

**PREDICTION OF PERFORMANCE AND
EMISSIONS, AND FAULT PROGNOSIS IN STEAM
BOILERS USING ARTIFICIAL INTELLIGENCE**

BY

AGUS HAERUMAN

A Thesis Presented to the
COLLEGE OF GRADUATE AND INTERDISCIPLINARY STUDIES
KING FAHD UNIVERSITY OF PETROLEUM & MINERALS
DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

In

MECHANICAL ENGINEERING

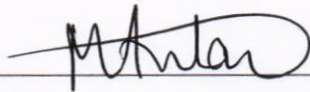
DECEMBER 2024

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

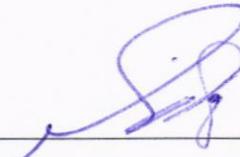
DHAHRAN- 31261, SAUDI ARABIA

DEANSHIP OF GRADUATE STUDIES

This thesis, written by **AGUS HAERUMAN** under the direction of his thesis advisor and approved by his thesis committee, has been presented and accepted by the Dean of Graduate Studies, in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE IN MECHANICAL ENGINEERING**.



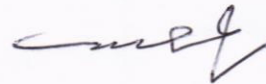
Dr. Mohamed Abdelkarim Antar
Department Chairman



Dr. Fahad Mohammed Alzahrani
(Advisor)



Prof. Suliman Saleh Al-Homidan
Dean of Graduate Studies



Prof. Mohammed Abdulaziz Habib
(Member)

Date



Dr. Maad Mohammed Al Owaifeer
(Member)



© Agus Haeruman

2024

Dedicated to my beloved wife, my children, my parents, and my family.

ACKNOWLEDGMENTS

All the praises to Allah ﷻ, The Most Gracious, The Most Merciful, for granting me the strength, and wisdom, to complete my master's degree at King Fahd University of Petroleum & Minerals (KFUPM). May blessings and peace be continuously bestowed upon Prophet Muhammad ﷺ, his families, his companions, and the people who follow his pathway.

I would like to express my profound gratitude to King Fahd University of Petroleum & Minerals (KFUPM) for granting me the opportunity to pursue my studies. Especially to the Mechanical Engineering Department, Prof. M. Abdelkarim Antar, and his team for their continuous support and encouragement throughout this journey.

My deepest gratitude goes to my advisor, Dr. Fahad M. Alzahrani, for his exceptional guidance, patience, and constant encouragement throughout my research. I am deeply thankful for your mentorship. To my academic advisor, Prof. Esmail Mokheimer, for advising me throughout my academic journey at KFUPM.

I would also like to thank my thesis committee members, Prof. Mohammed Abdulaziz Habib and Dr. Maad Mohammed Al Owaifeer, for their insightful feedback and valuable contributions to improving the quality of this work.

To my beloved wife, Lisda Maulani, your unconditional love, patience, and support have been my greatest source of strength. To my wonderful children, Arfan, Sarah, Hanna, and Haura, you are my pride and inspiration. To my parents and family for their boundless love, prayers, and unwavering support. This work is dedicated to you all with my deepest love.

TABLE OF CONTENTS

ACKNOWLEDGMENTS	v
TABLE OF CONTENTS	vi
LIST OF TABLES	ix
LIST OF FIGURES	x
ABSTRACT	xiv
الملخص.....	xvi
CHAPTER 1 INTRODUCTION.....	1
1.1 Background	1
1.2 Problem Statement	4
1.3 Objectives	5
CHAPTER 2 LITERATURE REVIEW.....	7
2.1 Boiler Systems	7
2.2 NO _x Emissions	9
2.3 Boiler Efficiency	12
2.4 Soft Sensor	14
2.4.1 Soft Sensor in Power Plant.....	15
2.4.2 Soft Sensor in Industrial Process	16
2.4.3 Soft Sensor for Emission and Efficiency Prediction	18
2.5 Fault Prognosis.....	19
2.6 Research Gaps.....	20

CHAPTER 3	DATA COLLECTION.....	22
3.1	Dataset from an Industrial Boiler Fueled with Natural Gas	24
3.2	Dataset from an Industrial Boiler Fueled with Hydrogen-Enriched Natural Gas	26
CHAPTER 4	METHODOLOGY.....	35
4.1	Data Pre-processing	36
4.2	Feature Engineering	39
4.3	Feature Selection.....	41
4.3.1	Variance Threshold.....	42
4.3.2	Pearson Correlation and Feature Importance	43
4.4	Building the AI Models	52
4.4.1	Linear Regression (LR).....	53
4.4.2	Decision Tree (DT)	54
4.4.3	Random Forest (RF).....	54
4.4.4	K Nearest Neighbors (KNN).....	55
4.4.5	Support Vector Regressor (SVR).....	57
4.4.6	Multi-Layer Perceptron (MLP)	58
4.4.7	Radial Basis Function (RBF)	59
4.4.8	Long Short Term Memory (LSTM).....	60
4.5	Performance Metrics	62
4.6	Data Augmentation for Model Validation.....	63
CHAPTER 5	RESULTS AND DISCUSSIONS	65
5.1	Industrial Boiler Fueled with Natural Gas.....	65
5.1.1	NO _x Emission Prediction	66

5.1.2	Efficiency Prediction.....	72
5.2	Industrial Boiler Fueled with Hydrogen-Enriched Natural Gas.....	79
5.2.1	NO _x Emissions Prediction Using Original Dataset.....	79
5.2.2	NO _x Emissions Prediction Using Reduced Dataset	85
5.2.3	NO _x Emissions Prediction Using Augmented Dataset.....	87
5.3	Fault Prognosis	91
CHAPTER 6 CONCLUSIONS AND RECOMMENDATIONS		100
6.1	Conclusions	100
6.2	Recommendations	102
REFERENCES.....		103
LIST OF ABBREVIATIONS		114
VITAE.....		119

LIST OF TABLES

Table 1	Siebert Constants.....	13
Table 2	Variable Description from Natural Gas Boiler Dataset.....	24
Table 3	CFD data generated with various hydrogen fractions	28
Table 4	CFD Data Generated with various Lambda	30
Table 5	Variable Descriptions of the CFD Dataset	32
Table 6	List of missing data in the natural gas boiler dataset	37
Table 7	The variance value of CFD Hydrogen-enriched boiler Dataset	42
Table 8	Input parameter for NO _x emission NG Boiler Dataset.....	47
Table 9	Input parameter for Efficiency Natural Gas Boiler Dataset	49
Table 10	Input parameter for NO _x emission CFD Dataset after feature selections.....	51
Table 11	Hyperparameters for MLP models	59
Table 12	Performance Metrics of NO _x Prediction Boiler NG Dataset (19 inputs)	67
Table 13	Performance Metrics of NO _x Prediction Boiler NG Dataset (12 inputs)	71
Table 14	Performance Metrics of Efficiency Prediction Boiler NG Dataset (18 inputs) 73	
Table 15	Performance Metrics of Efficiency Prediction Boiler NG Dataset (12 inputs) 78	
Table 16	Performance Metrics of Various ML/DL Models	80
Table 17	Performance Metrics of RBF with 13 input and 8 input features	85
Table 18	RBF Performance using Augmented Data	88
Table 19	MLP Performance using Augmented Data	89
Table 20	Test Accuracy of each LSTM Model	95

LIST OF FIGURES

Figure 1	Comparison of energy use in the last decade.....	2
Figure 2	Schematic of a fire-tube boiler.....	8
Figure 3	Schematic of a water-tube boiler	9
Figure 4	Design of the Boiler Furnace: (a) Front view, (b) Top view	23
Figure 5	Boiler radiation and convection sections	24
Figure 6	Visualization of boiler parameters from actual boiler: (a) Feed Water Flow Rate, (b) Steam Flow Rate, (c) Fuel Flow Rate, (d) Air Flow Rate, (e) Steam Drum Pressure, (f) Steam Drum Temperature	26
Figure 7	Temperature distribution on a vertical plane at lambda 1.15: (a) Fuel mixture 50% H ₂ and 50% CH ₄ , (b) Fuel mixture 100% H ₂ and 0% CH ₄	27
Figure 8	Thermal NO distribution on a vertical plane at lambda 1.15: (a) Fuel mixture 50% H ₂ and 50% CH ₄ , (b) Fuel mixture 100% H ₂ and 0% CH ₄	27
Figure 9	Effect of H ₂ Fraction on Temperature and NO _x Formation	29
Figure 10	Effect of lambda variations on temperature and NO formation	31
Figure 11	Visualization of Training and Testing Datasets: (a) Thermal NO _x , (b) Prompt NO _x	34
Figure 12	General methodology used in the current study	36
Figure 13	Various methods for handling missing data: (a) Original data, (b) Mean method, (c) Median method, (d) Mode method, (e) Linear Interpolation ...	38
Figure 14	NO _x emission obtained using feature engineering by Li and Thompson formula.....	40

Figure 15	Efficiency obtained using feature engineering by the Siegert formula.....	41
Figure 16	Pearson Correlation for NO _x of NG dataset variables	46
Figure 17	Feature Importance for NO _x of NG dataset variables.....	46
Figure 18	Pearson Correlation for Efficiency of NG Dataset Variables.....	48
Figure 19	Feature Importance for Efficiency of NG Dataset Variables	48
Figure 20	Pearson Correlation for the CFD dataset	50
Figure 21	Feature Importance using Extra Trees Algorithm	51
Figure 22	Illustration of Linear Regression	53
Figure 23	Illustration of Decision Tree Regression	54
Figure 24	Illustration of Random Forest.....	55
Figure 25	Illustration of K-nearest neighbors	56
Figure 26	Illustration of One-Dimensional Support Vector Regression.....	57
Figure 27	Structure of multi-layer perceptron (MLP).....	58
Figure 28	Structure of RBF Neural Networks.....	59
Figure 29	Structure of cell in long short term memory (LSTM) neural networks.....	61
Figure 30	Flowchart of Data Augmentation.....	64
Figure 31	Training and Testing Datasets of NO _x Emission from: (a) Training Set 80%, (b) Testing Set 20%	66
Figure 32	Metrics Performance of each model: (a) MAE, (b) R ²	68
Figure 33	Predicted vs Actual NO _x value: (a) LR, (b) DT, (c) RF, (d) KNN, (e) SVR, (f) MLP, (g) RBF	69
Figure 34	Testing Simulation for each model: (a) LR, (b) DT, (c) RF, (d) KNN, (e) SVR, (f) MLP, (g) RBF	70

Figure 35	Thermal NO _x comparison from Boiler NG dataset using RBF neural networks between actual (red), prediction 19 inputs (blue), and prediction 12 inputs (green).....	72
Figure 36	Training and Testing Datasets of Efficiency: (a) Training Set 80%, (b) Testing Set 20%	73
Figure 37	Metrics Performance of each model: (a) MAE, (b) R ²	74
Figure 38	Predicted vs Actual Efficiency value: (a) LR, (b) DT, (c) RF, (d) KNN, (e) SVR, (f) MLP, (g) RBF	76
Figure 39	Testing Simulation for each model: (a) LR, (b) DT, (c) RF, (d) KNN, (e) SVR, (f) MLP, (g) RBF	77
Figure 40	Efficiency comparison from Boiler NG dataset using RBF neural networks between actual (red), prediction 19 inputs (blue), and prediction 12 inputs (green).....	79
Figure 41	Metrics Performance of each model: (a) MAE, (b) PA	81
Figure 42	Predicted vs Actual NO _x value: (a) LR, (b) DT, (c) RF, (d) KNN, (e) SVR, (f) MLP, (g) RBF	83
Figure 43	Testing Simulation: (a) LR, (b) DT, (c) RF, (d) KNN, (e) SVR, (f) MLP, (g) RBF	84
Figure 44	Thermal NO _x comparison of RBF neural networks between actual (red), prediction 13 inputs (blue), and prediction 8 inputs (green).....	86
Figure 45	Prompt NO _x comparison of RBF neural networks between actual (red), prediction 13 inputs (blue), and prediction 8 inputs (green).....	87

Figure 46	RBF test simulation for Thermal NO _x : augmented data (red), prediction 13 input features (blue), and prediction 8 input features (green).....	89
Figure 47	RBF test simulation for Prompt NO _x : augmented data (red), prediction 13 input features (blue), and prediction 8 input features (green).....	90
Figure 48	MLP test simulation for Thermal NO _x : augmented data (red), prediction 13 input features (blue), and prediction 8 input features (green).....	90
Figure 49	MLP test simulation for Prompt NO _x : augmented data (red), prediction 13 input features (blue), and prediction 8 input features (green).....	91
Figure 50	Steam Flow Rate in normal condition (blue), and fault condition (red).....	92
Figure 51	State classification of the Steam Flow Rate: Blue is class 0, Red is class 1	93
Figure 52	Illustration of data transformation into sequential for LSTM model.....	93
Figure 53	Training and Validation loss of LSTM with 10s time steps	94
Figure 54	Confusion Matrix of LSTM with 10s time steps	94
Figure 55	Comparison of actual and predicted data in the time frame boiler operation	95
Figure 56	Comparison of Confusion Matrix of all time steps.....	97
Figure 57	Model performance to the first 20 data.....	98
Figure 58	Model performance to the last 20 data.....	98
Figure 59	Model performance during transition from normal to swings.....	99

ABSTRACT

Full Name : Agus Haeruman

Thesis Title : Prediction of Performance and Emissions, and Fault Prognosis in
Steam Boilers Using Artificial Intelligence

Major Field : Mechanical Engineering

Date of Degree : December 2024

This study focuses on developing soft sensor models using artificial intelligence to predict nitrogen oxides (NO_x) emissions, efficiency, and fault prognosis in steam boilers. Two datasets were employed for training and testing the models: the first was generated through computational fluid dynamics (CFD) simulations of a hydrogen-enriched steam boiler, and the second was collected from an industrial natural gas-fueled boiler. Machine learning (ML) models - including linear regression (LR), decision tree (DT), random forest (RF), k-nearest neighbors (KNN), and support vector regression (SVR), as well as deep learning (DL) models like multi-layer perceptron (MLP) and radial basis function (RBF), were evaluated. Results indicated that the DL-based models outperformed others for the hydrogen-enriched dataset with the prediction accuracy (PA) of 98.08% and 97.50%% for RBF and MLP, respectively, and the speed of RBF is 55 times faster than MLP's. For the natural gas boiler dataset, the tree-based DT and RF models performed best with the coefficient of determination/R-squared (R^2) of 99.32% with 0.728s response time, while RF has an R^2 of 99.68% with 53.632s response time. The efficiency prediction in the

natural gas boiler dataset also shows that RF and DT have the best accuracy with R^2 of 99.98% and 99.97%, respectively. Additionally, fault prognosis was conducted using long short-term memory (LSTM) neural networks to predict boiler swings. The results demonstrated that LSTM models, with the 30s and 60s time steps, had the best accuracy compared to other time steps time steps. This shows that the model can effectively forecast faults based on historical operational data, enabling proactive interventions to prevent major boiler damage.

Keywords: Artificial Intelligence, Soft Sensor, Steam Boiler, Hydrogen, NOx emission, Combustion Efficiency, Fault Prognosis.

الملخص

الاسم الكامل: أغوس هيرومان

عنوان الرسالة: التنبؤ بانبعاثات أكاسيد النيتروجين والكفاءة والتشخيص للأخطاء في غلايات البخار باستخدام الذكاء الاصطناعي

المجال الرئيسي: الهندسة الميكانيكية

تاريخ الحصول على الدرجة: جمادى الثاني 1446 هـ

تركز هذه الدراسة على تطوير نماذج مستشعرات برمجية باستخدام الذكاء الاصطناعي للتنبؤ بانبعاثات أكاسيد النيتروجين (NO_x)، والكفاءة، وتنبؤ الأعطال في الغلايات البخارية. تم استخدام مجموعتي بيانات لتدريب النماذج واختبارها: الأولى تم إنشاؤها من خلال المحاكاة الحاسوبية للديناميكا الحركية للسوائل (CFD) لغلاية بخارية تعمل باستخدام مزيج من الغاز الطبيعي والهيدروجين، والثانية تم جمعها من غلاية صناعية في أرامكو السعودية تعمل بالغاز الطبيعي. للتنبؤ بانبعاثات أكاسيد النيتروجين والكفاءة، تم استخدام عدد من نماذج التعلم الآلي (ML) مثل الانحدار الخطي (LR)، شجرة القرار (DT)، الغابة العشوائية (RF)، أقرب جار لمعامل K (KNN)، والانحدار باستخدام آلة المتجهات الداعمة (SVR)، بالإضافة إلى استخدام نماذج التعلم العميق (DL) مثل الإدراك متعدد الطبقات (MLP) ودالة الأساس الشعاعي (RBF). وقد أظهرت النتائج تفوق نماذج التعلم العميق على غيرها لمجموعة بيانات الغلاية التي تعمل باستخدام الهيدروجين، حيث بلغت دقة التنبؤ 98.08% و 97.50% لكل من RBF و MLP على التوالي، مع تفوق سرعة تنبؤ RBF بمقدار 55 مرة مقارنة بـ MLP. أما بالنسبة لمجموعة بيانات الغلاية التي تعمل بالغاز الطبيعي، فقد حققت نماذج DT و RF أفضل أداء، حيث بلغ معامل R^2 لـ DT نسبة 99.32% بزم من استجابة 0.728 ثانية، بينما حققت RF معامل R^2 99.68% بزم من استجابة 53.632 ثانية. كما أظهرت نتائج التنبؤ بالكفاءة أن نماذج RF و DT حققتا أعلى دقة، حيث بلغت قيمة R^2 لكل منهما 99.98% و 99.97% على التوالي. بالإضافة إلى ذلك، تم إجراء تنبؤ بالأعطال باستخدام الشبكات العصبية طويلة المدى (LSTM) للتنبؤ بالتغير المفاجئ في عمل الغلايات بعيداً عن حالة عملها الطبيعية. وأظهرت النتائج أن نماذج LSTM مع خطوات زمنية قدرها 30 و 60 ثانية حققت

أفضل دقة مقارنة بالخطوات الزمنية الأخرى، مما يثبت قدرة النموذج على التنبؤ بالأعطال بفعالية اعتمادًا على البيانات التشغيلية ، مما يتيح التدخل الاستباقي لتجنب حدوث أضرار كبيرة في الغلاية.

الكلمات المفتاحية: الذكاء الاصطناعي، المستشعرات البرمجية، الغلايات البخارية، الهيدروجين، انبعاثات أكاسيد النيتروجين، كفاءة الاحتراق، تنبؤ الأعطال.

CHAPTER 1

INTRODUCTION

1.1 Background

The availability of energy is a major challenge humankind faces nowadays due to the growth of the population which has increased the energy demand. At the same time, the use of fossil fuels negatively impacts the environment by releasing carbon emissions mainly during combustion. This situation is amplified because, in the last decade, most countries still use fossil fuels as an energy source. For example, Our World in Data reported, as depicted in **Figure 1**, that cumulative world energy consumption in 2010 from coal, oil, and natural gas was 79.41% compared to clean energy sources from hydropower, biomass, nuclear, wind, solar, and other green energy sources, which combined account for 20.59% of the global energy consumption. While the consumption of fossil-based fuels decreased slightly to 76.77% in 2020 (with only a 2.64% reduction), the overall energy consumption rose from 152,966 terawatt-hours in 2010 to 167,780 terawatt-hours in 2020 (about 10% increase), highlighting the slower pace of transition to cleaner energy sources, compared to the increase in energy consumption.

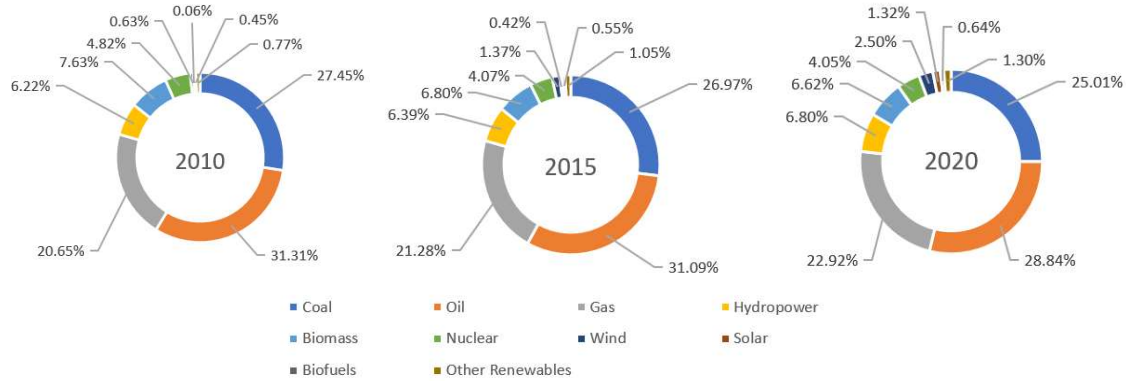


Figure 1 Comparison of energy use in the last decade [1].

To address these challenges and ensure sufficient energy supply, massive investments in cleaner energy sources, including hydrogen (H_2) as a carbon-free fuel, have emerged. Hydrogen has unique environmental and socio-economic characteristics and allows the possibility of escaping dependency on fossil fuels in the future. The energy obtained from hydrogen (i.e. 122 MJ/kg) is 2.75 times greater than that of fossil fuels [2]. According to the International Energy Agency (IEA), the global demand for hydrogen reached 94 million metric tons (Mt) in 2021, which is about 2.5% of global energy consumption and it is predicted to increase to 115 Mt in 2030 [3]. However, the use of hydrogen cannot directly replace the existing fossil-based fuels such as natural gas or methane. Hydrogen faces challenges in large-scale adoption, including production cost, storage, and distribution infrastructure. Further studies are needed to assist in the transition to a hydrogen-based economy using innovative techniques. One such strategy is hydrogen-natural gas blending, which is one of the promising techniques for transitioning to a neutral or low-carbon fuel energy system. Blending hydrogen with natural gas not only reduces carbon emissions but also leverages existing infrastructure, making it a cost-effective bridge to a hydrogen-based economy.

Recently, most combustion systems such as boilers and gas turbines operate on fossil fuels like coal and natural gas. In an attempt at decarbonization, blending hydrogen with fossil fuels – especially natural gas – holds promise for reducing carbon emissions. Some researchers have explored the effect of hydrogen enrichment on boiler efficiency and emissions. For example, Lamioni et al. [4] employed numerical simulations to analyze emissions from mixed hydrogen-natural gas in domestic condensing boilers, with hydrogen content varying from 0% to 50% by volume. Similarly, Celtek et al. [5] and Buyukakin et al. [6] employed numerical methods to examine the performance and emissions of an industrial boiler burning natural gas-hydrogen at levels of 25%, 50%, 75%, and 100% by volume. Wang et al. [7] also examined the effect of hydrogen-natural gas mixed fuel on thermal efficiency and emissions in a 108-ton-per-hour (t/h) steam boiler model, with hydrogen content ranging from 0% to 100%. These studies reveal that data on hydrogen blends up to 100% is mostly available from laboratory-scale studies or numerical simulations, which highlights data limitations regarding hydrogen utilization at the industrial scale. This is because using hydrogen as a fuel requires appropriate infrastructure and safety measures. Erdener et al. [8] discuss current regulations on hydrogen blending in several countries, where hydrogen content in existing natural gas pipelines is typically limited to 10% by volume.

Although blending hydrogen with natural gas can reduce carbon emissions, a trade-off with increasing nitrogen oxide (NO_x) is also present. The higher flame temperatures of hydrogen tend to increase NO_x formation, thus requiring effective control and monitoring to mitigate environmental impacts. NO_x are hazardous pollutants produced during fuel combustion. They contribute to several environmental problems, including ozone

depletion, acid rain, and smog formation [9]. Boilers, as steam generator equipment, widely used in industries for heating and electricity generation, are among the significant contributors to NO_x emissions. For instance, the United Nations Industrial Development Organization (UNIDO) reported in 2014 that industrial boilers accounted for about 30% of energy use in the global manufacturing industry [10]. Thus, monitoring NO_x emissions from boilers becomes crucial to reduce the negative impact on the environment. Reducing NO_x emissions can be achieved through primary and secondary control measures. Primary measures focus on controlling the combustion process to limit NO_x formation, such as optimizing fuel-air mixing, reducing excess air, and lowering flame temperatures. These approaches are often more cost-effective, as they focus only on controlling the input variables of the combustion process, avoiding the need to develop expensive after-treatment devices to prohibit NO_x emission. Secondary measures, on the other hand, involve post-combustion technologies, such as selective catalytic reduction to remove NO_x from exhaust gases. These methods typically require higher operational costs and capital infrastructure [11].

1.2 Problem Statement

Monitoring NO_x emissions to comply with local regulations is one of the main concerns for the industry. Meanwhile, maintaining efficiency is pivotal for productivity and environmental sustainability. Since increasing the efficiency results in the increase of NO_x emissions, and vice-versa, obtaining the best trade-off between these two parameters must be supported by accurate monitoring systems. The existing traditional analyzers such as the continuous emission monitoring system (CEMS), work based on an analogous hardware sensor, which has limitations due to capital investment and high operational and

maintenance costs. The lack of capability for real-time monitoring, due to transmission and data processing (i.e. sampling, analyzing, and data recording), is also a challenge faced by CEMS. Another alternative that has been introduced by many researchers is to employ a data-driven approach using artificial intelligence (AI) models (i.e. soft sensors), that are more cost-effective and reliable for real-time monitoring. Despite many studies related to soft sensors for emissions and efficiency prediction, most of them utilized fossil-based fuels as the subject of their research, while the study on soft sensors incorporating clean energy fuel, such as hydrogen, is still limited. This study focuses on employing AI-based soft sensors to predict NO_x and efficiency from an industrial boiler fueled with natural gas and H_2 -enriched natural gas fuels. A fault prognosis to detect the anomalies in boiler operation based on operational data will also be proposed as an early warning system to avoid major damage to the boilers. The data for this study is obtained from both an actual steam boiler and a computational fluid dynamics (CFD) model. The actual datasets contained historical time steps for boiler operations in the industrial boiler fueled by natural gas, while the CFD data provided information from boilers fueled by hydrogen-enriched natural gas. These data were then used to build an AI-based soft sensor for prediction and fault prognosis models using ML and DL methods.

1.3 Objectives

The main purpose of this work is to develop and validate AI-based data-driven soft sensor models based on actual and CFD boilers datasets for predicting NO_x emission, efficiency, and fault prognosis using ML and DL approaches. The specific objectives of this work are:

1. To develop AI-based soft sensor models to predict the NO_x and efficiency in an industrial steam boiler fueled by natural gas. Feature engineering to obtain NO_x and efficiency variables based on available formulas in the literature is employed, and various ML and DL models are investigated.
2. To develop AI-based soft sensor models for NO_x emission prediction of a hydrogen-enriched natural gas steam boiler. Data from a CFD model is used for this objective, and various ML and DL model performances were compared to obtain the most appropriate model for the boiler.
3. To build an AI-based fault prognosis model to predict anomalies in boiler operation based on the historical input data. The actual datasets from industrial boilers containing historical boiler parameters are employed, and the various time steps were evaluated using the deep learning method to obtain the best parameter to predict the fault/swing of the boiler.

CHAPTER 2

LITERATURE REVIEW

2.1 Boiler Systems

The boiler is a mechanical device that produces high-temperature steam by converting chemicals in the fuel to heat energy through the combustion process. The steam output can then be utilized for various applications. In a power generation plant, for example, high-pressure steam is utilized to rotate a mechanical turbine, which is connected to an electric generator to produce electricity [12]. In process industries, the steam from the boiler is used for many applications, such as a source of heat for distillation columns in the oil and gas, cleaning and sterilization in food industries, and many others [13].

Boilers can be categorized into two main types based on their configuration: fire-tube boilers and water-tube boilers. Fire-tube boilers are a type of boiler where hot gases from the combustion process pass through tubes that are surrounded by water. The heat from the gases is transferred to the water, which is then converted into steam. These boilers are known for their simple design, ease of operation, and low initial cost, making them popular in smaller-scale industries with lower pressure requirements. However, they are limited in terms of steam capacity and pressure capabilities when compared to water-tube boilers. A typical diagram of a fire-tube boiler is provided in **Figure 2**.

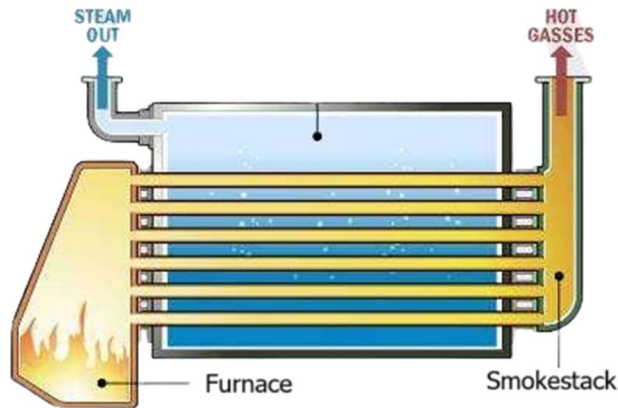


Figure 2 Schematic of a fire-tube boiler [14].

Water-tube boilers, on the other hand, operate by allowing water to flow through tubes that are externally heated by combustion gases. This design enables water-tube boilers to handle higher pressures and temperatures, making them suitable for large-scale industrial processes and power generation. Water-tube boilers are commonly used in chemical industries, power plants, and refineries. The advantages of water-tube boilers include high efficiency, faster steam generation, and the ability to operate under extreme conditions. A model of a typical water-tube boiler is provided in **Figure 3**. The feedwater is initially fed to the boiler system and circulated through the downcomer area. Heat is added to the system by burning fuel through the combustion chamber, and the resulting heat is used to vaporize water and naturally circulate it to the steam drum.

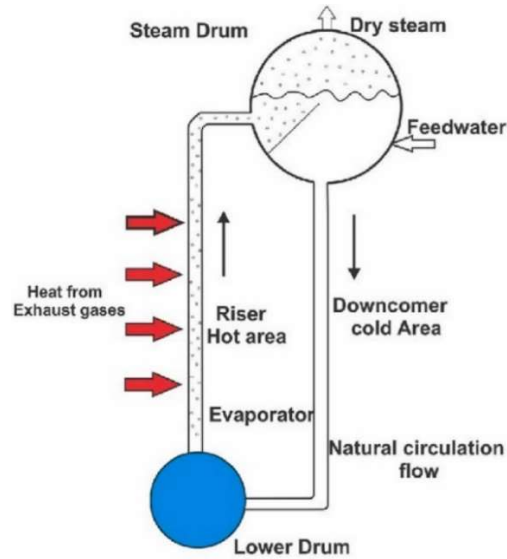
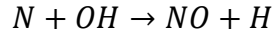
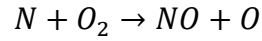
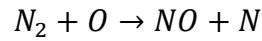


Figure 3 Schematic of a water-tube boiler [15].

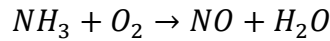
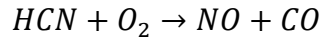
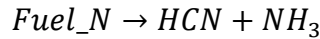
2.2 NO_x Emissions

NO_x is a collective term for nitrogen oxides, a combination of nitrogen monoxide (NO) and nitrogen dioxide (NO₂). NO_x is commonly produced during fuel combustion due to the reaction of nitrogen from the fuel with the oxygen from the air at a high temperature. There are three types of NO_x based on its source of formation: Thermal NO_x, Prompt NO_x, and Fuel NO_x. Thermal NO_x is generated due to the direct oxidation of nitrogen in the combustion air in regions where the highest local flame temperature occurs. Prompt NO_x is generated when molecular nitrogen in the air reacts with the fuel at the flame front. Fuel NO_x on the other hand is produced by the oxidation of nitrogen contained within the fuel [16]. A detailed explanation and key reaction of these three types of NO_x formation is provided in this section, which refers to Palash et al. [17].

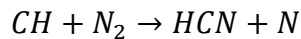
Thermal NO_x arises from the high-temperature reaction between atmospheric nitrogen and oxygen. Therefore, its formation is particularly significant in high-temperature systems, such as industrial boilers and gas turbines. Thermal NO_x is extensively described by the Zeldovich mechanism and involves a series of reactions where nitrogen molecules react with oxygen atoms to produce nitric oxide and nitrogen atoms, which further react with oxygen to form additional NO. The key reactions include:



Fuel NO_x is generated from nitrogen compounds that are present in the fuel source, such as coal or biomass. During combustion, these nitrogenous compounds are released and react with oxygen to form NO_x. The complexity of fuel NO_x formation is influenced by factors including fuel composition, combustion temperature, and oxygen availability. In low-nitrogen fuels, this mechanism contributes less to overall NO_x emissions. The process typically involves the following steps:



Prompt NO_x on the other hand, is formed in fuel-rich combustion zones where hydrocarbon radicals (e.g., CH and CH₂) react with atmospheric nitrogen. The formation of prompt NO_x is rapid and occurs via the following key reaction:



Subsequent reactions oxidize intermediate molecules like HCN to form NO. Although prompt NO_x typically constitutes a smaller portion of total NO_x emissions compared to thermal and fuel NO_x, it becomes significant in certain low-temperature or poorly mixed combustion processes.

