Research Article / Araştırma Makalesi GU J Sci, Part A, 8(4): 435-450 (2021) https://doi.org/10.54287/gujsa.997123

Gazi University
Journal of Science
PART A: ENGINEERING AND INNOVATION
http://dergipark.org.tr/gujsa

Data Mining and Application of Decision Tree Modelling on Electrochemical Data Used for Damaged Starch Detection

Nilufer YILDIRIM[1*], Niyazi Alper TAPAN[1]

[1]*Department of Chemical Engineering, Faculty of Engineering, Ankara, Turkey*

Keywords

Decision Tree
Damaged Starch
Electrochemical
Machine Learning

Abstract

In this study, unsupervised and supervised machine learning techniques, principal component analysis and classification tree modelling which could be improved with additional input variables were applied on iodine oxidation voltammetric data in order to determine routes and extract information about the electrochemical conditions leading to different damaged starch ratios in flour. For this purpose a database of 3542 observations which was normalized and filtered from outliers was used. It was seen that although it was almost impossible to generalize information or determine correlations from voltammetric data at different conditions, principal component analysis indicate that on platinum electrode UCD values of 16.5 mostly seen at high potentials, optimized decision tree indicate that the impact of variables on UCD values can be ordered as current density > potential > electrode type > KI concentration and give routes to UCD values with high class membership leaf nodes. Therefore machine learning with decision tree modelling could open perspectives for practical and fast prediction of damaged starch ratio which would help food industry to speed up and economize costs for analysis in flour.

Cite

Yıldırım, N. & Tapan, N. A. (2021). Data Mining and Application of Decision Tree Modelling on Electrochemical Data Used for Damaged Starch Detection. *GU J Sci, Part A, 8*(4), 435-450.

Author ID (ORCID Number)

N. Yıldırım, 0000-0003-0666-3182
N. A. Tapan, 0000-0001-8599-0450

Article Process

Submission Date 17.09.2021
Revision Date 10.11.2021
Accepted Date 28.11.2021
Published Date 01.12.2021

1. INTRODUCTION

In the world and in our country, most important flour based nutritional products like bread are brought to our daily lives by highly technological processes. In addition, there is an increasing demand on high quality food products and on development of standards every day. Therefore, it is utmost important to launch food product on the market from correctly processed flour.

During industrial flour production, one of the most important processing step is milling. In the milling process, the endosperm (nutritional tissue) is separated from the wheat bran and then the endosperm is ground as flour. The grinding process takes place in two stages as on gear and flat steel rollers, and the proper adjustment of the flat rolls during the grinding stage determines the damaged starch ratio in the ground wheat (Liu et al., 2017). If the chemical structure of starch which is a critical component and quality indicator of flour is analyzed, it is seen that starch is composed of straight and branched chain structures as amylose and amylopectin. Studies demonstrate that the application of mechanical energy or milling process deteriorate mostly the crystal structure of amylopectin rather than straight chain amylose molecules (Dhital et al., 2011; Li et al., 2014). The change of amylopectin / amylose ratio during wheat milling also affects the texture of the final product. In general, the damage of the starch structure during the milling of wheat is not only due to the size of the molecules but also depends on the branching structure. High branching density and short branch length in starch is also more prone to damage. A study on corn starch has shown that; the amylopectin is damaged more than amylose during milling not only due to the size of molecule but more short branches of

*Corresponding Author, e-mail: nilufer.yildirim@gazi.edu.tr

amylopectin (Liu et al., 2010). In addition, studies on types of starches with different amounts of amylopectin (99%, 75%, 20%) indicate that, after 15 min milling process of starch, damaged starch ratio reaches to 40% for 99% amylopectin ratio and the amylose ratio increase (Liu et al., 2017). Therefore, from the structural point of view, high damaged starch ratio which is affected by milling process time means high amylose content in the final wheat product. Beside the effect on the chemical structure, since the high or low proportion of damaged starch could cause quality defects like low volume and low viscosity in dough due to inefficient amylolytic activity during and after grinding process, fast and accurate determination and continuous monitoring of damaged starch in flour from a wide variety of wheat is very important for the desired end product (Zhu, 2016; Liu et al., 2017).

Today, for the determination of damaged starch, amperometric (or iodimetric) method (Medcalf and Gilles principle) which is based on the amount of iodine absorbed by starch granules is used (Medcalf & Gilles, 1965). In the amperometry technique which is based on iodine absorption kinetics and which the iodine absorption index is recorded as units of UCD (Chopin Dubois Unit (UCD) scales the Ai % (percentage of iodine absorption)), residual current after iodine fixing on the damaged starch is determined which takes about 7 to 10 minutes in a 120 ml electrolyte volume. Although the amperometry procedure and technique is well established, continuous and frequent analysis of damaged starch by this conventional method could bring burden of cost in the food industries in the world. At this point, in order to decrease analysis time and the amount of chemicals used for determination damaged starch during amperometry, it could be possible to use machine learning techniques based on the electrochemical data acquired by short electrochemical experiments like cyclic voltammetry of iodine oxidation in flour containing electrolyte. By the extraction of cyclic voltametric data performed on different type of electrodes, in different electrolyte concentrations and with different type of flours (with different UCD values), the degree of iodine absorption or UCD value can be modelled by supervised machine learning techniques like decision trees to determine routes and conditions leading to different damaged starch ratios. In addition, supervised learning procedure could help researchers to analyze damaged starch content with different experimental conditions easily and quickly.

Therefore, in this study, it is aimed to apply a decision tree machine learning strategy on determination of UCD values by selection of electrochemical features like electrode type, iodine concentration in the electrolyte, oxidation current density and applied potential. It is believed that machine learning strategy in this study can be extended and developed with the addition of more experimental variables like the wheat type or with different electrochemical techniques to increase the applicability or performance of the decision tree model.

2. MATERIAL AND METHOD

In this study, cyclic voltammetry technique in a three-electrode system were used to extract electrochemical data for decision tree analysis. During cyclic voltammetry, two different working electrodes as polycrystalline platinum (Pt) and glassy carbon (GC) discs, platinum counter electrode and standard calomel reference electrode were used for iodine oxidation in a three-electrode set up. For the electrolyte, sulphuric acid and 1.5 gr of boric acid (Merck Inc.) as an emulsifier, 1.5gr KI (Merck Inc.) as and iodine source and 60 ml of deionized water as a solvent were used in three neck glass electrochemical cell. In order to prevent formation of elemental iodine 1 drop of 0.1 M sodium thiosulphate (Merck Inc.) was added to the mixture. After stirring the electrolyte for 5 min, 15 ml electrolyte was transferred to a three-neck glass cell. Voltammetric scans were performed with and without flour addition to the electrolyte. After the first scan at 50 mV/sec between 0-1 V vs. SCE in flour free electrolyte, 0.5 gr of flour was added to the electrolyte and voltammetric scan was repeated. Cyclic voltammetric experiments were performed with different KI concentrations as 0.15M and 0.075M in the electrolyte with flour samples of three different UCD values of 16.5, 25 and 30 as well. Flour samples with different UCD values were collected from the local flour factories in central Anatolia region in Turkey. For the computations in machine learning only forward scan between 0-1V vs. SCE at different UCD values, KI concentrations and with different electrode types were used.

2.1. Principle Component Analysis (PCA) of Voltammetry Data

As a famous type of unsupervised machine learning where there is no training of the model with observed data, exploratory analysis and dimensionality reduction helps researchers to propose trends and provide initial insights in the data with less number of variables than during observations. At this point techniques like PCA

which is based on the covariance or correlation matrix can be used to assign latent variables that are linearly related with original variables with maximum variance (Comon, 1994).

For PCA analysis, initially a dataset of n observations with p numerical input variables is needed. During PCA, linear combinations of p vectors that would give us maximum variance (how wide the experimental data is distributed around mean value) is searched. The variance of linear combinations of p vectors is represented by a'Sa where a is a coefficient vector with dimensions p, a' is transpose of a and S is the covariance matrix (joint variability) between each pair of input variables which should be maximized (Jolliffe & Cadima, 2016).

