Effect of data augmentation on prediction performance of artificial neural network model for solid catalyst datasets

BY
Mr. Kittithat KONGTAWORN

A Thesis Submitted in Partial Fulfillment of the Requirements
for Master of Engineering CHEMICAL ENGINEERING
Department of CHEMICAL ENGINEERING
Silpakorn University
Academic Year 2025

Effect of data augmentation on prediction performance of artificial neural network model for solid catalyst datasets

By
Mr. Kittithat KONGTAWORN

A Thesis Submitted in Partial Fulfillment of the Requirements
for Master of Engineering CHEMICAL ENGINEERING
Department of CHEMICAL ENGINEERING
Silpakorn University
Academic Year 2025

Thesis Title	Effect of data augmentation on prediction performance of artificial neural network model for solid catalyst datasets
By	Mr. Kittithat KONGTAWORN
Field of Study	CHEMICAL ENGINEERING
Thesis Advisor	Assistant Professor Dr. Nutchapon Chotigkrai

Faculty of Engineering and Industrial Technology, Silpakorn University in Partial Fulfillment of the Requirement for the Master of Engineering

..... Dean of the Faculty of Engineering and Industrial Technology

(Assistant Professor Dr. Arunsri Leejeerajumnean)

Thesis COMMITTEE

..... Chairperson

(Assistant Professor Dr. Tarawipa Puangpetch)

..... Advisor

(Assistant Professor Dr. Nutchapon Chotigkrai)

..... Committee

(Dr. Sunthon Piticharoenphun)

..... External examiner

(Associate Professor Dr. Ekrachan Chaichana)

630920047: MAJOR CHEMICAL ENGINEERING
Keyword: Machine Learning, Stacking Ensemble Model (SEM), Data Augmentation, CO_2 Conversion, Small Datasets, Principal Component Analysis (PCA)

Mr. Kittithat KONGTAWORN : Effect of data augmentation on prediction performance of artificial neural network model for solid catalyst datasets Thesis Advisor : Assistant Professor Dr. Nutchapon Chotigkrai,

This research develops and evaluates machine learning models for predicting the outcomes of CO_2 conversion (to methane and methanol) and 5-HMF production, specifically addressing the challenge of limited experimental data. A comparison between Multilayer Perceptron (MLP) and Stacking Ensemble Model (SEM) revealed that the SEM consistently yielded superior accuracy across all datasets (e.g., CO_2 to methanol RMSE of 0.0728 vs. MLP's 0.0915). Data augmentation proved critical, transforming the non-viable 5-HMF model (13 data points, negative R^2) into a highly predictive one (R^2 = 0.9943). Moreover, the study discovered a model-dependent effect of Principal Component Analysis (PCA): it degraded the performance of the SEM while enhancing the accuracy of the MLP. These findings establish SEM as the superior architecture and demonstrate that data processing strategies must be carefully selected based on the dataset and model context.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to the many individuals who have provided invaluable support, guidance, and encouragement throughout the completion of this thesis.

My sincerest appreciation goes to Asst. Prof. Dr. Nutchapon Chotigkrai, my thesis advisor, for his insightful guidance, constructive criticism, and unwavering support throughout this research. His expertise and dedication were instrumental in bringing this work to fruition.

I am deeply grateful to Asst. Prof. Dr. Tarawipa Puangpetchm, Dr. Sunthon Piticharoenphun and Assoc. Prof. Dr. Ekrachan Chaichana for their valuable suggestions and recommendations, which greatly improved the quality and completeness of this thesis.

I would also like to express my sincere gratitude to Asst. Prof. Dr. Panyanat Aonpong for his invaluable assistance with the machine learning aspects of this research and for providing the initial framework for the data augmentation code.

I would like to extend my thanks to the Department of Chemical Engineering, Faculty of Engineering and Industrial Technology, Silpakorn University , for providing the necessary facilities, equipment, and a conducive environment for this research. I am also particularly grateful for the department's financial support for my tuition fees and for providing the travel grants that enabled the presentation of this work at academic conferences.

I am also indebted to the research groups of R. Yıldırım and Manu Suvarna for providing the foundational datasets, and to Mr. Poramathe Jarunothai for providing the 5-HMF dataset, which was crucial for this study.

I am thankful to my friends, colleagues, and fellow researchers in the department for their helpful discussions, shared knowledge, and constant encouragement.

Finally, I wish to express my heartfelt gratitude to my mother and family for their unconditional love, endless support, and profound encouragement throughout my academic journey.

Kittithat KONGTAWORN

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

Chapter I

Introduction

1.1 Motivation

Nowadays, disasters and natural disasters are becoming more serious. Many countries around the world are experiencing climate change and rising temperatures. This is a result of global warming caused by human activities that increase carbon dioxide in the atmosphere. The greenhouse effect causes the global average temperature to rise. The report of the study of The Intergovernmental Panel on Climate Change (2021) reveals that the global temperature has increased by 1.09 degrees Celsius from the highest concentration of carbon dioxide in 2 million years. Thailand in 2021 had the highest CO_2 emissions in the industrial sector, totaling 76.5 million tons of CO_2, an increase of 9.9% from the previous year, in line with the production of industrial products in 2021. expansion, especially the production of the main industries, namely the automotive industry steel industry. The capture and catalytic conversion of CO_2 to fuels or commodity chemicals (such as methane, methanol, ethanol, formic acid, etc.), as an effective approach, have been extensively studied. Among the numerous target products, methane and lower olefins (C_2-C_4) are a promising one as it is in great demand and feasible for cost-effective distributions. Furthermore, it was recently proposed that the hydrogen required for the hydrogenation reaction can be obtained by water electrolysis using renewable but intermittent electricity. In this respect, CO_2 methanation also provides a practical solution for the storage and transportation of low-grade energies. CO_2 utilization technologies can not only provide a pathway to reduce greenhouse gas emissions but may also enable renewable energy to be incorporated into important materials used in the society such as fuel and chemicals [1, 2].

The CO_2 methanation reaction ($CO_2 + 4H_2 + \longrightarrow CH_4 + 2H_2O$, $\Delta H_{298} = -164$ kJ/mol) [3]. In recent years, the potential of the CO_2 methanation reaction in both CO_2 abatement and hydrogen storage has attracted much attention. This reaction, being highly exothermic, is favorable in low-temperature regions. High temperatures can yield

the production of CO as the byproduct and accelerate the catalyst deactivation process, such as sintering. Therefore, operating the methanation reaction at a relatively low temperature to ensure higher methane selectivity and catalyst durability is one of the major. Lower olefins are important in the chemical industry. Traditional production routes use fossil feedstocks [4]. This is due to the increasing awareness of sustainable production. Non-fossil production routes and the transition to a circular economy is therefore gaining a lot of attention [5, 6]. Although a portion of olefin-based polymers can be recycled mechanically [7], the majority still needs to be recycled by other means [8]. Direct chemical recycling to olefin monomers is virtually impossible. Because selective breaking of the C–C bond and dehydrogenation are difficult [9], possible ways to close the carbon cycle are I) gasification or incineration of the polymer waste and II) Ethylene and propylene are synthesized via hydrogenation of the generated CO/CO_2 or CO_2 stream alone [10], while incineration of plastic waste is commercialized [11], and recently, studies The gasification of these wastes is intensive [12, 13]. Different routes have been proposed for hydrogenation from direct carbon dioxide to hydrocarbons. The feasibility is shown by the modified Fischer–Tropsch (MFT) synthesis. It consists of two main reactions in succession: a reverse water gasification (RWGS) reaction to produce CO from CO_2, followed by further conversion of CO to hydrocarbons via the Fischer– Tropsch reaction [14-16].

Machine learning (ML) is a branch of artificial intelligence that has been widely applied in recent years [17]. The main purpose of machine learning (ML) is to improve and optimize the performance of computer programs or algorithms by enabling them to automatically learn from data or past experience. Machine learning can create stable models by learning and mining existing data. and use these models to predict or classify unknown data. especially Since the advent of the big data era, ML has enabled built models to make more timely and accurate predictions than ever before [18]. Over the years, ML has also been used. In the field of organic chemistry to control the efficiency of chemical production [19], ML-based data analysis technology has become one of the most active research topics and development trends. The challenges of limited datasets in organic chemistry and experimental fields,

compounded by factors like cost and the COVID-19 pandemic, emphasize the need for a scientific algorithm addressing small dataset limitations in machine learning (ML) modeling. Traditional continuous training on large datasets is often impractical due to resource constraints, necessitating strategies to enhance ML model performance with limited data. Potential overreliance on original data and overfitting risks in small datasets require exploration of innovative algorithms or techniques to improve generalization. Researchers must actively seek solutions, incorporating domain knowledge, feature engineering, and advanced techniques amidst challenges in obtaining experimental data, ensuring robust ML models for small datasets amidst disruptions like the pandemic [20]. Numerous methods have been suggested to address the aforementioned challenges, with data augmentation standing out as the most prominent. Data augmentation involves generating additional data by transforming or characterizing the existing dataset. Primarily utilized to mitigate substantial errors associated with constructing prediction models using limited datasets, data augmentation has gained widespread recognition. Over the past decade, various data augmentation algorithms, including but not limited to the variational auto-encoder (VAE) and generative adversarial network (GAN), have demonstrated their effectiveness in enhancing model performance [21]. To extend and enhance the Finite Element Method (FEM)-based machine learning approach while reducing the need for experimental data collection, an efficient method known as transfer learning has been proposed, aiming to mitigate synthetic simulation effects [22]. While positive outcomes have been observed in the context of soft pneumatic actuators comprised of a single material or single structure, limited investigations have been conducted on multi-material Soft Fluidic Robotic Actuators (SFRBAs). Consequently, there is a pressing need to formulate a robust framework based on Machine Learning Algorithms (MLAs) for modeling SFRBAs, addressing challenges associated with high structural nonlinearity [23].

1.2 Objective of research

To develop enhanced predictive models for catalytic carbon dioxide methanation, carbon dioxide hydrogenation to methanol and 5-hydroxymethylfurfural using data processing and machine learning techniques.

1.3 Scope of research

In this study, machine learning was performed using a Multilayer Perceptron model, as well as a Stacking Ensemble Model incorporating Linear Regression, Ridge, Lasso, ElasticNet, Random Forest, Gradient Boosting, and MLP. These models were applied to carbon dioxide methanation, carbon dioxide hydrogenation to methanol, and 5-hydroxymethylfurfural production under different multiplier training datasets and beta values, using dimensionality reduction with Principal Component Analysis, as shown in **Table 1.1**.

Table 1.1 All experiments in machine learning

Dataset	CO_2 to Methane CO_2 to Methanol		5-HMF	
Model / Object	Multilayer Perceptron (MLP)			
	Stacking Ensemble Model (SEM): I) Linear Regression II) Ridge III) Lasso IV) ElasticNet V) RandomForest VI) GradientBoosting VII) MLP			
	Folds	Beta	Folds	Beta
Data augmentation	1x 2x 5x 10x	0.05 0.1 0.2	1x 10x	0.1
Dimensionality Reduction	Principal Component Analysis (PCA)		-	

1x is the original training dataset, 2x, 5x and 10x are 2-, 5- and 10-times size of original training dataset.

Chapter II

Literature review and Theory

2.1 Data analysis of catalytic carbon oxide conversion

2.1.1 Carbon oxide methanation

The methanation of CO_2 is a multifaceted process characterized by competitive side reactions and the inert nature of CO_2. Successful implementation of this process relies on the design of high-performance catalysts that can enhance selectivity and methane yield. Extensive research has been conducted to investigate the impact of various factors, including the choice of active metal, the role of support material, and the influence of the preparation method, all aimed at developing efficient catalysts [24].

In this paper of Beyza Yilmaz [24], the focus is on presenting and discussing simple descriptive statistics. These statistics are crucial for evaluating the potential impact of the aforementioned variables on catalytic performance. The objective is to gain insights into how the choice of active metal, the support material, and the preparation method contribute to the overall efficiency of the catalyst in promoting CO_2 methanation. Transition metals, particularly nickel (Ni), ruthenium (Ru), cobalt (Co), and palladium (Pd), have demonstrated remarkable catalytic prowess in methane reactions involving carbon dioxide. Extensive research has been dedicated to exploring the catalytic capabilities of noble and transition metals, given their crucial role in facilitating the conversion of CO_2 and selectivity CH_4 [25]. Notably, the metals Ni, Ru, Co, and Pd are commonly employed as base metals, as indicated in the dataset, with occurrences of 3227/4051, 396/4051, 110/4051, and 93/4051, respectively. Iron (Fe) is another metal discussed in the context of this study, sourced from papers [26, 27]. Additionally, rhodium (Rh) is referenced from papers [28, 29], highlighting its relevance in these catalytic processes. These metals, including Fe and Rh, play significant roles in methane reactions involving CO_2, as evidenced by various studies [30].

Nickel (Ni) emerges as the predominant and extensively researched metal for CO_2 methanation, attributed to its favorable characteristics, including high activity,

remarkable CH_4 selectivity, and cost-effectiveness, as depicted in **Figure 2.1(Left)** where the ball size signifies the number of data points featuring Ni as the base metal. Ni-based catalysts dominate CO_2 methanation research due to their high activity and cost-effectiveness, leading to the highest CO_2 conversion. Relying solely on overall average conversion can be misleading, as data clusters around near-zero and 75% conversion values, revealing variability not captured by the average. Despite challenges, Ni catalysts consistently outperform, while Ru and Rh catalysts, though pricier, show varied performance. Co (more expensive than Ni) and Fe (cheaper than Ni) catalysts exhibit decent catalytic activity. Pd-based catalysts, while used in numerous studies, tend to produce CO as the major product, suppressing CH_4 formation. The activity of catalysts can be enhanced by altering metal loading, supports, synthesis methods, and process conditions. Metal loading is one of the parameters that affects metal dispersion and in turn impacts CO_2 methanation; for instance, 20% Ni, as shown in **Figure 2.2(Left)**, increases CO_2 conversion while further increase does not affect performance.

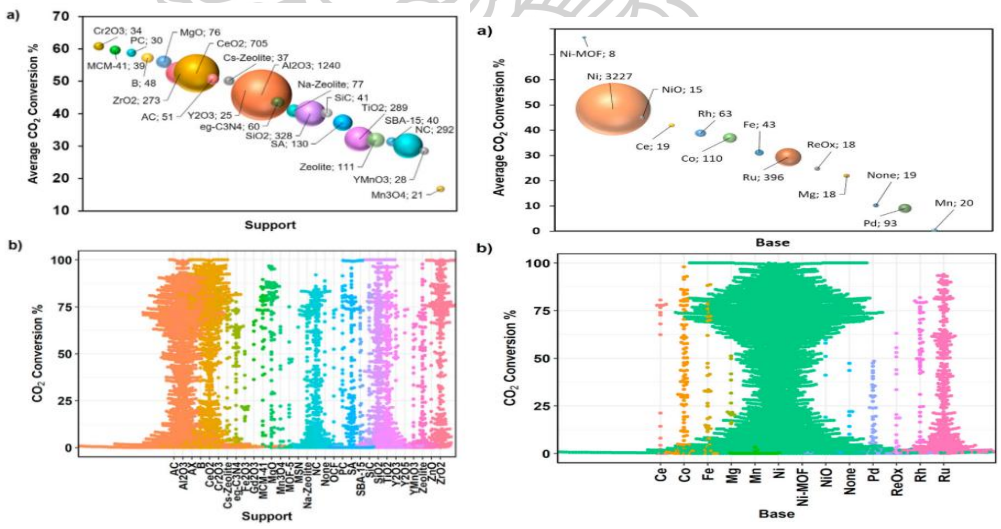

Figure 2.1 CO_2 conversion% for base materials (Left), and for support materials (Right) [24].

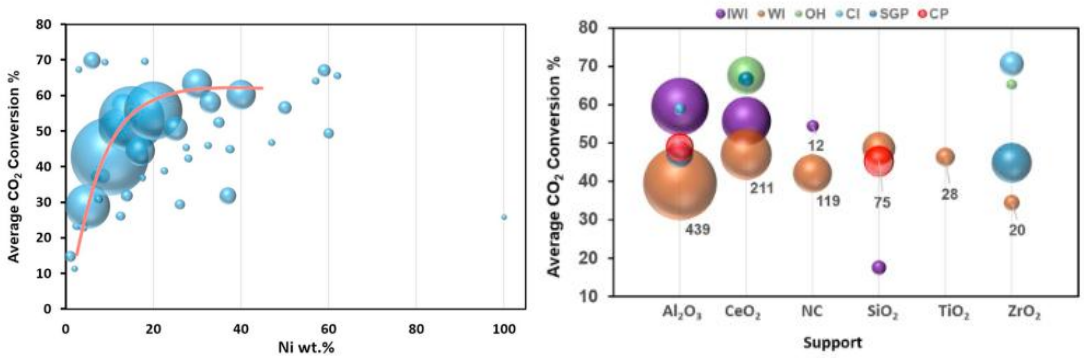

Figure 2.2 Effect of Ni wt.% (left), and Effect of catalyst preparation methods (Right) [24].

Support materials significantly impact the catalytic performance of methanation catalysts, enhancing activity through metal-support interactions, large surface area, extra active sites, high electron transfer, and improved metal dispersion. Metal oxides (Al_2O_3, CeO_2, SiO_2, TiO_2, ZrO_2), zeolites, and carbon materials (NC, AC) are widely investigated as supports (**Figure 2.1(Right)**). Al_2O_3, the most common support, attracts attention for its large surface area and favorable pore structure. However, Al_2O_3-supported catalysts may face issues like carbon formation and sintering at elevated temperatures, often addressed with promoters or additives. CeO_2 support offers properties conducive to CO_2 conversion, such as oxygen storage capacity, metal-support interaction, surface basicity, and advanced metal dispersion. ZrO_2 support leads to high catalytic activity due to improved Ni dispersion, basicity, and oxygen vacancies. TiO_2 shows lower CO_2 conversion favoring CO production. Ni emerges as the dominant choice for all supports except TiO_2 in base metal-support pairing, often modified with promoters (alkali metals, alkaline earth metals, transition metals, metal oxides) to enhance stability, activity, and selectivity.

Catalyst preparation methods, influencing crystal structure, metal dispersion, loading, and interactions, significantly impact CO_2 methanation performance. Wetness impregnation is most common (1507 data points), followed by incipient to wetness impregnation (IWI) with 1082 data points. IWI shows better performance with Al_2O_3, CeO_2, and NC supports, whereas WI performs less with TiO_2. Other methods include

sol-gel precipitation (SGP), co-precipitation (CP), one-pot hydrolysis (OH), CI, and H. Performance variations with different supports are depicted in **Figure 2.2(Right)**.

2.1.2 Carbon oxide to methanol

In this paper of Manu Suvarna [31], The literature database for CO_2 hydrogenation to methanol includes four main catalyst families—Cu-, Pd-, In_2O_3-, and ZnO-ZrO_2-based—accounting for 55%, 14%, 26%, and 5% of the dataset, respectively, with over 13 promoters and 24 supports, primarily metal or mixed oxides (**Figure 2.3**). Performance visualization via a 2D scatter plot reveals clustering of In_2O_3- and ZnO-ZrO_2-based catalysts with higher methanol selectivities and STY, whereas Cu-based catalysts display broader distribution (**Figure 2.4**). In_2O_3-based catalysts were mostly studied as bulk In_2O_3, with Pd, Pt, and Ni as the main promoters. Pd-based catalysts commonly employed CeO_2 or ZnO supports, and ZnO-ZrO_2 catalysts were mostly solid solutions with emerging promoters such as Cu, Ga, and Pd. Cu/ZnO and Cu/ZnO/Al_2O_3 systems, due to strong Cu–ZnO interactions and industrial relevance, served as reference catalysts, sometimes promoted with Pd, Ga, La, or Y.