To calculate NO_x emission in the boiler, Rehan et al. [9] introduced Li and Thompson formula by using three basic inputs: Fuel Flow Rate (FFR), actual Air Flow Rate (AFR_{act}), and burner tilt angle (ξ). The steps to calculate NO_x emission from oxygen concentration are explained through Equations (1) to (4) as follows:

Step 1: Calculate Excess Air using the following formula:

$$Excess\ Air\ (EA) = \frac{20.9}{20.9 - O_2} - 1 \quad (1)$$

Step 2: Calculate lambda from the excess air obtained in step 1.

$$lambda = 1 + EA \quad (2)$$

Step 3: Calculate actual AFR from lambda and AFR stoichiometric (AFR_{st}=16.71)

$$AFR_{act} = AFR_{st} * lambda \quad (3)$$

Step 4: Calculate NO_x by Li and Thompson formula using the above parameters.

$$\frac{d(NO)}{dt} = \alpha^0 \dot{m}_f^r \left(1 + \alpha^1 \frac{\xi - 55}{90} \right) \left(\frac{1}{AFR_{st}} - \frac{1}{AFR_{act}} \right)^{1/2} - (NO) \quad (4)$$

where, α^0 , α^1 , and r are empirical constants ($\alpha^0=1806$, $\alpha^1 = 0.438$, and $r = 0.25$), $\dot{m}_f$ is fuel mass flow rate, and ξ is burner tilt angle ($\xi=75$)

2.3 Boiler Efficiency

Efficiency is a measure of useful output energy compared to the total input energy. Boiler efficiency is one of the most important parameters to be considered in the industry because the highest operational cost mainly comes from fuel consumption. There are three types of efficiency from the boiler based on the scope of input and output parameters: combustion efficiency, thermal efficiency, and overall efficiency.

Combustion efficiency is a measure of how effectively a combustion system converts fuel energy into usable heat. This efficiency depends largely on proper air-to-fuel ratios, fuel quality, and the completeness of combustion. An incomplete combustion process leads to energy losses in the form of unburned fuel or the formation of combustible gases such as carbon monoxide. The flue gas temperature and oxygen concentration (or excess air) are primary indicators for combustion efficiency. For example, when the natural gas boiler operates with 15% excess air and with a small amount of unburnt fuel, less heat from the combustion is used for heating the excess air, thus maximizing the heat energy for the boiler load. The formula for combustion efficiency is provided in Equation (5),

$$\text{Combustion Efficiency} = \left(1 - \frac{\text{Heat Loss in Flue Gas}}{\text{Energy Input from Fuel}}\right) \times 100\% \quad (5)$$

where Heat Loss in Flue Gas is typically calculated using the temperature of the flue gases and the excess air levels, and Energy Input from Fuel is determined by the fuel's calorific value and the fuel consumption rate [18].

One method of calculating combustion efficiency is using the Siegert Formula [19]. The Siegert Formula is widely used in Europe to calculate boiler efficiency and utilizes

flue gas losses (qA) in its calculation. The step of calculating combustion efficiency using the Siegert Formula is started by calculating qA as presented in Equation (6),

$$qA = (Ts - Ta) \left(\frac{A2}{20.9 - O_2} + B \right) \quad (6)$$

and the efficiency was obtained by subtracting flue gas loss (qA) from 100% efficiency by Equation (7).

$$\text{Efficiency} = 1 - qA \quad (7)$$

where Ts is flue gas temperature, Ta is ambient temperature, $A2$ and B are Siegert Constant as described in **Table 1**.

Table 1. Siegert Constants [19].

Fuel Type	A2	B
Natural Gas	0.66	0.009
Fuel Oil	0.68	0.007
Town Gas	0.63	0.011
Coking Oven Gas	0.60	0.011
LPG (Propane)	0.63	0.008

Thermal efficiency, on the other hand, evaluates the overall heat conversion efficiency of the boiler, accounting for heat losses in the flue gas, radiation losses, and convection losses. It represents the ratio of the useful heat output absorbed by the water/steam to the total energy input from the fuel. This measure provides a more comprehensive assessment of the boiler's operational efficiency. The formula to calculate Thermal Efficiency is provided in Equation (8).

$$\text{Thermal Efficiency} = \frac{\text{Heat Absorbed by the Water/Steam}}{\text{Energy Input from Fuel}} \times 100\% \quad (8)$$

where Heat Absorbed by the Water/Steam can be calculated using the mass flow rate of the fluid, its specific heat capacity, and the temperature rise across the boiler [20].

Overall efficiency, on the other hand, is the most comprehensive measure of a boiler's performance, as it takes into account all types of energy losses and energy input in the system. Unlike thermal or combustion efficiency, overall efficiency considers all potential energy losses, including heat lost in the flue gas, radiation, convection, unburned fuel, and auxiliary power consumption. The formula of overall efficiency is depicted by Equation (9).

$$\text{Overall Efficiency} = \frac{\text{Useful Heat Output}}{\text{Total Energy Input}} \times 100\% \quad (9)$$

where Useful Heat Output is the heat energy transferred to the working fluid (water or steam), and Total Energy Input is the energy supplied by fuel plus heat credit (i.e. preheated combustion air, waste heat recovery, etc.) [21].

2.4 Soft Sensors

Soft sensors, also known as inferential sensors, play essential roles in modern industries for process monitoring, control, and optimization, particularly when direct measurement using hardware sensors is difficult due to harsh environments or high capital costs. A number of studies have been performed implementing intelligent systems for developing soft sensors in various industries, which are highlighted in this section.

2.4.1 Soft Sensor in Power Plant

In power plant applications, Tronci et al. [22] investigated soft sensors based on neural networks for emission monitoring in a 4.8 MW power plant. Four different types of fuels are used including pulverized coal, natural gas, oil, and orimulsion, and developed a soft sensor that successfully inferred three emission concentrations, including CO, NO_x, and O₂ based on various operating conditions and fuel switching. Liukkonen et al. [23] developed adaptive soft sensors to evaluate the quality of the bed by estimating the distribution of grain size of the circulating fluidized bed (CFB). The preliminary feature selection was performed using sequential forward selection (SFS), and then the dynamic model was built using multiple linear regression (MLR) and multi-layer perceptron (MLP) from three different datasets of CFB boilers. Furthermore, Sujatha et al [24] developed a Soft Sensor for monitoring flame temperature using flame images based on back propagation neural networks (BPNN) and ant colony optimization (ACO). The works integrated soft sensors with the embedded computing system and Internet of Things (IoT) making it possible for online monitoring and combustion control to improve power plant productivity. He et al. [25] demonstrated a soft sensor to predict the deformation of the rotor from boiler pre-heater in a thermal power plant. Coefficient variation and maximum mutual information coefficient (CV-MIC) are used for variable selection, and the stacked ensemble deep neural network (DNN) model combines three base learners using long short-term memory - gaussian process regression (LSTM-GPR), stacked denoising autoencoder & support vector regression (SDAE-SVR), and XGBoost. The experiment shows that the ensemble model outperformed the three single models. Li et al. [26] experimented with three types of soft sensor models, including static, segmentation, and

dynamic type soft sensors based on ordinary least square (OLS), SVR, XGBoost, random forest (RF), and neural networks (NN), to predict sulfur dioxide (SO₂) emission from the ultra-low emission 1000 megawatt (MW) coal-fired power plant. The study reported that the static model was not suitable due to the complexity of the process and the time delay present so segmentation and dynamic approach were used to improve performance and generalization ability. It showed that the dynamic NN-based soft sensor has the best performance compared to others. Wang et al. [27] used the same industrial boiler [26] to develop a soft sensor on the view of a typical non-stationary characteristic condition process in the boiler. Thus, condition-driven category boosting (CatBoost) was introduced in this paper to predict SO₂ content from the intake of flue gas desulfurization (FGD). Load-based sequential condition division (LSCD) and variational auto encoding (VAE) were used to divide the boiler process into load condition slices (LCS) based on their characteristics. The LCSs with identical structures are combined into five load condition modes (LCMs). The CatBoost then is used to develop a soft sensor for each LCM to predict SO₂ with different operating conditions.

2.4.2 Soft Sensor in Industrial Processes

In industrial process applications, a number of studies have been conducted to develop soft sensors for monitoring and controlling important variables. For example, Gonzaga et al [28] presented a soft sensor to measure the viscosity of polyethylene terephthalate (PET) in the industrial polymerization process using feed-forward neural networks (FFNN). The developed soft sensor model then is used for the PET supervisory control system to maintain optimal plant operation. De Lima et al. [29] presented a novel ensemble deep relevant learning soft sensor (EDRLSS) to predict butane concentration in

the industrial debutanizer column and SO₂ concentration in the Sulfur Recovery Unit (SRU) of the refining process. The authors used mutual information (MI)-based stacked autoencoder (SAE) to obtain relevant variables and a bagging strategy to construct a soft sensor model using two layers of neural networks. It is shown that the EDRLSS model performed best accuracy compared to other linear and non-linear models such as partial least squares (PLS), SVM, MLP, SAE, convolutional neural networks (CNN), and non-ensemble DRLSS. Rozanec et al. [30] developed an ML-based soft sensor to predict real-time energy consumption in a distillation crude oil unit using three different models including linear regression (LR), MLP, and Gradient Boosting Decision Tree Regression. It is found that the model demonstrated good prediction results even though they used a limited number of input features such as temperature, pressure, and volumetric flow of the system. Zhai et al. [31] developed a soft sensor based on the multisource transfer learning - knowledge distillation (MTL-KD) method based on temporal convolutional networks (TCN) to predict Billet temperature in different heating furnaces. The combination of Cloud Computing and Edge Computing process using the Teacher-Student architecture allows the model to extract important features from the cloud-based information (as a teacher) and transfer to edge-based computation (as a student) with more light computational cost. Sifakis et al. [32] investigated the soft sensor model in the Liquefied Petroleum Gas (LPG) refinery process using artificial neural networks (ANN) and ensemble random forest regressor (RFR). Real data was obtained from the LPG refinery process containing 500000 data points from 2 years of observation and 60s time steps. The dataset contains six input parameters and three outputs including C₂, C₅ concentration, and energy consumption.

2.4.3 Soft Sensor for Emission and Efficiency Prediction

Besides the above studies, several works have shown the application of soft sensors for monitoring and control processes, especially for emission and efficiency control in the industrial sector. For example, Shakil et al. [33] developed an inferential sensor for monitoring NO_x and O_2 in industrial boilers as an alternative to the traditional sensor, using two dynamic neural networks, namely focused-time lagged feed-forward neural network (FTLNN) and input-output recurrent neural network (IORNN). The authors used principal component analysis (PCA) to reduce input dimension and genetic algorithm (GA) to optimize time delays. The result shows that dynamic neural networks outperform static neural networks based on MLP. The improvement of this work was then conducted by Iliyas et al. [34] by using radial basis function neural networks (RBFNN), making the model applicable for online monitoring. The model was validated using real boiler data and showed high performance and accuracy compared to the traditional CEMS with significant cost reduction in equipment and maintenance. Lin et al. [35] presented a systematic approach using robust statistical techniques in developing soft sensors for free lime and NO_x emission monitoring in cement kilns. To improve data quality, several pre-processing methods were conducted, including applying univariate and multivariate PCA for detecting outliers, before developing a regression model based on dynamic partial least squares (DPLS).

Furthermore, some studies have also been performed on utilizing AI for control optimization and efficiency. For example, Elshafei et al. [36] developed a boiler dynamic control for maximum load change rate using GA-based optimization. The model could effectively address the problem of sudden and large steam load changes that usually happen

in industrial processes and, at the same time, maintain operational stability. This could prevent the boiler from tripping and shutting down during the process. Alzaharnah et al. [37] also successfully developed boiler control to minimize NO emissions by addressing the swing rate problem. The authors showed the effect of boiler swing and firing rate on the NO production. The result shows that a lower swing rate results in decayed NO formation, while a higher swing rate increases the total NO emission.

2.5 Fault Prognosis

Fault prognosis is the process of using past observation to predict the fault in the upcoming state, so that appropriate actions can be taken before the fault occurrence [38]. Many researchers have developed data-driven models for fault prognosis in boiler systems. For example, Salman et al. [39] proposed an intelligent system to detect leakage in the boiler water tube using ML-Based Optimal Sensor Selection. Correlation analysis was employed to identify the most relevant sensors and eliminate redundancy. The selected sensor data was then used to develop four models, namely support vector machine (SVM), k-nearest neighbors (KNN), naive bayes (NB), and linear discriminant analysis (LDA). Rostek et al. [40] proposed an ANN-based model for the early detection of leakages in fluidized-bed boilers. Two-stage ANN model for diagnosing and predicting faults is performed. The first stage involves early fault detection via virtual sensors, while the second stage uses classifiers to isolate leaks. The model demonstrated detection of 11 out of 12 faults, with at least two days' advance notice before shutdowns. This approach outperforms previous works using the neuro-fuzzy technique. Qin et al. [41] investigated advanced data-driven techniques for fault prediction in thermal boiler systems. It evaluated the effectiveness of three machine learning models - RF, least absolute shrinkage and

selection operator (Lasso) regression, and SVR for predicting key performance indicators (KPIs) related to faults. After pre-processing and selecting features using RF and support vector machine-recursive feature elimination (SVM-RFE), the study employed a stacking algorithm to integrate these models. The proposed stacked model significantly improves fault prediction performance compared to the original prediction approach. Li et al. [42] presented a few-shot fault diagnosis architecture for supercharged boilers using Siamese Neural Networks (SNN). The model incorporates feature engineering to enhance model performance. Results show that the Five-shot method outperforms traditional models such as KNN, MLP, RF, SVM, and CNN by up to 45.17% in accuracy with only 30 training samples. With 60 samples, the Five-shot method maintains a competitive edge over CNN, achieving up to 85% accuracy with 75 samples. This framework is effective for diagnosing faults in supercharged boilers with limited data.

2.6 Research Gaps

Despite a lot of research that has been performed in developing soft sensors and fault prognosis as mentioned above, the following gaps remain exist and need more exploration for further improvement:

- Most of the above studies focus on specific target variables, such as emissions [22], bed quality [23], flame temperature [24], or rotor deformation [25]. However, the study on the soft sensor combining the prediction of boiler efficiency and emissions is still limited. Combining both variables will enable real-time adjustments in combustion control.
- Most of the above studies incorporate traditional fossil-based fuels as their subject, such as coal [26], oil [30], and natural gas [33]. However, the research on soft sensors

specialized for monitoring emissions using more clean energy sources, such as hydrogen-enriched natural gas is still limited.

- While the study on soft sensor models based on data-driven is massively available, such as [25], [26], [27], the gap still exists in combining statistics and domain knowledge approaches, which is crucial for complex systems such as boilers. Physical understanding provides domain insights in selecting the most appropriate features for the models and thus leads to soft sensors that are more reliable, robust, and tailored to specific industrial applications.
- While advanced models like deep neural networks [25], ensemble methods [29], and multisource transfer learning [31], provide high accuracy, their model complexity causes high computational costs, and limited attention has been given to the simple model in evaluating computational efficiency without compromising accuracy. Evaluating simulation time has been performed by several studies, including [43], [44], [45] to compare the computational cost of the models.
- The current studies of fault prognosis, largely focus on predicting failures such as leakage [39], overheating [41], and abnormal combustion [42], while the research on boiler swings, a major concern that commonly happens in industrial boilers is still limited. The fault prognosis for the boiler swings, incorporating boiler operational parameters is required to enhance predictive accuracy and ensure early fault prognosis.

CHAPTER 3

DATA COLLECTION

Two datasets have been obtained for this study. The first dataset is from the actual data obtained from an industrial steam boiler at Saudi Aramco fueled with natural gas, and the second one is from a CFD simulation for the same boiler, fueled with hydrogen-enriched natural gas. Both datasets originated from the actual industrial boiler, a 200 MW water tube boiler from Saudi Aramco, manufactured by Mitsubishi Heavy Industries (MHI), featuring two vertically arranged burners: a furnace (radiation section) and a return tube bank (convection section).

The illustration of the boiler is provided in **Figure 4a** and **Figure 4b**, showing the front and top views of the boiler model. The radiation section consists of riser tubes formed by D-tubes, while the convection section comprises a 2-meter-wide tube bank constructed from downcomer tubes, as shown in **Figure 5**. The combustion chamber dimensions are 12.541 meters in flame length, 4.579 meters in front width, and 7.925 meters in length between the drums. The two burners are mounted at heights of 2.198 meters and 4.662 meters on the front wall. Each burner includes eight fuel nozzles, each containing 24 fuel jets, with a total fuel inlet area of 0.03563 m² per burner. Primary air, introduced with a 45-degree swirl, surrounds the fuel nozzles, while non-swirled secondary air envelops the primary air. The inlet areas for primary and secondary air are 0.172 m² and 0.62754 m² per burner, respectively.

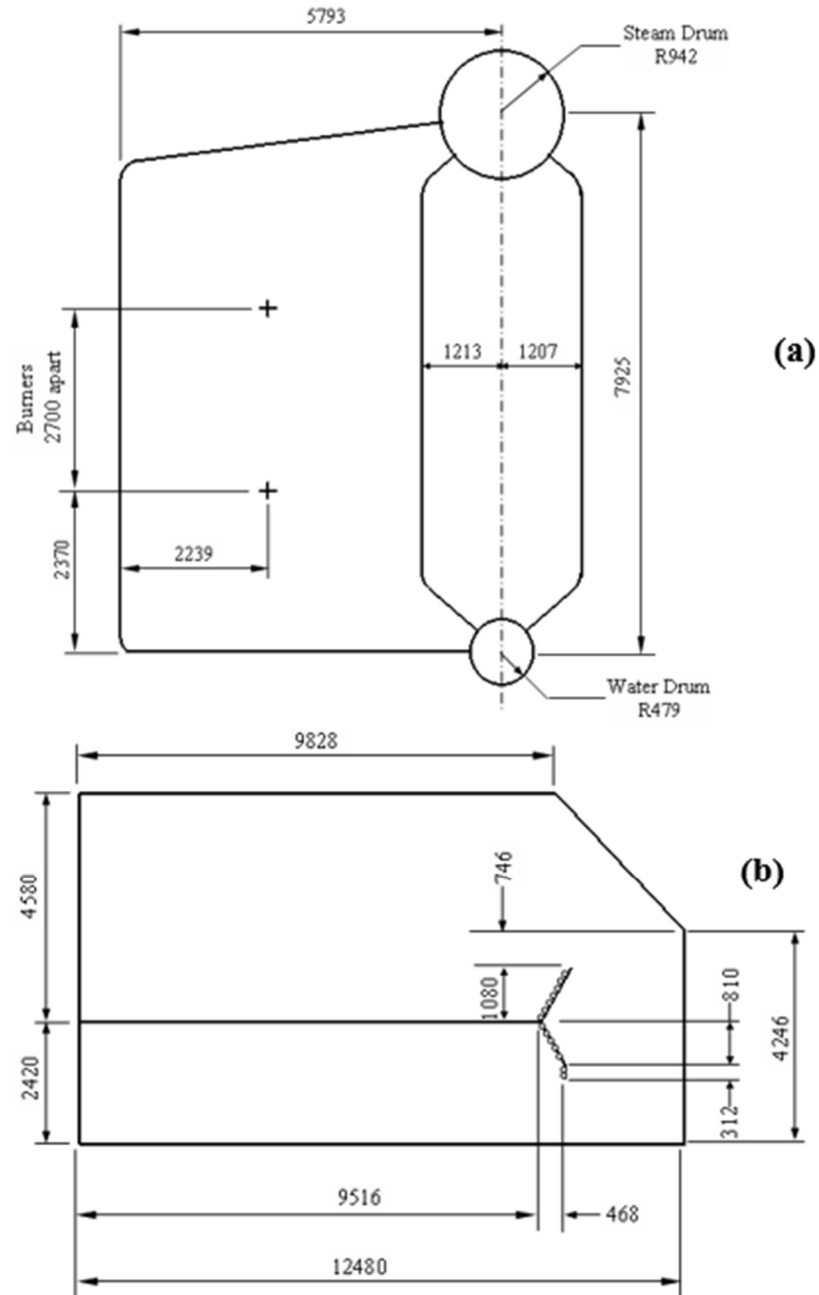


Figure 4. Design of the Boiler Furnace: (a) Front view, (b) Top view [46].

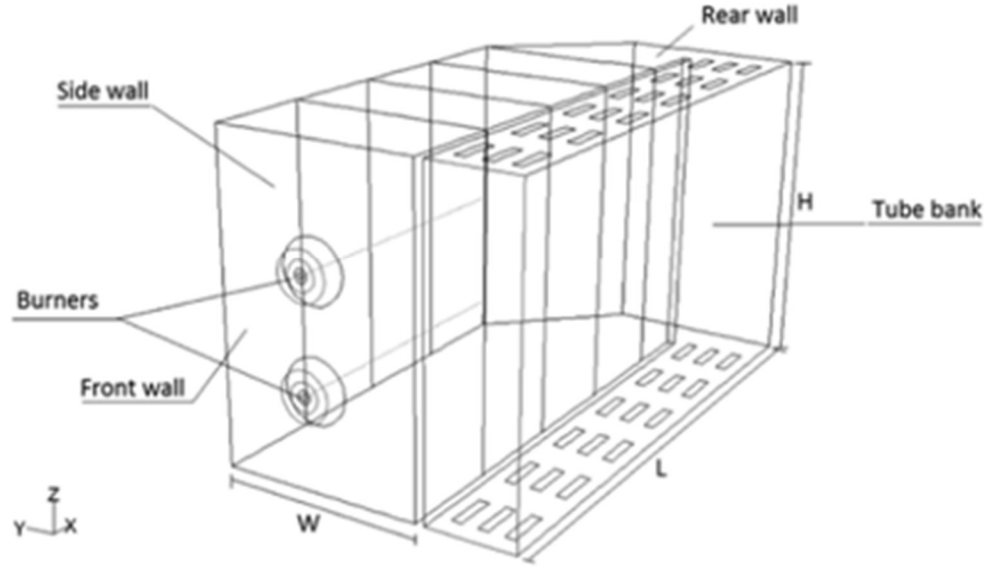


Figure 5. Boiler radiation and convection sections [46].

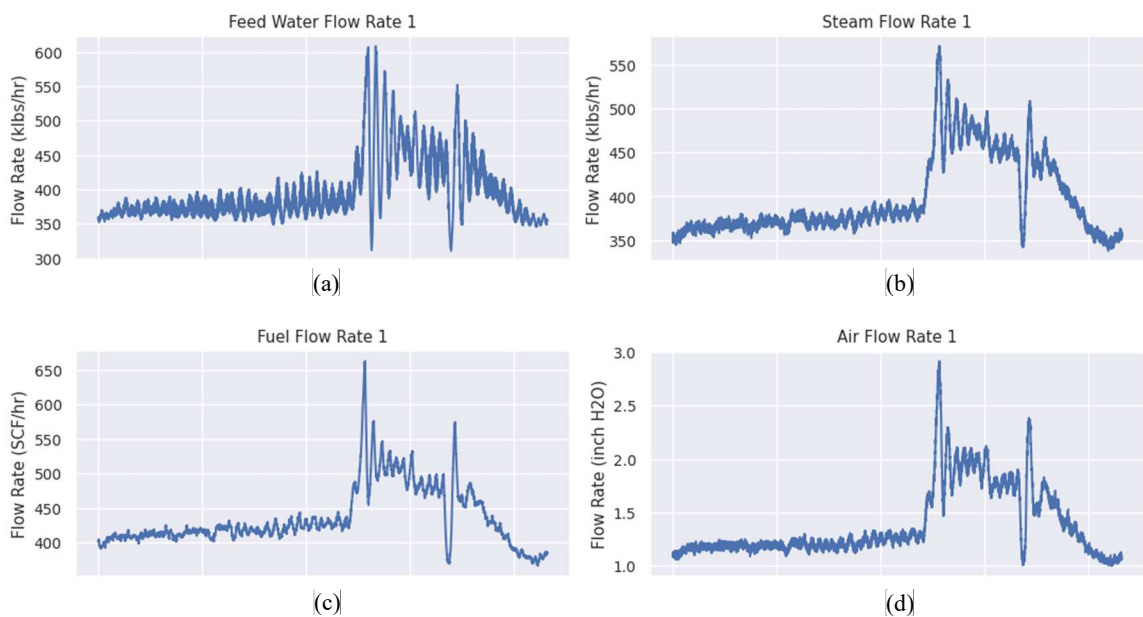
3.1 Dataset from an Industrial Boiler Fueled with Natural Gas

The dataset consists of 21,600 data points with a sampling rate of 1 second. This data contains a variety of real-time sensors that are used to measure boiler operation parameters. There are 21 parameters in total, which are variable no. 1 is Date and Time, while variable no. 2-20 are input parameters. Variable no. 21 is the oxygen concentration of the flue gas that will later on be used to calculate boiler efficiency as one of the target variables, other than NO_x . The description of each parameter can be found in **Table 2**, and the visualization of some boiler parameters, including feed water flow rate, steam flow rate, fuel flow rate, air flow rate, steam drum pressure, and steam drum temperature is provided in **Figure 6**.

Table 2. Variable Description from Natural Gas Boiler Dataset

Feature #	Feature Name	Description	Unit
1	DT	Date and time	Day, Hour, Minutes, Seconds
2	FWFR1	Feed Water Flow Rate 1	klbs/hr

3	FWFR2	Feed Water Flow Rate 2	klbs/hr
4	SFR1	Steam Flow Rate 1	klbs/hr
5	SFR2	Steam Flow Rate 2	klbs/hr
6	FFR	Fuel Flow Rate	kSCF/hr
7	AFR	Air Flow Rate	inch H ₂ O
8	FDP	Forced Draft Discharge Pressure	inch H ₂ O
9	FGP	Fuel Gas Pressure	psig
10	FLGP1	Flue Gas Pressure 1	inch H ₂ O
11	FLGP2	Flue Gas Pressure 2	inch H ₂ O
12	FUP	Furnace Pressure	inch H ₂ O
13	IGP	Ignitor Pressure	psig
14	SDP1	Steam Drum Pressure 1	psig
15	SDP2	Steam Drum Pressure 2	psig
16	DAT	Deaerator Temperature	Fahrenheit (°F)
17	SDT	Steam Drum Temperature	Fahrenheit (°F)
18	SHT1	Superheater Temperature 1	Fahrenheit (°F)
19	SHT2	Superheater Temperature 2	Fahrenheit (°F)
20	FLGT	Flue Gas Temperature	Fahrenheit (°F)
21	OXY	Oxygen Concentration in flue gas	%



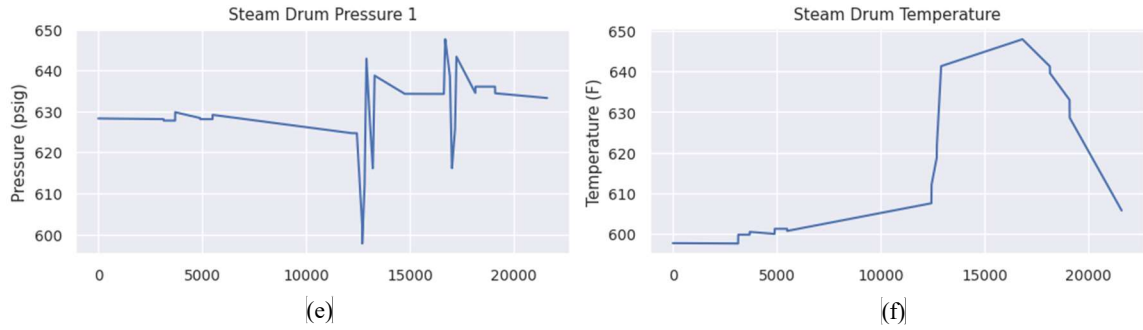


Figure 6. Visualization of boiler parameters from actual boiler: (a) Feed Water Flow Rate, (b) Steam Flow Rate, (c) Fuel Flow Rate, (d) Air Flow Rate, (e) Steam Drum Pressure, (f) Steam Drum Temperature

3.2 Dataset from an Industrial Boiler Fueled with Hydrogen-Enriched Natural Gas

Due to limited data available on the actual boiler fueled with hydrogen-enriched natural gas, the dataset was collected through CFD simulation using the Saudi Aramco boiler explained at the beginning of this chapter, with the change of fuel from natural gas to hydrogen-enriched natural gas [46]. **Figure 7** illustrates a sample of the temperature distribution on a vertical plane passing through the burners. Figure 7a shows the temperature distribution for a mixture of fuel of 50% H_2 and 100% CH_4 (by volume) while Figure 7b shows the distribution for a mixture of 100% H_2 and 0% CH_4 (by volume). **Figures 8a** and **8b** show the corresponding NO_x distribution for the same fuel mixture conditions. These figures show the correlation between temperature and NO_x formation. It is clearly shown that NO_x strongly correlates with temperature. The figures also show that

temperature and NO_x formation area (shown by red, orange, and yellow) are growing by adding hydrogen to the fuel mixture.

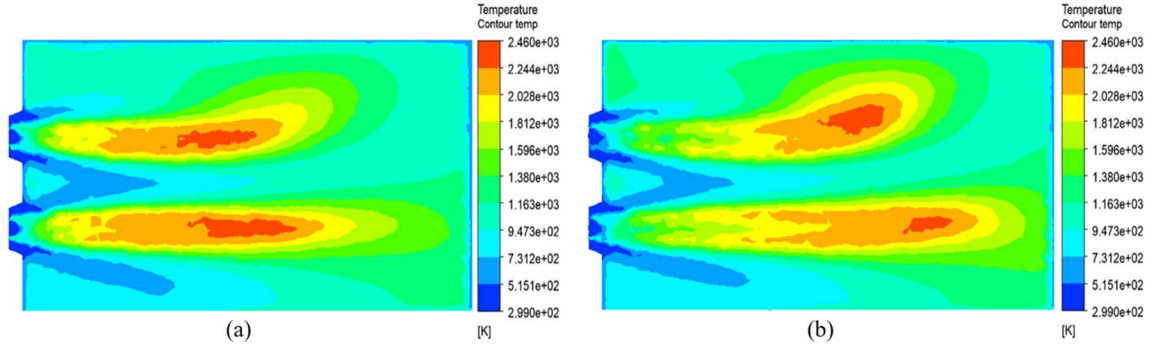


Figure 7. Temperature distribution on a vertical plane at lambda 1.15: (a) Fuel mixture 50% H₂ and 50% CH₄, (b) Fuel mixture 100% H₂ and 0% CH₄.

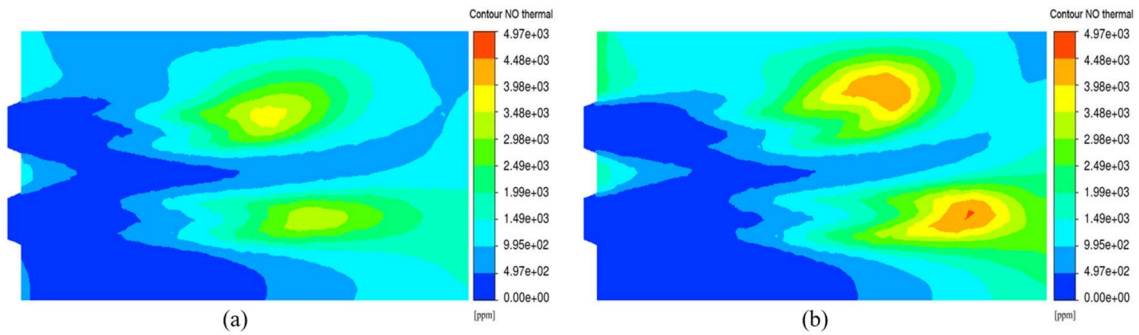


Figure 8. Thermal NO distribution on a vertical plane at lambda 1.15: (a) Fuel mixture 50% H₂ and 50% CH₄, (b) Fuel mixture 100% H₂ and 0% CH₄.

The simulation performed to collect the first dataset consists of two parametric studies that investigate the effect of different input parameters on the output flame temperature and NO_x emission. The first study considers the effect of hydrogen-enriched natural gas mixed fuel by changing the hydrogen fraction in the H₂/CH₄ mixture, while the second considers the effect of excess air by changing the lambda (excess air factor). In the first part, the composition of H₂ in the H₂/CH₄ mixture was changed from 0% up to 100% at a fixed lambda of 1.15. The result of this experiment, 16 different operational conditions

(samples), is depicted in **Table 3**. The correlation of H₂ variation with furnace/exhaust temperature and NO_x formation is also described in **Figure 9**. From this chart, it can be seen that a significant increase in temperatures and NO_x production has happened since the first addition of 1% H₂. The exhaust temperature increased from 962 K (0% H₂) to 993 K (1% H₂), followed by the increase of Thermal NO_x and Prompt NO_x from 37 ppm (0% H₂) to 1326 ppm (1% H₂) and from 2 ppm (0% H₂) to 28 ppm (1% H₂), respectively. The addition of the H₂ fraction in the subsequent ranges is followed by the increase of exhaust temperature from 1003 K (2% H₂) to 1070 K (100% H₂), thus increasing Thermal NO_x and Prompt NO_x production from 1284 ppm (2% H₂) to 1474 ppm (100% H₂) and from 28 ppm (2% H₂) to 90 ppm (100% H₂), respectively.

Table 3. CFD data generated with various hydrogen fractions.

