Politecnico di Torino

Master of Science in Automotive Engineering
Academic Year 2025/2026
Graduation Session July 2026

Data-Driven Modelling of Diesel Engine Performance, Combustion, and Emissions Metrics for Real-Time Applications: A Comparison of Regression Models and Neural Networks

Supervisors:
Prof. Stefano D’Ambrosio
Ing. Roberto Finesso

Candidate:
Davide Fazari

Summary

1. Introduction

Emission regulations over the last thirty years have forced automotive manufacturers to focus on energy efficiency and on the reduction of environmental impact of their internal combustion engines. Among these, diesel engines are still widely used in various heavy duty transport applications due to their high thermal efficiency and reliability. However, the increasing complexity of modern fuel injection systems, turbocharging technologies, and exhaust after-treatment devices has made advanced modelling and simulation tools essential for engine design, calibration, and control.

Among the various technologies under investigation, model-based control strategies play a crucial role in the development of advanced powertrain systems. However, high-fidelity physics-based engine models typically require computational resources and execution times that are not compatible with real-time implementation in modern Engine Control Units (ECUs). As a result, data-driven approaches have emerged as promising alternatives capable of simulating the behaviour of complex physical models while significantly reducing computational costs.

The availability of large amounts of experimental data acquired during engine test sessions has encouraged the development of data-driven modelling approaches, which are able to describe system behaviour without the need of a full physical representation. These methodologies can identify complex and nonlinear relationships between engine control variables and performance variables. They also provide effective tools for the estimation of quantities that are difficult or expensive to measure directly, reducing time and cost associated with experimental testing.

This thesis develops predictive models to estimate selected engine output variables from a set of measurements acquired during diesel engine test bench operations. A preliminary analysis of the available dataset was carried out in order to investigate its statistical properties and identify the correlations among the measured quantities.

Subsequently, different modelling approaches based on system identification and machine learning techniques were investigated and implemented. In a first stage, separate regression models were developed and trained to estimate each output variable individually. In a second stage, Multi-Layer Perceptron (MLP) neural networks were designed to simultaneously predict multiple output variables, exploiting the relationships among them.

The predictive performance of all developed models was assessed through appropriate quantitative metrics and validated against experimental measurements. Particular attention was devoted to evaluating the robustness and generalization capability of the model.

The results of this work demonstrate the potential of data-driven methodologies as effective tools for approximating complex engine simulations. Moreover, the developed models may provide valuable support for advanced applications in engine control, virtual sensing, diagnostics, performance optimization for automotive systems.

2. Experimental Data Preparation

2.1 Data Pre-processing

The first step of the procedure was to process and clean the experimental data that were acquired from a diesel engine during a bench test. The aim of this activity was to identify the main variables influencing the outputs and to prepare a dataset suitable for modelling.

This stage started by importing the entire dataset which was provided in Excel format and contained a mix of numerical values and text strings.

The dataset was then cleaned by keeping only numerical columns and the relative variable names, removing those with constant values or missing data. This was done because constant variables do not provide useful information for modelling purposes, while the presence of missing values can compromise the regression procedure.

Input variables were then selected based on physical knowledge of the engine system and their expected influence on the output variables of interest.

Specifically, quantities typically processed by ECU were selected, including:

- Engine speed
- Exhaust flap actuator position
- Injected fuel quantity
- Pilot fuel quantity
- Main injection timing
- Pilot injection timing
- Rail pressure
- Variable geometry turbocharger (VGT) position
- EGR valve position

For specific output variables, other inputs were considered but these are introduced in the following chapters.

It is important to specify that this feature selection was guided by knowledge of diesel engine and supported by statistical analysis.

In parallel, the system output variables were defined:

1. Brake torque
2. NOx emissions expressed in g/kWh
3. NOx emissions expressed in ppm
4. Indicated Mean Effective Pressure
5. Mass Fraction Burned 50%
6. Air mass flow rate
7. Maximum in cylinder pressure
8. Exhaust pressure
9. Exhaust gas temperature upstream of the turbine
10. Exhaust gas temperature downstream of the turbine

The problem was thus formalized as a supervised learning problem, in which the objective was to analyse the data and model the relationship between input and output variables.

Overall, the adopted methodology combines data cleaning techniques, variable selection based on statistical analysis and physical knowledge of the engine. This hybrid approach enabled the construction of a coherent and informative dataset, used for the development of predictive models.

2.2 Dataset preparation

The next phase consisted of preparing the dataset for training and testing the regression models. First, the dataset was split according to the total number of available observations. After determining the overall number of observations, approximately 100, 80% of them was dedicated for model training while the remaining portion was assigned to the test phase.

To ensure reproducibility of the results, the random number generator seed was fixed. In this way, the same split between training and test sets was ensured each time the script was executed.

The observation indices were then shuffled randomly, preventing any potential bias related to the temporal or experimental ordering of the measurements.

To prevent differences in variable scales from distorting the modelling process, each variable was normalized using the mean and the standard deviation computed exclusively on the training dataset. This step prevents the data leakage from the test set into the training phase.

Overall, this stage represents the preparation of the dataset for training supervised regression models.

3. Linear Regression Models

3.1 Full linear regression model

As a first modelling approach, a multiple linear regression model was developed using all the selected input variables to describe the relationship between engine parameters and the output variable.

The model was trained on the standardized training dataset using MATLAB's **“fitlm”** function, which estimates the regression coefficients through the least squares method. In this model, the output variable was expressed as a weighted sum of the input variables plus a constant term, with the coefficients determined to minimize the prediction error on the training set.

Once the training phase was completed the predicted values were compared with the measured values through a scatter plot, where the reference line $y=x$ represents the perfect prediction. The distance from the reference line gives a visual idea of the model accuracy.

To numerically evaluate the model performance, the Root Mean Square Error (RMSE) was computed. This metric measures the average magnitude of the prediction error and lower RMSE values indicate higher accuracy.

The same model was tested on the independent test set. The predictions were compared with the corresponding measured outputs using the same graphical representation adopted for the training. This step allows the assessment of the model's ability to generalize to unseen data.

The comparison between the RMSE values makes it possible to evaluate if the model exhibits good generalization capabilities or shows signs of overfitting.

Overall, this linear regression model allows to obtain a straightforward interpretation of the relationship between inputs and output, while providing a useful benchmark against which more advanced modelling techniques can be compared.

3.2 Multicollinearity analysis using the Variance Inflation Factor (VIF)

After building the full linear regression model, an analysis of multicollinearity among the input variables was performed using the Variance Inflation Factor (VIF).

The VIF quantify how linearly dependent each input variable is on the remaining ones. For each variable, the value is computed by fitting a regression in which the considered variable is treated as the dependent variable and all the others are used as predictors.

From the coefficient of determination R_i^2 obtained with this regression, the VIF coefficient is defined as:

$$VIF_i = \frac{1}{1 - R_i^2}$$

High VIF values indicate that a variable can be well explained by a linear combination of the others, suggesting the presence of multicollinearity.

In general, a commonly used threshold is 5, above which significant redundancy among variables is present.

Performing this analysis to the full model revealed a strong level of multicollinearity among the selected inputs. This result is consistent with the nature of the engine system, in which many ECU input variables are physically correlated leading to information redundancy.

The presence of multicollinearity does not necessarily compromise the predictive capability of the model. In fact, the full model still provides accurate predictions of the output variable during the testing phase. However, it reduces the reliability and interpretability of the regression coefficients. When strong correlations exist among predictors, small variations in the data can lead to significant changes in the estimated coefficients, making it difficult to attribute the contribution of each input to the output.

The full model can therefore be considered highly predictive but poorly interpretable, since regression coefficients are not sufficiently stable. This suggested that an additional variable selection was necessary.

3.3 Reduced linear regression model

The objective of this phase was also to reduce model complexity by selecting the most relevant variables while maintaining high predictive performance.

Variables were progressively removed based on physical interpretation of the system, prioritising those considered less informative or strongly correlated with others. The remaining variables were then used as regressors for the reduced linear model.

The model was evaluated on both the training and test sets to assess its predictive performance and generalization capability.

The mitigation of multicollinearity among the selected variables was then checked through the computation of the Variance Inflation Factor (VIF), to find any residual dependencies among the regressors in the reduced configuration.

3.4 Conditioning Number

An additional analysis of the correlation structure among the input variables of the reduced model was conducted to further assess the presence of multicollinearity and the overall stability of the model.

The conditioning number of the regressor matrix was computed. This index represents a global indicator of multicollinearity, as it quantifies the sensitivity of the system to variations in the explanatory variables only if they were previously standardized. A conditioning number below the commonly accepted threshold of ≈30 indicates that no significant multicollinearity is present.

Reducing the number of input variables lead to more stable and interpretable regression coefficients, improving the reliability of the individual variable contributions compared to the full model.

3.5 Residual analysis of the reduced model

Subsequently, a residual analysis was conducted to verify the statistical validity of the model assumptions.

The residuals were computed as the difference between the measured and predicted values, and their mean was calculated to check the absence of systematic bias.

The analysis was then extended through a graphical representation of the residuals versus the predicted values. This plot allows to detect any non-random structure in the errors, which could suggest nonlinear dependencies not captured. In the ideal case, residuals should be randomly distributed around zero, without any trends or structures.

Finally, a QQ-plot of the residuals was created to assess the normality of residual distribution. The good alignment of the points with the ideal diagonal indicates a residual distribution consistent with normality, supporting the validity of the linear model assumptions.

4. Non Linear Regression Models

4.6 Quadratic Regression Model

The quadratic model was applied only to those output variables for which the linear models did not provide satisfactory results in terms of RMSE and R^2, to improve predictive performance where a linear approximation was not sufficient.

The quadratic regression model was built from the variables selected in the reduced linear model, with the aim of capturing the main nonlinearities of the system while keeping complexity low.

The model was estimated on the training dataset using MATLAB's "**stepwiselm"** function with a quadratic specification, which automatically selects statistically significant linear, interaction, and quadratic terms through a stepwise procedure. In this way it was possible to eliminate all the terms that were not necessary for the prediction.

The quadratic model was validated on the test dataset to assess its generalization ability. Finally, a direct comparison between the reduced linear model and the quadratic model was performed using the same test dataset.

4.7 Power Law Regression Model

Specifically for the evaluation of NOx emissions, expressed both in ppm than g/kWh, a power law model was also considered.

This model was developed as an alternative to linear and quadratic models, with the aim of representing nonlinear relationships not captured before.

The model was built starting from the reduced variable set selected in the previous stages. Unlike the previous models, the non-standardised dataset was used, because the power law formulation requires positive quantities. For this reason, an additional data filtering was applied, keeping only observations where all input variables and the output are positive.

The model structure was defined in parametric form, in which the output is expressed as the product of input variables raised to exponents to be estimated.

$$y = \beta_0 \cdot x_1^{\beta_1} \cdot x_2^{\beta_2} \cdot x_3^{\beta_3} \cdots x_i^{\beta_i}$$

The model parameters were estimated through the MATLAB's "**lsqcurvefit"** function. Initial values for the exponents were set based on the physical dependence among the system variables.

Once the optimal parameters were estimated, the model was evaluated on the test dataset by comparing predicted and measured values. Performance was quantified using the RMSE to assess the model's ability to follow the output trend.

5. Neural Network Models

In addition to the previously developed linear and non-linear regression models, a neural network model one was implemented to further investigate the relationships between the selected input variables and each of the target output.

The same input variables and dataset partition adopted for the previous models were considered, ensuring identical training and test splits. This choice guarantees a fair and consistent comparison among the different modelling approaches.

A feedforward neural network for regression was then trained using MATLAB's "**fitrnet"** function. The architecture consisted of a single hidden layer, with the number of neurons chosen based on the expected complexity of the system to be modelled. The non-linearity of the system was introduced through a ReLu activation function in the hidden layer. The level of complexity was inferred from the performance and behaviour observed in the previously developed linear, quadratic, and power law models.

6. Multi-Output Neural Network Modelling

A subset of neural network models was developed by grouping output variables into physically related categories, to evaluate whether this strategy could improve prediction accuracy and reduce inference time.

Grouping outputs could allow the neural network to learn shared relationships between inputs and correlated outputs, exploiting implicit dependencies among variables within the same domain.

The available outputs were divided into four main groups:

- Performance: groups the engine torque, IMEP, MFB50 and peak pressure, representing performance indicators of the combustion process.
- NOx Emissions: includes NOx expressed both in g/kWh and ppm, which quantify pollutant emissions of NOx after the combustion.
- Thermal: is composed by thermodynamic variables of the system such as exhaust temperatures upstream and downstream of the turbine.
- Intake and Exhaust System: contains air mass flow rate and exhaust pressure, describing the intake – exhaust system behaviour.

The same input dataset used for the previously developed regression models was adopted for all neural network models. The dataset was split into training, validation, and test sets using a reproducible random partition. Both input variables and outputs were independently standardised based on training set statistics.

For each output group, a dedicated feedforward neural network was defined with an architecture adapted to the complexity of the corresponding subsystem.

The Performance group adopted a deeper network with larger hidden layers, reflecting the higher nonlinear complexity of combustion variables and the large number of variables to be estimated. The NOx and Thermal groups adopted a deep architecture to better capture the dependencies. Finally, the Air System group used a more compact network due to its simpler input-output relationship.

All models were trained using the Adam optimizer. The training procedure was kept consistent across all networks, with 500 epochs, a batch size of 32, and an initial learning rate tuned for each group. A validation set was used during training to monitor generalization performance and prevent overfitting, while the test set was kept completely unseen and for final evaluation.

Model performance was evaluated using the Root Mean Square Error (RMSE) and the coefficient of determination (R^2) for each output variable within each group.

7. Results Analysis

This chapter presents and discusses the results obtained from the different modelling models described in the previous sections. For each output variable, the predictive performance of the models is analysed through appropriate statistical indicators and graphical comparisons between measured and predicted values. Particular attention was devoted to the evaluation of model accuracy and generalization capability.

The analysis aims to identify the most suitable modelling approach for each output category and to assess the advantages and limitations of each technique.

7.1 Brake Torque

7.1.1 Full Linear Model

As a first modelling approach, a linear regression model was developed to describe the relationship between the engine torque and a set of variables available to the Electronic Control Unit (ECU). In particular, some actuator signals and parameters related to the combustion were selected as input variables. The application of the linear model provided a useful starting point for evaluating the contribution of each control variables to torque generation and for obtaining have an initial interpretation of the physical phenomenon.

The linear regression model resulted in the coefficients reported in the following table.

Linear regression model:
y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	197.56	0.30661	644.33	2.6562e-168
x1	-24.464	3.2783	-7.4624	4.8957e-11
x2	-8.2044	1.334	-6.1503	2.0317e-08
x3	122.48	0.60317	203.07	1.0455e-122
x4	3.4724	1.005	3.4551	0.00083743
x5	6.6081	2.3908	2.7639	0.0069127
x6	13.529	1.0024	13.497	1.93e-23
x7	-2.1847	1.3367	-1.6344	0.10564
x8	8.675	1.3039	6.6532	2.0995e-09
x9	4.3196	0.77089	5.6033	2.2221e-07

Number of observations: 101, Error degrees of freedom: 91
Root Mean Squared Error: 3.08
R-squared: 0.999, Adjusted R-Squared: 0.999
F-statistic vs. constant model: 1.91e+04, p-value = 5.05e-145

Table 1: Regression coefficients of the full linear model for torque estimation

The coefficient associated with the fuel quantity was significantly larger than all the others, indicating the dominant role of this variable in determining the engine torque.

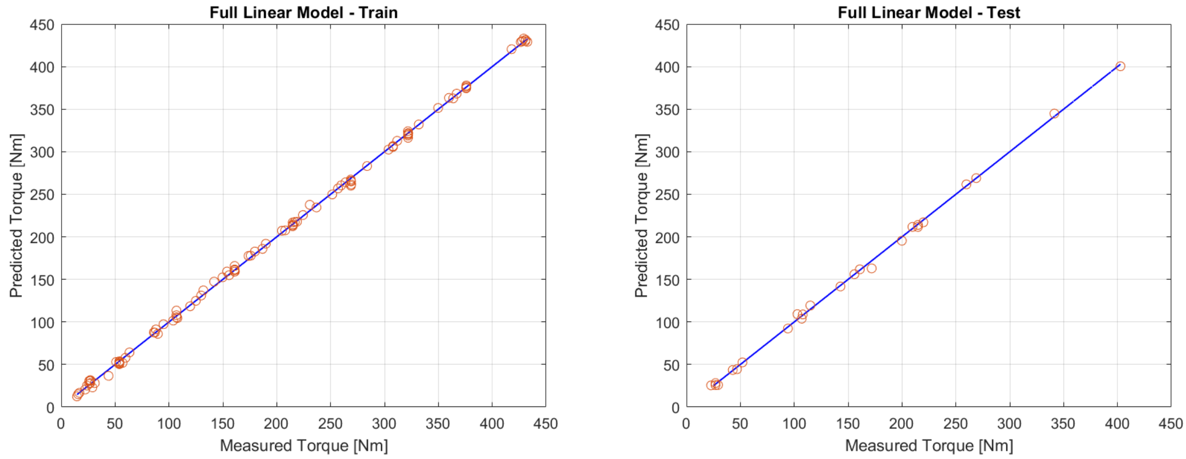

Figure 1: Training and testing plots of the full linear model for torque estimation

The predictive performance of the model on both the training and test sets is summarized below.

$RMSE_{TRAIN}$	$RMSE_{TEST}$	R^2
$2.92\ Nm$	$3.14\ Nm$	0.999

Table 2: Training and testing performance of the full linear model for torque estimation

These results highlight the excellent predictive capability of the model. The small difference between the training and test errors indicates that the model doesn't exhibit overfitting and generalises well to new data. Although the linear model exhibited very high predictive performance, it was important to assess whether the input variables were strongly correlated with one another. In general, if high correlation among the inputs is present, the model may have difficulty attributing the effect of each variable to the output and the estimated coefficients may become unstable.

The VIF analysis produced the following values for each input variable.

Input Variable	VIF
Engine Speed	113.19
Exhaust Flap	18.74
Fuel Quantity	3.83
Pilot Quantity	10.64
SOI Main	60.20
Rail Pressure	10.58
VGT Position	18.82
EGR Valve	17.91
SOI Pilot	6.26

Table 3: VIF values of the full linear model inputs for torque estimation

To assess the level of multicollinearity, the following commonly thresholds accepted in the literature were adopted:

- VIF < 5: no significant multicollinearity;
- 5 < VIF < 10: moderate multicollinearity;
- VIF > 10: high multicollinearity.

The results showed that several variables exhibited very high VIF values, indicating strong mutual correlation among the input variables.

To summarize, the full linear model estimated the engine torque with high accuracy, achieving a coefficient of determination of 0.999. However, the multicollinearity analysis revealed a correlation among the inputs, limiting the physical interpretability of the regression coefficients.

7.1.2 Reduced Linear Model

The multicollinearity analysis revealed a strong linear dependence among the input variables, as indicated by high VIF values. For this reason, a model was developed using as only the input variables considered physically relevant for engine torque estimation based on engineering knowledge of the system.

In particular, the following variables were selected:

1. **Injected Fuel Quantity**
   The variable most strongly correlated with torque, representing the energy introduced into the cylinder.
2. **Rail Pressure**
   It influences the fuel atomization and air–fuel mixture quality, affecting combustion speed and the energy release rate.
3. **Main Injection Timing**
   This variable determines the maximum in-cylinder pressure and influences both engine efficiency and torque output.
4. **EGR Valve**
   This valve controls the amount of recirculated exhaust gas, modifying oxygen concentration and therefore combustion temperature and reaction rates.

Engine speed was initially included in the set of input variables but later removed because the VIF values remained high.

$$VIF_{with\ engine\ speed} = \{42.45 \quad 3.61 \quad 30.35 \quad 6.30 \quad 3.05\}$$

When engine speed was excluded, the VIF values indicated that multicollinearity had been mitigated.

Input Variable	VIF
Injected Fuel	3.37
SOI Main	3.17
Rail Pressure	4.37
EGR valve	1.66

Table 4: VIF values of the reduced linear model inputs for torque estimation

The reduced linear regression model resulted in the coefficients reported in the following table.

```
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4

Estimated Coefficients:
                   Estimate      SE        tStat       pValue
                   ________    _______    _______    ___________

    (Intercept)     197.56     0.49242     401.2     1.3296e-156
    x1              122.31     0.90874    134.59     3.6589e-111
    x2             -11.697     0.88153   -13.268      1.9391e-23
    x3              10.965      1.0348    10.596      7.8481e-18
    x4               4.496     0.63842    7.0424      2.8403e-10


Number of observations: 101, Error degrees of freedom: 96
Root Mean Squared Error: 4.95
R-squared: 0.999,  Adjusted R-Squared: 0.999
F-statistic vs. constant model: 1.67e+04, p-value = 1.81e-135
```

Table 5: Regression coefficients of the reduced linear model for torque estimation

The model achieved comparable prediction accuracy on both the training and test sets as confirmed by the RMSE values.

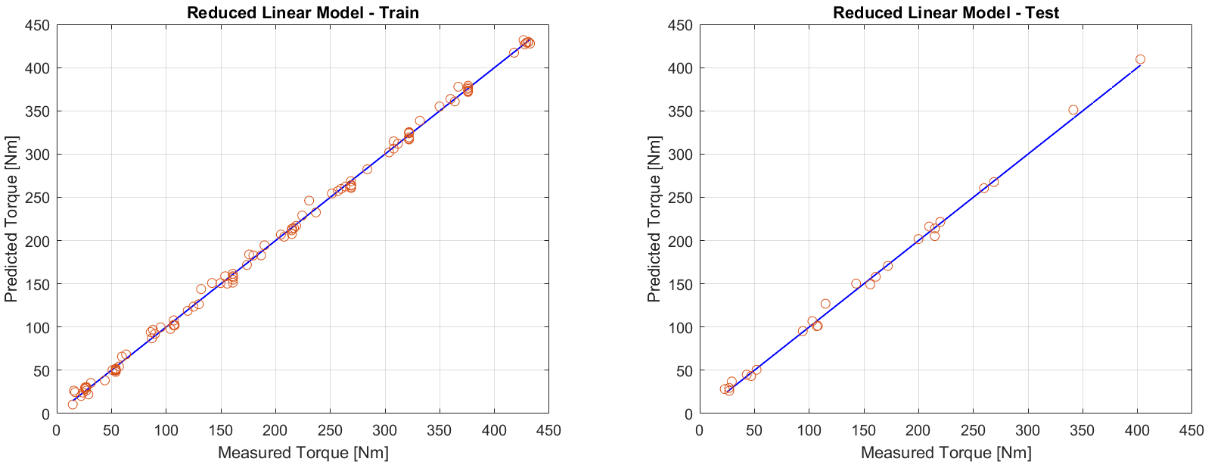

Figure 2: Training and testing plots of the reduced linear model for torque estimation

$RMSE_{TRAIN}$	$RMSE_{TEST}$	R^2
4.82 Nm	5.46 Nm	0.999

Table 6: Training and testing performance of the reduced linear model for torque estimation

To confirm the absence of multicollinearity in the reduced model, the condition number of the input matrix was analysed.

Since the input variables were standardized, this index represents a reliable measure of linear dependence among the explanatory variables.

The obtained value was:

$$\kappa = 4.43$$

As the condition number is significantly lower than the commonly accepted threshold in the literature (≈ 30), it can be concluded that no relevant multicollinearity issues were present among the set of variables.

Further confirmation was provided by the correlation matrix of the input variables.

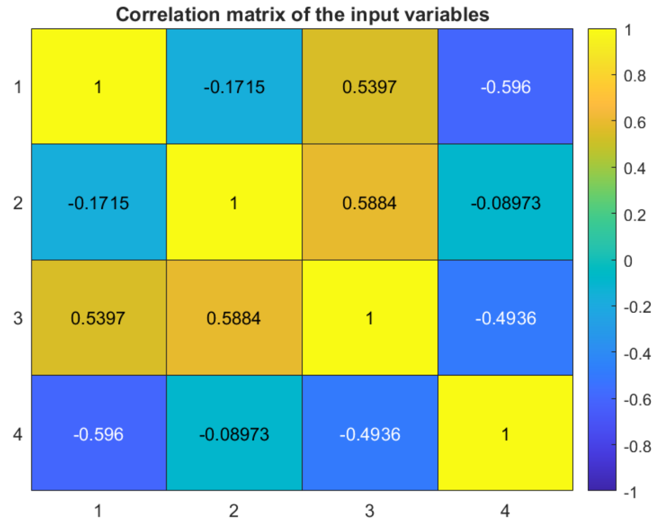

Figure 3: Correlation matrix of the selected input variables for torque estimation

The correlation matrix heatmap confirmed that no pair of variables exhibited correlations close to ± 1, indicating the absence of strong linear dependencies among the inputs.

The residual analysis revealed no systematic patterns, confirming the adequacy of the linear model. Actually, the residuals were randomly distributed around zero with approximately constant variance. However, slight deviations were observed, suggesting the presence of outliers.

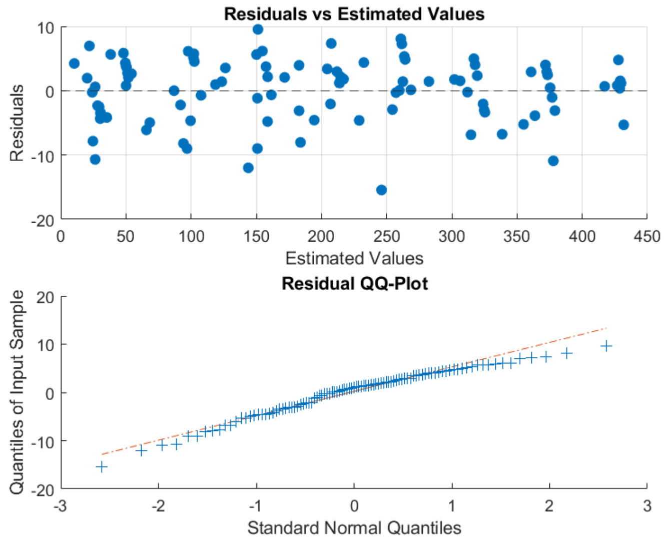

Figure 4: Residual plots of the reduced linear model for torque estimation

7.1.3 Neural Network

The same dataset and input variables used for the reduced linear model were adopted to ensure a fair comparison. Input variables were standardized using training set statistics, while the output variable remains the engine torque. A nonlinear regression model was developed using a Fully Connected Feedforward Neural Network, also known as Multilayer Perceptron MLP.