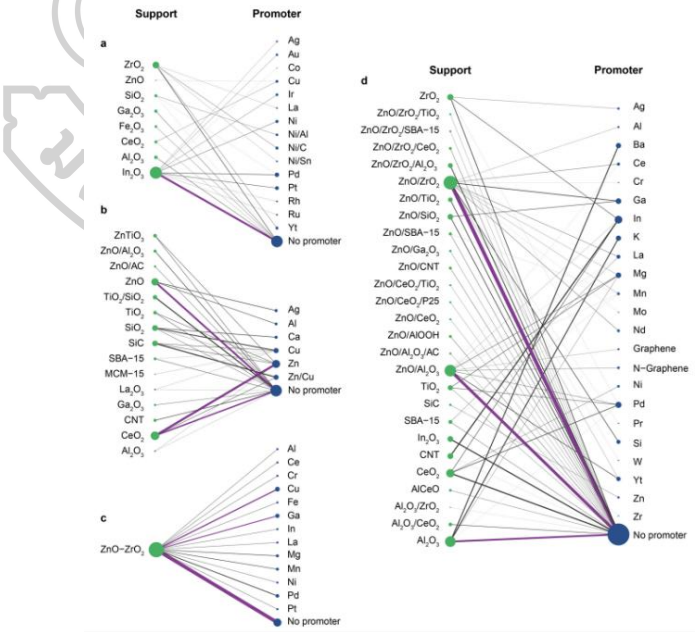

Figure 2.3 Network plots of supports (green) and promoters (blue) used in (a) In_2O_3-, (b) Pd-, (c) ZnO-ZrO_2-, and (d) Cu-based catalysts. Node size and link thickness indicate frequency, and prominent combinations are highlighted in purple [31].

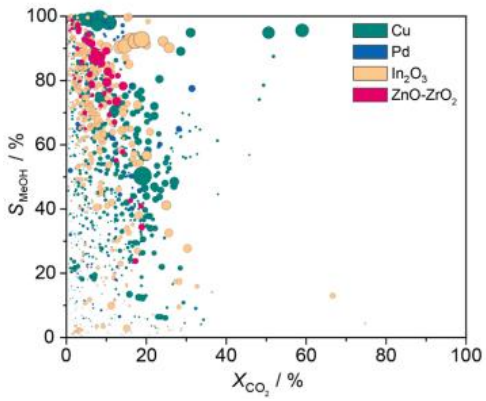

Figure 2.4 2D scatter plot of methanol selectivity versus CO_2 conversion for all catalyst families, with bubble color representing catalyst family and bubble size proportional to methanol STY [31].

Cu-based catalysts were predominantly prepared via precipitation (P) and coprecipitation (CP), accounting for ~43% of the dataset, suitable for Cu-rich systems (30–60 wt%) with additional components, allowing well-dispersed crystalline Cu within metal oxide domains. Impregnation (dry and wet, ~30%) facilitated surface dispersion of low promoter loadings, particularly in In_2O_3-based catalysts, while deposition-precipitation (~12%) selectively deposited promoters or active phases. Other methods, including flame spray pyrolysis, solvothermal, sol-gel, and microwave-assisted synthesis, were collectively classified as “others” (**Figure 2.5(a)**). Reaction temperature, pressure, and space velocity significantly affect methanol synthesis, with temperatures ranging from 433–673 K (**Figure 2.5(b)**), pressures mostly 3–5 MPa (**Figure 2.5(c)**), and GHSV varying from 2,000–24,000 $cm^3\ h^{-1}\ gcat^{-1}$ (**Figure 2.5(d)**). Additional reaction parameters, such as the feed H_2/CO_2 ratio and catalyst mass, were also analyzed, showing distributions across the dataset (**Figure 2.5(e-f)**). Inconsistencies in reporting formats and lack of standardized protocols pose challenges for machine learning applications, highlighting the need for uniform data collection and reporting practices.

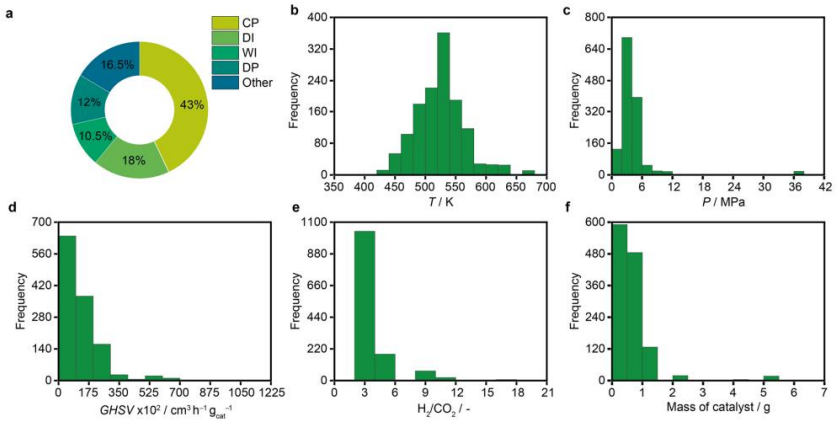

Figure 2.5 (a) Donut chart showing the percentage of commonly used synthesis methods: coprecipitation (CP), dry impregnation (DI), wet impregnation (WI), and deposition-precipitation (DP). Box plots illustrating the distribution, mean, and quartiles of reaction conditions: (b) temperature (T), (c) pressure (P), (d) gas-hourly space velocity (GHSV), (e) feed H_2/CO_2 ratio, and (f) catalyst mass, with x-axis representing the reaction parameters and y-axis representing data frequency [31].

2.2 Machine learning in catalysis

Machine learning, a subset of artificial intelligence, employs algorithms and statistical models to automatically learn patterns and make predictions. Through the training on data, machine learning models enhance their performance over time. This computational approach to data analysis and prediction finds applications in diverse fields like image recognition, natural language processing, and recommendation systems [23, 32]. In essence, machine learning is a facet of artificial intelligence that revolves around the development of algorithms capable of learning and making predictions or decisions without explicit programming. These algorithms learn from data, continually refining their performance. As a result, machine learning is extensively utilized in practical applications, including but not limited to image recognition, natural language processing, and predictive modeling [33].

2.2.1 Machine learning of carbon dioxide methanation

Beyza Yilmaz's research [24] compiled a CO_2 methanation dataset from 100 selected papers using Web of Science. With 4,051 data points from 527 experiments,

it includes 23 descriptors on catalyst, preparation, and conditions. Experiments were categorized based on catalyst variations to prevent data leakage. Simple descriptive statistics and random forest (RF) were used, and RF predicted CO_2 conversion profiles successfully. In **Figure 2.6**, The RF results show an RMSE and R^2 of 6.4 and 0.97 for the training set and 12.7 and 0.85 for the testing set, respectively.

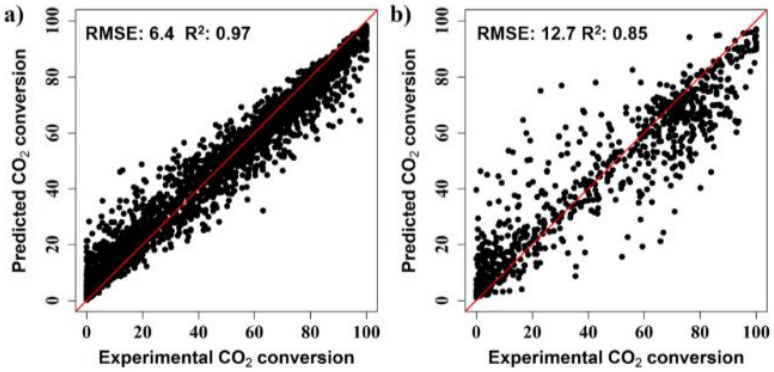

Figure 2.6 Predicted vs real conversion for a) training set b) testing set [24].

2.2.2 Machine learning of carbon dioxide to methanol

Suvarna et al [31]. conducted predictive analytics using three ensemble learning algorithms—Random Forest (RF), Gradient Boosting Decision Tree (GBDT), and Extreme Gradient Boosting (XGB)—to model methanol STY based on catalyst properties, synthesis conditions, and reaction parameters. The comparative evaluation of model performance is presented in joint scatter plots of actual versus predicted methanol STY (**Figure 2.7**), showing that while all algorithms performed well on both training and test datasets, XGB exhibited the highest predictive accuracy with superior R^2 and the lowest RMSE. Cross-validation using k-fold (k = 5) confirmed that the models were not overfitted, as the CV R^2 values closely matched the test R^2 for all algorithms. Simulation experiments varying metal covalent radii and promoter electronegativity, while keeping other features constant, indicated that these parameters alone did not significantly affect the predicted methanol STY, highlighting the dominant influence of catalyst composition and reaction conditions, which contributed to 97% of the model variance. Overall, **Figure 2.7** demonstrates the robust predictive capability of XGB for methanol STY based on literature-derived descriptors.

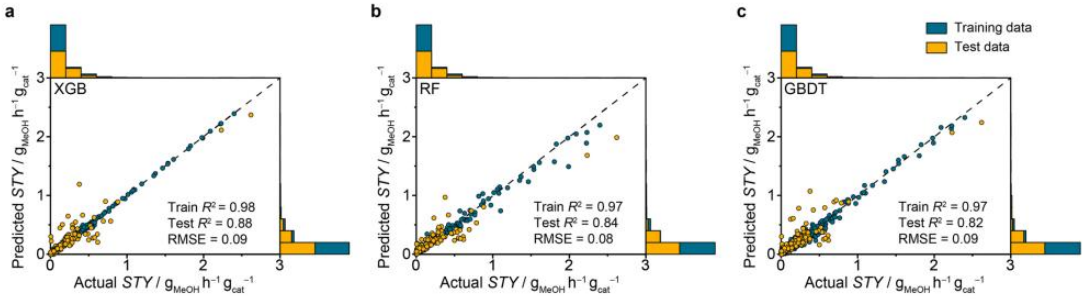

Figure 2.7 presents the predicted methanol STY by (a) XGB, (b) RF, and (c) GBDT. The comparison shows that XGB achieved the most accurate and generalized performance on both training and test datasets. Five-fold cross-validation was used during hyperparameter tuning to ensure model generalization, and the test RMSE values are reported in units of gMeOH h^{-1} $gcat^{-1}$ [31].

2.2.3 Multilayer perceptron algorithm (MLP)

The Multi-Layer Perceptron (MLP) is a powerful artificial neural network architecture widely recognized for its versatility and effectiveness in modeling and prediction tasks. Its strength lies in its proficiency in handling complex non-linear relationships between input and output variables, making it applicable across a diverse range of fields. With the capacity to learn and generalize from extensive datasets, MLP excels in making accurate predictions even in the presence of noise or uncertainty. MLP's ability for parallel processing enhances its efficiency, enabling the simultaneous handling of multiple tasks. Through the utilization of back-propagation techniques during training, MLP optimizes the network by adjusting weights and biases, leading to improved performance. Its ability to model and capture intricate patterns and relationships within the data positions it as a potent tool for data analysis and prediction. Furthermore, MLP excels in managing complex nonlinear relationships between input and output variables. Its flexibility in model configuration and optimization is attributed to the availability of various activation functions and solvers during training. This adaptability allows MLP to be tailored to the specific requirements of different tasks, further enhancing its versatility in diverse applications. In addition to its application in various domains, the Multilayer Perceptron (MLP) is a widely utilized form of Artificial Neural Network (ANN) known for its practicality and effectiveness. Its

inclusion of multiple hidden layers, an input layer, and an output layer allows it to adeptly learn intricate relationships between inputs and outputs. The utilization of nonlinear activation functions enables MLP to generate output based on the weighted sum of inputs, facilitating the modeling and prediction of complex systems or processes. The training process, involving algorithms like backpropagation, plays a crucial role in adjusting the weights to minimize errors, refining MLP's ability to accurately capture and represent underlying patterns in the data. This makes MLP a robust tool for various artificial intelligence and machine learning tasks [34-36].

In the study conducted by M. Anwar Hossain [35], the production of hydrogen-rich syngas from methane dry reforming over Ni/$CaFe_2O_4$ catalysts was modeled using artificial neural networks (ANN), specifically employing multi-layer perceptron (MLP) and radial basis function (RBF) architectures. The Ni/$CaFe_2O_4$ catalysts underwent synthesis and characterization through XRD, SEM, EDX, and FTIR techniques. For training and optimization of the ANN models, 70% of the experimental data were used for training, and the remaining 30% were allocated for testing and validation. The optimized conditions for the ANN-RBF model included 10 hidden neurons, while the ANN-MLP model had 12 hidden neurons. The evaluation of ANN model performance was based on the coefficient of determination (R^2) values for H_2 yield, CO yield, CH_4 conversion, and CO_2 conversion. The ANN-MLP model demonstrated higher R^2 values compared to the ANN-RBF model. Statistical analysis, including a t-test, was performed to assess the significance of the models. The results indicated that the MLP-based model outperformed the RBF-based model in predicting hydrogen-rich syngas production. The ANN models were deemed statistically significant based on the t-test results. The paper also presented a schematic diagram illustrating the experimental setup for hydrogen-rich syngas production from methane dry reforming. Parity plots were employed to compare observed and predicted CH_4 and CO_2 conversions, revealing that the MLP-based model exhibited better predictive capability than the RBF-based model.

In Xuejin Sun's paper [36], Machine Learning (ML) approaches, specifically K-Nearest Neighbor (KNN) regression, Decision Regression Tree (DT), and Multilayer

Perceptron (MLP), were employed to simulate and optimize the biodiesel production process. The study focused on predicting biodiesel production yield (%) from Soybean oil through the transesterification process, with input variables including the molar ratio of methanol to oil and catalyst loading. Comparisons of the KNN regression, DT, and MLP models revealed that all exhibited high accuracy, with R^2 values surpassing 0.9. The MLP model, in particular, demonstrated superior accuracy and was selected for optimizing biodiesel production yield. Through this optimization, the MLP model achieved a yield of 83.88% using specific input values. The DT model, structured as a binary tree, was utilized to hierarchically and sequentially represent conditions for predicting biodiesel production. MLP models, with hidden layers, were fine-tuned by adjusting hyperparameters such as hidden layer size, activation functions, and solver functions to optimize their performance. In the results of the paper, all ML models, including KNN, DT, and MLP, exhibited high accuracy with R^2 values exceeding 0.9, indicating their effectiveness in predicting biodiesel production yield. The MLP model, with a root mean square error (RMSE) of $4.9460E^{-01}$, demonstrated its ability to accurately predict biodiesel production yield. The paper showcased the utility of ML techniques, particularly MLP, in simulating and optimizing biodiesel production processes.

2.2.4 Stacking ensemble model (SEM)

The theory of the Stacking Ensemble Model, also known as Stacked Generalization, was first introduced in the research paper “Stacked Generalization” by Wolpert in 1992 [37]. It is an advanced ensemble technique designed to build models that minimize errors in generalization to unseen data (Generalization Error) [38]. The architecture of this model consists of two levels as defined by Wolpert: Level 0, which comprises diverse Base Learners, and Level 1 [39], a Meta-Learner responsible for learning how to optimally combine the predictions from all Base Learners [40, 41]. A key aspect of this theory is that the Meta-Learner is not trained on the original data but instead uses the predictions generated by the Base Models as attributes in a new dataset [42]. To prevent overfitting, these predictions are produced through a cross-validation process, ensuring that the Meta-Learner learns from data unseen by the

Base Models during training [43]. Consequently, Stacking can enhance both accuracy and robustness of predictions compared to any single model [44], as demonstrated in various studies such as “A Stacking Ensemble Learning-Based Financial Fraud Prediction Framework” [45] and “Stacking Ensemble Learning in Deep Domain Adaptation for Ophthalmic Image Classification” [46].

Abdallah Abdellatif [47] introduced an advanced machine learning model called “Stack-ETR” for accurate day-ahead photovoltaic (PV) power generation forecasting. The study addresses the challenges of uncertainty and variability in PV power output, which affects grid stability as PV adoption increases. The proposed Stack-ETR employs a stacking ensemble architecture, integrating Random Forest Regressor (RFR), Extreme Gradient Boosting (XGBoost), and Adaptive Boosting (AdaBoost) as base learners, with Extra Trees Regressor (ETR) serving as the meta-learner to refine final predictions. Using four years (2018–2021) of real meteorological data from three PV system types in Malaysia—thin-film, monocrystalline, and polycrystalline—the model demonstrated superior performance over other models. Compared to the base ETR, Stack-ETR reduced RMSE by 24.49%, 40.2%, and 27.95%, and MAE by 28.88%, 47.2%, and 40.88% for the respective PV systems. These results highlight Stack-ETR’s effectiveness and reliability in improving PV power forecasting accuracy, enhancing grid stability, and supporting renewable energy integration.

2.3 Data processing

2.3.1 Data augmentation

In the realm of experiment-based fields such as chemical reaction, the challenge of small sample sizes can be effectively addressed through the application of data augmentation techniques. These methods, such as variational auto-encoder and generative adversarial network algorithms, play a crucial role in preventing overfitting in machine learning models. By generating additional training data, data augmentation contributes to the creation of more robust and accurate models, particularly in scenarios with limited sample sizes [20]. The benefits of data augmentation extend beyond mitigating overfitting, encompassing improvements in

handling extreme values and enhancing prediction accuracy. Specifically, the augmentation of data proves instrumental in overcoming the limitations imposed by small sample sizes [48]. This process not only expands the volume of training data but also reduces randomness and elevates the quality of generated samples. Furthermore, the augmentation of data contributes to the enrichment of training datasets in terms of volume, quality, and diversity, thereby enhancing the overall performance of machine learning models. This is particularly valuable in situations where obtaining sufficient training data is challenging, a common scenario in real-world applications [49]. The advantages of data augmentation extend to various computer vision domains, including medical imaging, remote sensing, and agriculture. One notable aspect is the ability of data augmentation techniques to learn augmentation operations directly from the data, leading to increased computational efficiency compared to traditional methods of learning distributions over the input space [49]. The flexibility and adaptability of data augmentation make it a valuable tool in addressing challenges related to small sample sizes and improving the robustness and accuracy of machine learning models across diverse applications.

In the research conducted by Sizhou Wei [20], the application of data augmentation techniques, including variational auto-encoder and generative adversarial network algorithms, played a pivotal role in enhancing the modeling of the bio-polymerization process. The study employed both random forest and artificial neural network algorithms, with a particular focus on continuous correction and evaluation of hidden neurons. Additionally, the data underwent normalization using the standardized Z-score to alleviate dimensional effects and improve the stability and efficiency of the network. The results of the paper demonstrated the efficacy of data augmentation techniques in improving the performance of regression models within the bio-polymerization process. Specifically, the random forest model, augmented by the generative adversarial network technique, exhibited the most promising outcomes in predicting molecular weight. It achieved an R^2 of 0.94 on the training set and 0.74 on the test set, with a coefficient of determination of 0.74. This study underscores the significant impact of data augmentation techniques, particularly those involving

generative adversarial network algorithms, in substantially enhancing the performance of regression models in bio-polymerization processes. The combination of the random forest model with data augmentation emerged as the most accurate approach for predicting molecular weight. These findings underscore the potential of leveraging machine learning techniques and data augmentation to address challenges related to small sample sizes and to enhance control strategies in bio-polymerization processes.