Mole CH ₄	Mole H ₂	Lambda	Furnace Avg Temp (K)	Exhaust Avg Temp (K)	Avg Thermal NO _x (ppm)	Avg Prompt NO _x (ppm)
1.00	0.00	1.15	1033.34	962.64	37	2
0.99	0.01	1.15	1044.16	993.28	1,326	28
0.98	0.02	1.15	1052.59	1002.95	1,284	28
0.97	0.03	1.15	1046.15	998.81	1,220	28
0.96	0.04	1.15	1047.92	1001.56	1,198	28
0.95	0.05	1.15	1047.81	1001.68	1,233	29
0.93	0.07	1.15	1046.72	1001.44	1,197	30
0.90	0.10	1.15	1047.37	999.06	1,142	31
0.87	0.13	1.15	1049.90	1004.59	1,129	33
0.83	0.17	1.15	1046.43	997.55	1,074	34
0.80	0.20	1.15	1063.15	1018.90	1,068	36
0.70	0.30	1.15	1053.39	1003.59	1,136	41
0.50	0.50	1.15	1061.65	1012.10	1,036	50

0.35	0.65	1.15	1070.66	1025.41	1,115	57
0.15	0.85	1.15	1090.82	1043.15	1,231	71
0.00	1.00	1.15	1123.05	1070.83	1,474	90

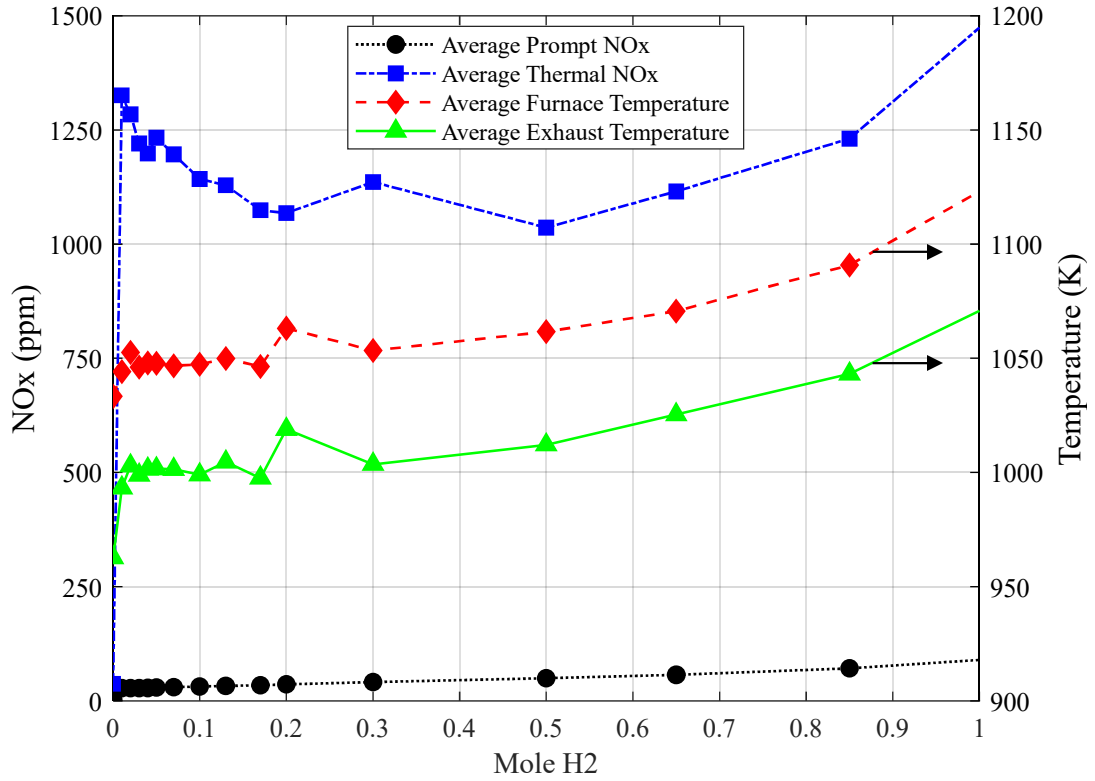


Figure 9. Effect of H₂ Fraction on Temperature and NO_x Formation.

As for the second part, the effect of lambda (excess air factor) on the temperature and NO_x production is investigated, by fixing an H₂ fraction of 10% and natural gas of 90% in the fuel mixture by volume, resulting in various lambda values of 1.05 to 1.45. Nine operating conditions (samples) were obtained from this simulation, as provided in **Table 4**. The correlation of lambda variation with temperature and NO_x formation is also described in **Figure 10**. The chart shows the furnace/exhaust temperature and NO_x production are consistently decreasing by increasing the lambda because the fuel flow rates

are kept constant while the airflow increases, resulting in the Thermal NO_x and Prompt NO_x decreasing from 1483 ppm (lambda=1.05) to 652 ppm (lambda=1.45), and from 58 ppm (lambda=1.05) to 1 ppm (lambda=1.45), respectively.

Table 4. CFD Data Generated with various Lambda.

Mole CH₄	Mole H₂	Lambda	Furnace Avg Temp (K)	Exhaust Avg Temp (K)	Avg Thermal NO_x (ppm)	Avg Prompt NO_x (ppm)
0.90	0.10	1.05	1095.96	1042.44	1,483	58
0.90	0.10	1.10	1068.15	1013.25	1,308	43
0.90	0.10	1.15	1047.37	999.06	1,142	31
0.90	0.10	1.20	1031.22	993.64	1,031	22
0.90	0.10	1.25	1004.53	964.77	947	15
0.90	0.10	1.30	973.72	938.34	804	10
0.90	0.10	1.35	941.84	869.15	745	6
0.90	0.10	1.40	934.58	895.41	709	3
0.90	0.10	1.45	909.27	891.08	652	1

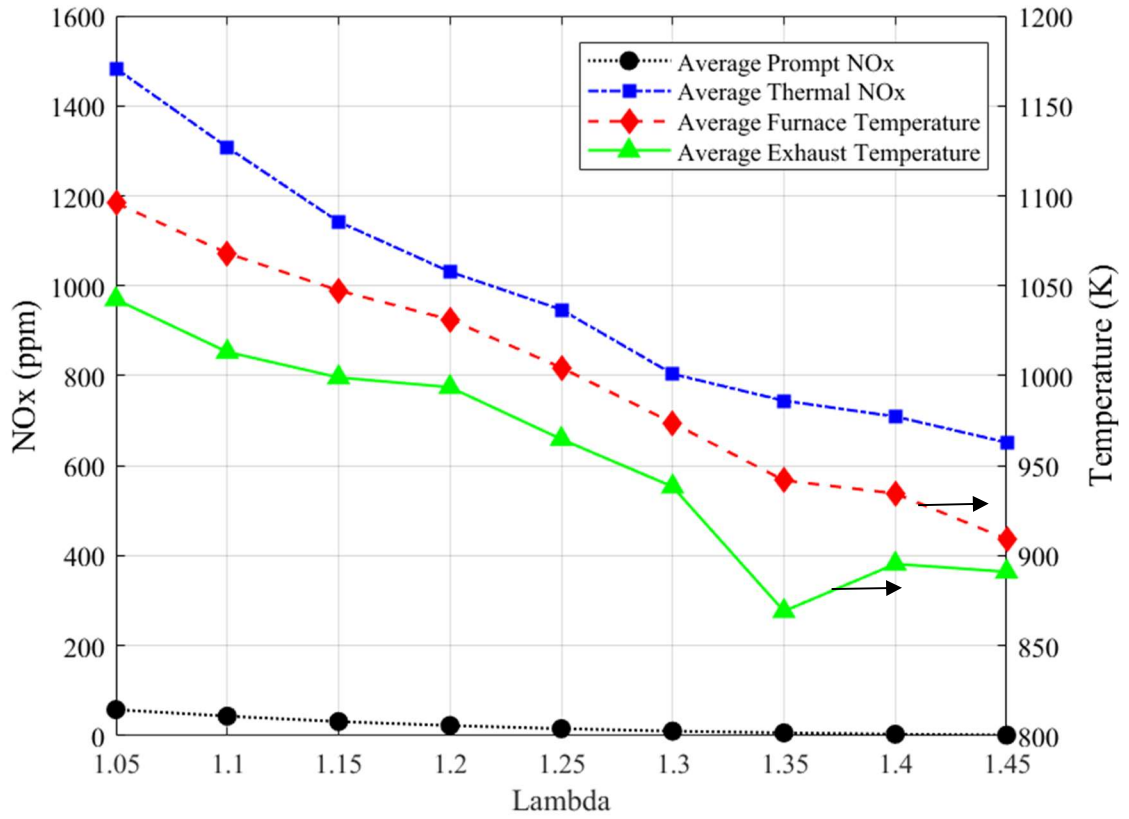


Figure 10. Effect of lambda variations on temperature and NO formation.

When examining the trend in Figure 9, the furnace average temperature exhibits a non-linear behavior, increasing from 1046.43 K (17% H₂) to 1063.15 K (20% H₂), and then decreasing to 1053.39 K (30% H₂). A similar trend is observed in the exhaust average temperature, which rises from 997.55 K (17% H₂) to 1018.90 K (20% H₂), before dropping to 1003.59 K (30% H₂). Additionally, Thermal NO_x also shows an anomaly at a hydrogen fraction of 30%, when there is a sudden increase from 1068 ppm to 1136 ppm, and then drops to 1036 ppm. The anomaly also happens in Figure 10 when the average exhaust temperature drops at a lambda value of 1.35. These anomalies were not explicitly explained by Nemitallah et al. [46]. However, some other studies also reported a similar phenomenon when hydrogen-enriched natural gas was experimented.

For instance, Celtek et al. [5] observed that the exhaust temperature decreases from 600 K (0% H₂) to 585 K (25% H₂), before increasing slightly to 590 K (50% H₂) and 595 K (100% H₂). In contrast, NO_x emissions in their study showed a consistent increase, rising from 50 ppm (0% H₂) to 455 ppm (100% H₂). Similarly, Buyukakin et al. [6] found that the flue gas temperature increases from 300°C (0% H₂) to 325°C (25% H₂) and remains steady at 325°C for higher hydrogen fractions (50% to 75%). Gersen et al. [47] identified a different anomaly in NO_x emissions, where values dropped from 80 mg/m³ (60% H₂) to 78 mg/m³ (62% H₂).

According to the literature, this variability is attributed to several factors, including turbulent mixing, combustion dynamics, and the interaction between flame stretch and chemical kinetics. These factors can lead to unexpected patterns in temperature and NO_x emissions under varying conditions. Additionally, the limitations of current modeling approaches further highlight the challenges in accurately capturing these phenomena.

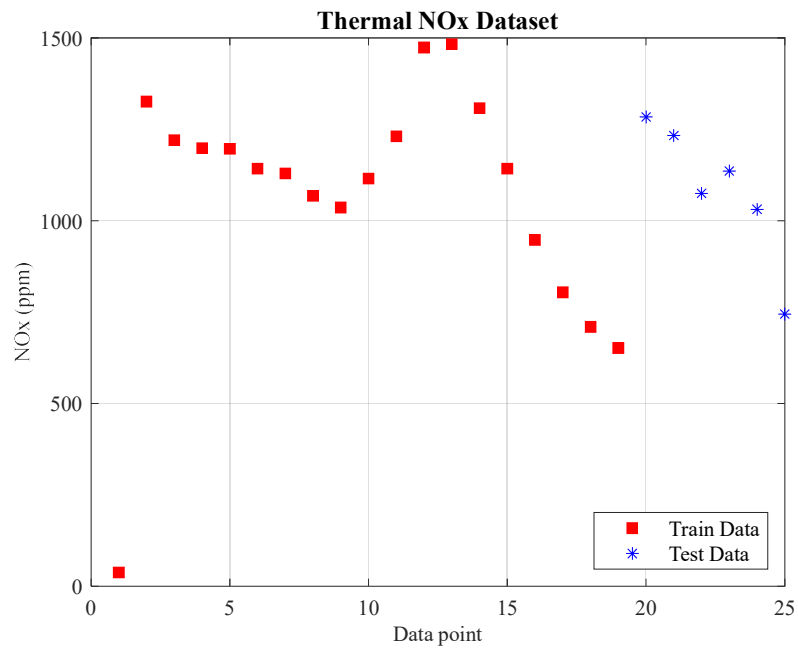
For developing data-driven soft sensors, both datasets are combined, resulting in 25 data points and 15 variables as shown in **Table 5**. Variables 1 to 13 are defined as input features (independent variables), while variables 14 and 15 are the target (dependent variable).

Table 5. Variable Descriptions of the CFD Dataset.

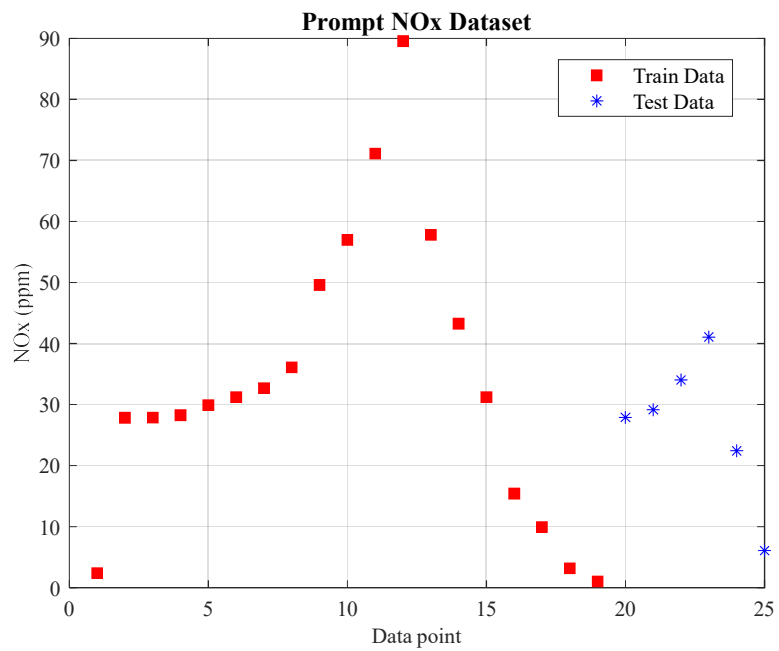
Feature #	Feature Name	Description	Unit
1	MCH4	The number of moles of methane	mole
2	MH2	The number of moles of hydrogen	mole
3	MSH2	Mass of hydrogen	kg
4	AFR	Air flow rate	kg/s

5	FFR	Fuel flow rate	kg/s
6	ATFR	Air-to-fuel ratio	dimensionless
7	STO	The number of stoichiometry (ideal fuel/air mixture)	dimensionless
8	LMD	The number of lambda (excess air factor)	dimensionless
9	MFHV	The Heating value of mixed fuel	kJ/kg
10	MFT	Maximum furnace temperature	K
11	AFT	Average furnace temperature	K
12	MET	Maximum exhaust temperature	K
13	AET	Average exhaust temperature	K
14	TNO _x	Average of thermal NO _x	ppm
15	PNO _x	Average of prompt NO _x	ppm

Before training the model, the data is randomly split into training and testing to avoid learning bias in the models, where 75% (19 samples) were used for training and 25% (6 samples) were used for testing. **Figure 11** shows the data visualization of Thermal NO_x (Figure 11a) and Prompt NO_x (Figure 11b) after splitting.



(a)



(b)

Figure 11. Visualization of Training and Testing Datasets: (a) Thermal NO_x, (b) Prompt NO_x.

CHAPTER 4

METHODOLOGY

In this work, AI-based soft sensor models were developed using ML and DL approaches to predict NO_x emission, efficiency, and boiler swings (i.e. fault prognosis). Five ML models including linear regression (LR), random forest (RF), k-nearest neighbors (KNN), and two DL algorithms, including multi-layer perceptron (MLP), and radial basis function (RBF) were used to predict NO_x emissions and efficiency. Additionally, a recurrent neural network-based long short-term memory (LSTM) was applied to build a fault prognosis model for predicting swings in the steam flow rate of the boiler. All datasets were pre-processed using feature engineering, and feature selection before building the models.

The general methodology used in the study is described in **Figure 12**. First, the data from the actual and CFD boiler was collected and then imported to the platform (i.e. Google Colab) for further processing. Data pre-processing was then performed to handle missing data and outliers, if any. The next step was to perform feature engineering, by converting data types into proper form (i.e. from string to integer) and creating new features as output by utilizing some parameters available in the original dataset. The next step was to perform feature selection, by applying several feature selection methods, including variance threshold, Pearson correlation, feature importance, and domain knowledge. The last part was to build the ML/DL models and evaluate them using performance metrics.

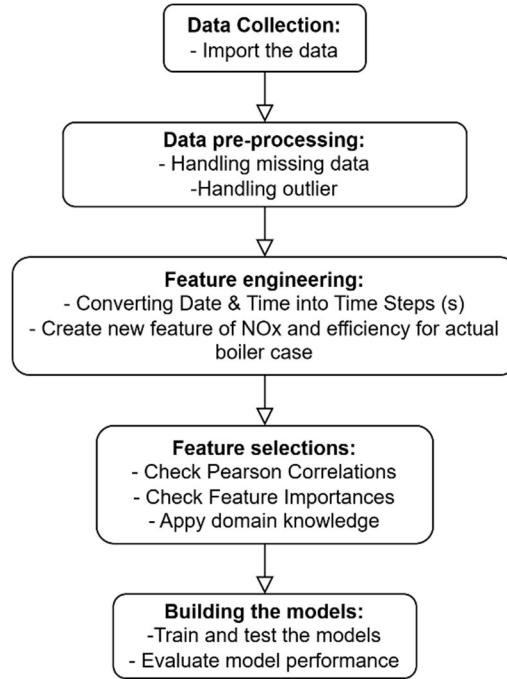


Figure 12. General methodology used in the current study

4.1 Data Pre-processing

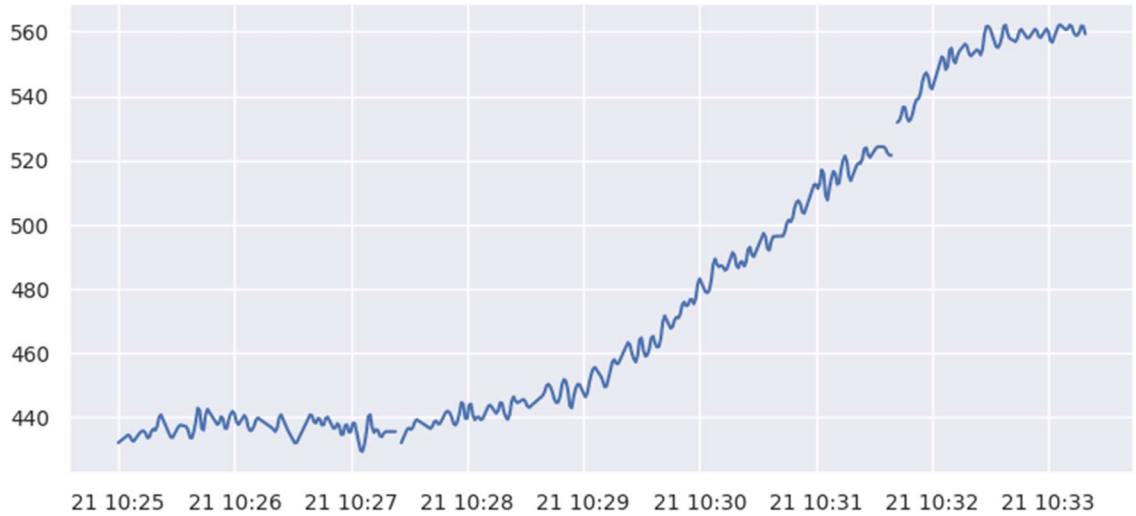
The first step was to explore the dataset and check for any missing values, outliers, or incorrect data types. For the natural gas boiler dataset, several data pre-processing techniques were employed. During the data exploration, it was found that the initial data type from the original datasets was presented as a string object, making it necessary to change the data type to the respective values. For example, data for Date and Time were converted into datetime64 format, and other parameter values were converted into floating-point numbers. This new data type format enabled the model to read and process the data during training and testing. Although there were no outliers in this dataset, the missing values were present in all variables, as shown in **Table 6**. These missing values likely occurred due to sensor malfunctions, reading errors, noise, or environmental disturbances during boiler operation.

Table 6. List of missing data in the natural gas boiler dataset

Feature #	Variable Name	Number of Missing Data
1	Date and time	90
2	Feed Water Flow Rate 1	90
3	Feed Water Flow Rate 2	18
4	Steam Flow Rate 1	18
5	Steam Flow Rate 2	18
6	Fuel Gas Flow Rate	18
7	Air Flow Rate	18
8	Forced Draft Discharge Pressure	18
9	Fuel Gas Pressure	18
10	Flue Gas Pressure 1	18
11	Flue Gas Pressure 2	562
12	Furnace Pressure	18
13	Ignitor Pressure	18
14	Steam Drum Pressure 1	18
15	Steam Drum Pressure 2	18
16	Deaerator Temperature	24
17	Steam Drum Temperature	18
18	Superheater Temperature 1	18
19	Superheater Temperature 2	18
20	Flue Gas Temperature	18
21	Oxygen Excess Air	18

In this study, several methods were used to evaluate the most appropriate techniques for filling in the missing values, including statistical approaches (i.e. mean, median, mode) and linear interpolation method. **Figure 13** presents the performance of various techniques for handling missing values, where Figure 13a shows the original data, and Figure 13b to Figure 13e depicts the data after filling by the methods. It is obviously

shown that linear interpolation provided a smoother fill to the data compared to the other methods and was thus considered the best approach for handling missing data in this dataset.



(a)



(b)



(c)



(d)



(e)

Figure 13. Various methods for handling missing data: (a) Original data, (b) Mean method, (c) Median method, (d) Mode method, (e) Linear Interpolation.

For the CFD hydrogen-enriched dataset, no missing values or outliers were found on this dataset, and all data were in the correct data types. The second step was to perform data standardization before training neural network models, a scaling technique used to transform each feature so that the data had a mean of zero and a standard deviation of one across all examples [48]. Data scaling was crucial to ensure that all features were within the same order of magnitude, avoiding large differences between the parameters that could significantly impact model convergence. Data scaling was performed on the dataset using Equation (10).

$$x_{scaled} = \frac{x - \mu}{\sigma} \quad (10)$$

where x is the input feature, μ is the mean, and σ is the standard deviation.

4.2 Feature Engineering

Feature Engineering is the process in machine learning to extract information from data, and create new features based on data analytics or domain knowledge [49]. In this study, feature engineering was applied to the NG Boiler dataset to create new features, namely NO_x and Efficiency, as output variables. NO_x and Emission were not originally available in the dataset; therefore, creating these new features using domain knowledge was required. The purpose was to use the available parameters in the original data and apply appropriate formulas to generate these new features. In this context, based on a study in the literature, the oxygen concentration available in the original data was sufficient to calculate NO_x emissions and Efficiency.

First, to calculate NO_x emission, this study employed Li and Thompson formula as explained in Section 2.2. After performing the calculation, a new parameter, NO_x , was obtained, and the visualization of this output variable is described in **Figure 14**.

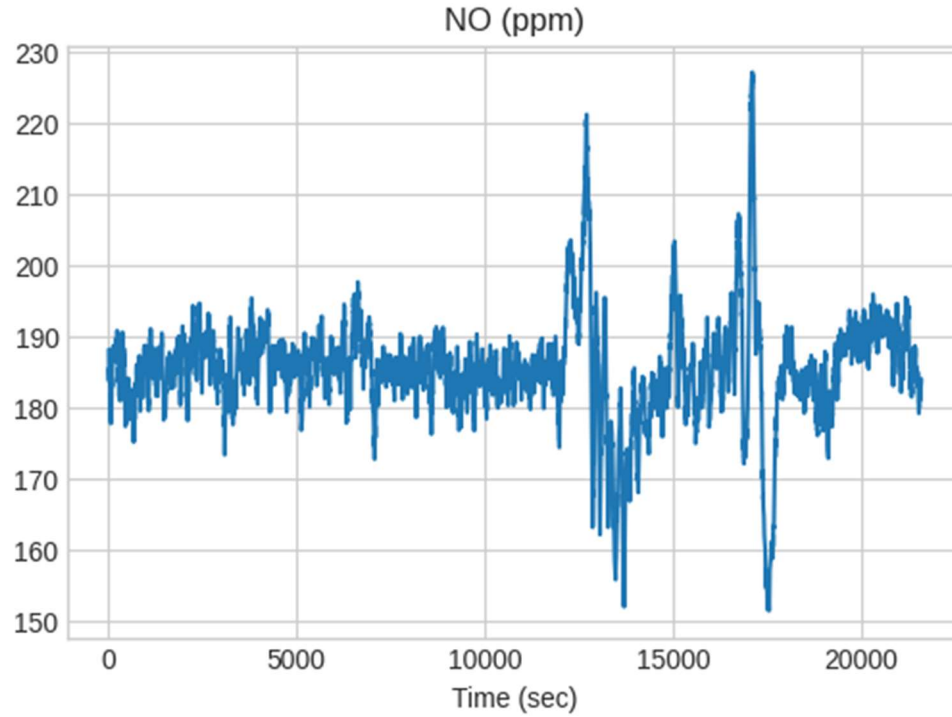


Figure 14. NO_x emission with feature engineering by using Li and Thompson formula.

The next stage was to calculate the Efficiency of the boiler. In this study, the Combustion Efficiency was calculated using the Siegert Formula, as explained in Section 2.3. After performing calculations using the Siegert formula, the new feature of efficiency was obtained and is described in **Figure 15**.

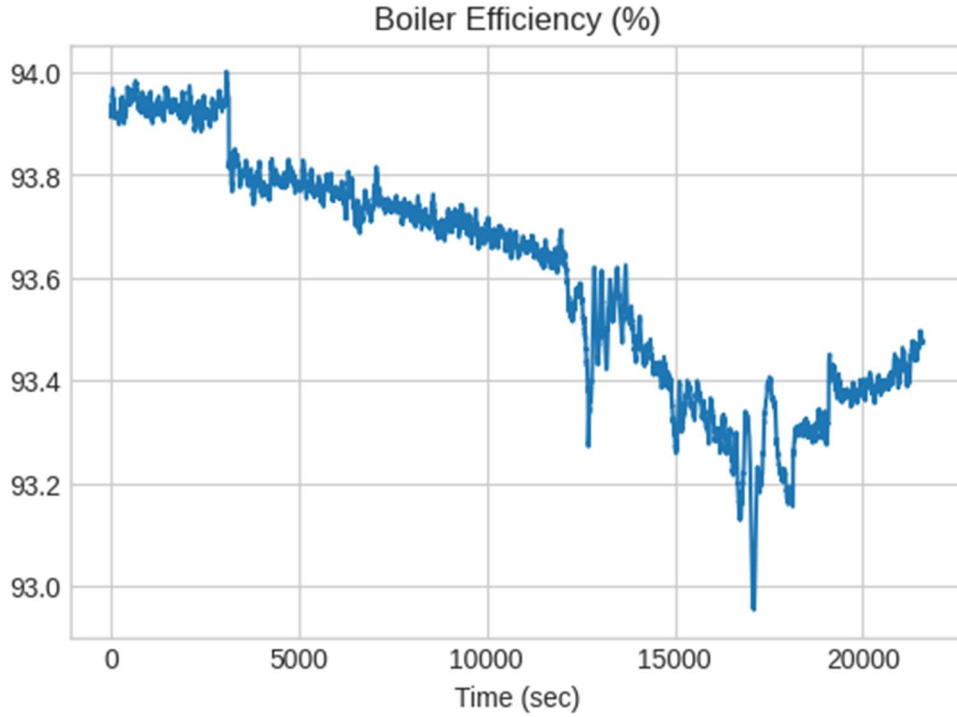


Figure 15. Efficiency obtained using feature engineering by the Siegert formula.

4.3 Feature Selections

Feature selection is one method for optimizing computational efficiency and minimizing the response time, and in some cases, it can increase the model performance. Some literature explored various techniques for effective feature selections. For example, Zhou et al. [50] used four different feature selection techniques to enhance model prediction, including Pearson correlation, RFE, and RF feature importance, to obtain the best input variable. Furthermore, Shao et al. [51] used total subset regression (TSR) and kernel principal component analysis (KPCA) to select the best features and reduce dimensionality before applying the model. Tang et al. [52] combined feature selection algorithms based on regression trees, random forest, extreme gradient boosting, and

maximal information coefficient (MIC) to select critical input variables. Similarly, Alfian et al. [53] employed Extra Tree-based feature selection for predicting breast cancer using the SVM model. In this section, a number of feature selection techniques are explained.

4.3.1 Variance Threshold

Variance is the measure of variability. It is calculated by taking the squared differences between each data x_i and the mean $\bar{x}$, and then dividing the sum of the squares by the number of values in the sample data set $n - 1$, as shown in Equation (11) below.

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} \quad (11)$$

Variance Threshold is a simple feature analysis technique by removes all features that have variance values below the designated threshold [54]. For example, in the case of a hydrogen-enriched boiler, when the threshold was set to zero, it removed features with zero variance, or in other words, it removed the features with similar values in all samples. In contrast, features with higher variance contained more useful information, which was needed for training the model. In this study, the variance threshold was set to zero. From **Table 7**, it was found that Stoichiometry had zero variance; thus, this variable was removed from the features because it did not affect the model's performance.

Table 7. The variance value of CFD Hydrogen-enriched boiler Dataset.

Feature #	Feature Name	Variance Value	Unit
1	MCH4	6.97e-02	mole
2	MH2	6.97e-02	mole
3	MSH2	4.44e-02	kg
4	AFR	6.01e+01	kg/s

5	FFR	3.14e-01	kg/s
6	ATFR	1.82e+01	dimensionless
7	STO	0.00e+00	dimensionless
8	LMD	8.65e-03	dimensionless
9	MFHV	2.24e+08	kJ/kg
10	MFT	1.35e+03	K
11	AFT	2.48e+03	K
12	MET	2.67e+03	K
13	AET	2.25e+03	K
14	TNO _x	9.15e+04	ppm
15	PNO _x	4.56e+02	ppm

4.3.2 Pearson Correlation and Feature Importance

Pearson Correlation and Feature Importance were used to identify the correlations between parameters and to evaluate the importance of the features to the output variable. Both methods were jointly employed to analyze input parameters and reduce possible parameters that had insignificant effects on the model. Pearson correlation measures the dependency between two random variables, allowing for the reduction of one input variable that had a strong correlation while keeping important information from the remaining variable. On the other hand, Feature Importance is used to identify the relative importance score of every input variable to the target variable. In this study, the extra tree-based feature importance method was performed [55].

The Pearson Correlation is the most popular measure and is defined as Equation (12) below [56]:

$$r_{xy} = \frac{\sum(x_i - \bar{x})\sum(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} \quad (12)$$

where $\bar{x}$ is the mean of x , $\bar{y}$ is the mean of y , the value of r_{xy} varies between -1 and 1, where $r_{xy} = 1$ when x and y have a positive correlation, $r_{xy} = 0$ when no obvious correlation between x and y , and $r_{xy} = -1$, when x and y are a negative correlation.

For the Actual Boiler Dataset, the Pearson Correlation between variables of the original data was depicted in **Figure 16**. From this figure, it was found that some of the features had a strong correlation with other features. For example, FFWR1 and FFWR2 had a strong correlation value of 1.0 (CV=1.0). Keeping only one of these features and removing the other was deemed harmless to the model because they essentially contained the same information. To decide which parameter to remove, feature importance in **Figure 17** indicated that FFWR1 was less important compared to FFWR2. Thus, FFWR2 was retained, and FFWR1 was removed from the input variables. Similarly, SFR1 and SFR2 also exhibited a strong correlation with a value of 1.0 (CV=1.0) because they measure the same properties. With the help of feature importance, it was determined that SFR2 was more important than SFR1, so SFR2 was retained and SFR1 was removed. The same method was applied to other pairs of correlated features, including SHT1-SHT2 (CV=0.8), FLGP1-FLGT (CV=0.9), IGP-FLGT (CV=-0.9), FUP-FFR (CV=1.0), and FGP-FFR (CV=1.0).

Domain knowledge was also applied during feature selections. For example, Flue Gas Pressure (FLGP1) and Flue Gas Temperature (FLGT) were identified as dependent variables according to the Ideal Gas Law ($PV=nR*T$). As pressure increases with temperature, the correlation value between these variables was high ($CV = 0.9$). For machine learning purposes, retaining only one of these parameters was sufficient for the model to capture the necessary information. Based on the feature importance, FLGP1 was less important than FLGT, so FLGT was retained, and FLGP1 was removed. A similar analysis was conducted using domain knowledge for Ignitor Gas Pressure (IGP) and Flue Gas Temperature (FLGT). The temperature of flue gas is an outcome of fuel combustion, and when the flue gas temperature becomes too high, fuel input is reduced to control the steam drum temperature. As a result, these two variables had a negative correlation ($CV = -0.9$). For machine learning purposes, only one of these variables was needed. Feature importance analysis revealed that IGP was less important, so it was removed from the input variables. After performing a thorough analysis of all parameters using a combination of Pearson Correlation, Feature Importance, and Domain Knowledge, the following parameters were removed from the input variables for the NO_x emission prediction problem: FWFR1, SFR1, SHT1, FLGP1, IGP, FUP, and FGP. This reduced the number of input variables from 19 to 12. The reduced parameters after feature selection for the NO_x emission problem are provided in **Table 8**.

