyst experimentally, and the results were also analyzed by using modular neural network. As usual in catalytic processes, reaction temperature, feed ratio, promoter type and amount, and catalyst weight are the main parameters. The modular neural network model gave successful predictions about promoter type and operational variables (Günay *et al.*, 2012).

As a novel approach, decision tree and modular neural network was used consecutively as data mining tools to model preferential CO oxidation over promoted Au/Al_2O_3 by Günay and Yıldırım, in 2013. The dataset was reduced by using decision tree then it was modeled using modular neural network. The effects of co-catalyst were examined. To use decision tree as complementary to the modular neural network, provided easily comprehensible information. It was also claimed that, the novel approach could be also applied for similar catalytic systems (Günay *et al.*, 2013).

The current article about data mining in the area of chemical engineering was published by Odabaşı, *et al.* in 2014. By using approximately 4360 experimental data

points on water gas shift reaction over Pt and Au based catalysts, a dataset was formed. Three different tools; decision tree, support vector machine and artificial neural network were used to extract and analyze dataset. It was concluded that all three models were successful in terms of deducing useful rules and correlations (Odabaşı *et al*, 2014).

3. COMPUTATIONAL DETAILS

3.1. Experimental Data Collection

The aim of this thesis was to extract knowledge from published articles about photocatalytic water splitting on literature, then to examine whether the result of an unperformed experiment can be estimated, or not. For this purpose a large number of articles published from 2005 to 2014 about PWS in the literature are studied and a comprehensive database was constructed from those articles which are available in the online sources including Elsevier, Wiley Online Library, and ACS (American Chemical Society) Publications. Figure 3.1 shows the increasing attention on PWS over the years; the number of publications on PWS has increased 1140% in 2014 in respect to 2005.

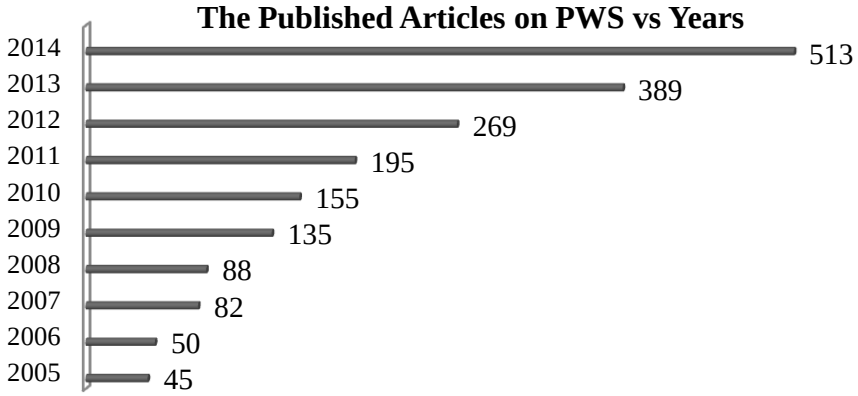

Figure 3.1. The Published Articles on PWS vs. Years.

PWS studies can be roughly divided into three categories; first group is the traditional studies in which titanium based catalysts are used and tried to be improved for better PWS activity. In the second group, the perovskite type semiconductors were used while the last group involves dye sensitization of present photocatalysts in first two groups. As it can be seen from Figure 3.2 which indicates the number of works in these three groups, there is a few number of publication in the literature about water splitting using dye-sensitized photocatalysts, thus that category was excluded from the database. To improve the performance of models described in remaining sections of thesis it was decided to divide database into sub-groups based on semiconductor type which affects the list of features most; the sub-groups used defined in Figure 3.3.

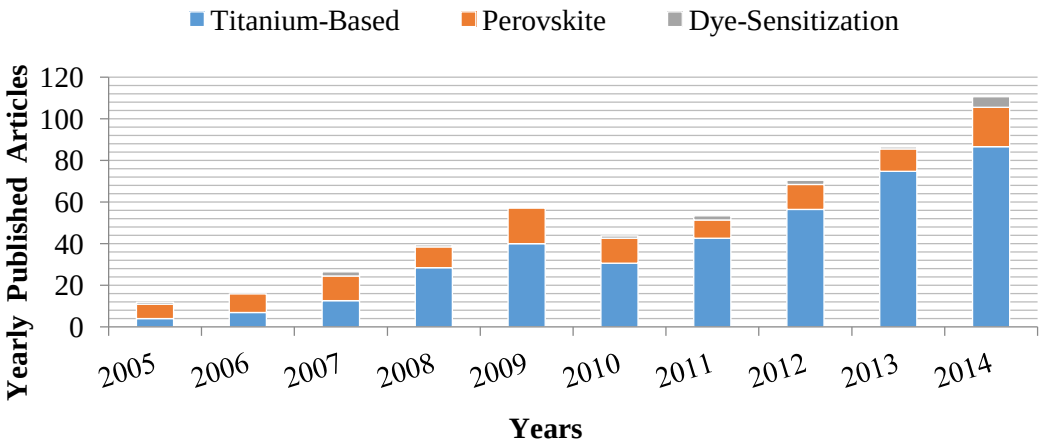

Figure 3.2. Relative Importance of PWS Study Branches.

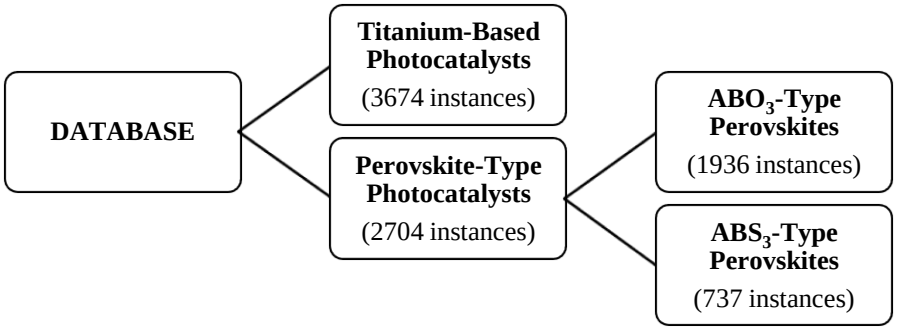

Figure 3.3. Subset Groups of Database.

While developing database, 196 articles were investigated in detail, and then 129 of them were decided as suitable to be included in data set. The excluded articles did not contain the variables used in the model or had some missing values of these variables. As a result, the total number of instances is recorded as 6378; the distribution of number of instances over sub-groups is also given in Figure 3.3.

The variables that affect activity of photocatalytic system vary according to divided groups. In the most general sense, semiconductor type, catalyst preparation method, promoter type and promoter weight percent, ingredients of reaction solution, light source properties such as power, intensity, wavelength and type of the light, and characteristics of catalyst material (like amount, band gap, crystal structure, and particle size) are recorded as the most significant factors.

As it is mentioned before, there are three subgroups in database, in the rest of thesis they are mentioned as “titanium database”, “ABO_3 database”, and “ABS_3 database”. As an exception, ABO_3 and ABS_3 databases can be also called together as “perovskite database”.

Table 3.1. Used Semiconducting Elements and Their Mole Percent in Perovskites.

PARAMETER	INPUT VARIABLES	RANGE (%mole)	NUMBER OF DATA
Semiconducting Materials (A and B)	**Ta**	0 - 0.767	1334
	Na	0 - 0.14	773
	Zn	0 - 0.845	653
	Nb	0 - 0.616	557
	Sr	0 - 0.559	498
	Ti	0 - 0.517	476
	La	0 - 0.564	473
	H	0 - 0.003	452
	Cd	0 - 0.778	341
	In	0 - 0.663	336
	Sn	0 - 0.467	173
	Ca	0 - 0.224	172
	K	0 - 0.159	162
	Al	0 - 0.289	155
	Bi	0 - 0.592	142
	Sb	0 - 0.717	137
	Ba	0 - 0.289	129
	Y	0 - 0.261	121
	Ni	0 - 0.86	111
	Yb	0 - 0.457	90
	Li	0 - 0.029	75
	Ce	0 - 0.357	70
	N	0 - 0.13	68
	Zr	0 - 0.509	58
	Cu	0 - 0.128	56
Semiconducting Materials (X_3)	**O**	0-0.536	1936
	S	0-0.379	787

Preferred semiconducting material is the major factor that affects activity directly. In Table 3.1 elements used in perovskite type photocatalysts as A, B and as X_3, and their mole percentages in one molecule of perovskite are given. Among included perovskites in database, 1936 of them are ABO_3 type while 787 of them contain S instead of O (ABS_3). Tantalum, niobium and strontium and titanium are the most preferred elements for B site of perovskites. Since titanium database includes only titanium as photocatalyst material, there is no need to give such table for that database.

Catalyst preparation method is as important as semiconducting materials in terms of their effects on photocatalytic activity. Table 3.2 and 3.3 shows the method used to prepare promoted or bare titanium based photocatalysts in articles.

Table 3.2. Preparation Methods of Titanium-Based Catalysts without Promoter.

PARAMETER	**INPUT VARIABLES**	**NUMBER OF DATA**
Titanium-Based Catalyst Preparation Methods	Sol-Gel Method	774
	Solvo-Thermal Route	231
	Deposition Precipitation	201
	Novel	61
	RF Magnetron Sputtering	23
	Commercial	17

Sol-gel method and impregnation are the most common ways to produce titanium-based photocatalyst and promote co-catalyst on it, respectively. There are some methods which are presence in both table, it means the related method can be suitable to synthesize bare or co-catalyzed materials.

Perovskite preparation methods with or without promoter are tabulated in Table 3.4 and Table 3.5, respectively. Solid state reaction and photo deposition are the most common methods for this case.

Table 3.3. Preparation Methods of Titanium-Based Catalysts with Promoter.

PARAMETER	INPUT VARIABLES	NUMBER OF DATA
Catalyst (with Promoter) Preparation Methods	Impregnation	974
	Photo-Deposition	330
	Citrate Method	248
	Solvo-Thermal Route	193
	Hydrothermal Method	175
	Sol-Gel Method	174
	Electro-Spinning Method	145
	Alcohol-Thermal Method	143
	Ion Exchange Method	90
	Ionic Liquid Template Approach	25
	Commercial	21

Table 3.4 Preparation Methods of Perovskite-Type Catalysts without Promoter.

PARAMETER	INPUT VARIABLES	NUMBER OF DATA
Perovskite-Type Catalyst Preparation Methods	Solid State Reaction	1044
	Hydrothermal	460
	Ion-Exchange	445
	Co-Precipitation	191
	Polymerizable Complex	174
	Sol-Gel	138
	Thermal Sulfuration	110
	Flux Synthesis	101
	Impregnation	75
	Solvo-Combustion	72
	Cirtic Acid Complex	51
	Precipitation	50
	Solvo-Thermal Route	16

Table 3.5. Preparation Methods of Perovskite-Type Catalysts with Promoter.

PARAMETER	INPUT VARIABLES	NUMBER OF DATA
Catalyst (with promoter) Preparation Methods	Photo deposition	869
	Dry Impregnation	292
	Photo reduction	155
	Stepwise Intercalation	123
	Incipient to Wetness Impregnation	80
	Wet Impregnation	54
	Deposition-Precipitation	15
	Ion-Exchange	13

Catalyst preparation process also includes heating, calcination, drying, oxidation, and reduction steps (thermal treatment process). The duration of these steps or temperature level used have great impacts on catalyst properties such as particle size, band gap, or surface area of the material,; hence those steps indirectly but considerably affects hydrogen production in PWS. Table 3.6 and Table 3.7 show the range of duration and temperatures of these steps in database. In tables “0 Kelvin” or “0 hour” stands for the cases that calcination, drying, and reduction were not performed. Also, specific to calcination, oxygen content is also considered in database as it is shown in tables.

Table 3.6. Temperature and Time Ranges of Treatment Processes in Titanium Database.

THERMAL TREATMENT PROCESS	INPUT VARIABLES	RANGE
Heating	Temperature (K)	0 - 453
	Time (h)	0 - 168
Calcination	Temperature (K)	0 - 1173
	Oxygen Content (%)	0 - 21
	Time (h)	0 - 12
Drying	Temperature (K)	0 - 383
	Time (h)	0 - 168
Reduction	Temperature (K)	0 - 623
	Time (h)	0 - 1
Oxidation	Temperature (K)	0 - 623

Table 3.7. Temperature and Time Ranges of Treatment Processes in Perovskite Database.

THERMAL TREATMENT PROCESS	INPUT VARIABLES	RANGE
Heating	Temperature (K)	0 - 1373
	Time (h)	0 - 24
Calcination	Temperature (K)	0 - 1873
	Oxygen Content (%)	0 - 100
	Time (h)	0 - 12
Drying	Temperature (K)	0 - 383
	Time (h)	0 - 24
Reduction	Temperature (K)	0 - 773
	Time (h)	0 - 2
Oxidation	Temperature (K)	0 - 623
	Time (h)	0 - 1

Among catalyst thermal treatment steps, calcination is the most important process for both titanium and perovskite based photocatalysts. The reason of this was discussed in Results and Discussion chapter. The detailed information about calcination temperatures are given in Figure 3.4 and 3.5 for titanium and perovskite database, respectively.

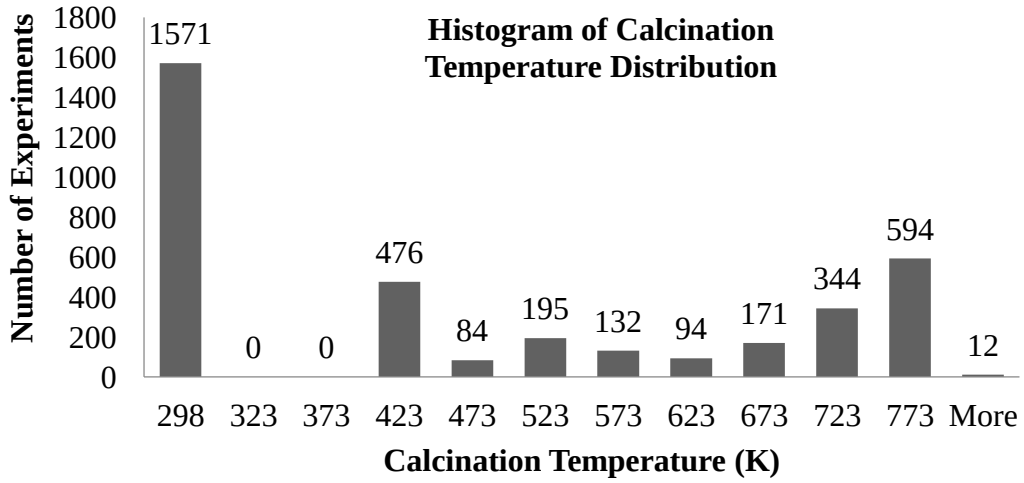

Figure 3.4. Distribution of Calcination Temperature for Titanium Database.

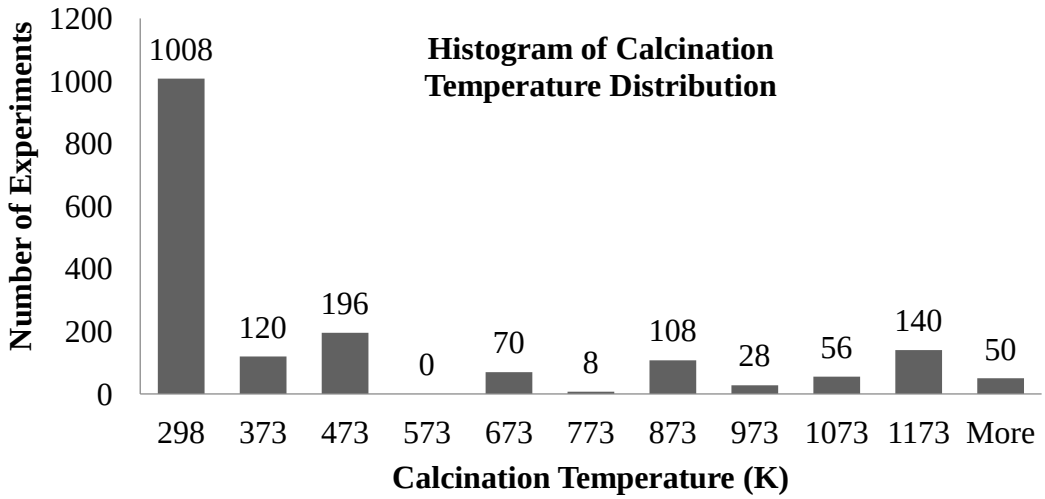

Figure 3.5. Distribution of Calcination Temperature for Perovskite Database.

Using co-catalyst materials with photocatalyst in PWS usually change the activity of reaction favorably (sometimes adversely). Therefore promoter type and amount used in the database can be seen from Table 3.10 and 3.11 in titanium and perovskite catalyst, respectively. “Pt” is at the top both of the lists and its range is wider than the others.

Table 3.8. Promoter Type and Amounts Involved in Titanium Database.

VARIABLE	INPUT VARIABLES	RANGE (% Weight)	NUMBER OF DATA
Promoter Type and Amount	Pt	0 - 14	1413
	Fe	0 - 5	368
	Ni	0 - 5	243
	Au	0 - 5	189
	C-Based	0 - 10	149
	Ag-Based	0 - 4.5	132
	Co	0 - 1.8	121
	Ir	0 - 1.8	76
	N-Based	0 - 1	35

The physical and optical properties of catalyst can be considered as another major group of variables and PWS activity can be also improved easily by manipulating these variables. Crystalline structure and band gap gains importance especially in PWS because of their direct relation with light absorption. Surface area, particle size and amount of

catalyst are also influential. Catalyst properties and crystal structures of titanium and perovskite database are given in Table 3.10-11-12-13, respectively.

Table 3.9. Promoter Type and Amounts Involved in Perovskite Database.

VARIABLE	INPUT VARIABLES	RANGE (% Weight)	NUMBER OF DATA
Promoter Type and Amount	Pt	0 - 8	958
	NiOx-Based	0 - 5.02	189
	Ru-based	0 - 2	182
	Rh	0 - 0.005	53
	Au	0 - 0.77	21

The physical and optical properties of catalyst can be considered as another major group of variables and PWS activity can be also improved easily by manipulating these variables. Crystalline structure and band gap gains importance especially in PWS because of their direct relation with light absorption. Surface area, particle size and amount of catalyst are also influential. Catalyst properties and crystal structures of titanium and perovskite database are given in Table 3.10-11-12-13, respectively.

It was realized that X-Ray Diffraction (XRD) results were also given in the reviewed articles, regularly. Different peaks read from XRD analysis represent a particular crystal structure. The collected XRD results from articles which are included in database are given in Table 3.14 and 3.15 for titanium and perovskite database, respectively.

Table 3.10. Catalytic Properties of Experiments in Titanium Database.

VARIABLE	INPUT VARIABLES	RANGE
Catalyst Properties	Band Gap (eV)	1.28 - 3.51
	Crystallite Size (nm)	1.8 - 180
	Surface Area (m^2/g)	2.2 - 569
	Catalyst Weight (g/L)	0.0042 - 1.67

Table 3.11. Crystal Structures of Catalysts in Titanium Database.

VARIABLE	INPUT VARIABLES	NUMBER OF DATA
Crystal Structure	Anatase	1844
	P25	1525
	Orthorombic	29
	Rutile	29

Table 3.12. Catalytic Properties of Experiments in Perovskite Database.

VARIABLE	INPUT VARIABLES	RANGE
Catalyst Properties	Band Gap (eV)	1.27 - 4.98
	Crystallite Size (nm)	3 - 2500
	Surface Area (m^2/g)	0.36 - 158.82
	Catalyst Weight (g/L)	0.2 - 4.08

Table 3.13. Crystal Structures of Catalysts in Perovskite Database.

VARIABLE	INPUT VARIABLES	NUMBER OF DATA
Crystal Structure	Orthorombic	864
	Cubic	683
	Layered-Perovskite	281
	Tetragonal	184
	Hexagonal	171
	Octahedral	131
	Rhombohedral	123
	Monoclinic	43
	Anatase	43

Since the allowable minimum value for band gap is defined in PWS thermodynamically and the increase in upper limit causes deviation from visible region, the range for band gap is quite narrow. Interval for other variables, on the other hand, displays a wider array. The most common crystal structure of titanium is anatase; for example, P25, which is a commercial product developed by Degussa Corporation, is one of the most commonly used material for this purpose and it includes 20% of rutile phase and 80% of anatase phase of titanium. There is larger spectrum of crystal structures for perovskite-type photocatalysts. It is logical since the motivation behind the fabrication of

Table 3.14 XRD Results of Photocatalysts Included in Titanium Database.

OTHER INPUT VARIABLES		RANGE
XRD Results	(101)	2418
	(200)	2190
	(004)	1973
	(105)	1261
	(211)	788
	(110)	759
	(220)	467
	(210)	145
	(130)	144
	(111)	70
	(002)	53
	(112)	25
	(103)	19

Table 3.15. XRD Results of Photocatalysts Included in Titanium Database.

OTHER INPUT VARIABLES		RANGE
XRD Results	(220)	1036
	(110)	550
	(020)	545
	(111)	544
	(002)	531
	(200)	367
	(400)	346
	(004)	305
	(101)	278
	(206)	243
	(222)	207
	(112)	202
	(100)	196
	(202)	165
	(001)	79
	(211)	68
	(115)	65
	(105)	64
	(129)	51
	(013)	24

different perovskites is to make them suitable for visible light harvesting by changing their crystal structures and therefore, their band gap values.

In published articles, using alcohol as sacrificial agent in reaction solution is reported as a promising way to suppress recombination rate. Most of these articles used aqueous solutions of ethanol and methanol in different compositions; Table 3.16 and 3.17 shows used type of alcohol and their volume percent in solution; methanol is obviously much more commonly used than ethanol.

Table 3.16. Ingredients of Reaction Solutions in Titanium Database.

VARIABLE	INPUT VARIABLES	RANGE	NUMBER OF DATA
Ingredients of Reaction Solution	Water (%v/v)	0 - 100	3660
	Methanol (%v/v)	0 - 100	3026
	Ethanol (%v/v)	0 - 72	464

Table 3.17. Ingredients of Reaction Solutions in Perovskite Database.

VARIABLE	INPUT VARIABLES	RANGE	NUMBER OF DATA
Ingredients of Reaction Solution	Water (%v/v)	20 - 100	2886
	Methanol (%v/v)	0 - 80	1836
	Ethanol (%v/v)	0 - 80	9

Light properties such as type, power, wavelength and intensity can be also considered as variables of PWS. Table 3.18 and 19 gives the tabulated forms of light properties in titanium and perovskite database, respectively. Hg lamps are generally used for UV-light illumination, xenon and halogen lamps are preferred as visible light source, and metal halide lamp can be used for either visible or UV illumination depending its wavelength. The most preferred type is Hg lamps among other in both databases. It derives probably from UV light illumination of those lamps which increases PWS activity.

Finally the range of time on stream of experiments (time passed until the data taken from the beginning of the experiment) in titanium and perovskite database is shown in Table 3.20 together as the last input variable that used in the database. It should be noted

Table 3.18. Properties of Light Used in Titanium Database.

VARIABLE	INPUT VARIABLES		NUMBER OF DATA
Light Properties	Light Type	Hg Lamp	1354
		Xenon Lamp	1086
		Metal Halide Lamp	837
		UV Light	406
	Physical Properties		**RANGE**
		Power (W)	2 - 1000
		Wavelenght (>nm)	254 - 500
		Light Intensity (mW/cm^2)	2.2 - 103.9

Table 3.19. Properties of Light Used in Perovskite Database.

VARIABLE	INPUT VARIABLES		NUMBER OF DATA
Light Properties	Light Type	Hg Lamp	1618
		Xenon Lamp	809
		Halogen Lamp	254
		UV Light	193
		Metal Halide Lamp	24
	Physical Properties		**RANGE**
		Power (W)	8-1000
		Wavelength (>nm)	200-430
		Light Intensity (mW/cm^2)	0.926-150

that this variable is somewhat different than the others; no other characteristics of the system changes with time on stream, hence building a model with data belonging some values of time on stream and verifying the model with the data at other time on stream values will overestimate the predictive power of the model.

Table 3.20. Time on Stream Ranges for Titanium and Perovskite Databases.

INPUT VARIABLES	RANGE
Time on Steam (Titanium Database)	0 - 150
Time on Steam (Perovskite Database)	0.18 - 32.5

The output variable in both databases is cumulative hydrogen production, and Table 3.21 and 3.22 gives the range of this variable included in databases.

Table 3.21. Cumulative Hydrogen Production as Output Variable in Titanium Database.

OUTPUT VARIABLE	RANGE
Cumulative Hydrogen Production (μmol/g-cat)	0 - 83700

Table 3.22. Cumulative Hydrogen Production as Output Variable in Perovskite Database.

OUTPUT VARIABLE	RANGE
Cumulative Hydrogen Production (μmol/g-cat)	0 - 155000

3.2. Modeling

The computational work was performed by using R, which is a free software environment for statistical computing and graphics. Various data mining algorithms and supplementary tools were used, various R codes were written for this work and they altered several times to reach best results. For instance, the consistency of predictions was tested by comparing random forest and neural network outcomes.

3.2.1. Preprocessing Data Set

The constructed database was not initially suitable for data mining; it had incomplete, inconsistent and noisy data. A great effort was required to clean and harmonize the data as the "pre-processing step".

First of all, articles with unsuitable variables were eliminated from database. For example, there are various techniques to prepare photocatalyst and some of them are unique to related article. Since each method is included in dataset as a column, those unique articles increase the number of column with a small contribution to number of rows) and it causes to degrees of freedom problem; the contribution of those articles to the model's learning was minor compare to additive effect to model complexity and data sparsity.

Catalyst treatment processes may vary among catalytic materials; some of them do not require any calcination, reduction or oxidation step. Inserting zero into the dataset to indicate that those experiments were untreated creates a problem. For example, through a calcination temperature column, if one researcher performed calcination under 973 K, then the corresponding cell was named as 973. Besides, in the case of no calcination, if a zero is put here, constructed model will perceive it as calcination at 0 K which is impossible, and the range of that parameter becomes too large. After thinking profoundly, it was decided to describe them by inserting 298 K into related cells, since they were exposed to normal conditions in any way even though they were not specially treated.

The most important part of the pre-processing step was to fill the missing values, especially for catalyst properties which have a big impact on photocatalyst activity. The catalyst amount, band gap, surface area and particle size are determined as indispensable variables. Articles which were unable to give catalyst amount and the ones which had missing information in maximum two of remaining three parameters (band gap, surface area, and particle size) were excluded at the beginning.