To allow for interpretation of voltammetry data, we have taken into account 4 input variables of potential, ln(1+current density), KI concentration and type of electrode (0 for glassy carbon, 1 for platinum electrode). Therefore, PCA identified new variables, principal components as linear combinations of these four electrochemical variables. In Matlab environment "score" together with "pca" command give principal components by centering each input variable to zero average level. For instance one principle component can be expressed as a_{11}*potential$_{centered}$ + a_{12}*ln(1+current density)$_{centered}$ + a_{13}*KI concentration$_{centered}$ + a_{14}*presence of Pt electrode$_{centered}$, where a_{11}, a_{12}, a_{13}, a_{14} are coefficients for the first principal component (Ringnér, 2008). After determination of the percentage of variances, principal components with highest variances were used for explanatory analysis of the voltammetric data.

2.2. Construction of Decision Tree

A decision tree is a model that shows routes to output by the help of input variables or features partitioned and predicts new data based on the trained tree. In decision tree, the subsets that are formed in the leaves of the tree by splitting observed data set should have a desired purity. In order to have a desired purity the three main splitting criteria used in decision tree are information gain, gini index and node error. In the decision tree the partition value is selected based on maximum information gain, minimum gini index or minimum node error. Information gain is given in equation 1.

$$\text{InfoGain} = \text{Info}(\text{Parent node}) - \sum_k(p_k)\text{Info}(\text{Childnode}_k) \quad (1)$$

Info and p_k in equation 1 is the information of the feature subspace (node), and the proportion of samples passed to the *k*th subspace (or node) as given in equation 2.

$$Info = -\sum_k \left(\frac{N_j(t)}{N(t)}\right) \cdot \ln \frac{N_j(t)}{N(t)} \quad (2)$$

In equation 2 where *N(t)* is the number of samples in node t, and $N_j(t)$ is the number of class *j* samples in node *t*. The other split criterion Gini index is an indication of node impurity as given in equation 3.

$$1 - \sum_i p^2(i) \quad (3)$$

In equation 3 above, the sum is over the classes *i* at the node, and *p(i)* is the observed fraction of classes with class *i* that reach the node. A node with just one class (a pure node) has Gini index 0; otherwise, the Gini index is positive, therefore minimum Gini index is desired for node purity. And finally, node error shows the fraction of misclassified observations at the node. *p(j)* in equation 4 below is the observed fraction of largest class (with largest number of observations).

$$1 - p(j) \quad (4)$$

In Matlab environment, other than Gini (diversity) index, two different split criterions 'twoing' and 'deviance' are used similar to equations 1-4. The deviance of a node is given in equation 5 below where a pure node has deviance 0; otherwise, the deviance is positive.

$$-\sum_k \left(\frac{N_j(t)}{N(t)}\right) \cdot ln \frac{N_j(t)}{N(t)} \quad (5)$$

As given in equation 6, twoing is expressed in terms of *L(i)* which denotes the fraction of members of class *i* in the left child node after a split, and *R(i)* which denotes the fraction of members of class *i* in the right child node after a split and *P(L)* and *P(R)* are the fractions of observations that are split to the left and right leaf nodes respectively. It is desired to maximize twoing to make child node purer.

$$P(L)P(R)(\sum_i |L(i) - R(i)|)^2 \quad (6)$$

A very important point about the splitting criterions mentioned above is that the choice of the criterion may affect the choice of best features for the root or branch nodes which may lead to different decision trees. Therefore, it is necessary to compare decision tree models from different scoring criteria and to identify certain features with more or less significance.

Since the decision tree algorithm is a recursive technique, running of algorithm continues until the selected criterion is optimized by selecting optimum partition of predictors (input variables) (Quinlan, 1986). Of course, there are different algorithms like ID3, CART, C4.5, C5.0 etc. in the computing history used for decision trees with different advantages of high classifying speed, strong learning ability and simple construction (Breiman et al., 1984). Although different algorithms exist, researchers still face difficulties during modeling like low accuracy and try to make improvements on the existing algorithms (Han et al., 2011). Matlab software environment (8.4.0.150421 (R2014b)) uses non-parametric CART (Classification and regression tree) algorithm (Breiman et al., 1984). This algorithm can canconstruct binary classification trees for categorical output variables. CART algorithm can be represented as a flow diagram as seen in Figure 1. In Matlab environment (8.4.0.150421 (R2014b)), binary decision trees for classification are built using "fitctree" command.

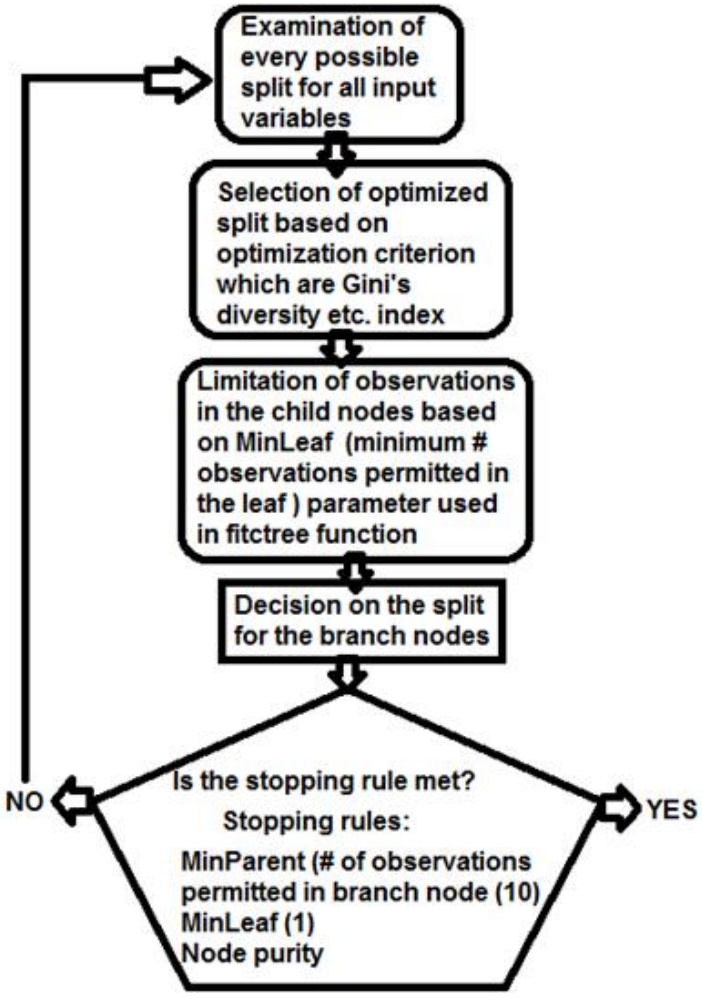


Figure 1. *Simple Representation of CART Algorithm in MATLAB Environment*

In this study, binary classification tree was used for the determination of routes to UCD values with potential, current density, KI concentration and type of electrode (platinum or glassy carbon disc electrode) as input variables. For the decision tree, a testing procedure was used to observe the accuracy of the model. First the database was randomly separated into training and testing subsets after shuffling the observations of the database randomly. Shuffling of observations was done in order to sample observations from the database in a wide range of input variables during training of the decision tree. The testing subset included 10% of the total number of observations. After the selection of splitting criterion, testing and training errors were compared at different tree sizes by the pruning the decision tree incrementally. The pruned tree size with the minimum testing error was selected as the optimal tree (Larose & Larose, 2014; Tapan et al., 2016; Baysal et al., 2017; Günay et al., 2018). Since it is known that the choice of optimization criterion may affect the choice of best features for the root or branch nodes which lead to different decision trees and therefore may affect the testing error, the testing procedure mentioned in Figure 1 was repeated to compare testing accuracy with different splitting criterion (Myles et al., 2004).