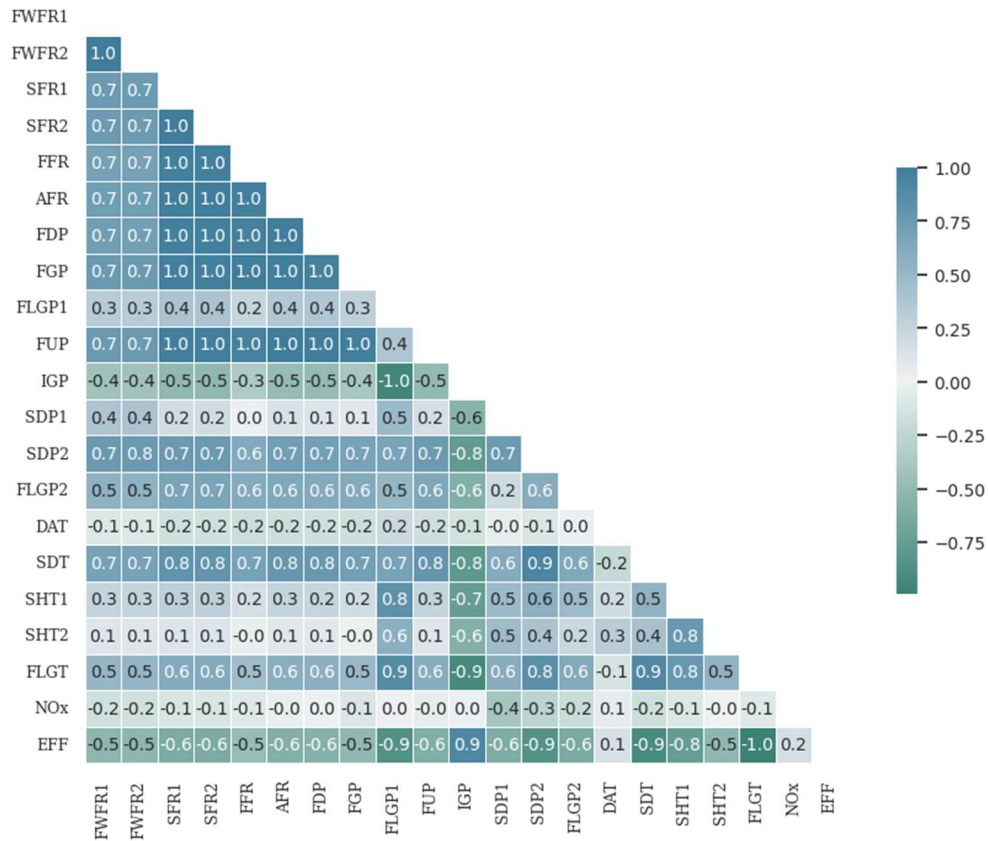


Figure 16. Pearson Correlation for NO_x of NG dataset variables.

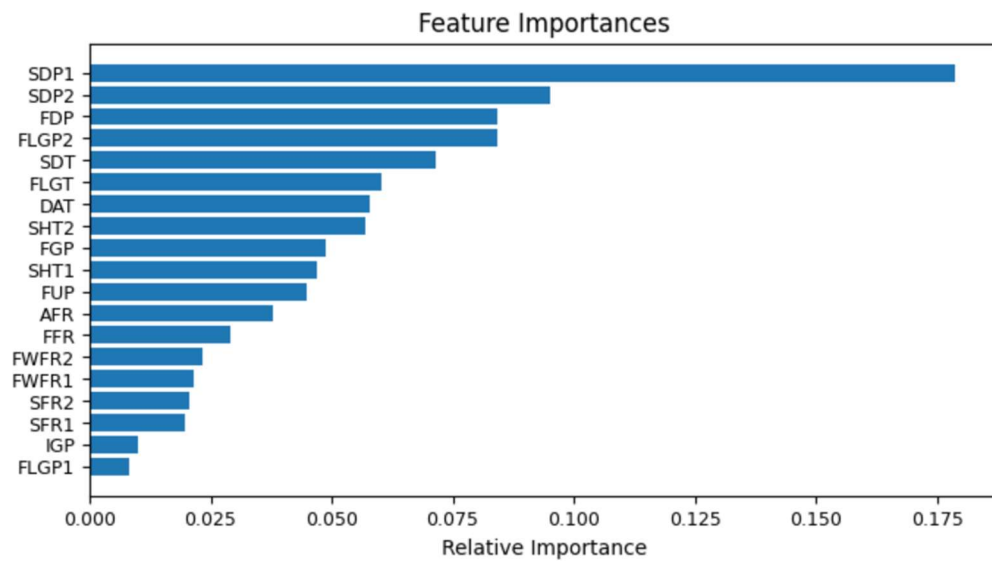


Figure 17. Feature Importance for NO_x of NG dataset variables.

Table 8. Input parameter for NO_x emission NG Boiler Dataset

Feature #	Feature Name	Description	Unit
1	FWFR2	Feed Water Flow Rate 2	klbs/hr
2	SFR2	Steam Flow Rate 2	klbs/hr
3	FFR	Fuel Flow Rate	kSCF/hr
4	AFR	Air Flow Rate	inch H2O
5	FDP	Forced Draft Discharge Pressure	inch H2O
6	FLGP2	Flue Gas Pressure 2	inch H2O
7	SDP1	Steam Drum Pressure 1	psig
8	SDP2	Steam Drum Pressure 2	psig
9	DAT	Deaerator Temperature	Fahrenheit (°F)
10	SDT	Steam Drum Temperature	Fahrenheit (°F)
11	SHT2	Superheater Temperature 2	Fahrenheit (°F)
12	FLGT	Flue Gas Temperature	Fahrenheit (°F)

For the Boiler NG efficiency problem, the same methods as described above were applied, combining Pearson Correlation, Feature Importance, and Domain Knowledge. Pearson Correlation and Feature Importance results for this problem are shown in **Figure 18**, and **Figure 19**, respectively. Notably, the total number of original features in these figures is 18, as Flue Gas Temperature (FLGT) had already been excluded during the feature engineering process. This exclusion was due to FLGT becoming dependent on the output variable (efficiency), as defined by Equations (6) and (7). After a thorough analysis, six variables were removed from the input variables: FWFR1, SFR1, SHT1, FDP, FGP, and FUP. This selection process reduced the total number of input variables to 12. The reduced parameters for the Boiler NG efficiency problem are presented in **Table 9**.

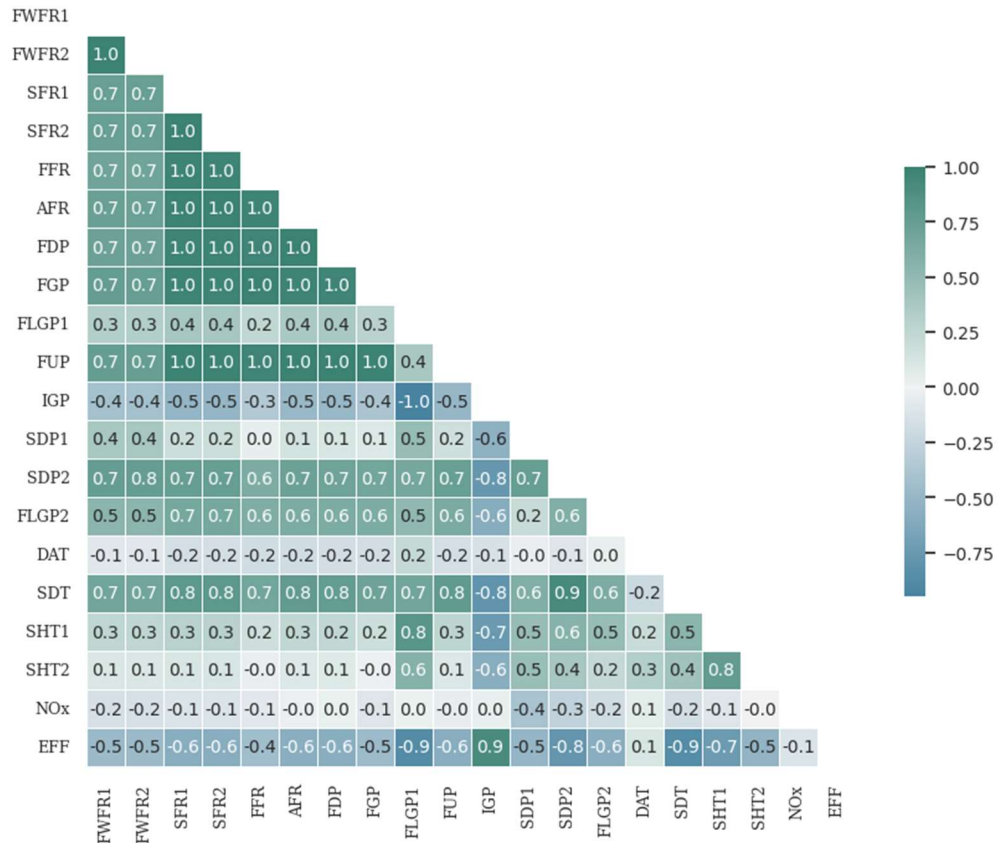


Figure 18. Pearson Correlation for Efficiency of NG Dataset Variables.

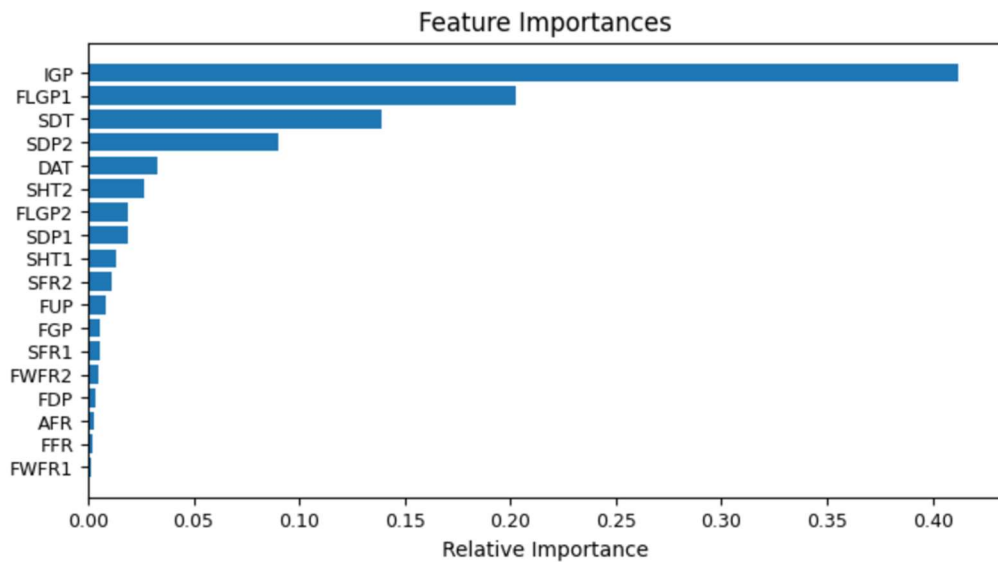


Figure 19. Feature Importance for Efficiency of NG Dataset Variables.

Table 9. Input parameter for Efficiency Natural Gas Boiler Dataset

Feature #	Feature Name	Description	Unit
1	FWFR2	Feed Water Flow Rate 2	klbs/hr
2	SFR2	Steam Flow Rate 2	klbs/hr
3	FFR	Fuel Flow Rate	kSCF/hr
4	AFR	Air Flow Rate	inch H2O
5	FLGP1	Flue Gas Pressure 1	inch H2O
6	FLGP2	Flue Gas Pressure 2	inch H2O
7	IGP	Ignitor Pressure	psig
8	SDP1	Steam Drum Pressure 1	psig
9	SDP2	Steam Drum Pressure 2	psig
10	DAT	Deaerator Temperature	Fahrenheit (°F)
11	SDT	Steam Drum Temperature	Fahrenheit (°F)
12	SHT2	Superheater Temperature 2	Fahrenheit (°F)

The same methods were applied to identify the CFD Dataset. First, the correlation between parameters was analyzed using Pearson Correlation, followed by the use of feature importance to determine which features to remove from the input parameters. For example, using Pearson Correlation in **Figure 20**, the number of moles of methane (MCH4) had a strong correlation with fuel flow rate (FFR) with a correlation value of 1.0. Using domain knowledge, it was understood that the number of moles of methane had a linearly dependent relationship with the fuel flow rate because methane is the main component of the fuel for the boiler, so one of these parameters could be kept while the other was removed. On the other hand, it was also found that the number of moles of hydrogen (MH2) had a strong correlation with the mass of hydrogen (MSH2) with a correlation value of 0.9, and with mixed fuel heating value (MFHV) with a correlation value of 0.9. This because the mass of hydrogen was linearly dependent on the number of moles of hydrogen, while

the mixed fuel heating value correlated with the increase in the number of moles of hydrogen added to the fuel. Furthermore, the value of air flow rate (AFR) had a strong correlation with lambda (LMD) with a correlation value of 0.9, because lambda is the ratio of the actual to the stoichiometric air-to-fuel ratio, thus lambda had a linear correlation with the airflow rate. Based on this investigation, four parameters were potentially removed.

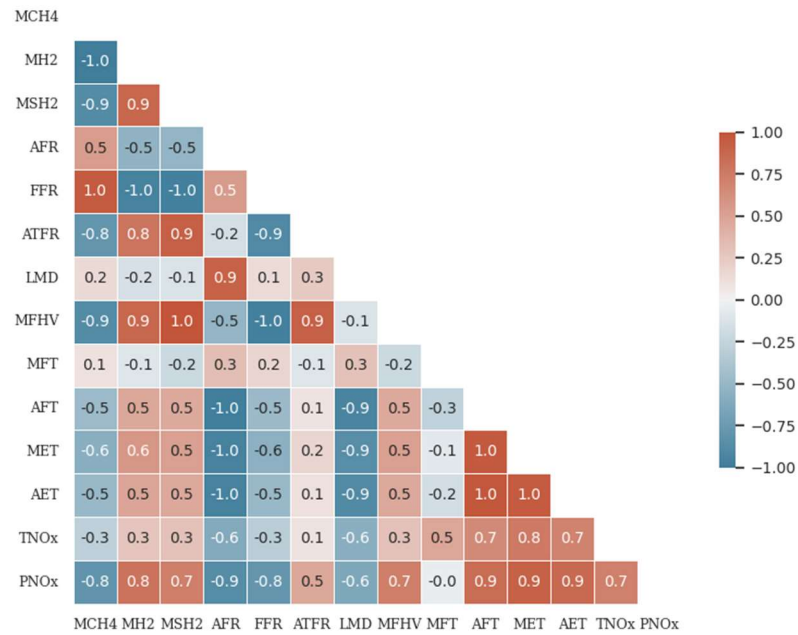


Figure 20. Pearson Correlation for the CFD dataset.

Feature importance in **Figure 21** was then used to determine which variables to remove based on the Pearson Correlation method. For example, according to the Pearson Correlation method, MCH4 and FFR had a strong correlation, and based on Feature Importance, FFR had a lower importance value; thus, FFR was removed from the features. The same approach was applied to MH2 with MSH2 and MFHV, where MSH2 and MFHV had lower importance values and were therefore removed from the features. This method was also applied to AFR and LMD, where AFR had the lower importance value and was

subsequently removed. These observations resulted in a total of four variables being removed from the features: MFHV, MSH2, FFR, and AFR.

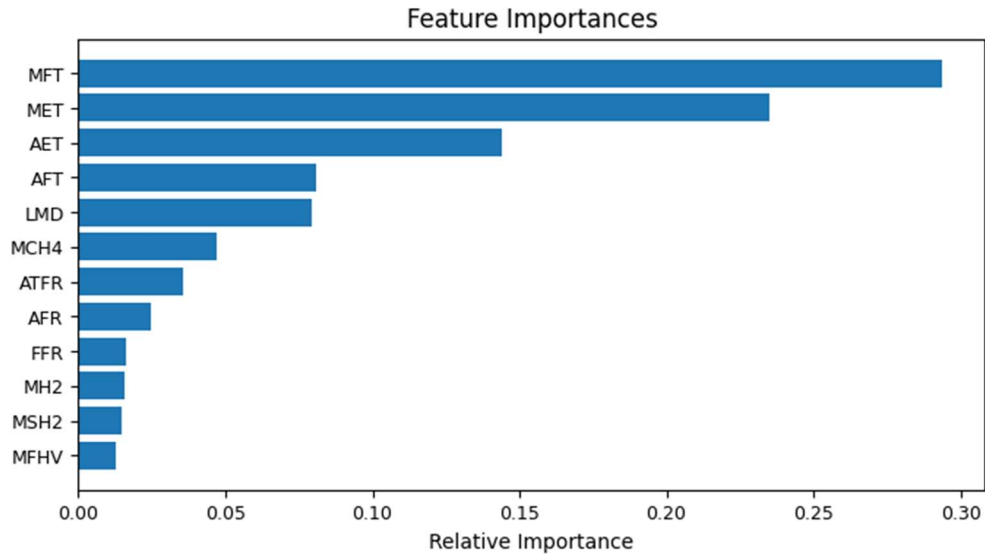


Figure 21. Feature Importance using Extra Trees Algorithm.

According to the feature selection processes described above (variance threshold, Pearson Correlation, and Feature Importance), five parameters were removed from the original CFD Dataset, including STO, FFR, AFR, MSH2, and MFHV. As a result, only eight input parameters were considered as input variables. The reduced dataset, after performing feature selection for the NO_x emissions problem of the hydrogen-enriched CFD boiler, is presented in **Table 10**.

Table 10. Input parameter for NO_x emission CFD Dataset after feature selections.

Feature #	Feature Name	Description	Unit
1	MCH4	The number of moles of methane	Mole
2	MH2	The number of moles of hydrogen	Mole
3	ATFR	Air-to-fuel ratio	Dimensionless

4	LMD	The number of lambda (excess air factor)	Dimensionless
5	MFT	Maximum furnace temperature	K
6	AFT	Average furnace temperature	K
7	MET	Maximum exhaust temperature	K
8	AET	Average exhaust temperature	K

4.4 Building the AI Models

In the development of soft sensor models for predicting NO_x emissions and efficiency, it is assumed that the time delay between the process input variables such as fuel flow rate, air flow rate, and temperature, and their propagation to the output variables (which are NO_x emissions and efficiency) is negligible. This assumption simplifies the modeling process by allowing the relationship between inputs and outputs to be treated as quasi-instantaneous (very short delay) that is small enough to be negligible for practical purposes and reducing the complexity of incorporating time lag effects [57].

Five ML models and two DL models were used to build the soft sensors for predicting NO_x emission and efficiency, as these seven models are suitable for regression problems. The models included linear regression (LR), decision tree (DT), random forest (RF), k-nearest neighbor (KNN), support vector regression (SVR), multi-layer perceptron (MLP), and radial basis function (RBF). Additionally, a long short-term memory (LSTM) model, which is a recurrent neural network-based approach, was employed for the fault prognosis model due to its capability of storing past information over specific time intervals. An introduction to each model is provided in this section.

4.4.1 Linear Regression (LR)

Linear regression is a machine learning algorithm to model the relationship between one or more independent variables and a continuous dependent variable by fitting a linear equation to observed data. The objective is to find the best-fitting line that minimizes the difference between the predicted values and the actual data points, typically using the least squares method. The general equation for linear regression can be expressed in Equation (13) below.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_n X_n + \epsilon \quad (13)$$

where Y is the predicted value, β_0 is the intercept, $\beta_1, \beta_2, \dots, \beta_n$ are the coefficients for the independent variables $X_1, X_2, \dots, X_n$, and ϵ is the residual error. Linear regression is widely used due to its simplicity and interpretability, making it a popular choice for predictive modeling and data analysis [57]. The illustration of simple linear regression is provided in **Figure 22**.

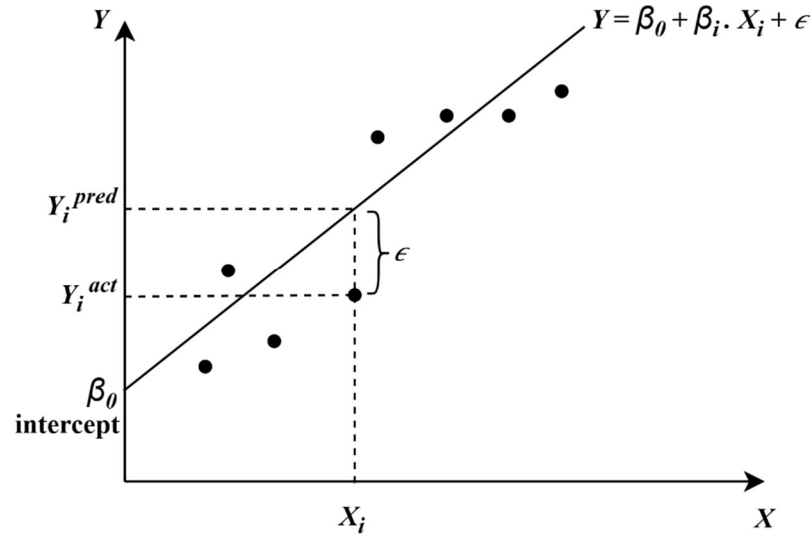


Figure 22. Illustration of Linear Regression.

4.4.2 Decision Tree (DT)

Decision Tree (DT) is a predictive modeling that employs a tree-like structure to predict continuous output. In this method, the dataset is split into subsets based on feature values, with each split aiming to minimize the variance of the target variable within the resulting groups. The algorithm creates branches for different feature thresholds, leading to terminal nodes (leaves) that provide the predicted value based on the average of the target variable in that subset [58]. The illustration of Decision Tree Regression is depicted in **Figure 23** below, to predict housing prices from the Kaggle Dataset [59].

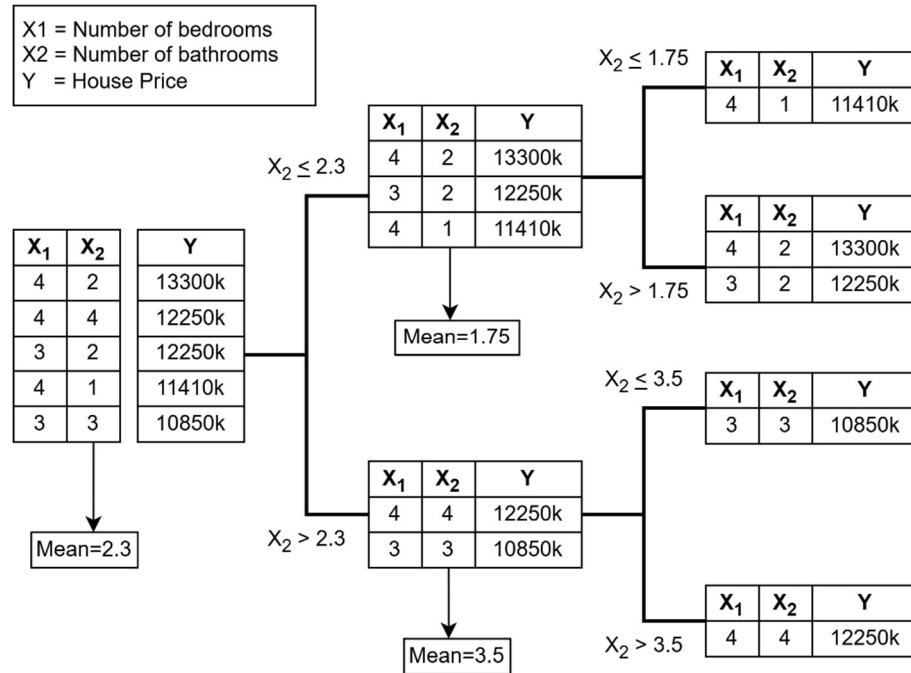


Figure 23. Illustration of Decision Tree Regression.

4.4.3 Random Forest (RF)

Random Forest (RF) is an ensemble method that involves multiple decision trees during training. Each tree is developed based on a random subset of data samples and features called bagging, which enhances accuracy and reduces overfitting. In this approach,

each tree is trained on a unique bootstrapped data sample, and random feature subsets are selected at each split to improve generalization. In the regression model, the final prediction is the average of all the outputs, reducing variance and increasing accuracy [60]. The Random Forest regression model's prediction is expressed in Equation (14) below,

$$\hat{y} = \frac{1}{B} \sum_{i=1}^B f_i(x) \quad (14)$$

where $f_i(x)$ represent the prediction from the i -th tree, and B is the total number of trees.

The simple illustration of Random Forest is provided in **Figure 24**.

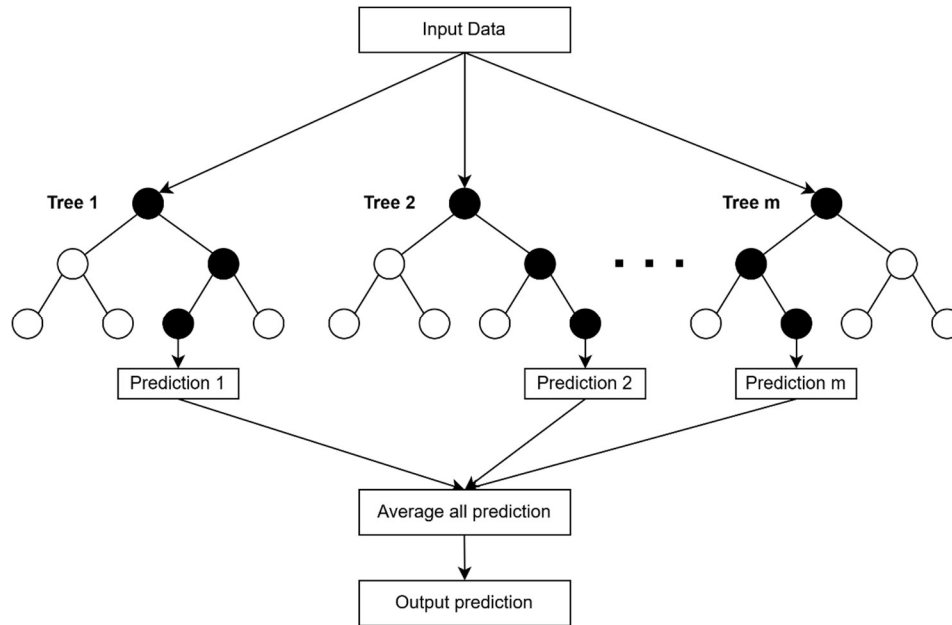


Figure 24. Illustration of Random Forest.

4.4.4 K Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is an instance-based learning for classification and regression tasks. In KNN, a new data point is classified based on the majority class of its nearest neighbors, typically determined by a distance metric like Euclidean distance. For

regression, the output is the average of the values of its nearest neighbors. The algorithm works by storing the entire training dataset, and predictions are made by finding the k closest points to the input based on the chosen metric [61]. The Euclidean distance, D between two points x_i and y_i in KNN is commonly calculated using Equation (15),

$$D = \sqrt{\sum_{i=1}^k (x_i - y_i)^2} \quad (15)$$

and the estimation value, V_x^{est} is obtained by averaging the output, V_y which is the estimation value of every k -number of neighbors, as shown in Equation (16), and the schematic of the KNN algorithm is illustrated in **Figure 25**.

$$V_x^{est} = \frac{1}{k} \sum_{y=1}^k V_y \quad (16)$$

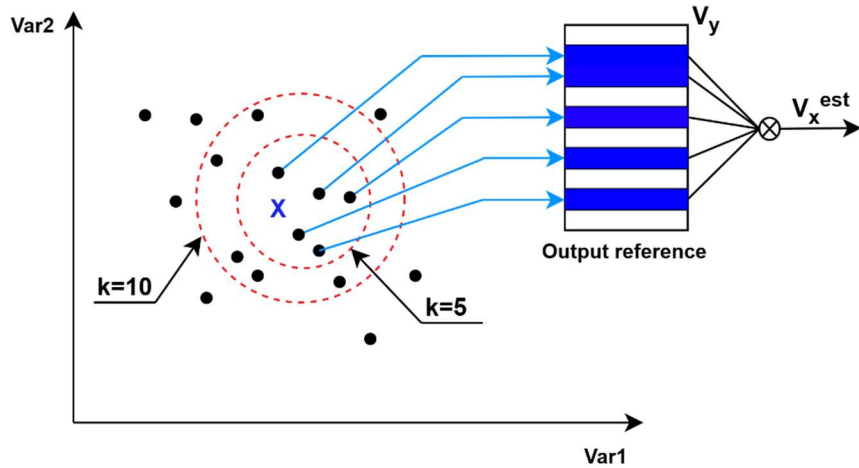


Figure 25. Illustration of K-nearest neighbors.

4.4.5 Support Vector Regressor (SVR)

Support Vector Regression (SVR) is a machine learning algorithm derived from Support Vector Machines (SVM). SVR aims to predict continuous outcomes by finding a function that has at most a specified deviation (denoted by ϵ) from the actual targets, while also maintaining model complexity at a minimum. It works by mapping the input data into a high-dimensional feature space and then finding a hyperplane that fits the data within a margin of tolerance. The algorithm minimizes an objective function that balances the error penalty and the flatness of the regression function [23]. The basic SVR optimization problem is formulated in Equation (17) below,

$$\min \frac{1}{2} ||w||^2 + C \sum_{i=1}^n (\xi_i - \xi_i^*) \quad (17)$$

subject to constraints $y_i - (w \cdot x_i + b) < \epsilon + \xi_i$ and $(w \cdot x_i + b) - y_i < \epsilon + \xi_i^*$, where w represents the weights, C is a regularization parameter, and ξ_i, ξ_i^* are slack variables.

Figure 26 given below illustrates a one-dimensional example of support vector regression.

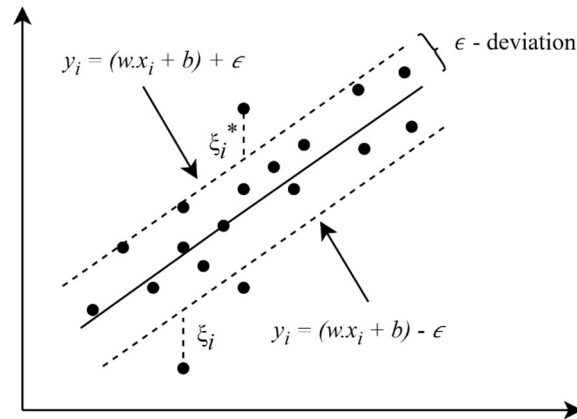


Figure 26. Illustration of One-Dimensional Support Vector Regression.

4.4.6 Multi-Layer Perceptron (MLP)

The multi-layer perceptron (MLP) is a conventional neural network model composed of multiple layers: an input layer, hidden layers, and an output layer. Each layer includes perceptron units, often referred to as neurons. The input layer corresponds to the number of input features, while the output layer represents the target output. Neurons in the hidden and output layers employ activation functions to introduce non-linearity into their outputs. To optimize the weights of the connections and reduce the model's loss function, backpropagation is typically used as the training method [24]. The structure of MLP is described in **Figure 27**.

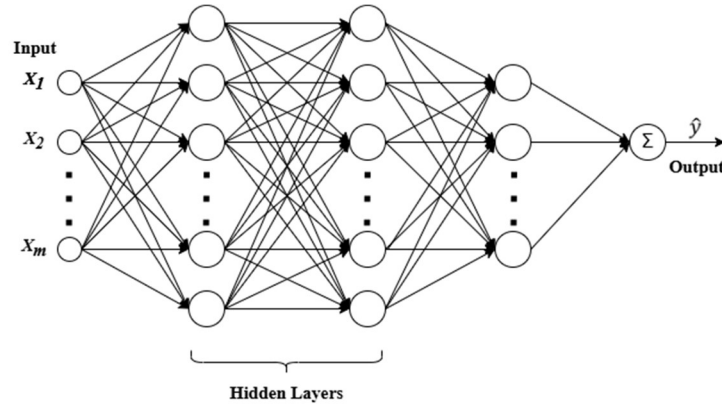


Figure 27. Structure of multi-layer perceptron (MLP).

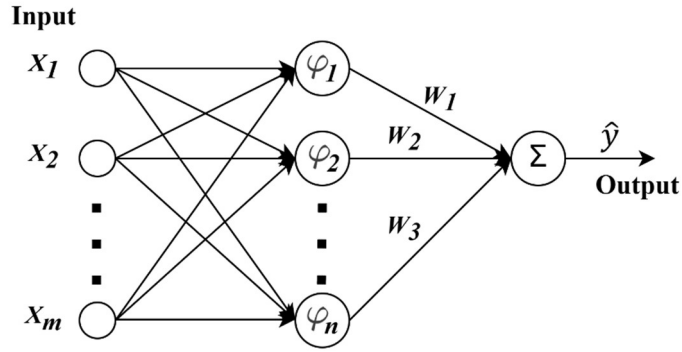
In this study, a backpropagation neural network was implemented with three hidden layers. The number of neurons in each layer and the hyperparameters were determined using the grid search cross-validation method (Grid Search CV). **Table 11** presents the optimal parameters obtained after hyperparameter tuning.

Table 11. Hyperparameters for MLP models.

Hyperparameter	Thermal NOx MLP	Prompt NOx MLP
Hidden layer size	30,30,30	20,20,20
Activation function	Rectified Linear Unit (ReLU)	Rectified Linear Unit (ReLU)
Optimizer	Adaptive Moment Estimation (Adam)	Adaptive Moment Estimation (Adam)

4.4.7 Radial Basis Function (RBF)

The RBF neural network is a neural network-based architecture suitable for online monitoring of the system and is proven to work with a small amount of data. This model has a simple structure with only one hidden layer, thus providing a fast-learning algorithm and less computational cost. The RBF neural network is a feedforward neural network with good universal approximation capabilities with no local minima problem [62].