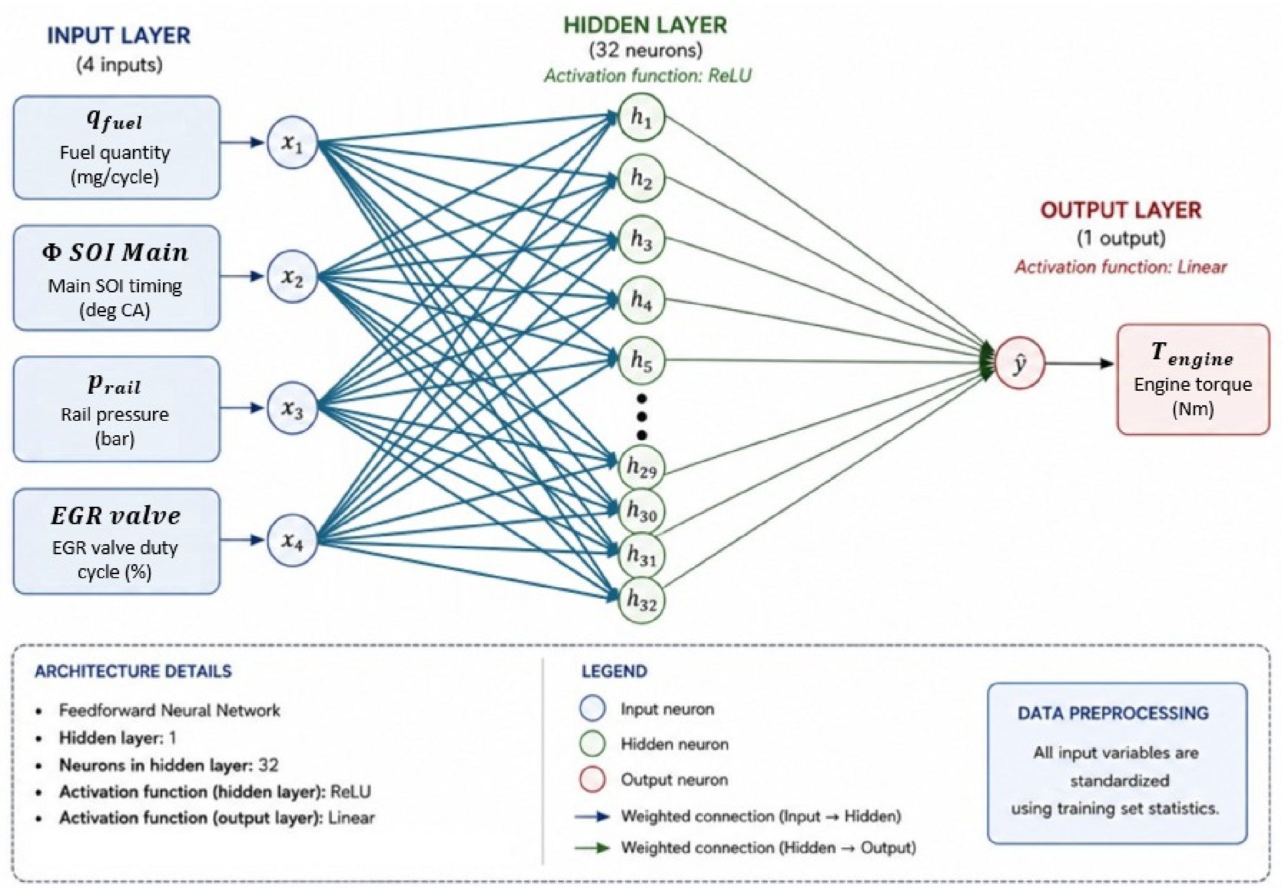

Figure 5: Neural network structure for torque estimation

$RMSE_{TRAIN}$	$RMSE_{TEST}$
$1.71\ Nm$	$2.55\ Nm$

Table 7: Training and testing RMSE of the neural network for torque estimation

These results indicated that the neural network accurately captured the nonlinear relationship between input variables and engine torque, achieving lower RMSE than linear models previously developed. The small difference between training and test errors confirms good generalisation.

7.1.4 Model Comparison

The full linear model, which included the entire set of available ECU variables, showed good overall performance but exhibited multicollinearity and redundancy among inputs. For these reasons, the model coefficients were less stable and more difficult to interpret.

The reduced linear model achieved good predictive performance, with an RMSE only slightly worse than the full model. The reduced input set successfully mitigated multicollinearity.

The neural network significantly increased prediction accuracy, reducing the test RMSE. The big improvement suggested that the relationship between the selected ECU variables and engine torque was not purely linear.

In particular, combustion-related phenomena introduce nonlinear behaviours that cannot be fully captured by a linear model. Through its nonlinear activation functions, the neural network was able to learn these complex interactions and achieve more accurate torque predictions.

On the other hand, the increased predictive capability came at the expense of interpretability. While the coefficients of the linear model can be directly related to physical effects, the neural network behaves as a black box model whose internal parameters do not have a clear physical meaning.

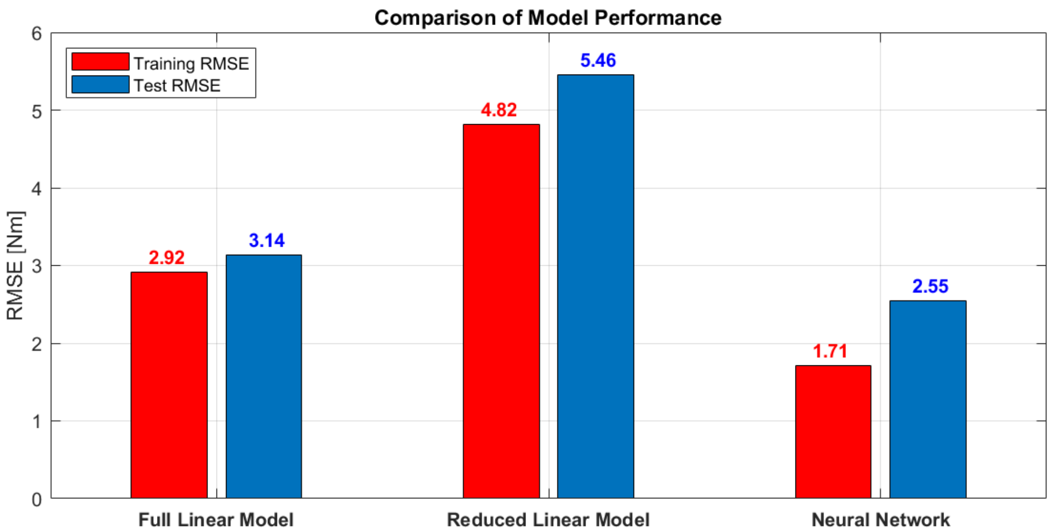

Figure 6: Performance comparison between models for torque estimation

7.2 IMEP

7.2.1 Full Linear Model

The modelling framework adopted for torque estimation was extended to the indicated mean effective pressure (IMEP), using the same input variables. The coefficients of the resulting model are summarized below.

```
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9

Estimated Coefficients:
                   Estimate       SE        tStat        pValue
                   ________    ________    _______    ___________

    (Intercept)      9.8654    0.014973      658.9     3.475e-169
    x1             -0.76478     0.16009    -4.7772     6.7812e-06
    x2             -0.40706    0.065142    -6.2488     1.3091e-08
    x3               5.3655    0.029454     182.16    2.0037e-118
    x4              0.13128    0.049076      2.675      0.0088609
    x5               0.3208     0.11675     2.7478      0.0072344
    x6              0.68343    0.048949     13.962     2.3889e-24
    x7             -0.12322    0.065276    -1.8876       0.062263
    x8              0.45235    0.063672     7.1045      2.616e-10
    x9              0.18168    0.037644     4.8261     5.5817e-06


Number of observations: 101, Error degrees of freedom: 91
Root Mean Squared Error: 0.15
R-squared: 0.999,  Adjusted R-Squared: 0.999
F-statistic vs. constant model: 1.56e+04, p-value = 5.28e-141
```

Table 8: Regression coefficients of the full linear model for IMEP estimation

The weight of the fuel quantity is significantly larger than the others, indicating its dominant contribution to the IMEP prediction. Actually, the fuel quantity directly influences the energy released per cycle and to the pressure inside the cylinder.

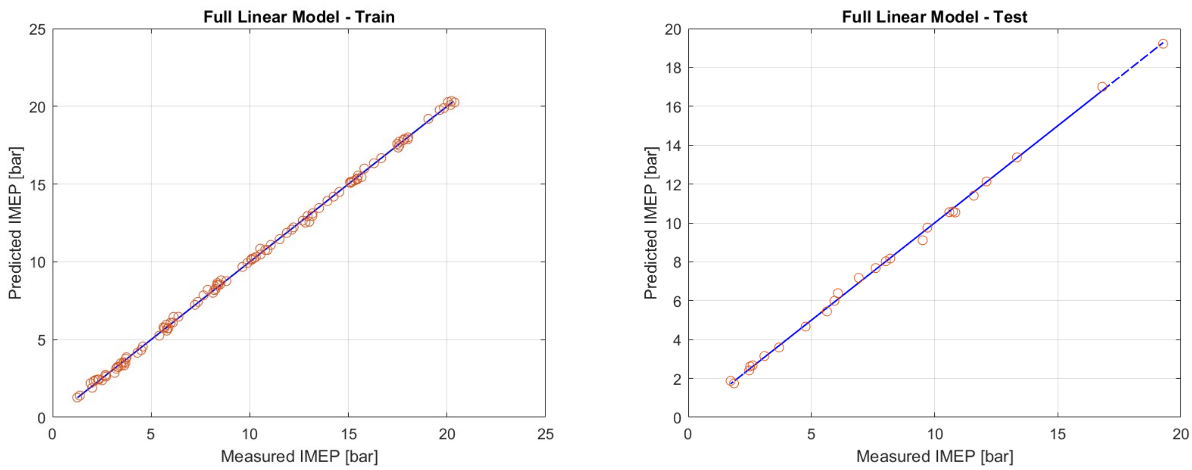

Figure 7: Training and testing plots of the full linear model for IMEP estimation

The predictive performance of the model on both the training and test sets is summarized below.

$RMSE_{TRAIN}$	$RMSE_{TEST}$	R^2
0.14 ¿	0.16 ¿	0.999

Table 9: Training and testing performance of the full linear model for IMEP estimation

The same observations made for torque are valid here: the model achieved excellent predictive performance but exhibited multicollinearity among the inputs, as indicated by the VIF values reported in the previous section.

7.2.2 Reduced Linear Model

A reduced model was developed using as inputs only those input variables considered physically relevant for the estimation of the IMEP prediction, based on engineering knowledge of the system.

The following variables were selected:

1. **Injected Fuel Quantity**
   It is the most influential variable, as higher fuel quantity injected increases the energy released during combustion and consequently the IMEP.
2. **Rail Pressure**
   It affects the injection conditions and fuel atomization, which has a relevant impact on the combustion efficiency and IMEP.
3. **Main Injection Timing**
   It determines the combustion phasing and the location of the peak pressure within the cycle. A proper calibration maximizes the useful work produced.
4. **EGR Valve**
   The EGR valve has a direct impact on IMEP, as it introduces exhaust gases into the intake charge, reducing the available oxygen and the combustion speed.

The VIF values indicated that the multicollinearity issue had been effectively mitigated.

Input Variable	VIF
Injected Fuel	3.37
SOI Main	3.17
Rail Pressure	4.37
EGR valve	1.66

Table 10: VIF values of the reduced linear model inputs for IMEP estimation

The reduced linear regression model yielded the coefficients reported in the following table.

```
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4

Estimated Coefficients:
                   Estimate       SE         tStat       pValue   
                   ________    ________    _______    ___________

    (Intercept)      9.8654    0.021348     462.12    1.7084e-162
    x1               5.3308    0.039397     135.31    2.2026e-111
    x2             -0.21304    0.038217    -5.5743      2.276e-07
    x3              0.66133    0.044863     14.741     2.1393e-26
    x4              0.16987    0.027677     6.1374     1.8781e-08


Number of observations: 101, Error degrees of freedom: 96
Root Mean Squared Error: 0.215
R-squared: 0.999,  Adjusted R-Squared: 0.999
F-statistic vs. constant model: 1.73e+04, p-value = 3.43e-136
```

Table 11: Regression coefficients of the reduced linear model for IMEP estimation

The model achieved comparable RMSE values on both the training and test datasets.

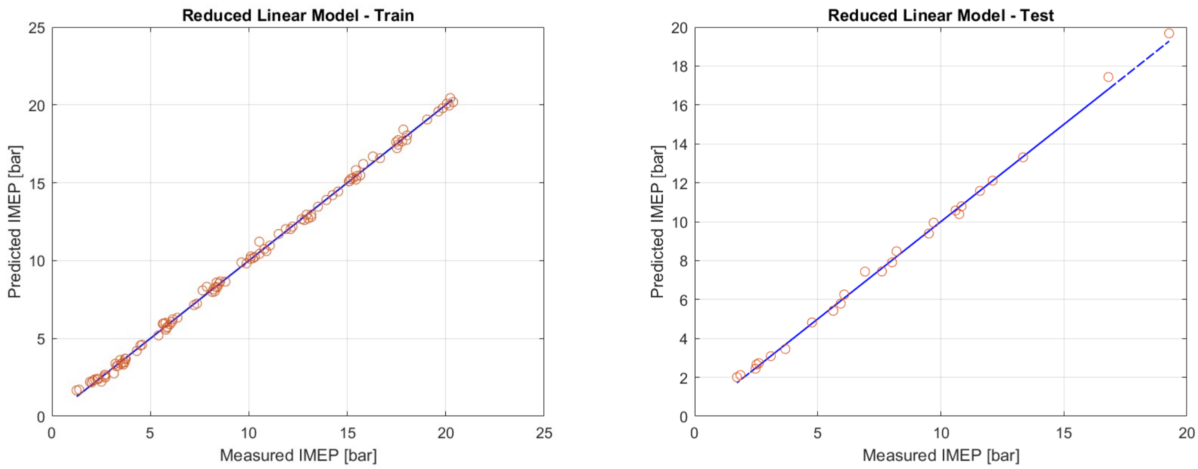

Figure 8: Training and testing plots of the reduced linear model for IMEP estimation

$RMSE_{TRAIN}$	$RMSE_{TEST}$	R^2
0.21 ¿	0.24 ¿	0.999

Table 12: Training and testing performance of the reduced linear model for IMEP estimation

Since the same set of input variables was used for both torque and IMEP models, the condition number and correlation matrix of the input were identical to those previously reported. Therefore, the multicollinearity assessment remains valid for this case.

The residual plot showed that the dispersion was overall contained within a range of approximately ± 0.4 bar. There were also some outliers, characterized by more pronounced deviations of up to approximately −0.6 and −0.7 bar.

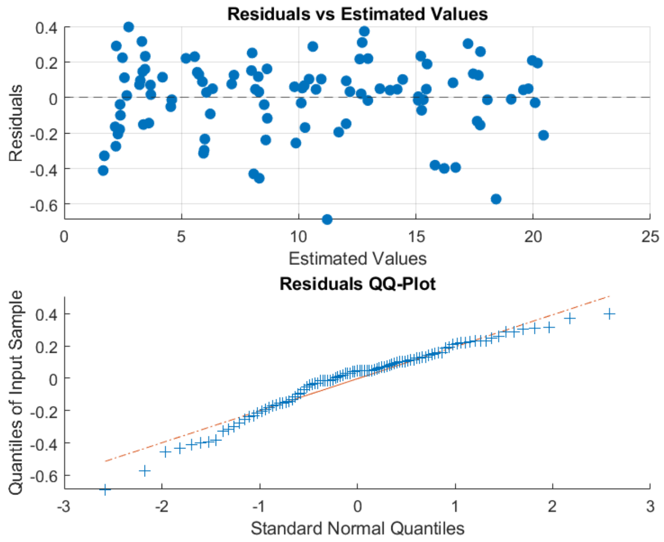

Figure 9: Residual plots of the reduced linear model for IMEP estimation

7.2.3 Neural Network

As it was done for the torque, a neural network model was developed for IMEP estimation.

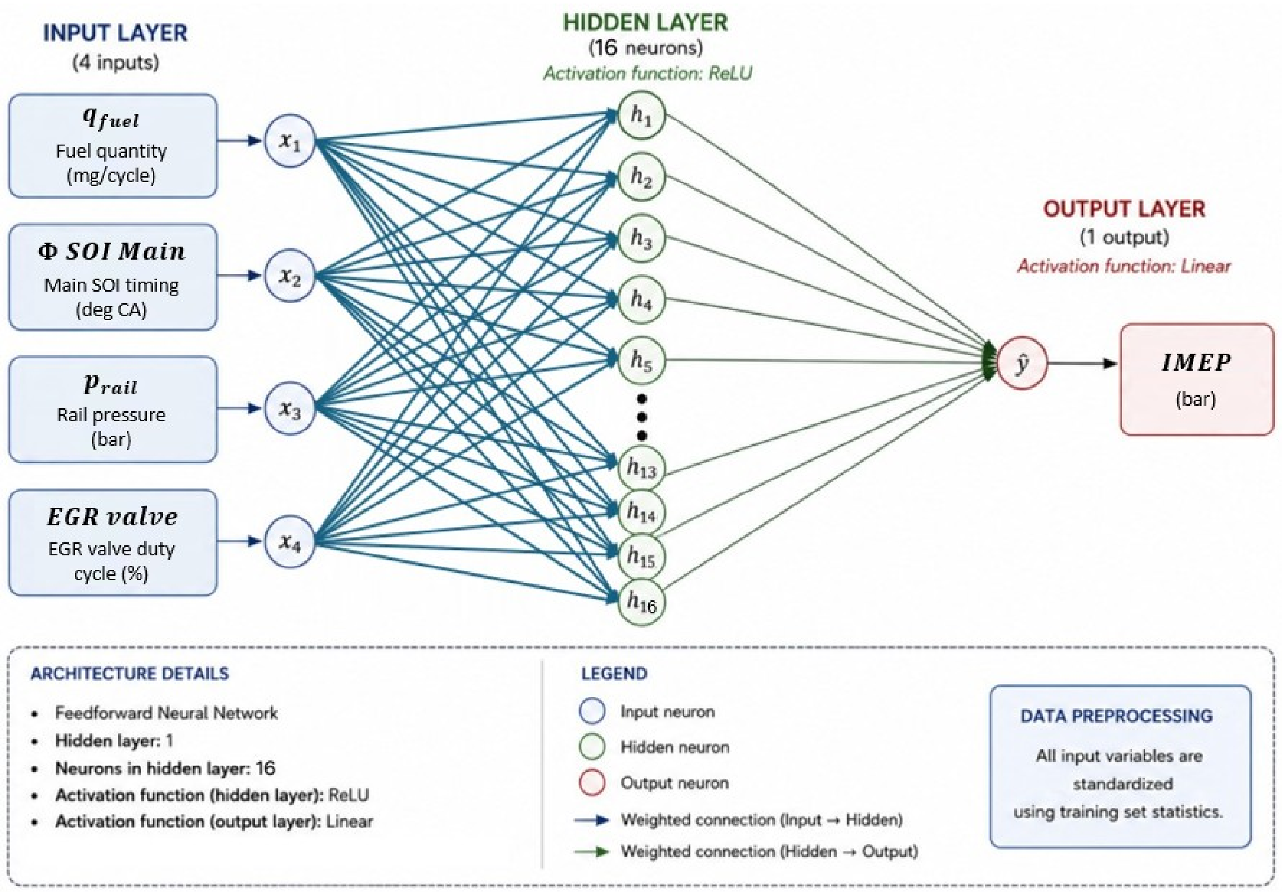

Figure 10: Neural network structure for IMEP estimation

The model performance was evaluated using the Root Mean Squared Error.

$RMSE_{TRAIN}$	$RMSE_{TEST}$
0.10 $\bar{¿}$	0.12 $\bar{¿}$

Table 13: Training and testing RMSE of the neural network for IMEP estimation

The neural network outperformed the linear models in prediction accuracy, demonstrating its capacity to capture the nonlinear dynamics governing IMEP. Moreover, training and test errors indicated satisfactory generalization performance.

7.2.4 Model Comparison

Similarly to the torque estimation case, the full linear model achieved high prediction accuracy for IMEP but suffered from multicollinearity, limiting the interpretability and stability of the estimated coefficients.

The reduced linear model demonstrated good predictive performance on the test set. Moreover, the reduced set of variables mitigated multicollinearity, obtaining stable coefficients that can be directly interpreted from a physical point of view.

The neural network significantly improved prediction accuracy compared to the linear regression model, reflecting its ability to capture nonlinear dependencies between the selected ECU variables and IMEP.

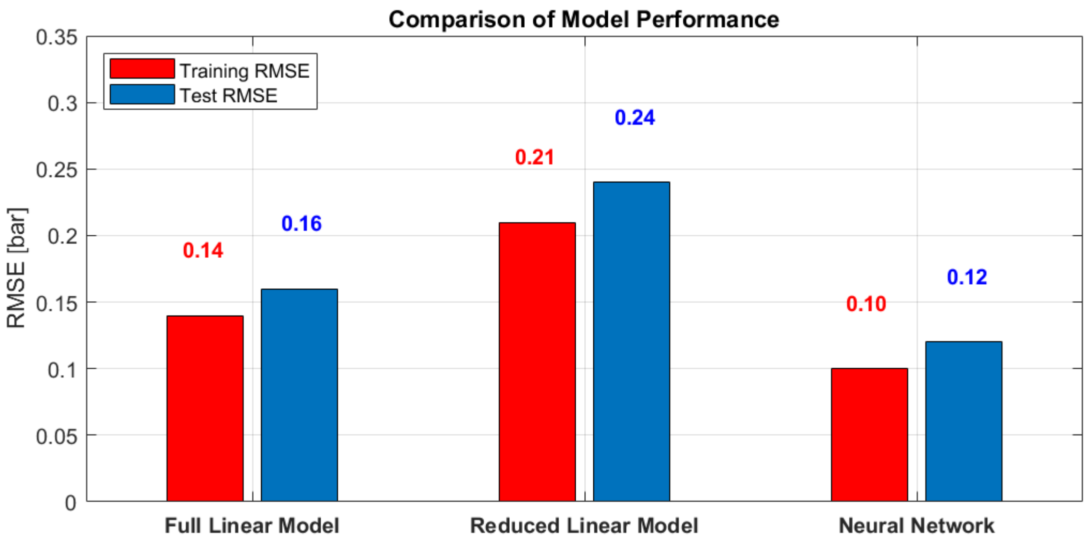

Figure 11: Performance comparison between models for IMEP estimation

7.3 MFB50

7.3.1 Full Linear Model

As it was done for the previous output variables, a full linear model was first constructed.

```
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9

Estimated Coefficients:
                   Estimate        SE         tStat        pValue   
                   _________    ________    ________    ___________

    (Intercept)       10.334    0.035937      287.57    1.9523e-136
    x1                 1.867     0.38424       4.859     4.8938e-06
    x2               0.52543     0.15635      3.3606      0.0011382
    x3                 3.439    0.070694      48.646     6.2909e-67
    x4               -1.0737     0.11779     -9.1153     1.8118e-14
    x5               -2.2297     0.28022     -7.9571     4.6988e-12
    x6              0.029179     0.11749     0.24836        0.80442
    x7               0.11506     0.15667     0.73439         0.4646
    x8              -0.49681     0.15282     -3.2509      0.0016136
    x9             -0.015867    0.090353    -0.17561        0.86099


Number of observations: 101, Error degrees of freedom: 91
Root Mean Squared Error: 0.361
R-squared: 0.992,  Adjusted R-Squared: 0.991
F-statistic vs. constant model: 1.22e+03, p-value = 8.5e-91
```

Table 14: Regression coefficients of the full linear model for MFB50 estimation

Overall, the results showed that MFB50 is primarily driven by the injected fuel quantity and the main injection timing. The other variables considered instead play a secondary in the estimation of this output.

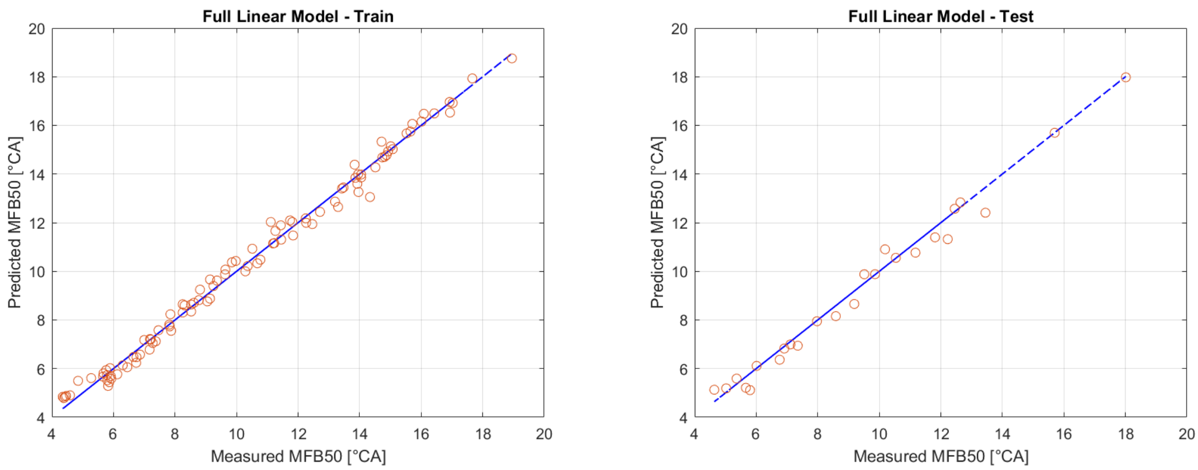

Figure 12: Training and testing plots of the full linear model for MFB50 estimation

$RMSE_{TRAIN}$	$RMSE_{TEST}$	R^2
$0.34\ °CA$	$0.44\ °CA$	0.992

Table 15: Training and testing performance of the full linear model for MFB50 estimation

The model achieved excellent predictive performance while generalising well to unseen data, as evidenced by the small gap between training and test errors and the absence of significant overfitting.

However, as seen before, the full linear model suffered from multicollinearity among the input variables. The VIF values were identical to those reported earlier, as the same set of regressors was used, and are omitted here for brevity.

7.3.2 Reduced Linear Model

Specifically for this output, the following variables were selected as inputs for the reduced linear regression model:

1. **Injected Fuel Quantity**
   The injected fuel quantity influences the rate of the combustion process and, consequently, the crank angle at which MFB50 is reached.
2. **Rail Pressure**
   The rail pressure governs the fuel atomization and the air-fuel mixture, making it a key driver for the combustion.
3. **Main Injection Timing**
   It is one of the most significant variables as it determines the onset of the main combustion event.
4. **Pilot injection Timing**
   It mainly affects the pre-combustion phase and the ignition delay.

As an initial attempt, engine speed was also included in the set of input variables; however, the resulting VIF values remained high, indicating that significant redundancy among the predictors was still present.

$$VIF_{with\ engine\ speed} = \{25.46 \quad 3.32 \quad 18.68 \quad 5.69 \quad 5.16\}$$

When engine speed was excluded from the set of input variables, the VIF values indicated that the multicollinearity issue had been effectively mitigated.

Only the main injection timing exhibited a VIF value slightly above 5 but this did not represent a critical concern, as the value remained relatively close to the threshold and all the other predictors showed acceptable VIF levels.

Input Variable	VIF
Injected Fuel	3.24
SOI Main	5.56
Rail Pressure	4.50
SOI P1	4.69

Table 16: VIF values of the reduced linear model inputs for MFB50 estimation

Based on these variables, the following reduced linear model was developed.

```
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   ________    ________    _______    ___________

    (Intercept)      10.334    0.062902     164.29    1.9335e-119
    x1               3.3386     0.11376     29.347     9.8144e-50
    x2              0.09284     0.14905    0.62286        0.53485
    x3              0.63851     0.13406      4.763     6.7569e-06
    x4             -0.17368     0.13683    -1.2693         0.2074


Number of observations: 101, Error degrees of freedom: 96
Root Mean Squared Error: 0.632
R-squared: 0.973,  Adjusted R-Squared: 0.972
F-statistic vs. constant model: 879, p-value = 1.13e-74
```

Table 17: Regression coefficients of the reduced linear model for MFB50 estimation

Compared to the full model, the linear one predicted the output variable with higher errors.