3. RESULTS AND DISCUSSION

3.1. Cyclic Voltammetry

As can be seen from the cyclic voltammetry experiments in Figure 2, oxidation and reduction peaks corresponding to triiodide/iodide redox couple (Boschloo & Hagfeldt, 2009) as given by the one step electrochemical reaction below in equation 7 appear close to 0.45 V and 0.25 V vs. SCE.

$$I_3^- + 2e^- \rightarrow 3I^- \quad (7)$$

Upon performing cyclic voltammetric experiments on 2 different types of electrodes, in two different electrolyte concentrations and with three different values of UCD, it is clearly seen that it is not possible to separate and analyze the effect of electrochemical features on current density or applied potential or to observe electrochemical conditions leading to different UCD values.

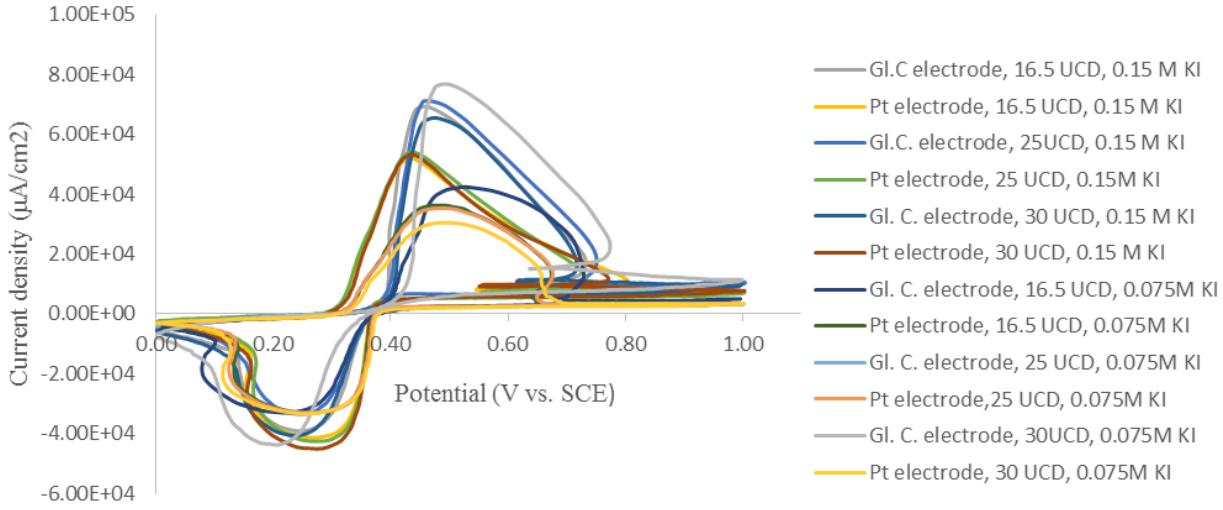


Figure 2. *Cyclic Voltammetry Experiments Performed with Pt and GC Disc Electrodes in Electrolytes with 0.075 M and 0.15M KI and with Flours of 16.5, 25 and 30 UCD Values (Scan Rate:50 mV/sec)*

For the purpose of extracting some general information and better visualization of electrochemical data used for computational purposes, voltammetry experiments were also classified with respect electrode type, UCD values and KI concentration in the electrolyte by the use of data visualization tool, ggplot2 package in R environment as seen in Figure 3. It was seen that although it not possible to differentiate UCD values in voltammetric data , higher oxidation peak current densities were observed on glassy carbon electrode and drop in KI concentration decreases peak oxidation currents for UCD values of 16.5 and 25 on both electrodes and secondly at UCD value of 30 highest oxidation peak current densities were seen on GC electrode which is just

the opposite of platinum electrode and which may indicate the different oxidation mechanisms on the two electrodes.

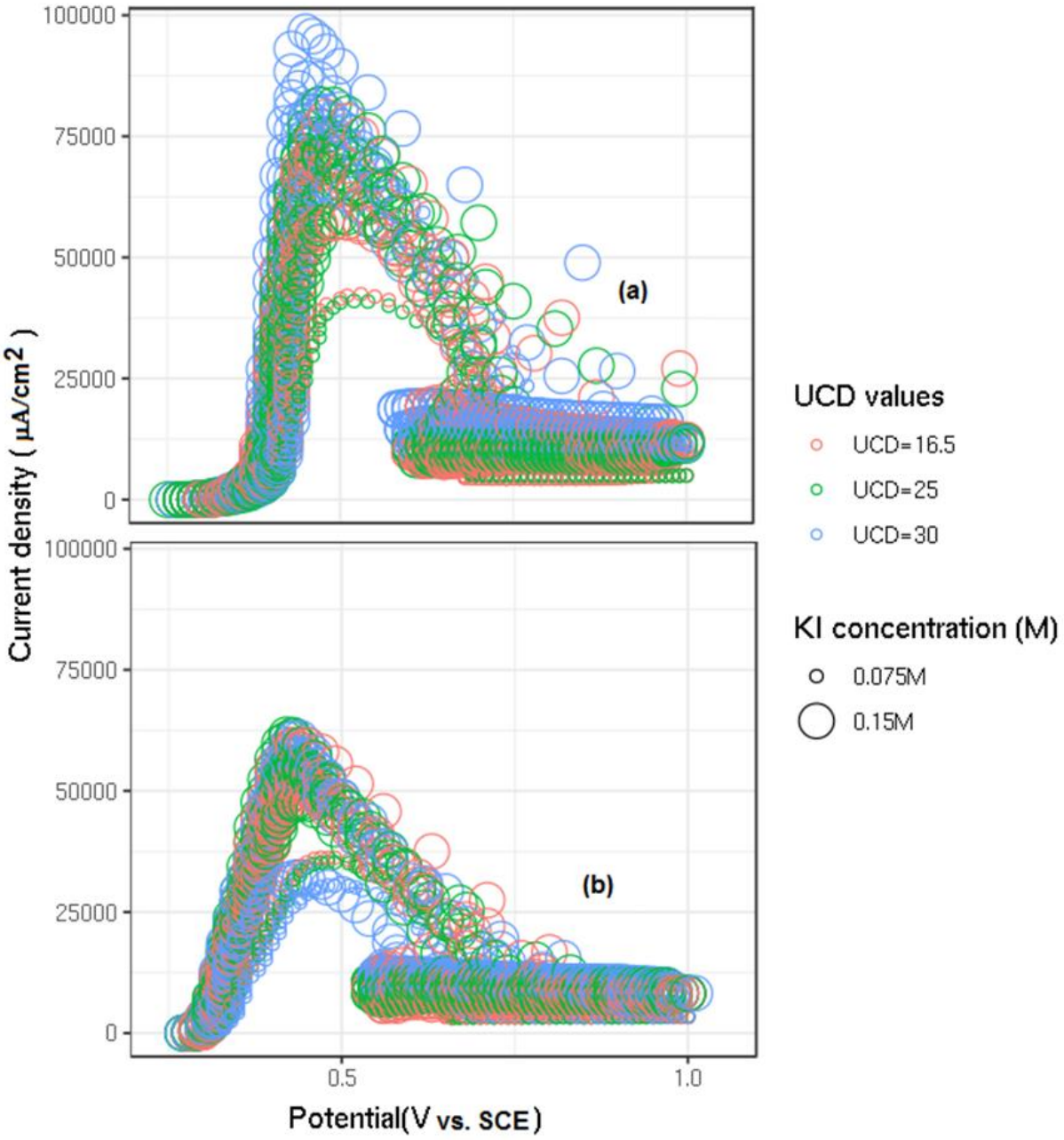
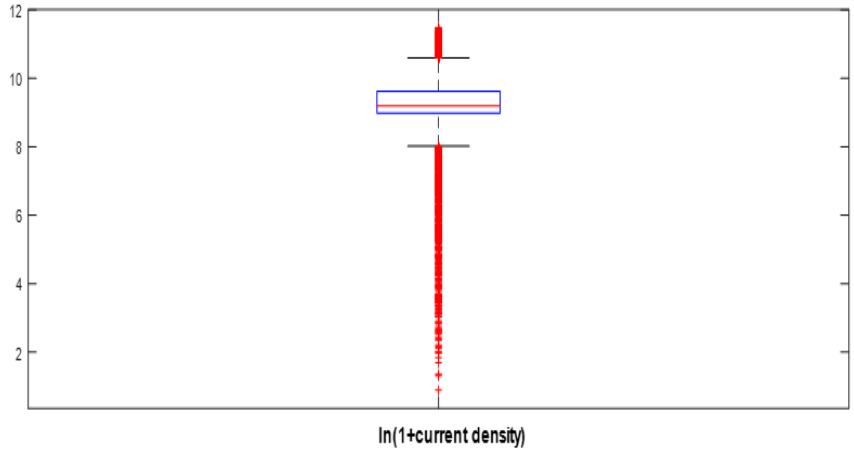