**Figure 28.** Structure of RBF Neural Networks.

The structure of the RBF neural network is described in **Figure 28**. Each node in the hidden layer is called a center. These centers take data from input and produce radially symmetrical responses. The output from these centers is then scaled by the weight parameters in each connection and summed up using the linear method in the output layer

to get the output result. The most commonly used radial base activation function, and also used in this study is the Gaussian function, as defined in Equation (18) below:

$$\varphi(x) = \exp\left(-\frac{\|x - c_i\|^2}{\sigma_i^2}\right) \quad (18)$$

where c_i is the center and σ_i is the spread of the RBF node.

In this study, the input layer represented the input parameters, consisting of 13 neurons. The input was propagated to the hidden layer, where a Gaussian function approximation calculated the proximity of the input to the center in each hidden node. The outputs from the hidden layer were then linearly summed in the output layer to produce the final estimated output value. The trained model was subsequently used to predict the testing set, with the output validated and measured using performance metrics. To determine the centers in the hidden layer, the K-means clustering method was employed, with the number of centers being optimized to minimize the loss function. The best performance was achieved when the number of centers and RBF widths were set to $c = 10$ and $\sigma = 20$ for the Thermal NO_x model, and $c = 12$ and $\sigma = 6$ for the Prompt NO_x model.

4.4.8 Long Short Term Memory (LSTM)

LSTM (Long Short-Term Memory) is an advanced architecture of recurrent neural networks (RNN), introduced by Schmidhuber in 1997, designed to address the challenges of vanishing and exploding gradients in traditional RNNs. RNNs process sequential data by connecting outputs from previous steps as inputs to the current step, forming a chain-like structure. However, their ability to retain information from earlier time steps diminishes over time due to the vanishing gradient issue. To overcome this, LSTM replaces

the standard RNN cells with LSTM cells, adding self-connected memory units and operations for storing, reading, writing, and judging information. Each LSTM cell includes three key gates: the input gate, which controls incoming data; the forget gate, which determines what historical information to retain; and the output gate, which regulates the release of processed data. This structure allows LSTM networks to effectively capture long-term dependencies in sequential data [63]. The structure of the LSTM cell is described in **Figure 29**.

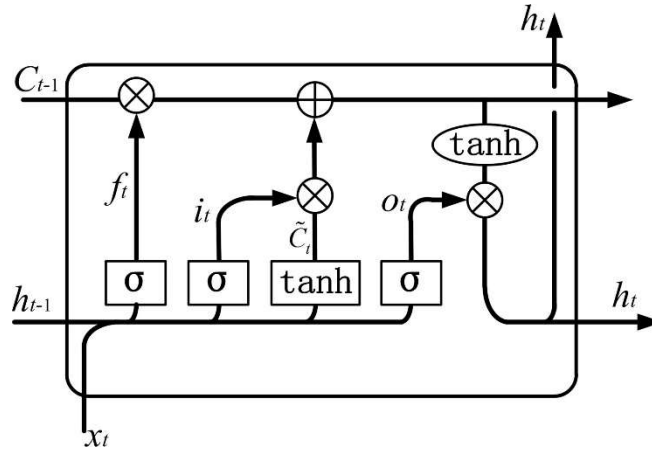


Figure 29. Structure of cell in long-short term memory (LSTM) neural networks [63].

where C_t represents the calculation rules of the LSTM cells at time t , while h_t denotes the output of the calculation unit at time t . The variables i_t , f_t , and o_t correspond to the input gate, forget gate and output gate. Each of these gates is connected to a multiplicative unit, which controls the flow of information and the state transitions within each LSTM cell. The LSTM structure was employed for the fault prognosis model, consisting of three hidden layers: the first hidden layer with LSTM (64 neurons, tanh), the second hidden layer with LSTM (32 neurons, tanh), and the third hidden layer with a linear layer (16 neurons,

ReLU). The output layer consists of 1 neuron with a sigmoid activation function, designed for a binary classification problem (normal vs. fault).

4.5 Performance Metrics

A number of metrics were used to evaluate the performance of each model. These metrics include mean absolute error (MAE), prediction accuracy (PA), the coefficient of determination (R-squared, R^2), and Fault Prognosis Accuracy [64].

The mean absolute error (MAE) is suitable to measure the costs associated with the average absolute error between predicted and actual values in the original unit. For example, here the value of Thermal NO_x is in the range of 37 to 1483 parts per million (ppm), thus the MAE will provide the error in the unit of ppm. The MAE is obtained using the formula as seen in Equation (19),

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (19)$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the total number of examples.

The prediction accuracy (PA) is to measure the model accuracy in percent and is obtained by subtracting 100% accuracy from the mean absolute percentage error (MAPE). Thus, the PA is presented using Equation (20),

$$PA = 100 - MAPE \quad (20)$$

where MAPE is the percentage of the average value of the absolute errors based on Equation (21).

$$MAPE = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (21)$$

The coefficient of determination or R-squared (R^2) is a measure to calculate the goodness of fit of a model. R^2 indicates the proportion of the variance in the actual variable (y_i) and the predicted variable ($\hat{y}_i$) as seen in Equation (22),

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (22)$$

The accuracy is to measure fault prognosis accuracy and defined as the comparison between the correct number prediction and the total number of predictions, as provided by Equation (23).

$$Accuracy = \frac{\text{Number of Correct Prediction}}{\text{Total Number of Prediction}} \quad (23)$$

4.6 Data Augmentation for Model Validation

Ensuring the reliability and applicability of the models required validation with extensive data from actual boiler systems. However, as highlighted in the Introduction, the availability of datasets for hydrogen-enriched natural gas-fueled steam boilers remained limited. Most industrial systems operated with hydrogen blends of no more than 20% by volume, making model validation with real-world data challenging, particularly for 100% hydrogen blends. Therefore, for predicting NO_x emissions using the CFD dataset, which had a limited number of sample data, this study employed augmented datasets as an alternative for model validation. The data augmentation method proposed by Iliyas in [34] was used with modifications, incorporating a pre-trained machine learning model to

generate target/output data. A total of 250 data points, including the original 25, were created by generating randomized input data with a 0.02 standard deviation for each operating condition. A pre-trained RBF model (with the original 13 input features) was then used to predict the outputs. New data points were generated around the original data, and only 9 data points, with a prediction error below 1% relative to the reference data, were retained. The schematic for this data augmentation process is illustrated in **Figure 30**.

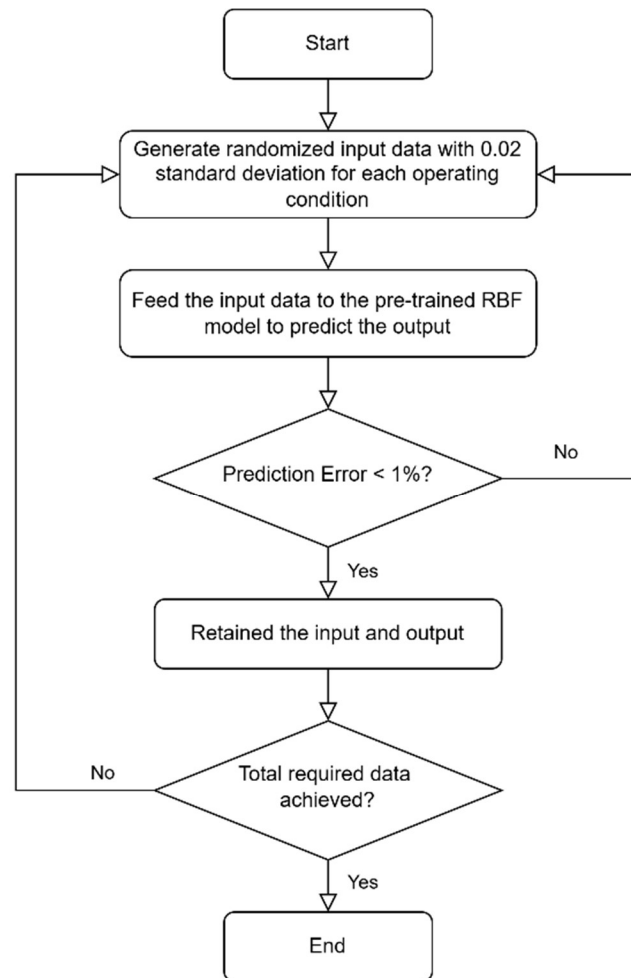


Figure 30. Flowchart of Data Augmentation.

CHAPTER 5

RESULTS AND DISCUSSIONS

This chapter presents the key findings of the three objectives. First, soft sensor models based on the actual natural gas boiler, to predict NO_x emission and efficiency are discussed, presenting various machine-learning model performances, using original and reduced datasets. The performance evaluation and simulation time of each model are also discussed. The second part of this chapter discusses the performance of soft sensor models based on CFD hydrogen-enriched datasets, presenting various machine-learning model performances using original and reduced datasets, and also validation using augmented datasets. Various Feature Engineering and Feature selection techniques are also investigated. The last section discusses the fault prognosis models using LSTM, by comparing various time steps with the model accuracy, to predict fault conditions based on previous boiler operating conditions.

5.1 Industrial Boiler Fueled with Natural Gas

The first objective of this study is to predict NO_x emission and efficiency from the actual boiler datasets. As discussed in section 4.2.1 Feature Engineering, the new features of NO_x and efficiency have been created using oxygen concentrations available in the original dataset. In this section, both NO_x and efficiency are predicted using various ML/DL approaches as introduced in section 4.3. Two datasets for each model will be employed, first using original datasets, and the second using reduced datasets after the

feature selections process as depicted in section 4.2.2 Feature Selections, to evaluate the effectiveness of feature reduction to the model performance and simulation time.

5.1.1 NO_x Emission Prediction

After performing data pre-processing, the data with a total of 21600 samples is split randomly into a training set (80%) and a testing set (20%). Before training the neural network models, the data is normalized using Standard Scaler to transform the data into the same magnitudes. The seven ML models and two DL models are evaluated and compared. The first part uses all original input features, and the next is to utilize the reduced dataset to investigate model performance and computational cost. The illustration of the training and testing for the NO_x emission dataset is provided in **Figure 31**.

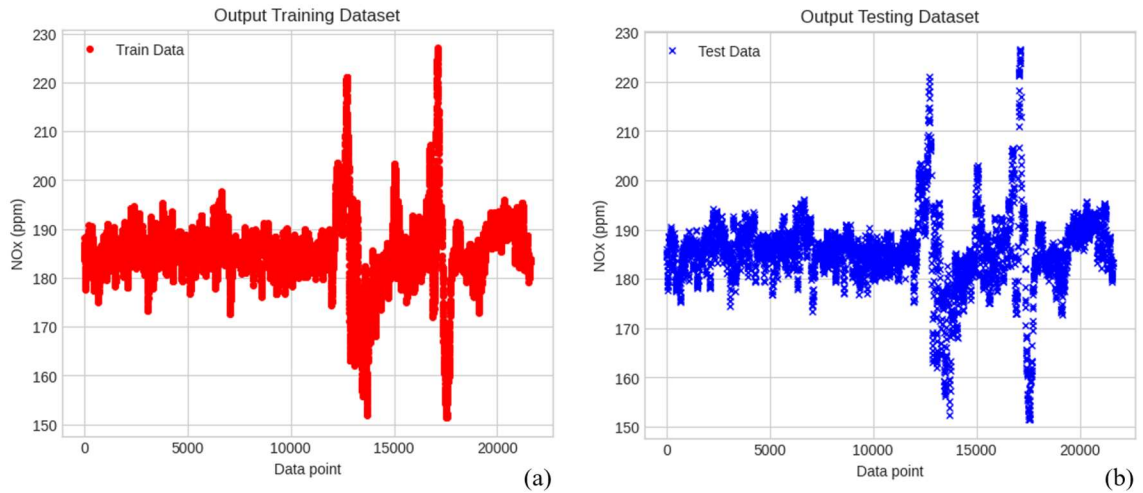


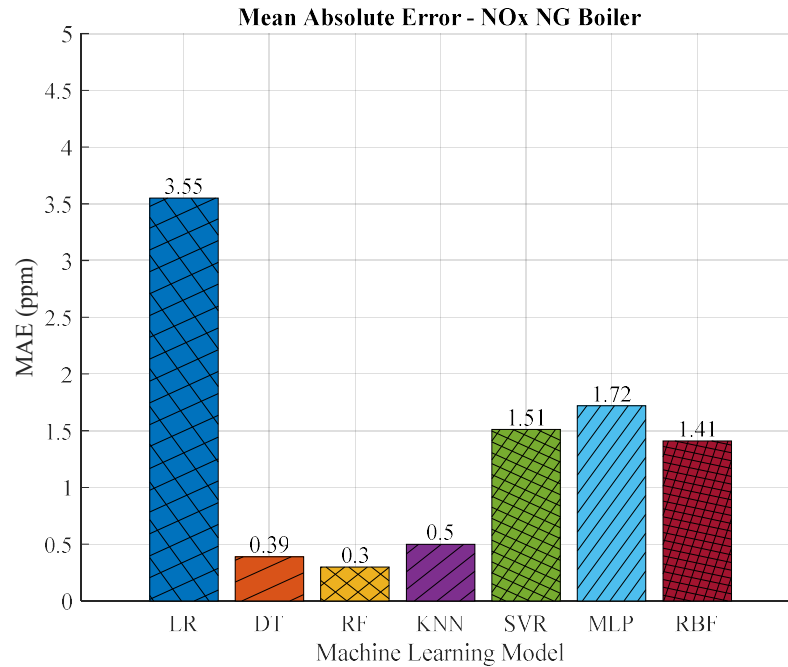
Figure 31. Training and Testing Datasets of NO_x Emission from: (a) Training Set 80%, (b) Testing Set 20%

a. Models Performance Using Original Dataset

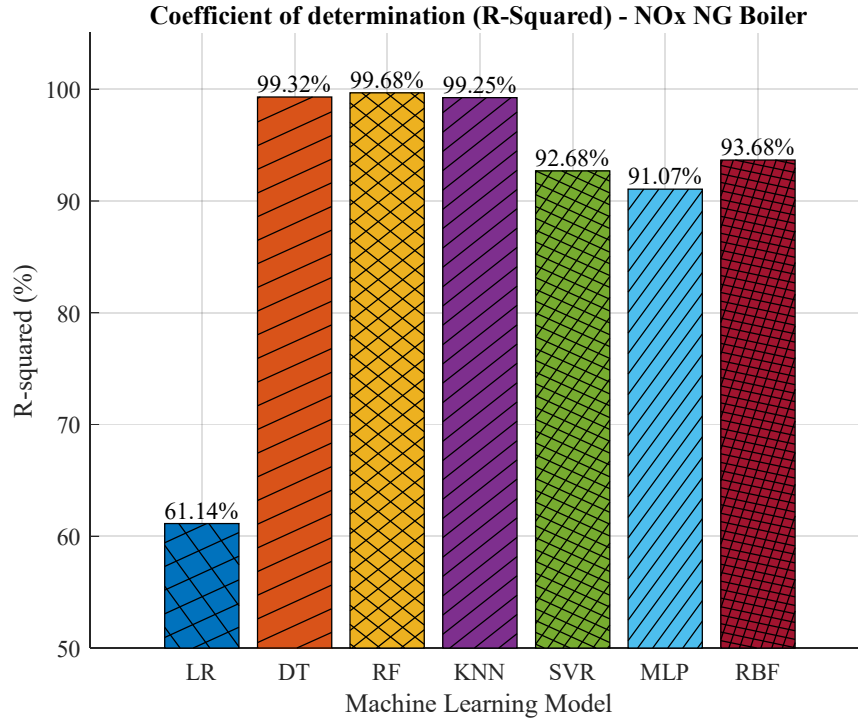
The seven AI models have been trained and tested using the original dataset of 19 input parameters. The summary of performance metrics is depicted in **Table 12**, and a comparison diagram of each model's performance is provided in **Figure 32**.

Table 12. Performance Metrics of NO_x Prediction Boiler NG Dataset (19 inputs).

Model ML/DL	MAE (ppm)	R ² (%)	Simulation Time (s)
Linear Regression	3.55	61.14	0.046
Decision Tree Regression	0.39	99.32	0.728
Random Forest	0.30	99.68	53.632
K Nearest Neighbor	0.50	99.25	4.683
Support Vector Regression	1.51	92.68	133.148
Multi Layer Perceptron	1.72	91.07	155.555
Radial Basis Function	1.41	93.68	10.503



(a)



(b)

Figure 32. Metrics Performance of each model: (a) MAE, (b) R^2 .

Based on the above evaluation metrics, the performance of every model shows high accuracy with an R^2 above 90% except for Linear Regression with an R^2 of 61.14%. RF and DT outperformed the other models with an R^2 of 99.68% and 99.32%, respectively, with DT being much superior in the response time with only 0.728s, compared to 53.632s for RF. Another observation is DL models perform good accuracy, with an R^2 of 93.68% for RBF, and 91.07% for MLP. However, compared to tree-based models, especially DT, the accuracy and response time are still less, because DL models require a large number of neurons to learn the data and to minimize the loss function. The illustration of actual vs predicted NO_x emission is provided in **Figure 33**, and the testing simulation for each model is provided in **Figure 34**.

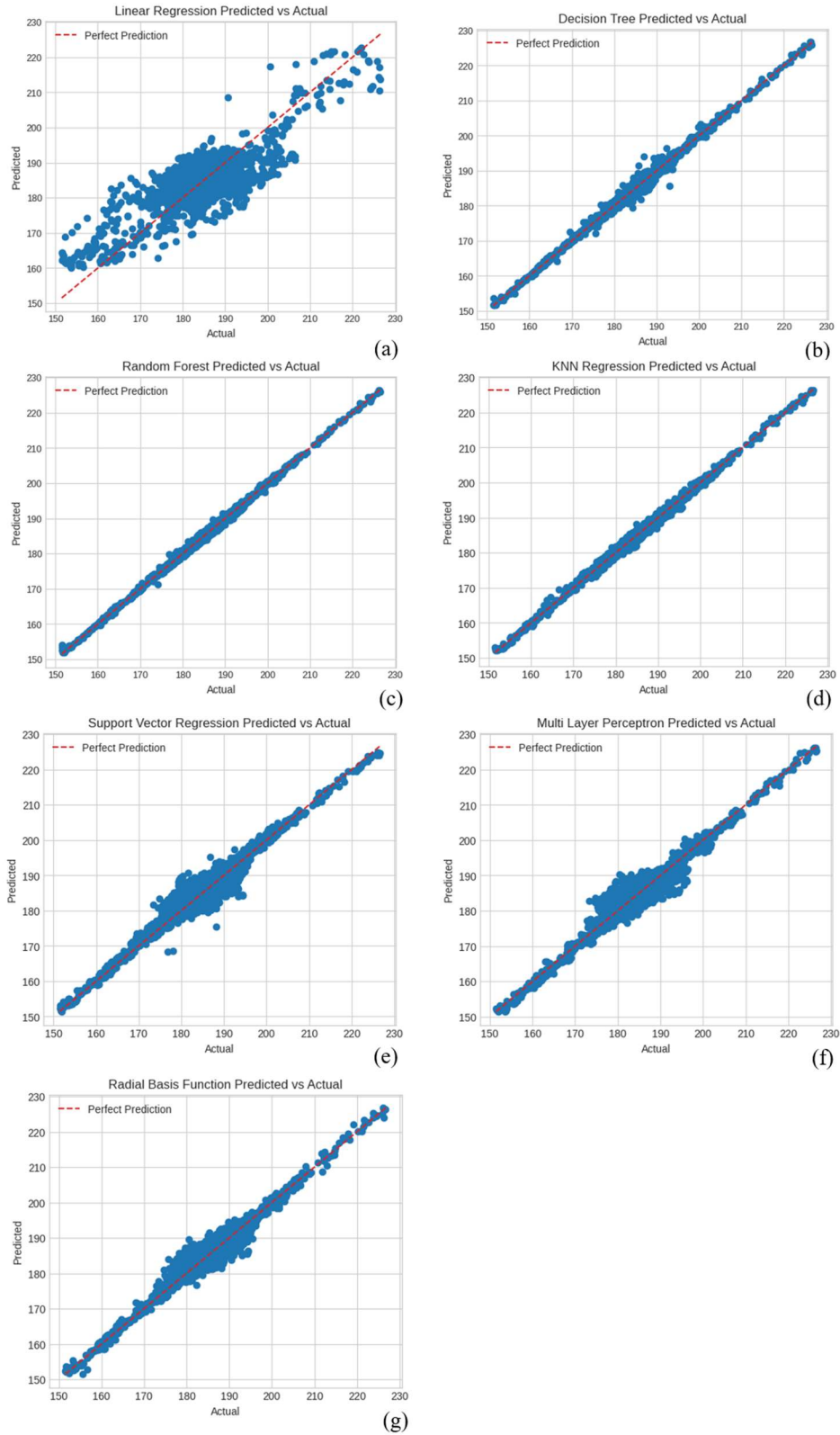


Figure 33. Predicted vs Actual NO_x value: (a) LR, (b) DT, (c) RF, (d) KNN, (e) SVR, (f) MLP, (g) RBF.

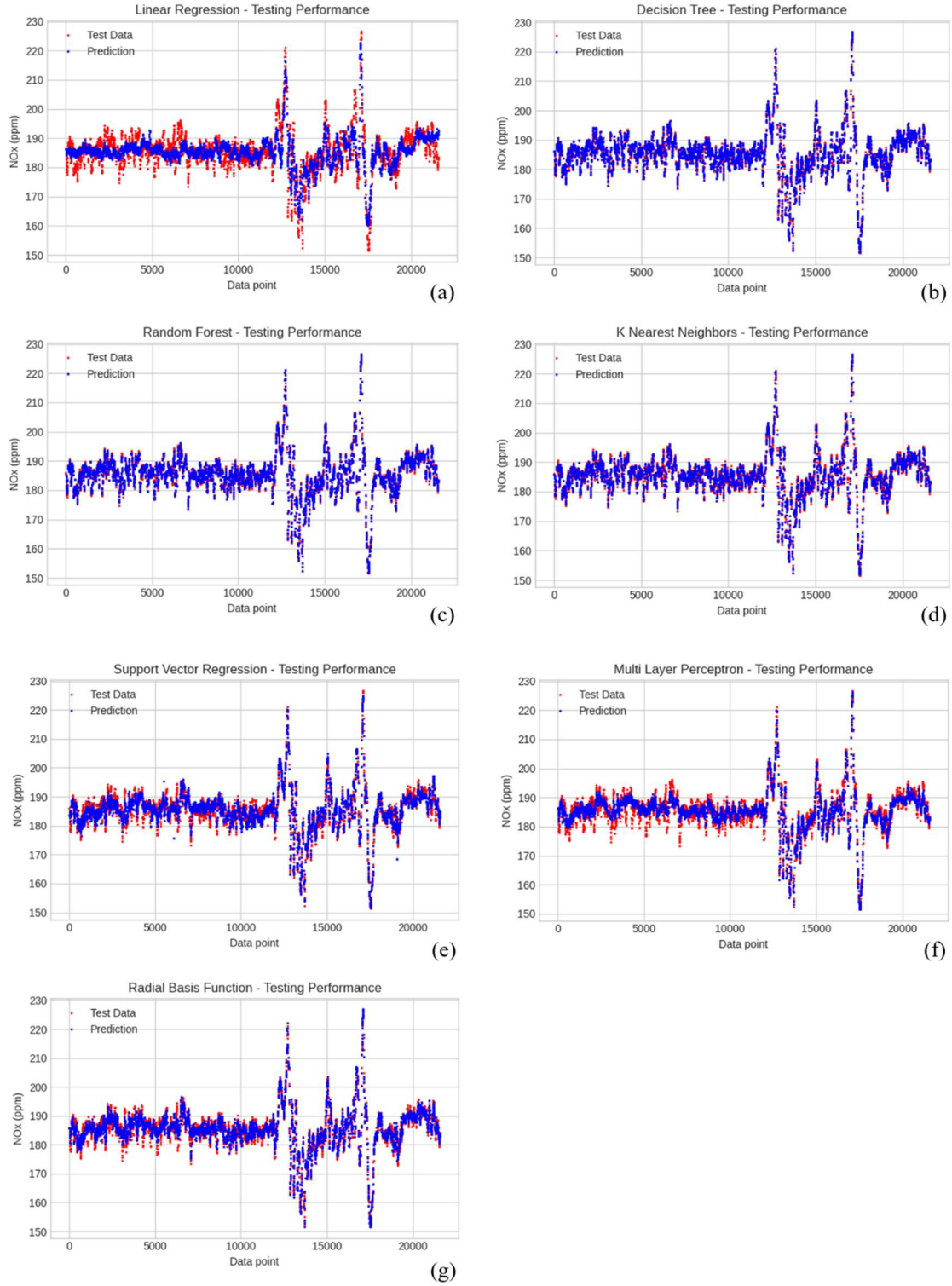


Figure 34. Testing Simulation for each model: (a) LR, (b) DT, (c) RF, (d) KNN, (e) SVR, (f) MLP, (g) RBF.

b. Models Performance Using Reduced Dataset

The second analysis is to develop models using the reduced dataset with 12 input parameters, and the performance metrics of all models are presented and compared in **Table 13**.

Table 13. Performance Metrics of NO_x Prediction Boiler NG Dataset (12 inputs).

Model ML/DL	MAE (ppm)	R ² (%)	Simulation Time (s)
Linear Regression	3.77	54.60	0.040
Decision Tree Regression	0.40	99.18	0.520
Random Forest	0.31	99.64	35.869
K Nearest Neighbor	0.49	99.30	0.600
Support Vector Regression	1.89	88.66	107.502
Multi Layer Perceptron	2.18	86.13	61.717
Radial Basis Function	1.39	93.67	5.463

By comparing **Table 12** and **Table 13**, it is shown that most of the models with the reduced dataset have less performance compared to the original datasets with an insignificant amount of difference, for example, MAE of DT increases from 0.39 to 0.40 ppm, and R² decreases from 99.32% to 99.18%, and for RF the MAE increases from 0.30 to 0.31 ppm, and R² decreases from 99.68% to 99.64%. The trend is the same for other models except for KNN where the accuracy increased from 99.25% to 99.30%. In terms of response time, all the models show the same improvement, which is that the model from the reduced dataset is faster than the original dataset. The comparison of testing simulation between actual and prediction using original and reduced datasets is presented in **Figure 35**.

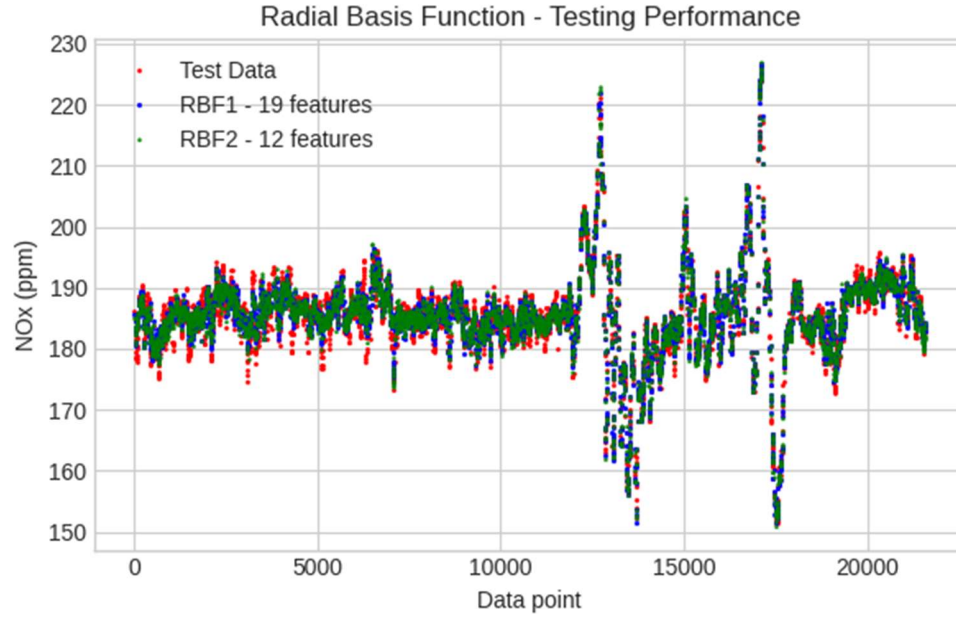


Figure 35. Thermal NO_x comparison from Boiler NG dataset using RBF neural networks between actual (red), prediction 19 inputs (blue), and prediction 12 inputs (green).

5.1.2 Efficiency Prediction

For Efficiency prediction, the same methodology as the NO_x emission prediction was also applied. The original input parameter is now 18 because the Flue Gas Temperature is excluded from the features as it has already been used to generate the output variable, which is efficiency, using the Siegert Formula, making this Flue Gas Temperature dependent on the output parameters. After performing data pre-processing, the data with a total of 21600 samples is split into a training set (80%), and a testing set (20%), and then the data is normalized using Standard Scaler before training the neural network models. The seven ML models and two DL models are again evaluated and compared, first using original input features, and the second using the reduced dataset. The illustration of the training and testing for the Efficiency dataset is provided in **Figure 36**.

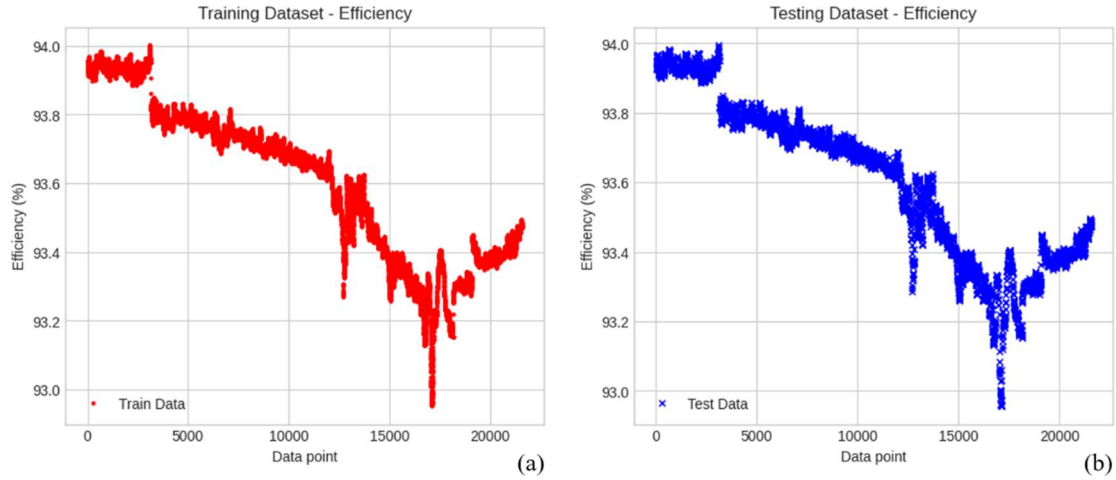


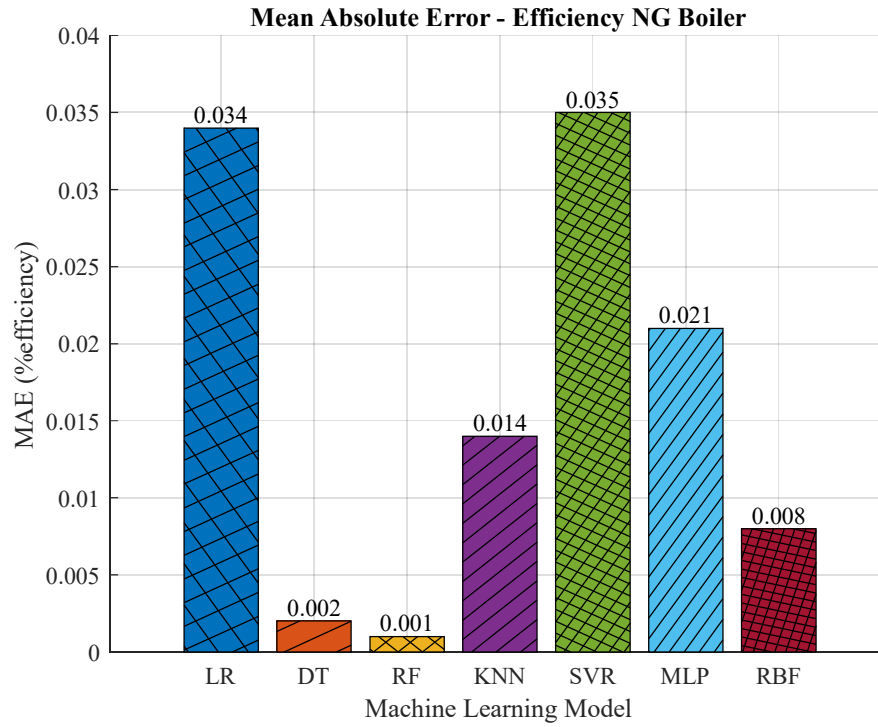
Figure 36. Training and Testing Datasets of Efficiency: (a) Training Set 80%, (b) Testing Set 20%

a. Models Performance Using Original Dataset

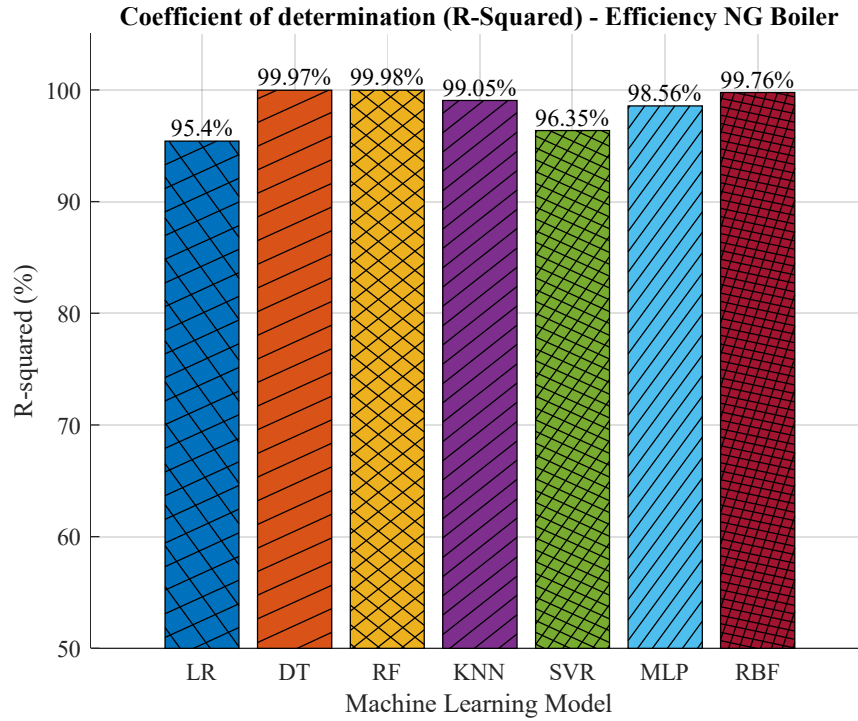
The performance summary of seven ML/DL models for Efficiency prediction in Natural Gas Boilers is depicted in **Table 14**, and a comparison metrics diagram of each model is provided in **Figure 37**.