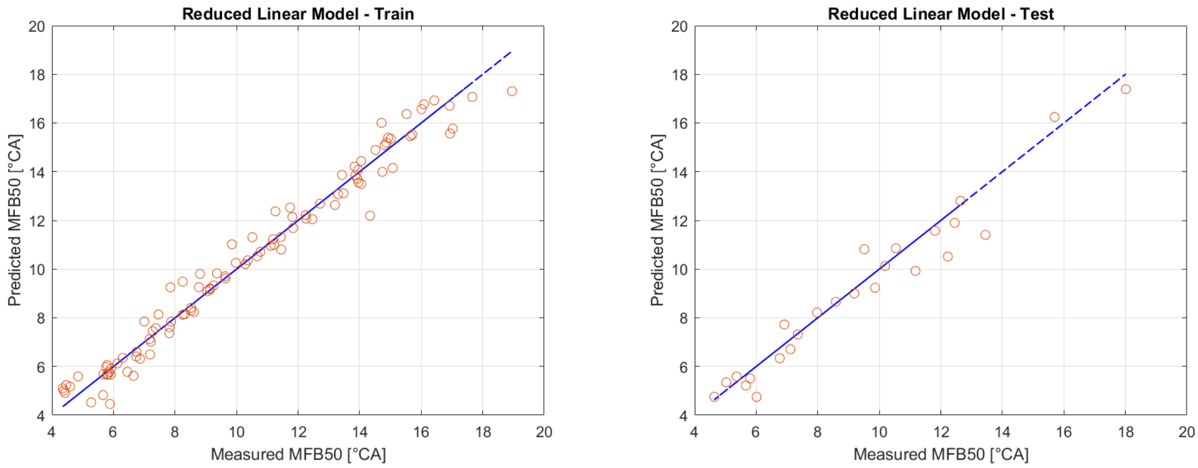

Figure 13: Training and testing plots of the reduced linear model for MFB50 estimation

$RMSE_{TRAIN}$	$RMSE_{TEST}$	R^2
$0.62\ {}^\circ CA$	$0.78\ {}^\circ CA$	0.973

Table 18: Training and testing performance of the reduced linear model for MFB50 estimation

The correlation matrix showed that only one pair of variables exhibited a correlation coefficient close to ±1, namely the start of injection timings of the main and pilot injections. The same correlation could also be observed in the VIF values, where these variables exhibited VIF values close to 5.

Nevertheless, the condition number of the input matrix, equal to 4.65, confirmed that this residual correlation did not constitute a critical multicollinearity issue.

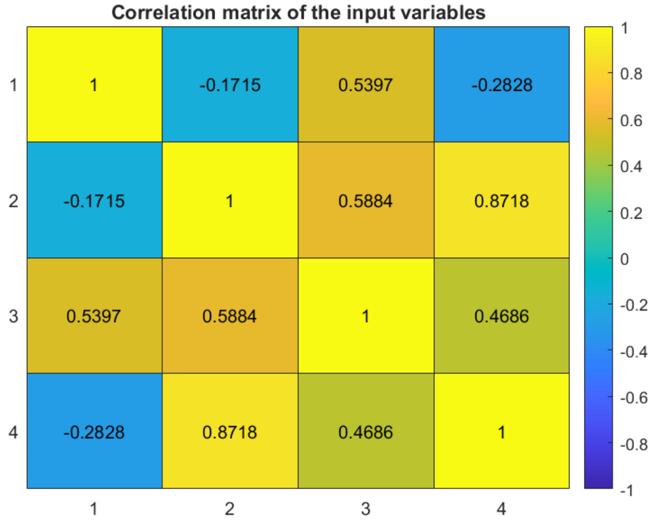

Figure 14: Correlation matrix of the selected input variables for MFB50 estimation

The residuals were randomly distributed around zero, with deviations confined within the range ±2 °CA.

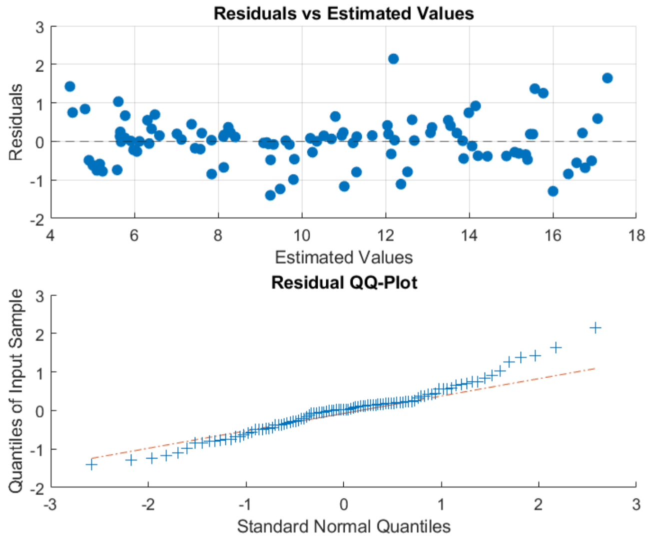

Figure 15: Residual plots of the reduced linear model for MFB50 estimation

7.3.3 Quadratic Model

To capture simple nonlinearities in the system, a quadratic model was estimated using the same inputs as the reduced linear one. The objective was to enhance the model's prediction capability by introducing nonlinear terms while maintaining the level of complexity low.

The model achieved satisfactory performance on the training set, with an RMSE of 0.43 °CA. indicating a good fit to the data. However, performance deteriorated significantly during the test phase, with an RMSE of 12.50 °CA. This sharp increase in the error clearly indicated the presence of overfitting, suggesting that the quadratic model was not appropriate for the MFB50.

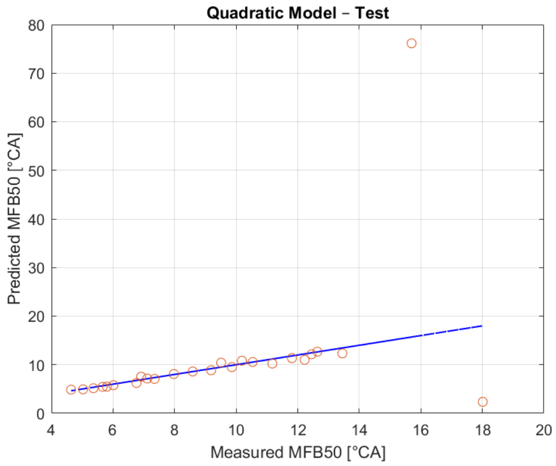

Figure 16: Testing plot of the quadratic model for MFB50 estimation

7.3.4 Neural Network

Finally, a Fully Connected Feedforward Neural Network was built, whose architecture is illustrated in the figure below.

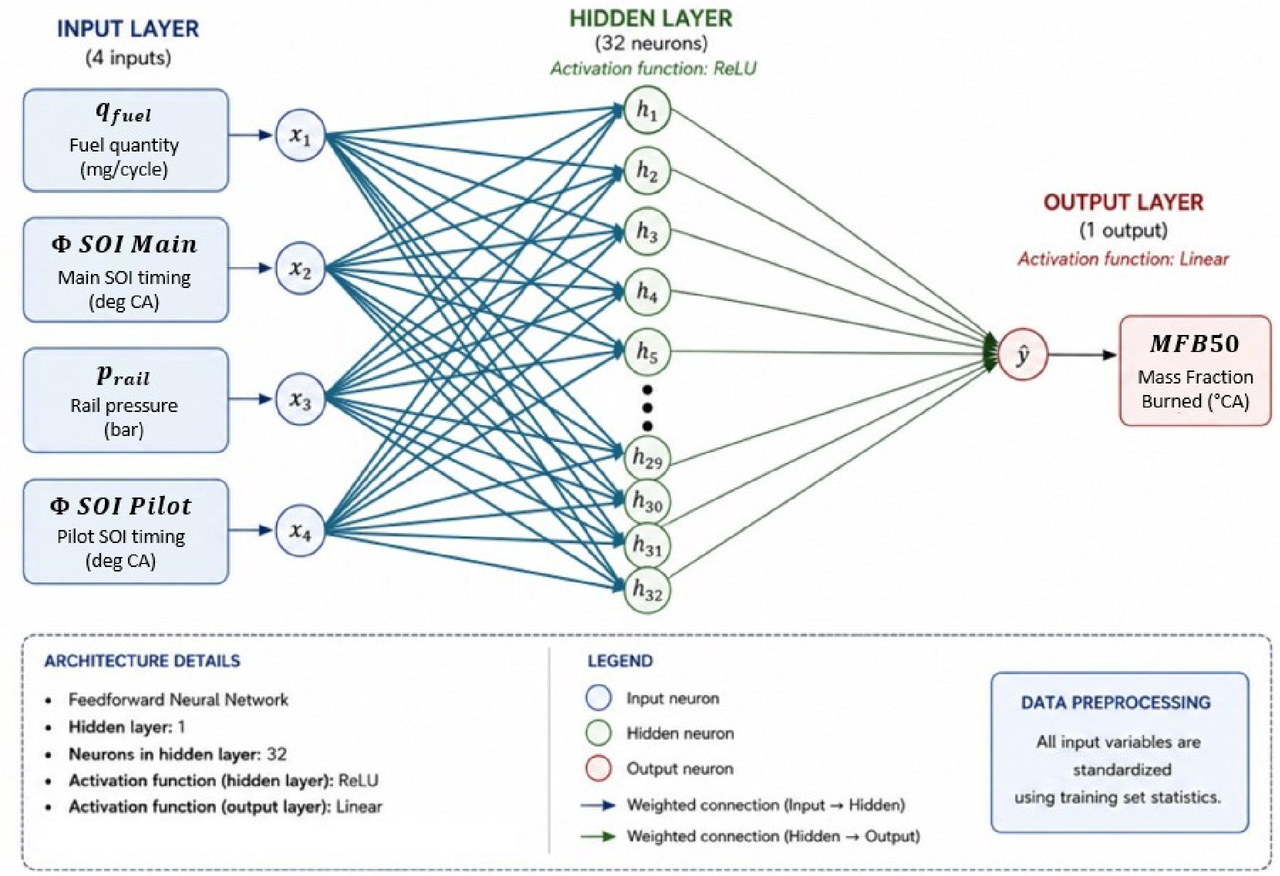

Figure 17: Neural network structure for MFB50 estimation

The model performance was evaluated using the Root Mean Squared Error (RMSE).

$RMSE_{TRAIN}$	$RMSE_{TEST}$
$0.14\ °CA$	$0.28\ °CA$

Table 19: Training and testing RMSE of the neural network for MFB50 estimation

The neural network achieved better predictive accuracy than the linear models, demonstrating its ability to capture the nonlinear relationships underlying MFB50. Furthermore, the small gap between training and test errors indicates good generalization performance and no evidence of overfitting.

7.3.5 Model Comparison

Similarly to the torque estimation case, the full linear model achieved high prediction accuracy for MFB50 but suffered from multicollinearity among the input variables, limiting the interpretability and stability of the estimated coefficients.

The reduced linear model achieved good predictive performance, with an RMSE of 0.78 °CA on the test dataset. Furthermore, the reduced set of variables successfully mitigated multicollinearity issues, resulting in stable coefficients that can be directly interpreted from a physical perspective.

Instead, the quadratic model was not suitable for the estimation of the MFB50 due to the poor performance on the test dataset and its overfit on the training set.

The neural network significantly improves prediction accuracy compared to the linear regression model. These nonlinearities are primarily associated with combustion-related phenomena, which cannot be fully captured by a linear modelling approach. By leveraging nonlinear activation functions in the hidden layer, the neural network is able to learn these complex interactions and provide more accurate MFB50 predictions.

However, this gain in predictive performance comes at the expense of interpretability. While the coefficients of the linear model can be directly associated with physical effects, the neural network operates as a black-box model, whose internal parameters do not have a straightforward physical interpretation.

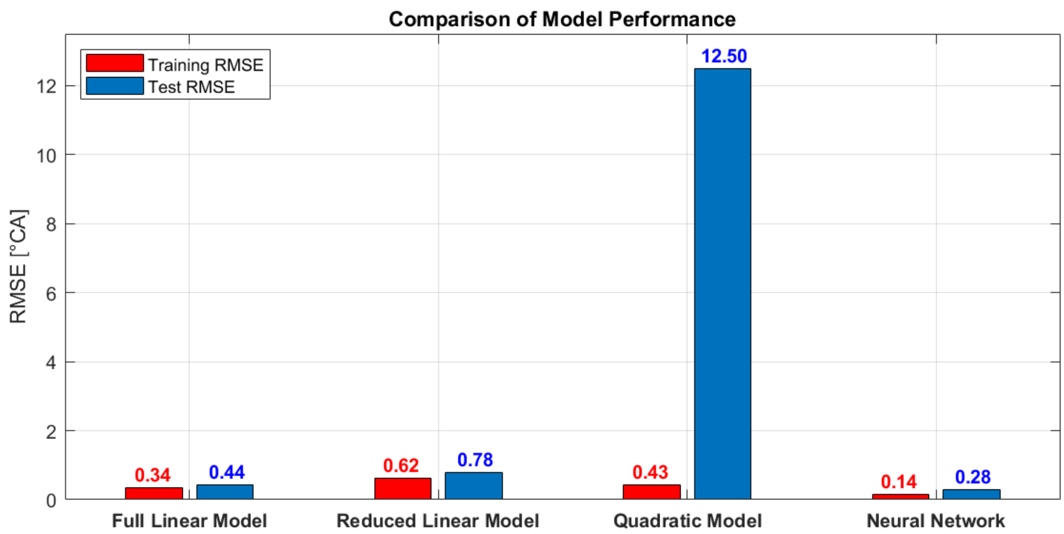

Figure 18: Performance comparison between models for MFB50 estimation

7.4 Peak Pressure

7.4.1 Full Linear Model

```
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9

Estimated Coefficients:
                   Estimate        SE        tStat        pValue
                   ________    _______     ________    ___________

    (Intercept)       97.52    0.14054       693.89    3.1361e-171
    x1              -10.172     1.5027      -6.7693     1.2334e-09
    x2              -3.3275    0.61145       -5.442     4.4193e-07
    x3                20.89    0.27647       75.559     6.6731e-84
    x4              0.50504    0.46065       1.0964        0.27582
    x5               9.6076     1.0959        8.767      9.697e-14
    x6               11.312    0.45946       24.621     5.1844e-42
    x7             -0.74635    0.61272      -1.2181        0.22634
    x8               2.7858    0.59766       4.6613     1.0715e-05
    x9             -0.10963    0.35335     -0.31026        0.75708


Number of observations: 101, Error degrees of freedom: 91
Root Mean Squared Error: 1.41
R-squared: 0.998,  Adjusted R-Squared: 0.997
F-statistic vs. constant model: 4.19e+03, p-value = 4.79e-115
```

Table 20: Regression coefficients of the full linear model for peak pressure estimation

Looking at the coefficients of the full linear model, it was clear that the injected fuel quantity was by far the most influential variable on peak pressure. Rail pressure and main injection timing also played a significant role.

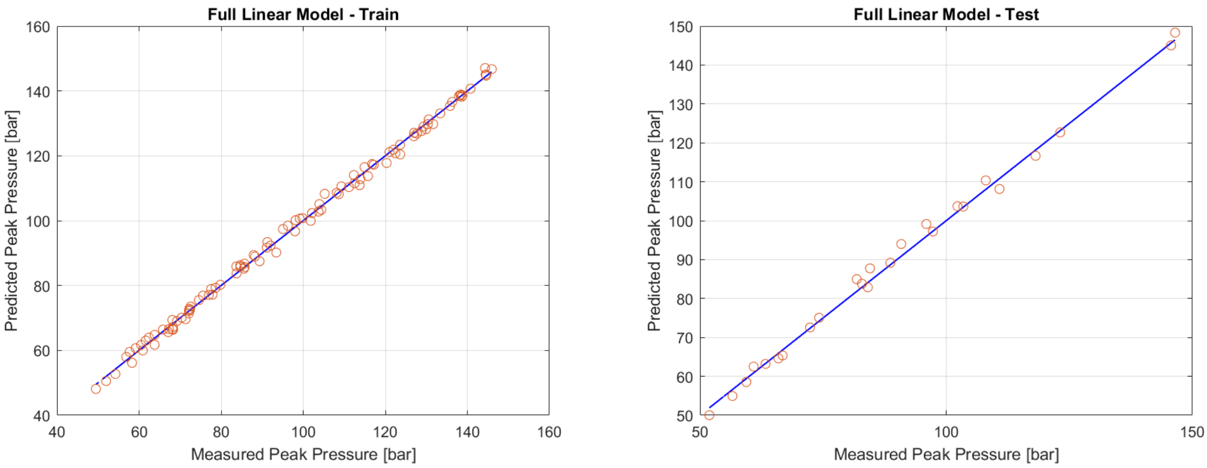

Figure 19: Training and testing plots of the full linear model for peak pressure estimation

$RMSE_{TRAIN}$	$RMSE_{TEST}$	R^2
1.34 $\bar{¿}$	1.76 $\bar{¿}$	0.998

Table 21: Training and testing performance of the full linear model for peak pressure estimation

The same considerations made for the other full linear models apply here as well. In particular, a further variable selection step is needed to improve the reliability of the regression coefficients.

7.4.2 Reduced Linear Model

Three input variables were selected to predict peak cylinder pressure, explaining for most of the output variability:

1. **Injected Fuel Quantity**
   It has a direct influence on the pressure reached in the combustion chamber of the engine.
2. **Rail Pressure**
   It governs the fuel atomization and the efficiency of the combustion process.
3. **Main Injection Timing**
   It strongly influences the rate of heat release and the resulting peak pressure in the cylinder.

The corresponding VIF values suggested that these three variables were suitable for building the reduced model.

Input Variable	VIF
Injected Fuel	2.92
SOI Main	3.16
Rail Pressure	4.33

Table 22: VIF values of the reduced linear model inputs for peak pressure estimation

The calculated coefficients of the reduced linear regression model are reported below.

```
Estimated Coefficients:
                   Estimate       SE        tStat       pValue   
                   ________    _______    ______    ___________

    (Intercept)      97.52     0.22885    426.14    1.5426e-160
    x1              20.581     0.39274    52.404     5.8005e-73
    x2              1.0419     0.40891    2.5479       0.012408
    x3              9.9242     0.47854    20.739     1.9843e-37


Number of observations: 101, Error degrees of freedom: 97
Root Mean Squared Error: 2.3
R-squared: 0.993,  Adjusted R-Squared: 0.993
F-statistic vs. constant model: 4.72e+03, p-value = 6.31e-105
```

Table 23: Regression coefficients of the reduced linear model for peak pressure estimation

The predicted values on the test set closely follow the experimental ones, confirming the good predictive capability of the model.

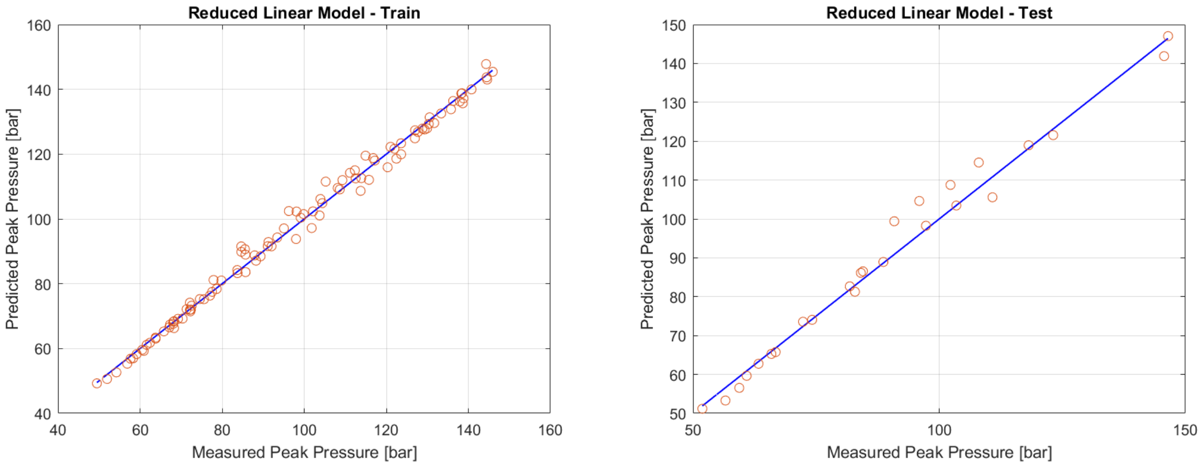

Figure 20: Training and testing plots of the reduced linear model for peak pressure estimation

$RMSE_{TRAIN}$	$RMSE_{TEST}$	R^2
2.25 ¿	3.54 ¿	0.993

Table 24: Training and testing performance of the reduced linear model for peak pressure estimation

The condition number of 3.93 and the correlation matrix analysis confirmed that the input variables were sufficiently independent of each other.

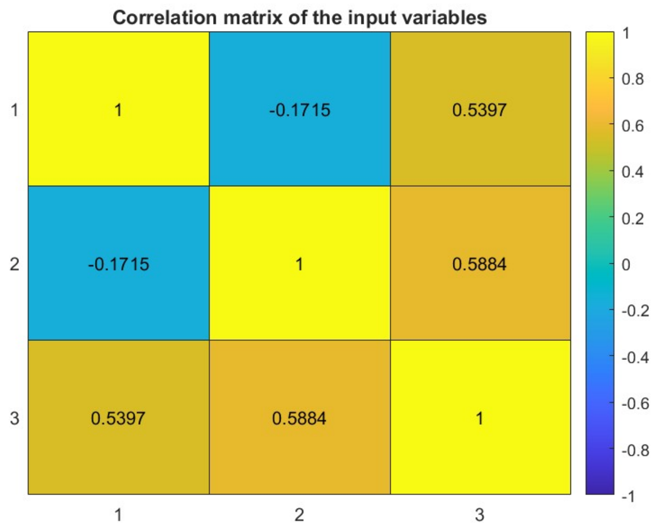

Figure 21: Correlation matrix of the selected input variables for peak pressure estimation

The residual plot revealed that the residuals were not randomly distributed. A higher dispersion was observed in the central region, roughly between 90 and 120 bar, suggesting heteroscedasticity and possible non-linear effects not fully captured by the model.

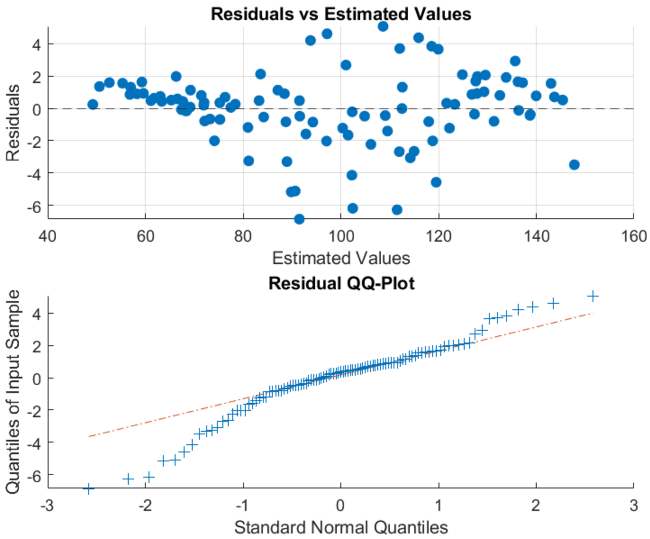

Figure 22: Residual plots of the reduced linear model for peak pressure estimation

7.4.3 Neural Network

A fully connected feedforward neural network was then trained to capture the nonlinear behaviour of the peak pressure.

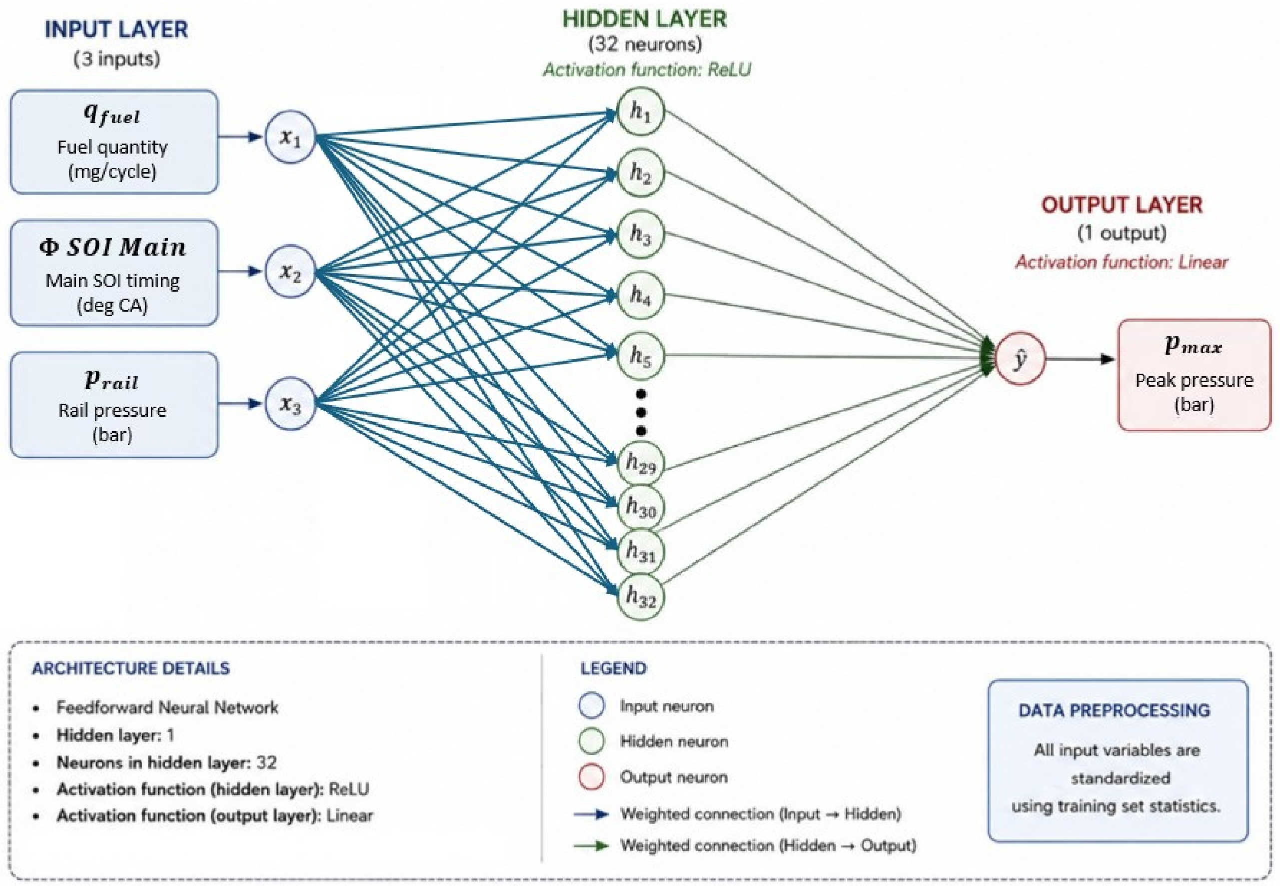

Figure 23: Neural network structure for peak pressure estimation

$RMSE_{TRAIN}$	$RMSE_{TEST}$
0.85 ¿	0.90 ¿

Table 25: Training and testing RMSE of the neural network for peak pressure estimation

The improvement achieved with the neural network over the linear models is especially evident when compared to the reduced linear model. The linear fit was already quite good, yet the neural network managed to capture the residual nonlinear relationships that the linear model had failed to account for.

7.4.4 Model Comparison

As in the torque case, the linear model already provided excellent peak pressure estimates. However, the neural network managed to push accuracy beyond the full linear model while avoiding multicollinearity, since it was trained on the reduced set of variables.

The reduced linear model represents a good trade-off between coefficient interpretability and predictive performance. Instead, the neural network is the best option when the focus is the prediction accuracy.