In titanium and perovskite database the missing values were observed as it is shown in Figure 3.6. Then dataset divided into seven parts as visualized in Figure 3.7. The yellow part represents parameters that affect catalyst properties and they are already completed. Each of other columns stands for band gap, surface area, and particle size; the green parts of each column refer to completed data while the red parts are missing parts. At each time, model was developed by using green parts and yellow parts to predict the missing values of in red parts. During the pre-processing step, neural network algorithm was used to predict missing values. The details of that algorithm are discussed in following sections.

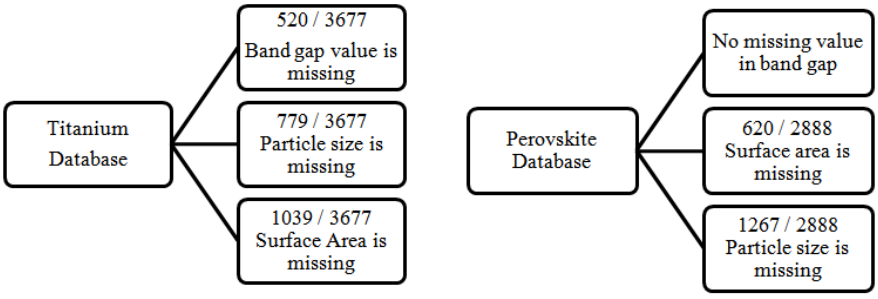

Figure 3.6. Missing Values in Catalyst Properties of Database.

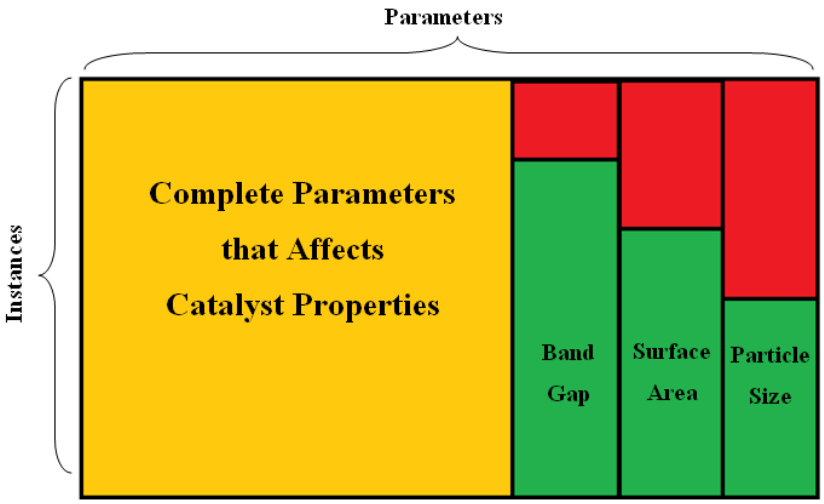

Figure 3.7. Visualization of Divided Dataset to Predict Missing Catalyst Properties.

It is unreasonable to build a model to predict the missing values for some variables such as power or wavelength of light because they are independent. Missing data in those variables were filled with mode of their complete part.

3.2.2. Linear Regression

R offers comprehensive support for multiple linear regression; in this work "lm" function, which is included in STATS package, was used. This was the simplest part of the modeling work. There are some arguments in that function such as *subset* which is an optional vector specifying a subset of observations to be used in the fitting process and *weights*, which refers to an optional vector of weight to be used in the fitting process again. The parameters were tried to be optimized; however it was obvious that dataset had strong non-linear behavior. Therefore in depth analysis using linear regression was regarded as useless and stopped.

3.2.3. Artificial Neural Network Algorithm

The range of variables in database was quite different from each other. For example weight percentage of promoter is likely between 0-10 wt. %, while, the power of light changes in the range of 2-1000 W; this significant difference may create a bias towards one side in mathematical modeling. To prevent this problem, the inputs and output of dataset are standardized separately. RSNNS package of R offers a scaling function which is called

as “normalizeData”. In that function three available options are normalization, standardization and centering data. In this work, the data was normalized to mean zero and variance one (which means standardization). That function uses following Equation 3.1 when processing data.

$$x_{new} = \frac{x - \mu}{\sigma} \tag{3.1}$$

Where x represents each data points to be standardized, μ is for the mean value and σ is standard deviation of these data points. After scaling the data, the neural network model was constructed and trained using k-fold cross validation (CV) method (10-fold and 20-fold CV was applied). To do that, a simple for loop was used in “R”, again. Let’s assume dataset includes 1600 instances and 10-fold CV will be used. By using the formula, which is shown by Equation 3.2, new indexes were determined to separate data.

$$index = \left\{\frac{number\ of\ instances}{number\ of\ folds} * (j^{th} - 1)\right\} + 1 \tag{3.2}$$

$$index = \left\{\frac{1600}{10} * (5 - 1)\right\} + 1 = 641 \qquad index = \left\{\frac{1600}{10} * (6 - 1)\right\} + 1 = 801$$

Where j^{th} represents the fold to be seperated. As it can be seen in the sample calculation given above, index number of 5th fold begins from 641 and continues up to 801. The instances which have indexes between these two numbers are collected as 5th fold. At each time, one fold was reserved as the test set and the remaining folds were used for training. Arguments of functions were tuned by using train set, then output of test set was tried to predict. In that way, the model was unable to see the test set during the construction step improving the reliability of the model.

To create models of neural network, again RSNNS package was used. The function “mlp” creates a multilayer perceptron (MLP) and trains it. In that function, the training was performed by error back propagation (as default). Apart from that, there are other arguments in “mlp” function to be decided. The number of units in the hidden layers, maximum iterations to learn, the activation function of all hidden units, the learning function to be used and type of activation function (the names are *size*, *maxit*, *hiddenActFunc*, *learnFunc, and linOut*, respectively) were tried to be optimized.

Optimizing number of neurons in the hidden layers is a very important step of deciding overall neural network architecture. Number of hidden layer and number of nodes in them have a tremendous effect on the final output. Therefore, hidden layer number was changed from 1 to 5 and number of nodes in hidden layers was changed from 1 to 100 simultaneously. Usually the training error decreases with increasing nodes; the testing error, on the other hand, decreases first but it start to increase after certain number of neuron as the indicator of overlearning (this is the point that one should stop increasing the complexity of the network).

A key feature of neural network is that it is an iterative learning algorithm. During the iterations, network learns the optimal weights, so increasing number of iterations gives algorithm better chance to learn. However there is again a limit for this at which the error rate of test set starts to increase as indicator of overlearning while the error rate of training set remains constant or decreases.

The activation functions were explained in previous chapter. It is very significant to choose correct activation function to reach reliable output values. In that sense, all available hidden layer and learning activation functions were applied one by one. Those functions which are available in R were given in the following Figure 3.8. and 3.9.

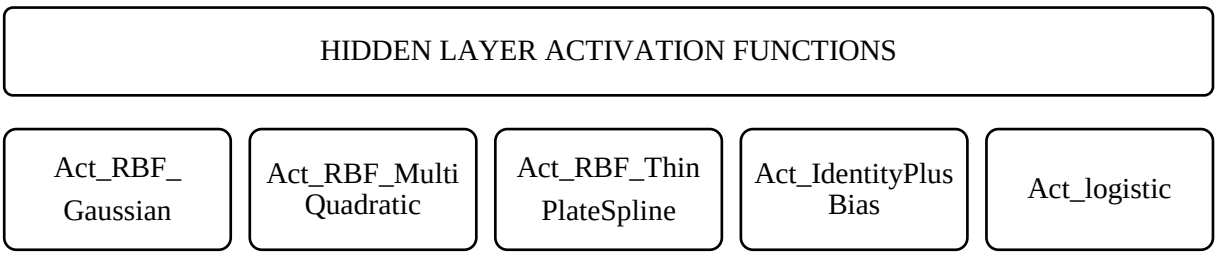

Figure 3.8. Applied Hidden Layer Activation Functions in MLP.

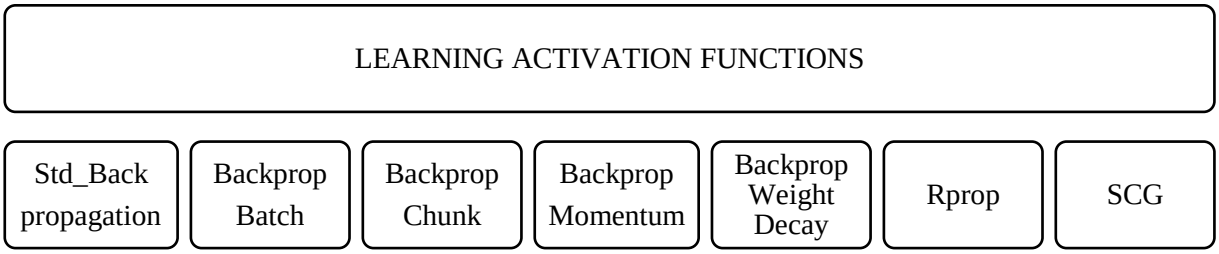

Figure 3.9. Applied Learning Activation Functions in MLP.

There is a supplementary package “NeuralNetTools” in R, and a function in this package, which is called as “garson”, was used to make the input significance analysis. This extracts variable importance measures as produced by model. It can use either mean decrease in accuracy or mean decrease in node impurity as importance measure. By using this function the input significance analysis was performed. First the parameters were sorted in decreasing order according to their relative importance in constructing network. If an attribute was in top 5 of the importance list, it was entitled as the most important parameters, and if an attribute was in bottom 5 of that list, it was named as the least important parameters. By using those returns of the analysis, some variables were excluded and the model was applied once again. Until the model stops improving itself, variables were continued to eliminate.

After deciding the most important variables to model the related data, sensitivity analysis was also performed. The difference of this technique from relative importance analysis is that it gives answer to the question of how much each variable changes the response. There is also a specialized function in R which is called “lek.fun” for that purpose. That function returns with a plot in which x-axis shows the change in an input variable, while y- axis represents the change in response variable. In each plot there is more than one curve in rainbow colors and they stand for incremental changes in other parameters. Generally bluish colors refer to minimum values of other parameters, and reddish colors are maximum values of other parameters. For example, a single blue curve in a single plot shows the trend of output variable while changing the related parameter and keeping constant other parameters at their maximum values. As a default, that function makes the analysis to all variable. However by changing *var.sens* which is an argument of that function, it may look at specific variables and also prevents creating busy plots; this analysis was applied to only most important five variables.

Once the model with optimum parameters was determined, the standardized residual analysis was also performed. Residual analysis extracts the relation, if there is any, between error and an input variable. Standardized residual is the ratio of the difference between the real and predicted value to the standard deviation of the real values. The importance of that analysis can be explained with a die-rolling analogy (Frost, 2012). A series of tosses can be assessed to determine whether there is a randomness or not in the

displaying of numbers. If a number shows much more than others and if this is proven statistically, the reliability of die tossing weakens. As it can be understood from the analogy, at the end of modeling the resultant errors should be uncorrelated with any variables.

As mathematical evaluation of model performance three criteria namely mean absolute error (MAE), root mean squared error (RMSE), and r-squared values were used. The formulas of them are given in Equation 3.3, 3.4, and 3.5.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|p_i - r_i| \tag{3.3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)^2} \tag{3.4}$$

$$R - squared = 1 - \frac{\sum_{i=1}^{n}(p_i - r_i)^2}{\sum_{i=1}^{n}(r_i - \bar{r})^2} \tag{3.5}$$

Where p_i is the predicted value, r_i is the real value, $\bar{r}$ is the mean of real values, and n is the number of instances. As it can be estimated higher r-squared and smaller MAE and RMSE values are preferable than others.

3.2.4. Random Forest

To apply random forest, randomForest package of R was used. Available function has also same name with package, "randomForest". The tuned parameters of that algorithm are *nodesize* and *ntree*. *Nodesize*, represents the allowable minimum size of terminal nodes. If this number sets larger, then it causes smaller grown trees and program gives answers quickly, and vice versa. Default node size for regression trees is 5 and in written code *nodesize* was changed from 1 to 10. Other parameter *ntree* stands for number of tree. Since random forest gets random rows from dataset, *ntree* value should not be set to too small, to ensure that every instance gets modeled more than one time. In algorithm that number was changed from 1 to 1000. However after analyzing the results with different

methods, it was decided to level down the upper limit because growing 1000 trees for the related dataset only cause complexity rather than improvement in model.

To make input significance analysis in randomForest, there is a function called “importance” works same as “garson” function in MLP. By following same steps as in MLP, input significance analysis was done. Also the standardized residual analysis was performed in random forest by applying same steps with MLP. Packages in R correlated with random forest are unable to make sensitivity analysis for most important variables, so it was skipped.

In the evaluation step of random forest, same criteria with MLP which are mean absolute error, root mean squared error and r-squared value, were used.

The general structure for all algorithms followed during this thesis is given in Figure 3.10.

3.2.5. Principal Component Analysis

Both databases include more than 50 variables, to visualize the relationship between them and to observe their relative importances, principal component analysis (PCA) was applied to databases. R provides a simple “promp” function to do that task. The only requirement of that function is variable names written in formula forms. Then using a supplementary “plot” function, both the observations and variables of multivariate databases are represented on the same plot.

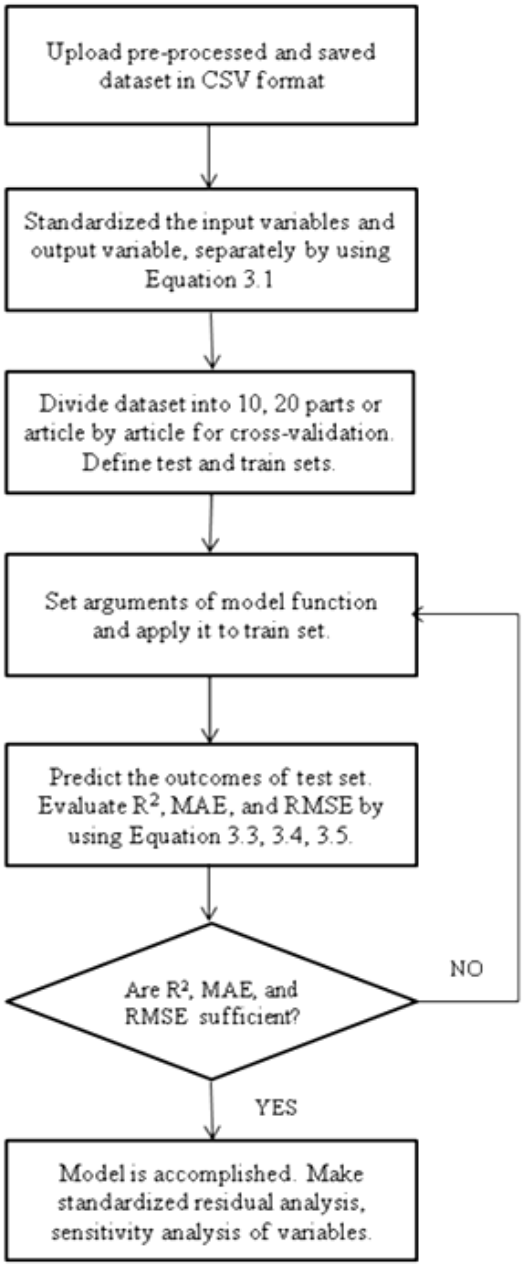

Figure 3.10. General Structures of Followed Algorithms.

4. RESULTS AND DISCUSSION

As it is mentioned in Chapter 3, data preprocessing was performed to increase the quality of data first; filling missing values, smoothing noisy data, removing outliers, normalization and reducing volume of dataset while retaining its content were the objectives of this step. Then the data set was analyzed using three exploratory data analysis methods, which are linear regression, decision tree and neural networks; the principal component analysis was also employed as a supplementary tool.

At first, the output variable was chosen as the cumulative hydrogen production rate, and the data was collected and saved that way. After some preliminary work, however, the results were not satisfactory probably due to the fact that the cumulative output variable had a very wide range from 0 to 150000 μmol/g-cat/h, and even after standardization, the constructed models were unable to make good predictions for extreme values. Then it was decided to change output variable from cumulative hydrogen production to rate of production, which was defined as the total hydrogen production divided by total reaction time. As a result of this new arrangement, the number of instances decreased from 6378 to 541; each experiment should be treated as one instance as far as rate concerned since the rate is almost constant at entire time period.

Another arrangement made after preprocessing was splitting database into further subsets as shown in Figure 4.1. In general sense, TiO_2 photocatalyst with UV light source shows higher performance than those employing visible light in PWS, and this causes a serious gap between average hydrogen production values of those two groups making a single model impossible. Therefore to separate the data accordingly into two groups was considered to be helpful for models to learn the general trends in the data. The perovskite database was also divided into two parts in terms of “X” group (of ABX_3), as it was explained before in Chapter 3.

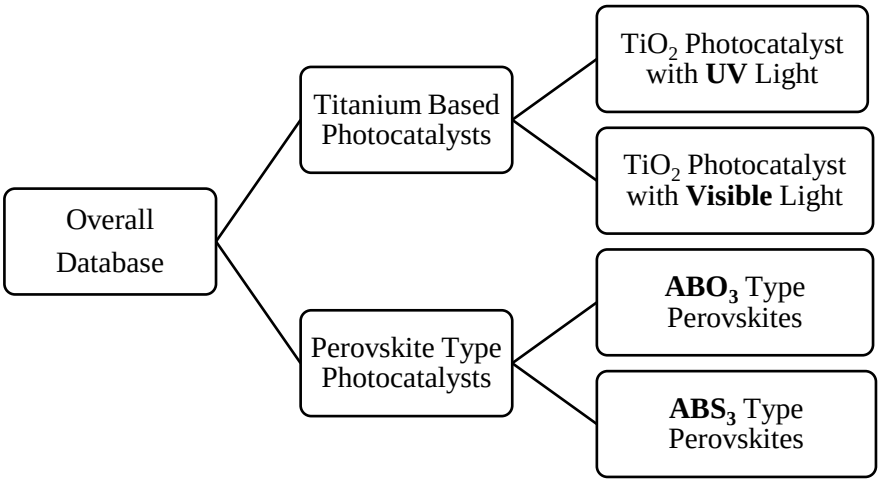

Figure 4.1. Subsets of Overall Database.

4.1. Analysis of Data for TiO_2 Photocatalysts with UV Light Source

Linear regression, neural network and random forest methods were performed for this database, which includes 67 catalytic and 11 operational input variables, and one output variable. However, to improve the models, some input variables were excluded due to their relatively low contribution as indicated by their relative low importance values. As a result, for this dataset 43 catalytic and 7 operational variables were used; this arrangement resulted in 160 instances, which means 160 experiments conducted under different conditions.

4.1.1. Linear Regression

Linear regression modeling was performed by using "lm" function in R, and the predictions of hydrogen production (μmol/gcat/h) with linear regression model are given in Figure 4.2 by comparing them with real output values. 10-fold CV was used to test the model; each color in figure represents a different fold. As it can be seen from the figure, the model was unable to predict the output variable properly and it resulted with constant values for different observations (prediction values fixated at 12000 and 8700). The reason of poor modeling ability of linear regression on this dataset may derive from the complex non-linear relationships between input variables that linear models cannot handle. Hence, no further analysis was conducted using this method; instead the other two methods were utilized.

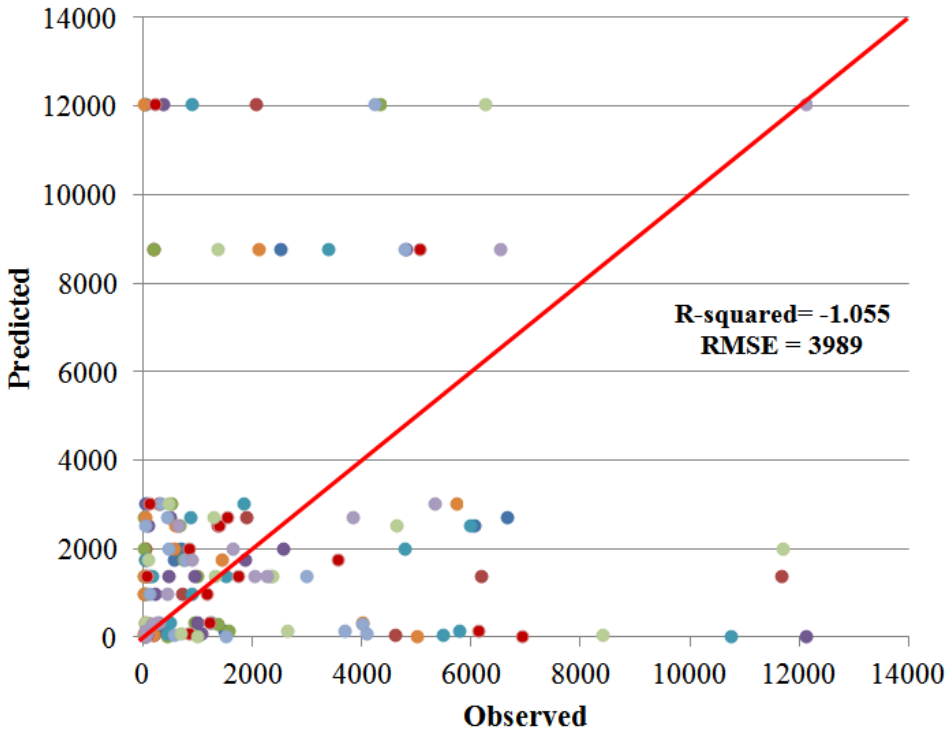

Figure 4.2. Predicted vs. Observed Output Values with Linear Regression.

4.1.2. Artifical Neural Network Modeling

In neural network modeling "mlp" function was applied and 5 different hidden layer and 7 different learning activation functions, which were mentioned in Chapter 3, were used. The number of neurons in hidden layers and number of hidden layers are also changed from 2 to 120 and from 1 to 3, respectively.

The plot of testing error versus the number of neuron is given in Figure 4.3 to see where the over fitting starts, and the global minimum error of 0.3 and minimum rmse of 0.47 were reached when 60 neurons in 1 hidden layer was chosen at which the r-squared value recorded as 0.78. After that point, the error started to elevate with increasing number of neurons; this means the model memorizes the train set and it is unable to predict test set properly after this point. These results were recorded by performing 10-fold CV. Since there are 160 instances, 10-fold CV means 144 instances were used as training set to form the model, and then the remaining 16 instances were used as test set for prediction. Figure 4.3 shows that there was no over fitting until 60 neuron considering that the test set contains the data never seen by the model before. However, one may still use a smaller but

satisfactory number of neurons to be in the safe side. It was found that three neurons in 1 hidden layer resulted in the r-squared value of 0.7 and rmse was 0.55; these results are satisfactory and close enough to the values obtained with 60 neurons so that they can be adopted as the model representing the data set. However, the analyses in the following sections were performed with the 60 neuron network (as the one with minimum testing error) for consistency. Figure 4.4 and 4.5, which give the predicted vs. observed hydrogen productions (μmol/gcat/h) for one hidden layer-three neurons and one hidden layer-60 neurons respectively. Each color appearing on Figure 4.3 and 4.4 represents a different fold.

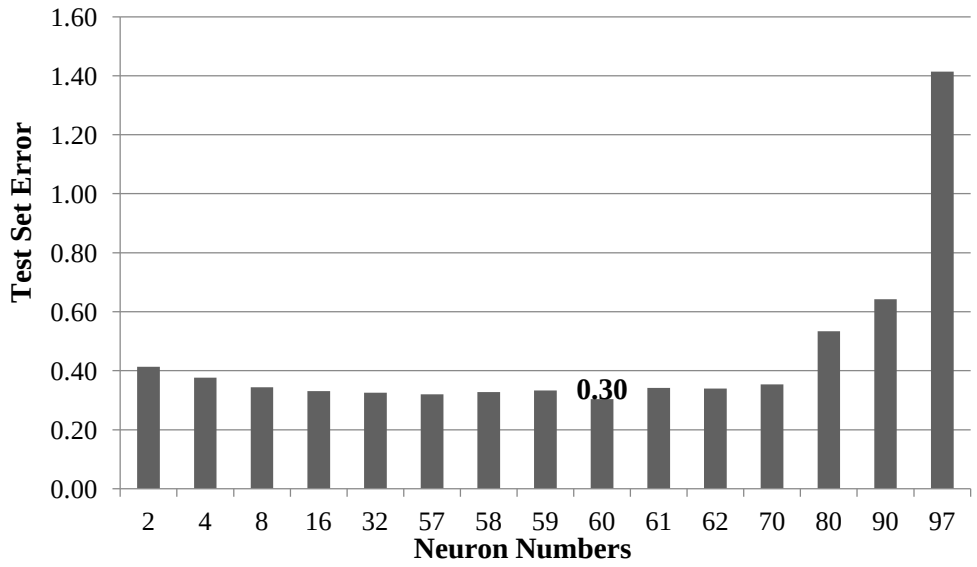

Figure 4.3. Test Set Error vs. Neuron Numbers in 1 Hidden Layer.

The motivation of this thesis was to help the researchers, who conduct experiments on PWS, and try to find a good experimental condition to start. To test the possibility of this, each experiment was excluded from the dataset, and a model was developed with the remaining experiments; then the output of that individual experiment was predicted using the model that did not see that instance. Figure 4.6 shows the predicted vs. observed hydrogen production values (μmol/gcat/h) given by constructed network. Standard error, rmse, and r-squared values were 0.32, 0.5, and 0.74, respectively; as expected they are quite close to the values obtained with 10-fold cross validation.

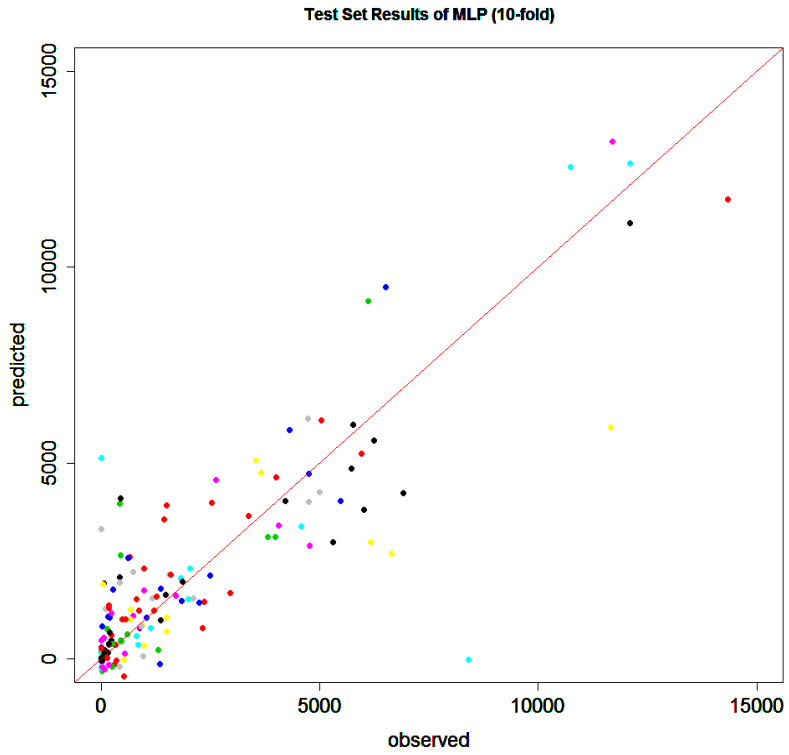

Figure 4.4. Predicted vs. Observed Output Values with NN-modeling (3 Neurons).