Table 14. Performance Metrics of Efficiency Prediction Boiler NG Dataset (18 inputs).

Model ML/DL	MAE (%Efficiency)	R ² (%)	Simulation Time (s)
Linear Regression	0.034	95.40	0.124
Decision Tree Regression	0.002	99.97	0.573
Random Forest	0.001	99.98	41.136
K Nearest Neighbors	0.014	99.05	4.865
Support Vector Regression	0.035	96.35	0.190
Multi Layer Perceptron	0.021	98.56	12.430
Radial Basis Function	0.008	99.76	11.030



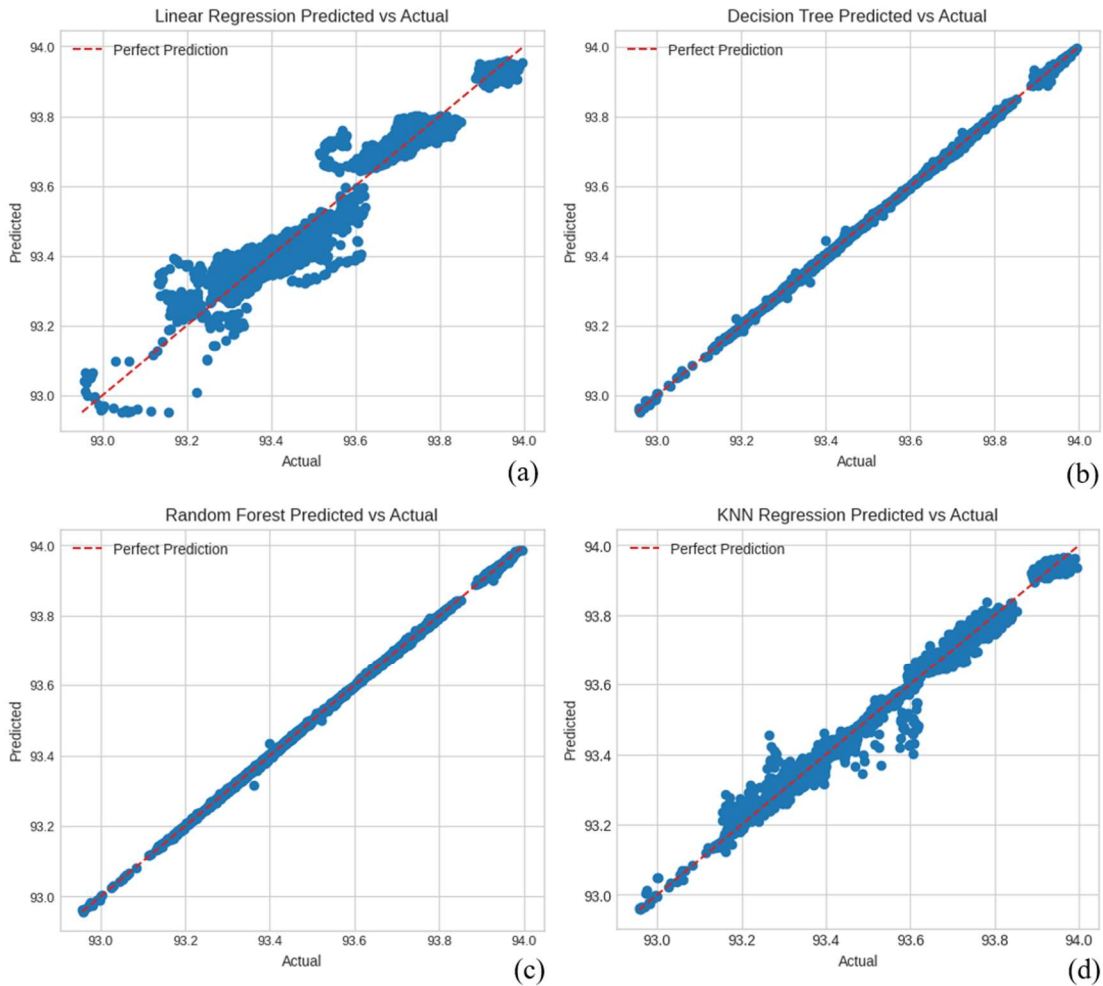
(a)



(b)

Figure 37. Metrics Performance of each model: (a) MAE, (b) R^2 .

As can be seen from the metrics above, all the prediction model shows high accuracy with an R^2 of above 95%, this again shows the benefit of having a large number of samples/data points making the models learn easily to pattern the data, and the inherent information in each variable. Here again, the tree-based model, RF and DT outperformed the other models with an R^2 of 99.98%, and 99.97, respectively, with DT having the fastest response time with only 0.573s, compared 41.136s for RF. The illustration of actual vs predicted Efficiency is provided in **Figure 38**, and the testing simulation for each model's performance is provided in **Figure 39**.



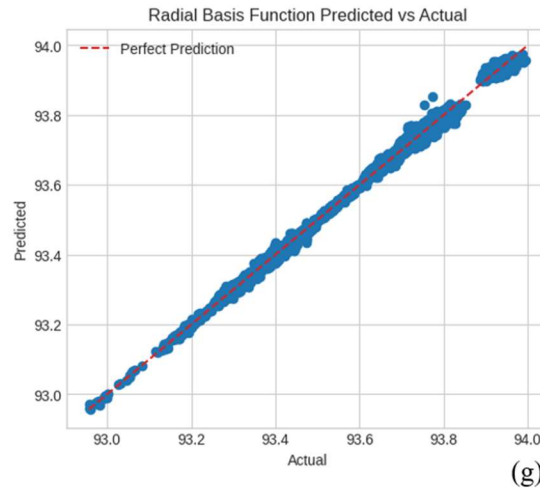
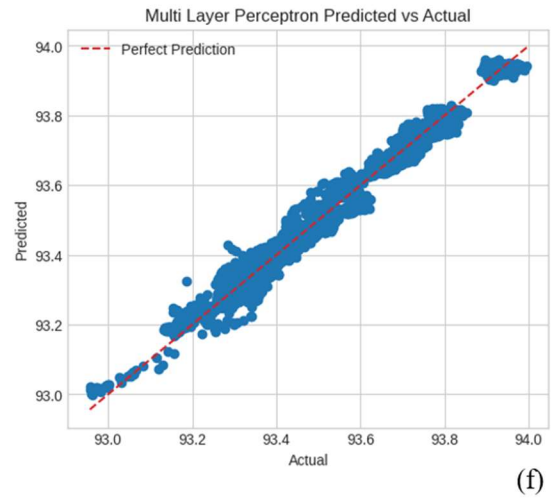
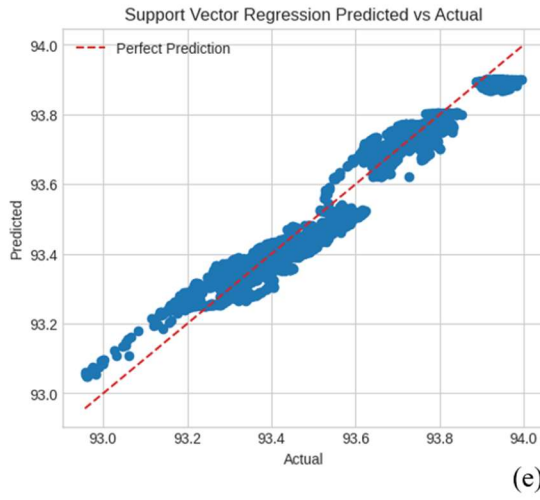
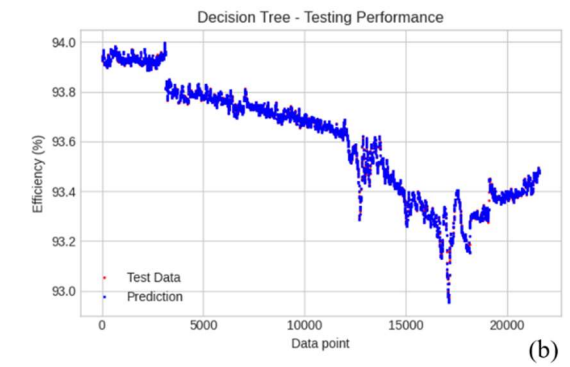


Figure 38. Predicted vs Actual Efficiency value: (a) LR, (b) DT, (c) RF, (d) KNN, (e) SVR, (f) MLP, (g) RBF.



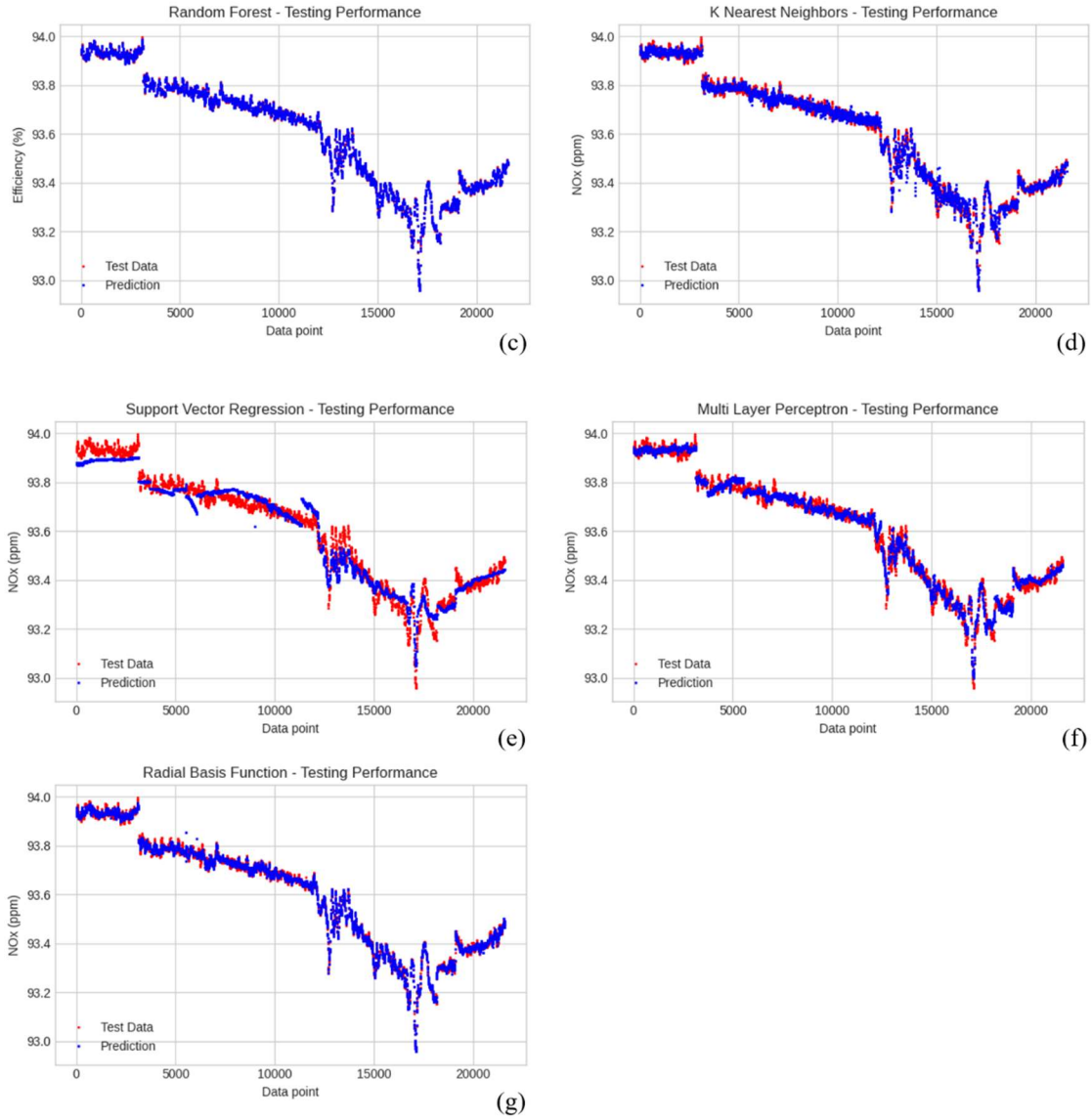


Figure 39. Testing Simulation for each model: (a) LR, (b) DT, (c) RF, (d) KNN, (e) SVR, (f) MLP, (g) RBF.

b. Models Performance Using Reduced Dataset

The second analysis uses a reduced dataset to develop ML/ML models for efficiency prediction, and the performance metrics of all models are presented and compared in **Table 15**.

Table 15. Performance Metrics of Efficiency Prediction Boiler NG Dataset (12 inputs).

Model ML/DL	MAE (%Efficiency)	R ² (%)	Simulation Time (s)
Linear Regression	0.036	94.47	0.078
Decision Tree Regression	0.002	99.97	0.424
Random Forest	0.002	99.98	27.158
K Nearest Neighbor	0.012	99.26	0.928
Support Vector Regression	0.033	96.55	0.158
Multi Layer Perceptron	0.024	97.89	6.997
Radial Basis Function	0.009	99.72	6.836

From **Table 14** and **Table 15**, it can be observed that the R^2 for all models is changing with a small amount of difference, except for DT and RF where the accuracy is steady. KNN and SVR accuracy increased from 99.05% to 99.26%, and from 96.35% to 96.55%. On the other hand, the response time is reduced by around 50% for all models. For example, the simulation time for RF was reduced from 41.136s to 27.158s. This concludes that with correct feature selection techniques, the model performance can be maintained while reducing the computational cost of the models. **Figure 40** shows the comparison of testing simulation between actual efficiency, predicted value using the original dataset, and predicted value using reduced datasets.

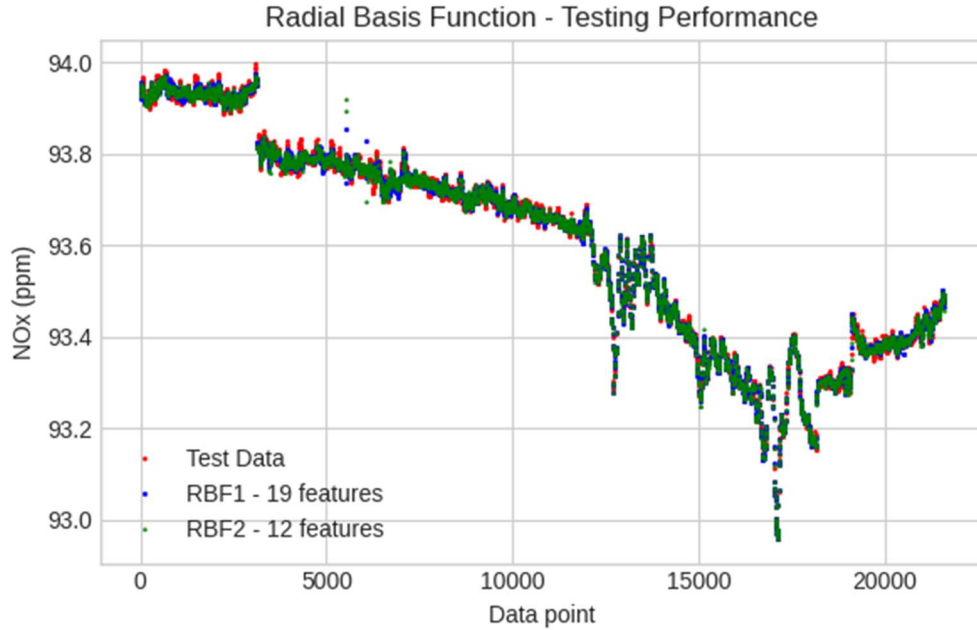


Figure 40. Efficiency comparison from Boiler NG dataset using RBF neural networks between actual (red), prediction 19 inputs (blue), and prediction 12 inputs (green).

5.2 Industrial Boiler Fueled with Hydrogen-Enriched Natural Gas

The second objective is to develop soft sensors to predict NO_x emissions based on CFD datasets. Two analyses using original and reduced datasets after feature selection are discussed in this section. The first analysis uses the original dataset that consists of 13 input variables to develop and compare the accuracy of the aforementioned AI-based algorithms, while the second analysis uses the reduced dataset with 8 inputs to study the effect of reducing the number of input parameters on model performance, by applying feature selection techniques.

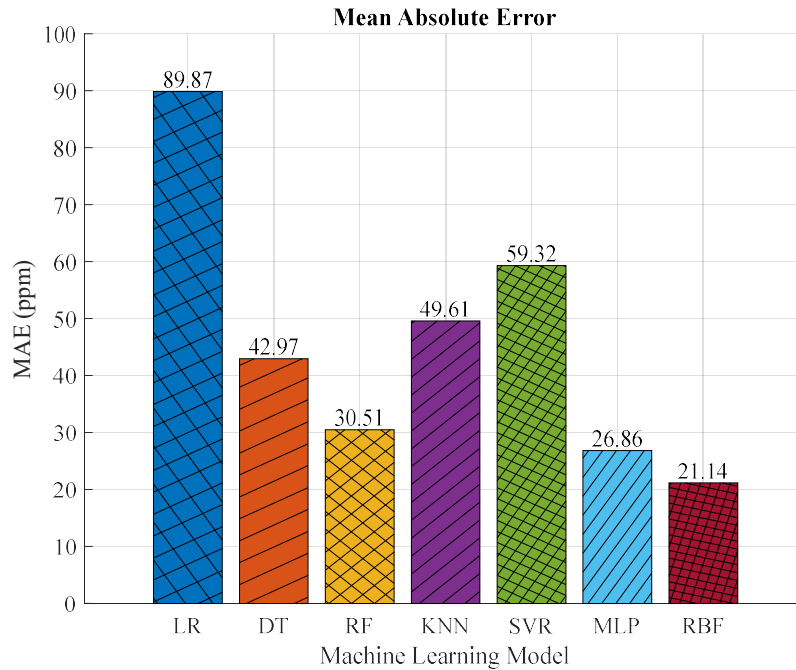
5.2.1 NO_x Emissions Prediction Using Original Dataset

Using the methodology mentioned in section 3.3 above, the seven AI models have been trained and tested using the original dataset of 13 input parameters. The complete

performance metrics summary is provided in **Table 16**, and the diagram of each model's performance is also compared in **Figure 41**. All the simulations were performed on 11th Gen Intel(R) Core(TM) i5-1135G7 @ 2.40GHz, 2419 Mhz, 4 Core(s), 8 Logical Processor(s), 8.00 GB RAM.

Table 16. Performance Metrics of Various ML/DL Models.

Model ML/DL	MAE (ppm)	PA (%)	Simulation Time (s)
Linear Regression	89.87	90.99	0.034
Decision Tree Regression	42.97	95.89	0.013
Random Forest	30.51	97.41	0.186
K Nearest Neighbor	49.61	95.30	0.021
Support Vector Regression	59.32	94.59	0.017
Multi Layer Perceptron	26.86	97.50	4.143
Radial Basis Function	21.14	98.08	0.076



(a)

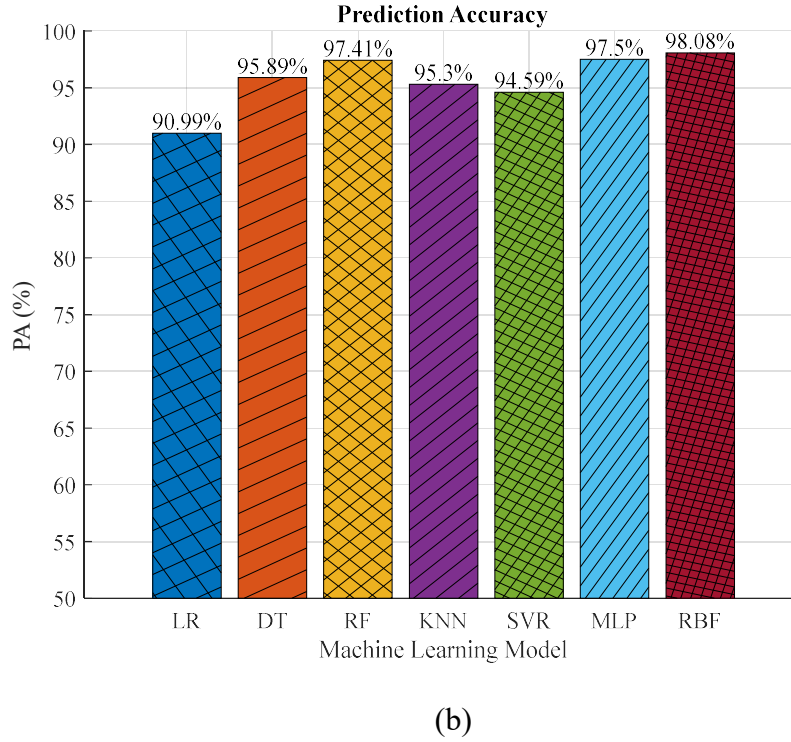
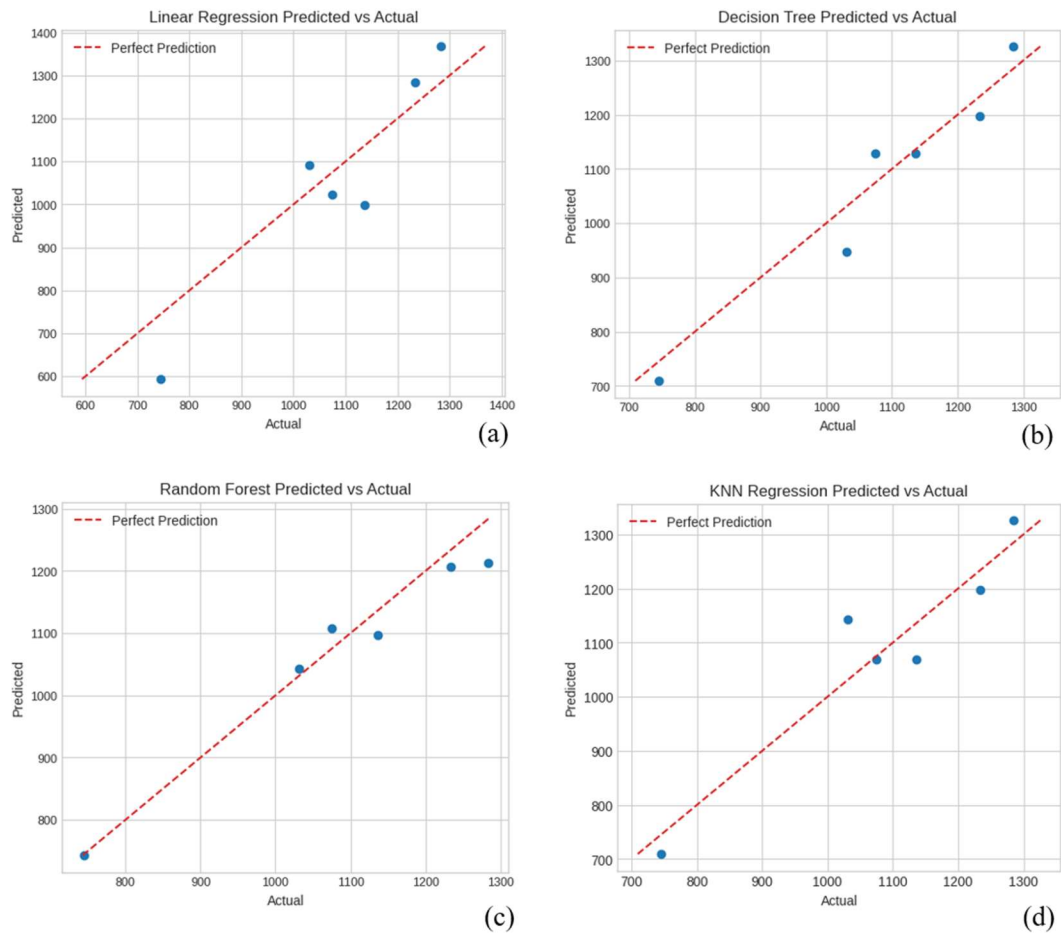


Figure 41. Metrics Performance of each model: (a) MAE, (b) PA.

From the above metrics, the RBF outperformed the other machine learning and deep learning models with the value of MAE 21.14 ppm, and PA 98.08%. Furthermore, MLP is the second-best model with performance values of MAE 26.86 ppm and PA 97.50%. In terms of simulation time, the RBF is 0.076s, which is 55 times faster compared to MLP with 4.143s, and 2.5 faster than RF with 0.186s. This implies that looking into overall performance, RBF shows superiority in the prediction of NO_x emission from hydrogen-enriched steam boilers compared to other ML/DL models. The superiority of RBF networks is supported by the simple structure and ability to perform localized basis functions around the centers, ensuring that neurons are active only for data points near these centers. For other neural network models, such as MLP, the model learns global approximations from all data points, which can make them less effective for solving complex problems. Additionally, RBF networks are less sensitive to hyperparameters

because selecting the number of neurons and their spread is relatively straightforward. On the other hand, MLP requires careful tuning of hyperparameters, such as the number of layers, activation functions, and learning rate, as these factors significantly impact model performance. Some studies have demonstrated the superiority of RBF neural networks amongst other models, including Iliyas et al. [34], Awadalla [65], Rahmati et al. [66], and Chang et al. [67].

The illustration of actual vs predicted NO_x is provided in **Figure 42** and the illustration of the testing simulation for each model's performance is provided in **Figure 43**.



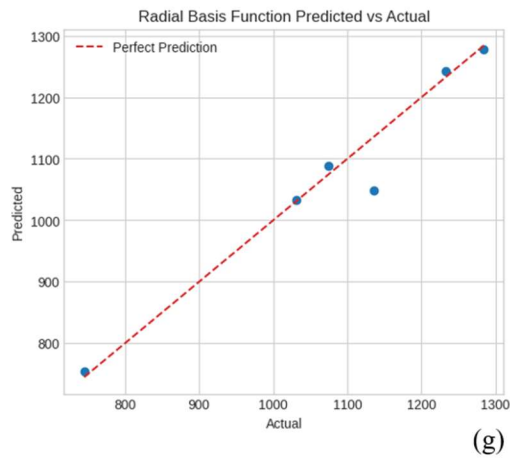
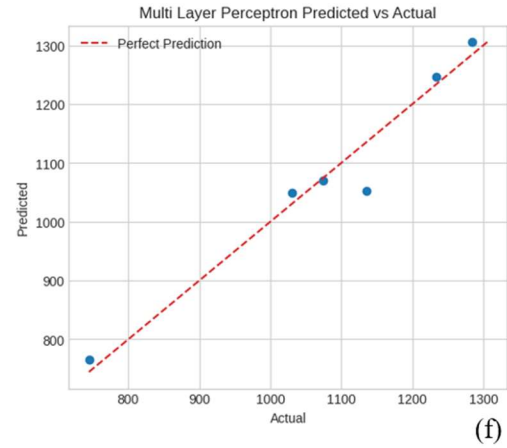
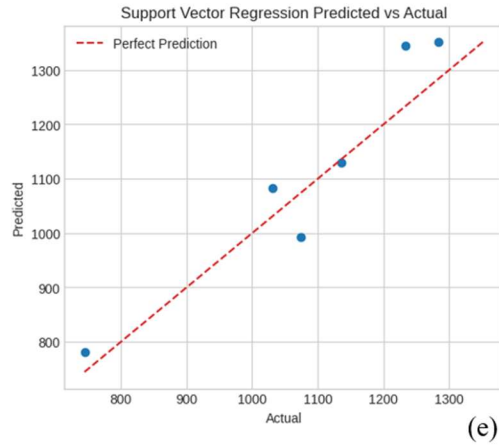
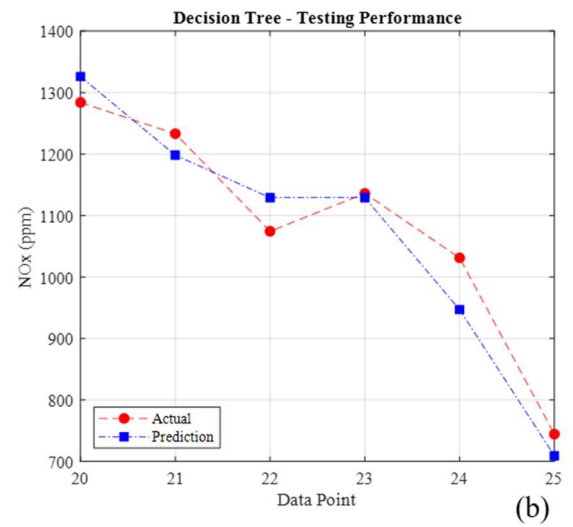
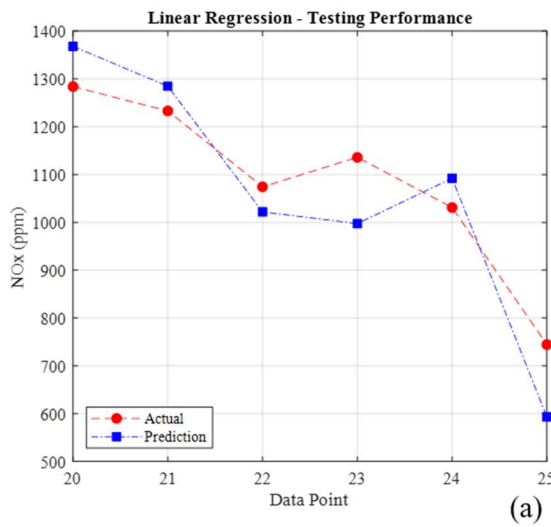


Figure 42. Predicted vs Actual NO_x value: (a) LR, (b) DT, (c) RF, (d) KNN, (e) SVR, (f) MLP, (g) RBF.



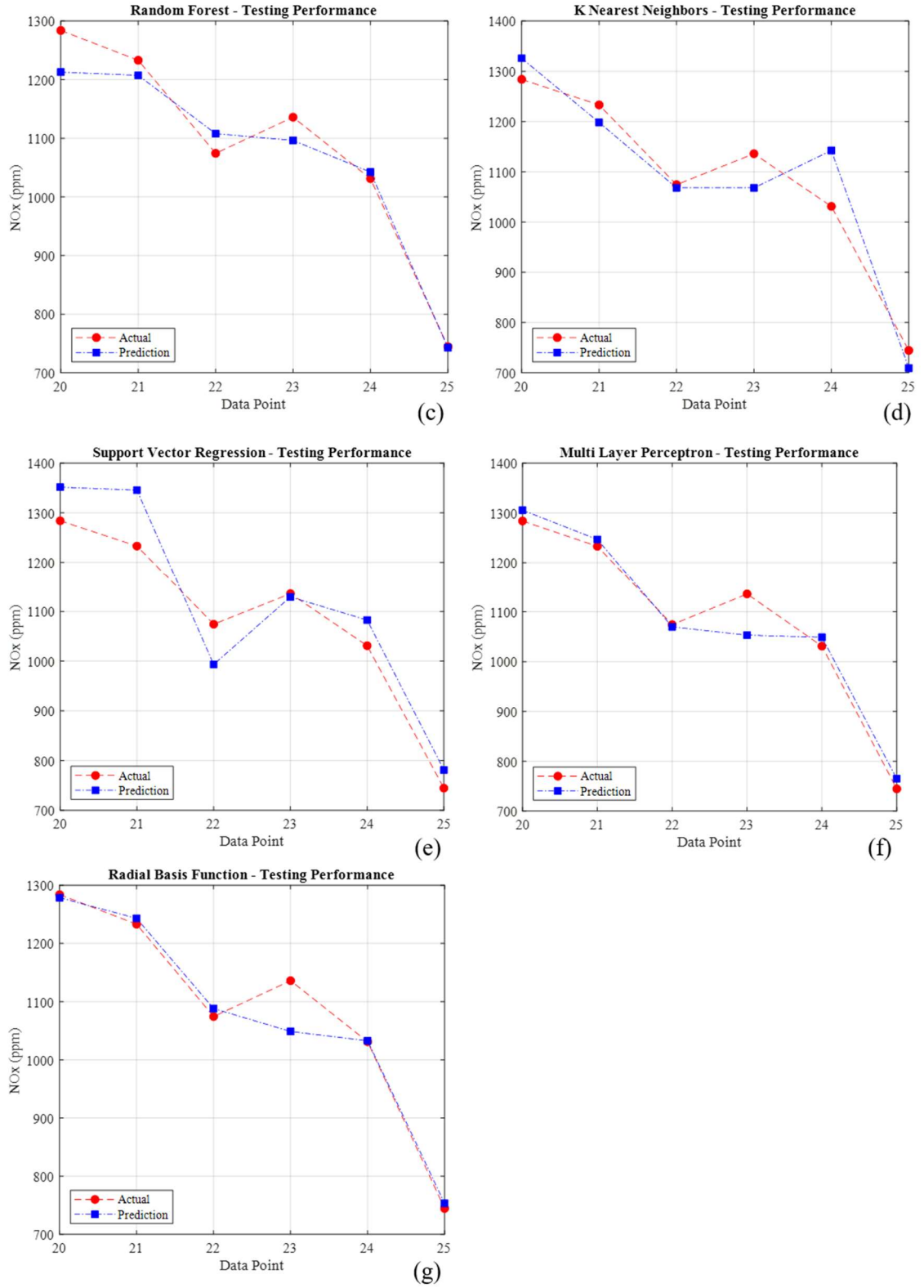


Figure 43. Testing Simulation: (a) LR, (b) DT, (c) RF, (d) KNN, (e) SVR, (f) MLP, (g) RBF.