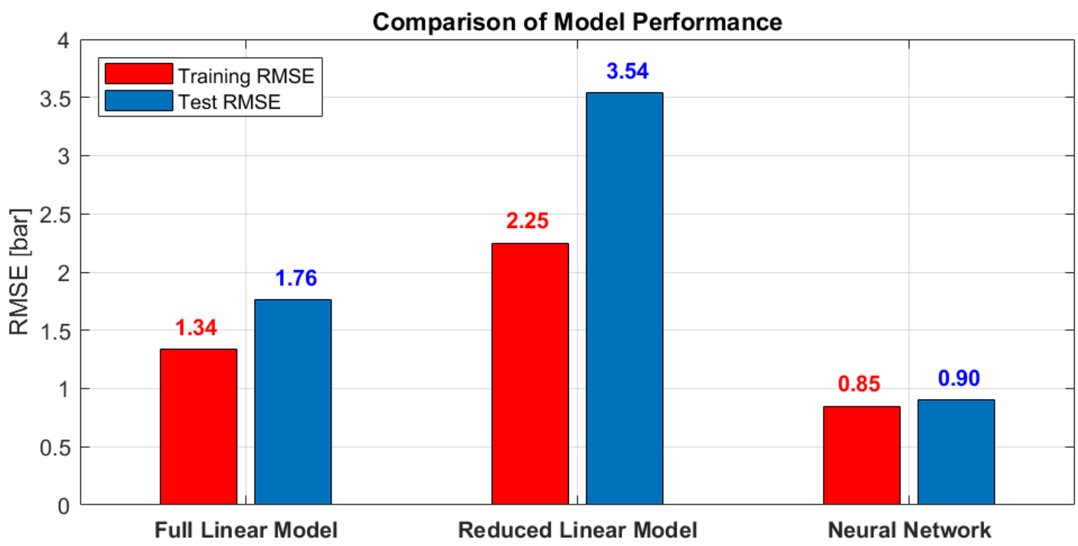

Figure 24: Performance comparison between models for peak pressure estimation

7.5 NOx Emissions (g/kWh)

As mentioned in the introduction, only for NOx emissions, two additional input variables were included in the model in order to better capture the NOx variability:

- **Lambda,** which is a measure of the air–fuel ratio that indicates whether the mixture is rich, stoichiometric, or lean;
- **EGR/lambda,** which captures the combined effect of exhaust gas recirculation and air–fuel ratio.

This choice was motivated by the strong influence of both the air-fuel mixture and exhaust gas recirculation on NOx formation.

7.5.1 Full Linear Model

```
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11

Estimated Coefficients:
                   Estimate       SE        tStat       pValue  
                   ________    ________    ________    __________

    (Intercept)      4.0282    0.077882      51.721    3.4025e-68
    x1              -2.7974      1.1624     -2.4065      0.018173
    x2              0.36235     0.45721     0.79251       0.43017
    x3              0.94555     0.15822      5.9761    4.6342e-08
    x4              0.71058      0.2813      2.5261      0.013303
    x5               1.3335     0.66739      1.9981      0.048759
    x6               1.0167     0.54181      1.8765      0.063859
    x7             -0.34033     0.37977    -0.89614       0.37259
    x8              -2.2806     0.86603     -2.6333     0.0099695
    x9              0.24551     0.19662      1.2486       0.21507
    x10              3.4099     0.54731      6.2302    1.5111e-08
    x11             0.19441     0.34574     0.56229       0.57533


Number of observations: 101, Error degrees of freedom: 89
Root Mean Squared Error: 0.783
R-squared: 0.846,  Adjusted R-Squared: 0.826
F-statistic vs. constant model: 44.3, p-value = 2.5e-31
```

Table 26: Regression coefficients of the full linear model for NOx g/kWh estimation

The linear regression analysis indicates that the NOx emissions were highly correlated with the lambda variable and the EGR as it was expected.

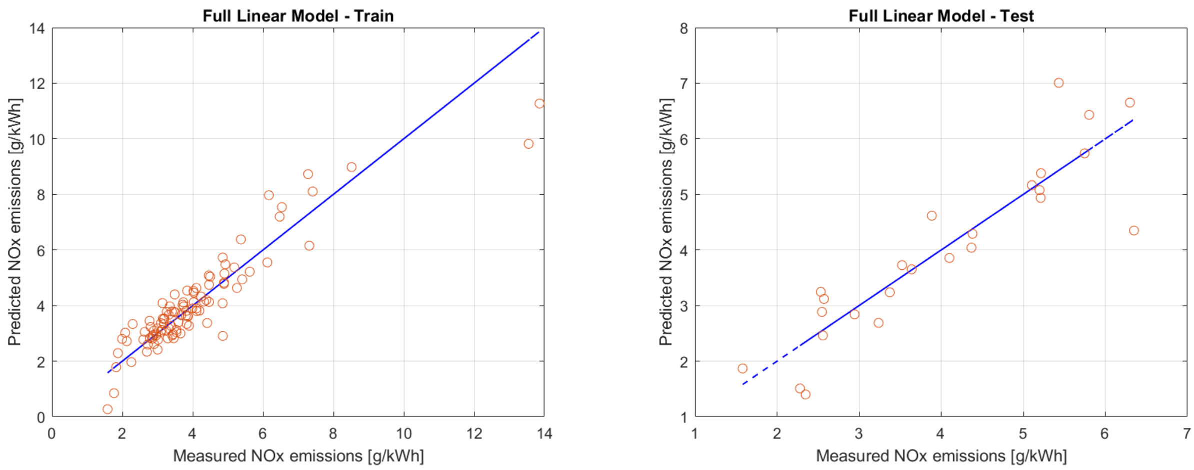

Figure 25: Training and testing plots of the full linear model for NOx g/kWh estimation

$RMSE_{TRAIN}$	$RMSE_{TEST}$	R^2
$0.73 \ g/kWh$	$0.65 \ g/kWh$	0.846

Table 27: Training and testing performance of the full linear model for NOx g/kWh estimation

The plots showed that part of the variability was not captured by the model, since the R^2 coefficient was not so high, and the RMSE error could be improved.

7.5.2 Reduced Linear Model

NOx formation in diesel engines is mainly driven by combustion temperature and oxygen availability. With this in mind, the following variables were selected:

1. **Fuel Quantity**
   This variable directly affects the amount of energy released in the combustion chamber and therefore the gas temperatures.
2. **Main Injection Timing**
   The injection advance governs the peak in-cylinder temperature, influencing NOx formation.
3. **EGR Valve**
   EGR affects the combustion temperature and the available oxygen concentration in the intake charge.
4. **Lambda**
   It directly determines oxygen availability during combustion.

In a fist moment, the EGR/Lambda ratio was also included among the candidate input variables for the model. However, this variable was then excluded from the final model because it slightly worsened the RMSE and significantly increased the Variance Inflation Factor (VIF).

```
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4

Estimated Coefficients:
                   Estimate       SE        tStat       pValue  
                   ________    ________    _______    __________

    (Intercept)      4.0282    0.081067     49.689    2.6949e-70
    x1              0.85241     0.15451     5.5167    2.9203e-07
    x2             -0.57772     0.11699    -4.9381    3.3262e-06
    x3              -1.5071     0.10947    -13.767    1.8805e-24
    x4               2.7521     0.18379     14.974    7.4608e-27


Number of observations: 101, Error degrees of freedom: 96
Root Mean Squared Error: 0.815
R-squared: 0.819,  Adjusted R-Squared: 0.812
F-statistic vs. constant model: 109, p-value = 8.29e-35
```

Table 28: Regression coefficients of the reduced linear model for NOx g/kWh estimation

The reduced model achieved similar results to the full one while using significantly fewer variables, though some margin for improvement remained.

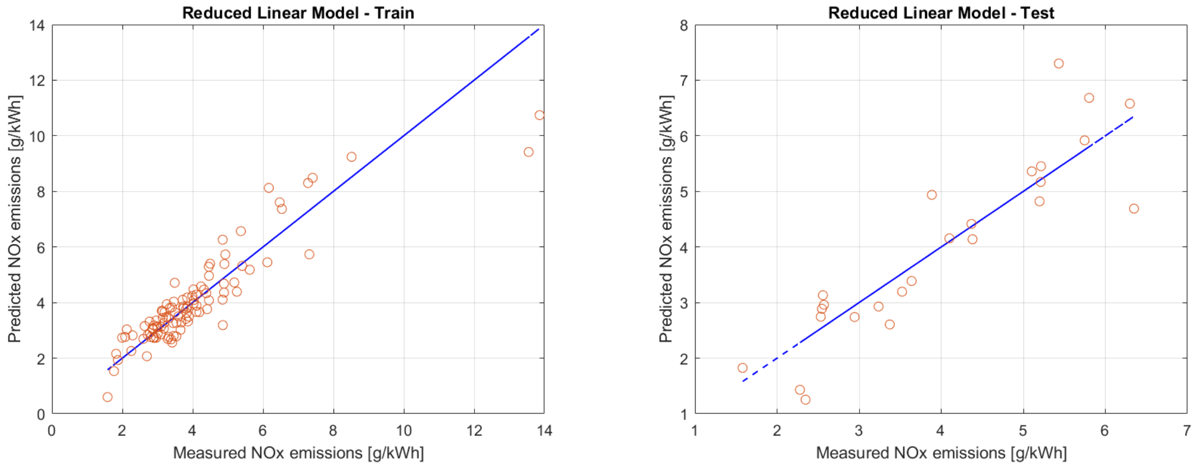

Figure 26: Training and testing plots of the reduced linear model for NOx g/kWh estimation

$RMSE_{TRAIN}$	$RMSE_{TEST}$	R^2
$0.79 \ g/kWh$	$0.69 \ g/kWh$	0.819

Table 29: Training and testing performance of the reduced linear model for NOx g/kWh estimation

The condition number of the input matrix was below 30, confirming that multicollinearity was successfully mitigated. The correlation matrix between the input variables further supported this conclusion.

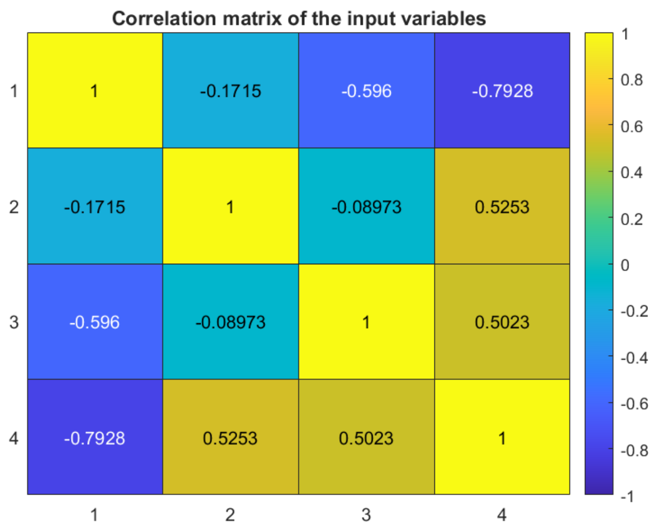

Figure 27: Correlation matrix of the selected input variables for NOx g/kWh estimation

From the residuals plot, few outliers with particularly large prediction errors were observed at high NOx levels, indicating that the model was less accurate under these operating conditions.

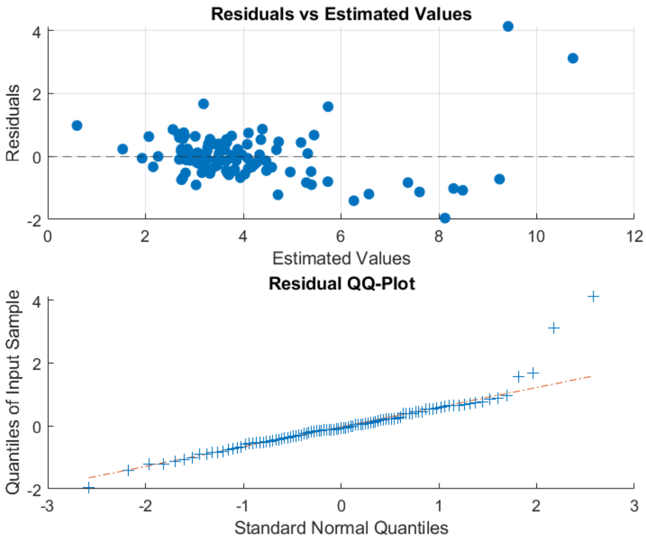

Figure 28: Residual plots of the reduced linear model for NOx g/kWh estimation

7.5.3 Quadratic Model

The quadratic model was explored with the aim of improving the overall RMSE and better capturing NOx emissions at high values.

```
Linear regression model:
    y ~ 1 + x1*x2 + x2*x4 + x3*x4 + x1^2 + x2^2 + x3^2 + x4^2

Estimated Coefficients:
                   Estimate      SE        tStat       pValue
                   ________    _______    _______    __________

    (Intercept)       1.762    0.32182     5.4751    3.9989e-07
    x1              -1.4724    0.45022    -3.2705     0.0015279
    x2               1.1327    0.21655     5.2305    1.1113e-06
    x3             -0.79351    0.17014    -4.6639    1.0864e-05
    x4              -2.2257    0.65707    -3.3873      0.001053
    x1:x2          -0.98574    0.17234    -5.7196    1.4104e-07
    x2:x4          -0.54721    0.21855    -2.5038      0.014109
    x3:x4          -0.26545     0.1085    -2.4466      0.016386
    x1^2            0.31822    0.15718     2.0245      0.045916
    x2^2            0.43566    0.13453     3.2385     0.0016897
    x3^2            0.45436    0.07948     5.7166    1.4288e-07
    x4^2             1.3324    0.19906     6.6934    1.8831e-09


Number of observations: 101, Error degrees of freedom: 89
Root Mean Squared Error: 0.531
R-squared: 0.929,  Adjusted R-Squared: 0.92
F-statistic vs. constant model: 106, p-value = 3.85e-46
```

Table 30: Regression coefficients of the quadratic model for NOx g/kWh estimation

The model included all the linear and quadratic terms of the explanatory variables and some of the interactions. The remaining terms were excluded from the final model because they were not statistically significant according to the stepwise selection.

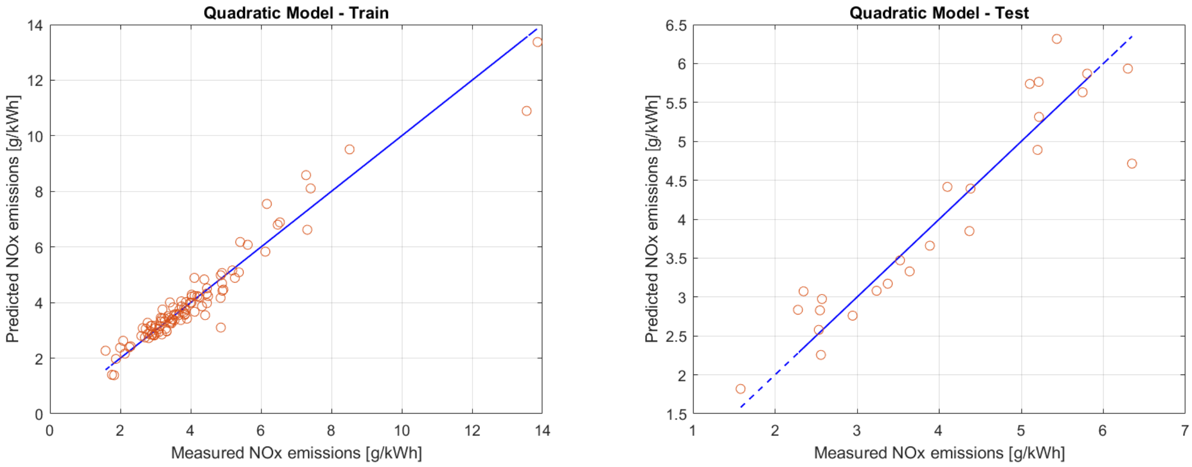

Figure 29: Training and testing plots of the quadratic model for NOx g/kWh estimation

$RMSE_{TRAIN}$	$RMSE_{TEST}$	R^2
$0.50\ \ g/kWh$	$0.52\ \ g/kWh$	0.929

Table 31: Training and testing performance of the quadratic model for NOx g/kWh estimation

Overall, the inclusion of quadratic and mixed terms allowed a substantially improved representation of the system behaviour compared with the linear model, while maintaining a relatively simple model structure.

7.5.4 Power Law Model

In this approach, NOx is expressed as the product of a multiplicative constant and the input variables raised to parametric exponents to be determined.

The MATLAB function requires an initial parameter vector to perform the regression. Each coefficient was initialized based on prior knowledge of the influence that each input variable has on the output.

Coefficient	*Initial Value*	*Input Variable*
β_0	1	*Scaling Factor*
β_1	1	q_{fuel}
β_2	1	ϕ *SOI Main*
β_3	-1	*EGR valve*
β_4	1	*Lambda*

Table 32: Initial parameter values of the power law model for NOx g/kWh estimation

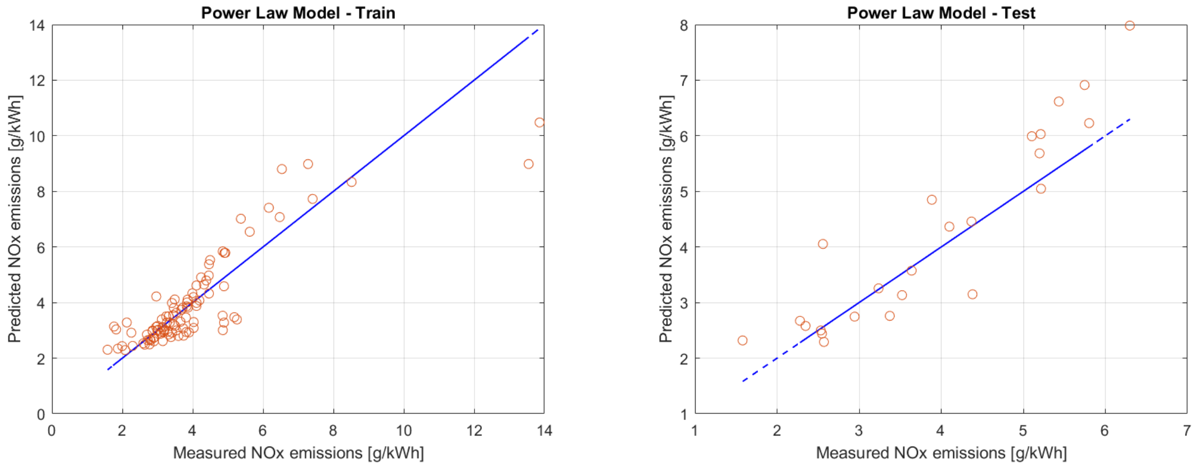

Figure 30: Training and testing plots of the power law model for NOx g/kWh estimation

The power law model performed poorly on both training and test sets, struggling to learn the input-output relationship even during training. Moreover, the RMSE value on the test set was higher than that obtained with the quadratic model, suggesting that the latter was more effective for this application.

$RMSE_{TRAIN}$	$RMSE_{TEST}$
$0.922\ \ g/kWh$	$0.755\ \ g/kWh$

Table 33: Training and testing RMSE of the power law model for NOx g/kWh estimation

7.5.5 Neural Network

The Neural Network in the image below was created to better approximate the strongly nonlinear relationship governing NOx formation.

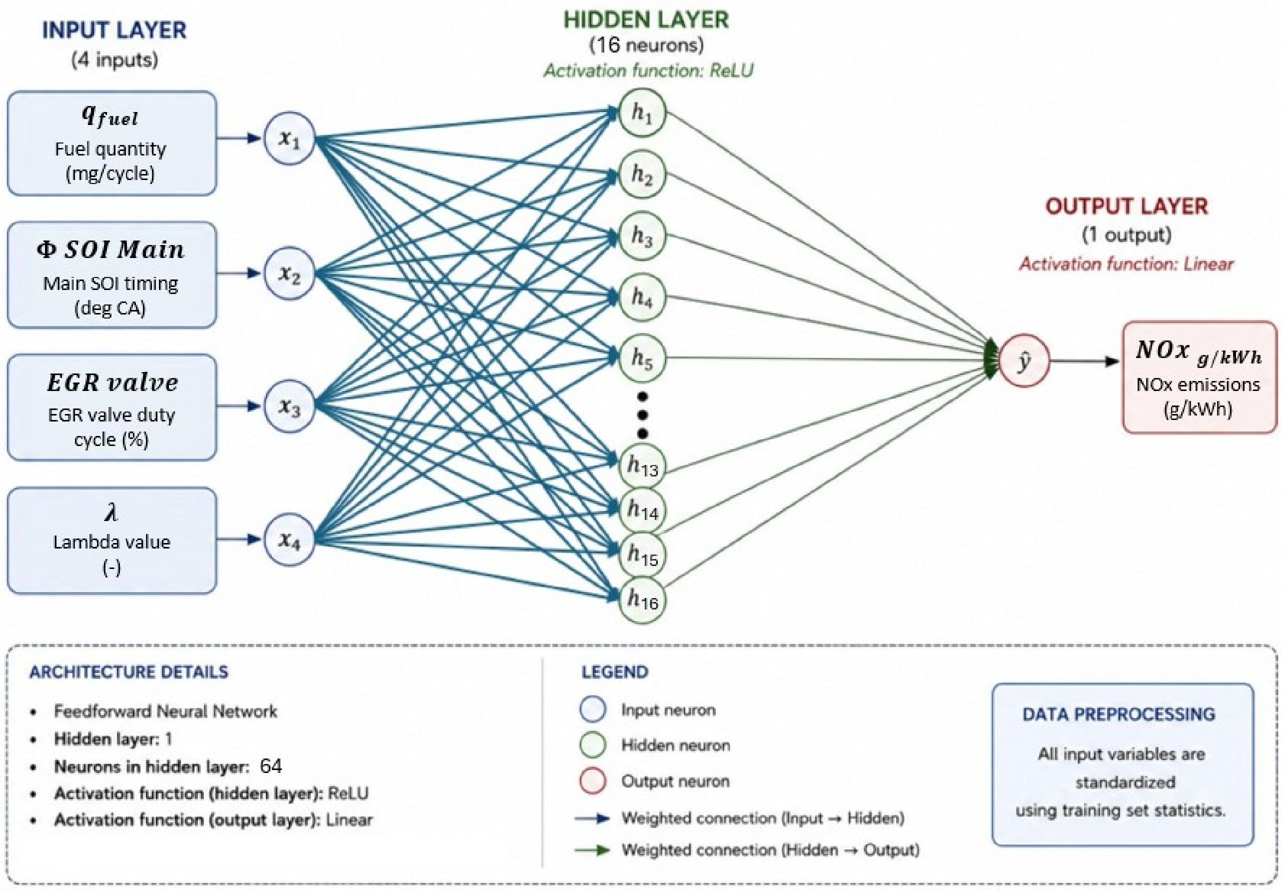

Figure 31: Neural network structure for NOx g/kWh estimation

$RMSE_{TRAIN}$	$RMSE_{TEST}$
$0.15\ g/kWh$	$0.32\ g/kWh$

Table 34: Training and testing RMSE of the neural network for NOx g/kWh estimation

Deeper networks could be explored, though the risk of overfitting would increase, given that even with the current architecture the test error was roughly twice the training error.

7.5.6 Model Comparison

The results showed a clear improvement moving from linear to nonlinear models. The Quadratic Model showed better performance compared to the linear models.

On the contrary, Power Law Model exhibited higher errors in both the training and test sets suggesting that its form was not well suited for this output estimation.

The Neural Network achieves the lowest training and testing errors demonstrating its good ability to learn the nonlinearities of the NOx emissions formation.

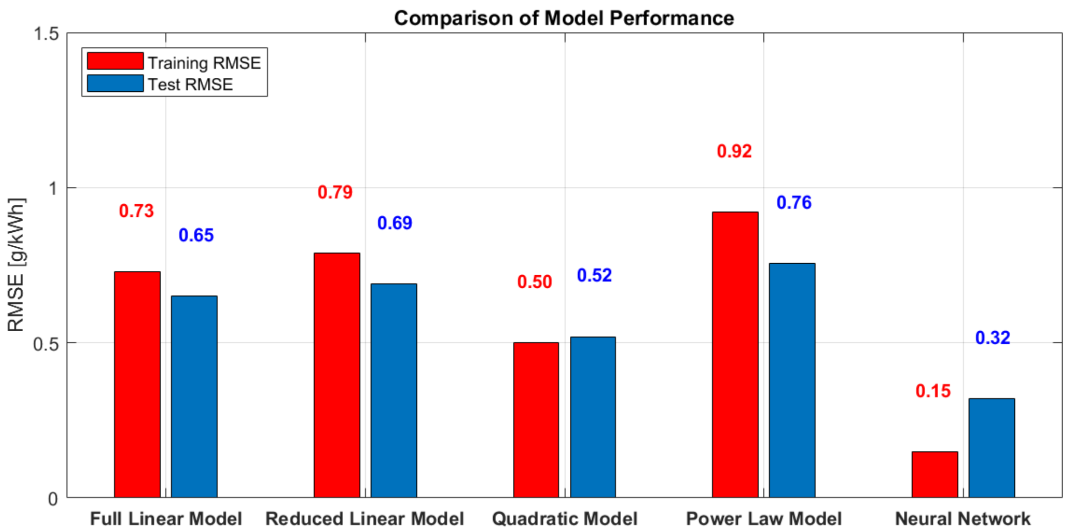

Figure 32: Performance comparison between models for NOx g/kWh estimation

7.6 NOx Emissions (ppm)

NOx concentration (ppm) model was developed using the same additional inputs adopted for NOx emissions (g/kWh), namely lambda and EGR/lambda.

7.6.1 Full Linear Model

Linear regression coefficients obtained considering all the input variables are reported in the following table.

```
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x1

Estimated Coefficients:
                   Estimate      SE        tStat       pValue
                   ________    ______    ________    __________

    (Intercept)      421.46    4.5786      92.051    4.7387e-90
    x1              -31.293    68.338    -0.45792       0.64813
    x2              -26.559    26.879    -0.98811       0.32578
    x3               102.74    9.3016      11.045    2.2877e-18
    x4               117.33    16.537      7.0949    2.9914e-10
    x5               43.386    39.235      1.1058       0.27179
    x6               27.206    31.852     0.85415       0.39532
    x7              -42.437    22.326     -1.9008      0.060566
    x8                204.4    50.913      4.0147    0.00012396
    x9             -0.39116    11.559    -0.03384       0.97308
    x10             -136.99    32.175     -4.2576    5.1067e-05
    x11             -216.85    20.325     -10.669    1.3439e-17


Number of observations: 101, Error degrees of freedom: 89
Root Mean Squared Error: 46
R-squared: 0.955,  Adjusted R-Squared: 0.949
F-statistic vs. constant model: 172, p-value = 6.55e-55
```

Table 35: Regression coefficients of the full linear model for NOx ppm estimation

The linear regression analysis indicates that the NOx emissions are highly correlated with the lambda variable and the ratio EGR/lambda.