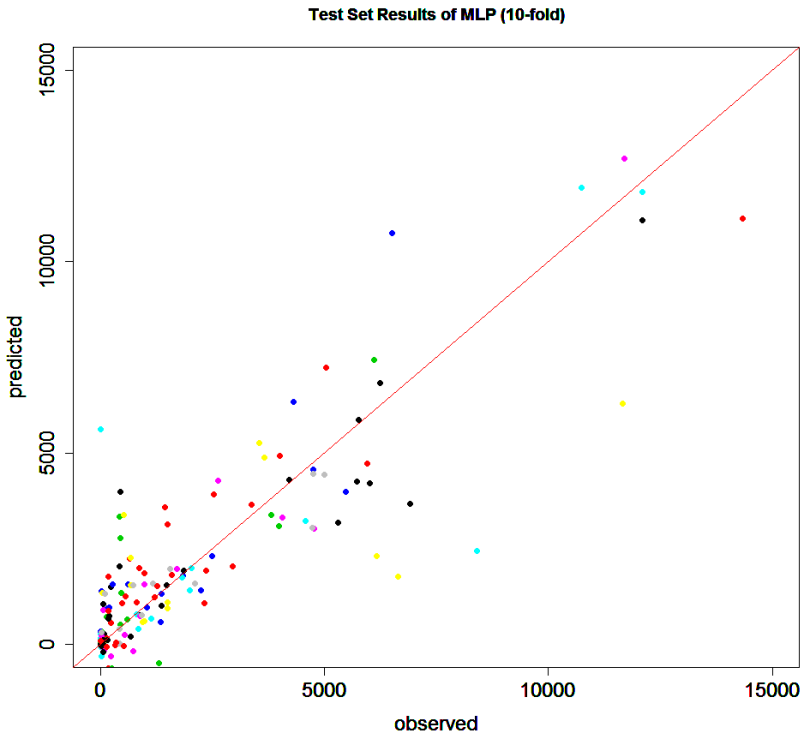

Figure 4.5. Predicted vs. Observed Output Values with NN-modeling (60 Neurons).

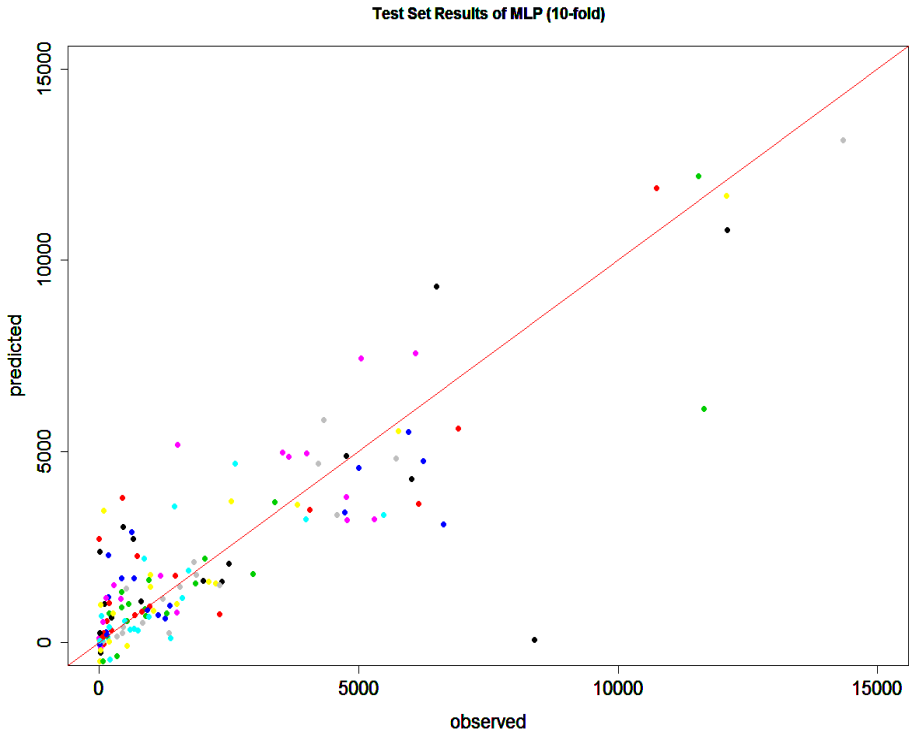

Figure 4.6. Predicted vs Observed Output Values of each Experiment with NN-modeling (60 Neurons).

As it is explained in Chapter 3, the input significance analysis was done by comparing their relative importance of input variables while constructing NN. Figure 4.7 shows the relative importance of input variables with radar map representation. Band gap was determined as the most deterministic variable in NN modeling for the related dataset, and calcination time, Pt loading, and surface area followed the band gap. It should be noted that the meaning of "relative importance" for the categorical variables in the radar chart is not as clear as that of continuous variables. For example, for computational easiness, all catalyst preparation methods were entered the database as separate variables (instead of the alternative values for a single variables as catalyst preparation), and the one used for each data point is labeled as "one" while the others had the value of "zero". If one catalyst preparation method is found to be important while the others are not, the result should be interpreted as follows: whether that method is used or not is important for the results but if that method is not used, it does not matter which other method is used. However, if these methods have varying level of importance, the interpretation of the results may be more complicated. For the continuous variables, on the other hand, the

relative importance simply shows whether the use of that variable in the model make a significant chances in results or not.

As the second step, the sensitivity analysis was done with four important variables determined by input significance analysis. In Figure 4.8, four plots representing Pt loading, band gap, surface area and calcination time are given together. By analyzing Figure 4.8, it can be said that increasing Pt loading favors PWS up to some point; after that Pt loading and production are inversely proportional with each other. The reason for that, excess Pt loading may block reaching light on semiconductor surface and reduces the activity. This result is also consistent with previous experimental studies (Sreethawong and Yoshikawa, 2006). Remaining three parameters have a direct proportion with output variable. Higher band gap values means UV light absorption of semiconductor, so it is logical that increment in band gap positively affects hydrogen production as it was also experimentally proved in literature (Liao *et al.*, 2013). Surface area and calcination time favors the PWS activity as it can be seen from Figure 4.8. and that result is also consistent with literature (Parida *et al.*, 2010, Kokporka *et al.*, 2013).

Standardized residual error analyses for the most important four parameters were also performed. As it can be shown in Figure 4.9, residual errors scatters randomly on plots; it means the related dataset has been modeled properly, and results obtained by this model were reliable. Besides, this analysis assumes that residuals are distributed normally, so about 95% data points should fall within 2σ (σ means standard deviation). Since standard deviation is 1 residuals should be between -2 and +2, remaining ones which are lower than -2 and higher than +2 refer to outliers (Wilcox, 2010). By considering this argument, it can be said that dataset had some outliers even after pre-processing step. Detecting and extracting those outliers from the dataset may improve model prediction ability.

4.1.3. Random Forest

In random forest modeling, randomForest function was used by tuning its parameters. Number of tree (t) was increased from 1 to 250 by 3, and minimum size of terminal nodes (n) changed from 1 to 5, simultaneously.

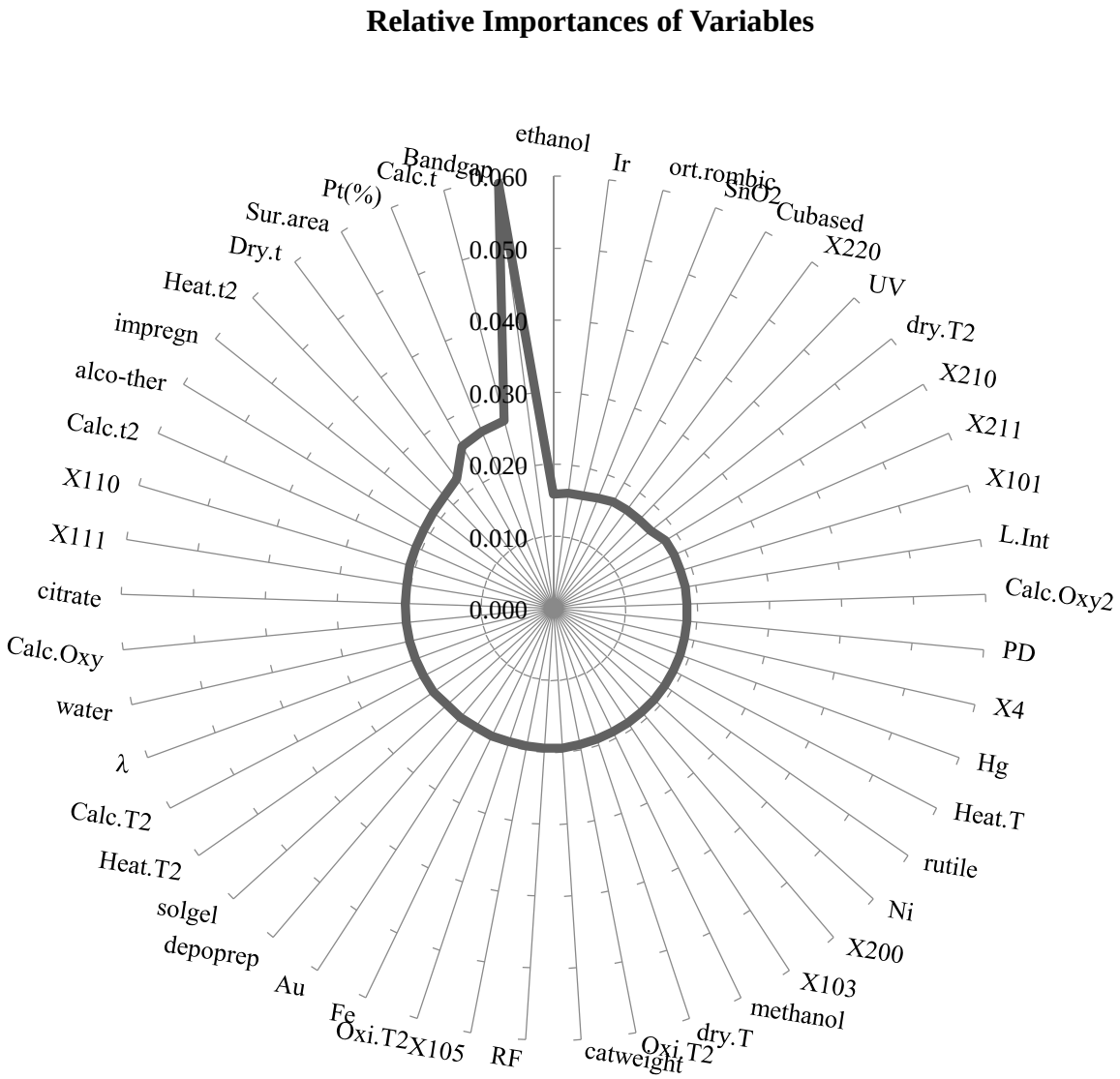

Figure 4.7. Relative Importances of Used Input Variables in NN-Modeling.

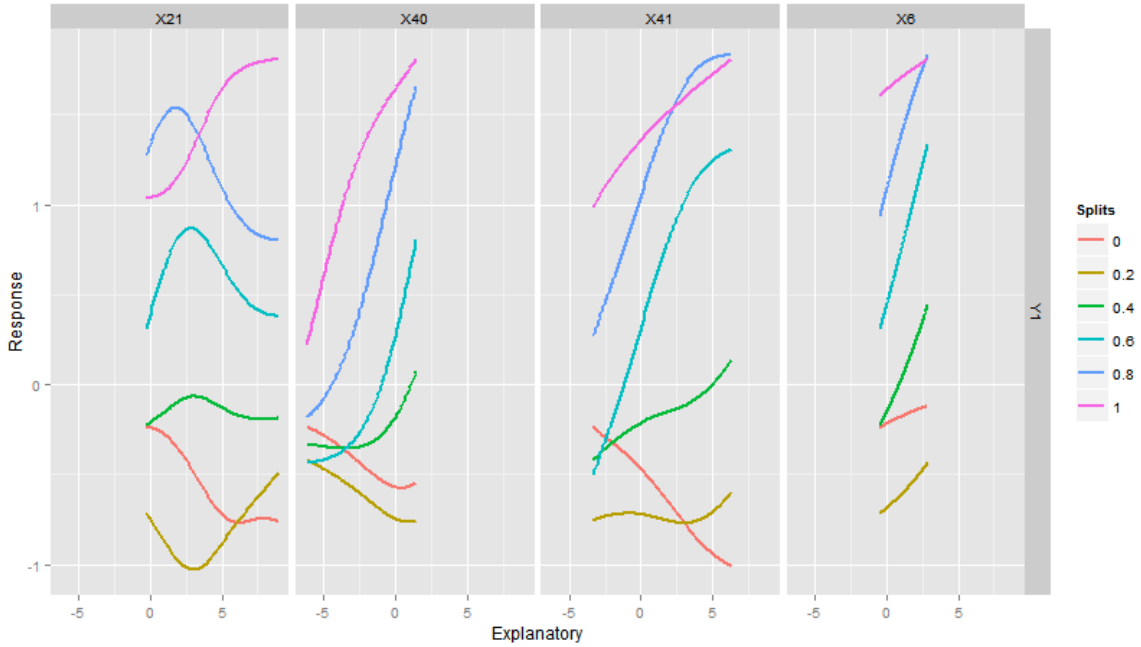

Figure 4.8. Sensitivity Analysis Plot for Four Most Important Variables.

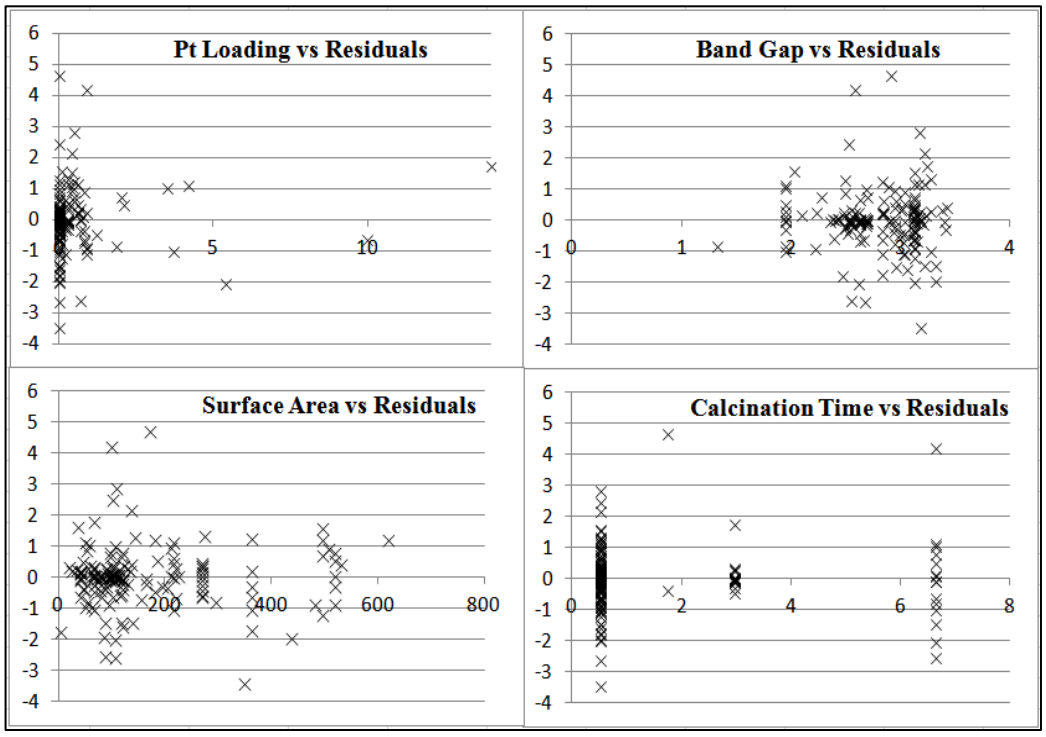

Figure 4.9. Most Important Input Variables vs. Residual Plots.

The following results were recorded by performing 10-fold CV on random forest modeling: The global minimum error of 0.11 and minimum rmse of 0.19 were reached when t equals to 19 and n is 2, and r-squared value recorded as 0.96. It should be also

mentioned that the lowest r-squared value observed in random forest modeling was 0.75 which is very close to the maximum r-squared value of NN-modeling. Therefore, it can be easily said that random forest performed much better than NN-modeling in this dataset. The best performance of that model is given in Figure 4.10; the y and x axes refer to predicted and observed hydrogen production (μmol/gcat/h) values respectively.

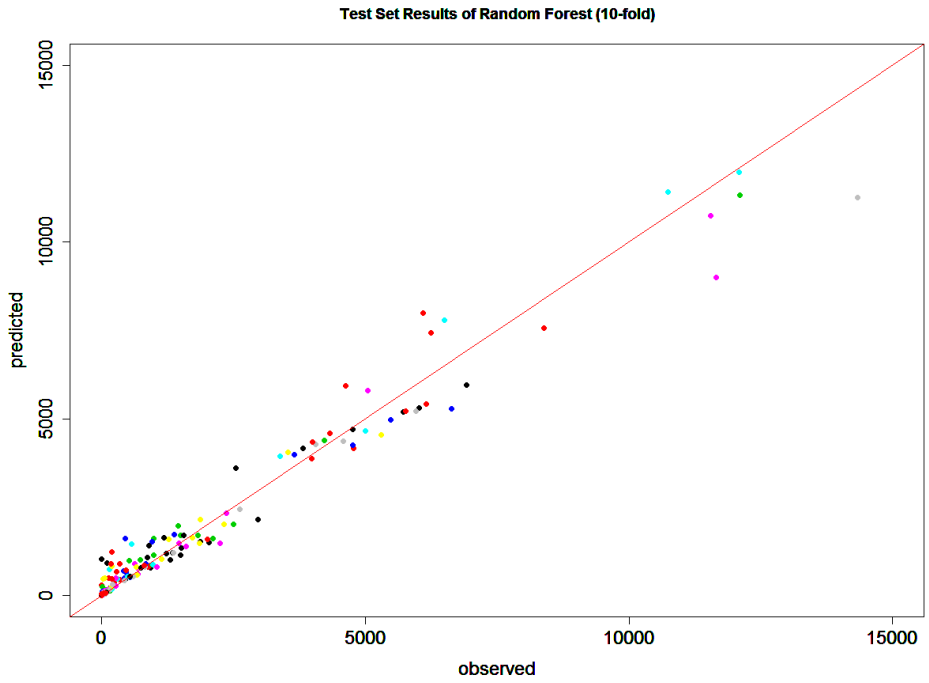

Figure 4.10. Predicted vs. Observed Output Values with Random Forest Modeling (t = 19, n=2).

In random forest algorithm, relative importance of input variables is determined based on two criteria: change in mean squared error and change in node purity. In both methods, surface area, calcination time and Pt loading were in the top 5 of important variable lists and band gap was determined as 5th and 8th most important parameter for two different methods. The results are generally consistent with NN-modeling even though band gap was found to be less significant in this case.

4.1.4. Residual Analysis

Residual analyses of those most important four parameters were done, again. As it is expected, the residual errors scatter randomly on plots Figure 4.11 and it proves the

related dataset has been modeled well and results obtained by this model were reliable, but the outliers were also detected in random forest modeling.

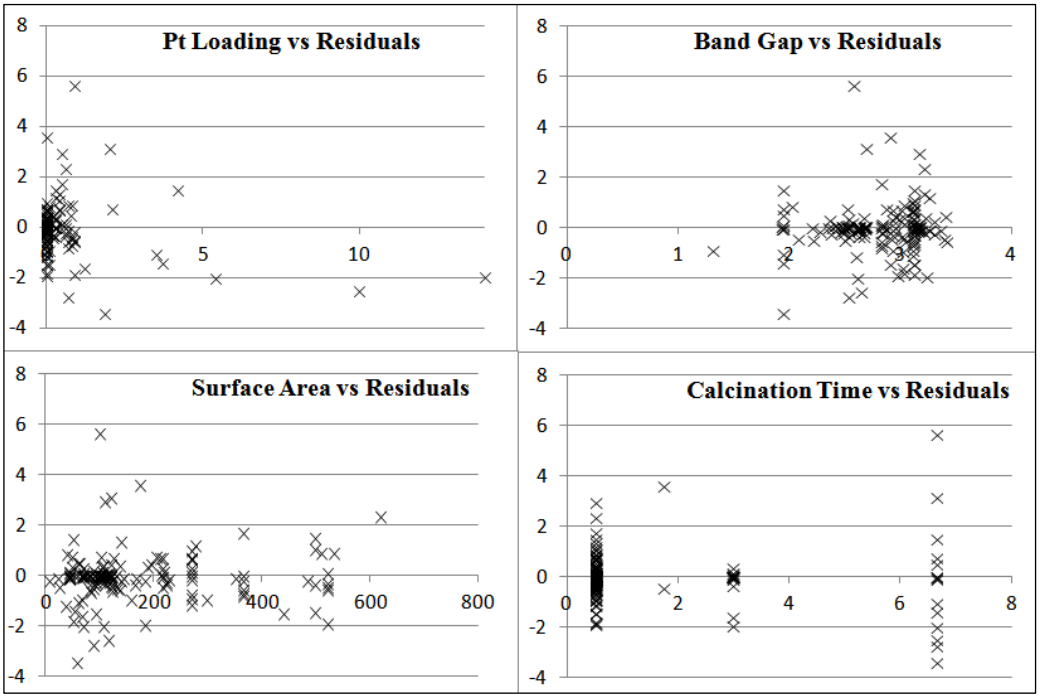

Figure 4.11. Most Important Input Variables vs. Residual Plots.

4.2. Analysis of data for TiO_2 Photocatalyst with Visible Light Source

Linear regression, neural network and random forest methods were also performed for this database. This dataset originally includes 66 catalytic and 12 operational input variables, and one output variable. As in previous dataset, however, some insignificant input parameters were excluded, and 24 catalytic and 3 operational variables were used for the models; the data set comprises of 130 experiments conducted under different conditions.

4.2.1. Linear Regression

Linear regression modeling was performed by using “lm” function again in R, and 10-fold CV was used to test the model. The performance of constructed model is given in Figure 4.12 showing the predicted vs. observed hydrogen productions (μmol/gcat/h). Again model was unable to generalize training data sets and give reliable prediction values for test sets.

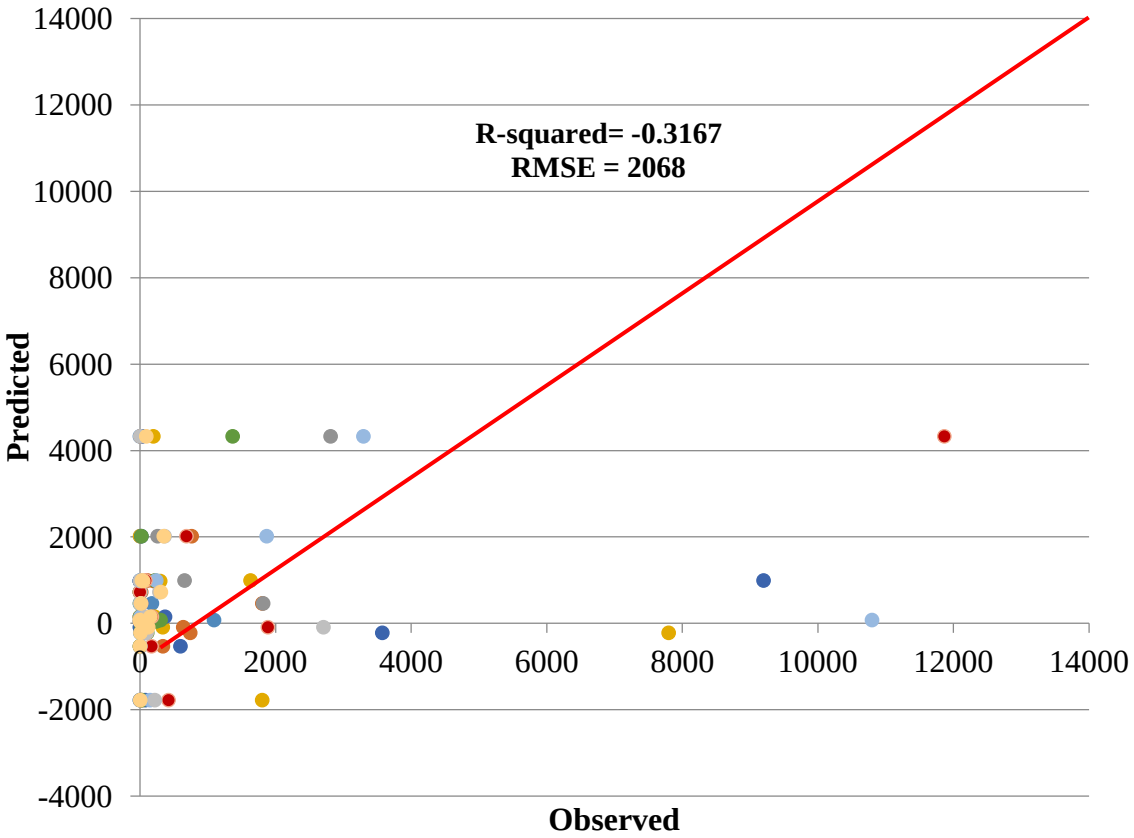

Figure 4.12. Predicted vs. Observed Output Values with Linear Regression.

4.2.2. Artificial Neural Network Modeling

In neural network modeling "mlp" function was applied with 5 different hidden layer and 7 different learning activation functions. The number of neuron in hidden layers and number of hidden layers also changed from 2 to 120 and from 1 to 3, respectively.

The testing error rate vs. neuron number is given in Figure 4.13 to show again where over fitting starts, and the global minimum error 0.2 and minimum rmse 0.33 were reached when 71 neurons in 1 hidden layer was set at which r-squared value recorded as 0.89. The predicted vs. observed hydrogen production (μmol/gcat/h) values at those conditions are shown in Figure 4.14. Again a smaller network could be used to be on the safe if it is required; for example neuron number of 10 gave the the error, rmse and r-squared values of 0.27, 0.49, and 0.76, respectively as it is given in Figure 4.15.