5.2.2 NO_x Emissions Prediction Using Reduced Dataset

The second analysis is tested on the reduced dataset using the RBF neural network, which has proven to have the best performance compared to the other models. During the training simulation using the 8 input features, it is found that the best performance is obtained when the number of centers, $c = 10$ and the RBF width, $\sigma = 20$ for Thermal NO_x, while for the Prompt NO_x, the best performance is obtained when the number of centers, $c = 12$ and RBF width, $\sigma = 5$. The performance metrics of both models are presented and compared in **Table 17**.

Table 17. Performance Metrics of RBF with 13 input and 8 input features.

RBF-25 samples	MAE (ppm)	PA (%)	Simulation Time (s)
Thermal NO_x			
Model-1 (13 input)	21.14	98.08	0.076
Model-2 (8 input)	24.29	97.81	0.021
Prompt NO_x			
Model-1 (13 input)	0.65	94.82	0.069
Model-2 (8 input)	0.70	95.70	0.027

As shown in Table 17, the Thermal NO_x prediction for the second model with 8 input features yields 97.81% PA, with a 0.021s simulation time, compared to 98.08% PA/0.076s for the first model with 13 input features. For the Prompt NO_x prediction, the second model yields a PA of 95.70%, with a simulation time of 0.027s, compared to the 94.82% PA/0.069s in the first model. The reduction of the PA is due to some information that has been removed from the original 13 input features and relies only upon information available in the new 8 input variables. This concludes that the second model with 8 input

features has comparable performance accuracy to the first model with 13 input features. However, the simulation time for the second model is faster than the first model. For example, the simulation time for Thermal NO_x prediction with 8 inputs is 3.6 times faster than the model with 13 inputs, while for Prompt NO_x prediction, the simulation time is 2.6 times faster. This implies that the feature selection methods work effectively in this study. The comparison of the testing output of the RBN neural network between actual and prediction using 13 and 8 input features is presented in **Figures 44** and **Figure 45**, where Figure 44 is for Thermal NO_x prediction, and Figure 45 is for Prompt NO_x prediction. From these Figures, it can be seen that predicted values of Thermal NO_x and Prompt NO_x fit well in almost all of the data points for both using both models with 13 and 8 input features, except for data point no. 23 in Thermal NO_x, where the model is unable to generalize well the target.

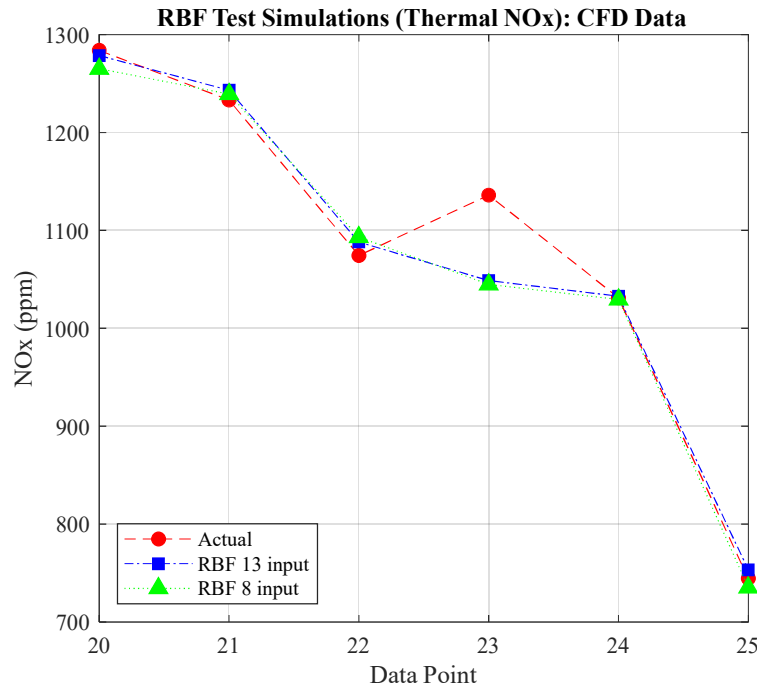


Figure 44. Thermal NO_x comparison of RBF neural networks between actual (red), prediction 13 inputs (blue), and prediction 8 inputs (green).

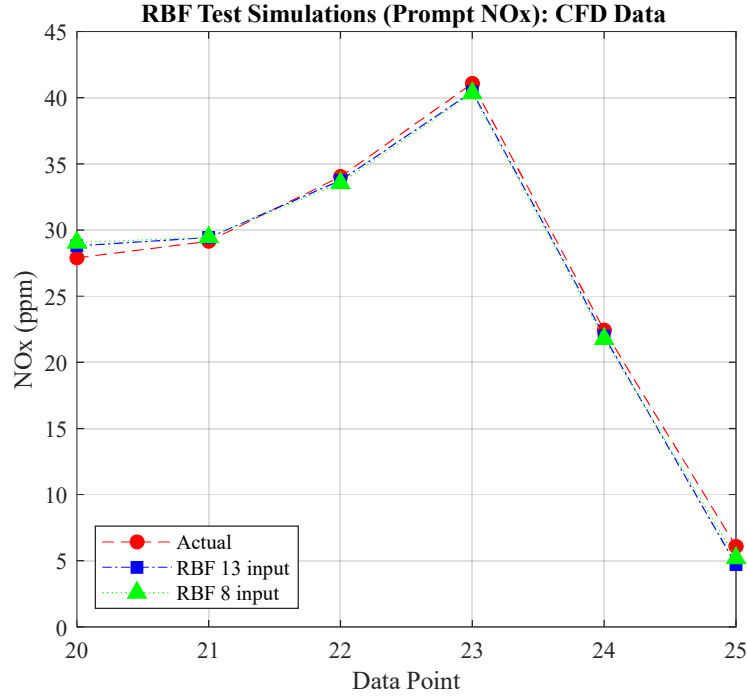


Figure 45. Prompt NO_x comparison of RBF neural networks between actual (red), prediction 13 inputs (blue), and prediction 8 inputs (green).

5.2.3 NO_x Emissions Prediction Using Augmented Dataset

The augmented dataset with a total of 250 data points is used to train and test the new RBF with the same structure as the above models. The data is split into 190 samples (75%) as a training set, and 60 samples (25%) as a testing set. Following the same procedure in section 3.3, the result of the training and testing simulation for both Thermal NO_x and Prompt NO_x emission prediction is depicted in **Table 18**. It is found that the new RBF models with augmented data have comparable accuracy with original CFD data. For example, the PA for Thermal NO_x decreased from 98.08% to 97.84% (-0.24%) for Model-1 with 13 input features, and from 97.81% to 97.08% (-0.73%) for Model-2 with 8 input features. For the Prompt NO_x, the PA increased from 94.82% to 97.28% (+2.46%) for Model-1 and decreased from 95.70% to 93.91% (-1.79%) for Model-2. In general, these

show the reliability of the built RBF Models by having error deviations around 0.24% to 2.46%. Furthermore, it also can be observed from the Augmented Data table, that simulation times of Model-2 are 5 times faster compared to Model-1 for both Thermal NO_x and Prompt NO_x, indicating the effectiveness of the Features Selections methods mentioned in section 3.2. The illustration of the Testing Simulations of the RBF Models using Augmented Data are also presented in **Figure 46** and **Figure 47**, where red points represent the augmented data, while blue and green points represent predicted values based on 13 and 8 input features, respectively.

Table 18. RBF Performance using Augmented Data.

RBF-250 samples	MAE (ppm)	PA (%)	Simulation Time (s)
Thermal NO_x			
Model-1 (13 input)	22.63	97.84	0.060
Model-2 (8 input)	30.84	97.08	0.011
Prompt NO_x			
Model-1 (13 input)	0.57	97.28	0.061
Model-2 (8 input)	0.73	93.91	0.012

For comparison, the traditional MLP models are also built using the augmented dataset. The performance of MLP models for both Thermal NO_x and Prompt NO_x emission prediction is shown in **Table 19**, and the illustration of the testing simulations of the MLP models using an augmented dataset is also presented in **Figure 48** and **Figure 49**. As can be seen, the MLP models have less accuracy compared with the RBF models, with significant differences in simulation time. This concludes the superiority of RBF models in terms of accuracy and response time because the RBF models have a simple structure

with only one hidden layer. As discussed in Section 5.2.1, RBF networks are advantageous due to their ability to utilize localized basis functions centered around specific points, ensuring that neurons are activated only for nearby data points. While the other models, like MLP rely on global approximations across all data points. Furthermore, RBF networks are less dependent on hyperparameter tuning, while the MLP requires meticulous adjustment of hyperparameters to have a great impact on their performance.

Table 19. MLP Performance using Augmented Data.

MLP-250 samples	MAE (ppm)	PA (%)	Simulation Time (s)
Thermal NO_x			
Model-1 (13 input)	31.84	96.87	19.044
Model-2 (8 input)	37.77	96.01	16.876
Prompt NO_x			
Model-1 (13 input)	0.92	94.59	16.350
Model-2 (8 input)	0.94	95.09	16.187

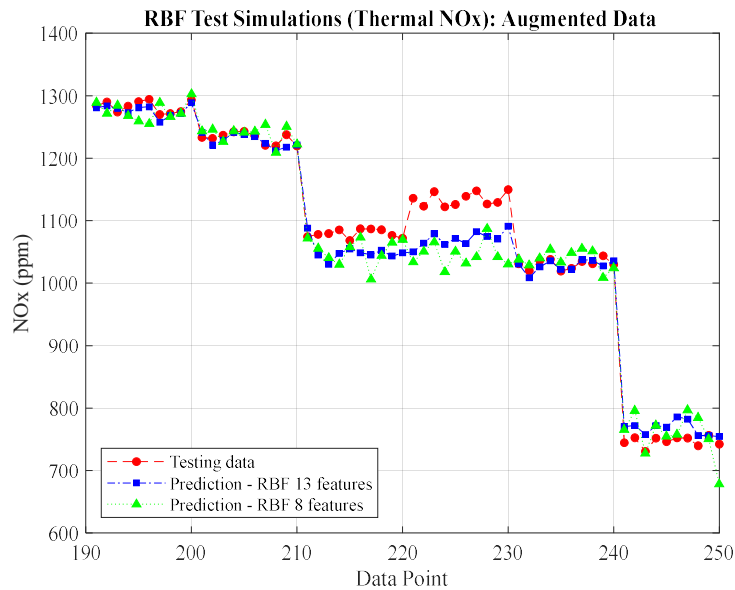


Figure 46. RBF test simulation for Thermal NO_x: augmented data (red), prediction 13 input features (blue), and prediction 8 input features (green).

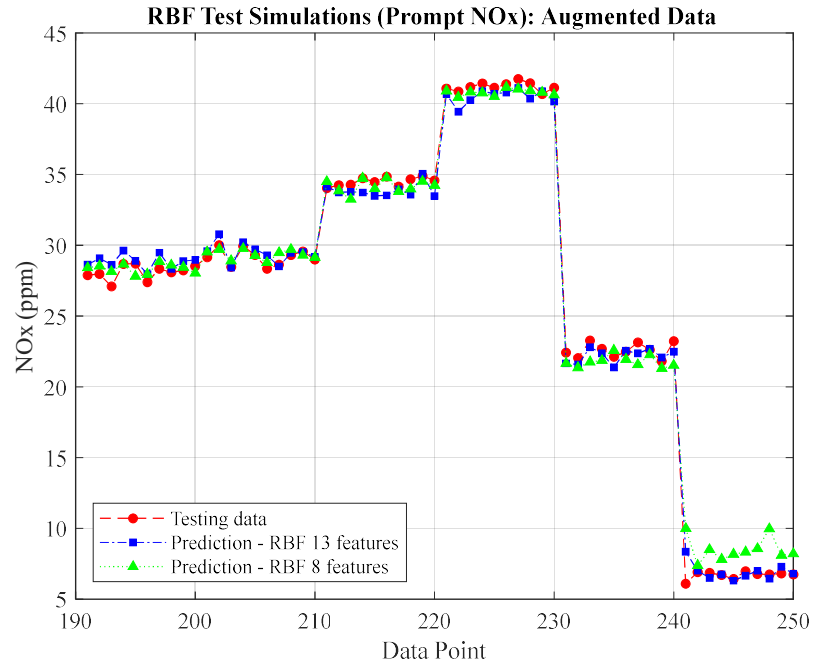


Figure 47. RBF test simulation for Prompt NO_x: augmented data (red), prediction 13 input features (blue), and prediction 8 input features (green).

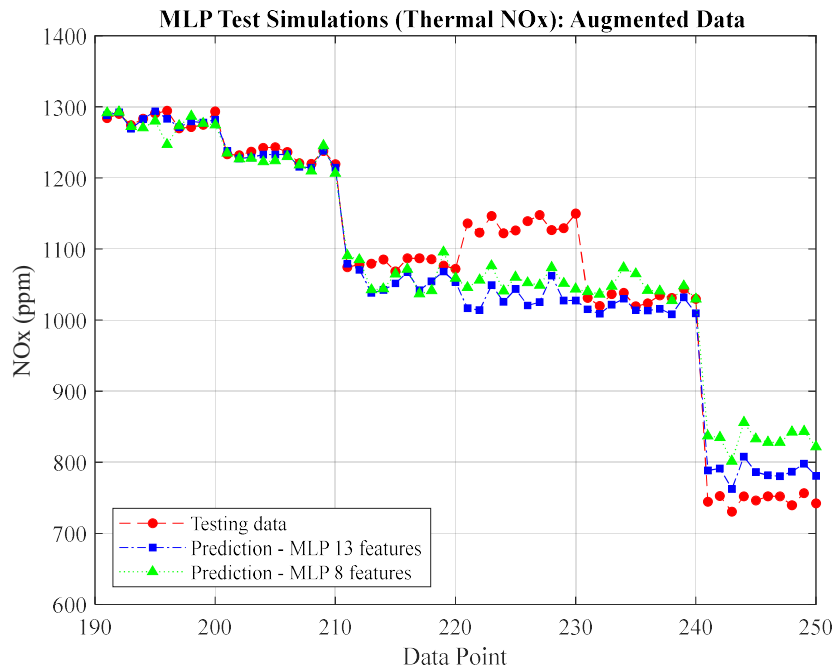


Figure 48. MLP test simulation for Thermal NO_x: augmented data (red), prediction 13 input features (blue), and prediction 8 input features (green).

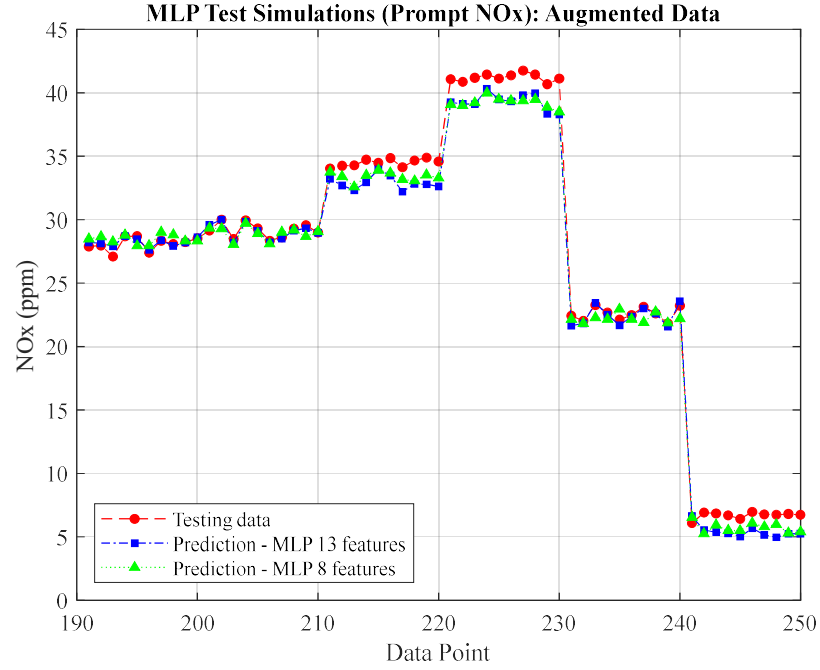


Figure 49. MLP test simulation for Prompt NO_x: augmented data (red), prediction 13 input features (blue), and prediction 8 input features (green).

5.3 Fault Prognosis

In this section, fault prognosis to predict boiler swing conditions based on various time steps is discussed. The dataset is based on the NG Boiler, employing 12 parameters used in NO_x emission prediction models, where the Steam Flow Rate is chosen as the output variable, while the other 11 variables are the input. The reason for choosing the Steam Flow Rate as the output variable is that the dynamic variations of all boiler parameters are due to the disturbance that occurs in the steam flow rate [9]. LSTM is chosen for this study due to the capability of this algorithm to store memory of the previous output in the LSTM cell, thus enhancing accuracy from the historical data, making it appropriate for time-dependent problems.

Before training the model, data pre-processing is required to define input and output data. Unlike the above works that utilize regression tasks to predict the number of NO_x and Efficiency, here the fault prognosis aims to perform a classification task in the output, thus the Steam Flow Rate behavior is classified into “normal” and “fault” conditions. The normal condition is when the Steam Flow Rate operates in the steady state condition, while the fault is when the Steam Flow Rate operates in the dynamic condition. In this study, the Steam Flow Rate is considered normal when operating under 400 klbs/hr, so the boiler swings happens when the fuel flow rate operates above this limit. The illustration Steam Flow Rate and the corresponding state condition are provided in **Figure 50**, and **Figure 51**. The blue line represents the normal condition (class 0), and the red line represents the fault condition (class 1).

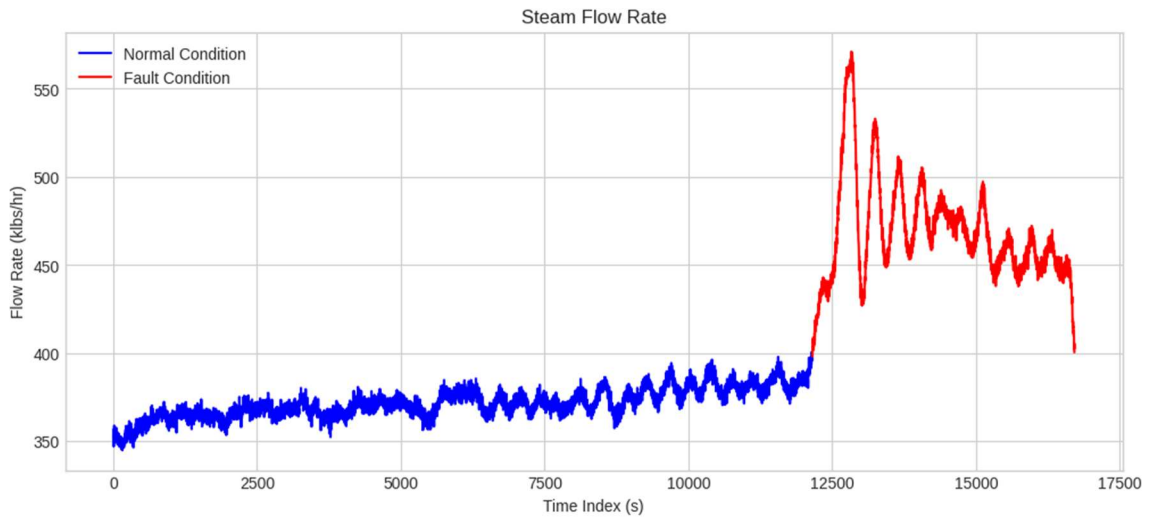


Figure 50. Steam Flow Rate in normal condition (blue), and fault condition (red).

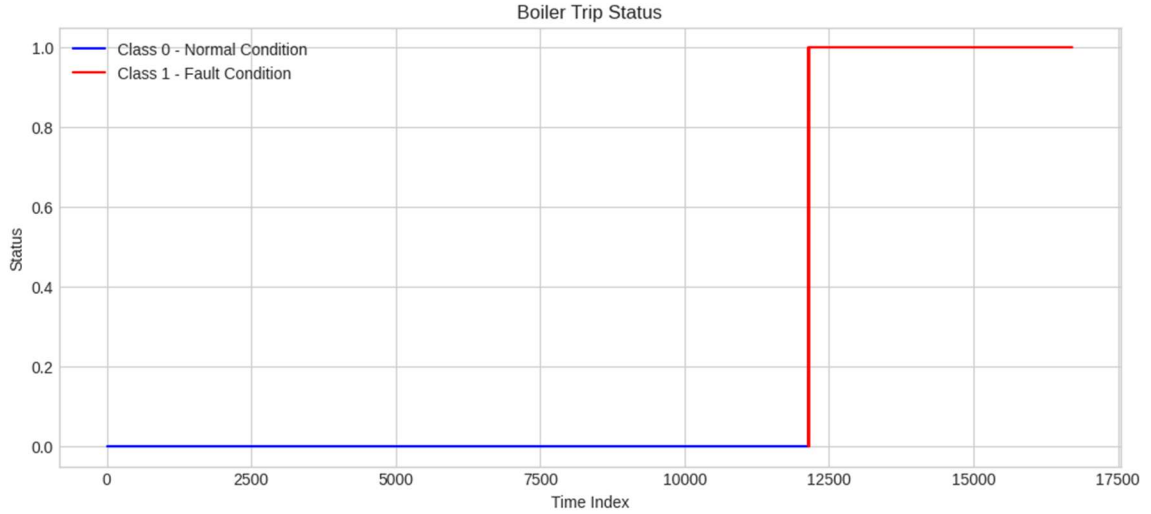


Figure 51. State classification of the Steam Flow Rate: Blue is class 0, Red is class 1.

Another data pre-processing process is required to define the input parameters. To make the model capable of predicting the event based on previous time steps, then the input parameter needs to be transformed into sequential data containing the required time steps. Here the time steps are experimented from 10s to 60s, to find the best accuracy of the model. The illustration of data transformation is described in **Figure 52**, and all this data transformation is applied for all 11 parameters because here we perform multivariate fault prognosis.

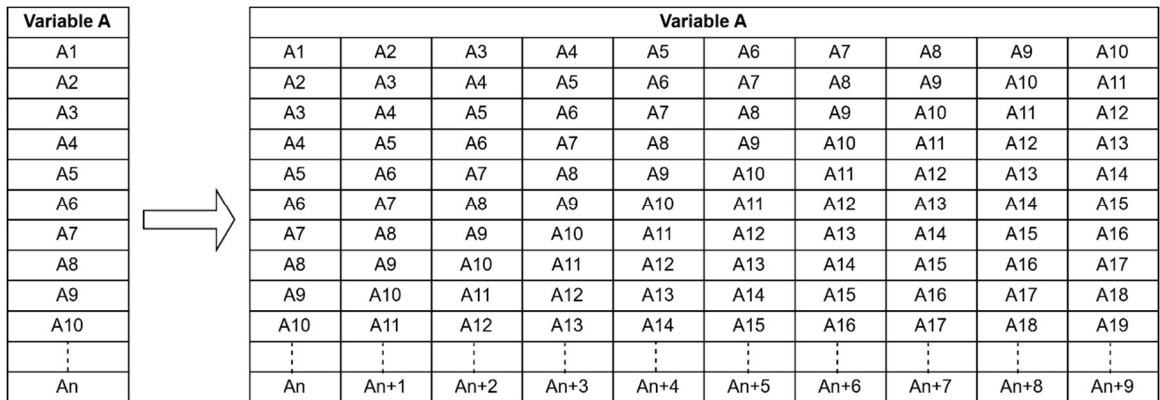


Figure 52. Illustration of data transformation into sequential for LSTM model.

First, the Model with 10-second time steps is performed and evaluated. After the input and output parameters have been defined, the next step is to normalize the data using Standard Scaler, and then train the data with LSTM. During the training, the input data is divided into Training and Validation Data with a portion of 80%-20%. LSTM is trained with 20 epochs, and the Training and Validation loss is provided in **Figure 53**.



Figure 53. Training and Validation loss of LSTM with 10s time steps.

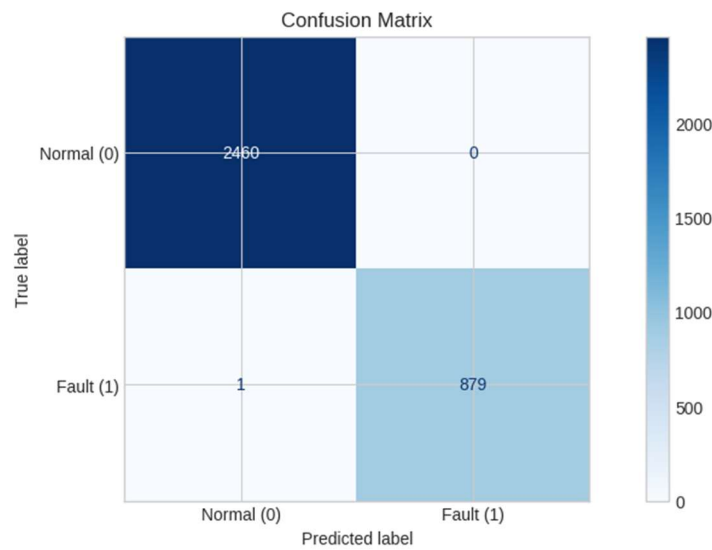


Figure 54. Confusion Matrix of LSTM with 10s time steps.

The model yields an accuracy of 99.96% for validation, and to understand more of the model performance, the Confusion Matrix is provided in **Figure 54**. From this matrix, it can be observed that only one miss-classification for the validation data, while the rest of the data is classified correctly as the actual data. Another representation to describe the model performance is to compare actual and predicted data along the actual time frame, as can be seen in **Figure 55**.

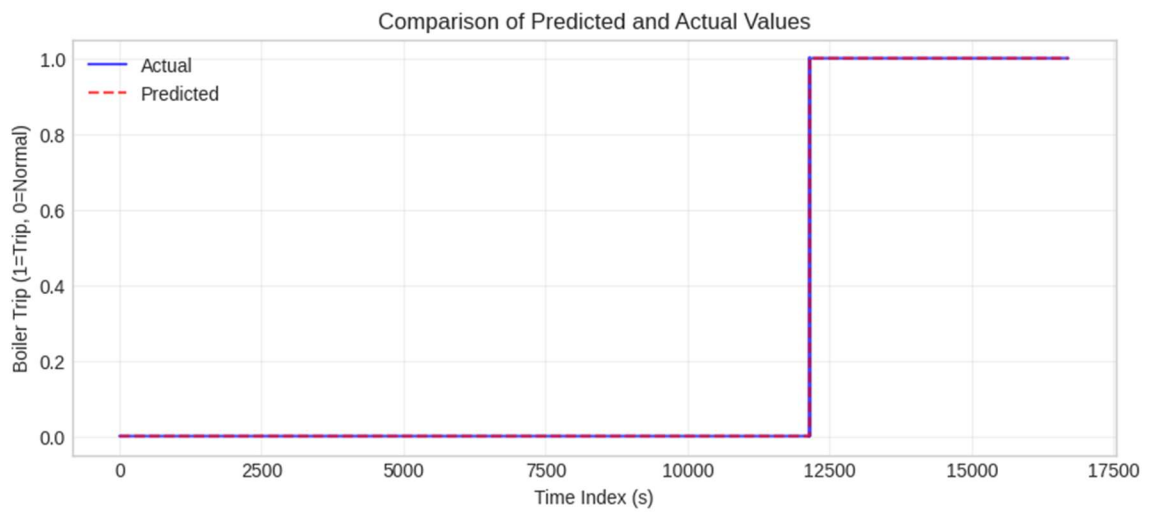


Figure 55. Comparison of actual and predicted data in the time frame boiler operation.

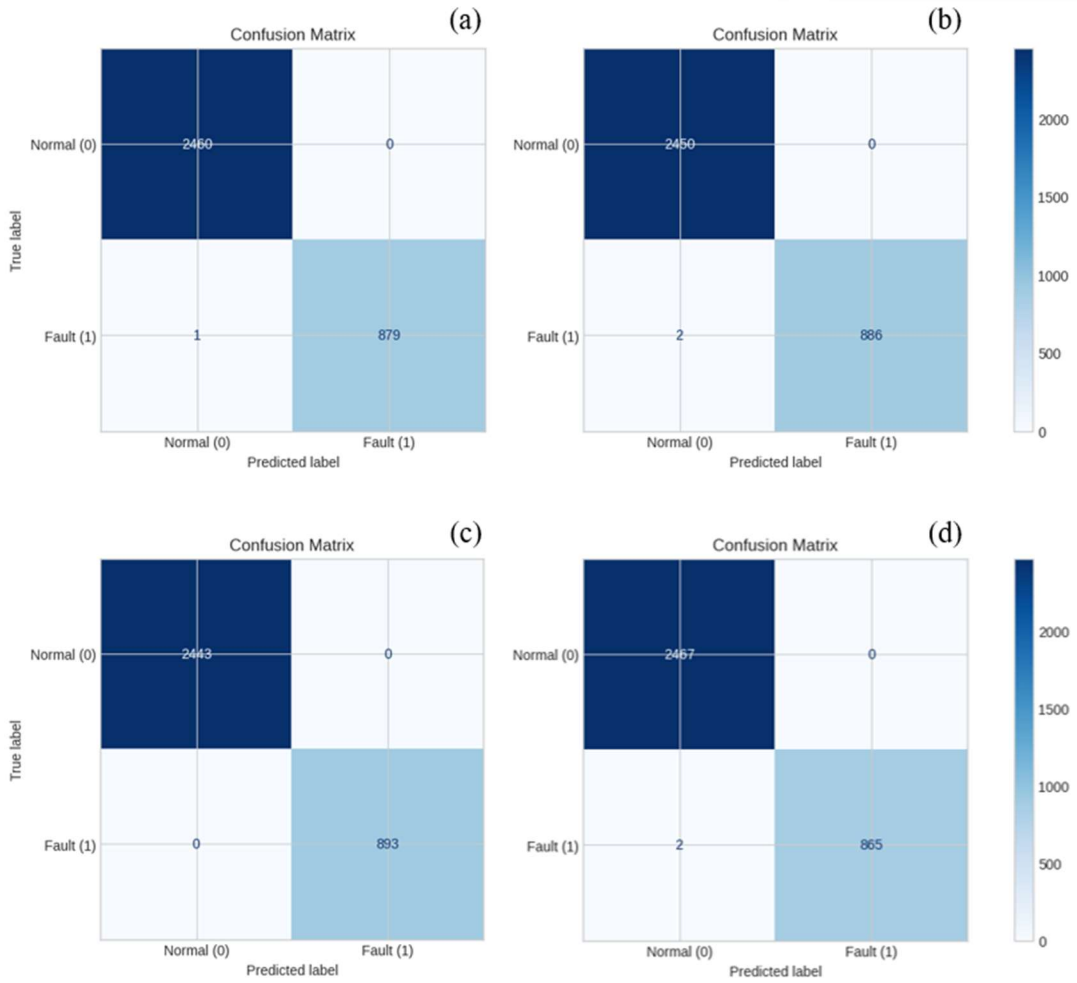
The next step is to investigate the model performance by varying time steps, starting from 10s, 20s, 30s, 40s, 50s, and 60s. The accuracy of each model is depicted in the **Table 20**.

Table 20. Test Accuracy of each LSTM Model.

Time Steps	Model Accuracy (%)
LSTM-10s	99.97
LSTM-20s	99.94
LSTM-30s	100.00

LSTM-40s	99.94
LSTM-50s	99.91
LSTM-60s	100.00

As can be seen from the above table, the LSTM performed well in all cases with an accuracy of more than 99.9%, and the best accuracy is obtained when the time step is 20s and 50s. The comparison of the Confusion Matrix is provided in **Figure 56**.