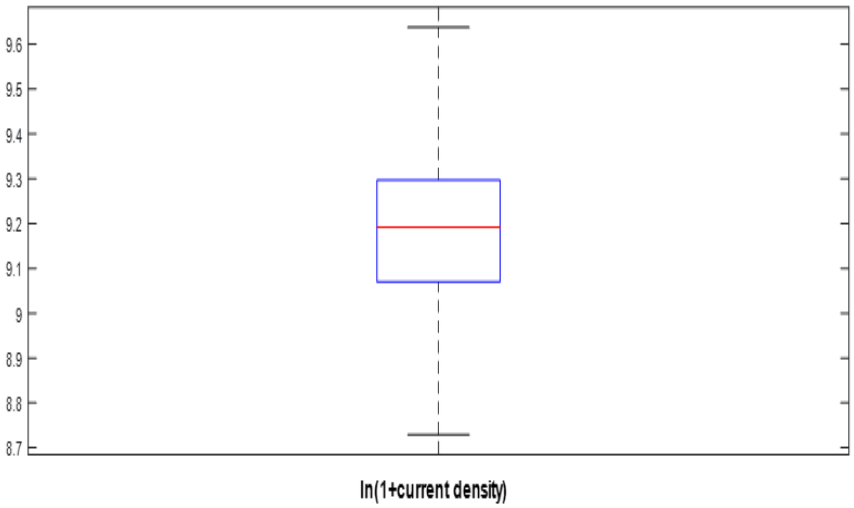
Figure 3. *Visualization of Voltammetric Experiments by Classifying with Respect to Electrode Type, UCD Values and KI Concentration* ***a)*** *glassy carbon electrode,* ***b)*** *platinum electrode (Scan Rate:50 mV/sec)*

3.2. Construction, Filtering and Analysis of Electrochemical Database

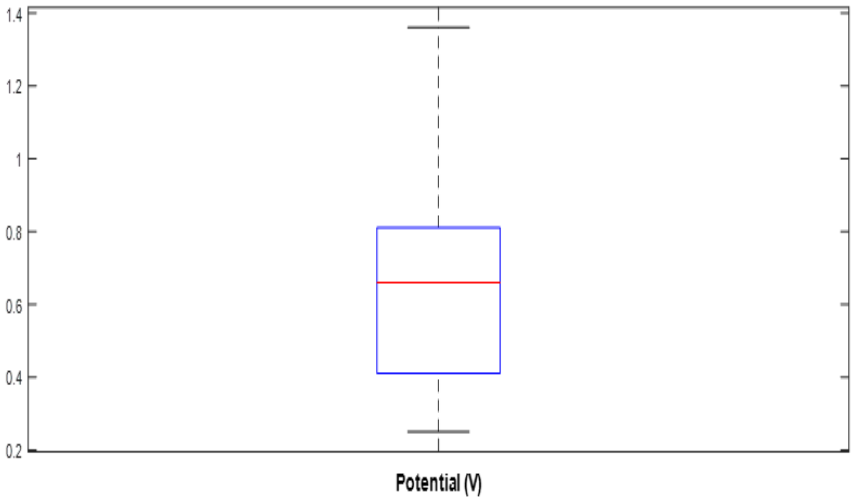
After cyclic voltammetry experiments, voltammetric data in the forward potential scan in range of 0 to 1V vs. SCE was filtered from outliers by analyzing box and whisker plots. In addition, in order to remove outliers from the voltammograms, Matlab code (MATLAB environment, version number 8.4; R2014b) was used and box and whisker plots (Walpole et al., 2012) before and after removal of outliers from the voltammetric data were compared as seen in Figure 4. After removal outliers from current density data, a filtered database of 3542 observations was used for the construction of decision trees. In Figure 4 natural logarithm transformation of current density data (ln (1+current density)) was performed in order to normalize data as much as possible after filtering from outliers.



(a)

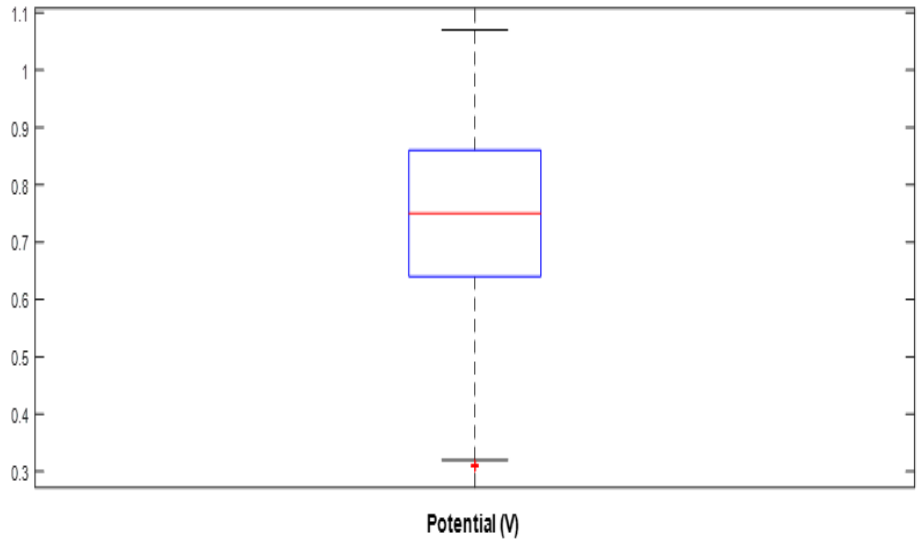


(b)

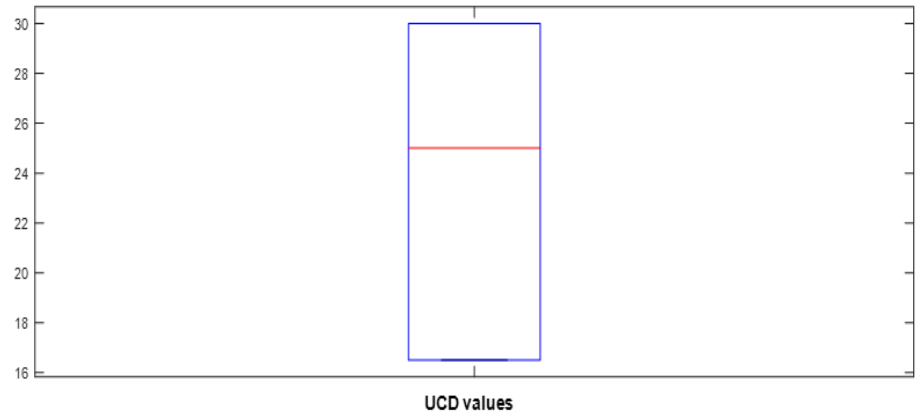


(c)

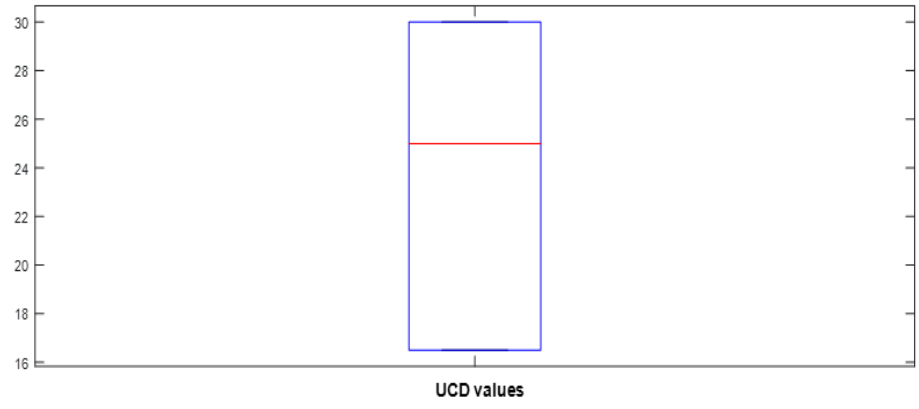
Figure 4. *Box and Whisker Plots of Electrochemical Features Before and After Removal of Outlier from the Current Density Data. Current Density Analysis were performed as the Natural Logarithm of (1+current density) [ln (1+current density)]* ***a)*** *before,* ***b)*** *after potential;* ***c)*** *before* ***d)*** *after UCD values;* ***e)*** *before,* ***f)*** *after removal of outliers*



(d)



(e)



(f)

Figure 4. *Continued*

It can be seen from Figure 4 that after removal of outliers which are the data points outside the whiskers, current density data was also normally distributed since the line in the box which shows the median of observations is almost in the middle of the box and the position of the box is almost in the center of upper and lower whiskers. Like current density, almost the similar behavior was observed for potential and UCD values. Therefore, after removal of outliers, no skewness in the filtered observations were observed.

