Dataset for the preparation and application of multi-component electrocatalysts in methanol oxidation based on non-precious metals for fuel cell and sensor under ambient-conditions

Nirwan Syarif [1]*, Dedi Setya Budidaya [2], Eliza[3]

Department of Chemistry, Faculty of Mathematics and Natural Sciences, Inderalaya, SouthnSumatera, Indonesia[1,3]
Department of Physics, Faculty of Mathematics and Natural Sciences, Inderalaya, South Sumatera, Indonesia[2]

*nsyarif@unsri.ac.id

Abstract. Artificial intelligence methods facilitate data exploration, application, and analysis, increasing significantly, including machine learning. Data must be reliable and accessible, which places high demands on its acquisition and storage. This complexity is partly due to the wide range of lengths and time scales involved in the many different processes. The data in this article refer to the materials prepared as nonprecious metal for methanol oxidation in fuel cells and sensors. Metal oxides are an alternative for such applications. Important parameters can be used to assess the performance of electrocatalysts: These four parameters are related to other quantities available as a dataset, namely castelli perovskite. The Castelli perovskite dataset contains data conduction band energy level value, heat of formation in eV, Fermi energy level, Fermi bandwidth, material, chemical formula, electronic band gap, magnetic moment, crystal structure, valence band energy level value). The Castelli perovskite dataset is then connected to experimental data through the material's chemical formula. Nonprecious metal oxides include 483 materials (electrocatalyst materials). Based on the results of KMeans data processing, 483 electrocatalyst materials were grouped into 4 types. Thus, non-noble metal oxides can be categorized into 4 types. Correlation data processing shows that the current density correlates with the proposed electrocatalyst type.

Keywords: DMFC, electrochemistry, oxidation, catalyst, data analyze

1. Introduction

The idea is to use the membrane electrode assembly (MEA) design for the sensor in addition to the facilities owned by the research center of excellence, Universitas Sriwijaya, and adequate mastery of MEA fabrications. Various energy materials called electrocatalysts are pinned to the devices in order to have new catalyst materials [1] with superior performance [2] higher energy density and higher energy conversion electrocatalysts for methanol oxidation with high product selectivity and maximum performance [3], high efficiency [4] used in energy conversion devices and storage.

ICIESC 2025, September 16, Medan, Indonesia

DOI 10.4108/eai.16-9-2025.2361146

The manufacture of chemicals, electrochemical conversion, the production of sustainable materials, and the transformation of energy systems all require catalysts. [5]. The use of catalysts in energy, environmental applications, and industrial processes represents a significant disruptive innovation. Mastering catalyst technology requires a deep understanding of physicochemical principles and a closer integration of experiments with theoretical frameworks, which are essential for determining optimal experimental conditions. Moreover, the growing application of artificial intelligence methods enhances data exploration, utilization, and analysis [6]. For this purpose, data must be reliable and accessible, which places high demands on its acquisition and storage.

Machine learning models already have a transformative impact on developing low-cost electrocatalysts [7] and one of the most widely used models [8]. Machine learning models for faradaic – nonfaradaic processes [9] were applied to predict the performance of electrochemical reactions of the materials in direct methanol fuel cells [10]. The community has collected and organized experiment data to the large data sets. The sets were reused in design, research, and development [11]. A set of the density functional theory calculations help researcher to identify electrocatalyst materials for the battery [12].

The models have been shown to predict the properties of crystalline materials much faster than quantum calculations [13], predict properties that are difficult to access through other computational tools [14], and guide the search for new materials [15]. In the modern information paradigm, data becomes an object of study called data science [16]. Reliable data is a valuable knowledge resource for research and development [17]. Data with its complexity and functionality [18] is the key to answering important questions related to it, such as the largest data, the smallest data, data distribution, average value, middle value, and relationships between data.

This article aims to build a data set of the current state of the data infrastructure in electrocatalysis of fuel cell and methanol research and propose and discuss solutions and future directions for data management, focusing on electrocatalysis experimental data. Approaches and experiences are also introduced. The lessons learned from this analysis will apply to other areas of catalysis and, even more broadly, to other disciplines in chemistry and physics, and benefits can be derived from mutual inspiration.

2. Methods

2.1. Data Description

The research focused on oxides of non-precious metals that have been functionalized as electrocatalysts for methanol oxidation in electrochemical devices. This approach aims to establish a relationship between the composition, crystallographic structure, and electrochemical activity of various catalytic materials. While single-component electrocatalysts can effectively catalyze methanol oxidation, there is a limitation to their use as they conform to the periodic table's constraints on materials. Therefore, we are exploring the potential of bi-oxides or tri-oxides as electrocatalysts to address the increasing energy demand and achieve the necessary power levels.

Methanol, as a liquid fuel, offers several advantages, including high energy density, ease of storage, and compatibility with existing infrastructure. During the oxidation process at the anode, methanol undergoes electrochemical oxidation, producing protons, carbon dioxide, and electrons. The

generated protons and electrons then participate in reactions at the electrodes, which facilitates the overall energy conversion process.

Understanding the requirements of methanol evolution reactions (MER) and methanol oxidation reactions (MOR) is essential for optimizing the design of electrocatalysts and electrodes in direct methanol fuel cells (DMFC) and sensors. This knowledge contributes to improved fuel cell efficiency and helps address challenges such as catalyst poisoning and the accumulation of intermediates. While platinum-based catalysts currently serve as the benchmark for both hydrogen evolution reactions (HER) and hydrogen oxidation reactions (HOR), their high cost and vulnerability to CO poisoning necessitate the exploration of alternative electrocatalysts that offer enhanced activity and durability for methanol oxidation and hydrogen oxidation reactions (HOR).

