Pontifícia Universidade Católica
do Rio de Janeiro

André Vitor Santana Souza

NOx Emission Prediction in Heavy-Duty Vehicles Using Deep Learning with Meteorological Data Integration

Dissertação de Mestrado

Dissertation presented to the Programa de Pós–graduação em Informática, do Departamento de Informática of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Informática.

Advisor: Prof. Paulo Ivson Netto Santos

Rio de Janeiro
September 2025

Pontifícia Universidade Católica
do Rio de Janeiro

André Vitor Santana Souza

NOx Emission Prediction in Heavy-Duty Vehicles Using Deep Learning with Meteorological Data Integration

Dissertation presented to the Programa de Pós–graduação em Informática of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Informática. Approved by the Examination Committee:

Prof. Paulo Ivson Netto Santos
Advisor
Departamento de Informática – PUC-Rio

Profª. Karla Tereza Figueiredo Leite
Departamento de Ciência da Computação – UERJ

Prof. Markus Endler
Departamento de Informática – PUC-Rio

Rio de Janeiro, September 2nd, 2025

André Vitor Santana Souza

Bachelor's degree in Computer Science from the Federal University of Sergipe (UFS).

Bibliographic data

Souza, André Vitor Santana

NOx Emission Prediction in Heavy-Duty Vehicles Using Deep Learning with Meteorological Data Integration / André Vitor Santana Souza; advisor: Paulo Ivson Netto Santos. – 2025.

73 f: il. color. ; 30 cm

Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2025.

Inclui bibliografia

1. Informática – Teses. 2. Deep Learning. 3. Veículos Pesados. 4. Emissão de NOx. 5. LSTM. 6. Dados Meteorológicos. 7. Krigagem. I. Santos, Paulo Ivson Netto . II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

To my parents, siblings, and family
for their support and encouragement.

Acknowledgments

To my advisor, Professor Paulo Ivson, for his encouragement and partnership in the development of this work.

To PUC-Rio, for the financial support provided, without which this work would not have been possible.

To God and to my family, for the gift of life, for all the blessings along my journey, and for their love, encouragement, support, patience, and wisdom, not only during this work but throughout my entire life.

To my colleagues at PUC-Rio.

To the professors who participated in the Examination Committee.

To all the professors and staff of the Department, for their teachings and assistance.

To all friends and relatives who, in one way or another, encouraged or helped me.

To Zane, represented by André Nunes, for providing the data used in this research.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Abstract

Souza, André Vitor Santana; Santos, Paulo Ivson Netto (Advisor). **NOx Emission Prediction in Heavy-Duty Vehicles Using Deep Learning with Meteorological Data Integration**. Rio de Janeiro, 2025. 73p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Nitrogen oxides (NOx) emissions from heavy-duty vehicles pose significant environmental and health risks, especially in urban regions with high traffic density. This study proposes a deep learning-based framework to forecast NOx emissions using real-world data collected from vehicles operating in Brazil, enriched with interpolated meteorological variables. A comprehensive preprocessing pipeline integrates onboard sensor readings, trip metadata, and weather station data through temporal and spatial interpolation, including kriging. The analysis evaluates four regression models: Linear Regression, Random Forest, XGBoost, and Long Short-Term Memory (LSTM). Across multiple multi-horizon forecasting strategies: single-step, recursive, multi-output and direct forecasting. Results indicates that LSTM consistently achieves the lowest error rates, particularly in recursive mode, with Mean Absolute Percentage Error (MAPE) dropping below 10% in the best configuration. Adjusted average speed emerged as the most relevant predictor, while meteorological variables further improved accuracy in some scenarios. Unlike previous studies, this work pioneers the integration of continuous onboard fleet data in Brazil with spatio-temporal meteorological interpolation via kriging, this study also advances the scientific frontier by proposing a methodological framework for NOx forecasting and benchmarking under multi-horizon strategies in real-world conditions

Keywords

Deep Learning; Heavy-duty Vehicles; NOx Emissions; LSTM; Meteorological Data; Kriging.