In the study conducted by Zherui Ma [48], a novel approach for sample generation in the context of supercritical water gasification (SCWG) hydrogen yield prediction is proposed. This method leverages a Generative Adversarial Network (GAN) and implements a hybrid strategy for data augmentation. The hybrid strategy incorporates cross-validation and looping iteration methods to mitigate the inherent randomness of GANs, thereby enhancing the quality of generated samples. To construct a hydrogen yield prediction model, the paper adopts the Least Squares Support Vector Regression (LSSVR). The evaluation of the model's performance improvement employs metrics such as the promoting percentage of the root mean square error (P_RMSE), the promoting percentage of the mean absolute error (P_MAE), and the promoting percentage of the R^2 coefficient of determination (P_R^2). Additionally, the Shapley Additive explanations (SHAP) theory is employed to interpret and analyze the prediction results, enhancing result interpretability. In the results of the paper, the proposed hybrid-GAN sample generation method effectively enhances the prediction accuracy of the SCWG model, particularly when dealing with small-sample datasets. After expanding the small sample dataset using the single GAN, the prediction model experiences a notable reduction in the average root mean square error (RMSE) and mean absolute error (MAE) by 20.43% and 11.58%, respectively. The hybrid-GAN samples significantly contribute to the training of the prediction model, leading to improved performance and decreased errors. Notably, the prediction model trained with the generated SCWG samples exhibits low average RMSE and MAE (1.1737 mol/kg and 1.0171 mol/kg, respectively) with an average R^2 of 0.9457. The hybrid strategy, incorporating cross-validation and looping iteration methods, further improves the quality and reliability of generated samples, thereby enhancing the training effect

of the model and improving its prediction performance. These findings underscore the efficacy of the proposed hybrid-GAN approach in addressing small-sample challenges in SCWG hydrogen yield prediction, showcasing advancements in both accuracy and interpretability.

Alhassan Mumuni's paper [49] provides a comprehensive review of data augmentation methods in computer vision. The study covers diverse strategies, including deep learning, meta-learning, and 3D graphics modeling. A comparative analysis of state-of-the-art methods is conducted, with insights into their effectiveness across various scenarios and datasets. The paper emphasizes the significance of fundamental geometric transformations and explores specific strategies, such as patch combinations and class-imbalance techniques, particularly relevant in object detection applications. The results section provides a comprehensive review of recent and advanced data augmentation techniques in computer vision. The authors meticulously compare the performance of state-of-the-art methods, offering insights into their effectiveness across diverse scenarios and datasets. The study evaluates the impact of geometric transformations, revealing instances where simultaneous application can detrimentally affect performance in controlled settings. Additionally, the paper underscores the effectiveness of specific strategies, such as polygon occlusion with random cropping, in improving image classification performance. Performance results on well-known datasets like CIFAR-10, CIFAR-100, and ImageNet demonstrate significant improvements over baseline configurations. The authors emphasize the value of combining traditional and novel augmentation methods, recognizing potential variations in gains across different datasets.

2.3.2 Principal Component Analysis

Principal Component Analysis (PCA) is a foundational multivariate statistical technique used for dimensionality reduction, designed to simplify complex datasets by transforming a large set of potentially correlated variables into a smaller, uncorrelated set of new variables called principal components. Its primary objective is to retain as much of the original dataset’s variability or statistical information as

possible while reducing the number of dimensions, making the data easier to interpret and analyze. PCA achieves this by creating a new coordinate system through an orthogonal linear transformation, where the first principal component (PC_1) lies in the direction of maximum variance in the data, and each subsequent component is orthogonal to the previous ones, capturing progressively less remaining variance. This hierarchical ordering ensures that the initial components summarize the most significant patterns in the dataset. Mathematically, PCA is equivalent to solving an eigenvalue-eigenvector problem for the data's covariance or correlation matrix, with the eigenvectors defining the directions of the principal components and the eigenvalues indicating the variance magnitude captured by each component. First introduced by Karl Pearson in 1901 and later independently extended by Harold Hotelling in the 1930s, PCA has become a cornerstone of exploratory data analysis and is widely applied to enhance data visualization, reduce multicollinearity, and improve the efficiency of machine learning algorithms by creating a more tractable representation of high-dimensional datasets. Its combination of dimensionality reduction, variance preservation, and orthogonal transformation makes PCA an essential tool for extracting meaningful structure from complex, multivariate data in fields ranging from statistics and engineering to finance and bioinformatics [50-52].

María Soledad Callén [53] conducted research aimed at investigating the influence of key gasification parameters and biomass composition on the quality of the output gas, specifically the content of methane (CH_4), ethylene (C_2H_4), and bitumen, using multivariate statistical analysis with principal component analysis (PCA) and partial least squares (PLS) regression. Experimental data were collected from a TRL-4 gasification plant operating under enhanced adsorption (SEG) gasification with steam and CaO. PCA results indicated that five factors could explain up to 89% of the total variance, highlighting the significant role of biomass composition, while increasing the reactor bed temperature (T_{bed}) and the CaO/C ratio was found to reduce CH4 and bitumen contents. Furthermore, CH4, C2H4, and bitumen contents exhibited positive correlations with each other. The developed PLS model successfully predicted the contents of these compounds, with tar showing an R^2 (Calibration) of 0.91 and an

external prediction error (OE_{ext}) of 26.2% ($RMSEP_{ext}$ = 5.74 g/m³N), CH_4 achieving an R^2 (Calibration) of 0.96 and OE_{ext} of 7.9% ($RMSEP_{ext}$ = 0.92 vol%), and C_2H_4 yielding an R^2 (Calibration) of 0.86 and OE_{ext} of 16.3% ($RMSEP_{ext}$ = 0.34 vol%). The average total prediction errors ranged from 8% to 26%, indicating that tar reduction was driven by high furnace bed temperature, low thermal input, and high VOC biomass content. To achieve optimal outlet gas quality, it is therefore necessary to balance the control of the average furnace base temperature, the CaO/C adsorption-to-mass ratio, and the chemical composition of the biomass, while using these statistical tools as a valuable first step to better understand and optimize the SEG gasification process.

Chapter III

Methodology

3.1 Dataset

The original dataset used in this work was collected by R. Yıldırım's group and the details were described elsewhere. In summary, 4,051 data points were manually extracted from 100 published experimental papers using "CO_2 methanation" and "catalytic CO_2 hydrogenation as keywords in web of science [24], Manu Suvarna's group conducted a comprehensive literature review on thermocatalytic CO_2 hydrogenation to methanol, covering studies published between 1996 and 2021. In total, 1,234 data points were manually extracted from 131 experimental papers [31]. and Poramathe Jarunothai conducted experiments on 5-hydroxymethylfurfural (5-HMF) production, yielding a total of 13 data points [54]. The input parameters (features) and output parameter (target) were summarized in **Table 3.1**, **Table 3.2**, and **Table 3.3**.

Table 3.1 Input features used in CO_2 methanation

Category	Parameters	Range
Catalyst properties	Base	Ce, Co, Fe, Mg, Mn, Ni, Ni-MOF, NiO, None, Pd, ReOx, Rh, Ru
	Base wt. %	0–100
	Base	Al_2O_3, Ba, C, Ca, Ce, CeO_2, Co, Cs, Cu, Eu, Fe, Gd, K, La, La_2O_3, Li, Mg, Mn, Na, NC, Ni, None, Pr, Ru, Sm, Sr, VOx, W, Y, Y_2O_3, Yb, Zr
	Base wt. %	0-37.5
	Base 2 (co metal)	AC, Al_2O_3, AX, B, CeO_2, Cr_2O_3, Cs-Zeolite, eg-C_3N_4, Fe_2O_3, Gd_2O_3, MCM-41, MgO, Mn_3O_4, MOF-5, MSN, Na-Zeolite, NC, None, OCF, PC, SA, SBA-15, SiC, SiO_2, TiO_2,

		Y_2O_3, Y_2O_5, $YMnO_3$, Zeolite, ZnO, ZrO_2
	Base 2 wt. %	0-100
	Support	Al_2O_3, Ba, BaO, CaO, CeO_2, Cr_2O_3, CTAB, Gd_2O_3, K, MgO, MgO-Nd_2O_3, Nd_2O_3, None, Pr_2O_3, Sm_2O_3, SrO_2, Y_2O_5, ZrO_2
	Support wt. %	0-100
	Support 2 (co support)	Al_2O_3, Ba, BaO, CaO, CeO_2, Cr_2O_3, CTAB, Gd_2O_3, K, MgO, MgO-Nd_2O_3, Nd_2O_3, None, Pr_2O_3, Sm_2O_3, SrO_2, Y_2O_5, ZrO_2
Synthesis conditions	Catalyst preparation method	AE, CC, CI, commercial, CP, CP-MSM, DBD, DI, DP, DPA, DPU, EISA, FM, FSP, H, HTP, IWI, IWI-DBD, MC, ME, ME-H, MI, OH, OUCI, SCT, SGP, UH, WI
	Calcination temperature (°C)	25-900
	Calcination time (h)	0-24
Reaction conditions	Reduction temperature (°C)	25-800
	Reduction pressure (bar)	1-30
	Reduction time (h)	0-24
	Reduction H_2%	0-100
	Temperature (°C)	20-800
	Pressure (bar)	1-40
	W/F (mgcat/minml)	0.15-72
	Time on stream (h)	6-500
	CO% in feed	0-4.9

	Inert% in feed	0-96
	CH_4% in feed	0-56
	H_2O% in feed	0-33
	H_2/CO_2 in feed	0.8-50
Catalytic performance	CO_2 conversion %	0-100

Table 3.2 Input features used in CO_2 to Methanol

Category	Parameters	Range
Catalyst properties	Base	In_2O_3, Cu, Pd, ZnO-ZrO_2
	Base Loading [wt. %]	0-100
	CR Metal [pm]	131-144
	Support 1	Fe_3O_4, ZnO, SBA-15, MCM-15, CeO_2, ZrO_2, TiO_2, AlCeO, SiO_2, CNT, Al_2O_3, In_2O_3, $ZnTiO_3$, Ga_2O_3, SiC
	MW Support 1 [g.mol^{-1}]	0-277.64
	Support 2	SBA-15, Al_2O_3, ZrO_2, SiO_2, AlOOH, CeO_2, ZnO, CNT, TiO_2, AC
	MW Support 2 [g.mol^{-1}]	0-172.12
	Support 3	SBA-15, MCF, Al_2O_3, TiO_2, CeO_2, ZrO_2, P25
	MW Support 3 [g.mol^{-1}]	0-248
	Total MW Support [g.mol^{-1}]	0-501.50
	Promoter 1	Mn, Cu, Pt, Ir, Ga, Y, Ni, Ce, Ru, Rh, In, Zr, Ag, Graphene, Nd, Zn, La, Pd, Al, Si, N_2, W, Ba, Mo, Cr, N doped Graphene, Pr, K, Mg, Fe

	Promoter 1 loading [wt. %]	0-57
	Promoter 2	Carbon, Ga, Al, Ca, Na, Sn, Zn
	Promoter 2 loading [wt. %]	0-7.14
Synthesis conditions	Calcination Temperature [K]	573-973
	Calcination duration [h]	2-5
	SBET [m^2 g^{-1}]	2-1102
Reaction conditions	H_2/CO_2	1-20
	GHSV [cm^3 h^{-1} $gcat^{-1}$]	900-120000
	Catalyst amount [g]	0-5
	Pressure [Mpa]	0.1-36
	Temperature [K]	433-673
Catalytic performance	STY [$gMeOHh^{-1}gcat^{-1}$]	0.000737-2.62

Table 3.3 Input features used in 5-HMF

Category	Parameters	Range
Catalyst properties	%Nb	0-10
	%Al	80-100
	%AA	0-10
	Surface area (cm^2g^{-1})	169-300
	Pore volume (cm^3g^{-1})	0.16-0.34
	Pore diameter (nm)	3.51-4.65
	%Weak	16.2050-33.1641
	%Medium	14.9134-40.3047
	%Strong	32.0896-65.6977
	Weak (mmol/g)	0.117-0.233

	Medium (mmol/g)	0.112-0.322
	Strong (mmol/g)	0.172-0.595
	Total (mmol/g)	0.536-1.012
	Weak D (mmol/g)	0.4021-1.0640
	Medium D (mmol/g)	0.3733-1.1459
	Strong D (mmol/g)	0.6255-2.6746
	Total D (mmol/g)	1.9491-4.0710
	Brønsted acidity (mmol/g)	0.249-0.564
	Brønsted basicity (mmol/g)	0.199-0.626
	BA D (μmol/m^2)	1.0767-2.9822
	BB D (μmol/m^2)	0.9803-2.5447
	BA/TA	0.3272-0.8910
Catalytic performance	5-HMF yield (%)	20.68-37.91
	Glucose conversion (%)	59.11-93.49
	Selectivity (%)	23.87-62.46

3.2 Data Preparation

3.2.1. One-Hot Encoding

Table 3.4 shows total categorical features, Categorical features were converted to numbers [0, 1] using the command pandas.get_dummies() before analysis because machine learning works well on numbers.

Table 3.4 Total categorical features

CO_2 methanation	CO_2 to Methanol
Base, Base 2, Support, Support 2, Catalyst preparation method	Base, Support 1, Support 2, Support 3, Promoter 1, Promoter 2

3.2.2. Missing Values

Missing values were filled with average values, as defined in Eq. (1).

$$Mean\ (\bar{x}) = \frac{\sum x}{n} \quad (1)$$

3.2.3. Data Partition

4,051 points of the dataset CO_2 methanation, 1,234 points of the dataset CO_2 to methanol and 13 points of the dataset 5-HMF were divided into 80:20, 85:15, and 100:0 as training and testing datasets, respectively. A data set into subsets is important when learning a model. The original dataset can be divided into three sub-sets: the training set, the validation set, and the testing set. The training set directly participates in the training process, which is used to learn the model and modify its parameters. Validation sets are used to adjust model hyperparameters and protect the model from overfitting. A test set is used to evaluate the performance and robustness of the model [55].

3.2.4. Data Augmentation

Data augmentation was used to increase the diversity of a dataset by applying various transformations or modifications to the existing data. The goal is to enhance the performance and robustness of machine learning models.

In this study, data augmentation generated training sets of the original dataset increasing to 2 times, 5 times and 10 times. The data was generated by adding noise that is randomly drawn from a normal distribution with mean equal to 0 and standard deviation equal to 1 as defined in Eq. (2).

$$output[i] = input[i] + noise \times beta \quad (2)$$

Where $output[i]$ represent the value obtained after adding noise in i loop, $input[i]$ represent the original data value at i in loop, $noise$ represent random values are obtained from the normal distribution, and $beta$ represent the size adjuster noise.

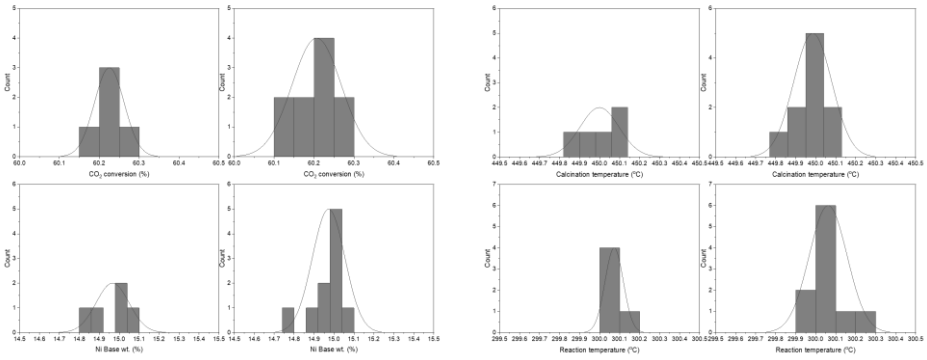

Figure 3.1 Example of distribution for 5x and 10x data points.

3.2.5 Principal Component Analysis

Principal Component Analysis (PCA) is a linear transformation technique used to project a set of correlated variables into a new coordinate system such that the greatest variance of the data lies on the first coordinate (called the first principal component), the second greatest variance on the second component, and so on.

Mathematically, given a dataset

$$X \in \mathbb{R}^{n \times p} \tag{3}$$

where n is the number of samples and p is the number of features

PCA seeks to find a projection matrix W that maximizes the variance of the projected data Z:

$$Z = XW \tag{4}$$

The optimization objective is to find W such that:

$$\max_{W} \ \text{Var}(Z) = \max_{W} \ W^T S W \tag{5}$$

where $S = \frac{1}{n-1} X^T X$ is the sample covariance matrix.

This leads to the eigenvalue decomposition of the covariance matrix:

$$S v_i = \lambda_i v_i \tag{6}$$

Where v_i is the i^{th} eigenvector (the principal component direction) and λ_i is the corresponding eigenvalue representing the explained variance

The explained variance ratio for each principal component is given by:

$$\text{Explained Variance Ratio}_i = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j} \quad (7)$$

In this study, the cumulative explained variance between 80% and 99% was analyzed to determine the optimal number of principal components that balance dimensionality reduction and information retention. The selected principal components corresponding to the best model performance were then used in subsequent machine learning model training.

3.2.6. Data Normalization

Data normalization was performed to standardize the scale of features in datasets, ensuring fair comparison, faster convergence in machine learning models, equal importance of features, improved interpretability, and mitigation of the impact of outliers. The standardized Z-score was adopted, as defined in Eq. (8).

$$Z = \frac{X - \mu}{\sigma} \quad (8)$$

Where X represents the individual data point, and μ and σ represent the mean and standard deviation of the dataset, respectively.

3.3 Model Evaluation

Quantifying the quality of the model can be calculated using root mean square error (RMSE), and the reliability of the model are obtained from R^2 score as defined in Eq. (9), (10), respectively.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2} \quad (9)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (10)$$

3.4 Model for learning

In this study, we utilized two different machine learning models—Multilayer Perceptron (MLP) and Stacking ensemble algorithm—to model three key reactions: CO_2 methanation, CO_2-to-methanol, and 5-HMF synthesis. The dataset used originated from the Origin dataset, with the training data augmented by factors of 2, 5, and 10, respectively.

The hyperparameter values for each algorithm, which play a crucial role in effective model learning, are summarized in **Table 3.5**. It is important to note that each model exhibits distinct learning characteristics, necessitating tailored tuning of the hyperparameters. Consequently, the tools and optimization procedures used for model training differ across the algorithms.

As shown in **Figure 3.2**, **Figure 3.3** and **Figure 3.4** illustrate the complete workflow, covering the entire process from the initial data preparation stages to the evaluation of the model's performance and reliability.