2.2. Descriptor Generation using Matminer

To complement the experimental data, we systematically generated theoretical descriptors using the Matminer Python library [19]. Matminer offers a curated collection of featurizers that transform chemical composition and crystallographic information into quantitative descriptors reflecting intrinsic materials properties, including elemental attributes, bonding characteristics, and structural motifs. This approach enables the integration of experimental observations with theory-informed representations, facilitating robust machine learning analysis. The resulting descriptor space provides a consistent and physically interpretable framework for exploring correlations between composition, structure, and electrocatalytic performance.

2.3. Composition-Based Descriptors

Composition-based descriptors were generated to capture intrinsic chemical information independent of long-range crystal order. These descriptors quantify elemental properties such as electronegativity, atomic radius, valence electron count, and ionization energy, and are statistically aggregated (e.g., mean, maximum, minimum, and variance) according to the material composition. Such features provide a compact yet physically meaningful representation of chemical complexity and are particularly suitable for systems with compositional disorder, solid solutions, or limited crystallographic information.

The descriptors were employed to encode intrinsic chemical characteristics that govern electrocatalytic activity independent of long-range crystal order. These descriptors statistically aggregate elemental properties—such as electronegativity, atomic radius, valence electron count, d-band filling, and ionization energy—according to the material composition. For HER and HOR, such features reflect trends in hydrogen binding strength and proton–electron transfer kinetics, while for methanol oxidation they capture the chemical propensity for intermediate adsorption and tolerance toward CO-like species. In composite catalysts, composition-based descriptors effectively represent elemental synergy, dopant effects, and metal–oxide interactions, providing a robust chemical fingerprint even when structural heterogeneity or partial amorphicity is present.

Composition-based descriptors provide a chemistry-driven baseline for HER/HOR and MOR activity by encoding elemental properties, whereas structure-based descriptors introduce geometric and bonding information that allows machine learning models to distinguish catalysts with similar compositions but different active-site architectures and reaction pathways.

Combining both descriptor types expands the feature space to capture complementary chemical and structural signals, thereby enhancing model expressiveness, reducing representation bias, and improving predictive performance across diverse electrocatalyst systems.

2.4. Data Integration and Processing

The composition-based and structure-based descriptors were combined with the experimental exchange current density, J_0, dataset by matching catalyst names. The merged dataset was then cleaned by removing features that had constant or missing values, normalizing numerical features when appropriate, and encoding categorical variables.

The type of electrocatalyst from the list of materials used, for example, Pt, MnO_2, NiMo, MoS_2 with the measured exchange current density (J_0), Matminer converts the simple text description ("PtRu") into a vector of numerical features describing physical and chemical properties, such as: mean and range of electronegativity, atomic radii, valence electron configurations, stoichiometry statistics, ionic properties, local structural fingerprints or crystal structures.

Matminer's featurizers were applied to systematically transform raw material information into a machine learning–ready dataset for predicting the exchange current density (J_0) of MER/MOR electrocatalysts.

Composition-based featurizers, such as ElementProperty, Stoichiometry, ValenceOrbital, and IonProperty, were utilized to generate descriptors that capture elemental averages—including electronegativity, atomic radius, and ionization energy—as well as valence-electron configurations, oxidation states, and stoichiometric metrics. Additionally, experiment-specific metadata, including current density and overpotential, was integrated into the feature matrix to account for operational conditions that significantly influence current density and to enhance the intrinsic descriptors. This featurization strategy facilitated the creation of a comprehensive, quantitative dataset that combines material properties with experimental context, providing a solid foundation for robust machine learning model training and interpretability analysis.

3. Results and Discussion

3.1. Data mining

Several important parameters can be used to assess the performance of electrocatalysts, namely (1) overpotential, which is the additional voltage required beyond the thermodynamic voltage. This additional voltage is used to drive chemical reactions. The general picture is that the electrocatalyst is more efficient in driving chemical reactions for low overpotential values. (2) exchange current density: a measure of the reaction rate at equilibrium voltage. If the exchange current density is high, the catalytic activity is better. (3) Turnover frequency (TOF): The number of catalytic cycles per active site per unit time. If the TOF value is high, the electrocatalysis is more efficient. (4) onset potential: the minimum voltage at which the catalytic reaction begins. A low onset voltage indicates higher activity.

Some computation and molecular simulation results are stored in a database whose data can be reused and analyzed to provide necessary conclusions not previously discussed in previous reports.

A dataset closely related to electrocatalysis should be mined from Matminer (https://hackingmaterials.lbl.gov/matminer) [20].

Several factors that affect the above parameters are material properties, namely electronic structure, area (electrochemistry), electrical conductivity, active sites, surface energy, and morphology; reaction environment - consisting of electrolyte pH, electrolyte composition, temperature; stability and durability - consisting of corrosion resistance; structural stability, chemical stability; selectivity - faradaic efficiency, reaction pathway control; conductive support material. If the performance parameters are determined from the measurement results, the influencing factors can be determined from the computation results. Computation and molecular simulation are carried out using calculation formulas at the ab initio, semi-empirical, density function, and force field levels.

Matminer is a Python library that mines data on material properties. Matminer has features ready-to-use datasets (matminer.datasets), covering a wide range of materials data domains used to create custom datasets for specific research purposes, such as in this study to study electrocatalysts from an online repository (matminer.data_retrieval).