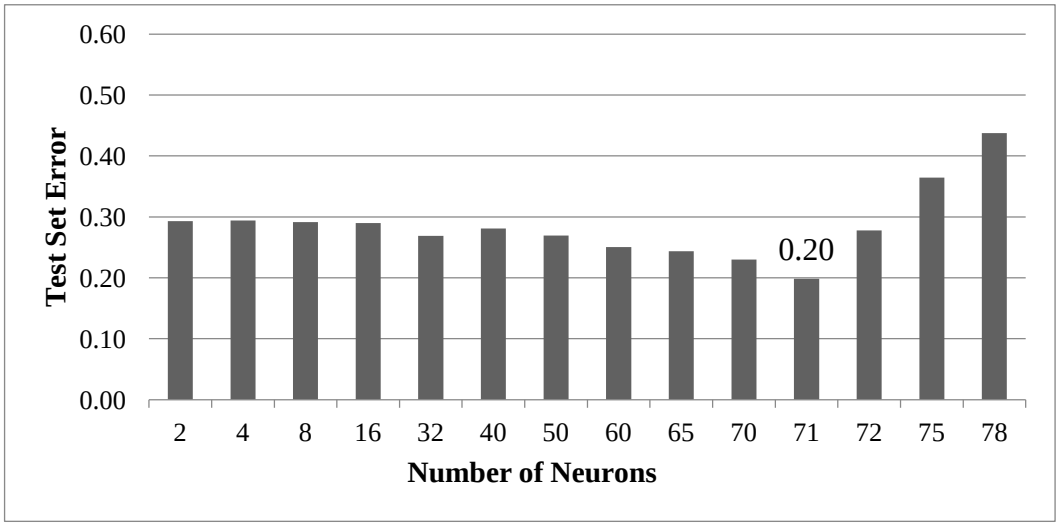

Figure 4.13. Test Set Error vs. Neuron Numbers in 1 Hidden Layer.

With the same object mentioned in previous section, the output value of individual experiments were also predicted by using NN-modeling; Figure 4.16 shows predicted vs. observed hydrogen production (μmol/gcat/h) values; the error, rmse and r-squared value were found as 0.21, 0.43 and 0.81, respectively.

By using the same function and following the same steps, input significance analysis was also performed for this dataset. Figure 4.17 shows the relative importance of all used input variables in model. As it is observed from the figure, there are not large differences in relative importance of top 10 most significant variables. Surface area was dedicated as the most important input variable in NN-modeling for the related dataset.

Heating temperature, photo-deposition method, wavelength of light follows the surface area in descending order. C-based, N-based and Fe promoters, and impregnation methods can be considered as almost ineffective on NN-modeling of that dataset. As explained above, the input significance of categorical variables are actually indicates the importance of that alternative (or level) rather than importance of variable.

By considering the results of input significance analysis, sensitivity analysis was done with four important variables. In Figure 4.18 four plots are given together which represents, photo-deposition method, surface area, wavelength of light and heating temperature, respectively. Although it is found that surface area favors the activity of UV-light driven titanium based photocatalysts, it appears that it has an inverse effects with the

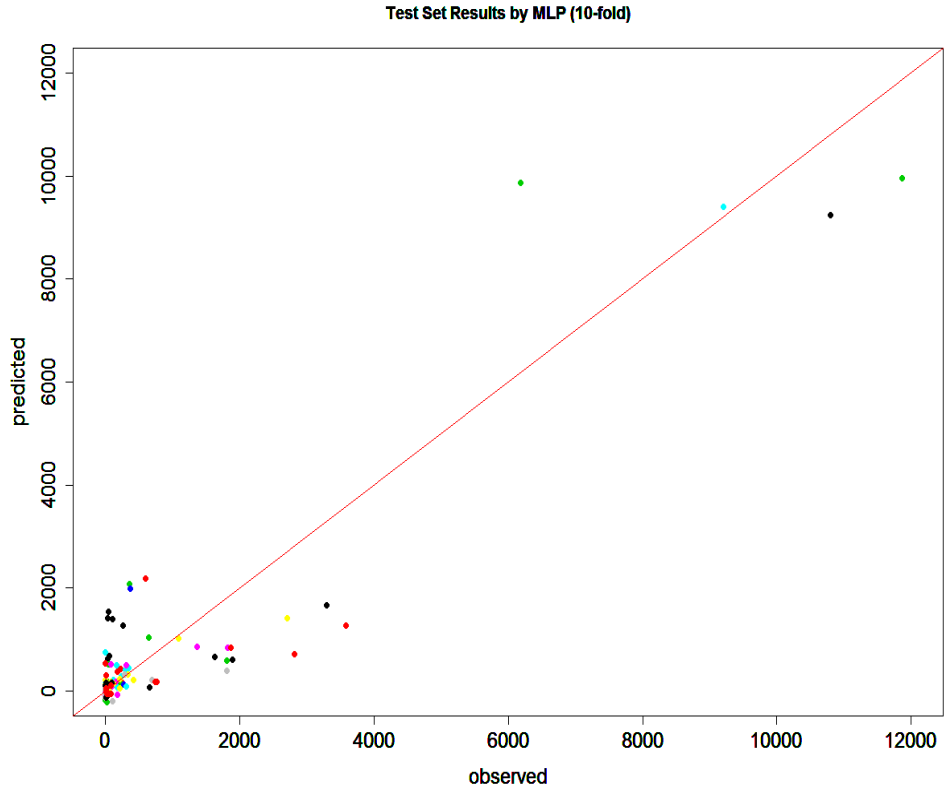

Figure 4.14. Predicted vs. Observed Output Values with NN-modeling (71 Neurons).

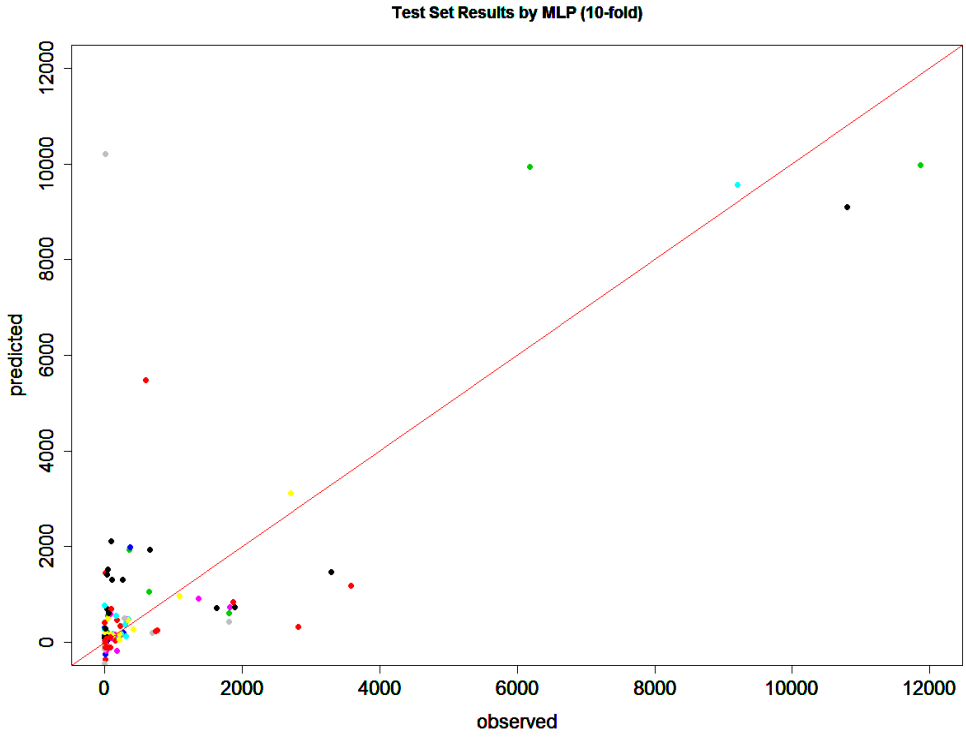

Figure 4.15. Predicted vs. Observed Output Values with NN-modeling (10 Neurons).

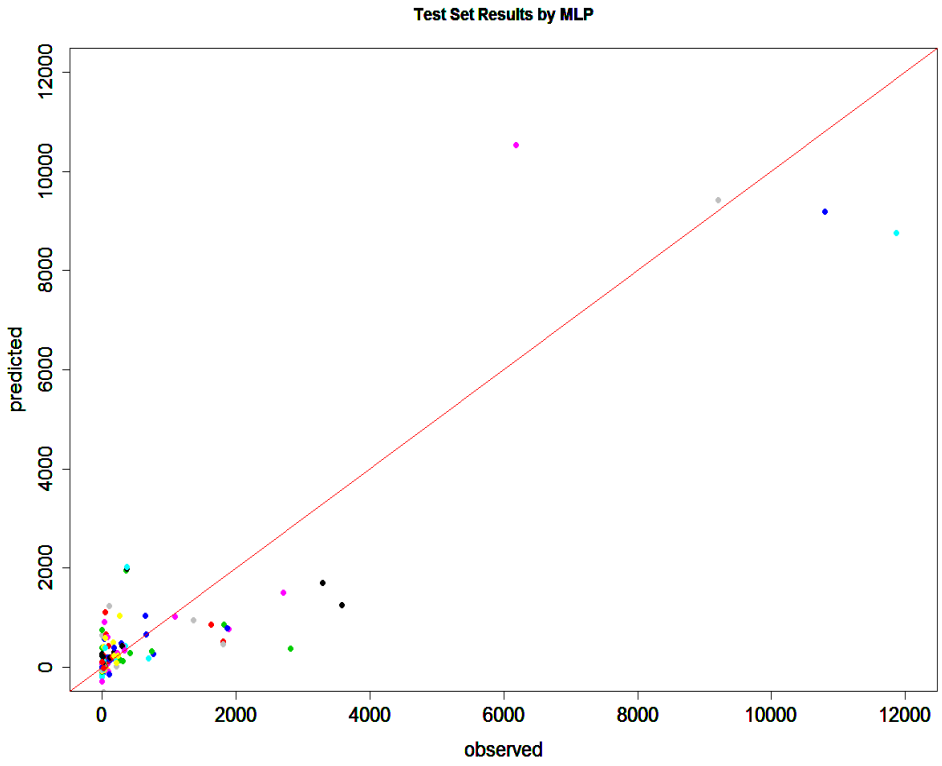

Figure 4.16. Predicted vs. Observed Output Values of each Experiment with NN-modeling (71 Neurons).

visible light driven works. Relevant to this matter, Zhao made PWS experiments by using Xenon light, which can be considered as visible light. In those experiments, titanium based photocatalysts with different surface area values were used and PWS activity failed to increase with increasing surface area (Zhao *et al.*, 2012). According to Figure 4.18, photo deposition always a good way to promote co-catalyst onto semiconductor due to its positive effect on PWS activity. Higher wavelength was also observed to favor the activity while the heating temperature does not have an obvious effect on hydrogen production.

As a last step, standardized residual analysis was also done for this dataset to see whether there is a trend or not in residual errors with respect to most important variables. Figure 4.19 shows the residual analysis vs. photo deposition method, surface area, wavelength of light, and heating temperature. The randomly distributed residuals are clearly seen from surface area vs. residuals plot; for example, in photo-deposition vs. residual plot it is impossible to see a distribution since it is a nominal variable, but to observe outliers is very easy. It can be seen that most of data points fall between -2 and 2 as it is explained previous section, but there are some points which are at -6 and 6 refer to

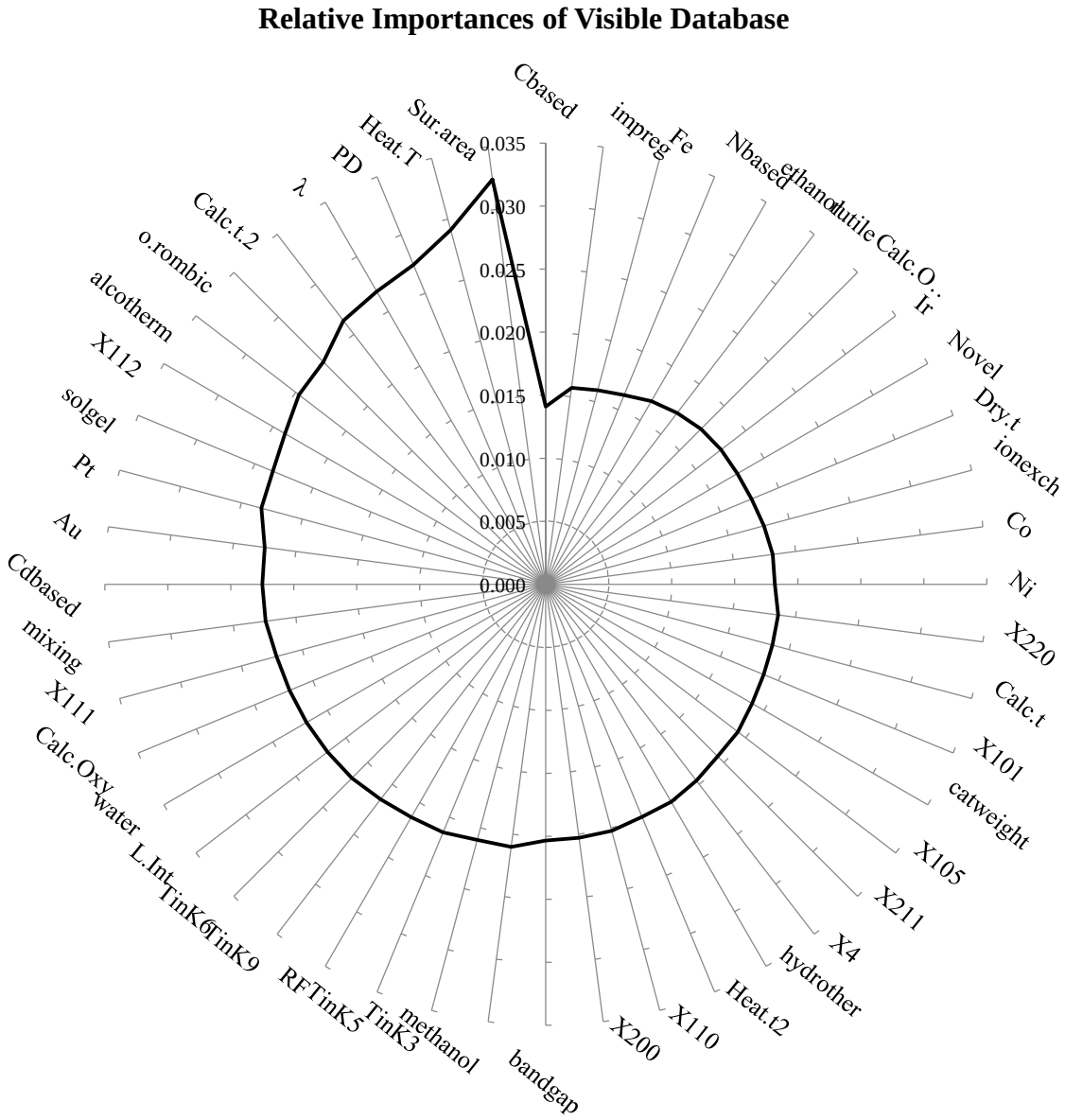

Figure 4.17. Relative Importance of Used Input Variables in NN-Modeling.

the outliers. Also, in surface area vs. residuals plot, the model had a tendency to predict smaller values for smaller surface area since residuals are positive.

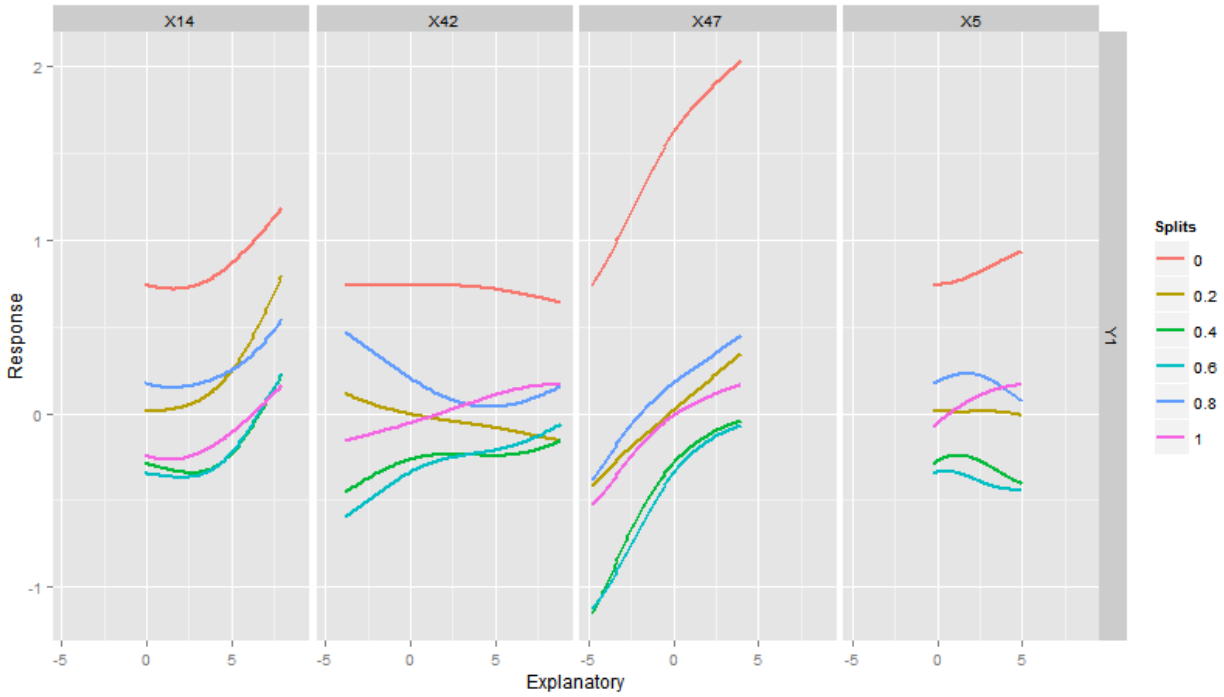

Figure 4.18. Sensitivity Analysis Plot for Four Most Important Variables.

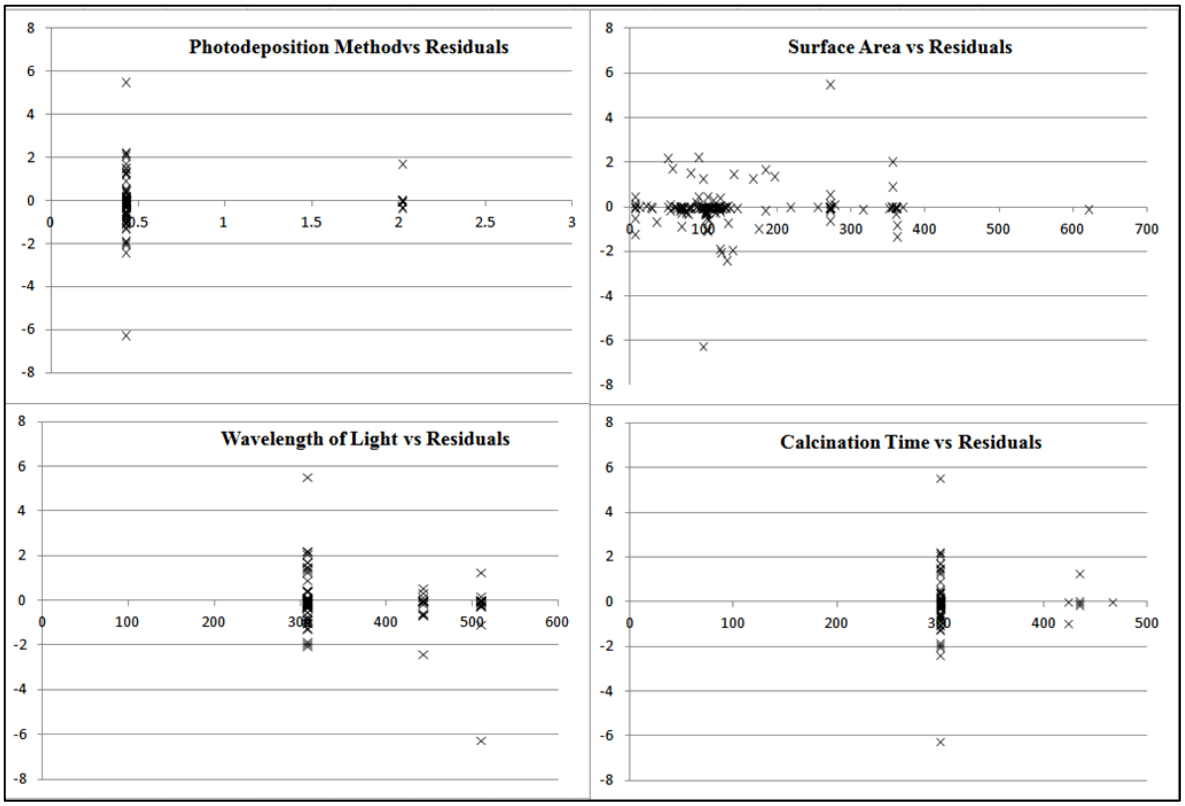

Figure 4.19. Most Important Input Variables vs. Residual Plots.

4.2.3. Random Forest

In random forest modeling, randomForest function was used by tuning its parameters. Number of tree (t) was increased from 1 to 250 by 3, and minimum size of terminal nodes (n) changed from 1 to 5, simultaneously.

The results were recorded by performing 10-fold CV again on random forest modeling. The global minimum error 0.12 and minimum rmse 0.24 were reached when t equals to 7 and n is 3, and r-squared value recorded as 0.94. Generally the r-squared values were so high in random forest modeling of related data and the smallest one was 0.58. Therefore, it can be concluded as the random forest performed better than NN-modeling in that dataset. The best performance of that model is given in Figure 4.20; the y and x axes refer to predicted and observed hydrogen productions (μmol/gcat/h), respectively.

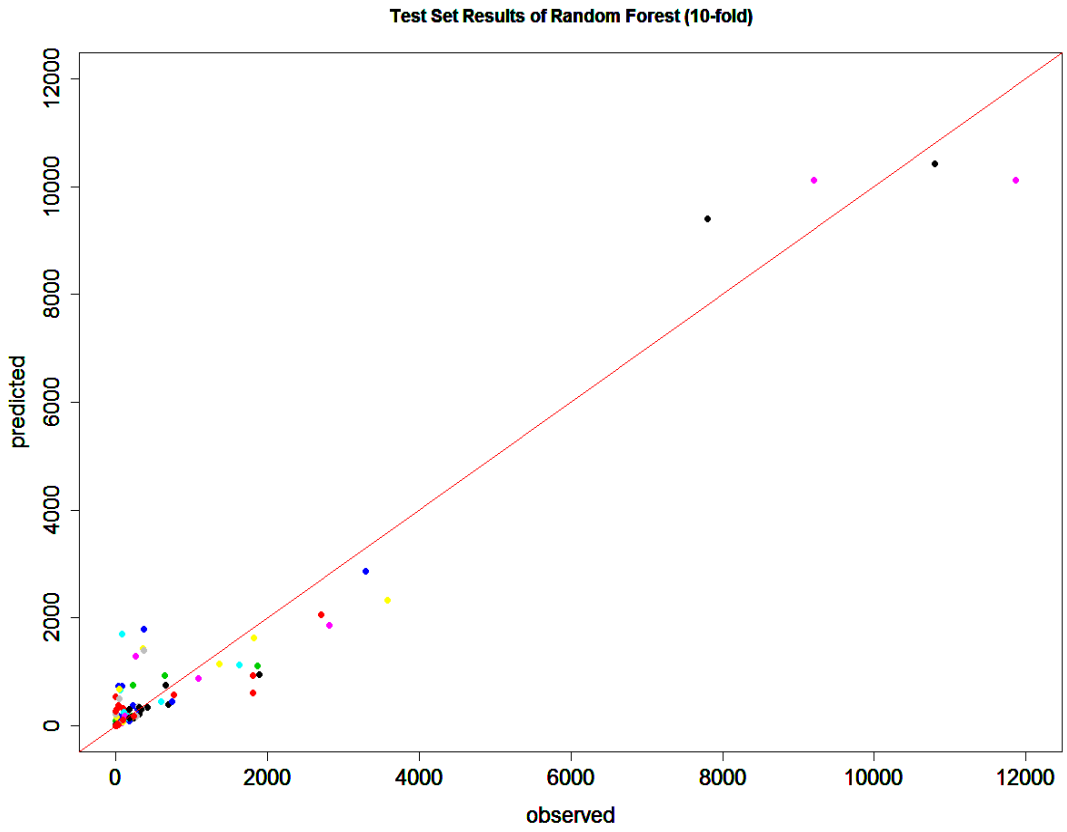

Figure 4.20. Predicted vs. Observed Output Values with Random Forest Modeling (t = 7, n=3).

As it is explained in previous section, there are two criteria for input significance analysis in random forest. If the mean squared errors were considered, hydrothermal method became the most important variable and anatase phase, light intensity, Cd-based promoter, and heating time followed it. If the node impurity took into consideration, top 5 most important variables recorded as hydrothermal method, calcination temperature, light intensity, heating time and power, in descending order. Hydrothermal method is one of the most common ways to prepare titanium-based photocatalysts, high number of instances in that category may create a bias in the model as it was expected. Besides, light intensity is known as more crucial than power, because in general power is used as excess amount but the energy exposures unit area is more important and it is one of the main variables affect PWS activity directly. The results of two criteria used in random forest "importance" function of R are consistent with each other, but results are different than found in neural network modeling.

The difference in input significance analyses of neural network and random forest modeling may derive from the dissimilar working principles of these algorithms.

Residual analyses of those most important four parameters were done, again. As it is expected there is not any trend between the residual errors values and variables in Figure 4.21 and the reliability of models was confirmed by this way. The data points fall between -5 and 5 which means there are certain outlier data points.

4.3. Analysis of Data for ABO_3-type Perovskite Photocatalyst

Linear regression, neural network and random forest methods were performed for this database. However the output value predictions with neural network was not reliable due to its high error and low r-squared values; so the results of NN-modeling are not included in this report. This dataset originally includes 95 catalytic and 11 operational input variables, and one output variable. After reduction in the number of inputs, 61 catalytic and 11 operational variables remained in dataset and they used for modeling. It also comprises of 179 experiments conducted under different conditions.