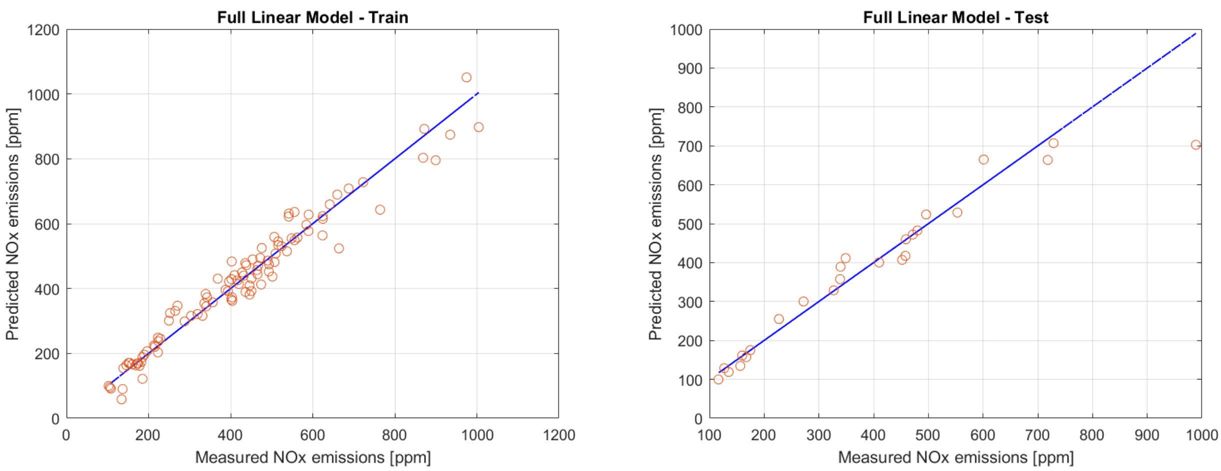

Figure 33: Training and testing plots of the full linear model for NOx ppm estimation

$RMSE_{TRAIN}$	$RMSE_{TEST}$	R^2
43.19 ppm	64.51 ppm	0.955

Table 36: Training and testing performance of the full linear model for NOx ppm estimation

Compared to the NOx emissions case in g/kWh, the linear model appeared to capture the NOx concentration in ppm reasonably well, with an R^2 higher than 0.95.

7.6.2 Reduced Linear Model

For NOx concentration in ppm, EGR/lambda was used as input variable instead of EGR valve because it captures the combined effect of EGR and air-fuel ratio. The set is composed as follows:

1. **Fuel Quantity**
2. **Main Injection Timing**
3. **Lambda**
4. **EGR/lambda**

In a fist moment, the EGR was also included among the candidate input variables for the model. However, this variable was then excluded from the final model because it increased the Variance Inflation Factor (VIF) and the RMSE was worse than considering Lambda and EGR/lambda.

The reduced regression model is reported below.

```
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4

Estimated Coefficients:
                   Estimate      SE        tStat       pValue   
                   ________    ______    ________    __________

    (Intercept)     421.46     6.3581      66.288    5.6917e-82
    x1              103.72     12.076      8.5887    1.5895e-13
    x2             -60.014     8.8948      -6.747    1.1388e-09
    x3             -42.071     13.782     -3.0525     0.0029354
    x4             -105.57     7.0619      -14.95    8.3257e-27


Number of observations: 101, Error degrees of freedom: 96
Root Mean Squared Error: 63.9
R-squared: 0.906,  Adjusted R-Squared: 0.902
F-statistic vs. constant model: 232, p-value = 1.96e-48
```

Table 37: Regression coefficients of the reduced linear model for NOx ppm estimation

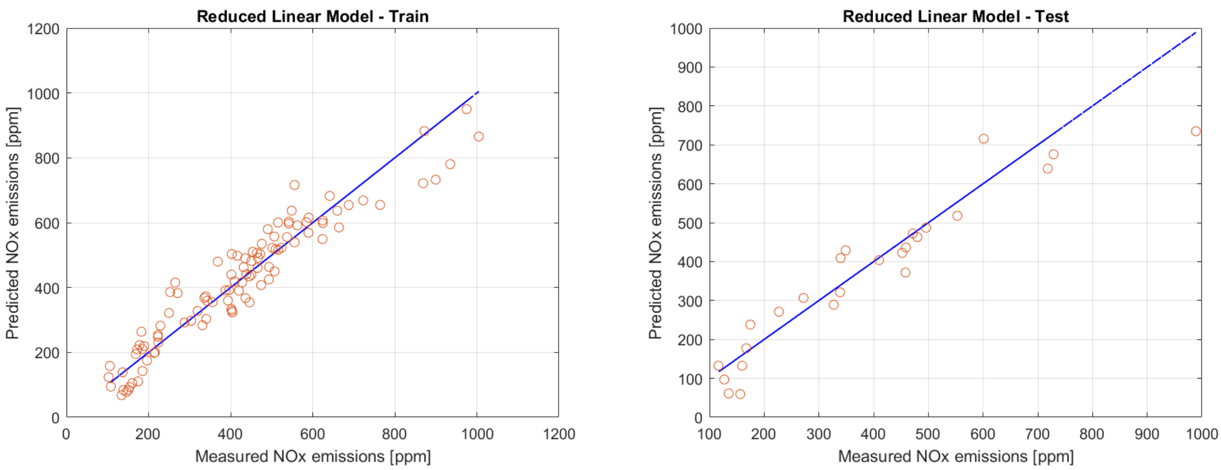

Figure 34: Training and testing plots of the reduced linear model for NOx ppm estimation

$RMSE_{TRAIN}$	$RMSE_{TEST}$	R^2
$62.30\ \ ppm$	$73.28\ \ ppm$	0.906

Table 38: Training and testing performance of the reduced linear model for NOx ppm estimation

Variable selection resolved the multicollinearity issue, as confirmed by the condition number and the correlation matrix, though at the cost of some accuracy.

$$\kappa = 4.10$$

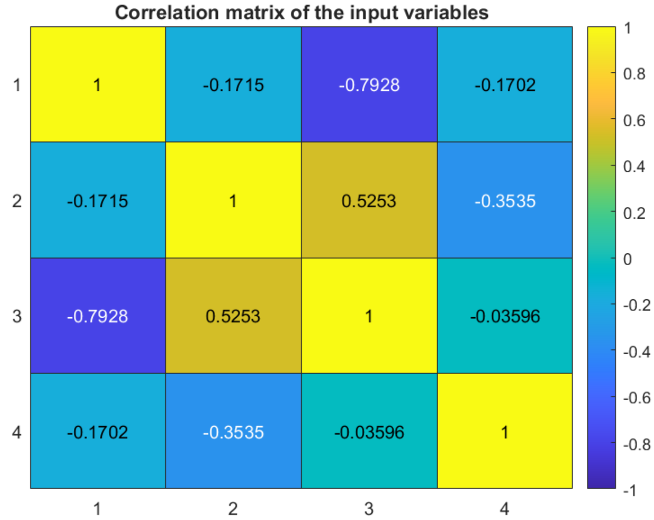

Figure 35: Correlation matrix of the selected input variables for NOx ppm estimation

The residuals were not symmetrically distributed around zero and their spread increased at higher fitted values, indicating heteroscedasticity. Several outliers with large residuals were also present.

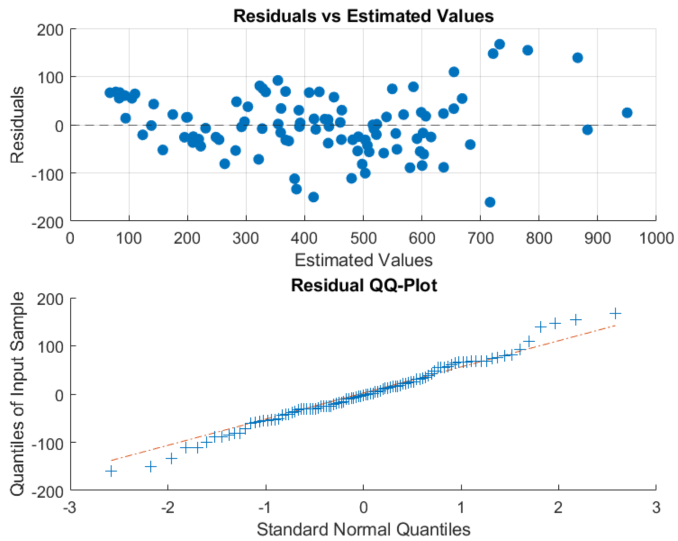

Figure 36: Residual plots of the reduced linear model for NOx ppm estimation

7.6.3 Quadratic Model

A quadratic model was built using the same variables as the reduced linear model, to improve the input-output relationship adding simple nonlinear relations.

```
Linear regression model:
    y ~ 1 + x1*x2 + x2*x3 + x3*x4 + x1^2 + x2^2

Estimated Coefficients:
                   Estimate      SE       tStat       pValue
                   ________    ______    _______    __________

    (Intercept)     395.39     13.654     28.958     1.054e-47
    x1               114.6     17.004     6.7399    1.4113e-09
    x2             -13.051      9.639     -1.354       0.17909
    x3             -57.106     18.642    -3.0633     0.0028797
    x4             -52.141     9.0944    -5.7333    1.2694e-07
    x1:x2          -85.242     10.788    -7.9019    6.1111e-12
    x2:x3          -22.609     11.885    -1.9023      0.060291
    x3:x4           48.436     6.9147     7.0048    4.1568e-10
    x1^2            -22.28     9.5079    -2.3433       0.02129
    x2^2            47.615     9.2159     5.1666    1.3989e-06


Number of observations: 101, Error degrees of freedom: 91
Root Mean Squared Error: 41.9
R-squared: 0.962,  Adjusted R-Squared: 0.958
F-statistic vs. constant model: 255, p-value = 1.63e-60
```

Table 39: Regression coefficients of the quadratic model for NOx ppm estimation

Model included all the linear terms of the explanatory variables, some selected interaction and quadratic contributions. The remaining terms were excluded from the final model because they were not statistically significant according to the stepwise function.

The model performance was evaluated on the training and test datasets.

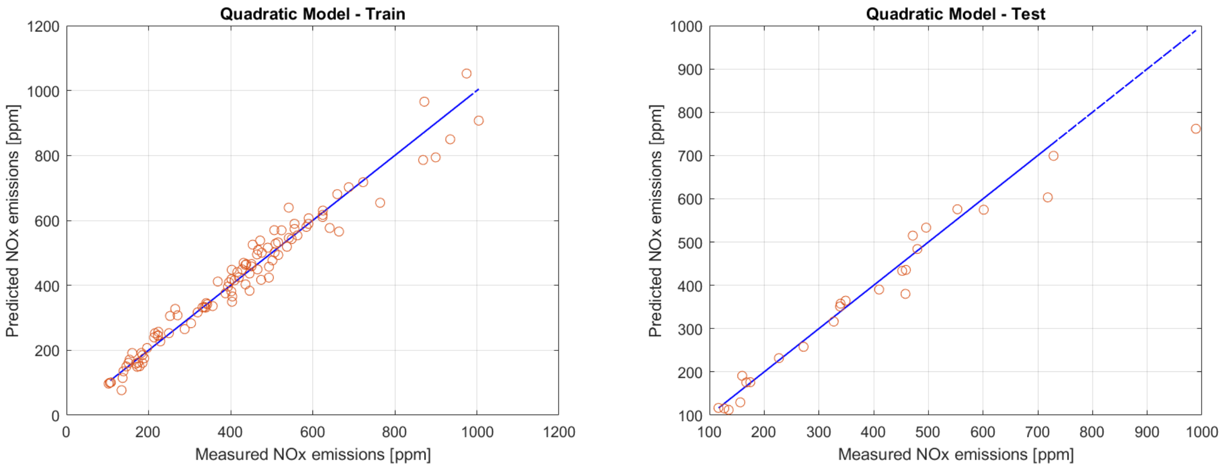

Figure 37: Training and test plots of the quadratic model for NOx ppm estimation

$RMSE_{TRAIN}$	$RMSE_{TEST}$	R^2
39.75 ppm	56.94 ppm	0.962

Table 40: Training and testing performance of the quadratic model for NOx ppm estimation

Overall, the quadratic model captured the system behaviour significantly better than the linear model on the training set, while keeping a relatively simple structure. However, the improvement on the test set was less pronounced, suggesting that there was still space for enhancement.

7.6.4 Power Law Model

Following the same approach used for NOx emissions in g/kWh, the initial parameter values were set before optimization.

Coefficient	*Initial Value*	*Input Variable*
β_0	1	*Scaling Factor*
β_1	1	q_{fuel}
β_2	1	ϕ *SOI Main*
β_3	1	*Lambda*
β_4	-1	*EGR / Lambda*

Table 41: Initial parameter values of the power law model for NOx ppm estimation

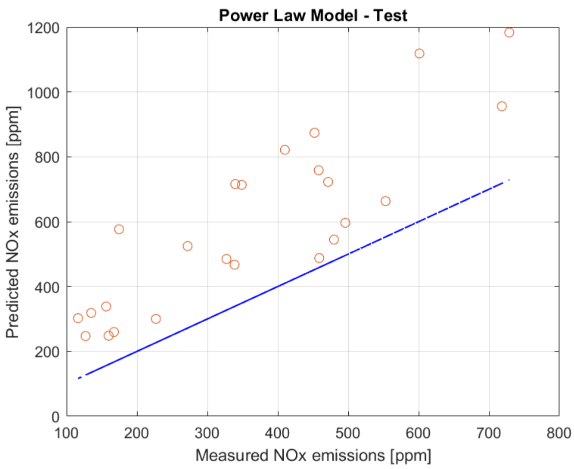

Figure 38: Testing plot of the power law model for NOx ppm estimation

The results obtained on the test set indicated that the Power Law model was not suitable for correctly simulating the output. In particular, the model showed clear signs of overfitting, as highlighted by the high test RMSE.

$RMSE_{TRAIN}$	$RMSE_{TEST}$
56.90 *ppm*	269.08 *ppm*

Table 42: Training and testing RMSE of the power law model for NOx ppm estimation

7.6.5 Neural Network

A fully connected feedforward neural network was developed to improve on the accuracy of the previous regression models

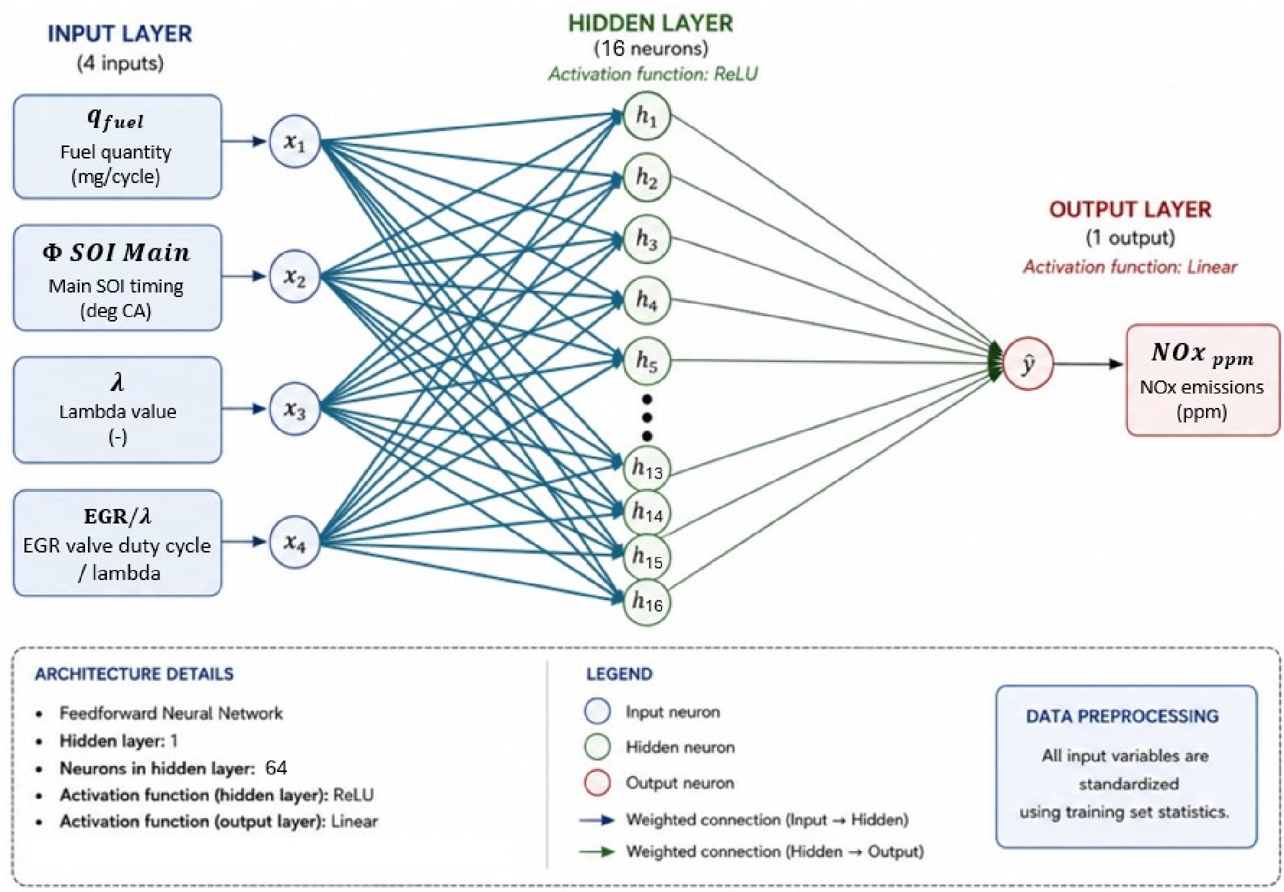

Figure 39: Neural network structure for NOx ppm estimation

The model performance was evaluated using the Root Mean Squared Error (RMSE).

$RMSE_{TRAIN}$	$RMSE_{TEST}$
17.56 ppm	29.76 ppm

Table 43: Training and testing RMSE of the neural network for NOx ppm estimation

The neural network outperformed all previous models, both linear and nonlinear, on training and test sets, confirming its ability to capture the complex relationships governing NOx emissions.

7.6.6 Model Comparison

Looking at the model performance comparison, a decreasing trend in the error was observed moving from linear to nonlinear models, except for the power law model, which performed poorly on the test set. The high test RMSE suggested either coefficient instability or overfitting, confirming that it was not suitable for the data. The quadratic model showed intermediate performance, though the test error remained relatively high. The neural network achieved the lowest errors on both sets, outperforming all other approaches.

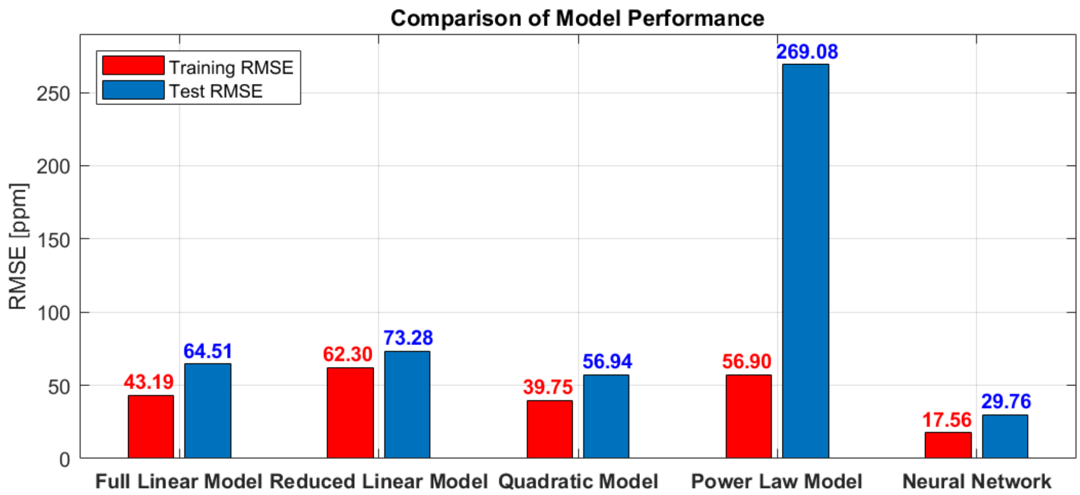

Figure 40: Performance comparison between models for NOx ppm estimation

7.7 Exhaust Temperature Before Turbine

7.7.1 Full Linear Model

The following table reports the regression coefficients of the full linear model. The additional input variables considered for the NOx emissions, for this output estimation were neglected.

```
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9

Estimated Coefficients:
                   Estimate      SE       tStat       pValue   
                   ________    ______    _______    ___________

    (Intercept)     689.32     1.1523     598.23    2.2798e-165
    x1             -30.662      12.32    -2.4888       0.014636
    x2               13.76     5.0132     2.7447      0.0072966
    x3              101.27     2.2667     44.676     1.0704e-63
    x4              4.6514     3.7768     1.2316        0.22129
    x5              12.822     8.9849      1.427          0.157
    x6              50.231      3.767     13.334     4.0334e-23
    x7              2.2588     5.0235    0.44964        0.65404
    x8             -41.875     4.9001    -8.5458     2.8095e-13
    x9              1.1723     2.8971    0.40466        0.68668


Number of observations: 101, Error degrees of freedom: 91
Root Mean Squared Error: 11.6
R-squared: 0.994,  Adjusted R-Squared: 0.994
F-statistic vs. constant model: 1.77e+03, p-value = 4.06e-98
```

Table 44: Regression coefficients of the full linear model for temperature before turbine estimation

Among all input variables, the injected fuel quantity showed the strongest influence on the temperature, followed by rail pressure and EGR valve.

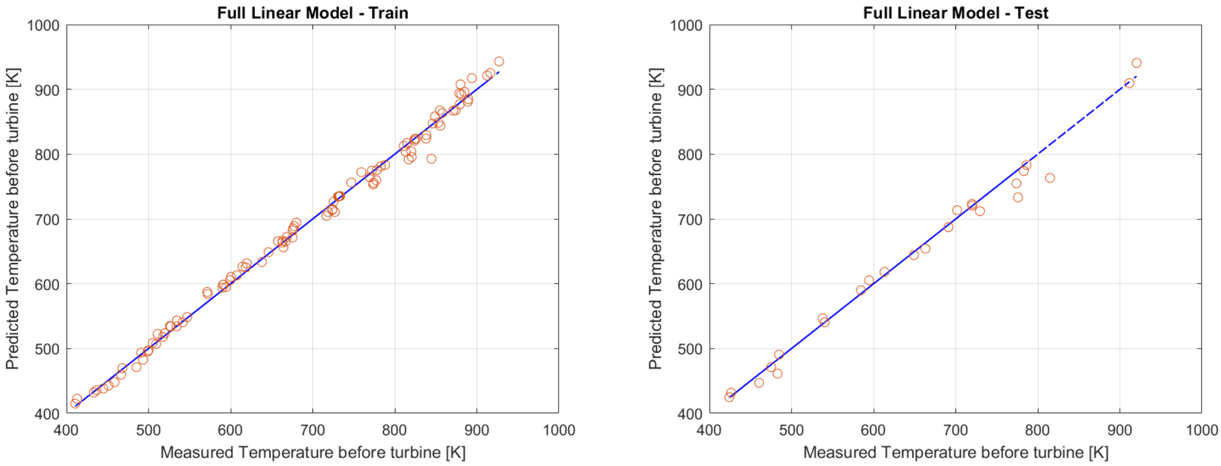

Figure 41: Training and testing plots of the full linear model for temperature before turbine estimation

The predictive performance of the model on the training and test sets is reported below.

$RMSE_{TRAIN}$	$RMSE_{TEST}$	R^2
$10.99\ K$	$16.63\ K$	0.994

Table 45: Training and testing performance of the full linear model

Since the test error was quite close to the training one, overfitting was not a concern. However, the multicollinearity problem among the input variables needed to be solved.

7.7.2 Reduced Linear Model

The variables selected based on physical relevance for predicting temperature before the turbine were the following:

1. **Injected Fuel Quantity**
   Exhaust temperature is strongly correlated with the heat released during combustion and therefore with the injected fuel quantity.
2. **Rail Pressure**
   Higher rail pressure improves fuel atomization and air-fuel mixing, promoting faster and more complete combustion.
3. **Main Injection Timing**
   It strongly influences the rate of heat release and the resulting temperature at the exhaust.
4. **EGR Valve**
   Since the EGR introduces gases that contain less oxygen, it dilutes the fresh charge and reduces the peak combustion temperature.

The selected variables were the same as those used for other outputs analysed previously, so the absence of multicollinearity was already guaranteed.

```
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4

Estimated Coefficients:
                   Estimate      SE       tStat       pValue
                   ________    ______    _______    __________

    (Intercept)     689.32     1.3032     528.96    4.0123e-168
    x1              104.59     2.4049     43.491     5.6343e-65
    x2             -13.918     2.3329     -5.966     4.0568e-08
    x3              37.424     2.7386     13.665      3.019e-24
    x4             -25.425     1.6895    -15.049     5.3303e-27


Number of observations: 101, Error degrees of freedom: 96
Root Mean Squared Error: 13.1
R-squared: 0.992,  Adjusted R-Squared: 0.992
F-statistic vs. constant model: 3.11e+03, p-value = 1.33e-100
```

Table 46: Regression coefficients of the reduced linear model for temperature before turbine estimation

The obtained model was then evaluated on the test set, and the results were comparable to the training ones.

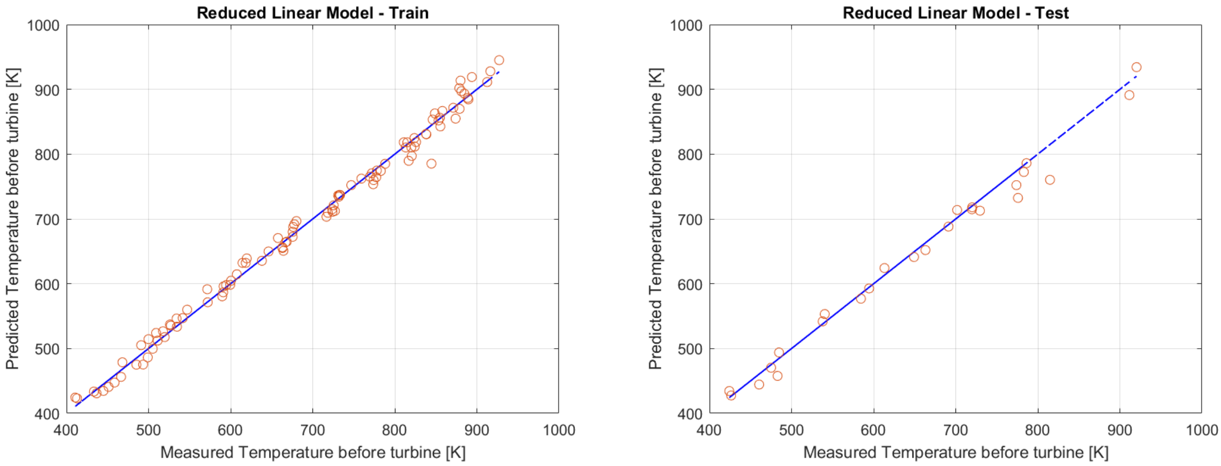

Figure 42: Training and testing plots of the reduced linear model for temperature before turbine estimation

$RMSE_{TRAIN}$	$RMSE_{TEST}$	R^2
12.77 K	18.09 K	0.992

Table 47: Training and testing performance of the reduced linear model for temperature before turbine estimation

The residual analysis showed a slight inverted "U-shaped" pattern. There were positive residuals at intermediate values in the range 700–800 K and negative residuals at temperatures higher than 850 K. It was also visible an outlier for a temperature around 880 K that had a high prediction error of 60 K.