The datasets module provides a growing collection of materials science datasets that have been collected, formatted as pandas data frames, and made available through a unified interface. Loading a dataset as a data frame uses pandas and the unified interface with the matminer module name and class to load the data into memory.

```
from matminer.datasets import load_dataset
df = load_dataset("jarvis_dft_3d")
from matminer.datasets.convenience_loaders import load_jarvis_dft_3d
df = load_jarvis_dft_3d(drop_nan_columns=["bulk modulus"])
```

Matminer's consistently formatted datasets make analyzing and visualizing initial data sets quick and easy:

```
from figrecipes import PlotlyFig
from matminer.datasets import load_dataset
df = load_dataset("elastic_tensor_2015")
pf = PlotlyFig(df, y_title='Bulk Modulus (GPa)', x_title='Shear Modulus
(GPa)', filename='bulk_shear_moduli')
pf.xy(('G_VRH','K_VRH'), labels='material_id', colors='poisson_ratio',
colorscale='Picnic', limits={'x': (0, 300)})
```

The downloaded data is presented in a data frame format to be processed with miners. featurizers. conversions. This module/library has several classes that enrich the meaning of the data being processed. This module defines a featurizer that can convert between various data formats.

Featurizers themselves do not produce features ready for machine learning. Instead, they must be used to preprocess the data through standalone transformations or as part of a Pipeline.

A utility feature to add oxidation states to pymatgen compositions. Oxidation states are determined using pymatgen's guess routine. The expected input is a pymatgen.core.composition. Composition

object. Note that this Featurizer does not produce features ready for machine learning but can be applied to preprocess data or as part of a Pipeline.

Regarding the electrocatalysts, where the relationship between experimental and computational data is important, a deeper look at the two datasets is needed. The search results suggest one dataset that can be linked to castelli_perovskites. The selection of the data set was based on the availability of information that could be linked to the information in the set related to the electrocatalytic performance. The following columns are available in Castelli perovskite. A 18,928 perovskites were generated by ABX combinatorics, calculating the glibsc band gap and pbe structure, and also reporting the absolute band edge positions and heats of formation with the number of entries: 18928 with the column descriptions as follows

Table 1. Descriptions of relationship between experimental and computational data

Column	Description
cbm	Conduction band energy level values based on the CBM method
e_form	The heat of formation in eV, where the reference state for oxygen is calculated from the chemical potential of oxygen in water vapor, not as molecular oxygen.
fermi level	Energy levels that indicate the relative position of the outermost electrons in relation to thermodynamic work, Gibbs free energy.
fermi width	The width of the fermi band, where the end of the fermi bandwidth is greatly affected by temperature. Some of the widths of the end of the band can rise, which means some electrons at the valence energy level can rise to the conduction band so that the material, previously a semiconductor, can change into a conductor.
formula	Chemical formula of material
gap gllbsc	The electronic band gap in eV is calculated through the gllbsc function
gap is direct	Boolean indicator for direct gap
mu_b	Magnetic moment in Bohr magneton units
structure	Crystal structures are represented by the Pymatgen Structure object
vbm	The valence band energy level values are calculated through gllbsc

If we look at the composition of the data set, some quantities are factors that affect the performance of the electrocatalyst. The following is data mining carried out on the Castelli data set (Table 2).

This data set contains several columns that contain information related to the formula (chemical formula) of the electrocatalyst material. The table displays a data set with the available columns of. factors that affect the performance

Table 2. The Castelli data set with the available factors that affect the performance

	fermi level	fermi width	e_form	gap is direct	structure	mu_b	formula	vbm	cbm	gap gllbsc
0	0.312138	0.001837	2.16	True	[[0. 0. 0.] Rh, [1.97726555 1.97726555 1.97726...	1.974478e-02	RhTeN3	6.187694	6.187694	0.0
1	0.297083	0.001837	1.52	True	[[2.54041798 0. 0.] Hf, [1.020...	-2.253054e-05	HfTeO3	6.033125	6.033125	0.0
2	0.191139	0.003675	1.48	True	[[0.60790913 0. 0.] Re, [2.186...	4.982109e+00	ReAsO2F	6.602253	6.602253	0.0
3	0.316346	0.001837	1.24	True	[[2.83091357 0. 0.] W, [2.6573...	-8.684496e-01	WReO2S	5.738462	5.738462	0.0
4	0.312658	0.003675	0.62	True	[[0.00518937 0. 0.] Bi, [2.172...	2.164069e-15	BiHfO2F	6.074736	6.074736	0.0

In the first mining we filter the data that is the object of this research, namely metal oxides from nonprecious metal elements. So that unnecessary elements contained in the alloy structure are set aside, namely by using the script attached below and the output

```
#mask = df["formula"].str.contains('Pt') & df["formula"].str.contains('Au')== false
filtered = asli_df['formula'].str.contains('Pt|Au|Te|Os|Hg|Tl|Ge|S|Hf|Re|As|F|Y|Ga|In|Rb|N|La|Ir|Rh|Ta') == False

#filtered = df.saring['gap gllbsc'] == 0.0
#

cleaned=df[filtered].drop(["cbm","gap is direct","structure","vbm"], axis=1)
cleaned.head(500)
```

Table 3. The output of the script

[17]:

	fermi level	fermi width	e_form	mu_b	formula
12	0.274061	0.001837	0.78	1.656502	PbRuO3
18	0.178276	0.001837	1.70	3.537787	RuMgO3
24	0.163969	0.001837	0.82	3.486010	KCoO3
31	0.279114	0.001837	0.66	-2.709618	VMoO3
42	0.304318	0.001837	1.40	3.818457	WMoO3
...	...	...	...	...	...
13804	0.142309	0.001837	1.56	-0.364134	CdAgO3
13820	0.035236	0.001837	1.32	3.000000	LiMgO3
13851	0.134259	0.001837	1.06	-5.451035	MnCdO3
13860	0.289975	0.001837	0.82	1.227465	TiCrO3
13869	0.187585	0.001837	1.30	2.000028	MoCdO3