Table 3.5 All hyperparameter of machine learning

Model	Hyperparameter		Optimize
Multilayer Perceptron	alpha		0.00001-1
	activation		"relu"
	solver		"adam"
	random state		42
	Hidden layer sizes	layer	1-10
		node	10-300
Stacking Ensemble	CV		5-20
Linear Regression	-		default
Ridge	-		default
Lasso	-		default
ElasticNet	-		default
RandomForest	n_estimators		100-1000

	max_depth	None-20
GradientBoosting	n_estimators	100-1000
	max_depth	None-20

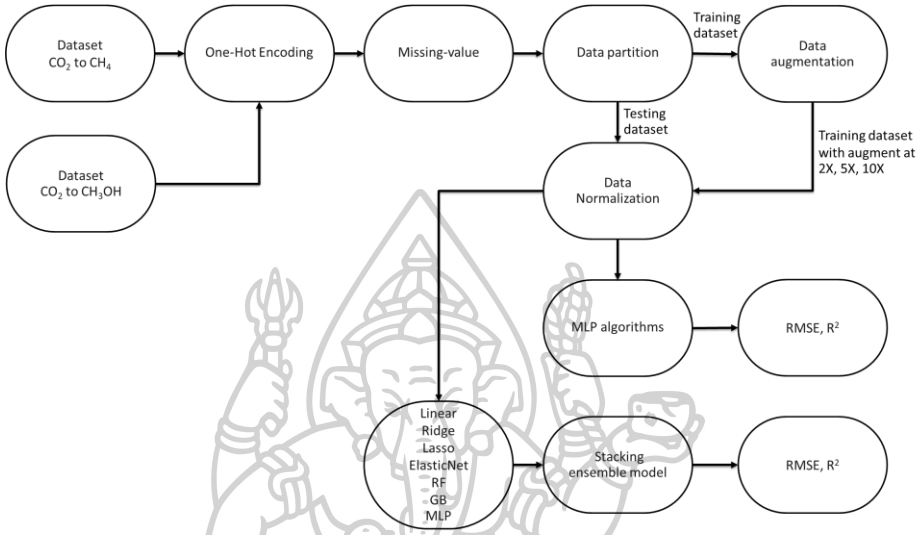

Figure 3.2 Workflow of dataset CO_2 methanation and CO_2 to methanol for data augmentation

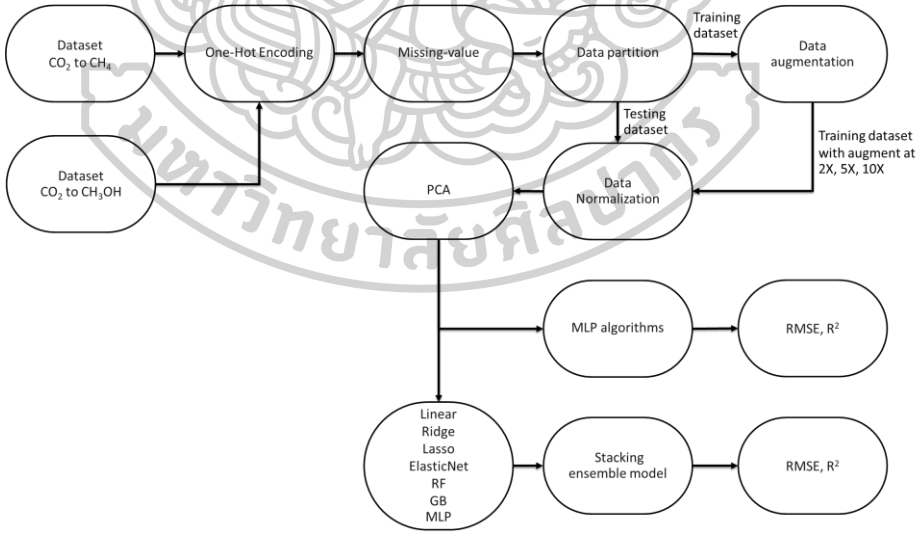

Figure 3.3 Workflow of dataset CO_2 methanation and CO_2 to methanol for principal component analysis

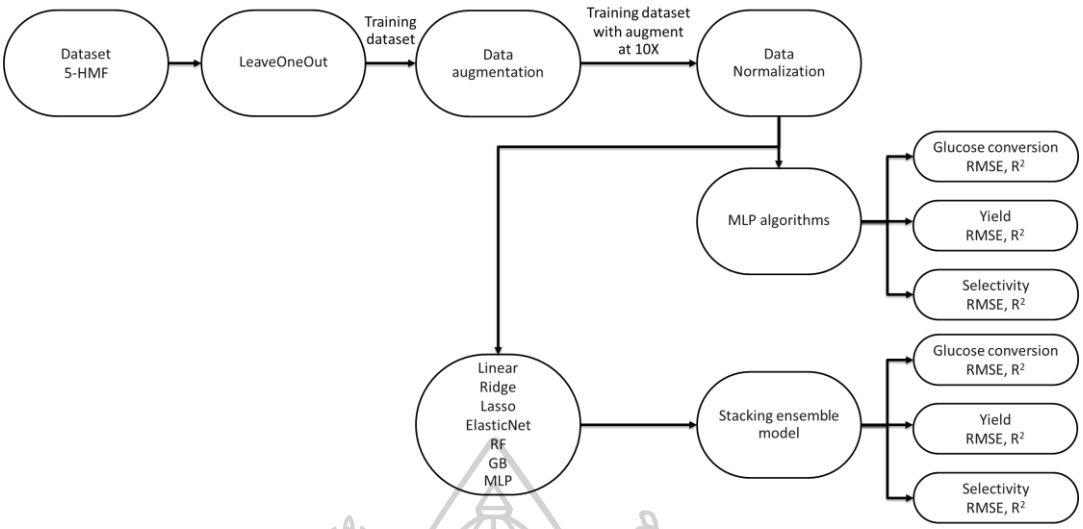

Figure 3.4 Workflow of dataset 5-HMF

CHAPTER IV

RESULTS AND DISCUSSION

4.1 Dataset character analytics

4.1.1 Dataset of carbon dioxide methanation

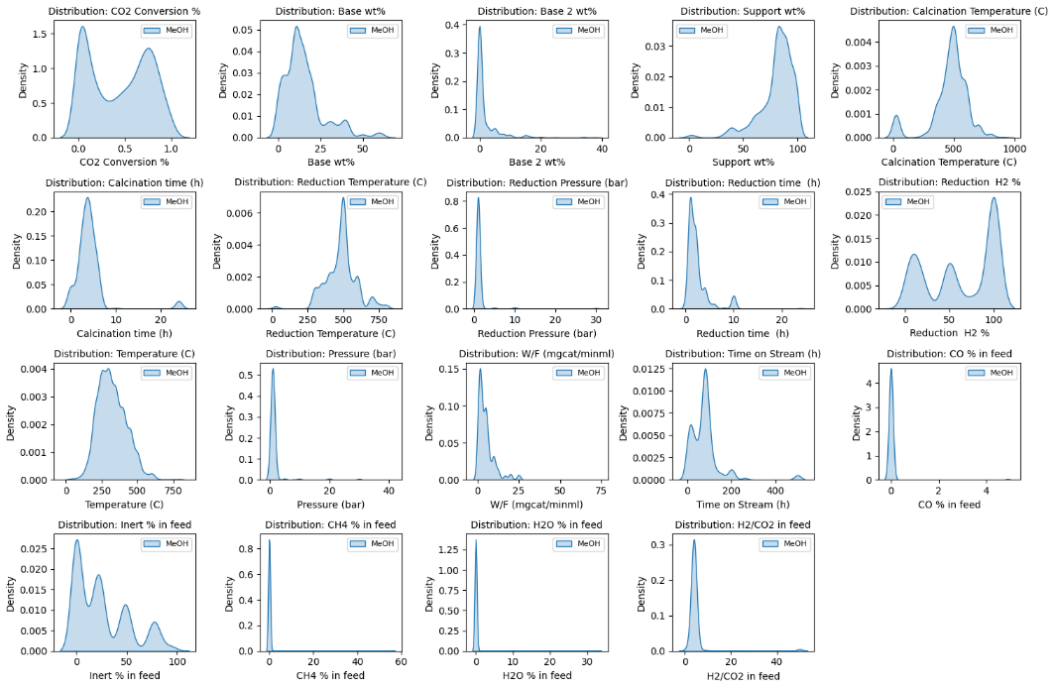

Figure 4.1 Plot seaborn of carbon dioxide methanation

In **Figure 4.1**, based on the data distribution plots generated using Seaborn for the CO_2 methanation dataset, it was observed that several features exhibited left-skewed distributions, including Base wt%, Base 2 wt%, Calcination time (h), Reduction pressure (bar), Reduction time (h), Pressure (bar), W/F (mgcat/min·mL), Time on stream (h), CO% in feed, Inert% in feed, CH_4% in feed, H_2O% in feed, and H_2/CO_2 in feed. Features showing approximately symmetric distributions were CO_2 conversion (%), Calcination temperature (°C), Reduction temperature (°C), and Temperature (°C). In contrast, support wt% and Reduction H_2% displayed right-skewed distributions.

4.1.2 Dataset of carbon dioxide to methanol

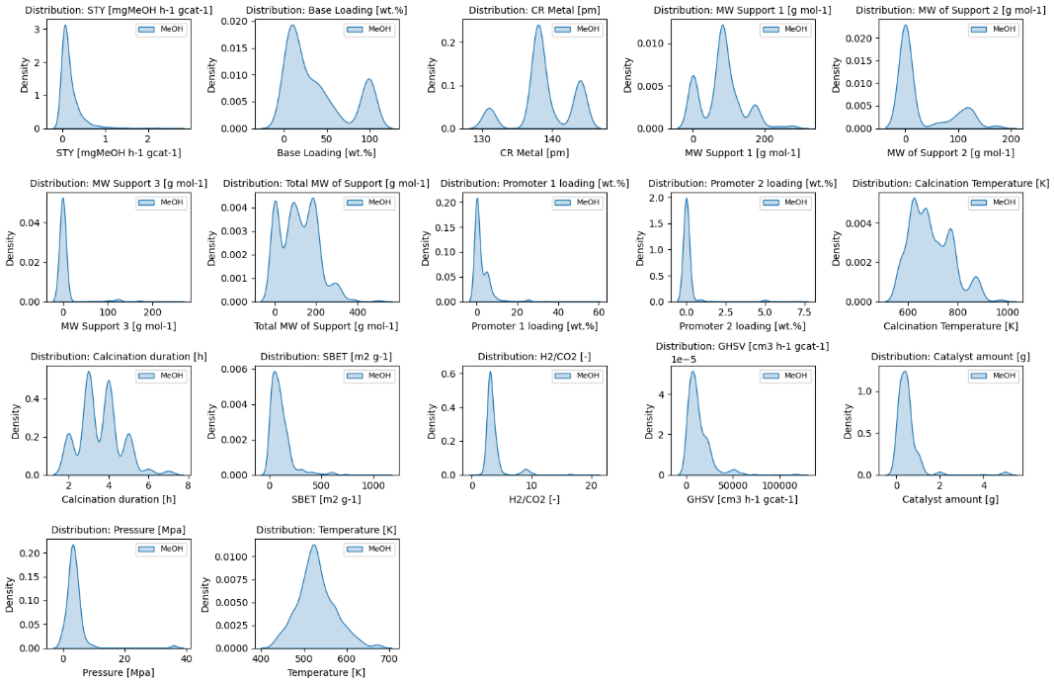

Figure 4.2 Plot seaborn of carbon dioxide to methanol

In **Figure 4.2**, based on Seaborn plots of the CO_2-to-methanol dataset, it was observed that several features exhibited left-skewed distributions, including STY (mg MeOH·h^{-1}·$gcat^{-1}$), Base Loading [wt.%], MW Support 1 (g·mol^{-1}), MW Support 2 (g·mol^{-1}), MW Support 3 (g·mol^{-1}), Total MW of Support (g·mol^{-1}), Promoter 1 Loading (wt.%), Promoter 2 Loading (wt.%), Calcination Temperature (K), SBET (m^2·g-1), H_2/CO_2, GHSV (cm^3· h^{-1}·$gcat^{-1}$), and Pressure (MPa). Features with approximately symmetric distributions included CR Metal (pm), Calcination Duration (h), and Temperature (K). Notably, no features exhibited right-skewed distributions.

4.1.3 Dataset of 5-hydroxymethylfurfural

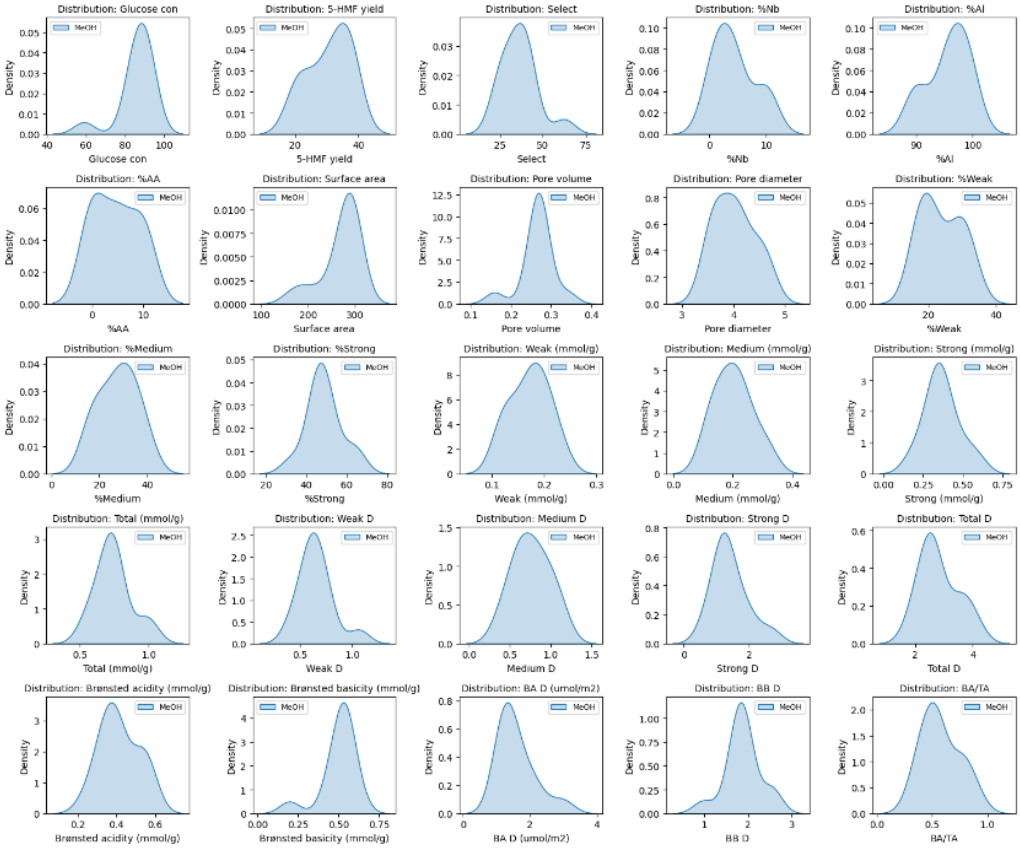

Figure 4.3 Plot seaborn of 5-hydroxymethylfurfural

In **Figure 4.3**, based on Seaborn plots of the 5-hydroxymethylfurfural dataset, it was observed that several features exhibited left-skewed distributions, including Selectivity, Weak D, Strong D, Total D, BA D (μmol/m²), and BA/TA. Features with approximately symmetric distributions included 5-HMF Yield, %Nb, %Al, %AA, Pore Diameter, %Weak, %Medium, %Strong, Weak (mmol/g), Medium (mmol/g), Strong (mmol/g), Total (mmol/g), Medium D, Bronsted Acidity (mmol/g), and BB D. In contrast, features with right-skewed distributions included Glucose Conversion, Surface Area, and Bronsted Basicity (mmol/g).

4.2 Comparison Between MLP and SEM Using Original Data

4.2.1 CO_2 methanation

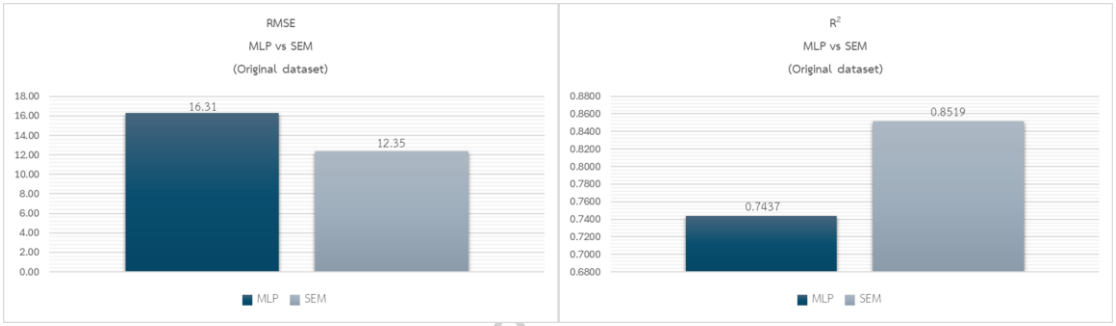

Figure 4.4 Compare MLP and SEM with Original Data

RMSE (Left), R^2 (Right)

Based on the experimental results, when the original dataset was trained using the MLP model and the SEM model—which integrates several sub-models including Linear Regression, Ridge, Lasso, ElasticNet, Random Forest, Gradient Boosting, and MLP—and evaluated based on the RMSE and R^2 metrics, it was found as shown **Figure 4.4** that the MLP model achieved an RMSE of 16.31 (R^2 = 0.7437). In comparison, the SEM model demonstrated an improvement in performance, yielding a lower RMSE of 12.35 (R^2 = 0.8519), indicating that the SEM model provided a better fit to the data.

4.2.2 CO_2 to methanal

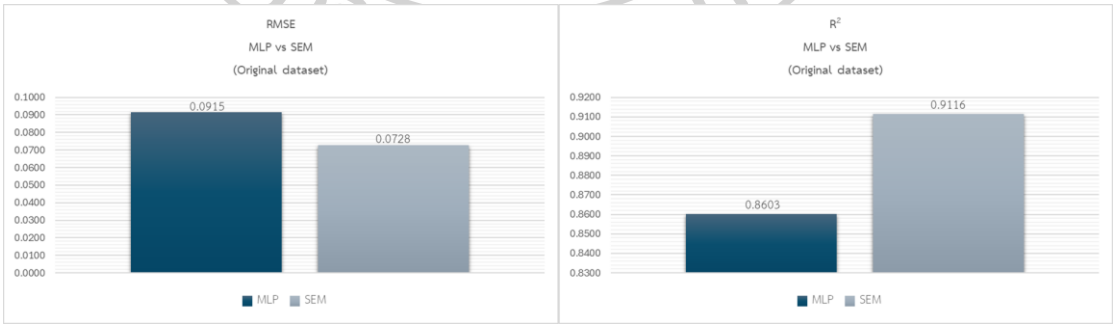

Figure 4.5 Compare MLP and SEM with Original Data

RMSE (Left), R^2 (Right)

Based on the experimental results, when the same data was used to train both an MLP model and a SEM (Stacking Ensemble Model), it was found that the SEM

performed better. The SEM internally consists of several models: Linear Regression, Ridge, Lasso, ElasticNet, RandomForest, GradientBoosting, and MLP. The models were evaluated using RMSE (Root Mean Square Error) and R-squared (R^2). From **Figure 4.5**, The MLP model achieved an RMSE of 0.0915 (with an R^2 of 0.8603). In contrast, the SEM showed a trend of decreasing RMSE, achieving a lower RMSE of 0.0728 (with a higher R^2 of 0.9116)

4.3 Comparison Between MLP and SEM Using Augmented Data

4.3.1 CO_2 methanation

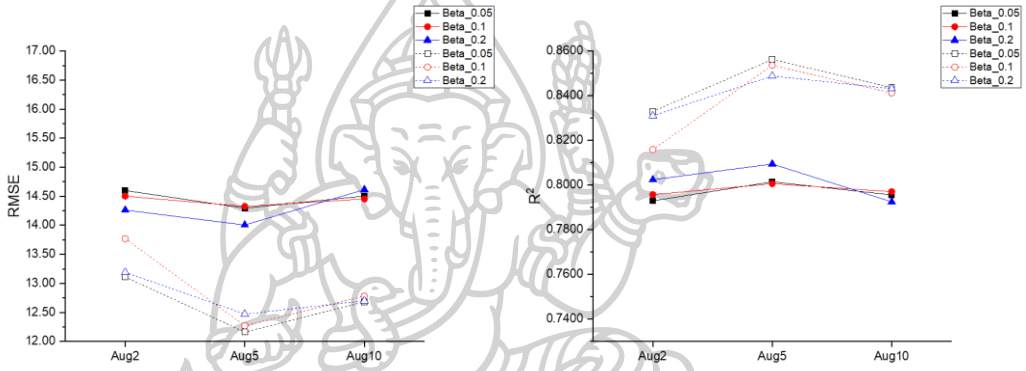

Figure 4.6 Compare MLP and SEM with Augmented Data

RMSE (Left), R^2 (Right)

Based on the experimental results, the original dataset was augmented by factors of 2, 5, and 10, and subsequently used to train two models: the Multilayer Perceptron (MLP) and the Stacked Ensemble Model (SEM). Model performance was evaluated using the Root Mean Square Error (RMSE). The results indicated that the SEM consistently achieved lower RMSE values compared to the MLP model. As shown in **Figure 4.6**, The lowest RMSE obtained from the SEM was 12.16 with an R^2 value of 0.8563, whereas the minimum RMSE of the MLP model was 14.00 with an R^2 value of 0.8095. Moreover, analysis of the β (beta) parameter revealed that, for the MLP model, RMSE tended to decrease when β = 0.2. In contrast, for the SEM model, RMSE showed a decreasing trend when β = 0.05. These findings suggest that the SEM model

demonstrates superior predictive performance and greater sensitivity to the influence of data augmentation compared to the MLP model.