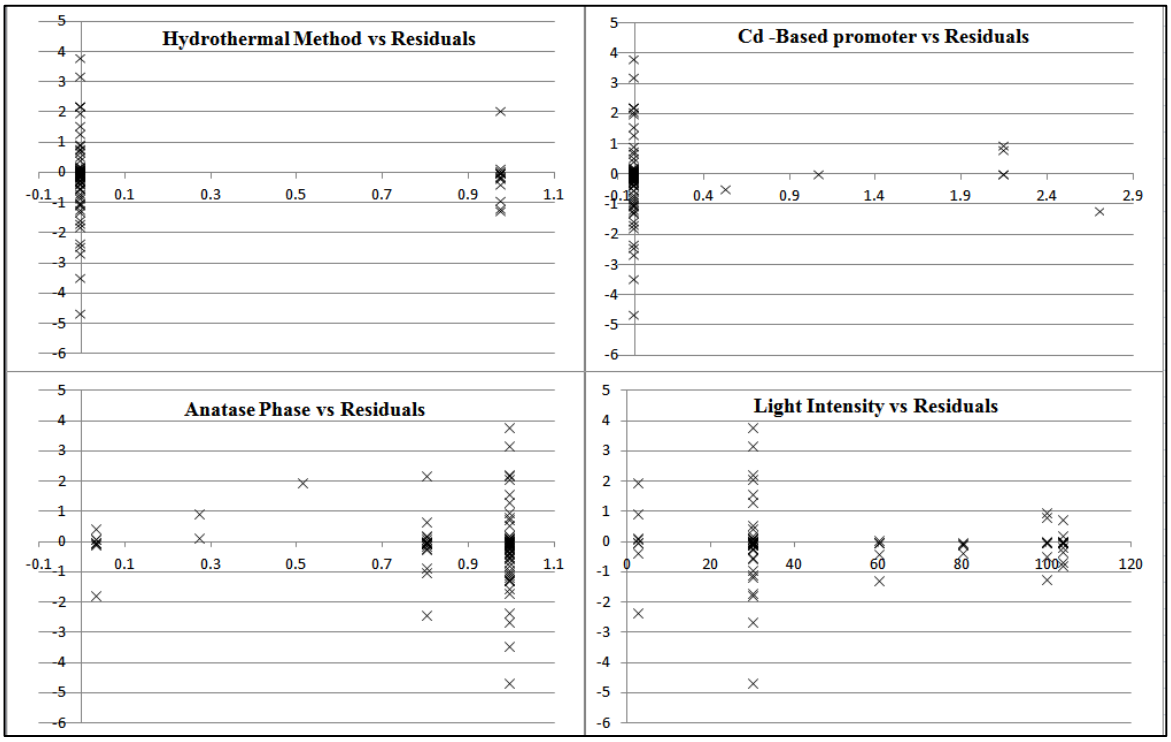

Figure 4.21. Most Important Input Variables vs. Residual Plots.

4.3.1. Linear Regression

Linear regression model was also used for the related dataset by using same function "lm" in R. Results of prediction of hydrogen production (μmol/gcat/h) on test set are given in Figure 4.22. Since model predicts total hydrogen production mostly as negative values and it gives very high rmse values; it can be concluded that the model was so poor. There were also so much fixated values around -6000, -2000, 3000, and 6000.

4.3.2. Random Forest

In random forest modeling, the same function, randomForest, was used and optimum parameters were tried to find, with that purpose; number of tree (t) was increased from 1 to 250 by 3, and minimum size of terminal nodes (n) changed from 1 to 5, simultaneously.

10-fold CV was applied to dataset and the following results were obtained by this method. The global minimum error 0.08 and minimum rmse 0.17 were reached when t equals to 22 and n is 1, and r-squared value recorded as 0.97. After analyzing the results, it

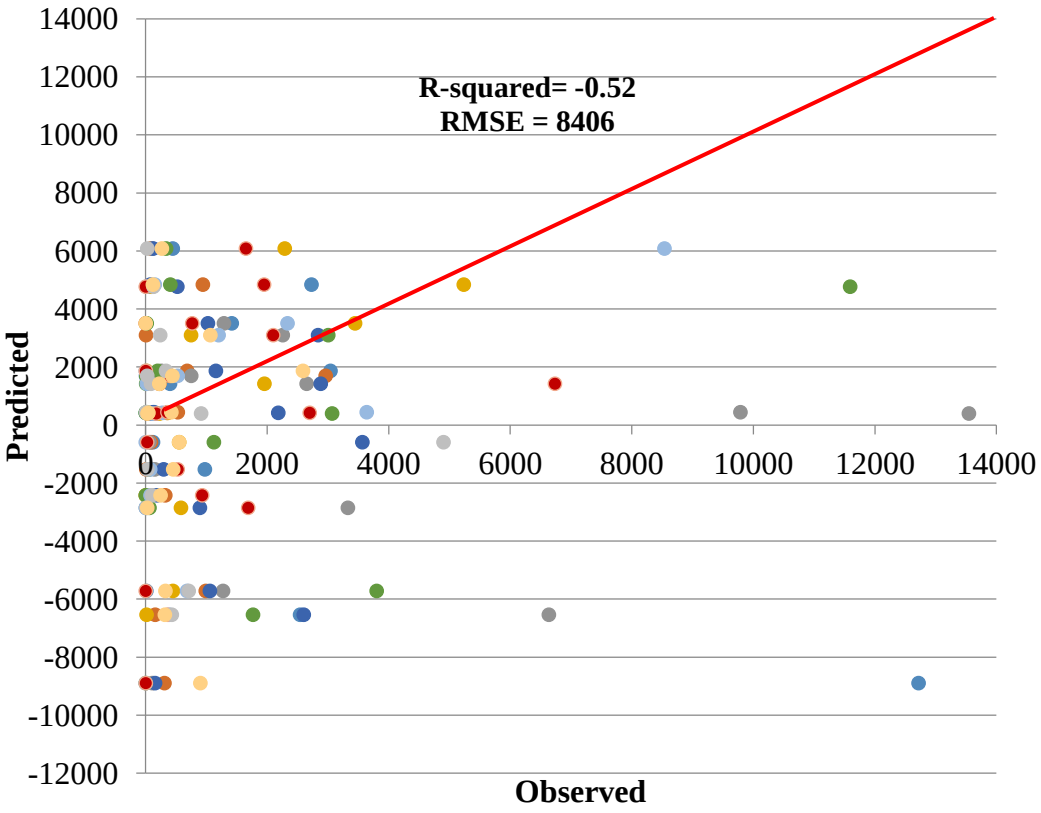

Figure 4.22. Predicted vs. Observed Output Values with Linear Regression.

was seen that the maximum value for test set error is 0.16; corresponding rmse and r-squared values are 0.75 and 0.43, respectively. This worst performance belonged to random forest of only 1 tree; r-squared values for models with more than one tree were recorded as greater than 0.7. Therefore, it can be concluded as random forest performed well for this dataset and the best performance of that model is given in Figure 4.23 and the y and x axes stand for predicted and observed hydrogen production (µmol/gcat/h) values.

Best model was also tested for each experiment and Figure 4.24 gives the plot of prediction vs. observed hydrogen production (µmol/gcat/h) values; the error, rmse, and r-squared value were recorded as 0.10, 0.24, and, 0.93, respectively. This model can be dedicated as a promising way to predict output of an unperformed experiment.

Relative importances of all variables were found by both of the criteria: mean squared error and node impurity. Since there are 72 used variables in that dataset, to give their relative importance in one plot is impossible. Besides, it was observed that about 20

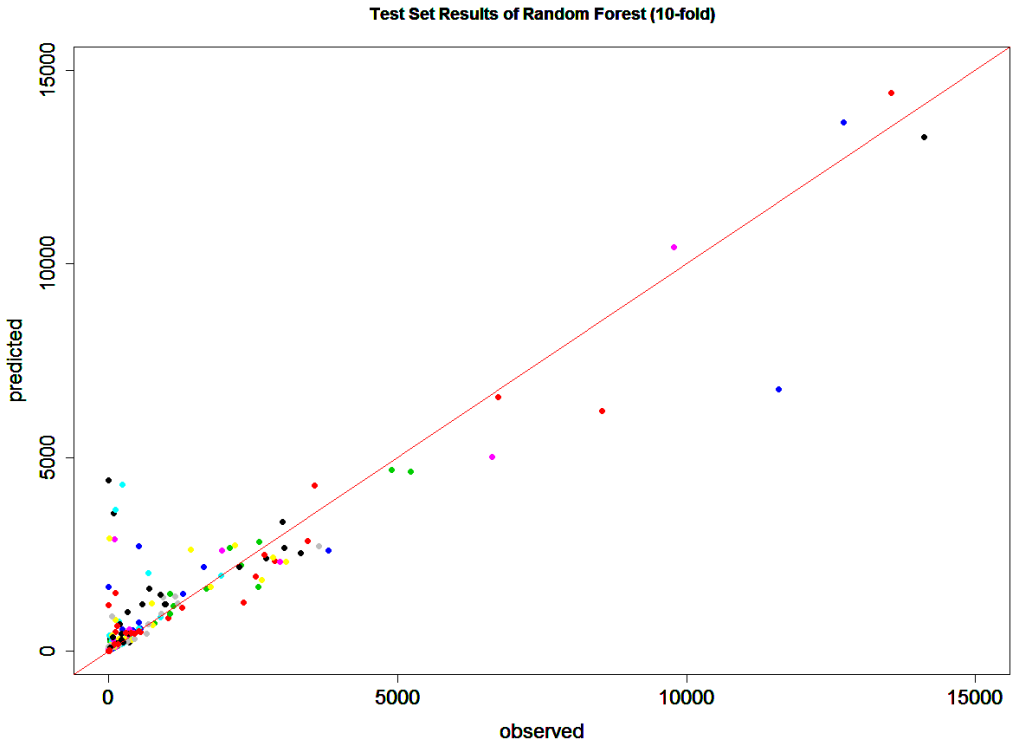

Figure 4.23. Predicted vs. Observed Output Values with Random Forest (t=22, n=1).

parameters have almost the same and relatively small effects on the model. Therefore, top 10 input variables were given with their relative importance values found by two criteria in Figure 4.25 and 4.26. The interesting point observed as a result of this analysis is that the elements were selected to produce a perovskite-type photocatalyst is much more important than catalyst properties such as band gap, crystal size, and surface area. These properties of photocatalyst were found as the main variables that affect PWS activity for titanium based catalysts in previous section.

Nb took place in both plot, it means that this variable is one of the deterministic variables, and as it is mentioned in Chapter 3, the number of niobium based perovskite in database is relatively higher than others so it may be the reason of finding Nb as significant variable. Total reaction time had minor effects on modeling of titanium based photocatalysts dataset, but in perovskites it seemed that it is an effective input variable. K and Na were the most preferred elements for "A" side of perovskites, that's why to see them in those plots is an anticipated result. Band gap, surface area, crystallite size, and heating conditions can be also denoted as significant input variables as in the modeling of titanium based photocatalysts experiments.

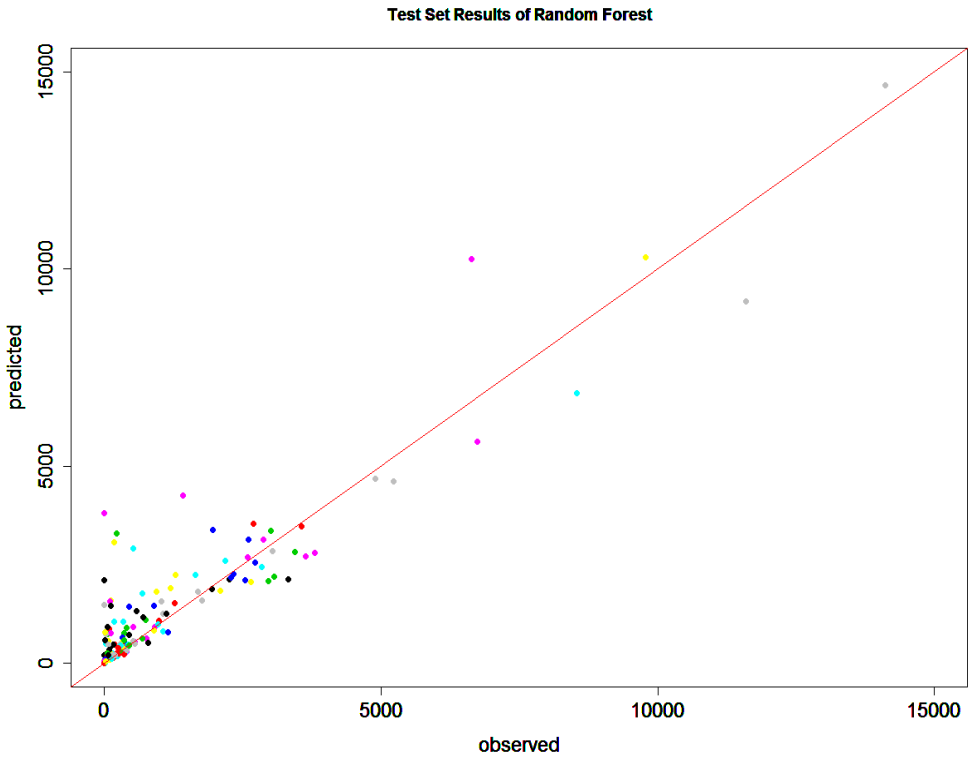

Figure 4.24. Predicted vs. Observed Output Values of Each Experiment with random forest (t=22, n=1).

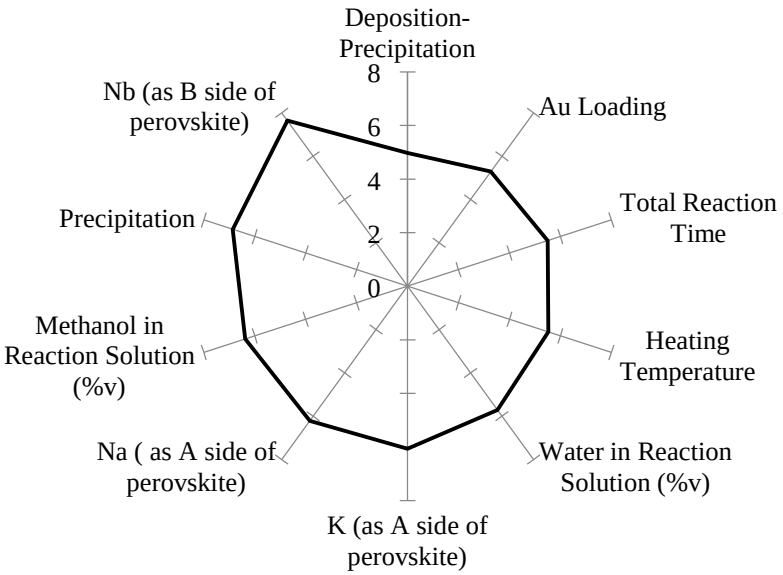

Figure 4.25. Relative Importance of Used Input Variables in Random Forest by Considering Change in Mean Squared Error.

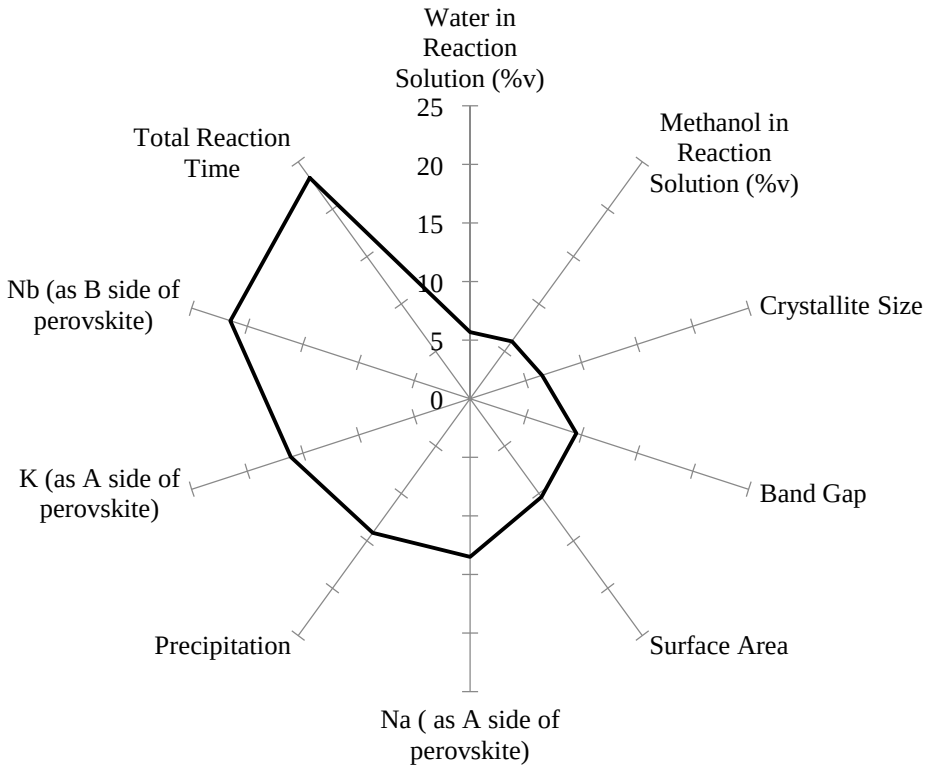

Figure 4.26. Relative Importance of Used Input Variables in Random Forest by Considering Change in Node Impurity.

Residual analysis was performed for top 5 input variables and the results were given in Figure 4.27. The randomly scattered residual errors with respect to input variables K, Nb, methanol, and reaction time can be seen obviously from the plots. However, precipitation is a nominal variable; all points were gathered around 0 in precipitation vs. residuals plot (Figure 4.27), it means there was a few number of experiments which used precipitation as preparation method, but the residuals are still randomly distributed between two sides. By following the same assumption that the residual errors should distribute normally and they should fall between -2 and 2 with 95% confidence interval, it can be said that this dataset has few outliers which are around 8 and -6. As a result of this analysis, the dataset can be considered as well-modeled and results were more likely reliable.

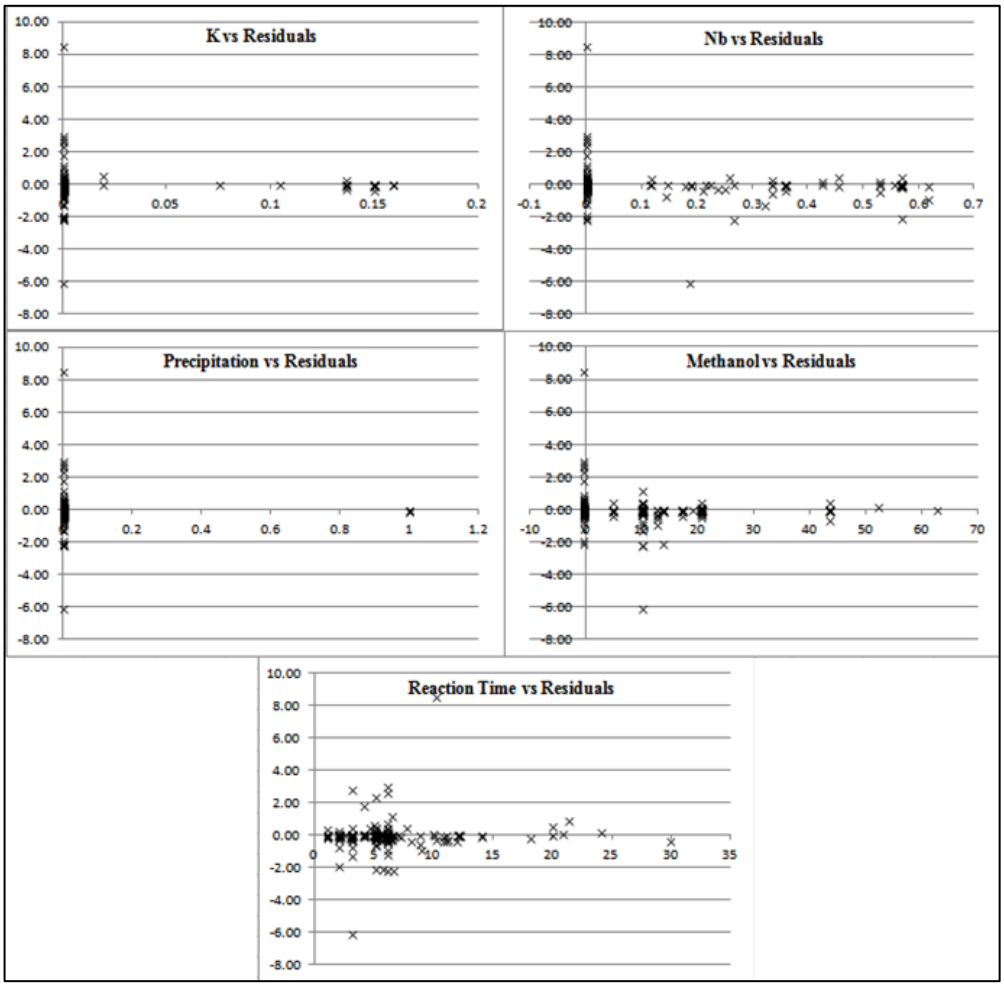

Figure 4.27. Most Important Input Variables vs. Residual Plots.

4.4. Analysis of Data for ABS_3-type Perovskite Photocatalyst

Linear regression, neural network and random forest methods were also performed for this database. However the output value predictions with neural network were not satisfactory as it was observed in ABO_3-type perovskite database. Therefore again the results of NN-modeling are not included in this report. This dataset originally includes 56 catalytic and 9 operational input variables, and one output variable. For specifically this dataset, decreasing number of input variables by extracting relatively unimportant variables did not improve the prediction ability of random forest modeling, so all variables were used for the models. It also comprises of 72 experiments conducted under different conditions.

4.4.1. Linear Regression

Linear regression model was tried with this dataset lastly by using “lm” function in R. The predicted vs. observed hydrogen production (µmol/gcat/h) values are given in Figure 4.28 and it can be observed that model gives only fixated values, so further development was assumed as unnecessary.

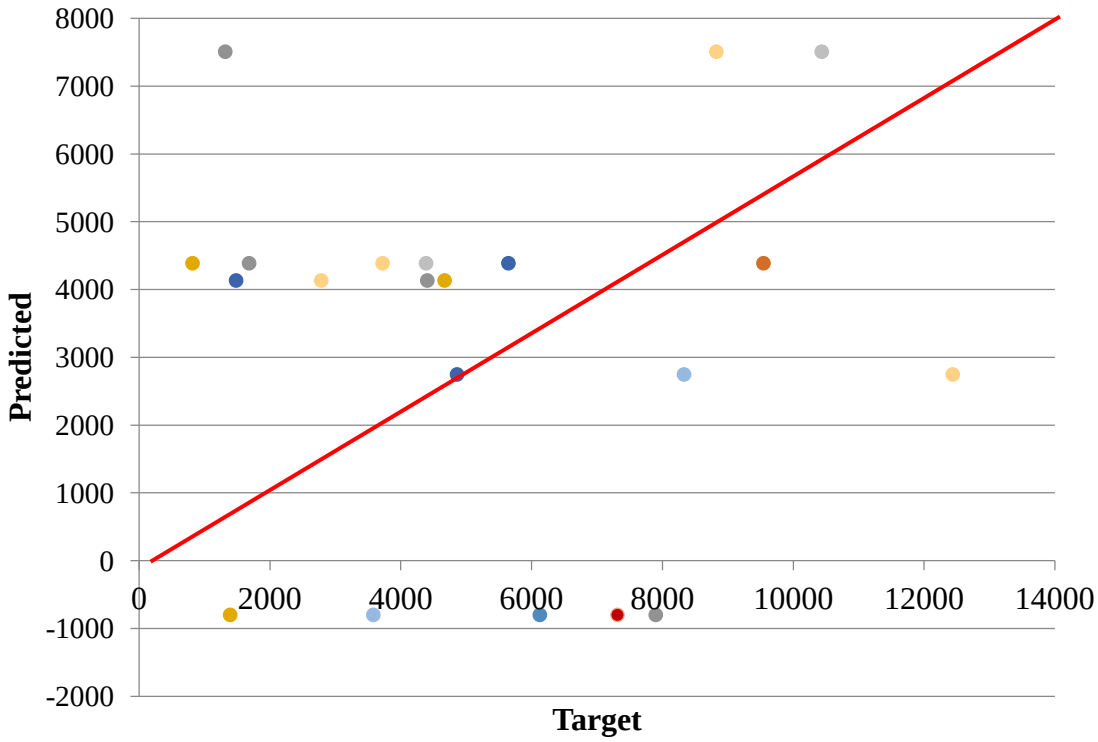

Figure 4.28. Predicted vs. Observed Output Values with Linear Regression.

4.4.2. Random Forest

In random forest modeling, the same function, randomForest, was used and optimum parameters were tried to find, with that purpose; number of tree (t) was increased from 1 to 250 by 3, and minimum size of terminal nodes (n) changed from 1 to 5, simultaneously.

10-fold CV was applied to dataset and the following results were obtained by this method; the global minimum error of 0.41 and minimum rmse of 0.60 were reached when t equals to 22 and n is 2, and r-squared value recorded as 0.64. Despite high r-squared values

of random forest modeling with ABO_3-type perovskite dataset, that dataset which comprises of ABS_3-type perovskites unable to reach r-squared values higher than 0.64. General prediction ability of that model was lower than others included in this thesis. The best performance of that model is given in Figure 4.29 and y and x axes represent predicted and observed hydrogen production (µmol/gcat/h) values. It can be seen from figure even with naked eyes that the model is weak, but it still may be considered as reasonable.

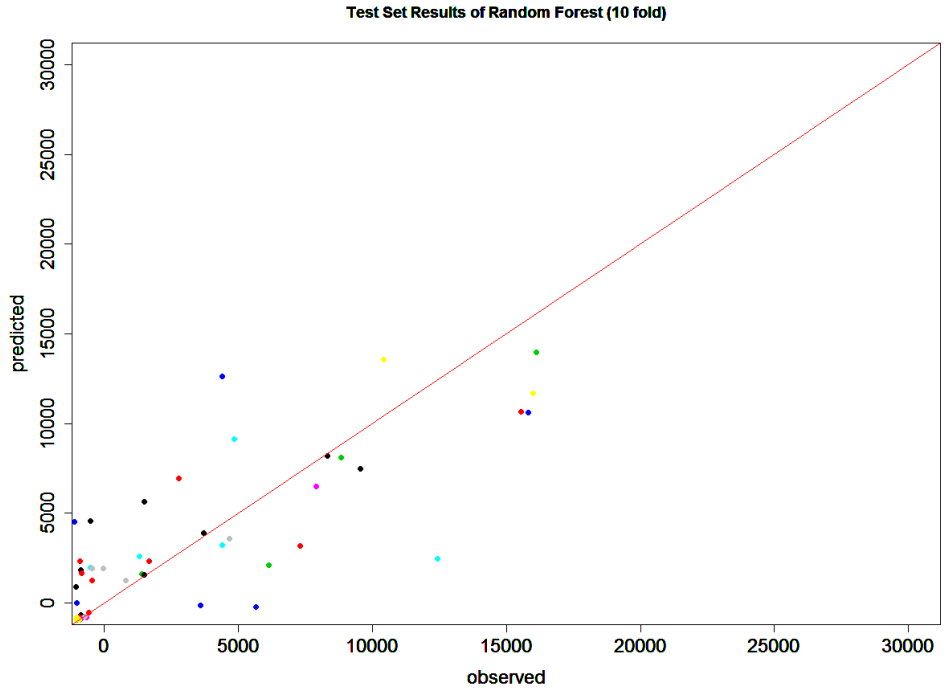

Figure 4.29. Predicted vs. Observed Output Values with NN-modeling (t=10, n=1).