The table above has sufficiently filtered the data so that it only contains oxides of nonprecious elements. In addition, columns containing unnecessary information also need to be set aside. If there were 10 columns previously, the filtering results set aside 5 columns not used in this study. In addition, a comparison was used, a data set that only contained oxides of precious metals or their alloys. Filtering was carried out using the following script. Nonprecious metal oxides contain a filtering script.

```
filtered = asli_df['formula'].str.contains('Pt|Au|Os|Ir|Rh|Pd|Ag|Te|Hg|Tl|Ge|S|Hf|Re|As|F|Y|Ga|In|Rb|N|La|Ta|Ru') == False


cleaned=df[filtered].drop(["cbm","gap is direct","structure","vbm"], axis=1)

cleaned.head(500)
```

A script is used to filter noble metal oxides.

```
filtered = asli_df['formula'].str.contains('Al|B|Ti|Li|Zn|Co|V|Mg|Pb|W|Cs|Cu|Ba|Ca|Mn|Mo|Cd|Cr|K|Zr|Bi|Te|Hg|Tl|Ge|S|Hf|Re|As|F|Y|Ga|In|Rb|N|La|Ta') == False

cleaned=df[filtered].drop(["cbm","gap is direct","structure","vbm"], axis=1)
cleaned.head(500)
```

Both scripts produce data sets that are stored in Excel files using scripts.

```
import pandas as pd
simpan=pd.DataFrame(cleaned)
cleaned.head(18928)
simpan.to_excel("output_castelli_precious.csv")
print('DataFrame is written to Excel File successfully.')
```

```
import pandas as pd
simpan=pd.DataFrame(cleaned)
cleaned.head(18928)
simpan.to_excel("output_castelli_non_precious_oxide.csv")
print('DataFrame is written to Excel File successfully.')
```

Screenshot of the table generated from the execution of the script (Table 4). The output_castelli_precious.csv file contains 64 columns or 64 types of electrocatalysts from binary metal oxides.

Table 4. The output of the executing script that filtering and cleaning data

	A	B	C	D	E	F
1	No	NoDB	fermi level	e_form	mu_b	formula
2	1	**24**	0.1639688	0.82	3.48601	KCoO3
3	2	**31**	0.2791143	0.66	-2.70962	VMoO3
4	3	**42**	0.3043177	1.4	3.818457	WMoO3
5	4	**196**	0.1025118	1.26	0.993863	KBaO3
6	5	**292**	0.2347449	0.48	1.927513	ZnCrO3
7	6	**301**	0.3041741	0.86	-0.07873	ZrZnO3
8	7	**313**	0.3362582	0.82	-1.91141	VBO3
9	8	**332**	0.1519019	1.68	-0.00069	CaCsO3
10	9	**385**	0.1502206	1.32	4.056302	CrKO3
11	10	**407**	0.2303465	1.26	0.692367	BeWO3
12	11	**612**	0.212938	1.28	1.980657	BeCrO3
13	12	**838**	0.215582	0.98	1.99922	CaBeO3
14	13	**952**	0.2467837	0.68	-0.00147	CaBiO3
15	14	**1005**	0.3448034	0.54	0.014509	BiMoO3
16	15	**1042**	0.2804981	0.14	2	PbCrO3
17	16	**1089**	0.2324525	0.18	-1.44445	VVO3
18	17	**1140**	0.1677221	1.16	-3.85279	MnZnO3
19	18	**1155**	0.22647	1.04	2.00334	MoZnO3
20	19	**1204**	0.0878629	2.16	1.999996	BeMgO3
21	20	**1218**	0.1664704	0.78	1.999999	BaCaO3
22	21	**1227**	0.2319997	1.08	0.354567	AlMoO3
23	22	**1323**	0.1062914	2.54	2.999995	BeLiO3
24	23	**1325**	0.1694516	0.54	0.264408	MgAlO3
25	24	**1377**	0.3951247	1.34	2.763076	MoBO3
26	25	**1407**	0.2053041	1.34	-3.08528	CrCuO3

The output_castelli_non_precious_oxide.csv file contains 484 types of electrocatalysts from binary metal oxides. Both datasets (files) are then equipped with experimental data and literature [21] search results, namely DA (current density) and mu_b (overpotential) which are added in the next column in the file. With the existence of the castelli_non_precious_oxide.csv file, preprocessing can be carried out on the data set before a model is created to optimize the performance of electrocatalysts using binary metal oxides available in the formula column.

The sklearn module is used for data preprocessing, including the electrocatalyst's initial classifier. Here is the script used to activate the sklearn module and several other modules, followed by loading the processed file.

```
import numpy as np
import pandas as pd
import pickle
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.utils import shuffle
from sklearn.preprocessing import LabelEncoder, OrdinalEncoder
from sklearn.ensemble import RandomForestClassifier
import matplotlib.pyplot as plt
```

```
import pandas as pd
df = pd.read_csv('C:/Users/EDLC/output_castelli_non_precious_oxide.csv', sep=',')
```

The dataset resulting from data mining and measurement results which are experimental data (mu_b and DA) are loaded into the application memory for preprocessing and clustering using pandas. After being loaded into memory, which means the process can be started, the first check is to check the data distribution (Table 5). The data distribution check is done on the FL (fermi level) column, e_form (formation energy), mu_b (excess potential), and DA (current density) visualize as histograms of each parameters (Figure 1)

```
[52]: df.describe()
```

Table 5. The distribution of data after being uploaded to memory

[52]:

	No	NoDB	FL	e_form	mu_b	DA
count	484.000000	484.000000	484.000000	484.000000	484.000000	484.000000
mean	242.500000	9441.039256	0.200722	1.141694	0.847973	5.609070
std	139.863028	5193.828445	0.086342	0.742622	1.936695	3.257473
min	1.000000	24.000000	-0.059384	-0.320000	-5.451035	0.000000
25%	121.750000	4984.000000	0.143233	0.620000	-0.000004	2.617500
50%	242.500000	9469.000000	0.201270	1.080000	0.511601	5.640000
75%	363.250000	13745.750000	0.256871	1.520000	2.000000	8.665000
max	484.000000	18910.000000	0.518511	3.640000	6.999998	10.990000