4.3.2 CO_2 to methanol

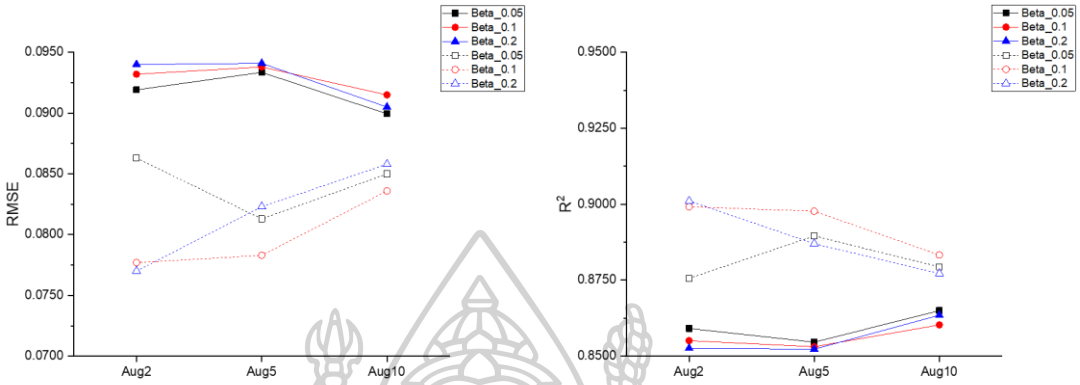

Figure 4.7 Compare MLP and SEM with Augmented Data

RMSE (Left), R^2 (Right)

Based on the experimental results, the original dataset was augmented by factors of 2, 5, and 10, and subsequently used to train two predictive models: the Multilayer Perceptron (MLP) and the Stacked Ensemble Model (SEM). The performance of both models was evaluated using the Root Mean Square Error (RMSE). The results revealed that the SEM achieved lower RMSE values compared to the MLP model. Specifically, as shown in **Figure 4.7**, the lowest RMSE obtained from the SEM was 0.0770 with an R^2 value of 0.9011, while the minimum RMSE from the MLP model was 0.0899 with an R^2 value of 0.8650. Furthermore, analysis of the β (beta) parameter indicated that, for the MLP model, the RMSE tended to decrease when β = 0.05. In contrast, for the SEM model, the RMSE showed a decreasing trend when β = 0.1. These findings suggest that the SEM exhibits superior predictive accuracy and greater robustness to data augmentation compared to the MLP model.

4.4 Comparison Between Original and Augmented Data Using MLP

4.4.1 CO_2 methanation

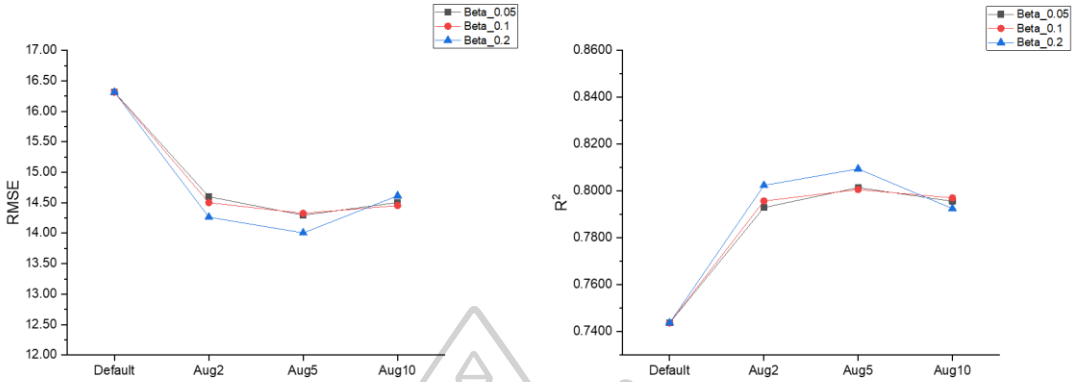

Figure 4.8 Compare Original and Augmented with MLP

RMSE (Left), R^2 (Right)

Based on the experimental results, it was observed that training the MLP model with the original dataset resulted in a root mean square error (RMSE) of 16.31 and a coefficient of determination (R^2) of 0.7437. When the dataset was augmented by factors of 2×, 5×, and 10×, and subsequently used to train the MLP model, the evaluation results indicated a decreasing trend in RMSE as the data volume increased. From **Figure 4.8**, The minimum RMSE value of 14.00 (R^2 = 0.8095) was obtained with 5× data augmentation. However, further increasing the data to 10× resulted in a slight increase in RMSE. Moreover, analysis of the beta (β) parameter demonstrated that β = 0.2 was the most effective value for reducing RMSE.

4.4.2 CO_2 to methanal

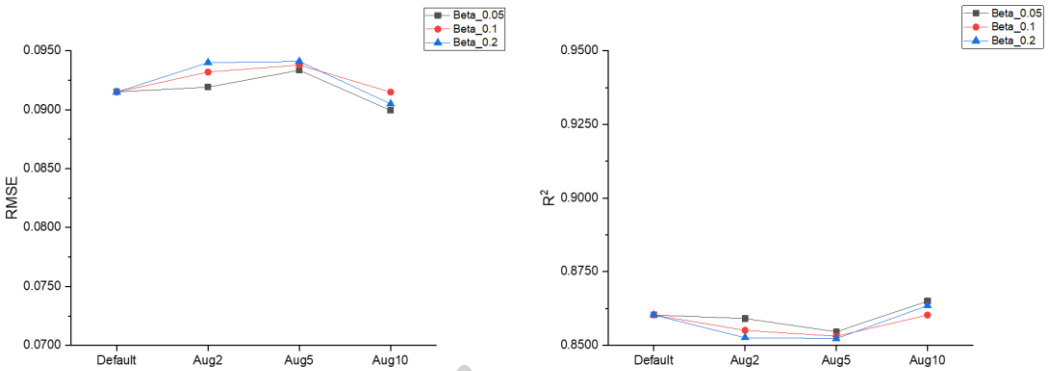

Figure 4.9 Compare Original and Augmented with MLP

RMSE (Left), R^2 (Right)

Based on the experimental results, it was found that training the MLP model using the original dataset yielded a root mean square error (RMSE) of 0.0915 and a coefficient of determination (R^2) of 0.8603. When the dataset was augmented by factors of 2×, 5×, and 10×, and subsequently used to train the MLP model, the evaluation results showed in **Figure 4.9** that the RMSE tended to increase when the data were expanded by 2× and 5×, while a slight decrease was observed at 10× augmentation. The lowest RMSE of 0.0899 (R^2 = 0.8650) was achieved with 10× data augmentation. Furthermore, analysis of the beta (β) parameter indicated that β = 0.05 was the most effective value in reducing the RMSE.

4.5 Comparison Between Original and Augmented Data Using SEM

4.5.1 CO_2 methanation

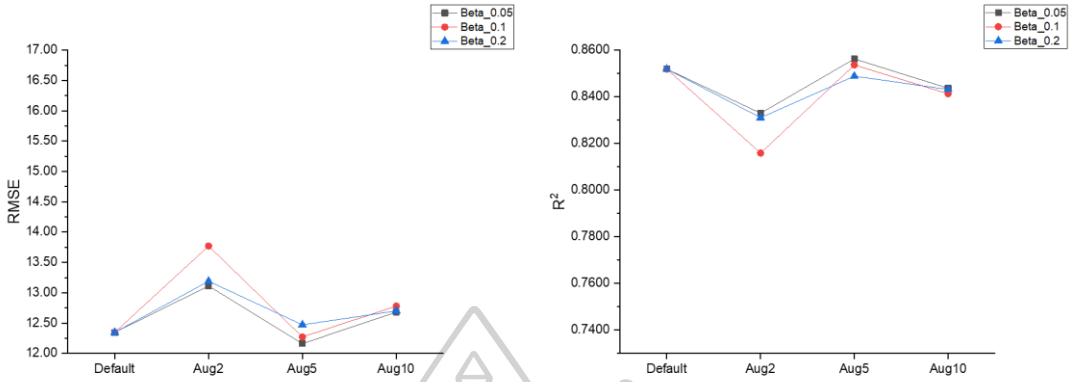

Figure 4.10 Compare Original and Augmented with SEM

RMSE (Left), R^2 (Right)

Based on the experimental results, it was observed that when the original dataset was used to train the SEM model—which integrates multiple sub-models including Linear Regression, Ridge, Lasso, ElasticNet, Random Forest, Gradient Boosting, and MLP—the model evaluation yielded a root mean square error (RMSE) of 12.35 and a coefficient of determination (R^2) of 0.8519. When the dataset was augmented by factors of 2×, 5×, and 10×, and subsequently trained using the SEM model, the evaluation results (**Figure 4.10**) indicated that the RMSE tended to increase with 2× and 10× data augmentation, while a decrease was observed at 5× augmentation. The lowest RMSE value of 12.16 (R^2 = 0.8563) was obtained when the dataset was augmented by 5×. Furthermore, analysis of the beta (β) parameter revealed that β = 0.05 tended to achieve the greatest reduction in RMSE.

4.5.2 CO_2 to methanal

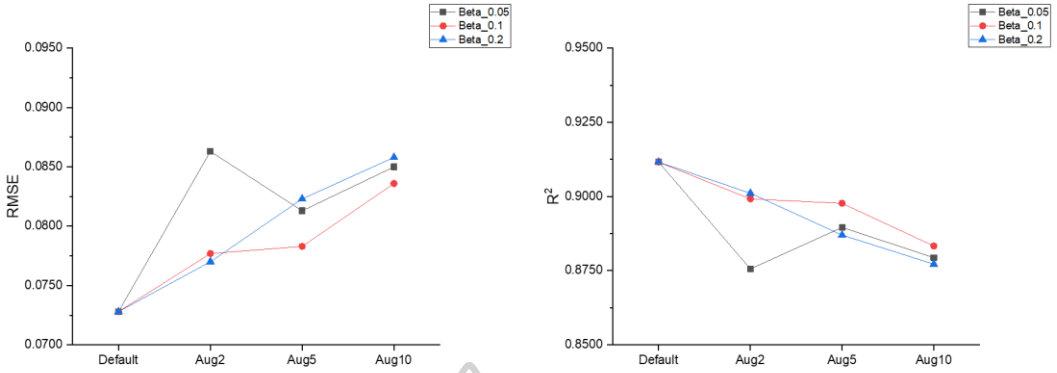

Figure 4.11 Compare Original and Augmented with SEM

RMSE (Left), R^2 (Right)

Based on the experimental results, it was observed that when the original dataset was used to train the SEM model—which integrates multiple sub-models including Linear Regression, Ridge, Lasso, ElasticNet, Random Forest, Gradient Boosting, and MLP—the model evaluation yielded a root mean square error (RMSE) of 0.0728 and a coefficient of determination (R^2) of 0.9116. When the dataset was augmented by factors of 2×, 5×, and 10×, and subsequently trained using the SEM model, the evaluation results (**Figure 4.11**) indicated that the RMSE tended to increase as the amount of augmented data increased. Furthermore, analysis of the beta (β) parameter revealed that β = 0.1 tended to yield the greatest reduction in RMSE.

4.6 Comparison of MLP and SEM Using Original and Augmented Data for 5-HMF

4.6.1 Glucose Conversion

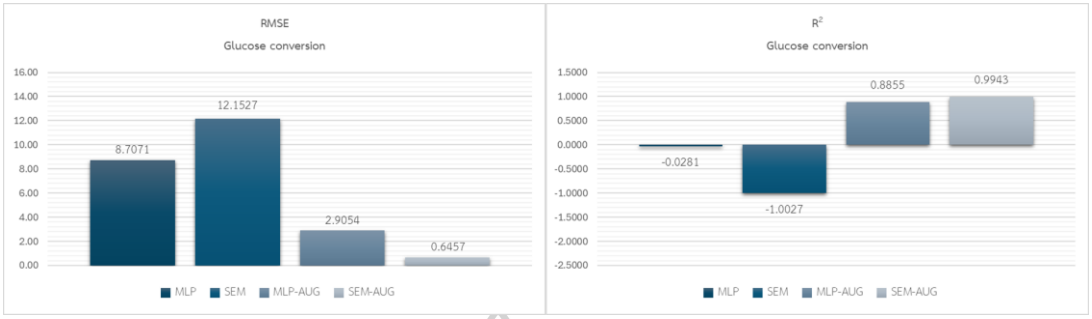

Figure 4.12 Compare Original vs Augmented Data with MLP and SEM

RMSE (Left), R^2 (Right)

In this experiment, a dataset of 13 samples of 5-HMF was augmented by a factor of 10×, and both the original and the 10× augmented datasets were used to train two models: MLP and SEM. The SEM model integrated several sub-models, including Linear Regression, Ridge, Lasso, ElasticNet, Random Forest, Gradient Boosting, and MLP, with Glucose conversion as the output variable. From **Figure 4.12** shown the experimental results indicated that the lowest RMSE was 0.6457 (R^2 = 0.9943), achieved by the SEM model trained with the 10× augmented dataset. The second-best performance was obtained by the MLP model trained with the same 10× augmented dataset, yielding an RMSE of 2.9054 (R^2 = 0.8855).

4.6.2 Selectivity

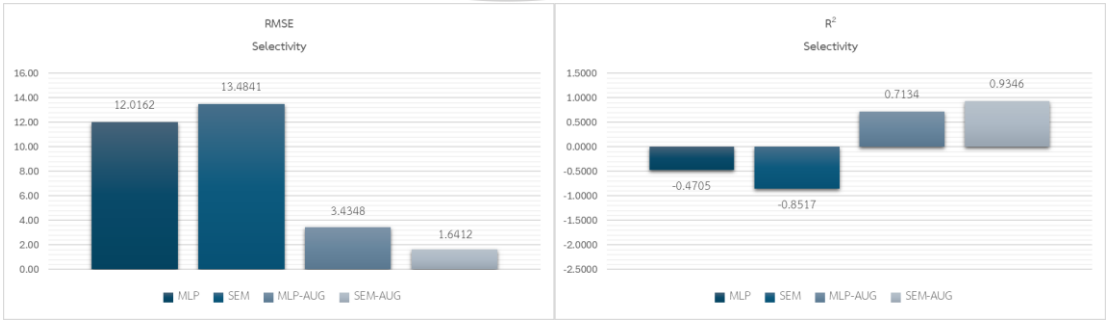

Figure 4.13 Compare Original vs Augmented Data with MLP and SEM

RMSE (Left), R^2 (Right)

In this experiment, a dataset consisting of 13 samples of 5-HMF was augmented by a factor of 10×, and both the original dataset and the 10× augmented dataset were used to train two models: MLP and SEM. The SEM model integrated several sub-models, including Linear Regression, Ridge, Lasso, ElasticNet, Random Forest, Gradient Boosting, and MLP, with Selectivity as the output variable. From **Figure 4.13** shown the experimental results revealed that the lowest RMSE was 1.6412 (R^2 = 0.9346), achieved by the SEM model trained with the 10× augmented dataset. The second-best performance was obtained from the MLP model trained with the same 10× augmented dataset, yielding an RMSE of 3.4348 (R^2 = 0.7134).

4.6.3 Yield

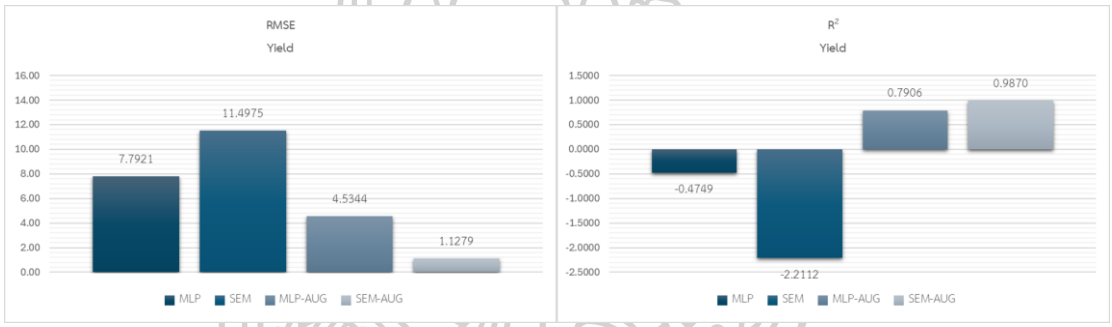

Figure 4.14 Compare Original vs Augmented Data with MLP and SEM RMSE (Left), R^2 (Right)

In this experiment, the original dataset of 5-HMF, consisting of 13 samples, was expanded by a factor of ten through a data augmentation process. Both the original and the augmented datasets were subsequently used to train two models: a Multilayer Perceptron (MLP) and a Stacked Ensemble Model (SEM). The SEM comprised multiple base learners, including Linear Regression, Ridge, Lasso, ElasticNet, Random Forest, Gradient Boosting, and MLP, with the target variable defined as the Yield. From **Figure 4.14** shown the experimental results revealed that the lowest root mean square error (RMSE) was obtained from the SEM trained with the tenfold augmented data, yielding an RMSE of 1.1279 and an R^2 of 0.9870. The second-best performance was achieved by the MLP trained on the augmented dataset, with an RMSE of 4.5344 and an R^2 of 0.7906. These findings indicate that data augmentation significantly enhances the

predictive performance of the SEM compared to models trained solely on the original dataset.

4.7 Comparison Between PCA and Non-PCA Using MLP

4.7.1 CO_2 methanation

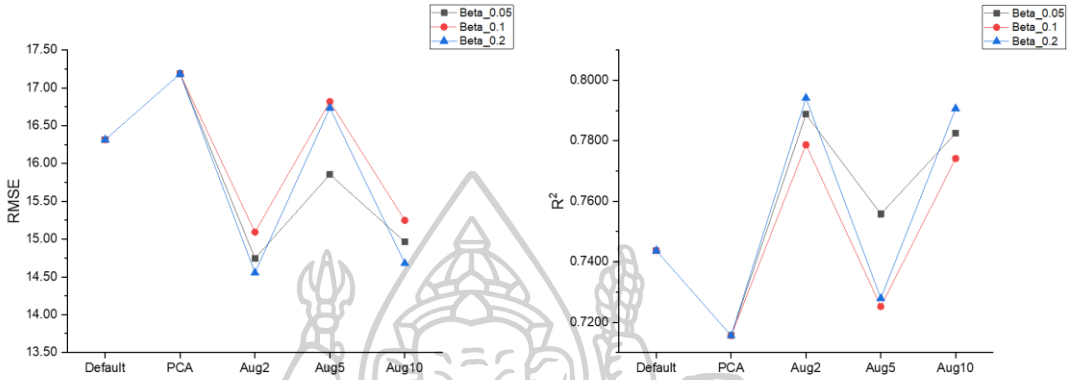

Figure 4.15 Compare PCA and Non-PCA with MLP

RMSE (Left), R^2 (Right)

Based on the experimental results, it was observed that when the original dataset—after undergoing dimensionality reduction using Principal Component Analysis (PCA)—was trained with the MLP model, the evaluation yielded a root mean square error (RMSE) of 17.19 and a coefficient of determination (R^2) of 0.7156. When the dataset was augmented by factors of 2×, 5×, and 10×, followed by PCA processing and training with the MLP model, the evaluation results (**Figure 4.15**) indicated that the RMSE tended to decrease with 2× data augmentation and further decrease with 10× augmentation. The lowest RMSE value of 14.56 (R^2 = 0.7942) was obtained with 2× data augmentation. Furthermore, analysis of the beta (β) parameter revealed that β = 0.2 tended to produce the greatest reduction in RMSE.