Resumo

Souza, André Vitor Santana; Santos, Paulo Ivson Netto . **Previsão de Emissões de NOx em Veículos Pesados Usando Deep Learning com Integração de Dados Meteorológicos**. Rio de Janeiro, 2025. 73p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

As emissões de óxidos de nitrogênio (NOx) provenientes de veículos pesados representam riscos ambientais e à saúde significativos, especialmente em regiões urbanas com alta densidade de tráfego. Este estudo propõe uma estrutura baseada em aprendizado profundo para prever emissões de NOx utilizando dados reais coletados de veículos em operação no Brasil, enriquecidos com variáveis meteorológicas interpoladas. Um pipeline abrangente de pré-processamento integra leituras de sensores embarcados, metadados das viagens e dados de estações meteorológicas por meio de interpolação temporal e espacial, incluindo krigagem. A análise avalia quatro modelos de regressão: Regressão Linear, Random Forest, XGBoost e Long Short-Term Memory (LSTM), em diferentes estratégias de previsão multi-horizonte: passo único, recursiva, multi-saída e previsão direta. Os resultados mostram que o LSTM alcança consistentemente as menores taxas de erro, particularmente no modo recursivo, com valores de Erro Percentual Absoluto Médio (MAPE) inferior a 10%. A velocidade média ajustada surgiu como o preditor mais relevante, enquanto as variáveis meteorológicas melhoraram a precisão em alguns cenários. Diferentemente de estudos anteriores, este trabalho integra, de forma inédita no Brasil, dados contínuos embarcados de frotas com interpolação meteorológica espaço-temporal por krigagem (kriging). Além disso, avança a fronteira científica ao propor um framework metodológico para previsão de NOx e avaliação comparativa (benchmarking) sob estratégias multi-horizonte em condições reais de operação

Palavras-chave

Deep Learning; Veículos Pesados; Emissão de NOx; LSTM; Dados Meteorológicos; Krigagem.

Table of contents

List of figures

List of tables

List of Abreviations

AI – Artificial Intelligence

CO – Carbon Monoxide

CONAMA – Conselho Nacional do Meio Ambiente

INMET – Brazilian National Institute of Meteorology

LSTM – Long Short-Term Memory

MAE – Mean Absolute Error

MAPE – Mean Absolute Percentage Error

MMSTNet – MacroMicro Spatiotemporal Neural Network

ML – Machine Learning

NGSIM – Next Generation SIMulation

NO – Nitric oxide

NO_2 – Nitrogen dioxide

NO_x – Nitrogen Oxide

O_3 – Ozone

OOB – Out-of-bag

OMPE – Online Model Parameter Estimation

$PM_{2.5}$ – particulate matter

PROCONVE – Programa de controle de emissões veiculares

R^2 – Coefficient of Determination

RMSE – Root Mean Squared Error

RNN – Recurrent Neural Networks

SVR – Support Vector Regression

WHO – World Health Organization

XGBoost – Extreme Gradient Boosting

Without data, you are just another person with an opinion.

W. Edwards Deming, *The New Economics for Industry, Government, Education. 1993.*

1
Introduction

Nitrogen oxides (NO_x) are toxic and colorless gases that contribute to both environmental and health problems. They are one of the main contributors to the formation of smog, producing the brownish haze often seen over cities, especially during the summer. When exposed to UV rays, these gases form ozone (O_3) at the tropospheric level, a major pollutant, in contrast to the ozone layer that protects the planet from solar radiation. Additionally, in the presence of rain, these gases contribute to the formation of acid rain (BRUNEKREEF; HOLGATE, 2002; AGENCY, 2022).

These gases can also penetrate deep into the lungs. While brief exposure can irritate the lungs of healthy individuals (THALLER et al., 2008), for people with conditions such as asthma, it may increase the risk of medical emergencies and necessitate professional health care (AGENCY, 2014). For residents of areas with continuous exposure to these pollutants, it can lead to diseases such as emphysema and bronchitis, worsen asthma and heart disease, and is also associated with a higher risk of premature death (JIANG; MEI; FENG, 2015).

Given the significant impacts of NOx on human health and the environment, it becomes essential to develop strategies for monitoring and forecasting its emissions, especially in highly polluting sectors like freight transportation. Heavy-duty vehicles, which are responsible for a significant share of urban emissions (WANG; YU; LI, 2020), frequently operate on long routes and under diverse climatic and topographic conditions, which can directly influence their pollution levels (PARK et al., 2023). Anticipating emissions from these vehicles based on onboard and environmental data not only helps mitigate their negative effects but also supports more effective public policies and sustainable logistics strategies.

In this context, the application of deep learning models has stood out as a powerful approach for time series forecasting, due to their ability to capture complex and nonlinear patterns in data. Recurrent Neural Networks (RNNs) are particularly well-suited for this task, as they are capable of retaining information from previous states to improve the prediction of future events (YU et al., 2024).

Among RNNs, Long Short-Term Memory networks (LSTMs) are widely used as they overcome the vanishing gradient problem common in standard RNNs (HOCHREITER; SCHMIDHUBER, 1997). LSTMs employ a specialized

cell architecture that allows the network to learn to store information over long periods, which is extremely useful for time series in which past events can strongly influence future ones (LINDEMANN et al., 2021).