In addition to filtering of electrochemical data, in order to search for any kind of correlation between each electrochemical feature, correlation matrix was constructed. Correlation coefficient matrix as dimensionless measure of linear dependence depends on the covariances (which is the joint variability between two random features) between any two features in the data set where covariance is divided by the product of standard deviation of selected two variables. The correlation matrix chart as shown in Figure 5 was constructed in Python 3 environment by importing "pandas" and "matplotlib" packages and using "corr ()" function.

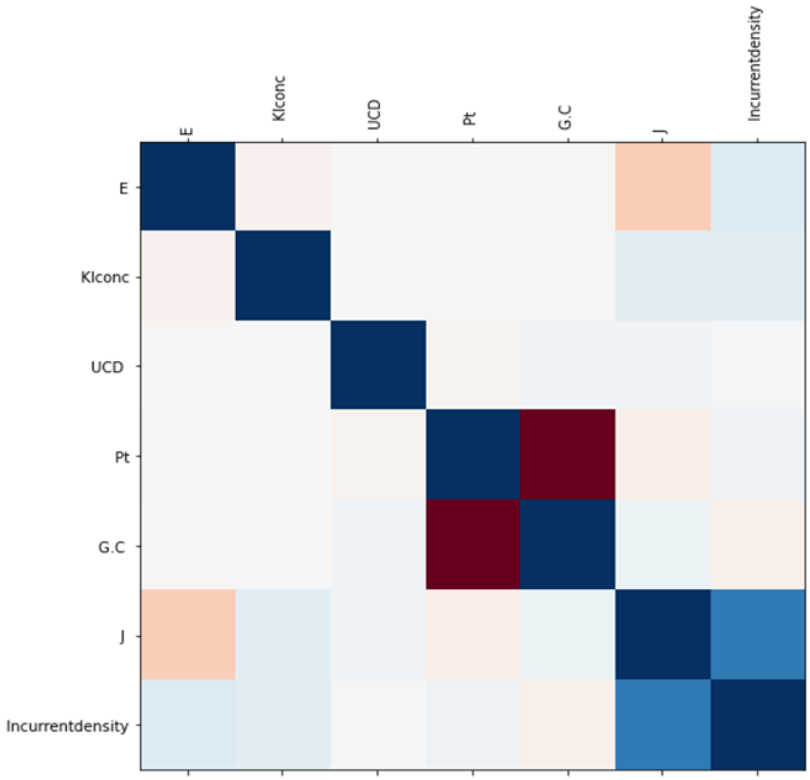


Figure 5. *Correlation Coefficient Matrix Chart between Features in the Electrochemical Data Set (**red box:** negative correlation, **blue box:** positive correlation)*

As seen in Figure 5, there is no strong correlation other than between current density and applied potential and other than a weak positive correlation between current density and KI concentration, there is almost no correlation between UCD values and other features.

3.3. Principal Component Analysis

In order to see the effect of each principal component in total variance (eigen values of covariance matrix) and to decide how many principal components (PC) to keep for explanatory analysis, "explained" together with "pca" was used to determine percent variance of each principal component as seen in Table 1.

Table 1. *Percentage of the Total Variance Explained by Each Principal Component*

Principal Component	Variance
PC1	**82.5**
PC2	**10.1**
PC3	**7.3**
PC4	**0.08**

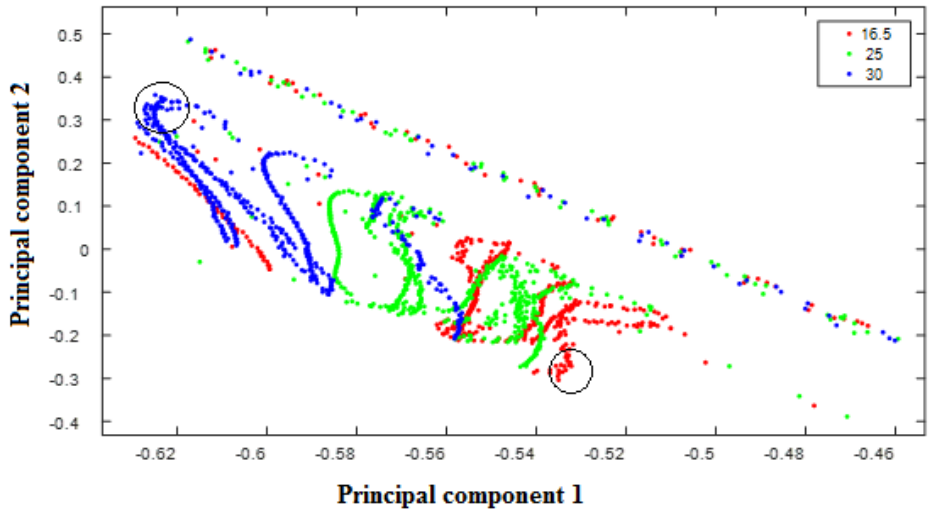
As can be seen from Table 1, the first two components have 92.6% of the original variance which is enough to retain all the original variance and dimensionality can be reduced to 2 without loss of information. In addition, Table 2 shows that among the coefficients of four input variables (or the weights of input variables) in the first two principal components, current density and type of electrode (1 for Pt electrode, 0 for GC

electrode) are dominant factors in the first principal component; and potential and current density are dominant factors for the second principal component.

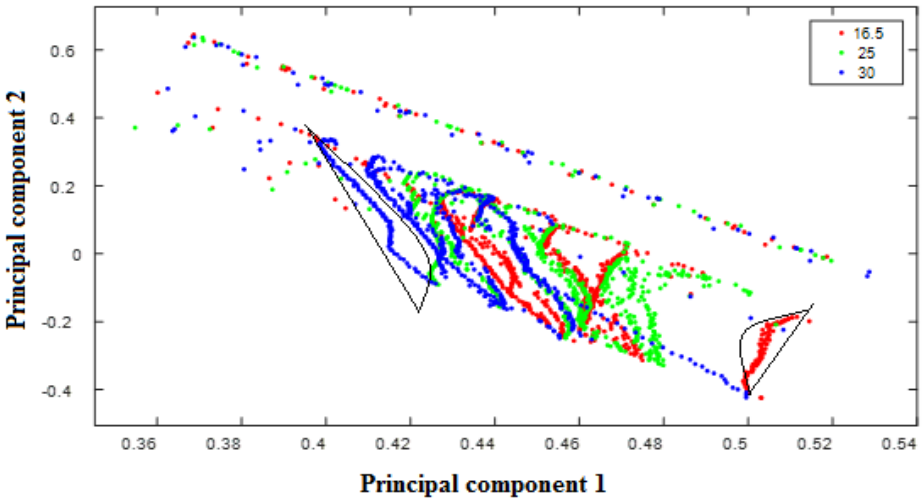
Table 2. *Coefficients of First Two Principal Components*