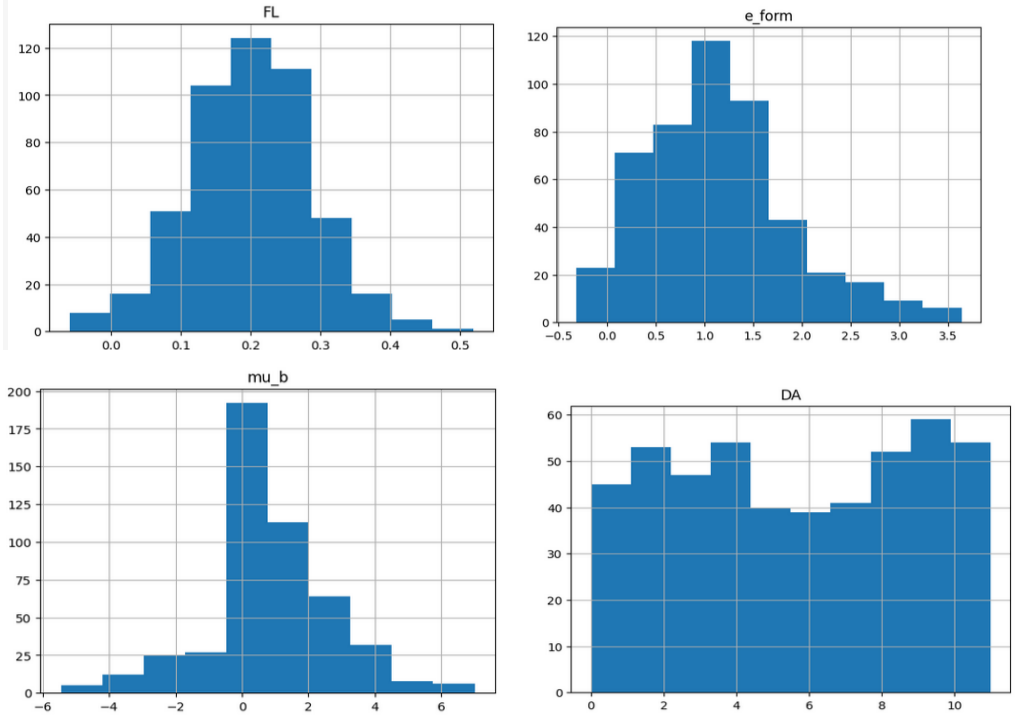

Figure 1. Data distribution histograms for (a) FL, (b) e_form, (c) mu_b and (d) DA

The dataset resulting from data mining and measurement results which are experimental data (mu_b and DA) are loaded into the application memory for preprocessing and clustering.

After being loaded into memory, which means the process standarization can be started by performing the data transformation using preprocessing module, i.e. StandarScaler utility class. The data transformation is done on the FL (fermi level) column, e_form (formation energy), mu_b (excess potential), and DA (current density) which obtain an array as we can see from Table 6.

Table 6. Array of data transformation on 4 descriptor

	No	NoDB	FL	e_form	mu_b	formula	DA	FL_T	e_form_T	mu_b_T	DA_T
0	1	24	0.163969	0.82	3.486010	KCoO3	6.43	-0.426108	-0.433635	1.363543	0.252275
1	2	31	0.279114	0.66	-2.709618	VMoO3	2.52	0.908867	-0.649311	-1.838840	-0.949284
2	3	42	0.304318	1.40	3.818457	WMoO3	2.17	1.201071	0.348189	1.535377	-1.056840
3	4	196	0.102512	1.26	0.993863	KBaO3	8.02	-1.138629	0.159473	0.075408	0.740888
4	5	292	0.234745	0.48	1.927513	ZnCrO3	7.79	0.394457	-0.891946	0.557990	0.670208
...	...	...	...	...	...	...	...	...	...	...	...
479	480	18623	0.175813	0.86	-1.342605	VZnO3	5.78	-0.288794	-0.379716	-1.132261	0.052527
480	481	18713	0.100716	1.98	0.002913	ZnBaO3	9.73	-1.159445	1.130013	-0.436793	1.266378
481	482	18823	0.210733	0.04	-2.630710	CrVO3	9.04	0.116072	-1.485054	-1.798054	1.054339
482	483	18887	0.341396	0.88	-0.053766	ZrMoO3	8.95	1.630952	-0.352757	-0.466089	1.026681
483	484	18910	0.187105	0.90	1.893716	ZnCoO3	0.78	-0.157869	-0.325797	0.540521	-1.483993

3.2. Optimizing and Clustering Data

Data clustering begins by optimizing the relationship between data, which is 483 rows of data as show in Table 6) The first clustering is the relationship between FL (fermi level) and DA (current density) using the KMeans method from the sklearn module to determine number of cluster on the data (Figure 2). Electrocatalysis technologies play an important role in both fuel cells and sensors. For example, fuel cells can be applied to electricity production in DMFC and methanol detection. The hydrogen produced by methanol electrolysis can play a role in electrical energy storage, and the amount of CO_2 and hydrogen in the atmosphere can be directly reduced by the electrochemical reduction of methanol. Moreover, these reaction processes can store the energy in chemical bonds, thereby improving the convenience of clean energy utilization.