4.7.2 CO_2 to methanol

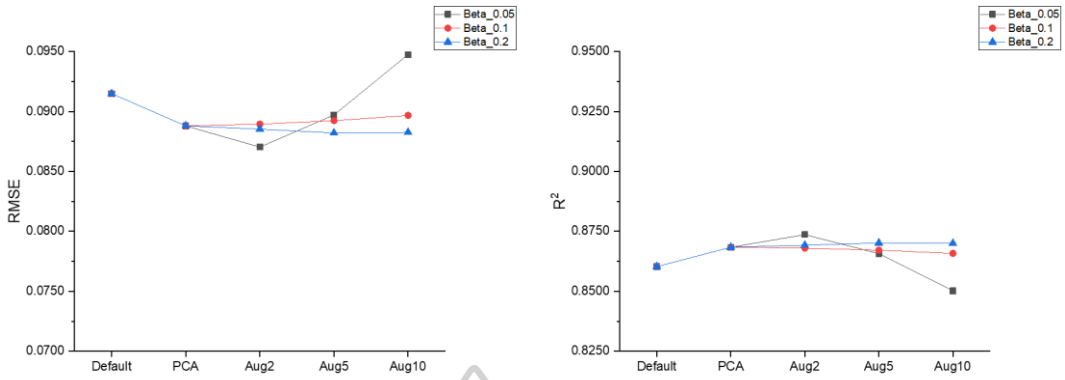

Figure 4.16 Compare PCA and Non-PCA with MLP

RMSE (Left), R^2 (Right)

Based on the experimental results, it was observed that when the original dataset, after undergoing dimensionality reduction using Principal Component Analysis (PCA), was trained with the MLP model, the evaluation yielded a root mean square error (RMSE) of 0.0888 and a coefficient of determination (R^2) of 0.8684. When the dataset was augmented by factors of 2×, 5×, and 10×, followed by PCA processing and training with the MLP model, the evaluation results (**Figure 4.16**) indicated that the RMSE tended to decrease when the data were augmented by 2×. The lowest RMSE value of 0.0870 (R^2 = 0.8736) was achieved with 2× data augmentation. Furthermore, analysis of the beta (β) parameter revealed that β = 0.05 tended to produce the greatest reduction in RMSE.

4.8 Comparison Between PCA and Non-PCA Using SEM

4.8.1 CO_2 methanation

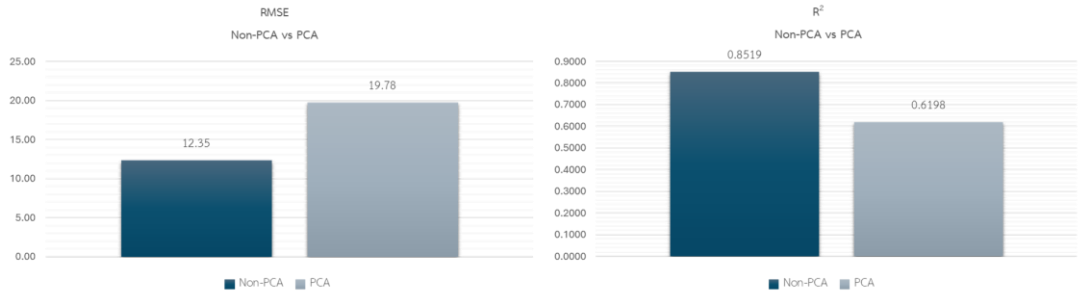

Figure 4.17 Compare PCA and Non-PCA with SEM

RMSE (Left), R^2 (Right)

Based on the experimental results, it was observed that when comparing the original dataset with the dataset that underwent dimensionality reduction using Principal Component Analysis (PCA), both were trained using the SEM model. As shown in **Figure 4.17**. The model evaluation indicated that the SEM trained without PCA achieved a root mean square error (RMSE) of 12.35 and a coefficient of determination (R^2) of 0.8519. In contrast, when PCA was applied, the RMSE increased to 19.78, with an R^2 value of 0.6198, indicating that the use of PCA tended to degrade the model's predictive performance.

4.8.2 CO_2 to methanol

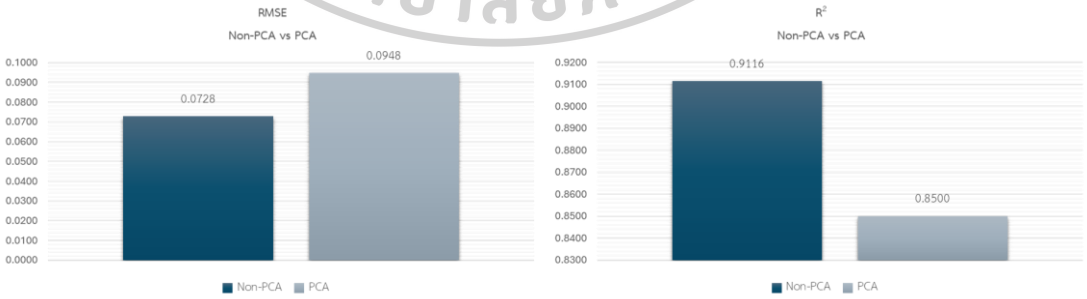

Figure 4.18 Compare PCA and Non-PCA with SEM

RMSE (Left), R^2 (Right)

Based on the experimental results, it was observed that when comparing the original dataset with the dataset that underwent dimensionality reduction using Principal Component Analysis (PCA), both were trained using the SEM model. As shown in **Figure 4.18**, The evaluation results showed that the SEM model trained without PCA achieved a root mean square error (RMSE) of 0.0728 and a coefficient of determination (R^2) of 0.9116. In contrast, when PCA was applied, the RMSE increased to 0.0948, with an R^2 value of 0.8500, indicating that the use of PCA tended to increase the prediction error and reduce the model's overall performance.

4.9 Analysis and Discussion

4.9.1 The Superiority of Stacking Ensemble Models (SEM)

A primary and consistent finding across all datasets was the superior performance of the Stacking Ensemble Model (SEM) over the Multilayer Perceptron (MLP) when trained on the original (non-augmented) data. For CO_2 methanation, the SEM achieved a significantly lower RMSE (12.35) compared to the MLP (16.31). This trend was even more pronounced in the CO_2 to methanol dataset, where the SEM's RMSE (0.0728) was substantially lower than the MLP's (0.0915), corresponding to a much higher R^2 value (0.9116 vs. 0.8603). This superiority can be attributed to the fundamental architecture of stacking, first introduced by Wolpert [37]. As described in the literature review, the SEM combines the predictive power of diverse 'Base Learners' (Level 0) and uses a 'Meta-Learner' (Level 1) to intelligently combine their predictions [39-41]. This ensemble approach is designed to enhance accuracy and robustness compared to any single model [44]. While a single MLP must find one complex, non-linear function to map all inputs to the output, the SEM leverages this "team of experts" 7 (Linear Regression, Ridge, Lasso, ElasticNet, Random Forest, Gradient Boosting, and MLP) which is inherently more robust and less prone to the specific biases of any single model.

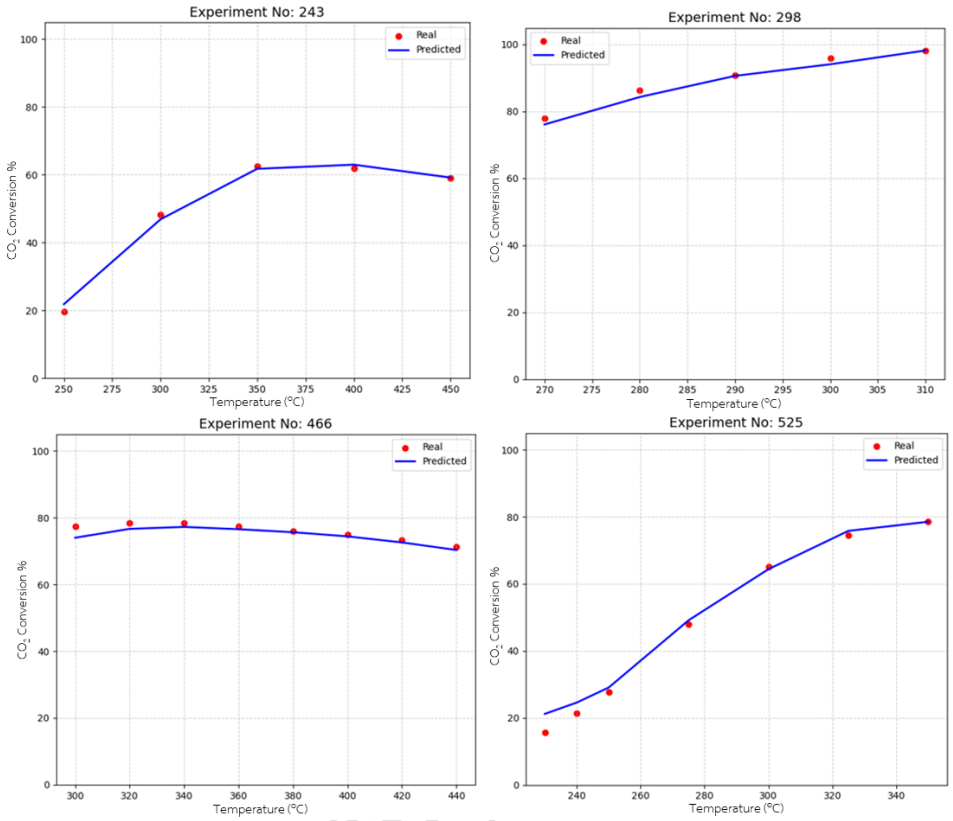

Figure 4.19 Comparison of Real vs. Predicted CO_2 Conversion as a Function of Temperature

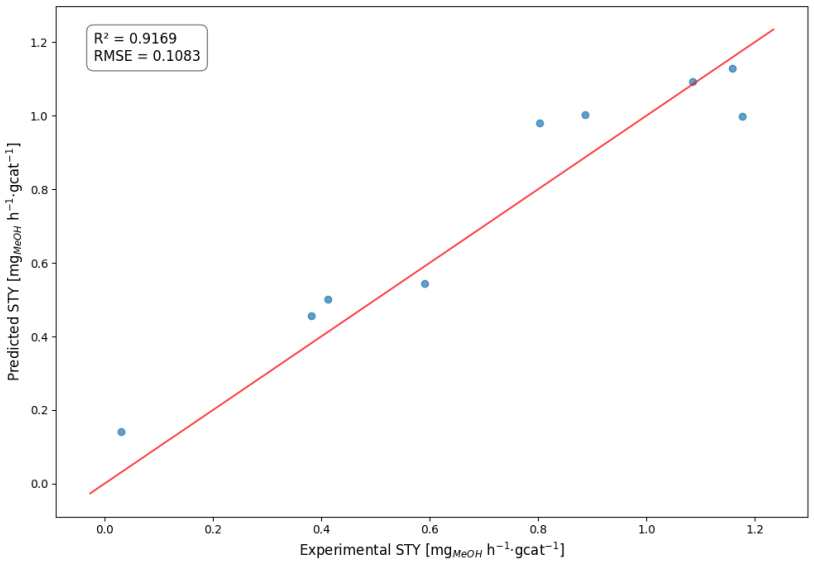

Figure 4.20 Parity Plot of Predicted STY vs. Experimental STY

In this research, the trained SEM models for both the CO_2 methanation and CO_2 to methanol datasets were utilized to predict unseen data. This was done to determine if the models could be used for preliminary trend analysis—for instance, if we wanted to study the effect of temperature on the CO_2 methanation reaction, or if we wanted to research the CO_2 to methanol reaction under specific catalyst and operating conditions beforehand. The findings show that Machine Learning can indeed be applied for preliminary analysis to help decide whether a full experiment should be conducted. **Figure 4.19** presents a comparison between the real experimental values and the predicted values for the CO_2 methanation reaction. It was found that the model can be used for preliminary study, such as predicting how changing the temperature might affect the output, or CO_2 conversion %. Data from experiments 243, 298, 466, and 525, which represented unseen data, were used. The results show that the predicted values are highly accurate when compared to the actual experimental data. Similarly, **Figure 4.20** shows a parity plot comparing the experimental and predicted STY [$mg_{MeOH}\ h^{-1}gcat^{-1}$] for the CO_2 to methanol reaction. Nine unseen data points from experiments were used for prediction, and the results confirm that the SEM model can accurately predict data it has not seen before. The model's performance evaluation yielded an RMSE of 0.1083 and an R^2 of 0.9169, which is considered highly accurate for unseen data.

4.9.2 The Critical and Nuanced Role of Data Augmentation

The most dramatic results of this study stem from the application of data augmentation, particularly on the 5-HMF dataset. The original 5-HMF dataset, with only 13 data points, was insufficient for any meaningful modeling. This was evident from the negative R^2 values for both MLP and SEM models trained on the original data, indicating that the models performed worse than simply predicting the mean. Upon applying 10x data augmentation, the performance improved drastically. The SEM trained on augmented data (SEM-AUG) became highly predictive, achieving R^2 values of 0.9943 (Glucose Conversion), 0.9346 (Selectivity), and 0.9870 (Yield). This directly confirms the hypothesis that data augmentation is a crucial strategy for overcoming the challenge of small experimental datasets, as it plays a key role in preventing

overfitting and creating more robust models [20]. Furthermore, the dataset character analytics from **section 4.1** provide a deeper insight into why this augmentation was so effective for 5-HMF but had mixed results on larger datasets. The augmentation method used involved adding noise from a normal distribution. The 5-HMF dataset, as shown in **Figure 4.3**, exhibits many symmetric or near-symmetric distributions (e.g., 5-HMF Yield, %Al, %Nb, Pore Diameter). Adding noise from a normal distribution (which is also symmetric) is a natural way to expand this dataset, filling gaps without distorting the underlying data structure. This explains the transformative performance leap. In contrast, the larger datasets (CO_2 methanation and CO_2 to methanol) are dominated by heavily left-skewed distributions for many key features (e.g., STY, GHSV, Pressure, H_2/CO_2). This skewness explains the nuanced results: The SEM, being a powerful model, had already learned the true, skewed distribution of the 1,234-point CO_2 to methanol dataset. When symmetric, noise-based data was added, it may have "polluted" the dataset with non-physical data points, conflicting with the real-world skewed pattern and thus increasing the RMSE. The MLP, being a single model, still benefited simply from more data to learn from, even if it was slightly distorted. This highlights that data augmentation is most effective when the augmentation method (e.g., normal distribution noise) is appropriate for the data's underlying distribution (as in 5-HMF) or when data is extremely scarce.

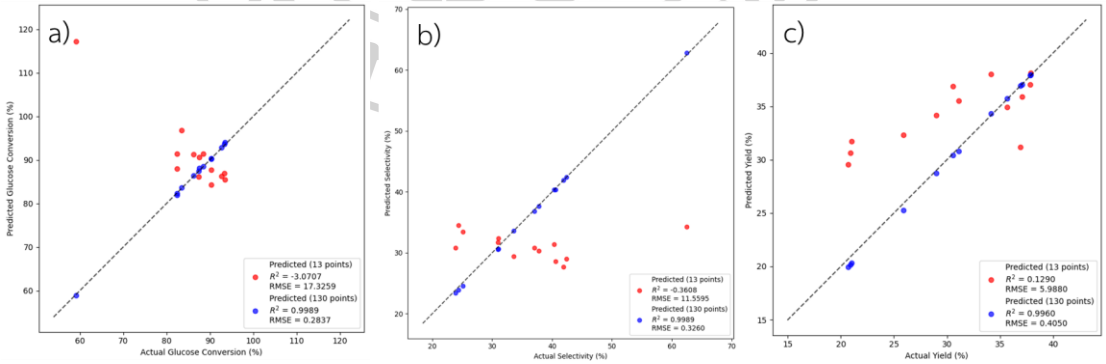

Figure 4.21 Parity Plots Comparing Model Performance Trained on Original (13 points) vs. Augmented (130 points) Data for 5-HMF Prediction: (a) Glucose Conversion, (b) Selectivity, and (c) Yield

The 5-HMF case study, which originally had only 13 data points, serves as the clearest proof of the importance of dataset size. We compared the performance of the SEM model trained on the original 13 data points (Predicted_13) with the model trained on data that was augmented 10-fold, resulting in 130 data points (Predicted_130). The results in **Figure 4.21** (Parity Plot) show a stark contrast: the model trained on 13 points (red dots) failed completely, yielding a negative R^2 value (e.g., -3.0707 for Glucose Conversion) and predicting values with significant error (e.g., actual Glucose Conversion was 59.11% while the model predicted 117.22%). In contrast, the model trained on 130 points (blue dots) was "revived" and provided remarkably accurate predictions. It achieved R^2 values approaching 1 (e.g., 0.9989 for Glucose Conversion and 0.9960 for Yield) and predicted values very close to the actual values (e.g., actual Glucose Conversion 59.11% predicted 58.929%). This case study, therefore, concludes that the Data Augmentation technique is an essential tool that can transform a dataset that is too small to be usable into one that is large enough to train a model to be highly efficient and reliably accurate.