Predictor	**PC 1**	**PC 2**
Potential	-0.03252	-0.6012
ln(1+current density)	-0.1771	0.7895
KI Concentration	0.00364	-0.0142
Pt Electrode (0 or 1)	0.9836	0.1223

In order to understand whether the variation retained in the selected components contains relevant information about the level of output (UCD values), each sample is projected onto these principal components and separate levels of the output (separate clusters of output) were searched on these components as seen in the Figure 6 below:



(a)



(b)

Figure 6. *Principal Component Analysis of Cyclic Voltammetry Data without Outliers and Distribution of UCD Values with Respect to the First Two Principal Components*
*(**red filled circles:** UCD=16.5; **green filled circles:** UCD=25; **blue filled circles:** UCD=30; pure clusters were surrounded by **black lines**)*
*Range of Principal Component Analysis with PC1 **a)** Between -0.64 and -0.46, **b)** Between 0.34 and 0.54*

In Figure 6, by the representation of UCD values (output levels) with 2 dimensional plots, clusters of UCD values can be visualized. In Figure 6, it was also decided to split the principal component analysis into two regions of first principal component in order to see clearly pure clusters of UCD values. Therefore, from Figure 6a at high current densities (negative PC1), UCD values of 30 and at high potentials and high currents, UCD values of 16.5; and from Figure 6b at high potentials on platinum electrode UCD values of 16.5 could be observed.

3.4. Decision Tree Analysis

For the construction decision tree, in order to see the effect of splitting criterion on the regions of overfit and underfit, initially, testing and training errors which are root mean square error of the difference between class predictions (from decision tree) and observations were compared at different pruning levels. As seen in Figure 7, different splitting criterions (Figure 7a,b,c) exhibit different regions of overfit and underfit; and minimum testing error was achieved at a pruning level of 21 for "deviance" splitting criterion.

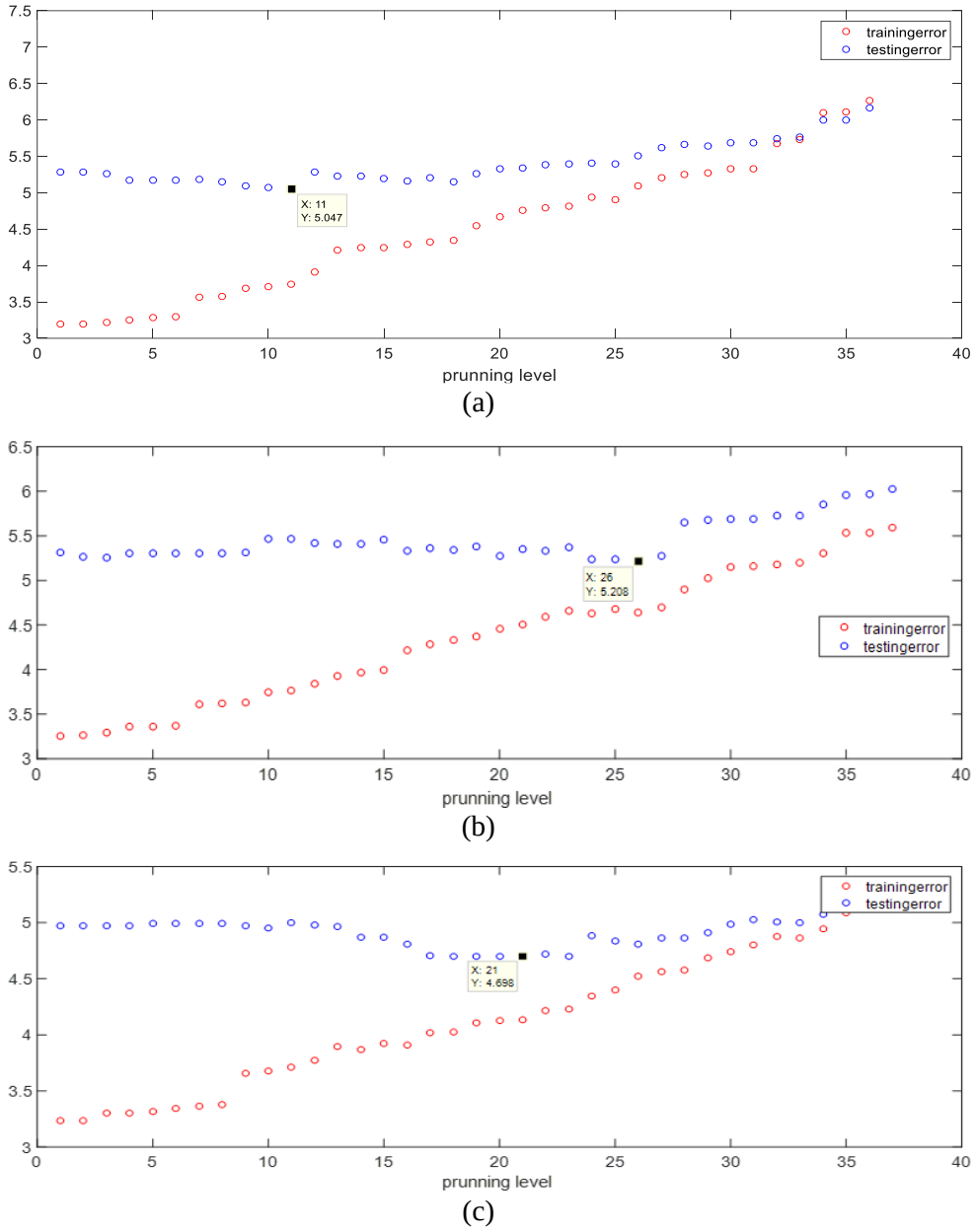


Figure 7. *Testing and Training Errors at Different Pruning Levels for Different Splitting Criterions* ***a)*** *gdi,* ***b)*** *twoing,* ***c)*** *deviance, testing subset:10% of observations*

After determination of training and testing errors for different pruning levels, fraction of misclassified observations were analyzed by determination of confusion matrices with Matlab command "confusionmat" for the three splitting criterion at the optimum pruning level. As seen in Table 3, deviance splitting criterion exhibit lowest number of misclassifications in the confusion matrix.

Table 3. *Confusion Matrix Charts of Testing Data for Different Splitting Criterions* ***a)*** *gdi,* ***b)*** *twoing,* ***c)*** *deviance (**red box:** misclassification, **blue box:** correct classification)*

gdi				
True	**UCD=16.5**	82	21	14
	UCD=25	30	87	12
	UCD=30	11	19	78
		UCD=16.5	**UCD=25**	**UCD=30**
(a)		**Predicted**		
twoing				
True	**UCD=16.5**	73	33	10
	UCD=25	27	92	11
	UCD=30	13	32	63
		UCD=16.5	**UCD=25**	**UCD=30**
(b)		**Predicted**		
deviance				
True	**UCD=16.5**	85	22	9
	UCD=25	20	90	19
	UCD=30	12	19	78
		UCD=16.5	**UCD=25**	**UCD=30**
(c)		**Predicted**		

After construction of binary decision tree with the optimum pruning level ,21, and optimum splitting criterion "deviance" based on minimum testing error and confusion matrix with minimum misclassifications, observations were split with 71 branch and 71 leaf nodes with these hyperparameters (splitting criterion and pruning level) as seen Figure 8.