Developing high-activity and high-stability electrocatalysts, such as metal oxides, is the key to realizing high-efficiency and low-cost electrochemical reactions. Thus, searching for practical approaches to design promising electrocatalysts is necessary.

```
from sklearn.cluster import KMeans

def optimise_k_means(data, max_k):
    means = []
    inertias = []

    for k in range (1, max_k):
        kmeans = KMeans(n_clusters=k)
        kmeans.fit(data)

        means.append(k)
        inertias.append(kmeans.inertia_)

    #Generate the elbow plot
    fig = plt.subplots(figsize=(10,5))
    plt.plot(means,inertias,'o-')
    plt.xlabel('Number of Clusters')
    plt.ylabel('Inertia')
    plt.grid(True)
    plt.show()

    optimise_k_means(df[['FL_T','DA_T']],10)
```

Figure 2. Practical approaches to design promising electrocatalysts

The typical electrocatalytic reaction consists of reactant adsorption, electron transfer, breakage and formation of chemical bonds, and product desorption. Moreover, many candidate electrocatalyst materials exist, each with a different adsorption capacity for intermediates on various surfaces or sites. Multi-scale modeling can be applied to approach variables that affect electrocatalyst performances. There are five stages related to the performances in (1) systems-level modeling, where we meet factors such as catalyst supporter processing methods, binder availability, test environment, current density, over potential; (2) meso-macro macro-level modelings, such as electrolyte decompositions, short circuit, nucleation, particles sizes, and shapes; porous structural (3) micro-level modeling, such as crack formation, nucleation, the energy of formation (4) atomic level modeling solvent polarity, dielectric constant, fermi level, Bohr constant, hydrophobicity. The development of catalysts faces difficulties due to complexity.

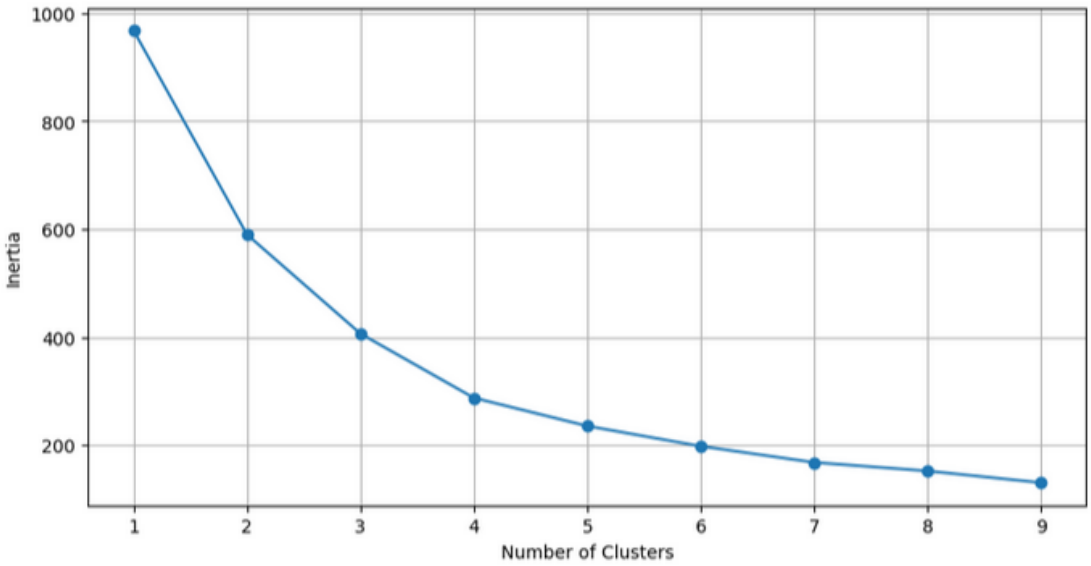

Figure 3. Clustering of electrocatalyst data using the relationship between FL (Fermi level) and DA (current density) using the KMeans class

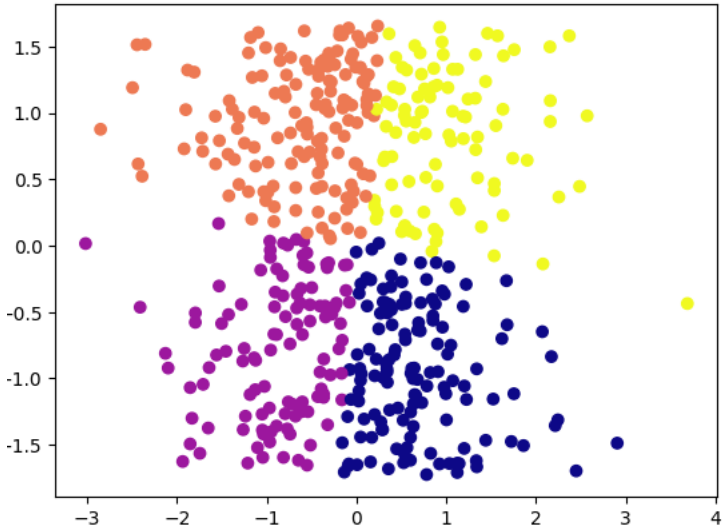

Figure 4. Clustering of binary metal oxide electrocatalysts data, non-precious metal

The clustering results show that the electrocatalysts can be divided into 4 types resulting from the KMeans analysis, where the optimal point of the clustering process occurs at "Number of Clusters" = 4. The following plot illustrates the clustering formed by KMeans (Figure 4).