By following the same logic, the predicted vs. experimental hydrogen production (µmol/gcat/h) values of each experiment included in dataset are given in Figure 4.30. The error, rmse, and r-squared value were recorded as 0.42 and 0.61, and 0.62 respectively.

Input significance analysis was also performed for this dataset, and the results are given in Figure 4.31 and 4.32 with mean squared error and node impurity criteria, respectively. Zn and Cd, which were recorded in dataset as B sides of perovskites, are seem as deterministic variables, and also according to Figure 4.31 and 4.32, S (the X side of perovskite) has a significant effect in random forest modeling. As it is expected from previous results, catalyst weight, crystal size, band gap, and surface area are the other variables that affect PWS activity considerably. Same knowledge with previous section which claims that semiconductor material is much more important than catalyst properties

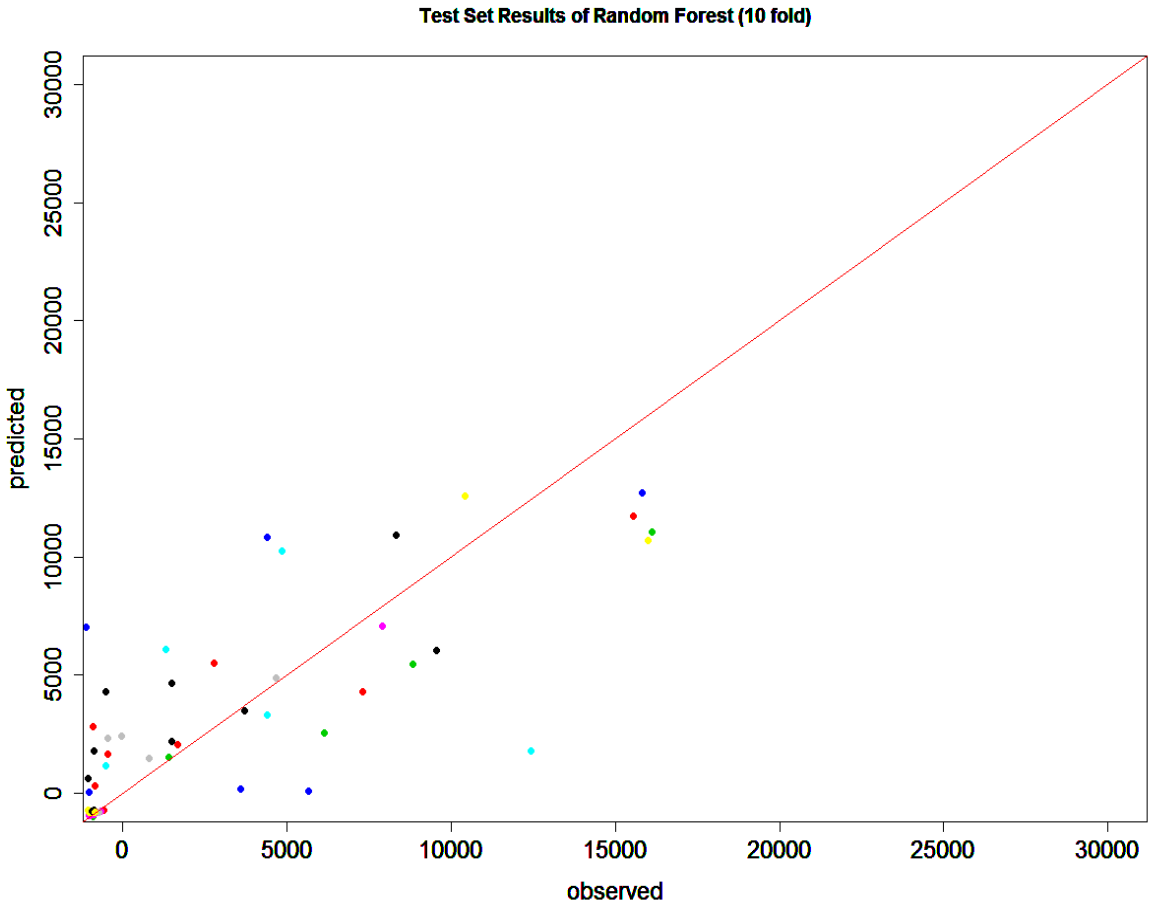

Figure 4.30. Predicted vs. Observed Output Values of each Experiment with random forest (t=10, n=1).

for modeling can be also extracted from these figures. However, since the results of model are not so good for this dataset; that knowledge can be supportive for the previous one but not proves it faithfully.

As the last step, the residual analysis was performed for top 5 input variables and the results are given in Figure 4.33. The random errors randomly scattered residuals with respect to input variables Cd, S, Zn, crystal size and catalyst weight can be seen obviously form plots. Also it can be easily understood that dataset involves fewer outlier than previous ones.

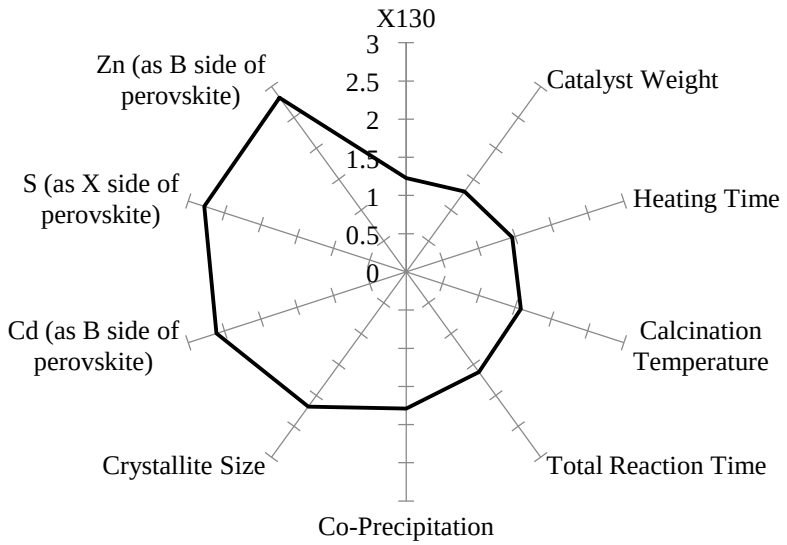

Figure 4.31. Relative Importances of Used Input Variables in Random Forest by Considering Change in Mean Squared Error.

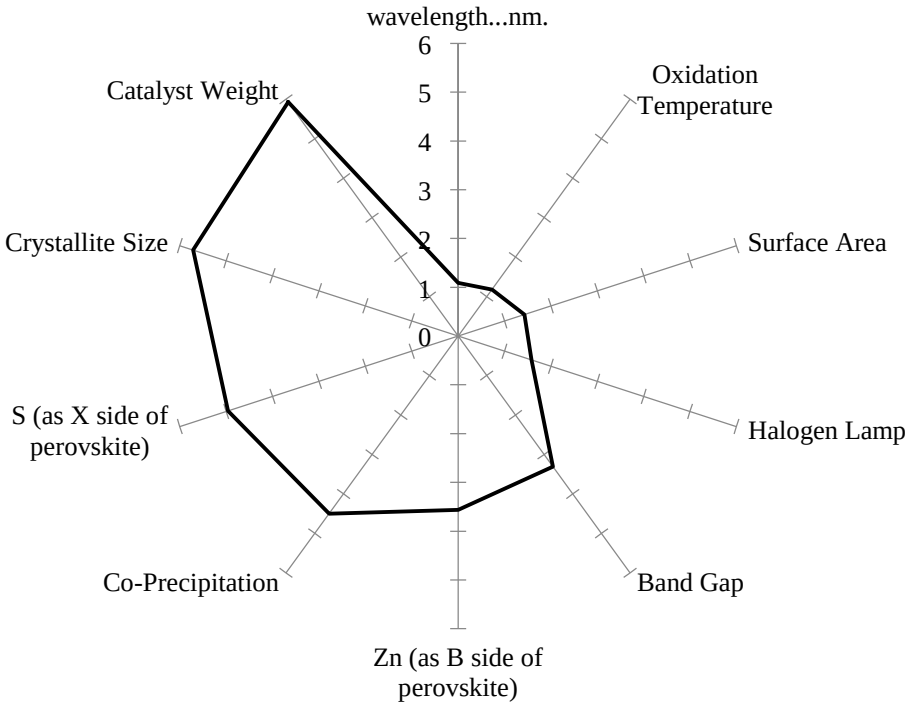

Figure 4.32. Relative Importances of Used Input Variables in Random Forest by Considering Change in Node Impurity

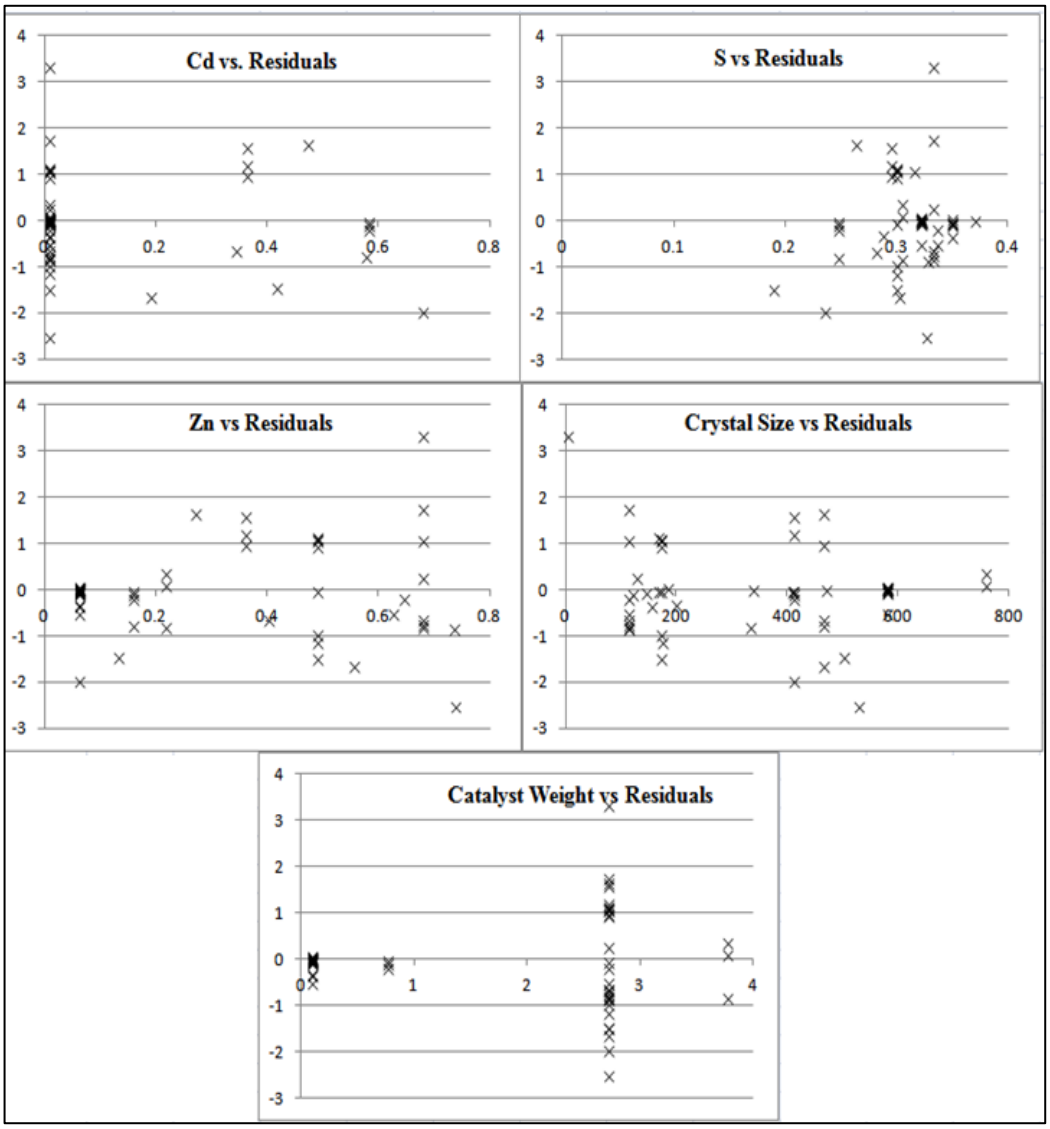

Figure 4.33. Most Important Input Variables vs. Residual Plots.

4.5. Principal Component Analysis

Principal component analysis (PCA) was performed with all four dataset mentioned above (titanium based catalysts under UV and visible light, ABO_3-type perovskites and ABS_3-type perovskites). The most important variables, which were identified with input significance analyses and some additional variables (light intensity, wavelength, band gap, surface area) whose effects are desired to observe were used to perform PCA. The datasets were divided into six classes according to their output values and then PCA was applied for all of these datasets to reduce their dimension to two by using "promp" function of R.

4.5.1. PCA for TiO2 Photocatalyst with UV Light Source

This dataset was divided into six subsets; as they can be seen from the legend of Figure 4.34, the calcination time, Pt loading, surface area, light intensity, wavelength, and band gap were the analyzed variables. Surface area vector are nearly orthogonal to all other vectors meaning that this variable has no relation with other variables. Along the vector specifying surface area, there are some light red points, which represent hydrogen evolution values between 2000 and 6000 μmol/gcat/h, so it can be said that surface area favors PWS activity. Light intensity, Pt loading and calcination time seemed to be correlated with each other and since the vectors, which represent them, are on the dark red points (hydrogen production above 6000 μmol/gcat/h), positive effects of these variables on PWS activity are obvious. Wavelength and band gap are inversely correlated with those three variables (calcination time, Pt loading, and light intensity), and they have also a negative influence on hydrogen evolution. This influence can be also observed from the figure as they point out data points with low hydrogen evolution rates. It is logical due to the fact that increase in band gap of semiconductor means more energy requirement to photogenerate electrons and holes, and decrease in wavelength means that light source radiates waves in visible more than UV region. On the other hand the correlation between wavelength and other variables is derived from their same effect on PWS activity, else there are no relation between those variables because wavelength of a light is independent from others.

4.5.2. PCA for TiO2 Photocatalyst with Visible Light Source

This dataset was divided into six subsets again, however the limits of subsets were lower than previous dataset, because as it was mentioned before, the hydrogen production rate becomes smaller when visible light source was used. Calcination time, Pt loading, surface area, light intensity, wavelength, and band gap used were reduced to 2D by PCA; the results are given in Figure 4.35. Surface area vector again can be considered as orthogonal to all other vectors, which indicates its irrelevance with other variables. On the contrary of previous dataset, the results show that surface area has a negative effect on PWS activity since the vector specifying surface area is on dark blue points (hydrogen production below 50 μmol/gcat/h). While light intensity has a moderate effect on hydrogen evolution, it cannot be said anything certain about Pt loading, calcination time, wavelength

and band gap because they are on both red and blue data points. Pt loading and calcination time was observed before as supportive variables while band gap and wavelength were determined as inhibitor variables, but according to the Figure 4.35 they are correlated with each other. Although the total variance of two PCs' (%69) is relatively higher, the results are unreliable due to its conflicting arguments.

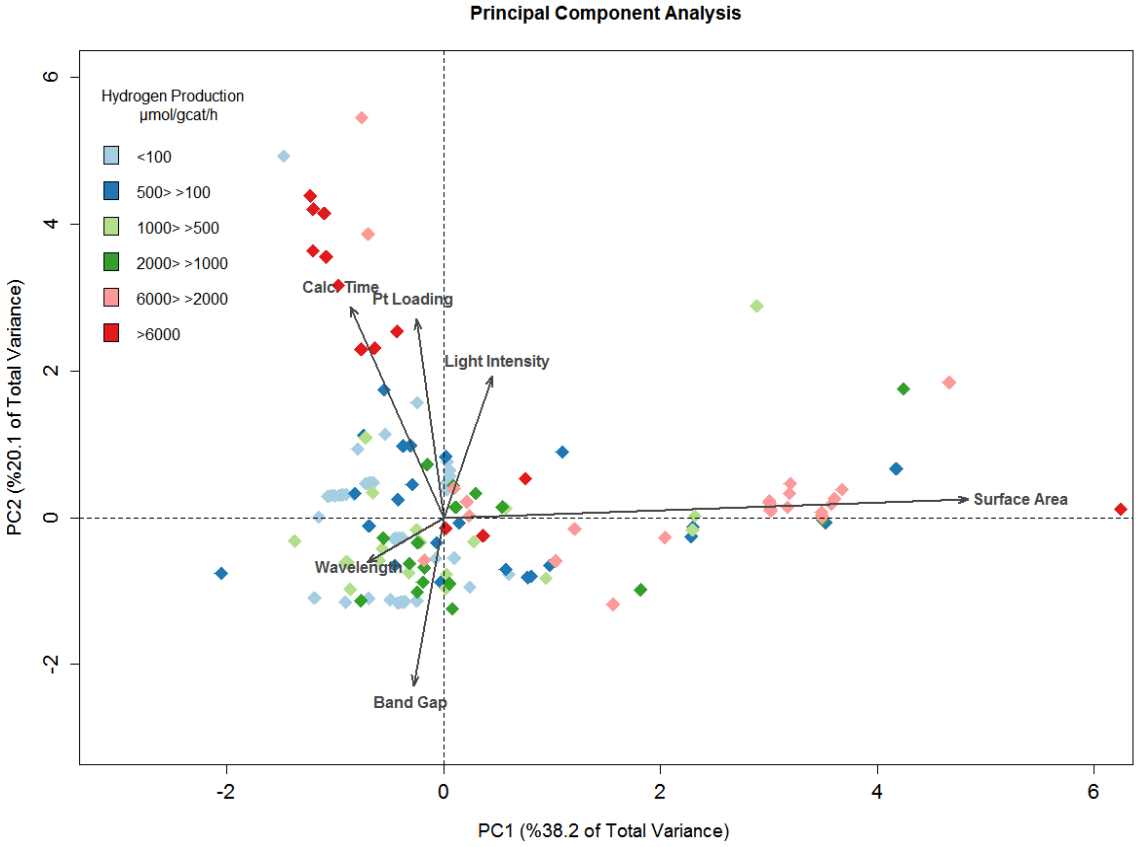

Figure 4.34. Principal Component Analysis of Titanium Based Photocatalysts under UV Light.

4.5.3. PCA for ABO_3-Type Perovskite Photocatalyst

Again six subsets were formed from this dataset and PCA was applied by using the same function, "promp" of R; Figure 4.36 shows the results. However as it can be seen from the figure, subsets could not be separated properly when they projected onto two dimensional space, so it cannot be said anything about the relation between variables used and their effects on PWS activity. Nevertheless, precipitation is seemed as positively correlated with Nb and K but it is uncorrelated with light intensity, band gap, wavelength, and reaction time. Precipitation is a method for promoting co-catalysts onto semiconductor

and so it is logical that it has no relation with light type (light intensity and wavelength) or reaction time at all. Reaction time is also an independent variable and it is orthogonal to some variables as it is expected, still the co-linearity of reaction time with wavelength, light intensity, and band gap can be explained by their possible similar effects on PWS activity, otherwise reaction time has no relation with those variables.

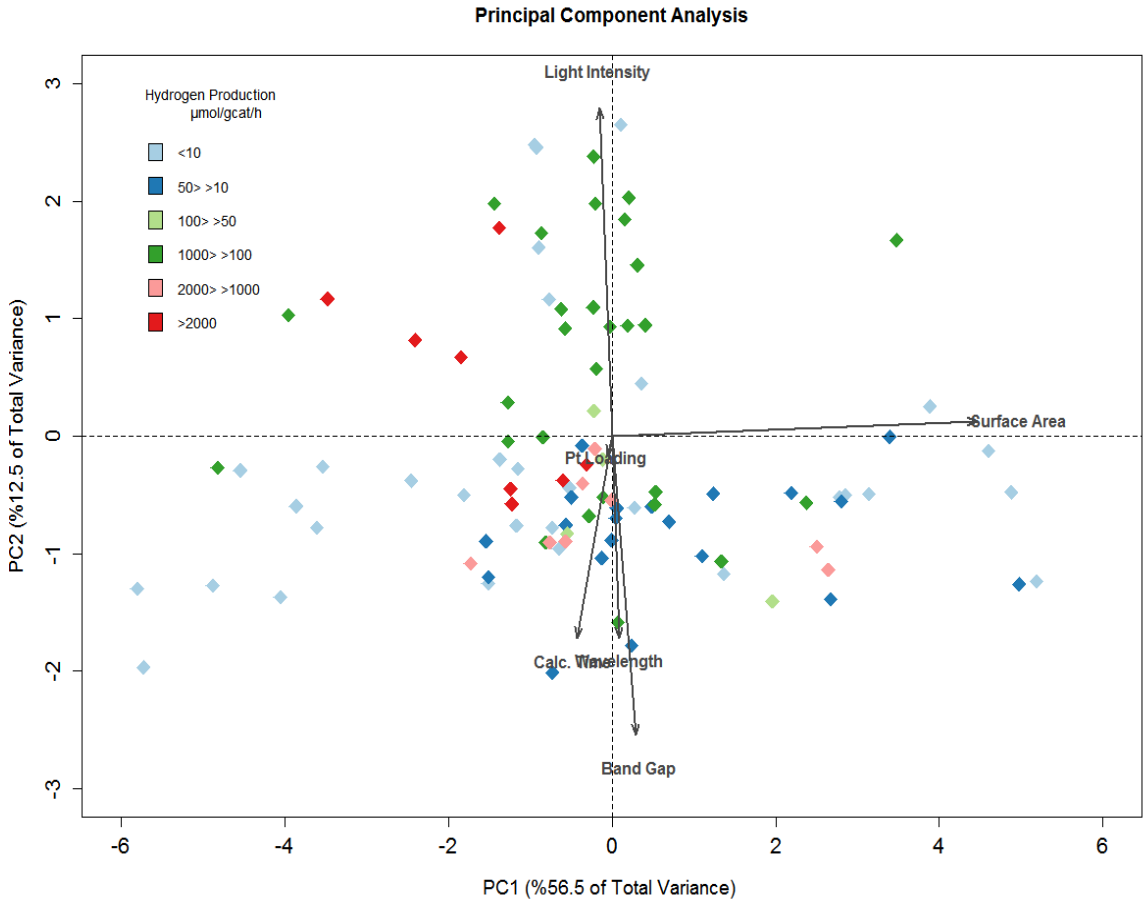

Figure 4.35. Principal Component Analysis of Titanium Based Photocatalysts Under Visible Light.

4.5.1. PCA for ABS_3-Type Perovskite Photocatalyst

The separation of this dataset into subsets was more successful than previous datasets as it is shown in Figure 4.37. Light intensity and Cd exhibits a positive effect on PWS activity since they are on red and green data points (hydrogen production rate between 1000 and 6000 µmol/gcat/h). Zn and catalyst weight has no relation with light intensity and Cd but they are seemed also as favoring variables of PWS reaction. Crystal

size, band gap, and wavelength are positively correlated with each other and they are observed as they have negative influence on PWS activity. Small crystal size is generally preferred in the literature, (Liu and Syu, 2013), so it is logical that it has negative relation with hydrogen production rate. As it is mentioned before greater wavelength means less energy of light so it reduces hydrogen production.

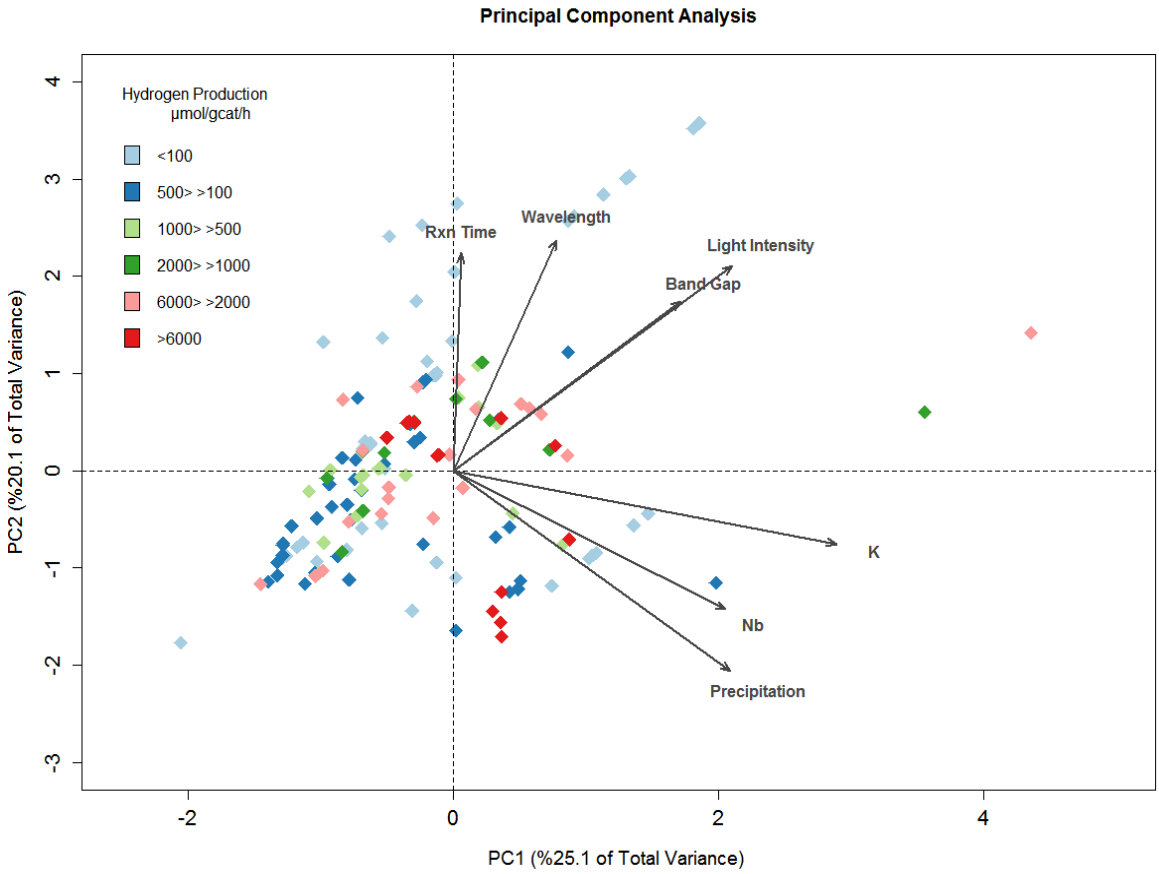

Figure 4.36. Principal Component Analysis of ABO_3-Type Perovskite Photocatalysts.