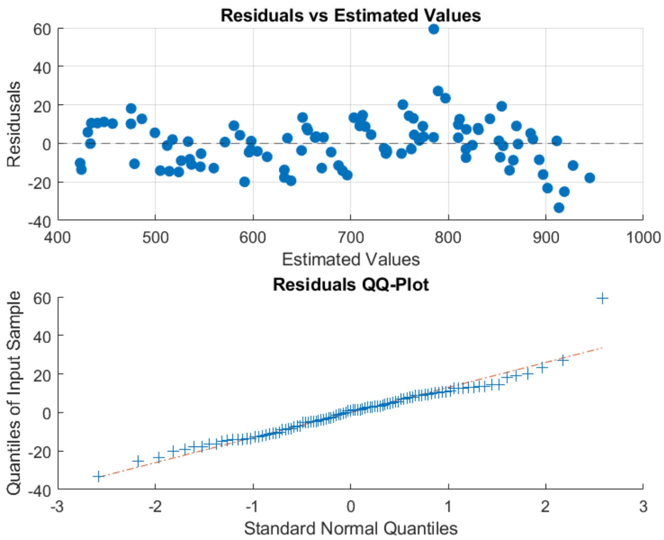

Figure 43: Residual plots of the reduced linear model for temperature before turbine estimation

7.7.3 Neural Network

Since the linear model already provided excellent results, a shallow neural network with few neurons was developed, as the relationship was expected to be not so complex.

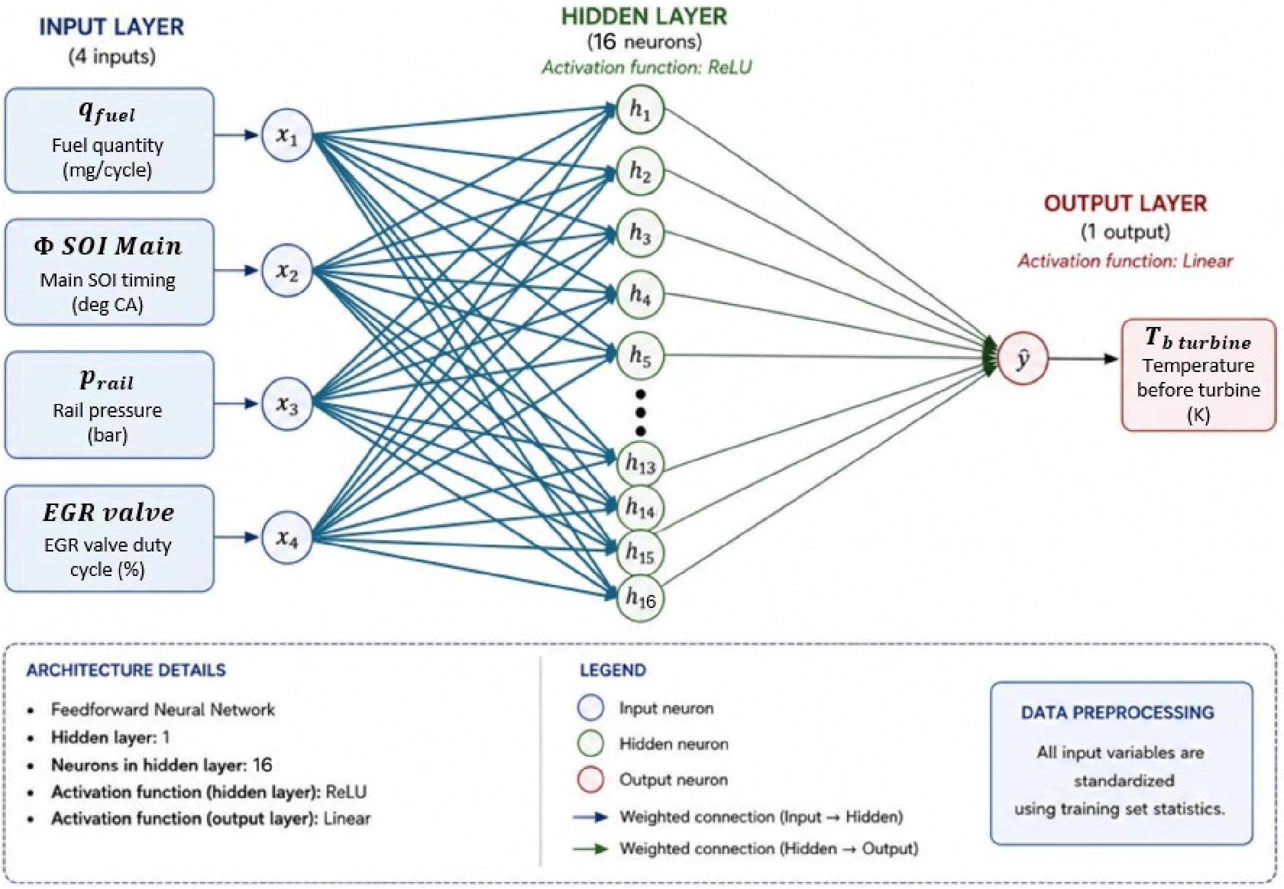

Figure 44: Neural network structure for temperature before turbine estimation

The RMSE metric was used to assess the model prediction ability.

$RMSE_{TRAIN}$	$RMSE_{TEST}$
10.48 K	12.27 K

Table 48: Training and testing RMSE of the neural network for temperature before turbine estimation

The nonlinearity introduced by the neural network further improved prediction accuracy, and the small gap between training and test errors confirmed good generalization.

7.7.4 Model Comparison

Despite its simplicity, the reduced linear model achieved good accuracy on the test set. The neural network further improved the results, with only a slight improvement in training error but a 30% reduction in test RMSE compared to the linear model

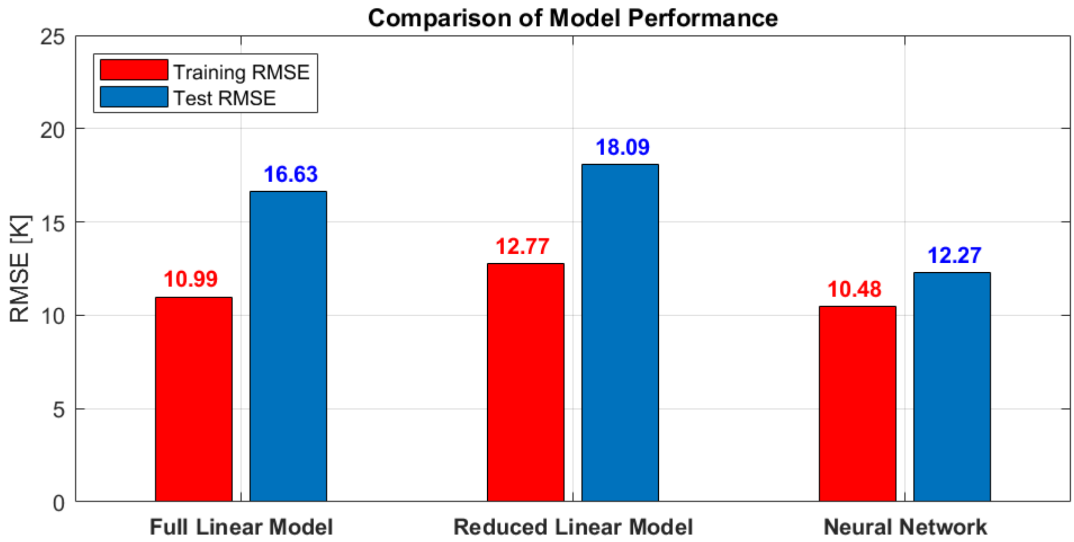

Figure 45: Performance comparison between models for temperature before turbine estimation

7.8 Exhaust Temperature After Turbine

7.8.1 Full Linear Model

```
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9

Estimated Coefficients:
                   Estimate      SE        tStat        pValue
                   ________    ______    ________    ___________

    (Intercept)     354.15     1.2026      294.48    2.2548e-137
    x1             -40.048     12.859     -3.1144      0.0024647
    x2              20.635     5.2323      3.9437      0.0001575
    x3              72.849     2.3658      30.792     6.5757e-50
    x4              3.4186     3.9419     0.86724        0.38809
    x5              11.037     9.3777      1.1769         0.2423
    x6               41.28     3.9317      10.499     2.3238e-17
    x7             0.13502     5.2432    0.025752        0.97951
    x8             -50.049     5.1143     -9.7861     7.1468e-16
    x9             0.51985     3.0237     0.17192        0.86388


Number of observations: 101, Error degrees of freedom: 91
Root Mean Squared Error: 12.1
R-squared: 0.99,  Adjusted R-Squared: 0.989
F-statistic vs. constant model: 977, p-value = 1.84e-86
```

Table 49: Regression coefficients of the full linear model for temperature after turbine estimation

From the table it is possible to assess that the temperature after the turbine is influenced mostly by the same variables considered for the temperature before the turbine.

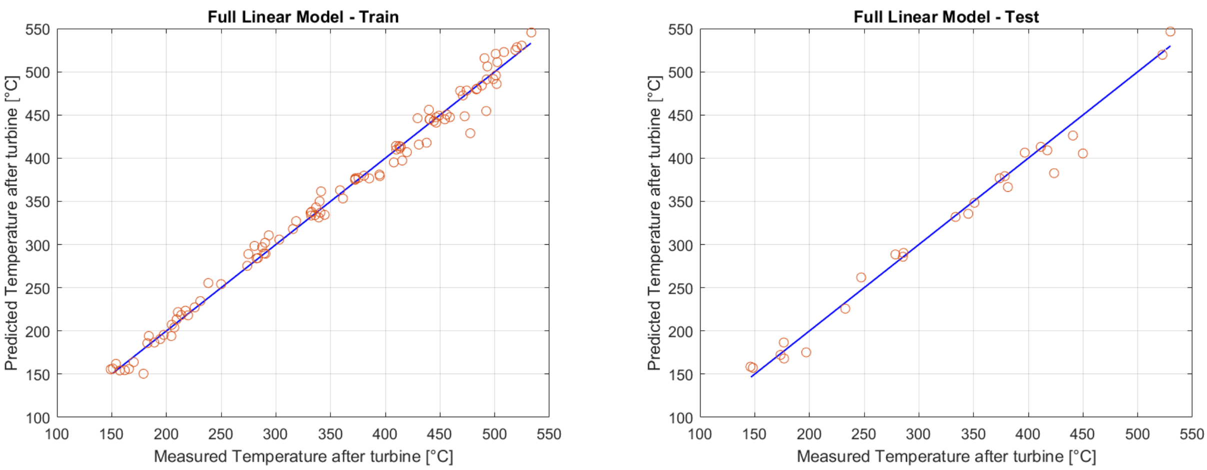

Figure 46: Training and testing plot of the full linear model for temperature after turbine estimation

$RMSE_{TRAIN}$	$RMSE_{TEST}$	R^2
11.47 $°C$	15.39 $°C$	0.990

Table 50: Training and testing performance of the full linear model for temperature after turbine estimation

These results show the excellent predictive capability of the model.

The model achieves strong generalization performance on unseen test data with small evidence of overfitting. Nevertheless, the correlation analysis highlights multicollinearity among the inputs that must be mitigated.

7.8.2 Reduced Linear Model

The physically significant variables for predicting this output were the same as those selected for the temperature before the turbine. This choice was made because the turbine does not introduce independent thermal dynamics, it mainly extracts energy from the exhaust flow.

1. **Injected Fuel Quantity**
2. **Rail Pressure**
3. **Main Injection Timing**
4. **EGR Valve**

Also, the considerations about the VIF coefficients done for the previous analysis are valid in this case. To be clear, the selection of these variables eliminates the multicollinearity issue.

The reduced regression model is reported below.

```
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4

Estimated Coefficients:
                   Estimate      SE        tStat        pValue
                   ________    ______    _______    __________

    (Intercept)     354.15     1.4934     237.15    1.0522e-134
    x1              77.476      2.756     28.112     3.9808e-48
    x2             -20.458     2.6735    -7.6521     1.5362e-11
    x3               24.66     3.1384     7.8575     5.6819e-12
    x4             -26.658     1.9362    -13.768     1.8686e-24


Number of observations: 101, Error degrees of freedom: 96
Root Mean Squared Error: 15
R-squared: 0.983,  Adjusted R-Squared: 0.983
F-statistic vs. constant model: 1.42e+03, p-value = 2.1e-84
```

Table 51: Regression coefficients of the reduced linear model for temperature after turbine estimation

The fitting of the train and test data are displayed in the images below.

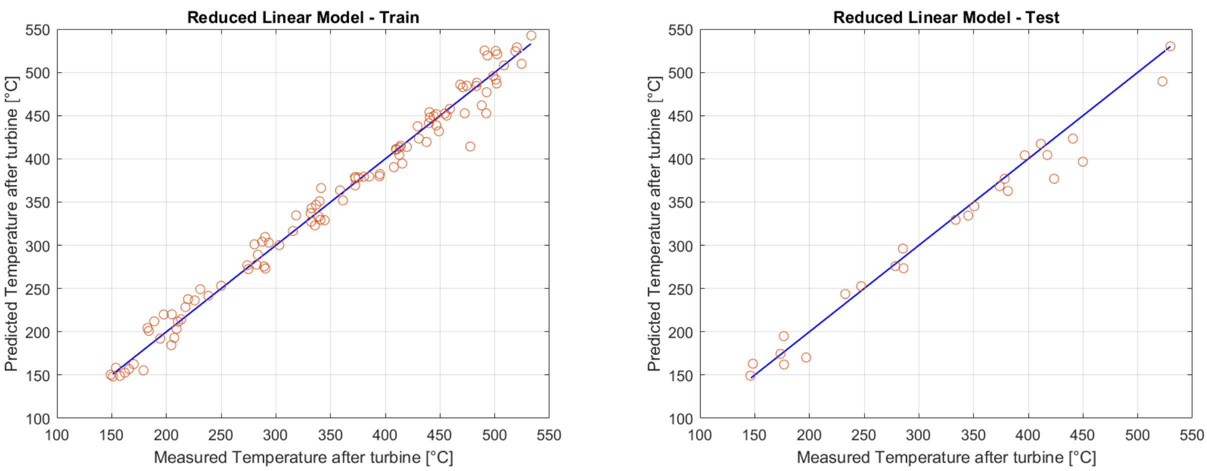

Figure 47: Training and testing plots of the reduced linear model for temperature after turbine estimation

$RMSE_{TRAIN}$	$RMSE_{TEST}$	R^2

14.63 $°C$	19.11 $°C$	0.983

Table 52: Training and testing performance of the reduced linear model for temperature after turbine estimation

The residuals didn't appear completely random but instead showed an oscillating pattern with alternating negative and positive values. This behaviour suggested the presence of a non-linear component that was not captured by the linear model.

In addition, some outliers are present, as certain points around 400°C - 450°C exhibited significantly larger residuals compared to the rest of the dataset.

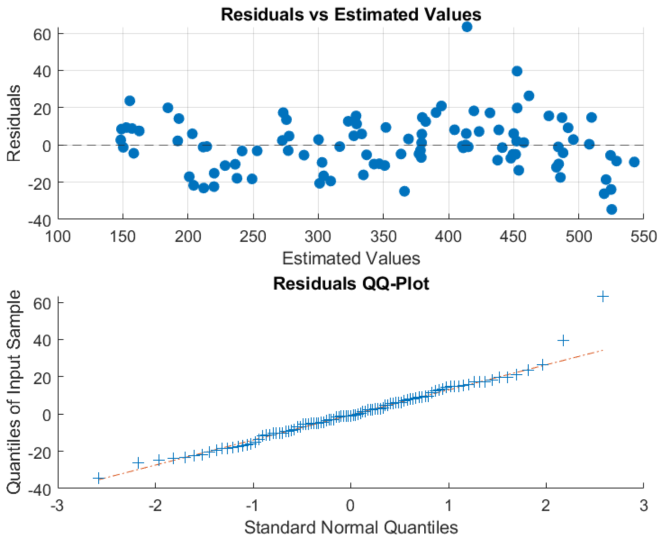

Figure 48: Residual plots of the reduced linear model for temperature after turbine estimation

7.8.3 Quadratic Model

A model was built using the same variables as the reduced linear model to assess if the quadratic contribution could model better the input-output relation. The objective was to enhance the model's descriptive capability by introducing nonlinear terms, while maintaining an acceptable level of complexity.

```
Linear regression model:
    y ~ 1 + x3 + x1*x4 + x2*x4 + x1^2 + x2^2

Estimated Coefficients:
                   Estimate      SE        tStat        pValue
                   ________    ______     _______    ___________

    (Intercept)     383.99     2.9305      131.03    2.2278e-106
    x1               109.1     3.5491      30.739      3.541e-50
    x2               5.761     2.7062      2.1289       0.035938
    x3             -11.583     3.7379     -3.0988      0.0025777
    x4             -32.312      2.592     -12.466     1.7982e-21
    x1:x4          -10.613     1.3898     -7.6358     2.0457e-11
    x2:x4          -9.0911     1.2245     -7.4241     5.5768e-11
    x1^2           -27.613     1.9971     -13.827     3.4771e-24
    x2^2           -9.6678     1.7038     -5.6741     1.6034e-07


Number of observations: 101, Error degrees of freedom: 92
Root Mean Squared Error: 7.78
R-squared: 0.996,  Adjusted R-Squared: 0.995
F-statistic vs. constant model: 2.67e+03, p-value = 2.19e-105
```

Table 53: Regression coefficients of the quadratic model for temperature after turbine estimation

The final form of the quadratic model obtained through the stepwise selection procedure included all the linear contributions, some of the interactions and the quadratic terms associated with fuel quantity and the main injection timing.

The fitting on the train and test set is illustrated in these two plots.

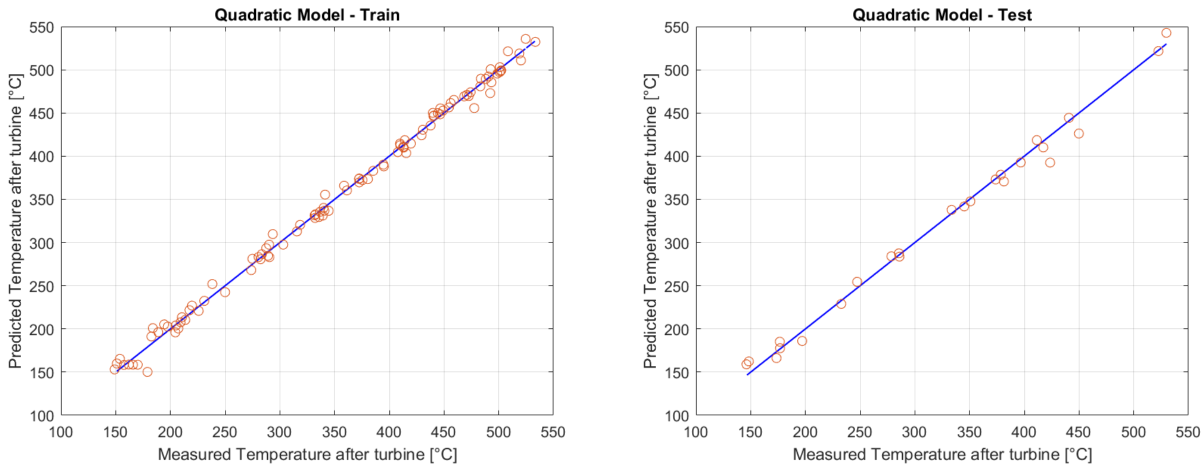

Figure 49: Training and testing plots of the quadratic model for temperature after turbine estimation

$RMSE_{TRAIN}$	$RMSE_{TEST}$	R^2
$7.42\ °C$	$10.39\ °C$	0.996

Table 54: Training and testing performance of the quadratic model for temperature after turbine estimation

Overall, the inclusion of quadratic and mixed terms allowed a substantially improvement in the system representation compared to the linear model.

7.8.4 Neural Network

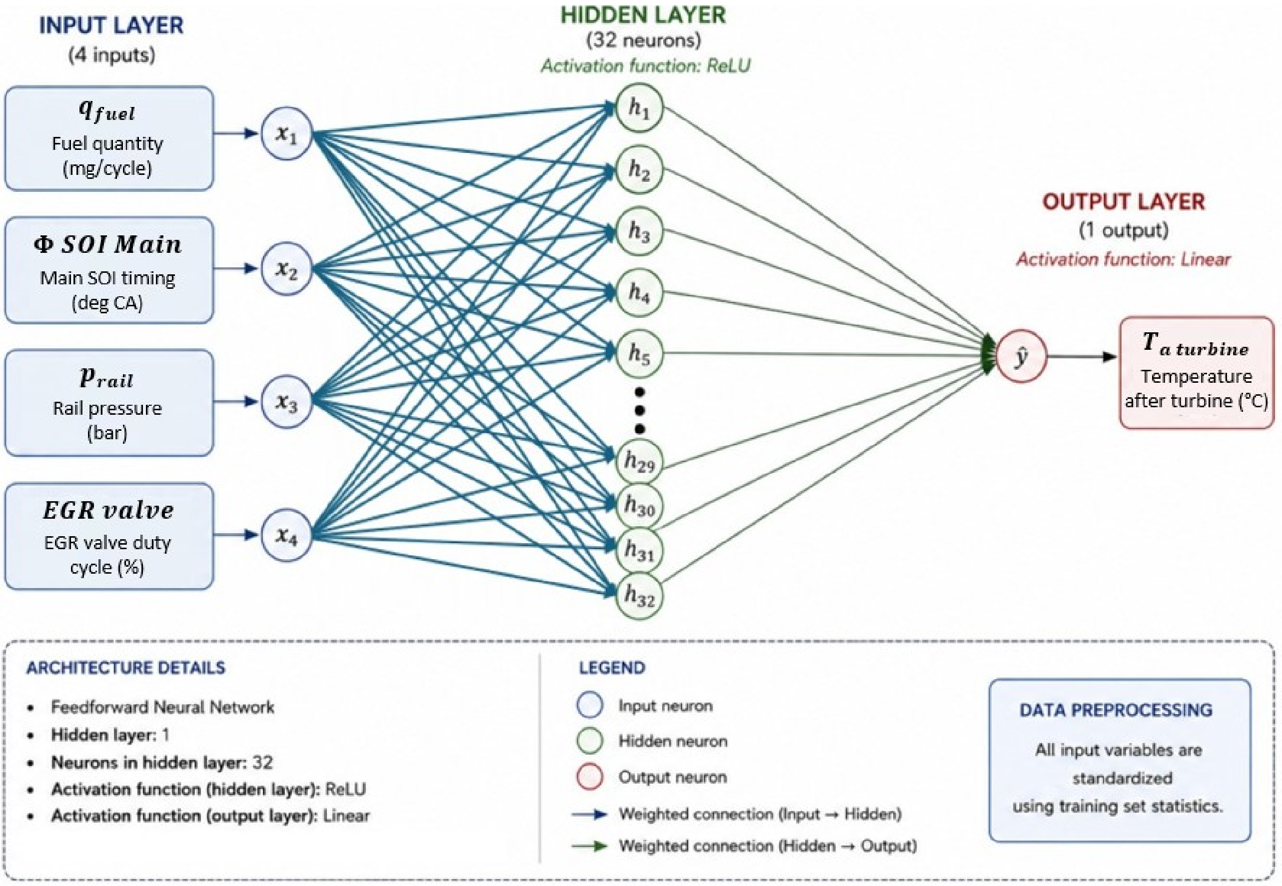

Figure 50: Neural network structure for temperature after turbine estimation

The model performance was evaluated using the Root Mean Squared Error (RMSE):

$RMSE_{TRAIN}$	$RMSE_{TEST}$
10.91 °C	13.67 °C

Table 55: Training and testing RMSE of the neural network for temperature after turbine estimation

The system relationship may be sufficiently simple to be well approximated by a quadratic model. In this case, the use of a neural network did not provide any benefit.

7.8.5 Model Comparison

The dominant nonlinearities of the process were effectively captured through quadratic and interaction terms, without requiring the additional complexity of a neural network. The linear model instead performed worse both compared the quadratic model and the neural network.

A more sophisticated neural network architecture could reduce the prediction error on train set, but it would increase the risk of overfitting.

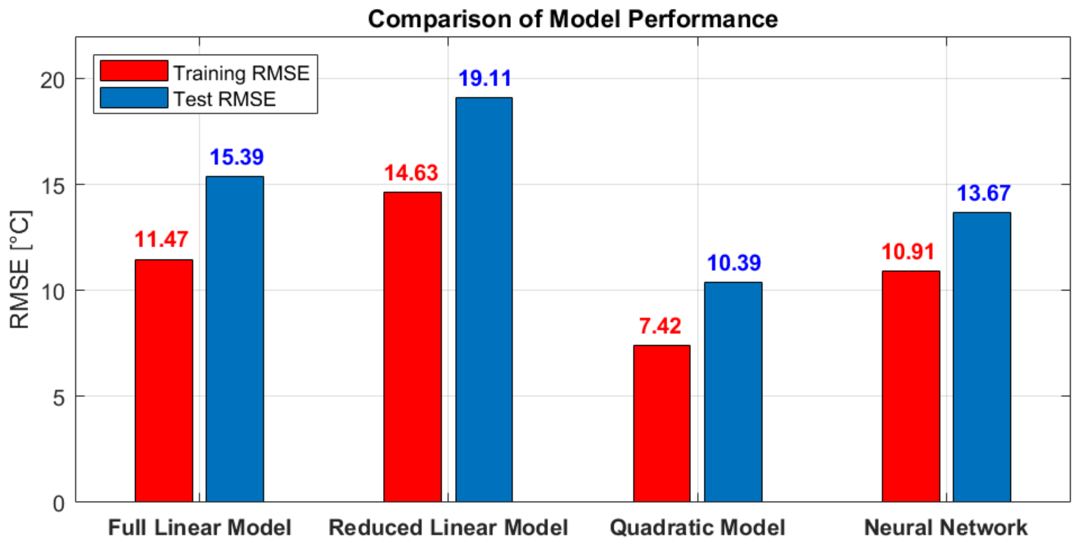

Figure 51: Performance comparison between models for temperature after turbine estimation

7.9 Air Flow Rate

7.9.1 Full Linear Model

For this specific output, it was decided to introduce the intake manifold pressure as input variable to better estimate the air flow rate.

```
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10

Estimated Coefficients:
                   Estimate      SE       tStat       pValue
                   ________    _______    _______    __________

    (Intercept)     251.05     0.82383     304.73    1.9658e-137
    x1              40.179      9.2416     4.3476     3.6143e-05
    x2              18.761      3.7668     4.9806     3.0407e-06
    x3              26.866      12.081     2.2239       0.028657
    x4              -25.46      3.6747    -6.9286     6.1652e-10
    x5              10.773      6.4355     1.6741        0.09759
    x6              22.765      3.4969       6.51     4.1713e-09
    x7             -28.478      3.6133    -7.8816     7.1468e-12
    x8              -23.17      5.2032    -4.4531     2.4235e-05
    x9             -7.9196      2.0781     -3.811     0.00025287
    x10             35.017      13.307     2.6314       0.010003


Number of observations: 101, Error degrees of freedom: 90
Root Mean Squared Error: 8.28
R-squared: 0.997,  Adjusted R-Squared: 0.997
F-statistic vs. constant model: 3.05e+03, p-value = 2.48e-109
```

Table 56: Regression coefficients of the full linear model for air flow rate estimation

The linear regression analysis indicated that engine speed and intake manifold absolute pressure are the most influential predictors for air mass flow. This is consistent with engine physics.