The process is continued to create an optimal model of binary metal oxide electrocatalysts and classify electrocatalysts into 4 clusters. In addition, a dataset of metal oxide electrocatalysts is also added as a data comparison. The results of the data processing are as in the following Table.7

Table 7. Optimal model of binary metal oxide electrocatalysts

[23]:

	No	NoDB	FL	e_form	mu_b	formula	DA	FL_T	e_form_T	mu_b_T	DA_T	kmeans_4
0	1	24	0.163969	0.82	3.486010	KCoO3	6.43	-0.426108	-0.433635	1.363543	0.252275	2
1	2	31	0.279114	0.66	-2.709618	VMoO3	2.52	0.908867	-0.649311	-1.838840	-0.949284	0
2	3	42	0.304318	1.40	3.818457	WMoO3	2.17	1.201071	0.348189	1.535377	-1.056840	0
3	4	196	0.102512	1.26	0.993863	KBaO3	8.02	-1.138629	0.159473	0.075408	0.740888	2
4	5	292	0.234745	0.48	1.927513	ZnCrO3	7.79	0.394457	-0.891946	0.557990	0.670208	3
...	...	...	...	...	...	...	...	...	...	...	...	...
479	480	18623	0.175813	0.86	-1.342605	VZnO3	5.78	-0.288794	-0.379716	-1.132261	0.052527	2
480	481	18713	0.100716	1.98	0.002913	ZnBaO3	9.73	-1.159445	1.130013	-0.436793	1.266378	2
481	482	18823	0.210733	0.04	-2.630710	CrVO3	9.04	0.116072	-1.485054	-1.798054	1.054339	2
482	483	18887	0.341396	0.88	-0.053766	ZrMoO3	8.95	1.630952	-0.352757	-0.466089	1.026681	3
483	484	18910	0.187105	0.90	1.893716	ZnCoO3	0.78	-0.157869	-0.325797	0.540521	-1.483993	0

The classification of 483 electrocatalysts is entered into the kmeans_4 column with details of the number of electrocatalysts in each group as follow:

Table 8. Classification of 483 electrocatalysts

kmeans_4	
1	143
0	134
2	108
3	99

3.3. Data Classification and Correlation

The data from the data grouping is loaded into memory using the panda's module. Several other modules must also be loaded into memory for the classification and data correlation process.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
from matplotlib import style
import seaborn as sns

from sklearn.model_selection import train_test_split, GridSearchCV, cross_val_score
from sklearn.praprocessing import standardScaler
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from skleran. metrics import confusion_matrix
```

Figure 5. Data Classification and Correlation

At this stage, sklearn is still used, which provides a special class used for data classification. The dataset contains columns with object types, namely containing chemical formula information, so it is set aside first.