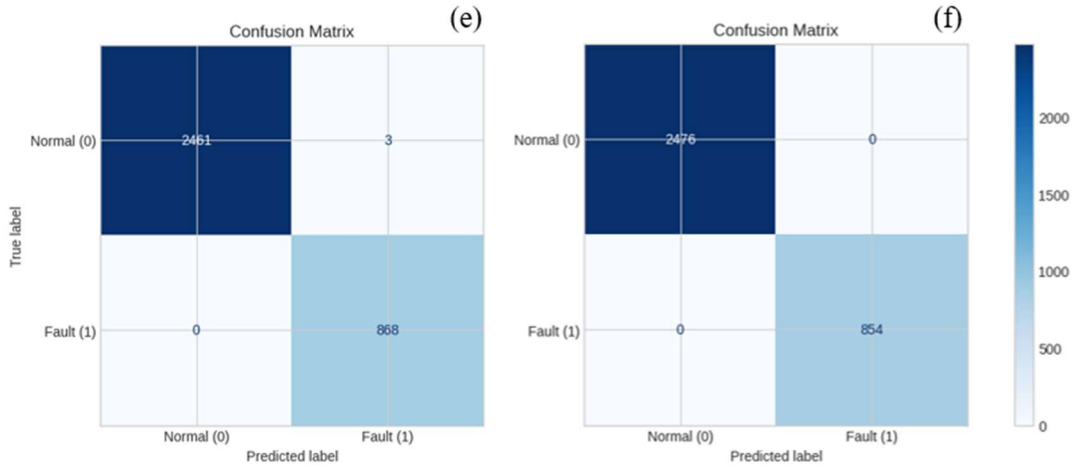


Figure 56. Comparison of Confusion Matrix of all time steps.

To see the model performance in the actual case, the simulation compares actual and predicted values in the following figures using the 60s time steps simulation. **Figure 57** shows the prediction of the first 20 data, **Figure 58** presents the model performance to the last 20 observation data, and **Figure 59** provides the prediction and actual data during the transition from normal to fault. It is clearly shown that during the steady state and the dynamic condition, the model can predict accurately the state condition, however, during the transition, there are some points that the model misclassified the state condition, here there are 6 points miss-classified, which is acceptable, because the fault is still on early stage, giving the chance to the operator to take the necessary action so that the potential damage of the boiler can be avoided.

```

# Simulations
print("Prediction during trip event")

print("\nIndex:",start_idx, "to index:",end_idx )

print("\nActual Outputs (Y):")
print(labels_original) # True labels (actual outputs)

print("\nPredicted Classes:")
print(predictions_class.flatten()) # Predicted class labels (binary 0 or 1)

1/1 ————— 0s 29ms/step
Prediction during trip event

Index: 0 to index: 20

Actual Outputs (Y):
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]

Predicted Classes:
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]

```

Figure 57. Model performance to the first 20 data.

```

# Simulations
print("Prediction during trip event")

print("\nIndex:",start_idx, "to index:",end_idx )

print("\nActual Outputs (Y):")
print(labels_original) # True labels (actual outputs)

print("\nPredicted Classes:")
print(predictions_class.flatten()) # Predicted class labels (binary 0 or 1)

1/1 ————— 0s 33ms/step
Prediction during trip event

Index: 16630 to index: 16650

Actual Outputs (Y):
[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]

Predicted Classes:
[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]

```

Figure 58. Model performance to the last 20 data.

```

# Simulations
print("Prediction during trip event")

print("\nIndex:",start_idx, "to index:",end_idx )

print("\nActual Outputs (Y):")
print(labels_original) # True labels (actual outputs)

print("\nPredicted Classes:")
print(predictions_class.flatten()) # Predicted class labels (binary 0 or 1)

```

1/1 ————— 0s 46ms/step
Prediction during trip event

Index: 12075 to index: 12105

Actual Outputs (Y):
[0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 1 0 0 1 1 0 0 1 1 1 1 1 1 1 1]

Predicted Classes:
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]

Figure 59. Model performance during transition from normal to swings.

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusions

This study demonstrated the effectiveness of utilizing artificial intelligence-based soft sensors to predict NO_x emission, efficiency, and fault prognosis in a steam boiler. By leveraging advanced machine learning (ML) and deep learning (DL) models, the research paves the way for intelligent, efficient, and sustainable boiler operations. The key findings are summarized below:

- For NO_x emission and efficiency prediction in industrial boilers fueled with natural gas, the results show that tree-based machine learning models, particularly decision tree (DT) and random forest (RF), achieved exceptional accuracy. The R² values for NO_x emission prediction were 99.32% (DT) and 99.68% (RF), while for efficiency prediction, they reached 99.97% (DT) and 99.98% (RF). Additionally, DT outperformed RF in computational efficiency, delivering simulation time in just 0.728s compared to 53.632s for RF in NO_x emissions prediction, and 0.573s (DT) versus 41.136s (RF) in efficiency prediction. This highlights the potential of DT-based soft sensors for real-time applications, where speed and accuracy are critical for NO_x emissions and process optimization in the industry.
- For NO_x emission prediction in industrial boilers fueled with hydrogen-enriched natural gas, results showed that the radial basis function (RBF) model outperformed others, achieving a prediction accuracy (PA) of 98.08% compared to multi-layer

- perceptron (MLP): 97.50%, with computational speed 55 times faster than that of MLP, making RBF an attractive option for hydrogen-enriched fuel applications. Validation on augmented datasets confirmed the consistency and reliability of these models under varying conditions, which shows promising results for clean energy source applications. The RBF outperformed the other models due to its simple structure and ability to perform localized basis functions around the centers, thus making it suitable for a limited no of datasets with high variability such as the H₂-enriched dataset.
- Fault prognosis to detect boiler swings using various time steps was also successfully investigated in this study using the long short-term memory (LSTM), and the models worked well to detect the fault based on multivariate boiler parameters. Results showed that the LSTM with the 30s and 60s time steps had 100% accuracy for the testing dataset. This demonstrates the viability of neural networks in accurately detecting boiler swings and preventing catastrophic failures (i.e. preventive maintenance), even under complex multivariate operating conditions.
 - This study also showed an effective feature selection process by combining data-driven techniques such as variance threshold, Pearson correlation, and feature importance, with domain knowledge. Feature selection based on domain knowledge significantly reduced simulation times without compromising accuracy. For instance, RF models achieved similar R^2 values (99.68% vs. 99.64%) but were 1.5 times faster when using reduced datasets. This demonstrates the value of efficient data preprocessing in industrial AI applications in utilizing the least number of features to achieve the most accurate result, which becomes highly significant when dealing with large volumes of data during online training and monitoring.

6.2 Recommendations

To further enhancement the accuracy and robustness of NO_x emission and efficiency prediction, as well as the fault prognosis model, the following visionary recommendations are proposed to drive advancements in the industry:

- With the expected growing application of hydrogen in the industry as a cleaner energy source, efforts should focus on acquiring real-world and large-scale datasets from industrial boilers using hydrogen-enriched fuels. This will enhance the generalizability and accuracy, as well as the applicability of AI-based soft sensors in the industry, and support the global transition toward sustainable energy systems.
- Future research should focus on creating universally applicable models capable of predicting emissions and efficiency across various fuel types (e.g., hydrogen, natural gas, ammonia, and their blends). Physics-informed neural networks (PINNs) should be explored to integrate domain knowledge, to leverage regularization and model interpretation, enabling robust predictions even under novel fuel types or extreme operational conditions.
- To capture fault events more effectively, it is crucial to significantly expand the dataset to include a broader range of operational scenarios and fault conditions. The limited dataset can restrict the model from learning patterns associated with diverse failures, thus ensuring the inclusion of multiple fault events and varying operational conditions will allow the models to capture temporal trends and rare anomalies. Advanced deep learning architectures, such as transformer-based models, are proposed to enhance anomaly detection and predict failures with greater accuracy.

REFERENCES

- [1] H. Ritchie, M. Roser, and P. Rosado, “Energy Production and Consumption - Our World in Data,” “Energy”, Published online at OurWorldInData.org. Accessed: May 31, 2023. [Online]. Available: <https://ourworldindata.org/energy-production-consumption>
- [2] A. Ahmed, A. Q. Al-Amin, A. F. Ambrose, and R. Saidur, “Hydrogen fuel and transport system: A sustainable and environmental future,” *Int J Hydrogen Energy*, vol. 41, no. 3, pp. 1369–1380, Jan. 2016, doi: 10.1016/J.IJHYDENE.2015.11.084.
- [3] IEA (2023), “Global Hydrogen Review 2023,” IEA, Paris. Accessed: Sep. 12, 2024. [Online]. Available: <https://www.iea.org/reports/global-hydrogen-review-2023>
- [4] R. Lamioni, C. Bronzoni, M. Folli, L. Tognotti, and C. Galletti, “Impact of H₂-enriched natural gas on pollutant emissions from domestic condensing boilers: numerical simulations of the combustion chamber,” *Int J Hydrogen Energy*, vol. 48, no. 51, pp. 19686–19699, 2023, doi: <https://doi.org/10.1016/j.ijhydene.2023.02.040>.
- [5] M. S. Cellek and A. Pınarbaşı, “Investigations on performance and emission characteristics of an industrial low swirl burner while burning natural gas, methane, hydrogen-enriched natural gas and hydrogen as fuels,” *Int J Hydrogen Energy*, vol. 43, no. 2, pp. 1194–1207, 2018, doi: <https://doi.org/10.1016/j.ijhydene.2017.05.107>.
- [6] M. K. Büyükakın and S. Öztuna, “Numerical investigation on hydrogen-enriched methane combustion in a domestic back-pressure boiler and non-premixed burner

- system from flame structure and pollutants aspect,” *Int J Hydrogen Energy*, vol. 45, no. 60, pp. 35246–35256, 2020, doi: <https://doi.org/10.1016/j.ijhydene.2020.03.117>.
- [7] T. Wang, H. Zhang, Y. Zhang, H. Wang, J. Lyu, and G. Yue, “Efficiency and emissions of gas-fired industrial boiler fueled with hydrogen-enriched nature gas: A case study of 108 t/h steam boiler,” *Int J Hydrogen Energy*, vol. 47, no. 65, pp. 28188–28203, 2022, doi: <https://doi.org/10.1016/j.ijhydene.2022.06.121>.
- [8] B. C. Erdener *et al.*, “A review of technical and regulatory limits for hydrogen blending in natural gas pipelines,” *Int J Hydrogen Energy*, vol. 48, no. 14, pp. 5595–5617, 2023, doi: <https://doi.org/10.1016/j.ijhydene.2022.10.254>.
- [9] A. Rehan, M. Elshafei, M. A. Habib, M. A. Nemitallah, and I. Alzaharnah, “Boiler dynamic control with optimized nitric oxides and efficiency,” *Proceedings of the Institution of Mechanical Engineers. Part I: Journal of Systems and Control Engineering*, vol. 231, no. 10, pp. 778–796, Nov. 2017, doi: [10.1177/0959651817729079](https://doi.org/10.1177/0959651817729079).
- [10] UNIDO, “Energy Efficiency Potentials in Industrial Steam Systems in China,” Vienna, 2014. Accessed: May 31, 2023. [Online]. Available: https://www.unido.org/sites/default/files/2015-09/EE_Potentials_Steam_Systems_China__0.pdf
- [11] G. Stupara *et al.*, “The influence of primary measures for reducing NO_x emissions on energy steam boiler efficiency,” *Proceedings of ECOS*, pp. 1–13, 2012.

- [12] J. M. Beér, “High efficiency electric power generation: The environmental role,” *Prog Energy Combust Sci*, vol. 33, no. 2, pp. 107–134, 2007, doi: <https://doi.org/10.1016/j.pecs.2006.08.002>.
- [13] Q. Chen *et al.*, “Condensing boiler applications in the process industry,” *Appl Energy*, vol. 89, no. 1, pp. 30–36, 2012, doi: <https://doi.org/10.1016/j.apenergy.2010.11.020>.
- [14] Marshall Brain & Yara Simón, “How Steam Engines Work,” howstuffworks. Accessed: Jan. 04, 2025. [Online]. Available: <https://science.howstuffworks.com/transport/engines-equipment/steam.htm>
- [15] M. A. Nemitallah *et al.*, “Artificial intelligence for control and optimization of boilers’ performance and emissions: A review,” *J Clean Prod*, vol. 417, p. 138109, 2023, doi: <https://doi.org/10.1016/j.jclepro.2023.138109>.
- [16] S. K. Hoekman and C. Robbins, “Review of the effects of biodiesel on NOx emissions,” *Fuel Processing Technology*, vol. 96, pp. 237–249, 2012, doi: <https://doi.org/10.1016/j.fuproc.2011.12.036>.
- [17] S. M. Palash, M. A. Kalam, H. H. Masjuki, B. M. Masum, I. M. Rizwanul Fattah, and M. Mofijur, “Impacts of biodiesel combustion on NOx emissions and their reduction approaches,” *Renewable and Sustainable Energy Reviews*, vol. 23, pp. 473–490, Jul. 2013, doi: [10.1016/J.RSER.2013.03.003](https://doi.org/10.1016/J.RSER.2013.03.003).
- [18] A. Gungor, “Analysis of combustion efficiency in CFB coal combustors,” *Fuel*, vol. 87, no. 7, pp. 1083–1095, Jun. 2008, doi: [10.1016/J.FUEL.2007.06.005](https://doi.org/10.1016/J.FUEL.2007.06.005).

- [19] Y. T. Villarino, L. P. Rial, and Á. Rodríguez-Abalde, “Assessment of a residual biomass micro-combined heat and power system based on an organic Rankine Cycle coupled to a boiler,” *J Environ Manage*, vol. 301, p. 113832, 2022, doi: <https://doi.org/10.1016/j.jenvman.2021.113832>.
- [20] Y. Li and H. Gao, “On-line Calculation for Thermal Efficiency of Boiler,” in *2010 Asia-Pacific Power and Energy Engineering Conference*, 2010, pp. 1–4. doi: 10.1109/APPEEC.2010.5449383.
- [21] S. Shah and D. M. Adhyaru, “Boiler efficiency analysis using direct method,” in *2011 Nirma University International Conference on Engineering*, 2011, pp. 1–5. doi: 10.1109/NUiConE.2011.6153313.
- [22] S. Tronci, R. Baratti, and A. Servida, “Monitoring pollutant emissions in a 4.8MW power plant through neural network,” *Neurocomputing*, vol. 43, no. 1, pp. 3–15, 2002, doi: [https://doi.org/10.1016/S0925-2312\(01\)00617-8](https://doi.org/10.1016/S0925-2312(01)00617-8).
- [23] M. Liukkonen, E. Hälikkä, T. Hiltunen, and Y. Hiltunen, “Adaptive soft sensor for fluidized bed quality: Applications to combustion of biomass,” *Fuel Processing Technology*, vol. 105, pp. 46–51, 2013, doi: <https://doi.org/10.1016/j.fuproc.2011.06.015>.
- [24] K. Sujatha, Nallamilli. P. G. Bhavani, S.-Q. Cao, and K. S. Ram Kumar, “Soft Sensor for Flame Temperature Measurement and IoT based Monitoring in Power Plants,” *Mater Today Proc*, vol. 5, no. 4, Part 3, pp. 10755–10762, 2018, doi: <https://doi.org/10.1016/j.matpr.2017.12.359>.

- [25] Z. He and H. Liu, “A Novel Soft Sensor Model Based on Stacking Ensemble Learning Framework,” in *2022 IEEE 11th Data Driven Control and Learning Systems Conference (DDCLS)*, 2022, pp. 507–512. doi: 10.1109/DDCLS55054.2022.9858462.
- [26] K. Li, Q. Li, H. Fan, Y. Wang, S. Chang, and C. Zhao, “Soft sensing of SO₂ emission for ultra-low emission coal-fired power plant with dynamic model and segmentation model,” *Fuel*, vol. 332, p. 125921, 2023, doi: <https://doi.org/10.1016/j.fuel.2022.125921>.
- [27] Y. Wang, X. Chen, and C. Zhao, “A data-driven soft sensor model for coal-fired boiler SO₂ concentration prediction with non-stationary characteristic,” *Energy*, vol. 300, p. 131522, 2024, doi: <https://doi.org/10.1016/j.energy.2024.131522>.
- [28] J. C. B. Gonzaga, L. A. C. Meleiro, C. Kiang, and R. Maciel Filho, “ANN-based soft-sensor for real-time process monitoring and control of an industrial polymerization process,” *Comput Chem Eng*, vol. 33, no. 1, pp. 43–49, 2009, doi: <https://doi.org/10.1016/j.compchemeng.2008.05.019>.
- [29] J. M. Moreira de Lima and F. M. Ugulino de Araujo, “Ensemble deep relevant learning framework for semi-supervised soft sensor modeling of industrial processes,” *Neurocomputing*, vol. 462, pp. 154–168, 2021, doi: <https://doi.org/10.1016/j.neucom.2021.07.086>.
- [30] J. M. Rožanec *et al.*, “Machine-Learning-Based Soft Sensors for Energy Efficient Operation of Crude Distillation Units,” in *2022 International Conference on*

- Electrical, Computer and Energy Technologies (ICECET)*, 2022, pp. 1–6. doi: 10.1109/ICECET55527.2022.9872983.
- [31] N. Zhai, X. Zhou, S. Li, and H. Shi, “Soft Sensor Model for Billet Temperature in Multiple Heating Furnaces Based on Transfer Learning,” *IEEE Trans Instrum Meas*, vol. 72, pp. 1–13, 2023, doi: 10.1109/TIM.2023.3267520.
- [32] N. Sifakis, N. Sarantinoudis, G. Tsinarakis, C. Politis, and G. Arampatzis, “Soft Sensing of LPG Processes Using Deep Learning,” *Sensors*, vol. 23, no. 18, 2023, doi: 10.3390/s23187858.
- [33] M. Shakil, M. Elshafei, M. A. Habib, and F. A. Maleki, “Soft sensor for NO_x and O₂ using dynamic neural networks,” *Computers and Electrical Engineering*, vol. 35, no. 4, pp. 578–586, Jul. 2009, doi: 10.1016/j.compeleceng.2008.08.007.
- [34] S. A. Iliyas, M. Elshafei, M. A. Habib, and A. A. Adeniran, “RBF neural network inferential sensor for process emission monitoring,” *Control Eng Pract*, vol. 21, no. 7, pp. 962–970, Jul. 2013, doi: 10.1016/j.conengprac.2013.01.007.
- [35] B. Lin, B. Recke, J. K. H. Knudsen, and S. B. Jørgensen, “A systematic approach for soft sensor development,” *Comput Chem Eng*, vol. 31, no. 5, pp. 419–425, 2007, doi: <https://doi.org/10.1016/j.compchemeng.2006.05.030>.
- [36] M. Elshafei, M. A. Habib, I. Al-Zaharnah, and M. A. Nemitallah, “Boilers Optimal Control for Maximum Load Change Rate,” *J Energy Resour Technol*, vol. 136, no. 3, Sep. 2014, doi: 10.1115/1.4027563.

- [37] I. Alzaharnah, M. A. Habib, M. Elshafei, and P. Ahmed, “Control of the Boiler Swing Rate for NO Emission Minimization,” *Energy & Fuels*, vol. 27, no. 10, pp. 6079–6086, 2013, doi: 10.1021/ef4017043.
- [38] A. T. Y. Watanabe, R. Sebem, A. B. Leal, and M. da S. Hounsell, “Fault prognosis of discrete event systems: An overview,” *Annu Rev Control*, vol. 51, pp. 100–110, 2021, doi: <https://doi.org/10.1016/j.arcontrol.2021.03.005>.
- [39] S. Khalid *et al.*, “Intelligent Steam Power Plant Boiler Waterwall Tube Leakage Detection via Machine Learning-Based Optimal Sensor Selection,” *Sensors*, vol. 20, no. 21, 2020, doi: 10.3390/s20216356.
- [40] K. Rostek, Ł. Morytko, and A. Jankowska, “Early detection and prediction of leaks in fluidized-bed boilers using artificial neural networks,” *Energy*, vol. 89, pp. 914–923, 2015, doi: <https://doi.org/10.1016/j.energy.2015.06.042>.
- [41] H. Qin, S. Yin, T. Gao, and H. Luo, “A Data-driven Fault Prediction Integrated Design Scheme Based on Ensemble Learning for Thermal Boiler Process,” in *2020 IEEE International Conference on Industrial Technology (ICIT)*, 2020, pp. 639–644. doi: 10.1109/ICIT45562.2020.9067216.
- [42] G. Li *et al.*, “Research on fault diagnosis of supercharged boiler with limited data based on few-shot learning,” *Energy*, vol. 281, p. 128286, 2023, doi: <https://doi.org/10.1016/j.energy.2023.128286>.

- [43] M. Lv, J. Zhao, S. Cao, and T. Shen, “Prediction of the 3D Distribution of NO_x in a Furnace via CFD Data Based on ELM,” *Front Energy Res*, vol. 10, 2022, doi: 10.3389/fenrg.2022.848209.
- [44] D. Zhang *et al.*, “Modeling and simulating of reservoir operation using the artificial neural network, support vector regression, deep learning algorithm,” *J Hydrol (Amst)*, vol. 565, pp. 720–736, Oct. 2018, doi: 10.1016/J.JHYDROL.2018.08.050.
- [45] G. Yang, Y. Wang, and X. Li, “Prediction of the NO_x emissions from thermal power plant using long-short term memory neural network,” *Energy*, vol. 192, p. 116597, 2020, doi: <https://doi.org/10.1016/j.energy.2019.116597>.
- [46] M. A. Nemitallah, M. Aliyu, M. Hamdy, and M. A. Habib, “Retrofitting Natural Gas-Fired Boiler for Hydrogen Combustion: Operational Performance and NO_x Emissions,” *J Energy Resour Technol*, vol. 146, no. 10, Jun. 2024, doi: 10.1115/1.4065205.
- [47] S. Gersena, B. Slima, R. Zeijlmakera, M. Van Essena, and R. Tichelaar, “The Development of a Natural Gas/Hydrogen Boiler System. The International Gas Union Research Conference, Muscat, Oman, 24–26 Feb., 2020,” URL: https://www.researchgate.net/publication/339899796_The_Development_of_a_Natural_GasHydrogen_Boiler_System, pp. 1–8, 2020.
- [48] K. Mahmud Sujon, R. Binti Hassan, Z. Tusnia Towshi, M. A. Othman, M. Abdus Samad, and K. Choi, “When to Use Standardization and Normalization: Empirical

- Evidence From Machine Learning Models and XAI,” *IEEE Access*, vol. 12, pp. 135300–135314, 2024, doi: 10.1109/ACCESS.2024.3462434.
- [49] T. Verdonck, B. Baesens, M. Óskarsdóttir, and S. vanden Broucke, “Special issue on feature engineering editorial,” *Mach Learn*, vol. 113, no. 7, pp. 3917–3928, 2024, doi: 10.1007/s10994-021-06042-2.
- [50] J. Zhou, H. Lin, H. Jin, S. Li, Z. Yan, and S. Huang, “Cooperative prediction method of gas emission from mining face based on feature selection and machine learning,” *Int J Coal Sci Technol*, vol. 9, no. 1, p. 51, 2022, doi: 10.1007/s40789-022-00519-8.
- [51] L. Shao and K. Zhang, “A Gas Emission Prediction Model Based on Feature Selection and Improved Machine Learning,” *Processes*, vol. 11, no. 3, 2023, doi: 10.3390/pr11030883.
- [52] Z. Tang, S. Wang, and Y. Li, “Dynamic NOX emission concentration prediction based on the combined feature selection algorithm and deep neural network,” *Energy*, vol. 292, p. 130608, 2024, doi: <https://doi.org/10.1016/j.energy.2024.130608>.
- [53] G. Alfian *et al.*, “Predicting Breast Cancer from Risk Factors Using SVM and Extra-Trees-Based Feature Selection Method,” *Computers*, vol. 11, no. 9, 2022, doi: 10.3390/computers11090136.
- [54] A. A. Dhumkekar, S. Ghorpade, and R. N. Awale, “Performance Analysis of Various Cancers Using Genetic Data with Variance Threshold,” in *2022 OITS*

- International Conference on Information Technology (OCIT)*, 2022, pp. 67–72. doi: 10.1109/OCIT56763.2022.00023.
- [55] G. Alfian *et al.*, “Predicting Breast Cancer from Risk Factors Using SVM and Extra-Trees-Based Feature Selection Method,” *Computers*, vol. 11, no. 9, 2022, doi: 10.3390/computers11090136.
- [56] H. Xu and Y. Deng, “Dependent Evidence Combination Based on Shearman Coefficient and Pearson Coefficient,” *IEEE Access*, vol. 6, pp. 11634–11640, 2018, doi: 10.1109/ACCESS.2017.2783320.
- [57] D. and H. T. and T. R. and T. J. James Gareth and Witten, “Linear Regression,” in *An Introduction to Statistical Learning: with Applications in Python*, Cham: Springer International Publishing, 2023, pp. 69–134. doi: 10.1007/978-3-031-38747-0_3.
- [58] E. Pekel, “Estimation of soil moisture using decision tree regression,” *Theor Appl Climatol*, vol. 139, no. 3, pp. 1111–1119, 2020, doi: 10.1007/s00704-019-03048-8.
- [59] M. Yasser, “Housing Prices Dataset,” Kaggle.
- [60] L. Breiman, “Random Forests,” *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [61] O. Kramer and others, *Dimensionality reduction with unsupervised nearest neighbors*, vol. 51. Springer, 2013.

- [62] M. Bianchini, P. Frasconi, and M. Gori, “Learning without local minima in radial basis function networks,” *IEEE Trans Neural Netw*, vol. 6, no. 3, pp. 749–756, 1995, doi: 10.1109/72.377979.
- [63] G. Yang, Y. Wang, and X. Li, “Prediction of the NO_x emissions from thermal power plant using long-short term memory neural network,” *Energy*, vol. 192, p. 116597, 2020, doi: <https://doi.org/10.1016/j.energy.2019.116597>.
- [64] M. Despotovic, V. Nedic, D. Despotovic, and S. Cvetanovic, “Evaluation of empirical models for predicting monthly mean horizontal diffuse solar radiation,” *Renewable and Sustainable Energy Reviews*, vol. 56, pp. 246–260, 2016, doi: <https://doi.org/10.1016/j.rser.2015.11.058>.
- [65] M. H. Awadalla, “Radial basis Function Neural Network for Predicting Flow Bottom Hole Pressure,” *International Journal of Advanced Computer Science and Applications*, 2019, [Online]. Available: <https://api.semanticscholar.org/CorpusID:86625054>
- [66] A. S. Rahmati and A. Tatar, “Application of Radial Basis Function (RBF) neural networks to estimate oil field drilling fluid density at elevated pressures and temperatures,” *Oil & Gas Science and Technology–Revue d’IFP Energies nouvelles*, vol. 74, p. 50, 2019.
- [67] G. W. Chang, C.-I. Chen, and Y.-F. Teng, “Radial-Basis-Function-Based Neural Network for Harmonic Detection,” *IEEE Transactions on Industrial Electronics*, vol. 57, no. 6, pp. 2171–2179, 2010, doi: 10.1109/TIE.2009.2034681.

LIST OF ABBREVIATIONS

A

ACO	: Ant Colony Optimization
AFR	: Air-to-fuel Ratio
AFRact	: Actual Ai-to-fuel Ratio
AFRst	: Stoichiometric Air-to-fuel Ratio
AI	: Artificial Antelligence
ANN	: Artificial Neural Networks

B

BPNN	: Back Propagation Neural Networks
------	------------------------------------

C

CatBoost	: Category Boosting
CEMS	: Continuous Emission Monitoring System
CFB	: Circulating Fluidized Bed
CFD	: Computational Fluid Dynamics
CNN	: Convolutional Neural Networks
CO	: Carbon Monoxide
CO ₂	: Carbon Dioxide
CV-MIC	: Coefficient Variation and Maximum Mutual Information Coefficient

D

DL	: Deep Learning
DNN	: Deep Neural Networks

DPLS : Dynamic Partial Least Squares

E

EA : Excess Air

EDRLSS : Ensemble Deep Relevant Learning Soft Sensor

Extra Tree : Extremely Randomized Trees

F

FFNN : Feed-Forward Neural Networks

FGD : Flue Gas Desulfurization

FTLNN : Focused-Time Lagged Feed-Forward Neural Network

G

GA : Genetic Algorithm

GPR : Gaussian Process Regression

H

HC : Hydrocarbons

H₂ : Hydrogen

H₂O : Water

I

IEA : International Energy Agency

IORNN : Input-Output Recurrent Neural Network

IoT : Internet of Things

K

K : Kelvin

klbs/hr : kilo pounds per hour

KNN	: K-Nearest Neighbors
KPIs	: Key Performance Indicators
L	
LASSO	: Least Absolute Shrinkage and Selection Operator
LCMs	: Load Condition Modes
LCS	: Load Condition Slices
LDA	: Linear Discriminant Analysis
LOCF	: Last Observation Carried Forward
LPG	: Liquefied Petroleum Gas
LR	: Linear Regression
LSCD	: Load-based sequential condition division
LSTM	: Long Short Term Memory
M	
MAE	: Mean Absolute Error
MHI	: Mitsubishi Heavy Industries
MI	: Mutual Information
MIC	: Maximal Information Coefficients
MJ	: Mega Joule
ML	: Machine Learning
MLP	: Multi-Layer Perceptron
MLR	: Multiple Linear Regression
Mt	: Million metric tons
MTL-KD	: Multisource Transfer Learning - Knowledge Distillation

MW	: Megawatt
N	
NB	: Naive Bayes
NG	: Natural Gas
NN	: Neural Networks
NOCB	: Next Observation Carried Backward
NO _x	: Nitrogen Oxides
O	
OLS	: Ordinary Least Squares
O ₂	: Oxygen
P	
PCA	: Principal Component Analysis
PET	: Polyethylene Terephthalate
ppm	: parts per million
psig	: Pounds per square inch gauge
Q	
qA	: Flue Gas Losses
R	
RBF	: Radial Basis Function
RBFNN	: Radial Basis Function Neural Networks
RF	: Random Forest
RFE	: Recursive Feature Elimination
RFR	: Random Forest Regressor

RNN	: Recurrent Neural Networks
R^2	: R-squared (Coefficient of Determination)
S	
SAE	: Stacked Autoencoder
SCF/hr	: Standard cubic feet of gas per hour
SDAE	: Stacked Denoising Auto Encoder
SFS	: Sequential Forward Selection
SNN	: Siamese Neural Networks
SRU	: Sulfur Recovery Unit
SVM	: Support Vector Machine
SVR	: Support Vector Regression
T	
Ta	: Ambient Temperature
TCN	: Temporal Convolutional Networks
Ts	: Flue Gas Temperature
TSR	: Total Subset Regression
t/h	: ton per hour
U	
UNIDO	: United Nations Industrial Development Organization
V	
VAE	: Variational Auto Encoding
X	
XGBoost	: Extreme Gradient Boosting

VITAE

Full Name : Agus Haeruman

Nationality : Indonesia

Email : agus.haeruman@gmail.com

Address : Jl. Mulawarman RT.22 Gang Sumber Tani Blok A.5 Lamaru,
Balikpapan, Kalimantan Timur, 76117

Research Interests:

Steam Boilers, Soft Sensors, Artificial Intelligence, Machine Learning, Deep Learning,
Neural Networks, Energy and Sustainability.

Academic Background:

- *Master of Science in Mechanical Engineering*

King Fahd University of Petroleum and Minerals, Dhahran, 31261, KSA

Year of Passing: 2024

CGPA: 3.09 (on a scale of 4)

- *Bachelor of Science in Chemical Engineering*

Universitas Gadjah Mada (UGM), Yogyakarta, Indonesia.

Year of Passing: 2009

CGPA: 3.31 (on a scale of 4)

Professional Experience:

- *Teaching Assistantship (TA)*

Aug 2022 – Dec 2024

Mechanical Engineering Department

King Fahd University of Petroleum and Minerals, Dhahran, 31261, KSA

- *Research Assistant (RA)*

Aug 2022 – Dec 2024

SDAIA-KFUPM Joint Research Center for Artificial Intelligence

King Fahd University of Petroleum and Minerals, Dhahran, 31261, KSA

- *Research Assistant (RA)*

Aug 2022 – Dec 2024

Interdisciplinary Research Center for Hydrogen Technologies and Carbon Management (IRC-HTCM)

King Fahd University of Petroleum and Minerals, Dhahran, 31261, KSA

- *Account Manager*

Jul 2018 to Aug 2022

Rolls Royce Solutions Indonesia

Balikpapan, Indonesia

Publications:

- Nemitallah MA, Nabhan MA, Alowaifeer M, Haeruman A, Alzahrani F, Habib MA, et al. Artificial intelligence for control and optimization of boilers'

performance and emissions: A review. J Clean Prod 2023;417:138109.
<https://doi.org/https://doi.org/10.1016/j.jclepro.2023.138109>.

- Harahap, M. A., Haeruman, A., and Mokheimer, E. M. A. Optimal Composition of Palm Oil Biomass to Minimize Biomass Power Plants' Greenhouse Gases Emission ASME Open J. Engineering ASME. January 2023 2 021034 doi: <https://doi.org/10.1115/1.4062627>.
- Haeruman A, Elshafei M, Aliyu M, Nemitallah MA, Habib MA, Alowaifeer M, Alzahrani F. Detection of NOx Emissions in a Steam Boiler Fueled by Hydrogen-Enriched Natural Gas Using AI-Based Soft Sensors. **(Ready for submission)**
- Haeruman A & Alzahrani F. Prediction of NOx Emissions and Efficiency in a Natural Gas Steam Boiler Using AI-Based Soft Sensors. **(Manuscript under preparation)**
- Haeruman A & Alzahrani F. AI-based Fault Prognosis to predict Boiler Swings in a Natural Gas Steam Boiler using LSTM . **(Manuscript under preparation)** |