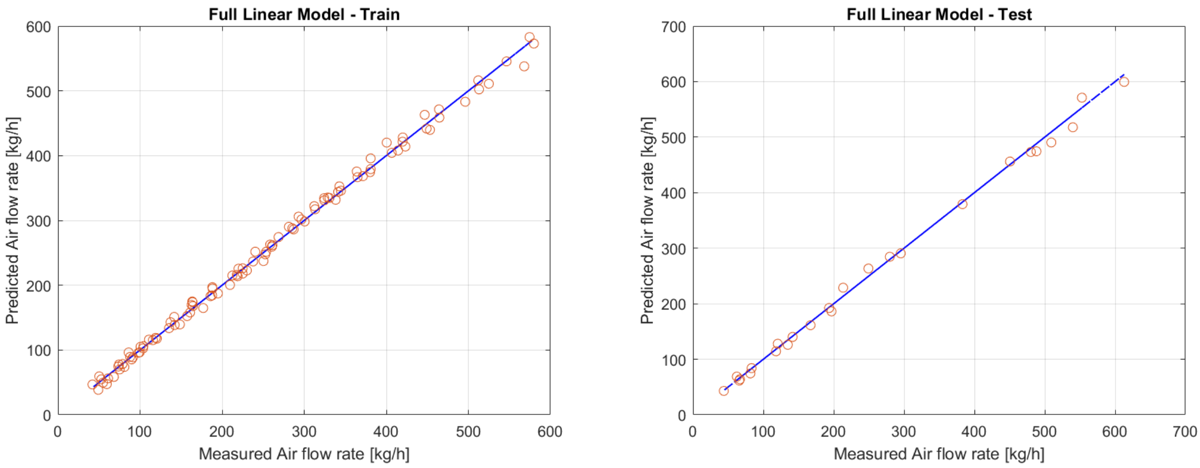

Figure 52: Training and testing plots of the full linear model for air flow rate estimation

$RMSE_{TRAIN}$	$RMSE_{TEST}$	R^2
$7.82 \ kg/h$	$10.09 \ kg/h$	0.997

Table 57: Training and testing performance of the full linear model for air flow rate estimation

These results showed the strong predictive capability of the model. Nevertheless, the correlation analysis highlighted multicollinearity among the inputs that must be reduced.

7.9.2 Reduced Linear Model

The physically significant variables for predicting the air flow rate were:

1. **Engine Speed:**
   The engine speed influences the number of intake cycles per unit time, changing the amount of air drawn into the cylinders.
2. **VGT position**
   It regulates the turbocharger boost pressure, thereby strongly affecting the intake air density and, ultimately, the intake air mass flow rate.
3. **EGR Valve**
   Exhaust Gas Recirculation reduces fresh air fraction in the intake charge.
4. **Intake Manifold Pressure**
   Higher values of intake pressure increase intake manifold air density and the mass.

The selection of these variables mitigated the multicollinearity issue, even if a small information redundancy was present because some VIF values were over 5. The results of the regression procedure are presented below.

```
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4

Estimated Coefficients:
                   Estimate      SE        tStat         pValue
                   ________    ______    ________    ___________

    (Intercept)     251.05     1.6647       150.8    6.9977e-116
    x1              40.981     4.1982      9.7614     4.8425e-16
    x2             -50.221     4.3655     -11.504     9.1627e-20
    x3             -1.4854     2.1023    -0.70657        0.48154
    x4              78.724      2.039      38.609     2.8092e-60


Number of observations: 101, Error degrees of freedom: 96
Root Mean Squared Error: 16.7
R-squared: 0.987,  Adjusted R-Squared: 0.987
F-statistic vs. constant model: 1.85e+03, p-value = 6.87e-90
```

Table 58: Regression coefficients of the reduced linear model for air flow rate estimation

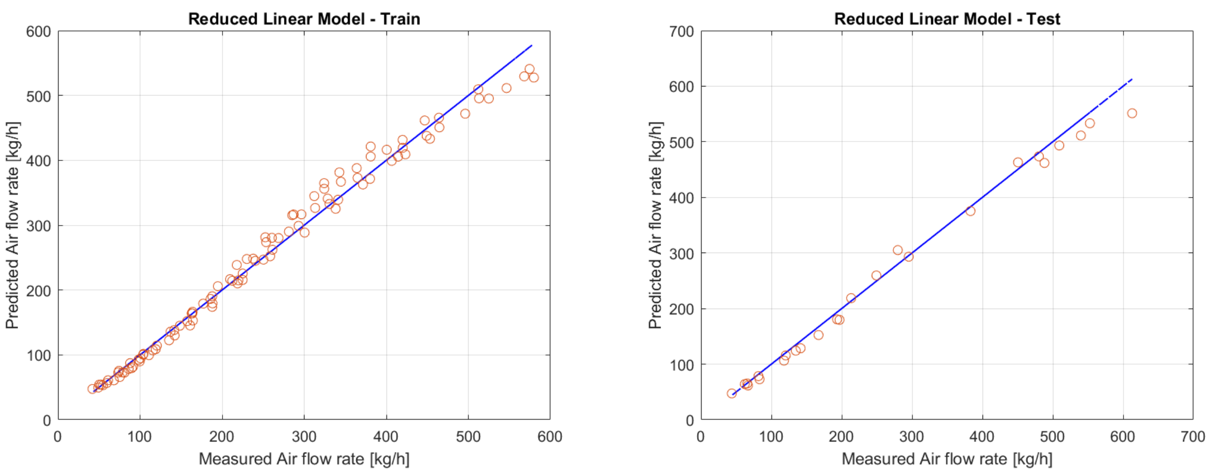

Figure 53: Training and testing plots of the reduced model for air flow rate estimation

$RMSE_{TRAIN}$	$RMSE_{TEST}$	R^2
$16.31\ kg/h$	$18.18\ kg/h$	0.987

Table 59: Training and testing performance of the reduced model for air flow rate estimation

The analysis of the correlation matrix made possible to identify the pairs of variables responsible for VIF values exceeding 5. In particular, x1 and x2 which are the engine speed and the VGT had a high absolute correlation.

However, the condition number was equal to 5.69 confirming the mitigation of the information redundancy in the input variables.

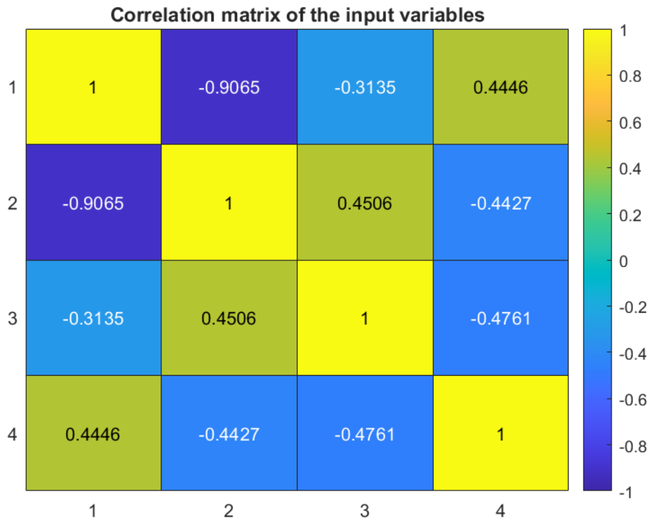

Figure 54: Correlation matrix of the selected input variables for air flow rate estimation

The residual plot showed a clear fan-shaped pattern: for low predicted values, the residuals were relatively small and concentrated, then spread progressively as prediction values increase. This indicated that the model suffered from heteroscedasticity.

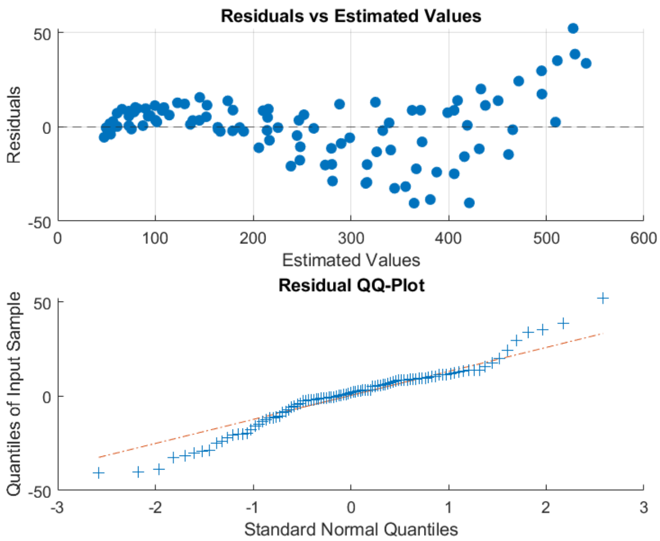

Figure 55: Residual plots of the reduced linear model for air flow rate estimation

7.9.3 Quadratic Model

With the aim to improve prediction accuracy, particularly at high values, and reduce heteroscedasticity, a quadratic model was adopted.

```
Linear regression model:
    y ~ 1 + x1*x4 + x2*x3 + x2*x4 + x3*x4 + x4^2

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   ________    _______     _______    __________

    (Intercept)     242.07     0.93023      260.23     1.716e-132
    x1              71.433      1.9294      37.022     1.1768e-56
    x2             -17.404      2.1203     -8.2082     1.4172e-12
    x3             -14.951      1.2903     -11.587     1.3199e-19
    x4              74.957     0.64241      116.68    6.7682e-101
    x1:x4           17.118      1.4754      11.602     1.2254e-19
    x2:x3           5.3561     0.77841      6.8808     7.3789e-10
    x2:x4          -7.0155      1.5155     -4.6292     1.2148e-05
    x3:x4          -3.2263       1.006     -3.2072      0.0018505
    x4^2           -5.6005     0.58895     -9.5093     2.7121e-15


Number of observations: 101, Error degrees of freedom: 91
Root Mean Squared Error: 2.57
R-squared: 1,  Adjusted R-Squared: 1
F-statistic vs. constant model: 3.52e+04, p-value = 4.34e-157
```

Table 60: Regression coefficients of the quadratic model for air flow rate estimation

The quadratic model included all the linear terms of the four explanatory variables, some of the interaction terms and only one quadratic term. The remaining terms were considered statistically non relevant.

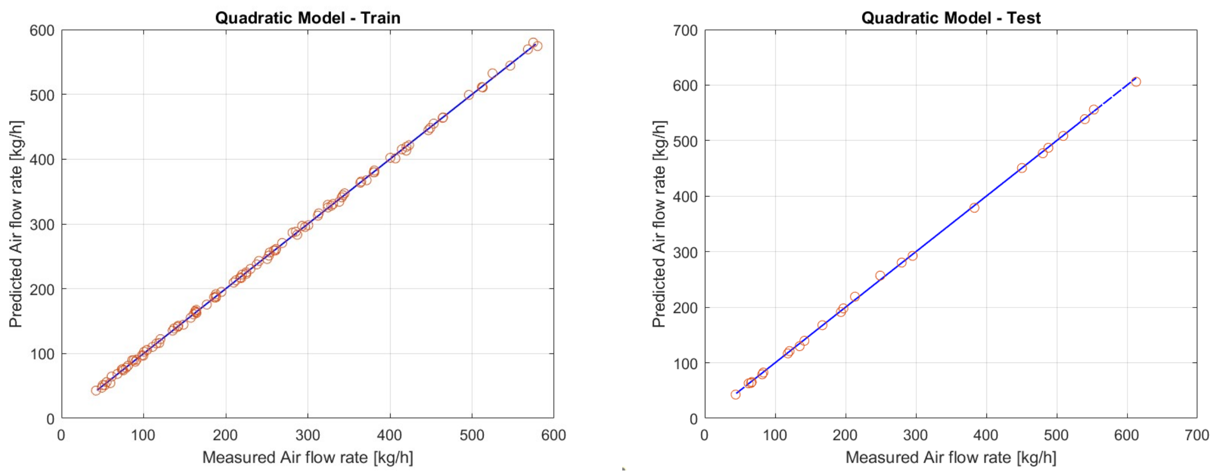

Figure 56: Training and testing plots of the quadratic model for air flow rate estimation

$RMSE_{TRAIN}$	$RMSE_{TEST}$	R^2
2.44 kg/h	3.07 kg/h	1

Table 61: Training and testing performance of the quadratic model for air flow rate estimation

The inclusion of quadratic and interaction terms yielded an R^2 of 1, indicating that the model explained all the variability of the output.

7.9.4 Neural Network

Despite the quadratic model perfectly capturing the output behaviour, a neural network was nonetheless developed.

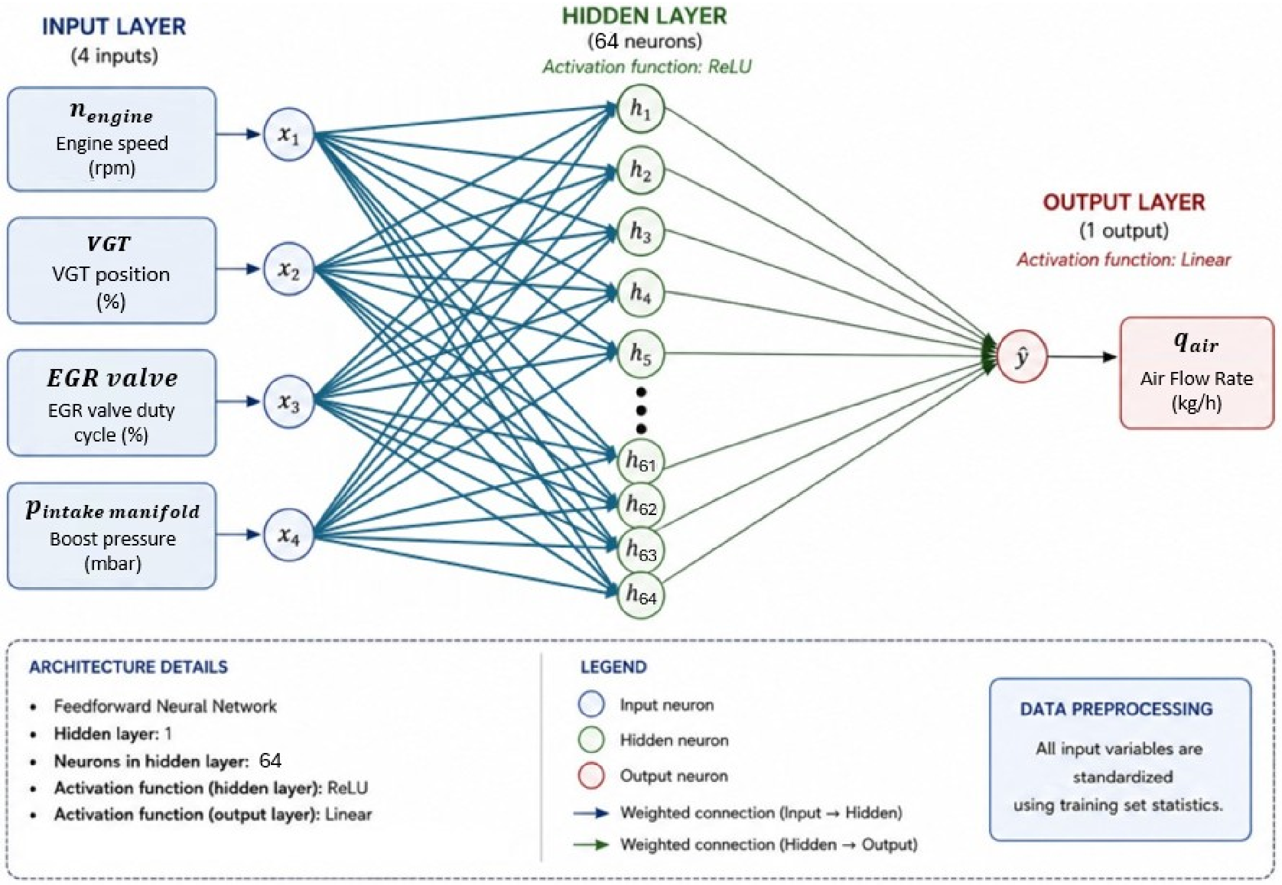

Figure 57: Neural network structure for air flow rate estimation

$RMSE_{TRAIN}$	$RMSE_{TEST}$
$1.18\ kg/h$	$2.28\ kg/h$

Table 62: Training and testing RMSE of the neural network for air flow rate estimation

The test RMSE of the neural network was only marginally lower than that of the quadratic model. The quadratic model, however, retained the physical interpretability of its coefficients and required far less complexity. Actually, the neural network result was obtained with 64 neurons in the hidden layer.

7.9.5 Model Comparison

The results comparison showed a progressive improvement from the linear models to the nonlinear approaches. From the reduced linear model to the quadratic one, the test RMSE dropped significantly. The neural network result showed that further improvement from the quadratic model was difficult, and the output was almost already explained.

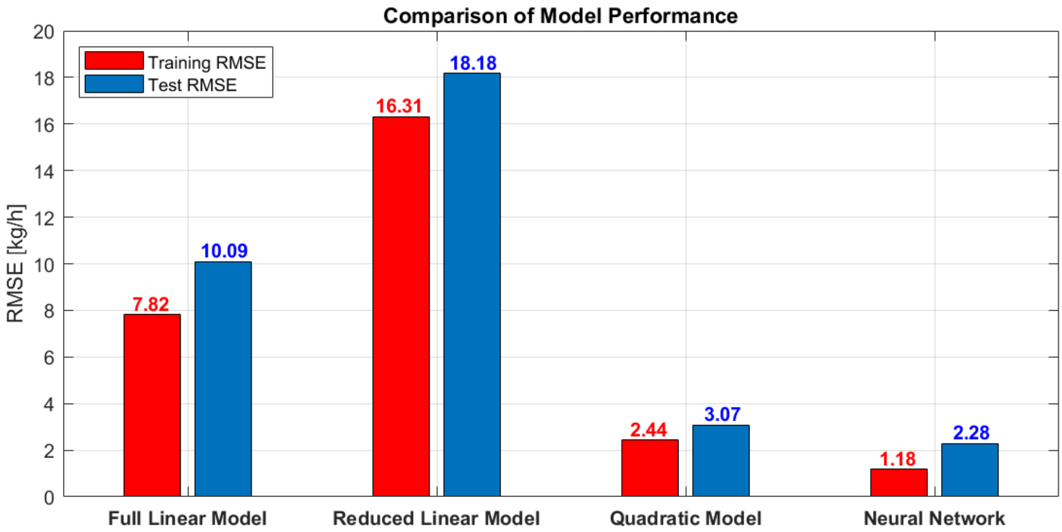

Figure 58: Performance comparison between models for air flow rate estimation

7.10 Exhaust Pressure

7.10.1 Full Linear Model

Linear regression coefficients obtained considering all the input variables are reported in the following table.

```
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9

Estimated Coefficients:
                   Estimate      SE        tStat        pValue
                   ________    ______     _______    __________

    (Intercept)      699.5     4.0252      173.78    1.4408e-116
    x1              85.134     43.038      1.9781       0.050939
    x2              34.672     17.512      1.9798       0.050744
    x3              379.29     7.9183        47.9     2.4366e-66
    x4             -116.11     13.193     -8.8007      8.248e-14
    x5              87.244     31.387      2.7796      0.0066118
    x6              103.61     13.159      7.8737      6.987e-12
    x7             -44.611     17.549     -2.5422       0.012708
    x8              40.193     17.117      2.3481       0.021035
    x9             -53.414      10.12      -5.278     8.8085e-07


Number of observations: 101, Error degrees of freedom: 91
Root Mean Squared Error: 40.5
R-squared: 0.994,  Adjusted R-Squared: 0.994
F-statistic vs. constant model: 1.76e+03, p-value = 5.91e-98
```

Table 63: Regression coefficients of the full linear model for exhaust pressure estimation

The linear regression analysis indicated that injected fuel quantity and the rail pressure were the most influential predictors for exhaust pressure.

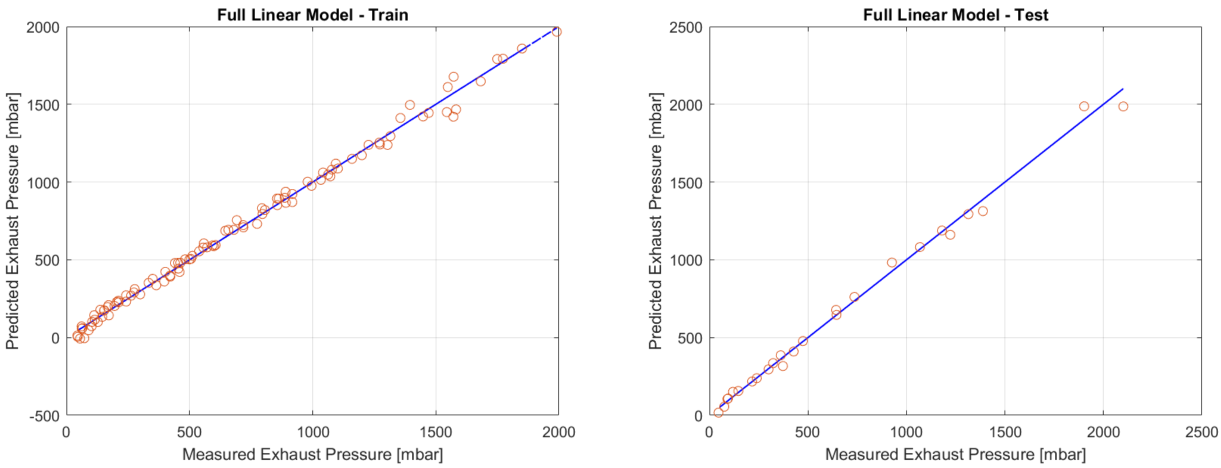

Figure 59: Training and testing plots of the full linear model for exhaust pressure estimation

The predictive performance of the model on the training and test sets is summarized below.

$RMSE_{TRAIN}$	$RMSE_{TEST}$	R^2
38.40 $mbar$	41.21 $mbar$	0.994

Table 64: Training and testing performance of the full linear model for exhaust pressure estimation

The model achieves strong generalization performance on unseen test data with no overfitting. Nevertheless, the multicollinearity problem was present in the input.

7.10.2 Reduced Linear Model

The physically significant variables for predicting the exhaust pressure were:

1. **Engine Speed:**
   The engine speed influences the number of intake cycles per unit time, which can affect the exhaust pressure.
2. **Exhaust flap**
   The more the flap is closed, the greater the restriction to the flow and therefore the higher the pressure upstream of the valve.
3. **Quantity of fuel**
   An increase in the amount of injected fuel leads to an increase in the temperature of the combustion gases, which tends to increase the exhaust pressure.
4. **VGT position**
   The VGT regulates the flow area of the exhaust gases through the turbine. By reducing the effective flow area, the resistance to gas flow increases, leading to higher backpressure upstream of the turbine and consequently an increase in exhaust pressure.

The selection of these variables mitigated the multicollinearity issue, even if a small information redundancy was still present highlighted by some VIF values over 5.

```
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4

Estimated Coefficients:
                   Estimate      SE       tStat       pValue  
                   ________    ______    _______    __________

    (Intercept)      699.5     10.484      66.72     3.091e-82
    x1              307.43     25.465     12.072    5.8152e-21
    x2              146.73     15.466      9.487    1.8824e-15
    x3               464.3     14.097     32.937    4.0904e-54
    x4             -60.508     25.529    -2.3701      0.019781


Number of observations: 101, Error degrees of freedom: 96
Root Mean Squared Error: 105
R-squared: 0.959,  Adjusted R-Squared: 0.957
F-statistic vs. constant model: 562, p-value = 1.16e-65
```

Table 65: Regression coefficients of the reduced linear model for exhaust pressure estimation

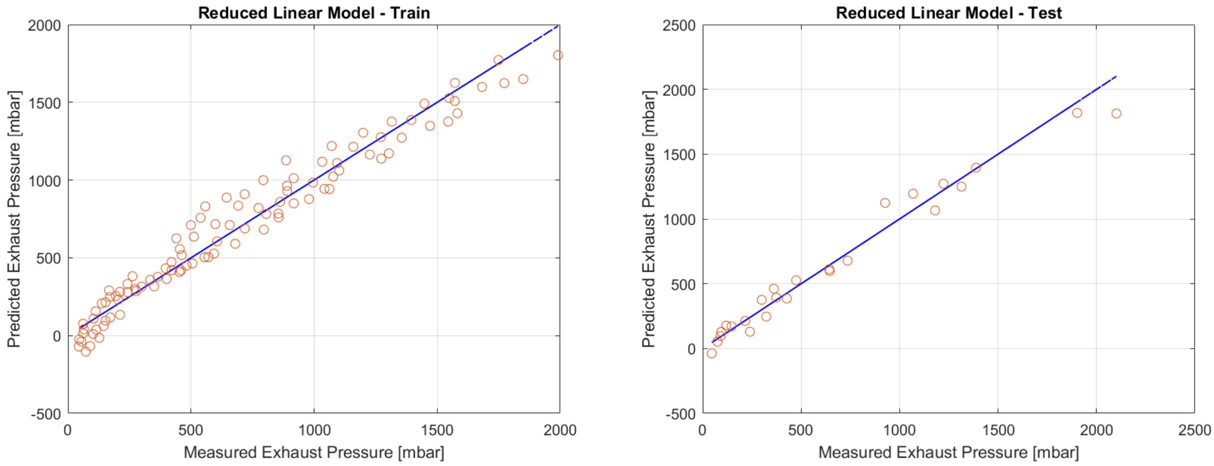

Figure 60: Training and testing plots of the reduced model for exhaust pressure estimation

$RMSE_{TRAIN}$	$RMSE_{TEST}$	R^2
$102.72\ mbar$	$94.21\ mbar$	0.959

Table 66: Training and testing performance of the reduced model for exhaust pressure estimation

The multicollinearity detected in the correlation analysis was driven by the correlation between x2, engine speed, and x4, VGT position. However, the condition number indicated that the issue was marginal and did not compromise model stability.

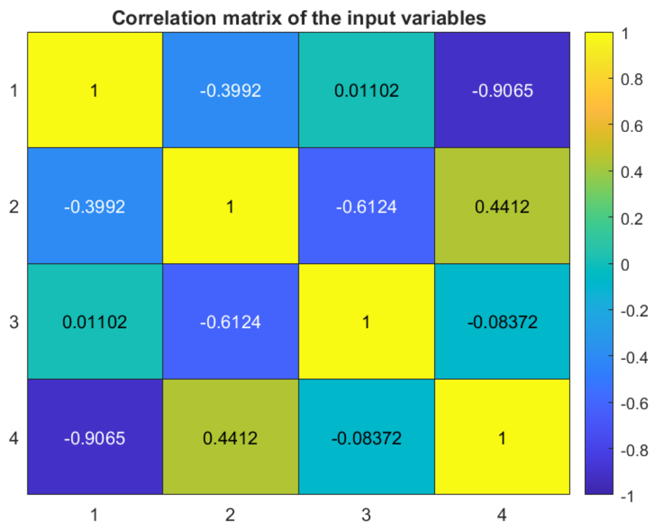

Figure 61: Correlation matrix of the selected input variables for exhaust pressure estimation

The residuals are not randomly distributed around zero, but instead showed a clear U-shaped pattern, which suggested potential non-linearity in the model.

In particular, the model underestimated at the high fitted values and overestimated in the intermediate region.

In addition, an increase in the variance of the residuals was observed as the fitted values increase, suggesting possible heteroscedasticity.

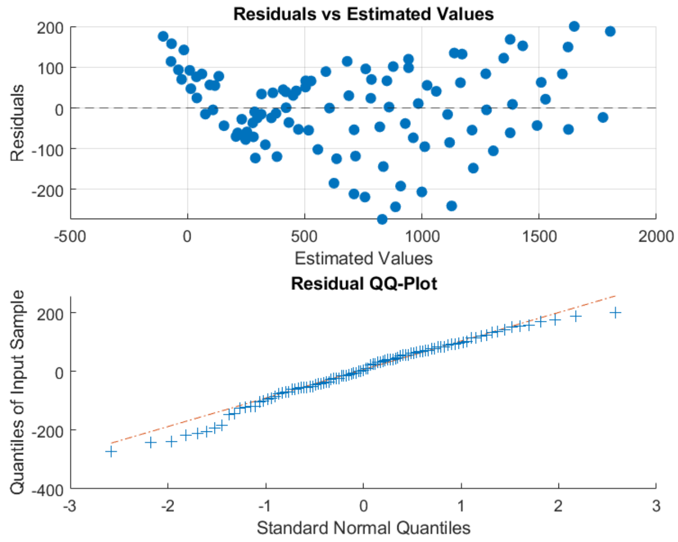

Figure 62: Residual plots of the reduced linear model for exhaust pressure estimation

7.10.3 Quadratic Model

A model was built using the same variables as the reduced linear model to assess if the quadratic contribution could capture better the input-output relation with a low complexity level.