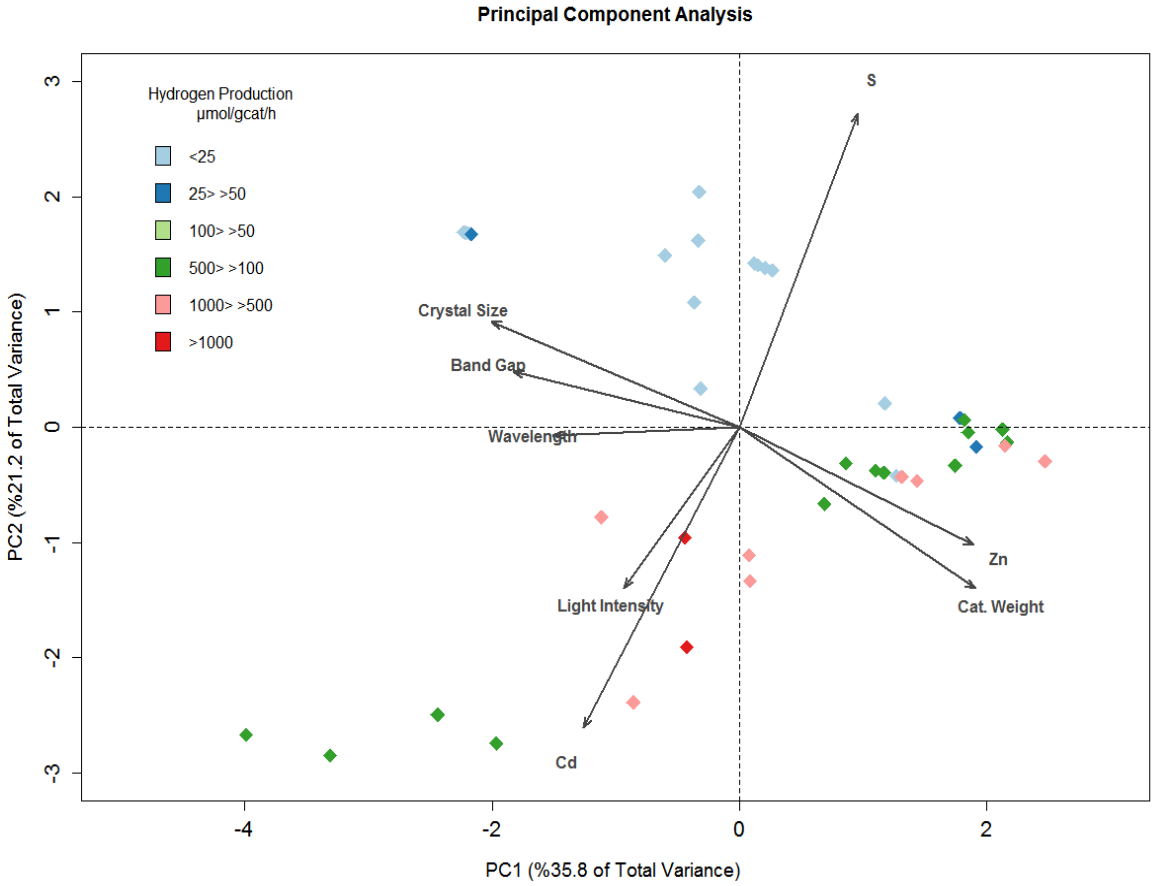

Figure 4.37. Principal Component Analysis of ABS_3-Type Perovskite Photocatalysts.

5. CONCLUSIONS AND RECOMMENDATIONS

5.1. Conclusions

In this thesis, a comprehensive database was formed by collecting experimental data from the published articles between the years 2005 and 2014 by using Science Direct, Wiley and ACS online databases. The database was originally consisted of 6378 instances and 129 articles; after pre-processing step, the database became more compact and suitable for modeling with 541 instances and 107 articles. Pre-processing step involved filling missing values of physical properties by predicting from the other given variables, smoothing the noisy data, removing outliers, normalization and reducing volume of data while keeping as much information as possible. The database was also divided into two subsets first by considering their semiconducting material types (perovskite type and titanium based semiconductors), and then further segmentation was done to perform better modeling. As a result, four small databases were formed and then linear regression, neural network, random forest and principal component analysis were applied to each of database. In all models 10-fold cross validation was used to evaluate the model.

For all four subsets linear regression resulted in poor predictions as it was expected, because there were complex non-linear relationships between input variables that cannot be learn by linear models.

In the analysis of data for TiO_2 photocatalyst with UV light source, the best model was obtained using random forest method. The global minimum error of 0.11 and minimum rmse of 0.19 were reached when number of tree (t) equals to 19 and minimum size of terminal nodes (n) is 2, and r-squared value recorded as 0.96. The results of NN modeling were also acceptable even though it was not as good as the random tree model. In both NN and random forest modeling, surface area, band gap, calcination time and Pt loading were in the top 5 of important variable lists which were obtained as a result of input significance analysis. Residual analysis was also performed and it was seen that residual errors scatter randomly, which proves the related dataset was modeled well and the results obtained by that model were reliable.

In the analysis of data for TiO_2 photocatalyst with visible light source, again random forest predicted the output variables better than NN-modeling. The global minimum error of 0.12 and minimum rmse of 0.24 were reached when t and n were set to 7 and 3, respectively, and r-squared value was 0.94. The input significance analysis performed at the end of NN-modeling and random forest and gave slightly different results; but in general terms, light intensity, surface area, heating temperature, and hydrothermal method which was the most common way of catalyst production, were chosen as the most deterministic variables. The goodness of model was also proved with residual analysis.

The highest r-squared value of 0.97, and the minimum error and rmse values of 0.08 and 0.17 were reached in random forest modeling of data for ABO_3-type perovskite photocatalyst; 22 trees and minimum 1 instance for terminal nodes were chosen while constructing the model. Nb (as B site of perovskite), K (as A site of perovskite), total reaction time and using precipitation method were determined as relatively more important input than the other input variables. In residual analysis, it was observed that residual errors scattered randomly within acceptable confidence interval. The value predictions with NN-modeling were not reliable due to its high error and low r-squared values for that dataset.

The data for ABS_3-type perovskite photocatalyst gave highest r-squared value of 0.64 by random forest modeling when t equals to 22 and n is 2; but the error and rmse were 0.42 and 0.61, respectively, so the predictions obtained by that model could not be considered as reliable. According to the results of input significance analysis, Zn and Cd which were recorded in dataset as B sites of perovskites were seem as deterministic variables.

As an interesting knowledge, it was observed that the elements used for production of perovskite type semiconductors are more significant than catalytic or operational variables in model building. On the other hand, in titanium based photocatalysts, the band gap, surface area, and particle size were dedicated as the most deterministic variables for modeling.

As the last step, PCA was applied for each small datasets as a supplementary analysis for input significance. The light intensity was determined as more important than the power of light, and the significance of Pt promoting for PWS activity was clearly seen. The effect of some input variables such as band gap, surface area or particle size, on the other hand, could not be identified properly due to conflicting results. The reason for this may be the high dimensions of datasets, reducing those datasets in two dimensions may cause data loss and PCA results became unreliable.

5.2. Recommendations

In the end of this work, some recommendations can be offered to develop better models and extract more detailed and accurate knowledge.

(i) To collect more data by analyzing more articles will improve the prediction ability of the models since it will be helpful to eliminate degrees of freedom problems, and while constructing the model, the training set will have a chance to see and learn data in a wider range.

(ii) There are some other machine learning tools and algorithms such as “support vector machine”, and “mixture of experts”; they can be also applied to reach better prediction results.

(iii) Output values can be converted into classes by working through and the classification methods such as decision tree that can be used for knowledge extraction and pattern recognition.

(iv) As another recommendation, some of the input variables (catalyst preparation methods, semiconducting elements, or crystal structure), can be converted into categorical variables instead of using as numeric variables; this may reduces dimension of data and may provide models better and quicker learning.

APPENDIX A: Articles Involved in Database

Table A.1 Articles Involved in Database.

Number of Article	Reference
1	Arney *et al.*, 2012
2	Baniasadi *et al.*, 2013
3	Biswal *et al.*, 2011
4	Chan *et al.*, 2014
5	Chen *et al.*, 2014
6	Chen *et al.*, 2012
7	Chen *et al.*, 2010
8	Chiou *et al.*, 2009
9	Cios *et al.*, 2009
10	Dang *et al.*, 2013
11	D'Elia *et al.*, 2011
12	Ding *et al.*, 2013
13	Fan *et al.*, 2010
14	Fan *et al.*, 2014
15	Fan *et al.*, 2013
16	Fang *et al.*, 2012
17	Gomathinaskar *et al.*, 2013
18	Gomez-Solis *et al.*, 2014
19	Guo and Han, 2014
20	Hao *et al.*, 2013
21	He and Guo, 2014
22	Hong *et al.*, 2014
23	Hu and Teng, 2007
24	Huang *et al.*, 2011
25	Ikeda *et al.*, 2006
26	Ismail and Bahnemann, 2014
27	Jana *et al.*, 2014
28	Jeong *et al.*, 2006
29	Jitputti *et al.*, 2007
30	Kanhere and Chen, 2014
31	Kanhere *et al.*, 2012
32	Khan and Akhtar, 2008
33	Khan *et al.*, 2008
34	Khan and Yang, 2009
35	Kim and Kang, 2012
36	Kim *et al.*, 2012

37	Kimi *et al.*, 2011
38	Kokporka *et al.*, 2013
39	Le *et al.*, 2012
40	Lee *et al.*, 2012
41	Lee eand Kang, 2013
42	Lee *et al.*, 2013
43	Lee *et al.*, 2012
44	Li *et al.*, 2010
45	Li *et al.*, 2008
46	Li *et al.*, 2010
47	Li *et al.*, 2011
48	Li *et al.*, 2009
49	Li *et al.*, 2009
50	Li *et al.*, 2008
51	Li *et al.*, 2010
52	Li *et al.*, 2014
53	Li *et al.*, 2011
54	Li *et al.*, 2010
55	Li *et al.*, 2014
56	Li *et al.*, 2009
57	Li *et al.*, 2013
58	Liao *et al.*, 2013
59	Lin *et al.*, 2011
60	Lin *et al.*, 2010
61	Lin *et al.*, 2012
62	Liu and Syu, 2013
63	Liu *et al.*, 2012
64	Liu *et al.*, 2008
65	Liu *et al.*, 2011
66	Martha *et al.*, 2012
67	Naik *et al.*, 2011
68	Onsuratoom *et al.*, 2011
69	Oros-Ruiz *et al.*, 2014
70	Parayil *et al.*, 2012
71	Parida *et al.*, 2010
72	Perez-Larios *et al.*, 2012
73	Puangpetch *et al.*, 2009
74	Rayalu *et al.*, 2013
75	Rosseler *et al.*, 2010
76	Rungjaroentawon *et al.*, 2012
77	Sathish *et al.*, 2007
78	Serrano *et al.*, 2014
79	Shen *et al.*, 2013
80	Sreethawong *et al.*, 2005

81	Sreethawong and Yoshikawa, 2006
82	Sun *et al.*, 2013
83	Sun *et al.*, 2012
84	Torres-Martinez *et al.*, 2010
85	Wang *et al.*, 2013
86	Wang *et al.*, 2012
87	Wang *et al.*, 2013
88	Wang *et al.*, 2013
89	Wang *et al.*, 2009
90	Wei *et al.*, 2013
91	Wei *et al.*, 2009
92	Wu *et al.*, 2012
93	Wu *et al.*, 2014
94	Yan *et al.*, 2009
95	Yao *et al.*, 2014
96	Yoshida *et al.*, 2014
97	Yoshioka *et al.*, 2005
98	Yu *et al.*, 2013
99	Zhang *et al.*, 2013
100	Zhang *et al.*, 2011
101	Zhang *et al.*, 2012
102	Zhang *et al.*, 2013
103	Zhao *et al.*, 2012
104	Zheng *et al.*, 2013
105	Zhou *et al.*, 2009
106	Zhou *et al.*, 2011
107	Zielinska *et al.*, 2012

REFERENCES

Arney, D., L. Fuoco, J. Boltersdorf, P. A. Maggard, 2012, "Flux Synthesis of $Na_2Ca_2Nb_4O_{13}$: The Influence of Particle Shapes, Surface Features, and Surface Areas on Photocatalytic Hydrogen Production", *Journal of American Ceramic Society*, Vol. 96, pp. 1148-1162.

Baniasadi, E., I. Dincer, G. F. Natere, 2013, "Hybrid Photocatalytic Water Splitting for an Expanded Range of the Solar Spectrum with Cadmium Sulfide and Zinc Sulfide Catalysts", *Applied Catalysis A: General*, Vol. 455, pp. 25-31.

Baydoğan, M. G., *IE 582 Statistical Learning for Data Mining*, Bogazici University, Istanbul, Turkey, 2014.

Biswal, N., D. P. Das, S. Martha, K. M. Parida, 2011, "Efficient Hydrogen Production by Composite Photocatalyst CdS-ZnS/Zirconium-Titanium Phosphate (ZTP) Under Visible Light Illumination", *International Journal of Hydrogen Energy,* Vol. 36, pp. 13452-13460.

Borgarello, E., J. Kiwi, M. Graetzel, E. Pelizetti, M. Visca, M., 1982, "Visible Light Induced Water Cleavage in Colloidal Solutions of Chromium-doped Titanium Dioxide Particles", *Journal of American Chemical Society*, Vol. 104, pp. 2996-3002.

Chan, C. C., C. C. Chang, C. H. Hsu, Y. C. Weng, K. Y. Chen, H. H. Lin, W. C. Huang, S. F. Cheng, 2014, "Efficient and Stable Photocatalytic Hydrogen Production from Water Splitting over $Zn_xCd_{1-x}S$ Solid Solutions Under Visible Light Irradiation", *International Journal of Hydrogen Energy*, Vol. 39, pp. 1630-1639.

Chen, W., X. Chen, Y. Yang, J. Yuan, W. Shangguan, 2014, "Synthesis and Performance of Layered Perovskite-Type H-$ABi_2Ta_2O_9$ (A = Ca, St, Ba, $K_{0.5}La_{0.5}$) for Photocatalytic Water Splitting", *International Journal of Hydrogen Energy*, Vol. 39, pp. 1-6.

Chen, W., C. Li, H. Gao, J. Yuan, W. Shangguan, J. Su, Y. Sun, 2012 "Photocatalytic Water Splitting on Protonated Form of Layered Perovskites $K_{0.5}La_{0.5}Bi_2M_2O_9$ (M=Ta; Nb) by Ion-Exchange", *International Journal of Hydrogen Energy*, Vol. 37, pp. 12846-12851.

Chen, X., S. Shen, L. Guo, S. Mao, 2010, "Semiconductor –based Photocatalytic Hydrogen Generation", *Chemical Reviews*, Vol. 110, pp. 6503-6570.

Chiou, Y. C., U. Kumar, J. C. S. Wu, 2009, "Photocatalytic Splitting of Water on $NiO/InTaO_4$ Catalysts Prepared by an Innovative Sol-Gel Method", *Applied Catalysis A: General*, Vol. 357, pp. 73-78.

Cios, K. J., W. Pedrycz, R. W. Swiniarski, L. A. Kurgan, *Data Mining A Knowledge Discovery Approach*, Springer Science & Business Media, New York, USA, 2007.

Dang, H., X. Dong, Y. Dong, J. Huang, 2013, "Facile and Green Synthesis of Titanate Nanotube / Graphene Nano-Composites for Photocatalytic H_2 Generation from Water", *International Journal of Hydrogen Energy*, Vol. 38, pp. 9178-9185.

Dang, H., X. Dong, Y. Dong, Y. Zhang, S. Hampshire, 2013, "TiO_2 Nanotubes Coupled with Nano-$Cu(OH)_2$ for Highly Efficient Photocatalytic Hydrogen Production", *International Journal of Hydrogen Energy*, Vol. 38, pp. 2126-2135.

DeBellis, M., *Neurons*, 2012, http://neuropsychologysketches.com/Neurons.html, [Accessed July 2015].

D'Elia, D. C., J. F. Beauger, J. F. Hochepied, A. Rigacci, M. H. Berger, N. Keller, V. Keller-Spitzer, Y. Suzuki, J. C. Valmalette, M. Benabdesselam, P. Achard, P. 2011, "Impact of Three Different TiO_2 Morphologies on Hydrogen Evolution by Methanol Assisted Water Splitting: Nanoparticles, Nanotubes and Aerogels", *International Journal of Hydrogen Energy*, Vol. 36, pp. 14360-14373.

Ding, J., S. Sun, W. Yan, J. Bao, C. Gao, 2013, "Photocatalytic H_2 Evolution on a Novel $CaIn_2S_4$ Photocatalyst Under Visible Light Irradiation", *International Journal of Hydrogen Energy*, Vol. 38, pp. 13153-13158.

Duonghong, D., E. Borgarello, M. Gratzel, 1981, "Dynamics of Light-Induced Water Cleavage in Colloidal Systems", *Journal of American Chemical Society*, Vol. 103, pp. 4685-4690.

Fan, W. J., Z. F. Zhou, W. B. Xu, Z. F. Shi, F. M. Ren, H. H. Ma, S. W. Huang, 2010, "Preparation of $ZnIn_2S_4$/Fluoropolymer Fiber Composites and its Photocatalytic H_2 Evolution From Splitting of Water Using Xe Lamp Irradiation", *International Journal of Hydrogen Energy*, Vol. 35, pp. 5525-6530.

Fan, X., J. Fan, X. Hu, E. Liu, E. Kang, L. Kang, 2014, "Preparation and Characterization of Ag Deposited and Fe Doped TiO_2 Nanotube Arrays for Photocatalytic Hydrogen Production by Water Splitting", *Ceramics International*, Vol. 40, pp. 15907-15917.

Fan, X., B. Lin, H. Liu, L. He, Y. Chen, B. Gao, 2013, "Remarkable Promotion of Photocatalytic Hydrogen Evolution from Water on TiO_2-Pillared Titanoniobate", *International Journal of Hydrogen Energy*, Vol. 38, pp. 832-839.

Fang, J., S. W. Cao, Z. Wang, M. M. Shahjamali, S. C. J. Lao, J. Barber, C. Xue, 2012, "Mesoporous Plasmonic Au-TiO_2 Nanocomposites for Efficient Visible-Light-Driven Photocatalytic Water Reduction", *International Journal of Hydrogen Energy*, Vol. 37, pp. 17853-17861.

Frost, J., *Why You Need to Check Your Residual Plots for Regression Analysis: Or, To Err is Human, To Err Randomly is Statistically Divine*, 2012, http://blog.minitab.com/blog/adventures-in-statistics/why-you-need-to-check-yourresidual-plots-for-regression-analysis, [Accessed June 2015].

Fujishima A., K. Honda, 1972, "Electrochemical Photolysis of Water at a Semiconductor Electrode", *Nature*, Vol. 238, pp. 37-38.

Fukuzumi S., D. Hong, Y. Yamada, 2013, "Bioinspired Photocatalytic Water Reduction and Oxidation with Earth-Abundant Metal Catalysts", *The Journal of Physical Chemistry Letters*, Vol. 4, pp. 3458-3467.

Gomathisankar, P., K. Hachisuka, H. Katsumata, T. Suzuki, K. Funasaka, S. Kaneco, 2013, "Enhanced Photocatalytic Hydrogen Production from Aqueous Methanol Solution Using ZnO with Simultaneous Photodeposition of Cu", *International Journal of Hydrogen Energy*, Vol. 38, pp. 11840-11846.

Gomez-Solis, C., M. A. Ruiz-Gomez, L. M. Torres-Martinez, I. Juarez-Ramirez, D. Sanchez-Martinez, 2014, "Facile Solvo-Combustion Synthesis of Crystalline $NaTaO_3$ and Its Photocatalytic Performance for Hydrogen Production", 2014, *Fuel*, Vol. 130, pp. 221-227.

Gunay M. E., F. Akpinar, I. Z. Onsan, R. Yıldırım, 2012, "Investigation of Water-Gas-Shift Activity of Pt-Mo_x-CeO_2/Al_2O_3 (M=K, Ni, Co) Using Modular Artificial Neural Networks", *International Journal of Hydrogen Energy*, Vol. 37, pp. 2094-2102.

Gunay M. E., R. Yıldırım, 2007, "Neural Network Aided Design of Pt-Co-Ce/Al_2O_3 Catalyst for Selective CO Oxidation in Hydrogen-Rich Streams", *Chemical Engineering Journal*, Vol. 140, pp. 324-331.

Gunay M. E., R. Yıldırım, 2010, "Analysis of Selective CO Oxidation over Promoted Pt/Al_2O_3 Catalysts Using Modular Neural Networks: Combining Preparation and Operational Variables", *Applied Catalysis A: General*, Vol. 377, pp.174-180.

Gunay M. E., R. Yıldırım, 2013, "Modeling Preferential CO Oxidation over Promoted Au/Al_2O_3 Catalysts Using Decision Trees and Modular Neural Networks", *Chemical Engineering Research and Design*, Vol. 91, pp. 874-882.

Guo, S.Y., S. Han, 2014, "Constructing a Novel Hierarchical 3D Flower-Like Nano/Micro Titanium Phosphate with Efficient Hydrogen Evolution from Water Splitting", Journal of *Power Sources*, Vol. 267, pp. 9-13.

Han, J., M. Kamber, J. Pei, *Data Mining Concepts and Techniques,* Elsevier Inc, Waltham, USA, 2012.

Hao, J., Y. Wang, X. Tong, G. Jin, X. Guo, 2013, "SiC Nanomaterials with Different Morphologies for Photocatalytic Hydrogen Production Under Visible Light Irradiation", *Catalysis Today,* Vol. 212, pp. 220-224.

He, K., L., Guo, 2014, "A Novel CdS Nanorod with Stacking Fault Structures: Preparation and Properties of Visible-Light-Driven Photocatalytic Hydrogen Production from Water Splitting", *Energy Procedia*, Vol. 61, pp. 245-2455.

Hong, E., D. Kim, J. H. Kim, 2014, "Heterostructured Metal Sulfide (ZnS-CuS-CdS) Photocatalysts for High Electron Utilization in Hydrogen Production from Solar Water Splitting", *Journal of Industrial and Engineering Chemistry*, Vol. 20, pp. 3869-3874.

Hu, C.C., H. Teng, 2007, "Influence of Structural Features on The Photocatalytic Activity of NaTaO3 Powders from Different Synthesis Methods", *Applied Catalysis A: General*, Vol. 331, pp. 44-50.

Huang, Y., Y. Li, Y. Wei, M. Huang, J. Wu, 2011, "Photocatalytic Property of Partially Substituted Pt-intercalated Layered Perovskite $ASr_2Ta_xNb_{3-x}O_{10}$ (A=K, H; x=0, 1, 1.5, 2 and 3)", *Solar Energy Materials & Solar Cells*, Vol. 95, pp. 1019-1027.

Ikeda, S., M. Fubuki, Y. K. Takahara, M. Matsumura, 2006, "Photocatalytic Activity of Hydrothermally Synthesized Tantalate Pyrochlores for Overall Water Splitting", *Applied Catalysis A: General*, Vol. 300, pp. 186-190.

Ismail, A. A., D. W. Bahnemann, 2014, "Photochemical Splitting of water for hydrogen production by photo-catalysis: A review", *Solar Energy Materials & Solar Cells*, Vol. 128, pp. 85-101.

Jana, P., C. M. Montero, P. Pizarro, J. M. Coronado, D. P. Serrano, V. A. P. O'Shea, 2014, "Photocatalytic Hydrogen Production in the Water/Methanol System Using Pt /

RE:$NaTaO_3$ (RE = Y, La, Ce, Yb) Catalysts, "*International Journal of Hydrogen Energy*, Vol. 39, pp. 5283-5290.

Jeong, H., T. Kim, D. Kim, K. Kim, 2006, "Hydrogen Production by the Photocatalytic Overall Water Splitting on $NiO/Sr_3Ti_2O_7$: Effect of Preparation Method", *International Journal of Hydrogen Energy*, Vol. 31, pp. 1142-1146.

Jitputti, J., S. Pavasupree, Y. Suzuki, S. Yoshikawa, 2007, "Synthesis and Photocatalytic Activity for Water-Splitting Reaction of Nanocrystalline Mesoporous Titania Prepared by Hydrothermal Method", *Journal of Solid State Chemistry*, Vol. 180, pp. 1743-1749.

Kanhere, P., Z. Chen, 2014, "A Review on Visible Light Active Perovskite-Based Photocatalysts", *Molecules*, Vol. 19, pp. 19995-20022.

Kanhere, P., J. Zheng, Z. Chen, 2012, "Visible Light Driven Photocatalytic Hydrogen Evolution and Photophysical Properties of Bi^{3+} doped $NaTaO_3$", *International Journal of Hydrogen Energy*, Vol. 37, pp. 4889-4896.

Kantardzic, M., *Data Mining Concepts, Models, Methods, and Algorithms*, John Wiley & Sons Inc. USA, 2003.

Khan, M. A., M. S. Akhtar, S. I. Woo, O. B. Yang, 2008, "Enhanced Photoresponse Under Visible Light in Pt Ionized TiO_2 Nanotube for the Photocatalytic Splitting of Water", *Catalysis Communications*, Vol. 10, pp. 1-5.

Khan, M. A., S. I. Woo, O. Yang, 2008, "Hydrothermally Stabilized Fe(III) Doped Titania Active Under Visible Light for Water Splitting Reaction", *International Journal of Hydrogen Energy*, Vol. 33, pp. 5345-5351.

Khan, M. A., O. Yang, 2009, "Photocatalytic Water Splitting for Hydrogen Production Under Visible Light on Ir and Co Ionized Titania Nanotube", *Catalysis Today*, Vol. 146, pp. 177-182.