4.9.3 The Ineffectiveness of Principal Component Analysis (PCA)

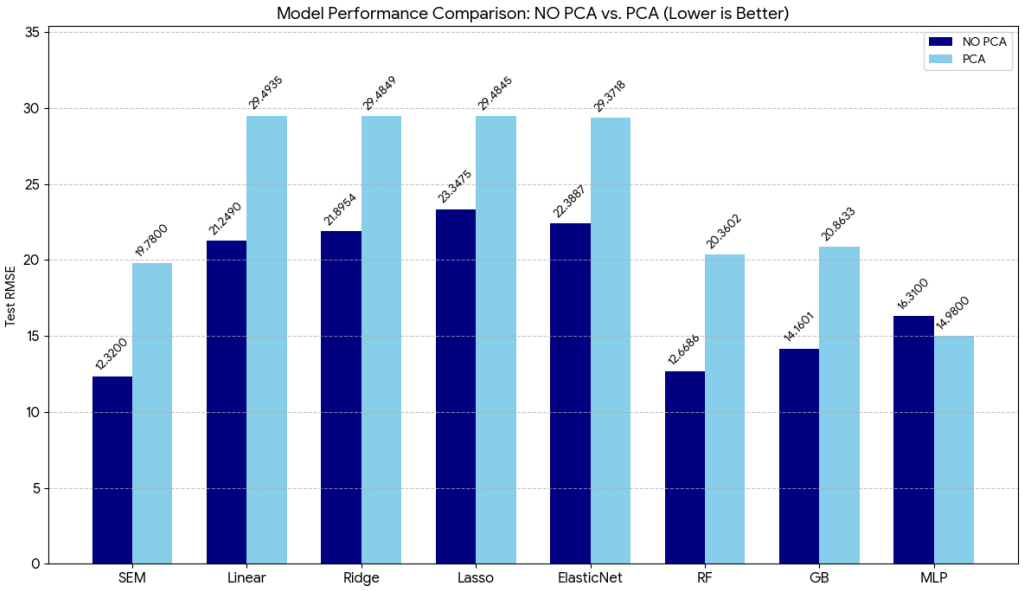

Figure 4.22 Comparison of Test RMSE for Various Models with and without PCA (CO_2 Methanation)

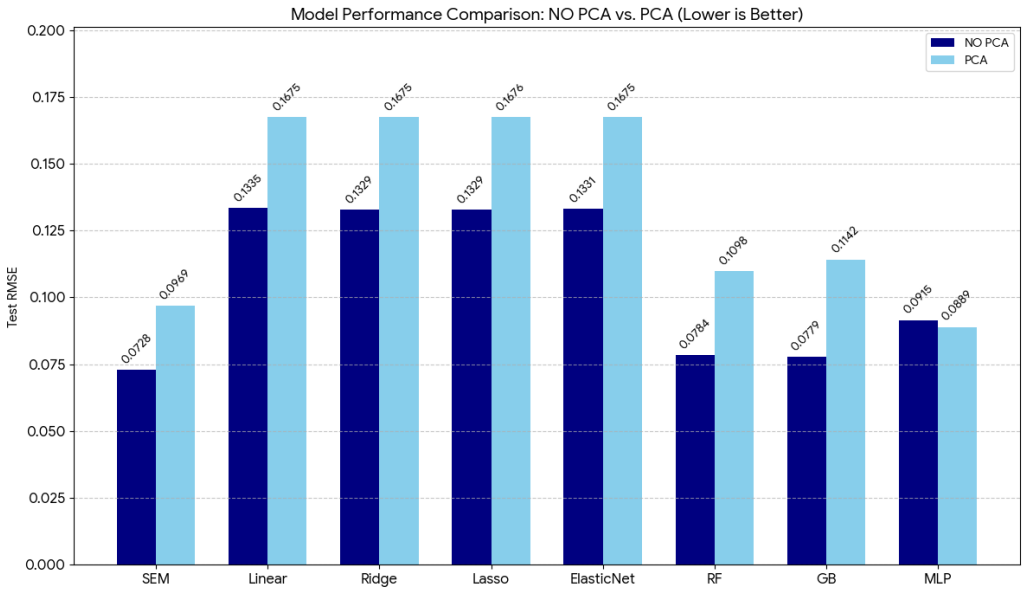

Figure 4.23 Comparison of Test RMSE for Various Models with and without PCA (CO_2 to methanol)

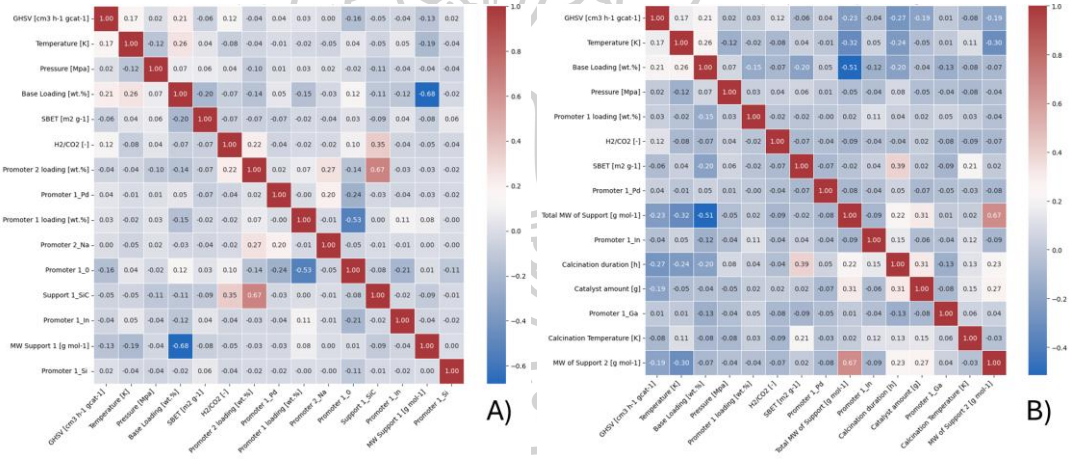

Figure 4.24 Correlation Heatmap Comparing Feature Inter-correlation for the CO_2 to Methanol Dataset: A) NO PCA vs. B) PCA

A highly significant finding from this study was that the effect of PCA was not universally negative; rather, it was entirely dependent on the model architecture. As shown in the performance comparison graphs (**Figure 4.22 and 4.23**), applying PCA before training the SEM resulted in a severe degradation of performance. The Test RMSE for CO_2 Methanation increased from 12.32 (NO PCA) to 19.78 (PCA), and for CO_2 to Methanol, it increased from 0.0728 (NO PCA) to 0.0969 (PCA). This suggests that the

SEM—particularly its tree-based components (Random Forest, Gradient Boosting)—excels at finding complex, non-linear feature interactions in the original, high-dimensional space. This is supported by the Correlation Heatmap (**Figure 4.24**), where the NO PCA heatmap (left) is predominantly white (correlation near 0), indicating the original features were "clean" and ideal for SEM to find interactions. PCA, being a linear transformation focused on variance, likely removed these critical non-linear relationships; the PCA heatmap (right) even shows new, artificial blue correlations being introduced, "dumbing down" the dataset and preventing the ensemble from leveraging its primary strength. In stark contrast, the MLP model showed the opposite effect, consistently performing better (lower RMSE) with PCA. For CO_2 Methanation, the MLP RMSE decreased from 16.31 (NO PCA) to 14.98 (PCA), and for CO_2 to Methanol, it decreased from 0.0915 (NO PCA) to 0.0889 (PCA). This suggests that a single MLP model is more sensitive to the "curse of dimensionality" and multicollinearity. By reducing the number of features and ensuring they are uncorrelated, PCA simplified the problem, likely to help the MLP to converge to a better solution more efficiently, even though the MLP itself is a non-linear model.

This discovery underscores that PCA is not a universally "good" or "bad" technique. Its effectiveness hinges on whether the model it is paired with benefits more from dimensionality reduction (like MLP) or suffers more from the loss of non-linear interactions (like SEM).

CHAPTER V
CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

The comparative study on the performance of machine learning models for predicting chemical reaction outcomes clearly demonstrates that the Stacking Ensemble Model (SEM) outperforms the Multilayer Perceptron (MLP) across all tested datasets, including CO_2 Methanation, CO_2 to Methanol, and 5-HMF. These findings indicate that the ensemble architecture, which integrates multiple base learners, is capable of generating models with higher accuracy and greater stability compared to complex single models. The "expert team" approach of SEM has proven effective in capturing the underlying relationships within chemical data more efficiently than individual models. Moreover, this study highlights two important lessons. First, data augmentation plays a critical role, particularly in cases involving small datasets, significantly enhancing model performance. Second, dimensionality reduction using Principal Component Analysis (PCA) does not always lead to improved model performance; in this study, applying PCA actually reduced predictive accuracy, emphasizing the need for careful consideration when selecting data preprocessing techniques appropriate for the problem context. In conclusion, this research not only establishes SEM as a superior modeling approach for predicting chemical reaction outcomes but also provides valuable insights into data management strategies. These findings offer practical guidance for data scientists and chemists in developing highly effective predictive models for future applications.

5.2 Recommendations

The performance of the Stacking Ensemble Model (SEM) in predicting catalytic reaction outcomes should be further examined under a wider variety of datasets to validate its robustness. Additional experiments incorporating diverse reaction conditions, catalyst compositions, and broader parameter ranges should be integrated to strengthen the structure–property relationship embedded within the model.

The dataset size for catalytic systems, particularly those with limited experimental points such as the 5-HMF dataset, should be expanded or systematically augmented. Controlled data augmentation with optimized noise factors should be further investigated to enhance the representativeness of small datasets while preserving the underlying physicochemical trends.

The influence of dimensionality reduction on machine learning performance should be further studied by testing PCA or other feature extraction techniques across different model architectures. Since PCA improved MLP performance but reduced SEM accuracy, additional analysis is recommended to clarify how latent variables correlate with catalyst descriptors and reaction responses.

Future studies should explore the transferability of learned representations between related catalytic systems, such as CO_2 hydrogenation to methane and methanol. Transfer learning approaches may provide improved predictive capability for new catalyst classes with limited data availability.

Advanced machine learning architecture includes gradient-boosted ensembles, deep neural networks, and graph-based models—should be examined to capture complex nonlinear relationships between catalyst composition, reaction conditions, and performance. Integration within situ or operando catalyst characterization data may further refine predictive accuracy and deepen mechanistic understanding.

REFERENCES

1. Lee, W.J., et al., Recent trend in thermal catalytic low temperature CO2 methanation: A critical review. Catalysis today, 2021. 368: p. 2-19.

2. Shen, L., et al., Essential role of the support for nickel-based CO2 methanation catalysts. ACS Catalysis, 2020. 10(24): p. 14581-14591.

3. Zhang, J., et al., Mechanistic understanding of CO2 hydrogenation to methane over Ni/CeO2 catalyst. Applied Surface Science, 2021. 558: p. 149866.

4. Elvers, B., Ullmann's encyclopedia of industrial chemistry. Vol. 17. 1991: Verlag Chemie Hoboken, NJ.

5. Dittmeyer, R., et al., Crowd oil not crude oil. Nature Communications, 2019. 10(1): p. 1818.

6. De, S., et al., Advances in the design of heterogeneous catalysts and thermocatalytic processes for CO2 utilization. ACS Catalysis, 2020. 10(23): p. 14147-14185.

7. Crippa, M., et al., A circular economy for plastics: Insights from research and innovation to inform policy and funding decisions. 2019.

8. Dogu, O., et al., The chemistry of chemical recycling of solid plastic waste via pyrolysis and gasification: State-of-the-art, challenges, and future directions. Progress in Energy and Combustion Science, 2021. 84: p. 100901.

9. Coates, G.W. and Y.D. Getzler, Chemical recycling to monomer for an ideal, circular polymer economy. Nature Reviews Materials, 2020. 5(7): p. 501-516.

10. Lechleitner, A., et al., Chemisches Recycling von gemischten Kunststoffabfällen als ergänzender Recyclingpfad zur Erhöhung der Recyclingquote. Österreichische Wasser-und Abfallwirtschaft, 2020. 72(1-2): p. 47-60.

11. Beckmann, M. and K. Thomé-Kozmiensky, Energie aus Abfall. Band, 2012. 2: p. 319-344.

12. Han, S.W., et al., Gasification characteristics of waste plastics (SRF) in a bubbling fluidized bed: Effects of temperature and equivalence ratio. Energy, 2022. 238: p. 121944.

13. Materazzi, M., et al., Production of BioSNG from waste derived syngas: Pilot plant operation and preliminary assessment. Waste Management, 2018. 79: p. 752-762.

14. Visconti, C.G., et al., CO2 hydrogenation to lower olefins on a high surface area K-promoted bulk Fe-catalyst. Applied Catalysis B: Environmental, 2017. 200: p. 530-542.

15. Numpilai, T., et al., Structure–activity relationships of Fe-Co/K-Al2O3 catalysts calcined at different temperatures for CO2 hydrogenation to light olefins. Applied Catalysis A: General, 2017. 547: p. 219-229.

16. Wu, T., et al., Porous graphene-confined Fe–K as highly efficient catalyst for CO2 direct hydrogenation to light olefins. ACS applied materials & interfaces, 2018. 10(28): p. 23439-23443.

17. Jordan, M.I. and T.M. Mitchell, Machine learning: Trends, perspectives, and prospects. Science, 2015. 349(6245): p. 255-260.

18. Chan, W.M., et al., Resource allocation in multiple energy-integrated biorefinery using neuroevolution and mathematical optimization. Process Integration and Optimization for Sustainability, 2021. 5: p. 383-416.

19. Arumugasamy, S.K., et al., Comparison between artificial neural networks and support vector machine modeling for polycaprolactone synthesis via enzyme catalyzed polymerization. Process Integration and Optimization for Sustainability, 2021. 5: p. 599-607.

20. Wei, S., et al., Data augmentation and machine learning techniques for control strategy development in bio-polymerization process. Environmental Science and Ecotechnology, 2022. 11: p. 100172.

21. Perez, L. and J. Wang, The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621, 2017.

22. Wiese, M., et al. Transfer learning for accurate modeling and control of soft actuators. in 2021 IEEE 4th International Conference On Soft Robotics (RoboSoft). 2021. IEEE.

23. Ye, Y., et al., Modelling of Soft Fiber-Reinforced Bending Actuators Through Transfer Learning from a Machine Learning Algorithm Trained from FEM Data. Sensors and Actuators A: Physical, 2024: p. 115095.

24. Yılmaz, B., B. Oral, and R. Yıldırım, Machine learning analysis of catalytic CO2 methanation. International Journal of Hydrogen Energy, 2023.

25. Tan, C.H., et al., Current Developments in Catalytic Methanation of Carbon Dioxide—A Review. Frontiers in Energy Research, 2022. 9: p. 795423.

26. Frontera, P., et al., The role of Gadolinia Doped Ceria support on the promotion of CO2 methanation over Ni and NiFe catalysts. International Journal of Hydrogen Energy, 2017. 42(43): p. 26828-26842.

27. Kirchner, J., et al., CO2 methanation on Fe catalysts using different structural concepts. Chemie Ingenieur Technik, 2020. 92(5): p. 603-607.

28. Büchel, R., A. Baiker, and S.E. Pratsinis, Effect of Ba and K addition and controlled spatial deposition of Rh in Rh/Al2O3 catalysts for CO2 hydrogenation. Applied Catalysis A: General, 2014. 477: p. 93-101.

29. Swalus, C., et al., CO2 methanation on Rh/γ-Al2O3 catalyst at low temperature:“In situ” supply of hydrogen by Ni/activated carbon catalyst. Applied Catalysis B: Environmental, 2012. 125: p. 41-50.

30. Kuznecova, I. and J. Gusca, Property based ranking of CO and CO2 methanation catalysts. Energy Procedia, 2017. 128: p. 255-260.

31. Suvarna, M., T.P. Araujo, and J. Pérez-Ramírez, A generalized machine learning framework to predict the space-time yield of methanol from thermocatalytic CO2 hydrogenation. Applied Catalysis B: Environmental, 2022. 315: p. 121530.

32. Chandana, K.S., et al., Machine learning aided catalyst activity modelling and design for direct conversion of CO2 to lower olefins. Journal of Environmental Chemical Engineering, 2023. 11(2): p. 109555.

33. Bhardwaj, A., et al., A principal component analysis assisted machine learning modeling and validation of methanol formation over Cu-based catalysts in direct CO2 hydrogenation. Separation and Purification Technology, 2023. 324: p. 124576.

34. Sumayli, A. and S.M. Alshahrani, Modeling and prediction of biodiesel production by using different artificial intelligence methods: Multi-layer perceptron (MLP), Gradient boosting (GB), and Gaussian process regression (GPR). Arabian Journal of Chemistry, 2023. 16(7): p. 104801.

35. Hossain, M.A., et al., Artificial neural network modeling of hydrogen-rich syngas production from methane dry reforming over novel Ni/CaFe2O4 catalysts. International Journal of Hydrogen Energy, 2016. 41(26): p. 11119-11130.

36. Sun, X., et al., Modeling and optimization of vegetable oil biodiesel production with heterogeneous nano catalytic process: Multi-layer perceptron, decision regression tree, and K-Nearest Neighbor methods. Environmental Technology & Innovation, 2022. 27: p. 102794.

37. Wolpert, D.H., Stacked generalization. Neural networks, 1992. 5(2): p. 241-259.

38. Pokrass, K. Stacked Generalization. 2019 [cited June 12, 2019; Available from: https://kpokrass.github.io/stacked_generalization.

39. Ting, K.M. and I.H. Witten, Issues in stacked generalization. Journal of artificial intelligence research, 1999. 10: p. 271-289.

40. Alexandropoulos, S.-A.N., et al. Stacking strong ensembles of classifiers. in IFIP International Conference on Artificial Intelligence Applications and Innovations. 2019. Springer.

41. Seewald, A.K., Meta-learning for stacked classification. audiology, 2002. 24(226): p. 69.

42. Ting, K.M. and I.H. Witten, Stacked Generalization: when does it work? 1997.

43. Nukui, T. and A. Onogi, An R package for ensemble learning stacking. Bioinformatics Advances, 2023. 3(1): p. vbad139.

44. Nguyen Van, L. and G. Lee, Optimizing Stacked Ensemble Machine Learning Models for Accurate Wildfire Severity Mapping. Remote Sensing, 2025. 17(5): p. 854.

45. Zhu, S., et al., A Financial Fraud Prediction Framework Based on Stacking Ensemble Learning. Systems, 2024. 12(12): p. 588.

46. Madadi, Y., et al. Stacking Ensemble Learning in Deep Domain Adaptation for Ophthalmic Image Classification. 2021. Cham: Springer International Publishing.

47. Abdellatif, A., et al., Forecasting Photovoltaic Power Generation with a Stacking Ensemble Model. Sustainability, 2022. 14(17): p. 11083.

48. Ma, Z., et al., Hydrogen yield prediction for supercritical water gasification based on generative adversarial network data augmentation. Applied Energy, 2023. 336: p. 120814.

49. Mumuni, A. and F. Mumuni, Data augmentation: A comprehensive survey of modern approaches. Array, 2022: p. 100258.

50. Williams, L., Principal component analysis, Wiley Interdisciplinary Reviews: Computational Statistics. 2010.

51. Jolliffe, I.T., Springer series in statistics. Principal component analysis, 2002. 29: p. 912.

52. Hotelling, H., Analysis of a complex of statistical variables into principal components. Journal of educational psychology, 1933. 24(6): p. 417.

53. Callén, M.S., et al., Principal component analysis and partial least square regression models to understand sorption-enhanced biomass gasification. Biomass Conversion and Biorefinery, 2024. 14(2): p. 2091-2111.

54. JARUNOTHAI, P. and N. Chotigkrai, Solution combustion synthesis of Nb-Al and W-Al oxide catalysts for direct production of 5-HMF from glucose. 2022, Silpakorn University.

55. Liaw, A. and M. Wiener, Classification and regression by randomForest. R news, 2002. 2(3): p. 18-22.

APPENDIX

APPENDIX I

Reduce input with principal analysis component

As a result of the PCA process code shown below, the feature space was reduced from 105 original inputs to 72 principal components, retaining 95% of the data's variance. **Table I.1** displays a sample of these newly generated inputs (Principal Components), revealing the loading coefficients that determine the contribution of each original variable to the new components.

Previews code

```
pca = PCA(n_components=0.95)

X_train_pca = pca.fit_transform(X_train_scaled)

X_test_pca = pca.transform(X_test_scaled)
```

Table I.1 Example of PCA

	Base Loading	CR Metal	MW Support 1	MW of Support 2	...	Temperature
PC1	-0.1913	-0.2038	0.1535	0.2687	...	-0.1843
PC2	0.2901	0.2649	-0.1956	0.1142	...	0.0795
PC3	0.0337	0.0852	-0.0824	-0.0844	...	0.1020
PC4	0.1484	0.0727	-0.1811	0.2226	...	-0.0540
PC5	-0.1460	0.2163	0.2427	0.0268	...	0.0271
...	...	...	...	...	...	...
PC68	0.0162	0.0514	-0.0042	0.0908	...	0.1151
PC69	-0.0351	-0.0297	-0.0114	0.0773	...	0.1291
PC70	-0.0145	0.0647	0.0388	-0.0529	...	-0.1855
PC71	-0.0390	-0.0030	-0.0122	0.0063	...	-0.1285
PC72	0.0088	0.0642	-0.0030	0.0217	...	0.0519

Example

PC1 = (-0.1913*Base Loading) + (-0.2038*CR Metal) + (0.1535*MW Support 1) + ...

(factor * end of features)

APPENDIX II

A. Final Optimized Hyperparameters

- **The best parameters for CO_2 methanation**

Fixed parameters activation='relu'

solver='adam'

max_iter=1000000

random_state=42

Table II.1 Optimized hyperparameters for the MLP model applied to the CO_2 Methanation dataset.