```
Data columns (total 6 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   formula   484 non-null    object
 1   FL_T      484 non-null    float64
 2   e_form_T  484 non-null    float64
 3   mu_b_T    484 non-null    float64
 4   DA_T      484 non-null    float64
 5   tipe      484 non-null    int64
dtypes: float64(4), int64(1), object(1)
```

Figure 6. Chemical formula information

Correlation between variables can be done using several methods, namely Pearson, Spearman, and Kendall. This study used Pearson, which produced the following results.

Table 9. Pearson correlation between the variables of data set

	FL_T	e_form_T	mu_b_T	DA_T	tipe
FL_T	1.000000	-0.561174	-0.063669	-0.031965	0.017081
e_form_T	-0.561174	1.000000	0.125967	-0.015515	0.070941
mu_b_T	-0.063669	0.125967	1.000000	-0.030594	-0.027956
DA_T	-0.031965	-0.015515	-0.030594	1.000000	0.386316
tipe	0.017081	0.070941	-0.027956	0.386316	1.000000

Correlation determines the closeness of the relationship between two or more different variables, as described by the correlation coefficient. The correlation coefficient is a coefficient that describes the

closeness of the relationship between two or more variables. The size of the correlation coefficient does not indicate a causal relationship between two or more variables but only describes the linear relationship between them. Correlation analysis is a statistical method used to determine a quantity that indicates the strength of the relationship between one variable and another without considering whether a particular variable is dependent on the other (Sekaran, 2010). According to Guilford (1956), correlation not only facilitates the analysis of linear relationships between variables but can also help identify the strength and direction of those relationships.

The correlation coefficient ranges from $-1 < 0 < 1$. If $r = -1$, the correlation is perfectly negative, meaning the significant influence of variable X on variable Y is very weak. If $r = 1$, the correlation is perfectly positive, meaning the significant influence of variable X on variable Y is powerful (Sudjana, 2005).

Based on the definition above, Pearson correlation is used to determine the closeness of the relationship between the dependent and independent variables. Pearson correlation ranges from -1 to 1. A positive value indicates a unidirectional and increasing relationship, while a negative value indicates a unidirectional and decreasing relationship. The level of closeness can be described as follows:

4. Conclusions

It can be concluded from this research that.

- Important parameters can be used to assess the performance of electrocatalysts: (1) overpotential, (2) current density or completeness (exchange current density), (3) turnover frequency (TOF), and (4) onset potential, the minimum voltage at which the actual catalytic reaction begins.
- These four parameters are related to other quantities available as a dataset, namely castelli perovskite. The Castelli perovskite dataset contains data cbm (conduction band energy level value), e_form (heat of formation in eV), Fermi energy level, Fermi bandwidth, material, chemical formula, gap gllbsc (electronic band gap), mu_b (magnetic moment), structure (crystal structure), vbm (valence band energy level value).
- The Castelli perovskite dataset is then connected to experimental data through the material's chemical formula. Nonprecious metal oxides include 483 materials (electrocatalyst materials).
- Based on the results of KMeans data processing, 483 electrocatalyst materials were grouped into 4 types. Thus, non-noble metal oxides can be categorized into 4 types. Correlation data processing shows that the current density correlates with the proposed electrocatalyst type.

References

[1] S. N. Steinmann, Q. Wang, and Z. W. Seh, "How machine learning can accelerate electrocatalysis discovery and optimization," *Mater. Horiz.*, vol. 10, no. 2, pp. 393–406, 2023, doi: 10.1039/d2mh01279k.

[2] Z. A. Che Ramli, S. K. Kamarudin, S. Basri, and A. M. Zainoodin, "Synthesis, Characterization and Potential of Pt-Ru Supported Carbon Nanocage (CNC) Electrocatalystfor Future DMFC," *Int. J. Integr. Eng.*, vol. 11, no. 7, Aug. 2019, doi: 10.30880/ijie.2019.11.07.025.

[3] B. Baruah and P. Deb, "Performance and application of carbon-based electrocatalysts in direct methanol fuel cell," *Mater. Adv.*, vol. 2, no. 16, pp. 5344–5364, 2021, doi: 10.1039/D1MA00503K.

[4] K.-C. Cheung, W.-L. Wong, D.-L. Ma, T.-S. Lai, and K.-Y. Wong, "Transition metal complexes as electrocatalysts—Development and applications in electro-oxidation reactions," *Coord. Chem. Rev.*, vol. 251, pp. 2367–2385, 2007.

[5] Y. Zhao, J. Raj, X. Xu, J. Jiang, J. Wu, and M. Fan, "Carbon Catalysts Empowering Sustainable Chemical Synthesis via Electrochemical CO_2 Conversion and Two-Electron Oxygen Reduction Reaction," *Small*, Feb. 2024, doi: 10.1002/smll.202311163.

[6] A. M. Rahmani *et al.*, "Artificial intelligence approaches and mechanisms for big data analytics: a systematic study," *PeerJ Comput. Sci.*, vol. 7, p. e488, Apr. 2021, doi: 10.7717/peerj-cs.488.

[7] Y. Hu, J. Chen, Z. Wei, Q. He, and Y. Zhao, "Recent advances and applications of machine learning in electrocatalysis," *J. Mater. Inform.*, vol. 3, no. 3, Aug. 2023, doi: 10.20517/jmi.2023.23.

[8] R. Cheng, Y. Min, H. Li, and C. Fu, "Electronic structure regulation in the design of low-cost efficient electrocatalysts: From theory to applications," *Nano Energy*, vol. 115, p. 108718, Oct. 2023, doi: 10.1016/j.nanoen.2023.108718.

[9] K. Tran and Z. W. Ulissi, "Active learning across intermetallics to guide discovery of electrocatalysts for CO2 reduction and H2 evolution," *Nat. Catal.*, vol. 1, no. 9, pp. 696–703, Sep. 2018, doi: 10.1038/s41929-018-0142-1.

[10] Y. Gu *et al.*, "Design and Application of Electrocatalyst Based on Machine Learning," *Interdiscip. Mater.*, vol. 4, no. 3, pp. 456–479, May 2025, doi: 10.1002/idm2.12249.

[11] R. Ding, J. Chen, Y. Chen, J. Liu, Y. Bando, and X. Wang, "Unlocking the potential: machine learning applications in electrocatalyst design for electrochemical hydrogen energy transformation," *Chem. Soc. Rev.*, vol. 53, no. 23, pp. 11390–11461, 2024, doi: 10.1039/d4cs00844h.

[12] X. Wu, F. Kang, W. Duan, and J. Li, "Density functional theory calculations: A powerful tool to simulate and design high-performance energy storage and conversion materials," *Prog. Nat. Sci. Mater. Int.*, vol. 29, no. 3, pp. 247–255, Jun. 2019, doi: 10.1016/j.pnsc.2019.04.003.

[13] W. Chen *et al.*, "Predicting Crystalline Material Properties with AI: Bridging Molecular to Particle Scales," *Ind. Eng. Chem. Res.*, vol. 63, no. 43, pp. 18241–18262, Oct. 2024, doi: 10.1021/acs.iecr.4c03224.

[14] M. Lu, S. Rao, H. Yue, J. Han, and J. Wang, “Recent Advances in the Application of Machine Learning to Crystal Behavior and Crystallization Process Control,” *Cryst. Growth Des.*, vol. 24, no. 12, pp. 5374–5396, Jun. 2024, doi: 10.1021/acs.cgd.3c01251.
[15] G. Huang, Y. Guo, Y. Chen, and Z. Nie, “Application of Machine Learning in Material Synthesis and Property Prediction,” *Materials*, vol. 16, no. 17, p. 5977, Aug. 2023, doi: 10.3390/ma16175977.
[16] M. Wolski and A. Gomolińska, “Data meaning and knowledge discovery: Semantical aspects of information systems,” *Int. J. Approx. Reason.*, vol. 119, pp. 40–57, Apr. 2020, doi: 10.1016/j.ijar.2020.01.002.
[17] J. Wang, Y. Liu, P. Li, Z. Lin, S. Sindakis, and S. Aggarwal, “Overview of Data Quality: Examining the Dimensions, Antecedents, and Impacts of Data Quality,” *J. Knowl. Econ.*, vol. 15, no. 1, pp. 1159–1178, Mar. 2024, doi: 10.1007/s13132-022-01096-6.
[18] C. P. Marshall, J. Schumann, and A. Trunschke, “Achieving Digital Catalysis: Strategies for Data Acquisition, Storage and Use,” *Angew. Chem. Int. Ed.*, vol. 62, no. 30, p. e202302971, Jul. 2023, doi: 10.1002/anie.202302971.
[19] X. Jia and H. Li, “Data mining of stable, low-cost metal oxides as potential electrocatalysts,” *Artif. Intell. Chem.*, vol. 2, no. 1, p. 100065, Jun. 2024, doi: 10.1016/j.aichem.2024.100065.
[20] L. Ward *et al.*, “Matminer: An open source toolkit for materials data mining,” *Comput. Mater. Sci.*, vol. 152, pp. 60–69, Sep. 2018, doi: 10.1016/j.commatsci.2018.05.018.
[21] X. Jia and H. Li, “Machine learning enabled exploration of multicomponent metal oxides for catalyzing oxygen reduction in alkaline media,” *J. Mater. Chem. A*, vol. 12, no. 21, pp. 12487–12500, 2024, doi: 10.1039/d4ta01884b.