Kim, J., M. Kang, 2012, "High Photocatalytic Hydrogen Production over The Band Gap-Tuned Urchin-Like Bi_2S_3-Loaded TiO_2 Composites System", *International Journal of Hydrogen Energy*, Vol. 37, pp. 8249-8256.

Kim, S. H., S. Park, S. W. Lee, B. S. Han, S. W. Seo, J. S. Kim, I. S. Cho, K. S. Hong, 2012, "Photophysical and Photocatalytic Water Splitting Performance of Stibiotantalite Type-Structure Compounds, $SbMO_4$ (M= Nb, Ta)", *International Journal of Hydrogen Energy*, Vol. 37, pp. 16895-16902.

Kimi, M., L. Yuliati, M. Shamsuddin, 2011, "Photocatalytic Hydrogen Production under Visible Light Over $Cd_{0.1}Sn_xZn_{0.9-2x}S$ Solid Solution Photocatalysts", *International Journal of Hydrogen Energy*, Vol. 36, pp. 9453-9461.

Kokporka, L., S. Onsuratoom, T. Puangpetch, S. Chavadej, 2013, "Sol-Gel Synthesized Mesoporous-Assembled TiO_2-ZrO_2 Mixed Oxide Nanocrystals and Their Photocatalytic Sensitized H_2 Production Activity under Visible Light Irradiation", *Materials Science in Semiconductor Processing*, Vol. 16, pp. 667-678.

Le, T. T., M. S. Akhtar D. M. Park J. C. Lee, O. B. Yang, 2012, "Water Splitting on Rhodamine-B Dye Sensitized Co-Doped TiO_2 Catalyst Under Visible Light", *Applied Catalysis B: Environmental*, Vol. 111-112, pp. 397-401.

Lee, C. W., D. W. Kim, I. S. Cho, S. Park, S. S. Shin, S. W. Seo, K. S. Hong, 2012, "Simple Synthesis and Characterization of SrSnO3 Nanoparticles with Enhanced Photocatalytic Activity", *International Journal of Hydrogen Energy,* Vol. 37, pp. 10557-10563.

Lee, G., M. Kang, 2013, "Physicochemical Properties of Core / Shell Structured Pyrite FeS_2 / Anatase TiO_2 Composites and Their Photocatalytic Hydrogen Production Performances", *Current Applied Physics,* Vol. 13, pp. 1482-1489.

Lee, H., Y. Park, M. Kang, 2013, "Synthesis of Characterization of Zn_xTi_yS and Its Photocatalytic Activity for Hydrogen Production from Methanol/Water Photo-Splitting", Journal of Industrial and Engineering Chemistry, Vol. 19, pp. 1162-1168.

Lee, K., W. S. Nam, G. Y. Han, 2004, "Photocatalytic Water-Splitting in Alkaline Solution Using Redox Mediator. 1: Parameter Study", *International Journal of Hydrogen Energy*, Vol. 29, pp. 1343-1347.

Lee, S. S., H. Bai, Z. Liu, D. D. Sun, 2012, "Electrospun TiO_2/SnO_2 Nanofibers with Innovative Structure and Chemical Properies for Highly Efficient Photocatalytic H_2 Generation", *International Journal of Hydrogen Energy*, Vol. 37, pp. 10575-10584.

Li, C., J. Yuan, B. Han, L. Jiang, W. Shangguan, 2010, "TiO_2 Nanotubes Incorporated with CdS for Photocatalytic Hydrogen Production from Splitting Water Under Visible Light Irradiation", *International Journal of Hydrogen Energy*, Vol. 35, pp. 7073-7079.

Li, G., T. Kako, D. Wang, Z. Zou, J. Ye, 2008, "Synthesis and Enhanced Photocatalytic Activity of $NaNbO_3$ Prepared by Hydrothermal and Polymerized Complex Methods", *Journal of Physics and Chemistry of Solids*, Vol. 69, pp. 2487-2491.

Li, J., J. Zeng, L. Jia, W. Fang, 2010, "Investigations on The Effect of Cu^{2+}/Cu^{1+} Redox Couples and Oxygen Vacancies on Photocatalytic Activity of Treated $LaNi_{1-x}Cu_xO_3$ (x = 0.1, 0.4, 0.5)", *International Journal of Hydrogen Energy*, Vol. 35, pp. 12733-12740.

Li, Q., T. Kako, J. Ye, 2011, "Facile Ion-Exchange Synthesis of Sn^{2+} Incorporated Potassium Titanate Nanoribbons and Their Visible-Light-Responded Photocatalytic Activity", *International Journal of Hydrogen Energy*, Vol. 36, pp. 4716-4723.

Li, Y., G. Chen, H. Zhang, Z. Li, 2009, "Electronic Structure and Photocatalytic Water Splitting of Lanthanum-Doped Bi_2AlNbO_7", *Materials Research Bulletin*, Vol. 44, pp. 741-746.

Li, Y., G. Chen, H. Zhang, Z. Li, 2009, "Photocatalytic Water Splitting of La_2AlTaO_7 and The Effect of Aluminum on The Electronic Structure", *Journal of Physics and Chemistry of Solids*, Vol. 70, pp. 536-540.

Li, Y., G. Chen, H. Zhang, Z. Li, J. Sun, 2008, "Electronic Structure and Photocatalytic Properties of $ABi_2Ta_2O_9$ (A = Ca, Sr, Ba)", *Journal of Solid State Chemistry*, Vol. 181, pp. 2653-2659.

Li, Y., G. Chen, H. Zhang, Z. Lv, 2010, "Band Structure and Photocatalytic Activities for H_2 Production of $ABiNb_2O_9$ (A = Ca, Sr, Ba)", *International Journal of Hydrogen Energy*, Vol. 35, pp. 2652-2656.

Li, Y., H. Gou, J. Lu, C. Wang, 2014, "A Two-Step Synthesis of $NaTaO_3$ Microspheres for Photocatalytic Water Splitting", *International Journal of Hydrogen Energy*, Vol. 39, pp. 13481-13485.

Li, Y., F. He, S. Peng, G. Lu, S. Li, S. 2011, "Photocatalytic H_2 Evolution from NaCl Saltwater Over ZnS_{1-x}-$0.5_yO_x(OH)_y$-ZnO Under Visible Light Irradiation", *International Journal of Hydrogen Energy*, Vol. 36, pp. 10565-10573.

Li, Y., Y. Huang, J. Wu, M. Huang, J. Lin, 2010, "Photocatalytic Activities for Hydrogen Evolution of New Layered Compound Series $HLaTa_{x/3}O_7$/Pt (x = 0, 2, 3, 4, and 6)", 2010, *Journal of Hazardous Material*, Vol. 177, pp. 458-464.

Li, Y., S. Jiang, J. Xiao, Y. Li, 2014, "Photocatalytic Overall Water Splitting Under Visible Light over an In-Ni-Ta-O-N Solid Solution without an Additional Cocatalyst", *International Journal of Hydrogen Energy*, Vol. 39, pp. 731-735.

Li, Y., J. Wu, Y. Huang, M. Huang, J. Lin, 2009, "Photocatalytic Water Splitting on New Layered Perovskite $A_{2.33}Sr_{0.67}Nb_5O_{14.335}$ (A = K, H)", *International Journal of Hydrogen Energy*, Vol. 34, pp. 7927-7933.

Li, Y., Y. Zhengmin, J. Meng, Y. Li, 2013, "Enhancing The Activity of a SiC-TiO_2 Composite Catalyst for Photo-Stimulated Catalytic Water Splitting", *International Journal of Hydrogen Energy*, Vol. 38, pp. 3898-3904.

Liao, C. H., C. W. Huang, J. C. S. Wu, 2013, "Visible-light-active Photocatalytic Thin Film by RF Sputtering for Hydrogen Generation", *Asia-Pacific Journal of Chemical Engineering*, Vol. 8, pp. 283-291.

Lin, H., H. Yang, W. Wang, 2011, "Synthesis of Mesoporous Nb_2O_5 Photocatalysts with Pt, Au, Cu, and NiO cocatalyst for water splitting", *Catalysis Today*, Vol. 174, pp. 106-113.

Lin, H. Y., Y. S. Chang, 2010, "Photocatalytic Water Splitting for Hydrogen Production on $Au/KTiNbO_5$", *International Journal of Hydrogen Energy*, Vol. 35, pp. 8463-8471.

Lin, Y. C., S. H. Liu, H. R. Syu, T. H. Ho, 2012, "Synthesis, Characterization and Photocatalytic Performance of Self-Assembled Mesoporous TiO_2 Nanoparticles", *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, Vol. 95, pp. 300-304.

Liu, S., H. Syu, 2013, "High Visible-Light Photocatalytic Hydrogen Evolution of C, N-Codoped Mesoporous TiO_2 Nanoparticles Prepared via an Ionic-Liquid-Template Approach", *International Journal of Hydrogen Energy*, Vol. 38, pp. 13856-13865.

Liu, S. H., H. R. Syu, 2013, "High Visible-Light Photocatalytic Hydrogen Evolution of C, N-Codoped Mesoporous TiO2 Nanoparticles Prepared via an Ionic-Liquid-Template Approach", *International Journal of Hydrogen Energy*, Vol. 38, pp. 13856-13865.

Liu, S. H., H. R. Syu, 2012, "One-Step Fabrication of N-Doped Mesoporous TiO_2 Nanoparticles by Self-Assembly for Photocatalytic Water Splitting Under Visible Light", *Applied Energy*, Vol. 100, pp. 148-154.

Liu, X. S., *Microarray Dimension Reduction*, Harvard University, Cambridge, USA, 2008.

Liu, Y., L. Xie, Y. Li, R. Yang, J. Qu, Y. Li, X. Li, 2008, "Synthesis and High Photocatalytic Hydrogen Production of $SrTiO_3$ Nanoparticles from Water Splitting Under UV Irradiation", *Journal of Power Sources*, Vol. 183, pp. 701-707.

Liu, Z., H. Bai, S. Xu, D. D. Sun, 2011, "Hierarchical CuO/ZnO "Corn-Like" Architecture for Photocatalytic Hydrogen Generation", *International Journal of Hydrogen Energy*, Vol. 36, pp. 13473-13480.

Mallick, J., P. M. Prangyan, A. Dulal, 2013, "Back Propagation Neural Network for Recognition of Dynamic Objects", *International Journal of Computer Applications in Engineering Sciences*, Vol 3, pp.141-147.

Martha, S., K. H. Reddy, K. M. Parida, P. K. Satapathy, 2012, "Enhanced Photocatalytic Activity over N-Doped GaZn Mixed Oxide Under Visible Light Irradiation", *International Journal of Hydrogen Energy*, Vol. 37, pp. 115-124.

Naik, B., S. Martha, K. M. Parida, 2011, "Facile Fabrication of $Bi_2O_3/TiO_{2-x}N_x$ Nanocomposites for Excellent Visible Light Driven Photocatalytic Hydrogen Evolution", *International Journal of Hydrogen Energy*, Vol. 36, pp. 2794-2802.

Odabaşı Ç., M. E. Günay, R. Yıldırım, 2014, "Knowledge Extraction for Water Gas Shift Reaction over Noble Metal Catalysts from Publications in The Literature Between 2002 and 2012", *International Journal of Hydrogen Energy*, Vol. 39, pp. 5733-5746.

Onsuratoom, S., S. Chavadej, T. Sreethawong, 2011, "Hydrogen Production from Water Splitting Under UV Light Irradiation over Ag-Loaded Mesoporous-Assembled TiO_2-ZrO_2 Mixed Oxide Nanocrystal Photocatalysts", *International Journal of Hydrogen Energy*, Vol. 36, pp. 5246-5261.

Oros-Ruiz, S., R. Zanella, S. E. Collins, A. Hernandez-Gordillo, R. Gomez, 2014, "Photocatalytic hydrogen production by Au-M_xO_y (M=Ag, Cu, Ni) catalysts supported on TiO_2", *Catalysis Communications*, Vol. 47, pp. 1-6.

Parayil, S. K., H. S. Kibombo, C. M. Wu, R. Peng, J. Baltrusaitis, R. T. Koodali, 2012, "Enhanced Photocatalytic Water Splitting Activity of Carbon-Modified TiO_2 Composite Materials Synthesized by a Green Synthetic Approach", *International Journal of Hydrogen Energy*, Vol. 37, pp. 8257-8267.

Parida, K. M., N. Biswal, D. P. Das, S. Martha, 2010, "Visible Light Response Photocatalytic Water Splitting over CdS-Pillared Zirconium-Titanium Phosphate (ZTP)", *International Journal of Hydrogen Energy*, Vol. 35, pp. 5262-5269.

Perez-Larios, A., R. Lopez, A. Hernandez-Gordillo, F. Tzompantzi, R. Gomez, L. M. Torres-Guerra, 2012, "Improved Hydrogen Production from Water Splitting Using TiO_2-ZnO Mixed Oxide Photocatalysts", *Fuel*, Vol. 100, pp. 139-143.

Puangpetch, T., T. Sreethawong, S. Yoshikawa, S. Chavadej, 2009, "Hydrogen Production from Photocatalytic Water Splitting over Mesoporous-Assembled $SrTiO_3$ Nanocrystal-Based Photocatalysts", *Journal of Molecular Catalysis A: General*, Vol. 312, pp. 97-106.

Rayalu, S. S., D. Jose, M. V. Joshi, P. A. Mangrulkar, K. Shrestha, K. Klabunde, 2013, "Photocatalytic Water Splitting on Au/TiO_2 Nanocomposites Synthesized Through Various Routes: Enhancement in Photocatalytic Activitiy due to SPR Effect", *Applied Catalysis B: Environmental*, Vol. 142-143, pp. 684-693.

Rosseler, O., M. V. Shankar, M. K. Du, L. Schmidlin, N. Keller, V. Keller, 2010, "Solar Light Photocatalytic Hydrogen Production from Water over Pt and Au/TiO_2 (anatase/rutile) photocatalysts: Influence of Noble Metal and Porogen Promotion", *Journal of Catalysis*, Vol. 269, pp. 179-190.

Rungjaroentawon, N., S. Onsuratoom, S. Chavadej, 2012, "Hydrogen Production from Water Splitting Under Visible Light Irradiation Using Sensitized Mesoporous-Assembled TiO_2-SiO_2 Mixed Oxide Photocatalysts", *International Journal of Hydrogen Energy*, Vol. 37, pp. 11061-11071.

Sadhu, T., *Machine Learning: Introduction to the Artificial Neural Network,* 2012, http://www.durofy.com/machine-learning-introduction-to-the-artificial-neural-network/, [Accessed June 2015].

Sathish, M., R. P. Viswanath, 2007, "Photocatalytic Generation of Hydrogen over Mesoporous CdS Nanoparticle: Effect of Particle Size, Noble Metal and Support", *Catalysis Today,* Vol. 129, pp. 421-427.

Seni, G., *How to Create Predictive Models in R Using Ensembles,* Santa Clara University, New York, USA, 2013.

Serrano, D. P., G. Calleja, P. Pizarro, P. Galvez, 2014, "Enhanced Photocatalytic Hydrogen Production by Improving the Pt Dispersion over Mesostructured TiO_2", *International Journal of Hydrogen Energy,* Vol. 39, pp. 4812-4819.

Shen, P., J. C. Lofaro Jr. W. R. Woerner, M. G. White, D. Su, A. Orlov, 2013, "Photocatalytic Activity of Hydrogen Evolution over Rh doped $SrTiO_3$ Prepared by Polymerizable Complex Method", *Chemical Engineering Journal*, Vol. 223, pp. 200-208.

Shibata, M., A. Kudo, A. Tanaka, K. Domen, K. Maruya, T. Onishi, T. 1987, "Photocatalytic Activities of Layered Titanium Compounds and Their Derivatives for H_2 Evolution from Aqueous Methanol Solution", *Chemistry Letters*, Vol. 16, pp. 1017-1018.

Sorzano C. O. S., 2014, *Bioengineering Laboratory*, http://biolab.uspceu.com/ index.php, [Accessed July 2015].

Sreethawong, T., Y. Suzuki, S. Yoshikawa, 2005, "Photocatalytic Evolution of Hydrogen over Mesoporous TiO_2 supported NiO Photocatalyst Prepared by Single-Step Sol-Gel Process with Surfactant Template", *International Journal of Hydrogen Energy*, Vol. 30, pp. 1053-1062.

Sreethawong, T., S. Yoshikawa, 2006, "Enhanced Photocatalytic Hydrogen Evolution over Pt Supported on Mesoporous TiO2 Prepared by Single-Step Sol-Gel Process with Surfactant Template", ", *International Journal of Hydrogen Energy*, Vol. 31, pp. 786-796.

Sun, T., E. Liu, J. Fan, X. Hu, F. Wu, W. Hou, Y. Yang, L. Kang, 2013, "High Photocatalytic Activity of Hydrogen Production from Water over Fe doped and Ag Deposited Anatase TiO2 Catalyst Synthesized by Solvothermal Method", *Chemical Engineering Journal*, Vol. 228, pp. 896-906.

Sun, T., J. Fan, E. Liu, L. Liu, Y. Wang, H. Dai, Y. Yang, W. Hou, X. Hu, Z. Jiang, 2012, "Fe and Ni Co-Doped TiO_2 Nanoparticles Prepared by Alcohol-Thermal Method: Application in Hydrogen Evolution by Water Splitting Under Visible Light Irradiation", *Powder Technology*, Vol. 228, pp. 210-218.

Torres-Martinez, L. M., R. Gomez, O. Vazquez-Cuchillo, I. Juarez-Ramirez, A. Cruz-Lopez, F. J. Alejandre-Sandoval, 2010, "Enhanced Photocatalytic Water Splitting Hydrogen Production on RuO_2/La:$NaTaO_3$ Prepared by Sol-Gel Method", *Catalysis Communications*, Vol. 12, pp. 268-272.

Wang, C., Q. Hu, J. Huang, L. Wu, Z. Deng, Z. Liu, Y. Liu, Y. Cao, 2013, "Efficient hydrogen production by photocatalytic water splitting using N-doped TiO_2 film", *Applied Surface Science*, Vol. 283, pp. 188-192.

Wang, L., W. Wang, 2012, "Photocatalytic Hydrogen Production from Aqueous Solutions over Novel $Bi_{0.5}Na_{0.5}TiO_3$ Microspheres", *International Journal of Hydrogen Energy*, Vol. 37, pp. 3041-3047.

Wang, Q., N. An, Y. Bai, H. Hang, J. Li, X. Lu, Y. Liu, F. Wang, Z. Li, Z. Lei, 2013, "High Photocatalytic Hydrogen Production from Methanol Aqueous Solution Using the Photocatalysts CuS/TiO_2", *International Journal of Hydrogen Energy*, Vol. 38, pp. 10739-10745.

Wang, Q., J. Li, N. An, Y. Bai, X. Lu, J. Li, H. Ma, R. Wang, F. Wang, Z. Lei, W. Shangguan, 2013, "Preparation of a Novel Recyclable Cocatalyst Wool-Pd for Enhancement of Photocatalytic H_2 Evolution on CdS", *International Journal of Hydrogen Energy*, Vol. 38, pp. 10761-10767.

Wang, X, G. Liu, Z. G. Chen, F. Li, A. Q. M. Lu, H. M. Cheng, 2009, "Efficient and Stable Photocatalytic H2 Evolution from Water Splitting by ($Cd_{0.8}Zn_{0.2}$) Nanorods", *Electrochemistry Communications*, Vol. 11, pp. 1174-1178.

Wei, P., J. Liu, Z. Li, 2013, "Effect of Pt Loading and Calcination Temperature on the Photocatalytic Hydrogen Production Activity of TiO_2 Microspheres", *Ceramics International*, Vol. 39, pp. 5387-5391.

Wei, Y., J. Li, Y. Huang, M. Huang, J. Lin, J. Wu, 2009, "Photocatalytic Water Splitting with In-Doped H2LaNb2O7 Composite Oxide Semiconductors", *Solar Energy Materials & Solar Cells*, Vol. 93, pp. 1176-1181.

Wilcox, W. R., *Explanation of Results Returned by The Regression Tool in Excel's Data Analysis*, Clarkson University, 2010, http://people.clarkson.edu/~wwilcox/ES100/regrint.htm, [Accessed June 2015].

Witten, I. H., E. Frank, M. A. Hall, *Data Mining Practical Machine Learning Tools and Techniques*, Elsevier Inc, Burlington, USA, 2011.

Wu, P., J. Shi, Z. Zhou, W. Tang, L. Guo, 2012, "$CaTaO_2N$-$CaZrO_3$ Solid Solution: Band Structure Engineering and Visible-Light-Driven Photocatalytic Hydrogen Production", *International Journal of Hydrogen Energy*, Vol. 37, pp. 13704-13710.

Wu, Z., G. Li, F. Zhang, W. Zhang, 2014, "Photocatalytic Activity of $NaTaO_3$:La Prepared Under Different Atmospheres", *Applied Surface Science*, Vol. 319, pp. 372-375.

Yamada M., K. Omata, Y. Watanabe, M. Hashimoto, T. Umegaki, 2004, "Simultaneous Optimization Composition of the Methanol Synthesis Catalyst by an All-Encompassing Calculation on an Artificial Neural Network", *Industrial & Engineering Chemistry Research*, Vol. 43, pp. 3282-3288.

Yan L., J. Zhang, X. Zhou, X. Wu, J. Lan, Y. Wang, G. Liu, J. Yu, L. Zhi, 2013, "Crystalline Phase-Dependent Photocatalytic Water Splitting for Hydrogen

Generation on $KbNO_3$ Submicro-Crystals", *International Journal of Hydrogen Energy*, Vol. 38, pp. 3554-3561.

Yan, S. C., Z. Q. Wang, Z. S. Li, Z. G. Zou, 2009, "Photocatalytic Activities for Water Splitting of La-Doped-$NaTaO_3$ Fabricated by Microwave Synthesis", *Solid State Ionics*, Vol. 180, pp. 1539-1542.

Yao, X., T. Liu, X. Liu, L. Lu, 2014, "Loading of CdS Nanoparticles on The (101) Surface of Elongated TiO2 Nanocrystals for Efficient Visible-Light Photocatalytic Hydrogen Evolution from Water Splitting", *Chemical Engineering Journal*, Vol. 255, pp. 28-39.

Yoshida, H., M. Takeuchi, M. Sato, L. Zhang, T. Teshima, M. G. Chaskar, 2014, "Potassium Hexatitanate Photocatalysts Prepared by a Flux Method for Water Splitting", *Catalysis Today*, Vol. 232, pp. 158-164.

Yoshioka, K., V. Petrykin, M. Kakihana, H. Kato, A. Kudo, 2005, "The Relationship between Photocatalytic Activity and Crystal Structure in Strontium Tantalates", *Journal of Catalysis*, Vol. 232, pp. 102-107.

Yu, Z., J. Meng, Y. Li, 2013, "Efficienct Photocatalytic Hydrogen Production from Water over a CuO and Carbon Fiber Comodified TiO_2 Nanocomposite Photocatalyst", *International Journal of Hydrogen Energy*, Vol. 38, pp. 16649-16655.

Yuret, D., *Machine Learning in 10 Pictures*, 2014, http://www.denizyuret.com/2014/02/machine-learning-in-5-pictures.html, [Accessed June 2015].

Zavyalova U., M. Holena, R. Schlögl, M. Baerns, 2011, "Statistical Analysis of Past Catalytic Data on Oxidative Methane Coupling for New Insights into The Composition of High-Performance Catalysts", *ChemCatChem*, Vol. 3, pp. 1935-1947.

Zhang, G., W. Zhang, D. Minakata, Y. Chen, J. Crittenden, P. Wang, P. 2013, "The pH Effects on H_2 Evolution Kinetics for Visible Light Water Splitting Over the

Ru/$(CuAg)_{0.15}In_{0.3}Zn_{1.4}S_2$ Photocatalyst", *International Journal of Hydrogen Energy*, Vol. 38, pp. 11727-11736.

Zhang, H., G. Chen, X. Li, Q. Wang, Q. 2009, "Electronic Structure and Water Splitting Under Visible Light Irradiation of $BiTa_{1-x}Cu_xO_4$ (x = 0.00 - 0.004) Photocatalysts", *International Journal of Hydrogen Energy*, Vol. 34, pp. 3631-3638.

Zhang, K., Z. Zhou, L. Guo, 2011, "Alkaline Earth Metal as a Novel Dopant for Chalcogenide Solid Solution: Improvement of Photocatalytic Efficiency of $Cd_{1-x}Zn_xS$ by Barium Surface Doping", *International Journal of Hydrogen Energy*, Vol. 36, pp. 9469-9478.

Zhang, X., Y. Sun, X. Cui, Z. Jiang, 2012, "A Green and Facile Synthesis of TiO2/Graphene Nanocomposites and Their Photocatalytic Activity for Hydrogen Evolution", *International Journal of Hydrogen Energy*, Vol. 37, pp. 811-815.

Zhang, Y. J., L. C. Liu, D. P. Chen, 2013, "Synthesis of CdS / Bentonite Nanocomposite Powders for H_2 Production by Photocatalytic Decomposition of Water", *Powder Technology*, Vol. 241, pp. 7-11.

Zhao, Y., P. Chen, B. Zhang, D. S. Su, S. Zhang, L. Tian, J. Lu, Z. Li, X. Cao, B. Wang, M. Wei, D. G. Evans, X. Duan, 2012, "Highly Dispersed TiO_6 Units in a Layered Double Hydroxide for Water Splitting", *Chemistry A European Journal*, Vol. 18, pp. 11949-11958.

Zheng, H., H. Yong, T. Ou-Yang, Y. Fan, H. Hou, 2013, "A New Photosensitive Coordination Compound $[RuL(bpy)_2](PF_6)_2$ and its Application in Phototocatalytic H_2 Production Under the Irradiation of Visible Light", *International Journal of Hydrogen Energy*, Vol. 38, pp. 12938-12945.

Zhou, C., G. Chen, Y. Li, H. Zhang, J. Pei, 2009, "Photocatalytic Activities of $Sr_2Ta_2O_7$ Nanosheets Synthesized by a Hydrothermal Method", *International Journal of Hydrogen Energy*, Vol. 34, pp. 2113-2120.

Zhou, C., G. Chen, Q. Wang, 2011, "High Photocatalytic Activity of Porous $K_4Nb_6O_{17}$ Microsphere with Large Surface Area Prepared by Homogeneous Precipitation Using Urea", *Journal of Molecular Catalysis A: Chemical*, Vol. 339, pp. 37-42.

Zielinska, B., E. Mijowska, R. J. Kalenczuk, 2012, "Synthesis, Characterization and Photocatalytic Properties of Lithium Tantalate", *Materials Characterization,* Vol. 68, pp.71-76.