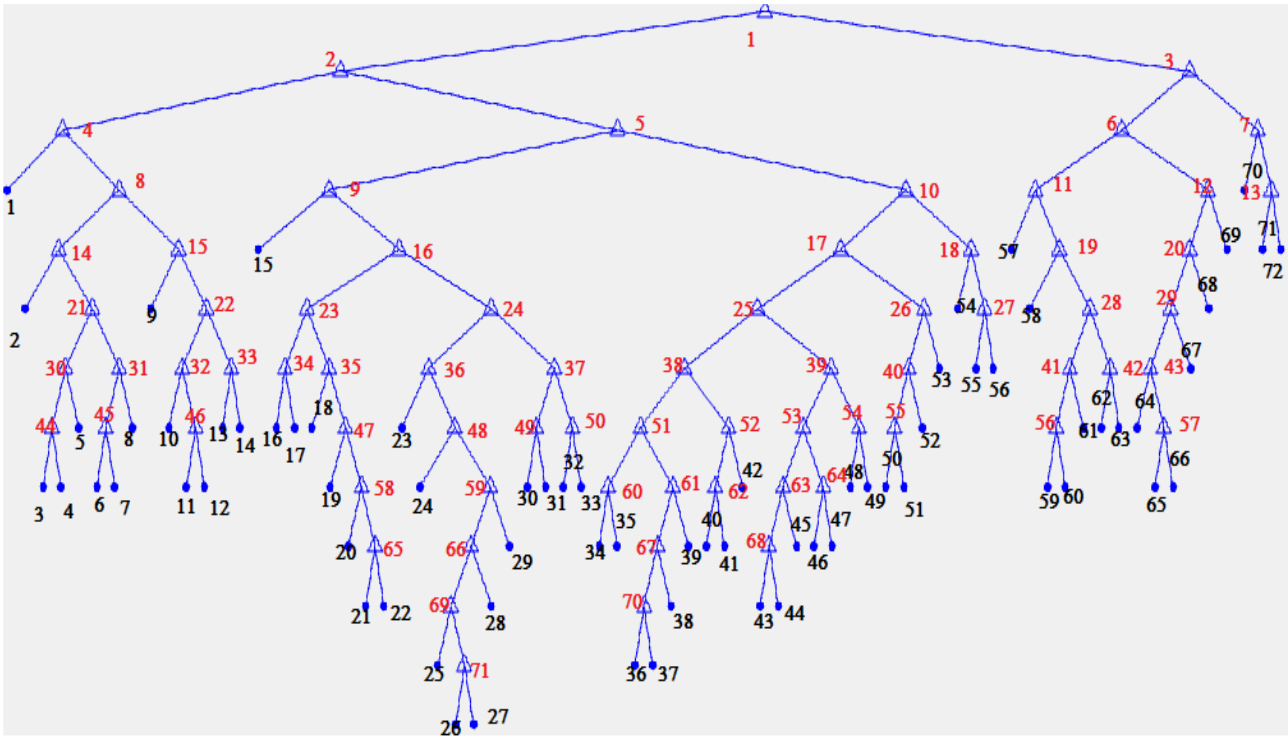


Figure 8. *Binary Classification Tree Constructed by "Deviance" Splitting Criterion*

The rules of branch nodes seen in Figure 8 is also presented in Table 4. As seen in Table 4, the decision tree grows with the root node which uses current density as a rule and the first two main branches are electrode type and current density again. In addition, Table 4 shows that most of the leaf nodes involve "current density" as a rule. It also important to note that branch nodes 7 and 13 are not taken into consideration for further analysis since "electrode type" is not used as a rule in these branch nodes.

Table 4. *Branch Node Rules Based on Binary Decision Tree with "Deviance" Criterion and 21 Pruning Level*

Node#	Rule	Node#	Rule
1	current density (μA/cm^2)<10350.33	36	potential (V) <0.545
2	electrode type (Pt=1,GC=0)=0	37	current density (μA/cm^2)<8392.39
3	current density (μA/cm^2)<12066.73	38	current density (μA/cm^2)<9761.56
4	potential (V) <0.515	39	current density (μA/cm^2)<9735.43
5	current density (μA/cm^2)<9020.36	40	current density (μA/cm^2)<10320.25
6	electrode type (Pt=1,GC=0)=0	41	potential (V) <0.815
7	potential (V) <0.49	42	potential (V) <0.85
8	current density (μA/cm^2)<9794.22	43	potential (V) <0.675
9	current density (μA/cm^2)<6638.42	44	current density (μA/cm^2)<8992.97
10	potential (V) <0.895	45	potential (V) <0.625
11	potential (V) <0.515	46	current density (μA/cm^2)<9817.96
12	current density (μA/cm^2)<11122.55	47	current density (μA/cm^2)<7841.63
13	potential (V) <0.965	48	current density (μA/cm^2)<8439.36
14	current density (μA/cm^2)< 8940.69	49	current density (μA/cm^2)<8189.17
15	potential (V) <0.815	50	potential (V) <0.845
16	current density (μA/cm^2)<8180.00	51	potential (V) <0.615
17	current density (μA/cm^2)<9867.47	52	potential (V) <0.665
18	current density (μA/cm^2)<9842.83	53	potential (V) <0.845
19	KI concentration (M) =0.075	54	potential (V) <0.765
20	potential (V) <0.735	55	potential (V) <0.635
21	current density (μA/cm^2)<9219.37	56	current density (μA/cm^2)<10671.97
22	current density (μA/cm^2)<9846.87	57	current density (μA/cm^2)<10606.54
23	current density (μA/cm^2)<7362.86	58	current density (μA/cm^2)<8139.77
24	potential (V) <0.765	59	potential (V) <0.745
25	potential (V) <0.765	60	current density (μA/cm^2)<9115.58
26	potential (V) <0.735	61	current density (μA/cm^2)<9443.80
27	current density (μA/cm^2)<9995.90	62	current density (μA/cm^2)<9833.28
28	current density (μA/cm^2)<11652.13	63	current density (μA/cm^2)<9167.23
29	current density (μA/cm^2)<10901.09	64	current density (μA/cm^2)<9254.28
30	potential (V) <0.895	65	potential (V) <0.835
31	potential (V) <0.885	66	current density (μA/cm^2)<9009.81
32	potential (V) <0.925	67	current density (μA/cm^2)<9240.23
33	potential (V) <0.925	68	potential (V) <0.785
34	current density (μA/cm^2)<6922.94	69	potential (V) <0.675
35	potential (V) <0.805	70	potential (V) <0.69
		71	current density (μA/cm^2)<8844.20

In order to determine the significance of input variables in the decision tree model, importance of the input variables were also determined by "predictorImportance" function in Matlab environment. In Matlab, predictor importance function sums up changes in the risks of the nodes related to each predictor in the pruned decision tree and divides the sum by the number of branch nodes. The change in the node risk is the difference between the risk for the parent node (related with specific predictor) and sum of risks of child nodes as given in equation 8 where R_p, R_c and N_{branch} is the is parent risk, child node risk and total number of branch nodes. The risk of a node is defined as multiplication of node impurity (equation 3) by the probability of the node which is the ratio of observations that are classified in that node to total number of observations.

$$(R_p - \Sigma R_{ci})/N_{branch} \quad (8)$$

Table 5 shows the importance of the predictors used in the pruned decision tree where current density has the highest impact on UCD. The importance of the predictors can be ordered as current density (J) > potential > electrode type > KI concentration in Table 5.

Table 5. *Importance of Predictors in the Optimum Decision Tree*

Predictor (or Input Variable)	**Importance**
Current Density (J)	$1.03x10^{-3}$
Potential	$7.25x10^{-4}$
Electrode Type (Pt or GC Electrode)	$9.57x10^{-5}$
KI Concentration	$8.78x10^{-5}$

Table 6 shows leaf node number, highest class membership and the leaf node rules for the optimized binary decision tree. High class memberships in the leaf nodes indicate that voltammetric data with different electrode types, damaged starch content flour (UCD values) and electrolyte concentration (KI concentration) can be successfully modeled by classification tree. If the rules in Table 6 are examined based on different UCD values on the same electrode type, it is seen that although rules indicate same potential ranges maximum current density (J) increase with UCD values on the GC electrode and UCD values of 30 give highest current ranges for both GC and platinum electrode. At UCD value of 16.5 and 25, higher currents were seen on GC electrode compared to platinum electrode.

As a final word, although it is difficult to extract physical interpretation from the rules of binary decision tree, performance indicators like confusion matrix and highclass memberships in the leaf nodes demonstrate the high accuracy of predicted UCD values just from voltammetric experiments.