Linear regression model:
    y ~ 1 + x2 + x1*x3 + x1*x4 + x3*x4 + x1^2 + x2^2 + x3^2 + x4^2

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	634.22	8.2148	77.204	2.4578e-83
x1	491.28	10.975	44.765	8.1253e-63
x2	-2.9357	9.7334	-0.30161	0.76365
x3	405.57	5.4722	74.115	8.8103e-82
x4	149.3	14.681	10.169	1.4349e-16
x1:x3	225.3	7.9413	28.371	2.3077e-46
x1:x4	45.454	23.723	1.916	0.058574
x3:x4	54.803	5.7604	9.5137	3.2435e-15
x1^2	61.426	11.744	5.2305	1.1116e-06
x2^2	8.2803	2.8359	2.9198	0.0044365
x3^2	17.426	3.8146	4.5681	1.5756e-05
x4^2	22.112	11.3	1.9568	0.053502

Number of observations: 101, Error degrees of freedom: 89
Root Mean Squared Error: 17.2
R-squared: 0.999, Adjusted R-Squared: 0.999
F-statistic vs. constant model: 7.96e+03, p-value = 4.11e-128

Table 67: Regression coefficients of the quadratic model for exhaust pressure estimation

For this output, the stepwise procedure retained almost all of the regressor terms, as can be seen from the number of regression coefficients.

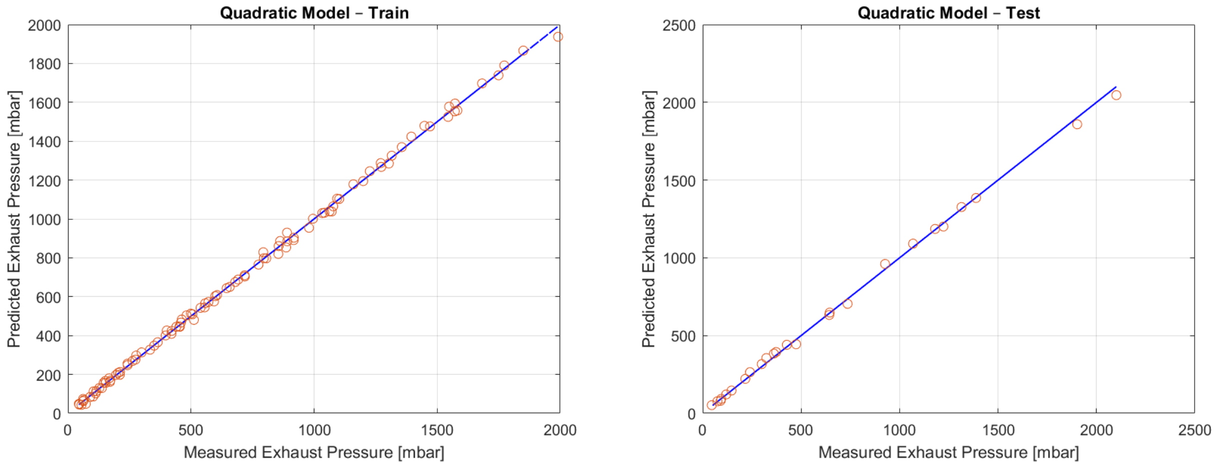

Figure 63: Training and testing plots of the quadratic model for exhaust pressure estimation

$RMSE_{TRAIN}$	$RMSE_{TEST}$	R^2
16.18 $mbar$	22.38 $mbar$	0.999

Table 68: Training and testing performance of the quadratic model for exhaust pressure estimation

Overall, the inclusion of quadratic and mixed terms led to a substantially improved representation of the system compared to the linear models, while maintaining a relatively simple structure.

7.10.4 Neural Network

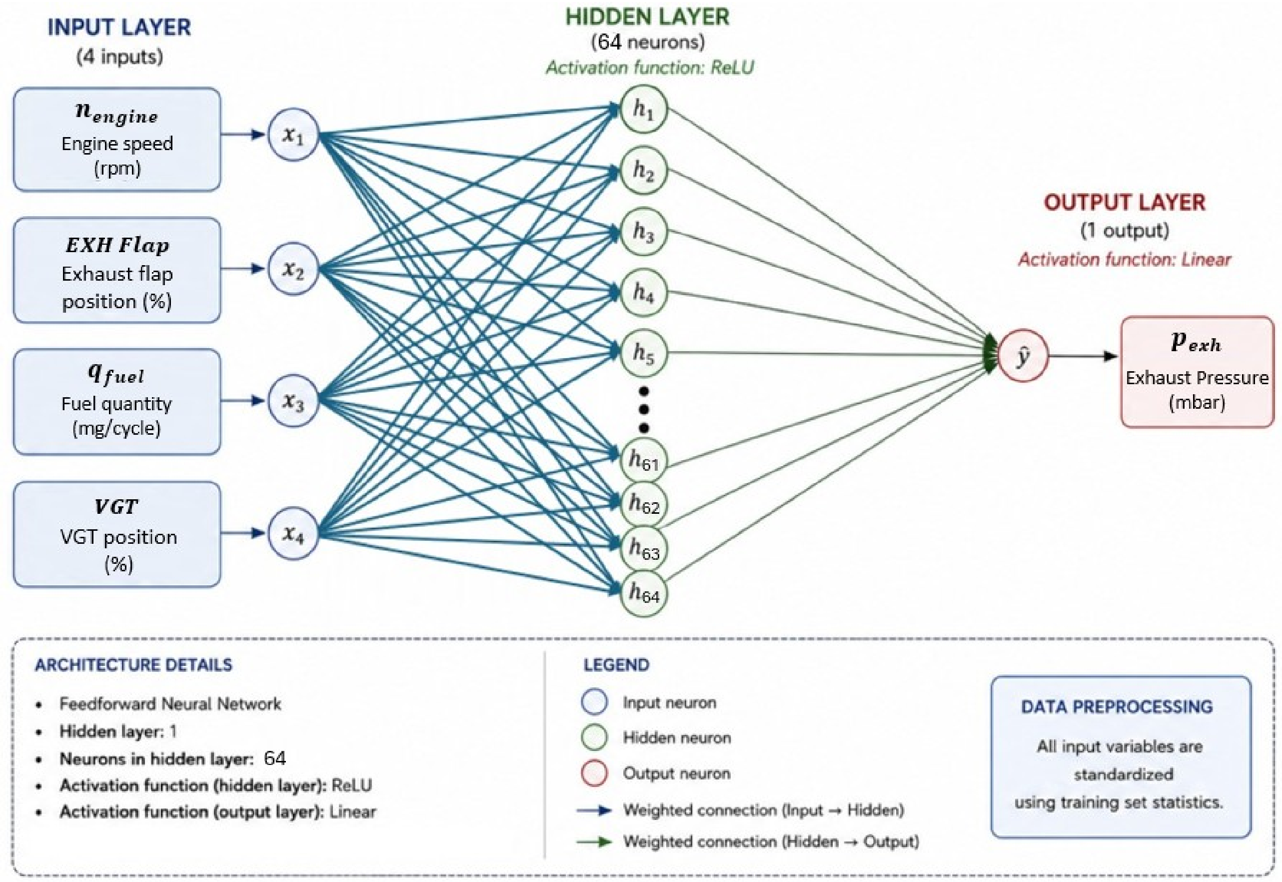

Figure 64: Neural network structure for exhaust pressure estimation

$RMSE_{TRAIN}$	$RMSE_{TEST}$
$4.73\ mbar$	$16.60\ mbar$

Table 69: Training and testing RMSE of the neural network for exhaust pressure estimation

The model fitted the training data very well, but the error increased significantly on unseen data. The gap between training and test RMSE suggested overfitting. However, the neural network achieved the lowest test RMSE, outperforming even the quadratic model.

7.10.5 Model Comparison

The results show a progressive improvement from the linear models to the more advanced nonlinear approaches. Actually, the dominant nonlinearities of the system were captured very well by the quadratic model and the neural network. The Neural Network achieves the lowest training error, but shows a more pronounced increase in test error, indicating a tendency toward overfitting.

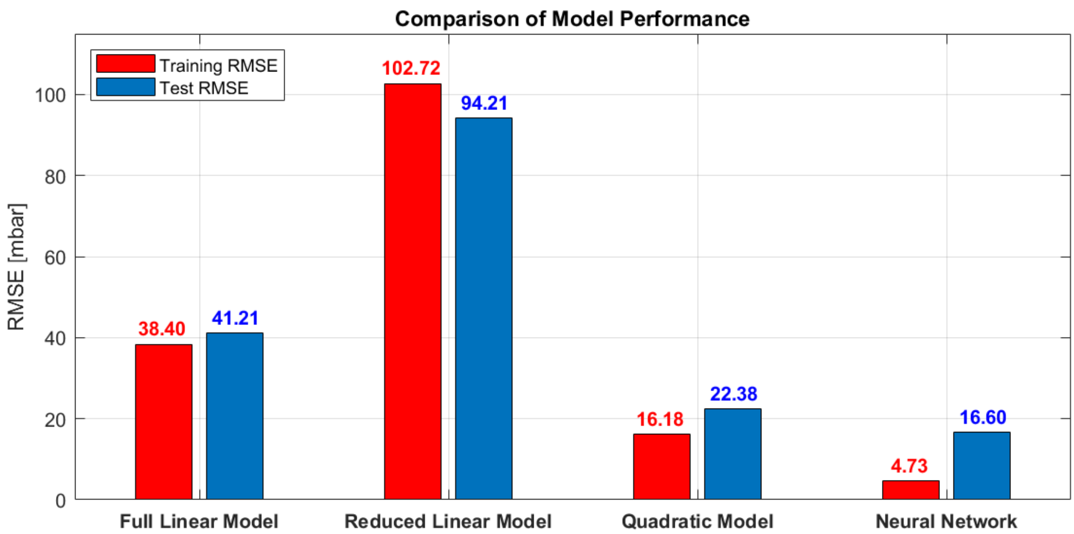

Figure 65: Performance comparison between models for exhaust pressure estimation

7.11 Multioutput Neural Network

This section presents the results of four neural networks, each trained to estimate a specific group of output variables. The networks differ in terms of architecture and learning rate, which were selected according to the characteristics of the variables being modelled. Instead, the number of epochs and the batch size were kept constant across all networks.

The table below summarizes the main characteristics of the neural networks.

GROUP	N° OUTPUTS	LEARING RATE	NEURAL NETWORK STRUCTURE
1 – PERFORMANCE ❖ Torque ❖ IMEP ❖ MFB50 ❖ Max Pressure	4	3e-4 (3×10^{-4})	512 (ReLU) → 256 (ReLU) → 128 (ReLU)
2 – NOx EMISSIONS ❖ NOx (g/kWh) ❖ NOx (ppm)	2	3e-4 (3×10^{-4})	256 (ReLU) → 128 (ReLU) → 64 (ReLU)
3 – THERMAL ❖ Temperature Before Turbine ❖ Temperature After Turbine	2	1e-3 (1×10^{-3})	512 (tanh) → 256 (tanh) → 128 (tanh) → 64 (tanh)
4 – AIR & EXHAUST SYSTEM ❖ Air Flow Rate ❖ Exhaust Pressure	2	3e-4 (3×10^{-4})	256 (tanh) → 128 (tanh)

Table 70: Overview of the multioutput neural network architectures

7.11.1 Performance Group

The figure below shows the training and validation loss curves. The two curves remained close, and both decreased in a steady way throughout training. There was no sign of overfitting, indicating good generalisation.

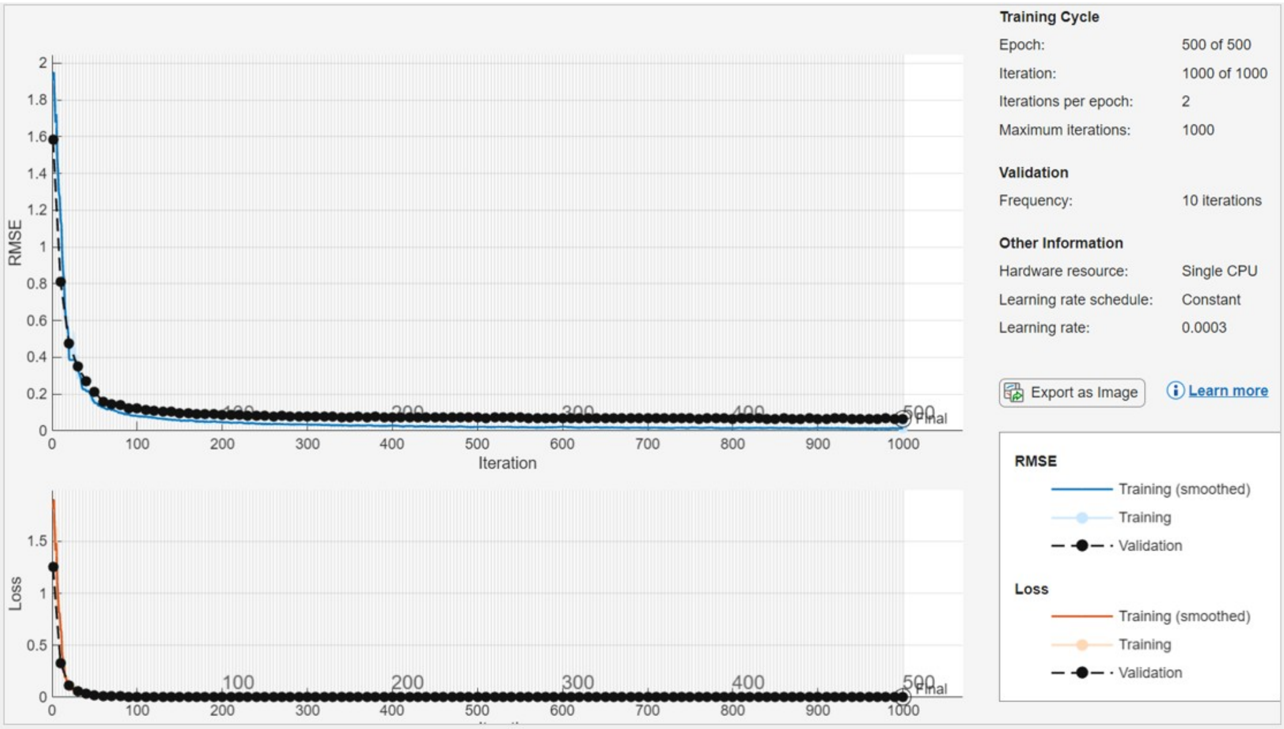

Figure 66: Training and validation curves of the performance group NN

The table shows a comparison between a neural network trained separately for each output (NN) and a multi-output approach (Group NN).

	RMSE_NN	RMSE_Group	Delta %
Torque	2.55	2.77	8.6275
IMEP	0.12	0.11	-8.3333
MFB50	0.28	0.21	-25
P Max	0.9	0.95	5.5556

Table 71: Comparison between single-output and multi-output neural networks for the performance group

The Group NN achieved performance comparable to the individual networks, showing both improvements and slight degradations depending on the analysed output. In particular, a significant improvement was observed in the estimation of MFB50, and a slight improvement was also observed for IMEP, suggesting that the group modelling of the targets was beneficial for their estimation. On the other hand, for torque and peak pressure, a slight increase in the mean squared error was observed, indicating choosing a multi output approach led to a trade-off between the different outputs.

Overall, the multi-output model was able to reach a performance comparable to that of specialized networks. The training phase was more computationally demanding due to the increased network complexity; however, during inference the multi-output model reduced computation time by estimating all four outputs simultaneously in a single pass.

7.11.2 NOx Group

The training phase is shown in the figure below. It can be observed that there is a saturation of the validation RMSE indicating that no further improvement was possible.

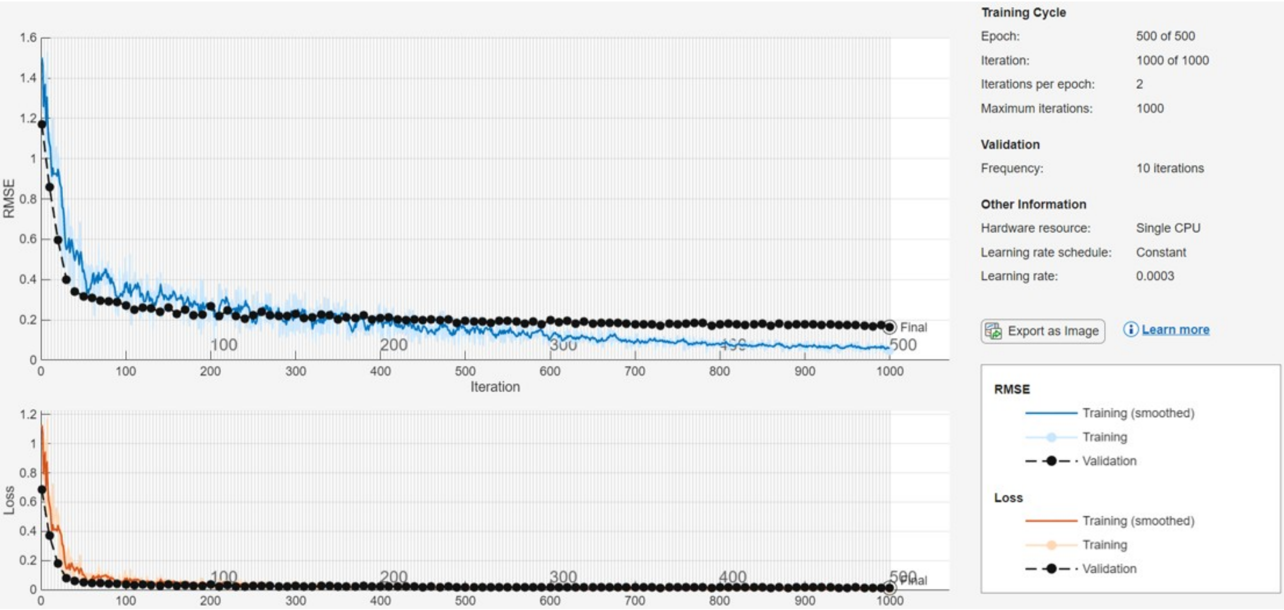

Figure 67: Training and validation curves of the NOx group NN

	RMSE_NN	RMSE_Group	Delta %
NOx g/kWh	0.32	0.39	21.875
NOx ppm	29.76	44.11	48.219

Table 72: Comparison between single-output and multi-output neural networks for the NOx group

For both representations of NOx, a significant increase in the mean squared error was observed, equal to approximately +21.9% and +48.4% compared to the single neural network, respectively.

These results indicated that, for NOx emissions, the multi-output strategy was not able to effectively minimize simultaneously both targets and instead introduced a performance trade-off.

The Group NN is more efficient during inference compared to using multiple separate neural networks, as it allows all outputs to be obtained through a single forward pass. This advantage is particularly relevant in real-time applications, where response time is a critical constraint.

7.11.3 Thermal Group

As in the other cases, training was successful, with both loss functions decreasing during the training process.

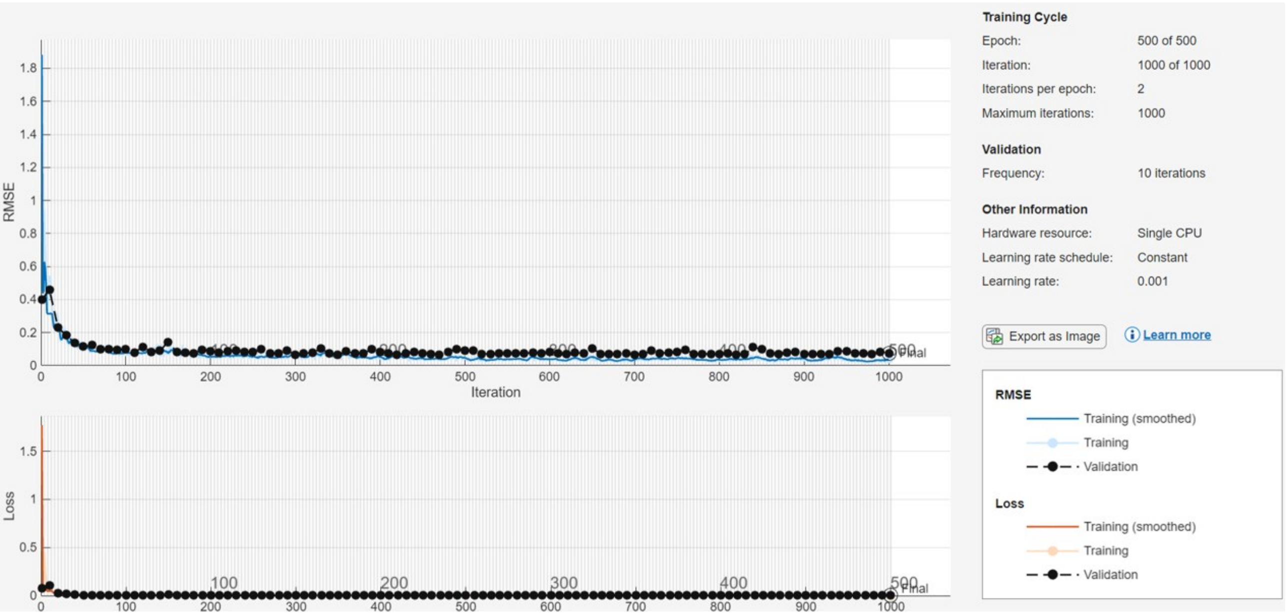

Figure 68: Training and validation curves of the thermal group NN

	RMSE_NN	RMSE_Group	Delta %
Temp. Before Turbine	12.27	3.88	-68.378
Temp. After Turbine	13.67	6.62	-51.573

Table 73: Comparison between single-output and multi-output neural networks for the thermal group

In both cases, the use of the Group NN led to a significant reduction in the mean squared error compared to the single neural networks for both output targets.

In particular, a reduction in error of more than 50% was observed for both outputs. This highlights how group modelling of the target allows an improvement in the prediction of physically correlated variables.

7.11.4 Intake and Exhaust System Group

The training phase is shown in the figure below.

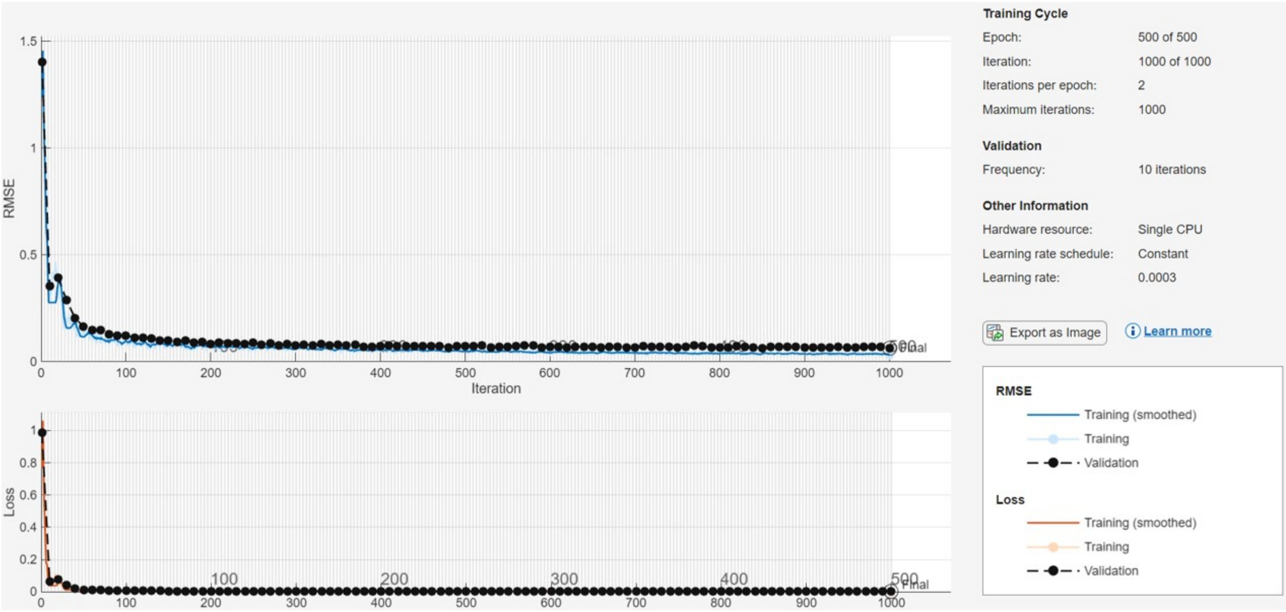

Figure 69: Training and validation curves of the intake-exhaust group NN

	RMSE_NN	RMSE_Group	Delta %
Air Flow Rate	2.28	5.14	125.44
Exhaust Pressure	16.6	15.34	-7.5904

Table 74: Comparison between single-output and multi-output neural networks for the intake-exhaust group

A significant deterioration in performance was observed for the air flow rate, with an increase in the mean squared error of more than 100%, indicating grouping strategy was not beneficial. On the other hand, a slight improvement was observed for the exhaust pressure compared to the single neural network.

Overall, these results suggest that the multi-output strategy is not always advantageous and that its effectiveness depends on the degree of correlation between the considered output variables and on the complexity of the physical phenomena being modelled.

8. Conclusions

This thesis investigated different data-driven modelling approaches for the prediction of key engine performance, thermal, intake-exhaust, and emission variables using experimental data from a diesel engine test bench.

Overall, the results of this work demonstrate that data-driven models are effective tools for approximating complex engine behaviour.

Linear models are able to capture general trends for some outputs, but they are not sufficient to fully describe the complex nonlinear behaviour of engine-related phenomena. In particular, their predictive accuracy was limited for strongly nonlinear variables, and their performance degraded when complex interactions between inputs and outputs were present.

Quadratic models allowed to capture nonlinear dependencies and interaction effects among variables, leading to better predictive performance compared to linear regression, particularly for variables having simple nonlinearities.

Despite these improvements, the quadratic model still showed limitations when dealing with highly complex phenomena, where higher-order nonlinearities and dynamic interactions could not be fully represented.

For NOx emissions specifically, a power-law model was also developed in order to better represent the pollutant formation mechanisms. However, the model showed low prediction accuracy and poor generalization capability and was therefore considered not suitable for this application.

Neural networks significantly improved predictive performance compared to classical regression techniques, demonstrating their ability to capture complex nonlinear relationships without requiring explicit functional assumptions.

Multi-output neural network architectures (Group NN) were introduced to exploit correlations between physically related outputs.

Results showed that the multi-output approach was particularly effective for strongly correlated variables, such as thermodynamic and combustion indicators, where significant error reductions were achieved.

However, for some variables, such as NOx emissions and air mass flow rate, this strategy didn't work.

The table below reports a qualitative comparison among the models developed in this work in terms of model complexity and accuracy.

Model	***Complexity***	***Accuracy***

Linear Regression	Low	Low–Medium
Quadratic Regression	Low–Medium	Medium
Power-Law Model	Medium	Low
Single-output NN	High	High
Multi-output NN	High	High

Table 75: Overall comparison of the predictive models