Dataset	Adjustable parameters	
	alpha	hidden layer sizes
Original	2.76E-03	(220, 73, 276, 193, 19, 144)
0.05_aug2	1.04E-04	(217, 165, 254, 193)
0.05_aug5	2.76E-03	(217, 165, 254, 193)
0.05_aug10	5.36E-04	(260, 103, 175, 67, 221, 21, 84, 181)
0.1_aug2	2.76E-03	(260, 103, 175, 67, 221, 21, 84, 181)
0.1_aug5	1.66E-04	(217, 165, 254, 193)
0.1_aug10	8.55E-04	(217, 165, 254, 193)
0.2_aug2	1.66E-04	(217, 165, 254, 193)
0.2_aug5	4.08E-05	(260, 103, 175, 67, 221, 21, 84, 181)
0.2_aug10	3.22E-05	(260, 103, 175, 67, 221, 21, 84, 181)

Table II.2 Optimized hyperparameters for the MLP model with PCA applied to the CO_2 Methanation dataset.

Dataset	Adjustable parameters		
	n_components	alpha	hidden layer sizes
Original	0.93	6.51E-05	(298, 82, 22, 143, 15, 107, 78, 226)
0.05_aug2	0.92	1.00E-05	(179, 156, 104, 226)
0.05_aug5	0.89	3.63E-02	(298, 34, 282, 87, 121, 105, 12, 115, 202, 111)
0.05_aug10	0.91	1.00E-05	(217, 96, 122, 133)
0.1_aug2	0.93	6.77E-04	(142, 271, 258)
0.1_aug5	0.89	8.55E-04	(298, 82, 22, 143, 15, 107, 78, 226)
0.1_aug10	0.93	1.37E-03	(220, 73, 276, 193, 19, 144)
0.2_aug2	0.91	1.26E-05	(193, 186, 145, 32, 245, 284, 73, 203, 50)
0.2_aug5	0.89	5.36E-04	(229, 183, 240, 127, 86, 187)
0.2_aug10	0.96	8.23E-05	(152, 89, 120, 182, 62, 57, 204, 59)

Table II.3 Optimized hyperparameters for the SEM model applied to the CO_2 Methanation dataset.

Base learners	Parameters
LR	Default
Ridge	max_iter=1000000, random_state=42
ElasticNet	max_iter=1000000, random_state=42
Lasso	max_iter=1000000, random_state=42
RF	n_estimators=1150, max_depth=25, random_state=42
GB	n_estimators=1150, max_depth=5, random_state=42
MLP	The best from previous experiments
Ridge (Meta learner)	max_iter=1000000, random_state=42
SEM	cv=20 for original data and PCA cv=10 for augmented data

- **The best parameters for CO_2 to methanol**

  Fixed parameters activation='relu'

  solver='adam'

  max_iter=1000000

  random_state=42

Table II.4 Optimized hyperparameters for the MLP model applied to the CO_2 to methanol dataset.

Dataset	Adjustable parameters	
	alpha	hidden layer sizes
Original	9.25E-02	(225, 30)
0.05_aug2	9.25E-02	(217, 96, 122, 133)
0.05_aug5	1.42E-02	(62, 289, 54, 226)
0.05_aug10	8.23E-05	(229, 183, 240, 127, 86, 187)
0.1_aug2	4.58E-02	(217, 96, 122, 133)
0.1_aug5	2.87E-02	(62, 289, 54, 226)
0.1_aug10	1.80E-02	(135, 124, 81)
0.2_aug2	9.25E-02	(100, 240, 262, 286, 21, 100, 15, 83)
0.2_aug5	4.58E-02	(47, 179, 251, 296, 61, 191, 232)
0.2_aug10	1.42E-02	(290, 17, 57, 214, 10, 262, 180, 134, 176)

*Table II.**5** Optimized hyperparameters for the MLP model with PCA applied to the CO_2 to methanol dataset.*

Dataset	Adjustable parameters		
	n_components	alpha	hidden layer sizes
Original	0.95	9.25E-02	(142, 271, 258)
0.05_aug2	0.99	1.17E-01	(47, 179, 251, 296, 61, 191, 232)
0.05_aug5	0.91	3.63E-02	(268, 23, 297, 111, 289, 224)

Dataset	Adjustable parameters		
	n_components	alpha	hidden layer sizes
0.05_aug10	0.96	3.35E-04	(58, 138, 282, 85, 168, 60)
0.1_aug2	0.91	1.17E-01	(98, 157, 173, 294, 175)
0.1_aug5	0.91	5.79E-02	(225, 30)
0.1_aug10	0.91	2.87E-02	(103, 230, 69, 174, 130, 228, 281)
0.2_aug2	0.91	1.48E-01	(251, 49, 295, 48, 104, 189, 91, 203, 73, 56)
0.2_aug5	0.95	4.58E-02	(245, 197, 200, 14, 296, 79, 15, 142, 228)
0.2_aug10	0.89	4.58E-02	(100, 240, 262, 286, 21, 100, 15, 83)

Table II.6 Optimized hyperparameters for the SEM model applied to the CO_2 to methanol dataset.

Base learners	Parameters
LR	Default
Ridge	max_iter=1000000, random_state=42
ElasticNet	max_iter=1000000, random_state=42
Lasso	max_iter=1000000, random_state=42
RF	For original data and PCA n_estimators=65, max_depth=15, random_state=42 For augmented data n_estimators=1150, max_depth=25, random_state=42
GB	For original data and PCA n_estimators=1000, max_depth=4, random_state=42 For augmented data n_estimators=1150, max_depth=5, random_state=42
MLP	The best from previous experiments
Ridge (Meta learner)	max_iter=1000000, random_state=42
SEM	cv=20 for original data and PCA cv=5 for augmented data

- **The best parameters for 5-HMF**

  Fixed parameters activation='relu'

  solver='adam'

  max_iter=1000000

  random_state=42

Table II.7 Optimized hyperparameters for the MLP model in the 5-HMF dataset.

Dataset		Adjustable parameters	
		alpha	hidden layer sizes
Original	G	2.98E-01	(54,)
Augmented 10x		4.76E-01	(260,)
Original	S	1.48E-01	(58, 138, 282, 85, 168, 60)
Augmented 10x		7.61E-01	(290, 17, 57, 214, 10, 262, 180, 134, 176)
Original	Y	1.48E-01	(58, 138, 282, 85, 168, 60)
Augmented 10x		6.02E-01	(290, 17, 57, 214, 10, 262, 180, 134, 176)

G is Glucose conversion%, S is Selectivity% and Y is yield%

Table II.8 Optimized hyperparameters for the MLP model in the 5-HMF dataset.

Base learners	Parameters
LR	Default
Ridge	max_iter=1000000, random_state=42
ElasticNet	max_iter=1000000, random_state=42
Lasso	max_iter=1000000, random_state=42
RF	n_estimators=100, random_state=42
GB	n_estimators=100, random_state=42
MLP	The best from previous experiments
Ridge (Meta learner)	max_iter=1000000, random_state=42
SEM	cv=5

B. Key Python Snippets

- **Data Augmentation Procedure**

1. Library Importing

   The Python libraries pandas and numpy were imported to support data manipulation and numerical operations.

2. Data Loading

   The original dataset was loaded from a CSV file into a pandas DataFrame, which served as the base data for augmentation.

3. Parameter Definition

   Two parameters were specified prior to augmentation:

   - the number of augmentation rounds (aug_mult), and
   - the noise intensity coefficient (β), which controls the magnitude of injected Gaussian noise.

4. Preparation of Output Structure

   An empty matrix was created to store both the original dataset and all augmented samples. Its size was determined by the number of rows, number of features, and the total number of augmentation rounds.

5. Gaussian Noise Augmentation

   For each augmentation round, the following steps were applied:

   - In the first round, the unmodified original data were stored.
   - For all subsequent rounds, random noise was generated from a standard normal distribution ($\mu = 0$, $\sigma = 1$).
   - The noise was scaled by β and added to each numerical feature.
   - The absolute value of each resulting feature was taken to ensure all values remained non-negative and physically meaningful.

6. Combining Augmented Data

   All generated data—including the original and augmented samples—were concatenated into a single output matrix (output_mult), which was later used for model training.

```python
import pandas as pd
import numpy as np

# Load dataset
PATH_MEOH = 'local/path/to/data.csv'
data = pd.read_csv(PATH_MEOH)

# Define augmentation parameters
aug_mult = 5            # Number of augmentation rounds
beta = 0.1              # Noise intensity coefficient

num_rows = data.shape[0]
num_cols = data.shape[1]

# Output matrix for original + augmented data
output_mult = np.zeros((num_rows * aug_mult, num_cols))

# Augmentation loop
for aug in range(aug_mult):
    data_cpy = []

    for i in range(num_cols):
        original_column = data.values[:, i]

        # Augmented data (skip augmentation for round 0)
        if aug != 0:
            noise = np.random.normal(0, 1, num_rows)
            augmented_values = np.abs(original_column + beta * noise)
            data_cpy.append(augmented_values)
        else:
            # First round = original data
            data_cpy.append(original_column)

    # Convert list → matrix
    data_cpy = np.array(data_cpy).T

    # Insert into output matrix
    output_mult[num_rows * aug : num_rows * (aug + 1), :] = data_cpy
```

- **Stacking Ensemble Modeling Procedure**

1. Importing Required Libraries

   Machine learning libraries from scikit-learn were imported to construct the base models, meta-model, and the final stacking regressor. Additional libraries, such as NumPy and metrics modules, were used for numerical operations and performance evaluation.

2. Defining Base Models

   A set of diverse regression algorithms was defined as base learners to capture different patterns within the data. The base models included:

   - Linear Regression
   - Ridge Regression
   - Lasso Regression
   - Elastic Net
   - Random Forest Regressor

- Gradient Boosting Regressor
- Multilayer Perceptron (MLP) Regressor

Each model was configured with specific hyperparameters (e.g., number of estimators for Random Forest, learning settings for MLP, and maximum iterations for linear models).

3. Defining the Meta-model

   A Ridge Regression model was selected as the final estimator (meta-model). Its role is to learn from the predictions of the base models and combine them optimally into a final output.

4. Constructing the Stacking Regressor

   A StackingRegressor was created with the following key configurations:

   - Base learners: the list of models defined earlier
   - Meta-model: Ridge Regression
   - Cross-validation (cv = 20) to ensure stable internal model training
   - Passthrough = False, the meta-model receives only the predictions from the base models, excluding the original feature set.
   - n_jobs = -1, enabling parallel processing for faster computation

5. Model Training

   The stacking model was trained using the scaled training dataset (X_train_scaled, y_train). The .fit() function trained all base models followed by the meta-model on cross-validated predictions.

6. Generating Predictions

   After training, predictions were obtained for both:

   - Training set – to assess how well the model fits the known data
   - Testing set – to evaluate generalization performance

7. Model Evaluation

   Performance metrics were calculated:

   - Coefficient of Determination (R^2)
   - Root Mean Squared Error (RMSE)

These metrics were computed for both training and test datasets to assess model accuracy and detect potential overfitting.

8. Output Reporting

The evaluation results were formatted and printed, allowing for direct comparison between training and testing performance.

```python
# Base models
base_models = [
    ('lr', LinearRegression()),
    ('ridge', Ridge(max_iter= 1000000, random_state=42)),
    ('lasso', Lasso(alpha= 0.0001, selection= "random", max_iter= 1000000, random_state=42)),
    ('elastic', ElasticNet(alpha= 0.0001, l1_ratio= 0.5, max_iter= 1000000, random_state=42)),
    ('rf', RandomForestRegressor(n_estimators=65,
                                 max_depth=15,
                                 random_state=42)),
    ('gb', GradientBoostingRegressor(n_estimators=1000,
                                     max_depth=4,
                                     random_state=42)),
    ('MLP',MLPRegressor(hidden_layer_sizes=(225,30),
                        activation='relu',
                        solver='adam',
                        alpha=0.092491473,
                        max_iter=1000000,
                        random_state=42))
]
```

```python
# Meta model
meta_model = Ridge(max_iter= 1000000,
                   random_state=42)
# StackingRegressor
stack = StackingRegressor(estimators=base_models,
                          final_estimator=meta_model,
                          cv=20,                    # ทำ cross-validation ภายใน stacking
                          passthrough=True,         # True = รวม feature ดิบเข้าด้วยกับ prediction ของ base model (optional)
                          n_jobs=-1                 # รันหลาย core ได้
                          )
# Train the stacking model
stack.fit(X_train_scaled, y_train.ravel())
# Predict
stack_train_pred = stack.predict(X_train_scaled)
stack_test_pred = stack.predict(X_test_scaled)
# Evaluate
stacking_metrics = {'Train R²': r2_score(y_train, stack_train_pred),
                    'Train RMSE': np.sqrt(mean_squared_error(y_train, stack_train_pred)),
                    'Test R²': r2_score(y_test, stack_test_pred),
                    'Test RMSE': np.sqrt(mean_squared_error(y_test, stack_test_pred)),
                    }
print("📊 Stacking Regressor Performance:")
for k, v in stacking_metrics.items():
    print(f"{k}: {v:.4f}")
```

APPENDIX III

Comprehensive Results

Table III.1 All results for MLP model in the CO_2 methanation datasets.

Data Condition	Augmentation	MLP (Train Set)		MLP (Test Set)	
Beta	Multiplier	RMSE	R^2	RMSE	R^2
Original	1x	9.01	0.9201	16.31	0.7437
0.05	2x	4.18	0.9834	14.60	0.7929
	5x	3.45	0.9887	14.30	0.8015
	10x	1.99	0.9963	14.50	0.7956
0.1	2x	4.42	0.9815	14.50	0.7957
	5x	3.03	0.9913	14.33	0.8006
	10x	2.56	0.9938	14.45	0.7971
0.2	2x	4.89	0.9774	14.26	0.8024
	5x	3.22	0.9902	14.00	0.8095
	10x	3.40	0.9891	14.61	0.7925

Table III.2 All results for MLP model with PCA in the CO_2 methanation datasets.

Data Condition	Augmentation	PCA + MLP (Train Set)		PCA + MLP (Test Set)	
Beta	Multiplier	RMSE	R^2	RMSE	R^2
Original	1x	8.47	0.9294	14.98	0.7820
0.05	2x	5.46	0.9718	14.75	0.7888
	5x	5.14	0.9750	15.85	0.7558
	10x	4.04	0.9846	14.96	0.7825
0.1	2x	4.90	0.9773	15.09	0.7787
	5x	5.31	0.9733	16.82	0.7252
	10x	3.44	0.9888	15.25	0.7741
0.2	2x	4.64	0.9796	14.56	0.7942
	5x	5.24	0.9740	16.73	0.7279
	10x	4.04	0.9845	14.68	0.7907

Table III.3 All results for MLP model in the CO_2 methanation datasets.

Data Condition	Augmentation	SEM (Train Set)		SEM (Test Set)	
Beta	Multiplier	RMSE	R^2	RMSE	R^2
Original (NO PCA)	1x	5.83	0.9679	12.35	0.8519
Original (PCA)	1x	5.59	0.9704	19.78	0.6198
0.05	2x	3.11	0.9909	13.11	0.8330
	5x	2.40	0.9945	12.16	0.8563
	10x	1.36	0.9982	12.68	0.8437
0.1	2x	3.61	0.9877	13.77	0.8159
	5x	2.57	0.9938	12.27	0.8537
	10x	1.54	0.9978	12.78	0.8413
0.2	2x	3.73	0.9869	13.19	0.8311
	5x	2.46	0.9943	12.47	0.8489
	10x	1.70	0.9973	12.70	0.8432

Table III.4 All results for MLP model in the CO_2 to methanol datasets.

Data Condition	Augmentation	MLP (Train Set)		MLP (Test Set)	
Beta	Multiplier	RMSE	R^2	RMSE	R^2
Original	1x	0.0719	0.9393	0.0915	0.8603
0.05	2x	0.0630	0.9535	0.0919	0.8591
	5x	0.0509	0.9696	0.0934	0.8546
	10x	0.0540	0.9658	0.0899	0.8650
0.1	2x	0.0539	0.9658	0.0932	0.8551
	5x	0.0550	0.9645	0.0938	0.8531
	10x	0.0563	0.9628	0.0915	0.8603
0.2	2x	0.0632	0.9531	0.0940	0.8526
	5x	0.0613	0.9559	0.0941	0.8523
	10x	0.0548	0.9647	0.0905	0.8635

Table III.5 All results for MLP model with PCA in the CO_2 to methanol datasets.

Data Condition	Augmentation	PCA + MLP (Train Set)		PCA + MLP (Test Set)	
Beta	Multiplier	RMSE	R^2	RMSE	R^2
Original	1x	0.0789	0.9270	0.0888	0.8684
0.05	2x	0.0650	0.9504	0.0870	0.8736
	5x	0.0650	0.9503	0.0897	0.8657
	10x	0.0505	0.9700	0.0948	0.8502
0.1	2x	0.0694	0.9435	0.0890	0.8680
	5x	0.0744	0.9350	0.0892	0.8671
	10x	0.0589	0.9593	0.0897	0.8658
0.2	2x	0.0820	0.9210	0.0885	0.8692
	5x	0.0662	0.9485	0.0882	0.8702
	10x	0.0674	0.9466	0.0883	0.8700

Table III.6 All results for SEM model in the CO_2 to methanol datasets.

Data Condition	Augmentation	SEM (Train Set)		SEM (Test Set)	
Beta	Multiplier	RMSE	R^2	RMSE	R^2
Original (NO PCA)	1x	0.0281	0.9907	0.0728	0.9116
Original (PCA)	1x	0.0593	0.9587	0.0948	0.8500
0.05	2x	0.0271	0.9914	0.0863	0.8756
	5x	0.0140	0.9977	0.0813	0.8896
	10x	0.0077	0.9993	0.0850	0.8794
0.1	2x	0.0263	0.9919	0.0777	0.8992
	5x	0.0138	0.9978	0.0783	0.8977
	10x	0.0079	0.9993	0.0836	0.8833
0.2	2x	0.0273	0.9913	0.0770	0.9011
	5x	0.0147	0.9975	0.0823	0.8870
	10x	0.0108	0.9986	0.0858	0.8772

Table III.7 All results for MLP model in the 5-HMF datasets.

Data Condition	Augmented	Output	LOO + MLP (Train set)		LOO + MLP (Test set)	
			RMSE	R^2	RMSE	R^2
Original	1x	Glucose conversion%	0.0291	1.0000	8.7071	-0.0281
0.1	10x		0.0457	1.0000	2.9054	0.8855
Original	1x	Selectivity%	0.0921	0.9999	12.0162	-0.4705
0.1	10x		0.0567	0.9999	3.4348	0.7134
Original	1x	Yield%	0.0736	0.9999	7.7921	-0.4749
0.1	10x		0.0526	1.0000	4.5344	0.7906

Table III.8 All results for SEM model in the 5-HMF datasets.

Data Condition	Augmented	Output	LOO + SEM (Train set)		LOO + SEM (Test set)	
			RMSE	R^2	RMSE	R^2
Original	1x	Glucose conversion%	20.8594	-4.9004	12.1527	-1.0027
0.1	10x		0.3956	0.9979	0.6457	0.9943
Original	1x	Selectivity%	8.8561	0.2012	13.4841	-0.8517
0.1	10x		0.3517	0.9970	1.6412	0.9346
Original	1x	Yield%	7.7701	-0.4666	11.4975	-2.2112
0.1	10x		0.2604	0.9993	1.1279	0.9870

VITA

NAME	**Mr. Kittithat Kongtaworn**
INSTITUTIONS ATTENDED	Bachelor's degree : Department of chemical engineering, Faculty of Engineering and Industrial Technology, Silpakorn University
PUBLICATION	Removal of paraquat from aqueous media via HFSLM and mathematical modeling
AWARD RECEIVED	-