Table 6. *Leaf Node Rules in the Optimum Decision Tree*

Leaf Node #	**Class Membership**	**Branch Nodes**
2	UCD 16.5=93/105	Glassy carbon electrode, potential >0.515V, J(μA/cm^2) <8940.69
10	UCD 16.5=40/43	Glassy carbon electrode, 0.925>potential (V) >0.815, 9794.22<J (μA/cm^2) <9846.87
14	UCD 16.5=14/16	Glassy carbon electrode, potential (V) >0.925, 9846.87<J (μA/cm^2) <10350.33
60	UCD 25=89/121	Glassy carbon electrode, 0.815>potential (V) >0.515, 10671.97<J (μA/cm^2)<11652.13, KI concentration (M) =0.15
61	UCD 25=109/115	Glassy carbon electrode, potential (V) >0.815, 10350.33<J (μA/cm^2)<11652.13, KI concentration (M) =0.15
7	UCD 25=88/90	Glassy carbon electrode,0.625<potential (V) <0.885, 9794.22>J (μA/cm^2)>9219.37
62	UCD 25=39/39	Glassy carbon electrode, 0.515<potential (V) <0.85, 11652.13<J (μA/cm^2)<12066.73, KI concentration (M) =0.15
3	UCD 25=11/14	Glassy carbon electrode, 0.895V>potential >0.515V, 8992.97>J (μA/cm^2)>8940.69
5	UCD 25=13/13	Glassy carbon electrode, 0.895V<potential, 8992.97>J(μA/cm^2) >8940.69
63	UCD 30=25/26	Glassy carbon electrode, potential (V) >0.85, 11652.13<J (μA/cm^2)<12066.73, KI concentration (M) =0.15
58	UCD 30=16/17	Glassy carbon electrode, potential (V) >0.515, 10350.33<J (μA/cm^2)<12066.73,KI concentration (M) =0.075
59	UCD 30=12/13	Glassy carbon electrode, 0.515<potential (V) <0.815, 10350.33<J (μA/cm^2)<10671.97, KI concentration (M) =0.15
15	UCD 16.5=69/80	Platinum electrode, J (μA/cm^2) <6638.42
20	UCD 16.5=43/58	Platinum electrode, potential (V) >0.805, 8139.77> J(μA/cm^2) >7841.63
25	UCD 16.5=43/56	Platinum electrode,0.675>potential (V) >0.545, 9009.81>J (μA/cm^2) >8439.36
24	UCD 16.5=44/51	Platinum electrode,0.545<potential (V) <0.765, 8180.00<J (μA/cm^2) <8439.36
45	UCD 16.5=36/39	Platinum electrode,0.765<potential (V) <0.845, 9167.23<J (μA/cm^2)<9735.43
44	UCD 16.5=15/15	Platinum electrode,0.785<potential (V) <0.845, 9020.36<J (μA/cm^2)<9167.23
30	UCD 16.5=12/14	Platinum electrode, potential (V) >0.765, 8189.17>J (μA/cm^2) >8180.00
18	UCD 25=89/110	Platinum electrode, potential (V) <0.805, 7362.86<J(μA/cm^2)<8180.00
17	UCD 25=25/35	Platinum electrode, 6922.94<J (μA/cm^2)<7362.86
27	UCD 25=16/18	Platinum electrode, 0.675<potential (V) <0.745, 8844.20<J (μA/cm^2) <9009.81
28	UCD 25=12/16	Platinum electrode, 0.545<potential (V) <0.745, 9009.81<J (μA/cm^2) <9020.36
65	UCD 25=13/14	Platinum electrode, 0.675<potential (V) <0.735, 10606.54>J (μA/cm^2) >10350.33
66	UCD 25=13/14	Platinum electrode, 0.675<potential (V) <0.735, 10606.54<J (μA/cm^2) <10901.09
55	UCD 25=12/13	Platinum electrode, potential (V) >0.895, 9842.83<J(μA/cm^2)<9995.90
69	UCD 30=72/92	Platinum electrode, 11122.55<J (μA/cm^2)<12066.73
68	UCD 30=46/47	Platinum electrode, Potential>0.735V, 11122.55<J (μA/cm^2)<12066.73
38	UCD 30=23/24	Platinum electrode, 0.615<potential (V) <0.765, 9240.23<J (μA/cm^2)<9443.80
67	UCD 30=12/14	Platinum electrode, potential (V) <0.735, 10901.09<J (μA/cm^2)<11122.55

4. CONCLUSION

In this study, classification tree was successfully applied on iodine oxidation voltammetric data. High class memberships and testing error in the pruned tree indicate that UCD values (damaged starch ratio) can be predicted from the selected electrochemical variables during voltammetric study. The machine learning strategy can be extended to other electrochemical techniques or with higher number of electrochemical variables for further studies.

ACKNOWLEDGEMENT

We would like to thank Gazi University Scientific Research Projects, BAP #: 06/2018-12 for the financial support.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Baysal, M., Günay, M. E., & Yıldırım, R. (2017). Decision tree analysis of past publications on catalytic steam reforming to develop heuristics for performance: A statistical review. *International Journal of Hydrogen Energy*, *42*(1), 243-254. doi:10.1016/j.ijhydene.2016.10.003

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC. doi:10.1201/9781315139470

Boschloo, G., & Hagfeldt, A. (2009). Characteristics of the Iodide/Triiodide Redox Mediator in Dye-Sensitized Solar Cells. *Accounts of Chemical Research*, *42*(11), 1819-1826. doi:10.1021/ar900138m

Comon, P. (1994). Independent component analysis, A new concept?. *Signal Processing*, *36*(3), 287-314. doi:10.1016/0165-1684(94)90029-9

Dhital, S., Shrestha, A. K., Flanagan, B. M., Hasjim, J., & Gidley, M. J. (2011). Cryo-milling of starch granules leads to differential effects on molecular size and conformation. *Carbohydrate Polymers*, *84*(3), 1133-1140. doi:10.1016/j.carbpol.2011.01.002

Günay, M. E., Türker, L., & Tapan, N. A. (2018). Decision tree analysis for efficient CO_2 utilization in electrochemical systems. *Journal of CO_2 Utilization*, *28*, 83-95. doi:10.1016/j.jcou.2018.09.011

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Elsevier.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*, 20150202. doi:10.1098/rsta.2015.0202

Larose, D. T., & Larose, C. D. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining* (2nd ed.). John Wiley & Sons.

Li, M., Hasjim, J., Xie, F., Halley, P. J., & Gilbert, R. G. (2014). Shear degradation of molecular, crystalline, and granular structures of starch during extrusion. *Starch-Stärke*, *66*(7-8), 595-605. doi:10.1002/star.201300201

Liu, W-C., Halley, P. J., & Gilbert, R. G. (2010). Mechanism of Degradation of Starch, a Highly Branched Polymer, during Extrusion. *Macromolecules*, *43*(6), 2855-2864. doi:10.1021/ma100067x

Liu, X., Xiao, X., Liu, P., Yu, L., Li, M., Zhou, S., & Xie, F. (2017). Shear degradation of corn starches with different amylose contents. *Food Hydrocolloids*, *66*, 199-205. doi:10.1016/j.foodhyd.2016.11.023

Medcalf, D. G., & Gilles, K. A. (1965). Determination of Starch Damaged by Rate of Iodine Absorption. *Cereal Chemistry*, *42*, 546-557.

Myles, A. J., Feudale, R. N, Liu, Y., Woody, N. A., & Brown, S. D. (2004). An Introduction to Decision Tree Modeling. *Journal of Chemome*trics, *18*(6), 275-285. doi:10.1002/cem.873

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, *1*, 81-106. doi:10.1007/BF00116251

Ringnér, M. (2008). What is Principal Component Analysis?. *Nature Biotechnology*, *26*(3), 303-304. doi:10.1038/nbt0308-303

Tapan, N. A., Günay, M. E., & Yildirim, R. (2016). Constructing Global Models from Past Publications to Improve Design and Operating Conditions for Direct Alcohol Fuel Cells. *Chemical Engineering Research and Design*, *105*, 162-170. doi:10.1016/j.cherd.2015.11.018

Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (2012). *Probability & Statistics for Engineers & Scientists* (9th ed.). Prentice Hall, Boston.

Zhu, F. (2016). Buckwheat Starch: Structures, Properties and Applications. *Trends in Food Science & Technology*, *49*, 121-135. doi:10.1016/j.tifs.2015.12.002