Although there are several models for forecasting emissions and environmental time series, few studies use real onboard data from heavy vehicles in continuous operation. This work proposes a NOx concentration forecasting model based on time series and machine learning, using data from vehicle sensors and meteorological stations with heterogeneous sampling rates (e.g., minute-level NOx versus hourly meteorology) are handled via temporal alignment and, when appropriate, kriging-based spatialization to vehicle positions. The study proposes a framework evaluating different approaches to multi-step forecasting, with and without explicit NOx variables and proposes a variable importance analysis.

Using real data collected from heavy-duty vehicles operating in Brazil, NOx emissions are predicted using four regression models: Linear Regression (LR), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and LSTM. The study also explores configurations such as input window size, multi-horizon forecasting strategies, and different combinations of input features.

The contribution of this work lies in proposing a methodological framework for NOx forecasting in a case study that integrates onboard data with interpolated meteorological variables, evaluating multi-horizon strategies that remain underexplored in the literature, especially in the real-world context of Brazilian fleets.

The main contributions of this work are:

- A comparative evaluation of multiple regression models under different scenarios for forecasting NOx in real vehicles;
- An analysis of the influence of meteorological and geographic variables using spatio-temporal interpolation (kriging);
- An investigation into multi-horizon forecasting using different strategies (recursive, multi-output, and direct forecasting);
- A novel integration of continuous onboard data with meteorological interpolation (kriging), establishing a methodological framework for emission forecasting in real-world fleet operations.

This document is structured as follows. Chapter 2 presents some previous work relevant to our problem. Chapter 3 details the theoretical foundations

that support the study. Chapter 4 outlines the proposed methodology. Chapter 5 reports and discusses the experimental results. Finally, Chapter 6 provides the concluding remarks and discusses potential directions for future research.

2
Related Work

Numerous studies have been conducted to understand the sources of air pollutant emissions, their temporal trends, and the resulting impacts on air quality and human health. This section presents related work on vehicle-emissions forecasting and time-series modeling using machine-learning techniques.

Regarding NOx emissions forecasting, (TOSCANO; MURENA, 2019) investigated atmospheric emissions from ships in port areas, with real-time data to quantify pollutant levels, correlating emissions (NOx, PM) with ship-traffic indicators and using a fit-robust model through the statistical software JMP. (HE et al., 2022) employed deep neural networks to estimate NOx emissions across the United States, leveraging surface ozone forecasts to infer temporal patterns, The analysis indicates that satellite-based NOx trends align with surface NO_2 in high NOx emission regions. (PARK et al., 2023) design a lightweight Random Forest (RF) focused specifically on NOx emissions prediction in diesel engines, the final RF has 11-feature (30% of the Base model), was selected with SHAP (XAI) and Pearson correlation and attains $R^2 = 0.965$ and RMSE $= 15.76\%$ on test data. (SUN et al., 2025) proposed an LSTM-based architecture, incorporating combined cost functions and a multi-head attention, a combined MAE+Huber loss, to improve predictive performance called **LSTM-CCF-MA** to predict NOx emissions and exhaust gas temperature in diesel engines, reaching a minimum of $R^2 = 0.971$ in train/test.

Similarly, (WANG; YU; LI, 2020) utilized LSTM networks to predict transient NOx emissions from diesel vehicles based on real-world operational data, reporting mean absolute deviation and RMSE reductions of about 23.6% and 8.3% versus RF and support vector regression(SVR) baselines. (QING; HONGLI, 2020) also adopted LSTM to forecast NOx emissions in thermal boilers, demonstrating the model's robustness under varying operational conditions. In an urban monitoring context, (FERNáNDEZ-AVILéS; MATTERA; SCEPI, 2024) employed meteorological data and lagged NOx values in a hybrid architecture combining autoregressive neural networks with latent factors to predict NOx levels at monitoring stations in Madrid. (KUMARI; SINGH, 2023) examined CO_2 emissions forecasting in India, showing that LSTM models outperformed traditional statistical and machine learning methods with a MAPE of 3.1%.

In the Brazilian context, (CHAVES et al., 2024) explored the forecasting of particulate matter in urban regions using RF and LSTM, reaching 99.37% with RF in daily prediction and 99.71 % with LSTM using five hours of lagged inputs in 1-hour forecast horizon.

Complementarily, (ZHANG et al., 2017) proposed a probedata approach that forecasts travel time by matching similar spatiotemporal patterns using gray-level co-occurrence matrices (GLCM) with a normalized-squared-difference metric and a negative-exponential weighted combination for multi-step prediction. While (FENG et al., 2023) introduced a macromicro spatiotemporal neural (MMSTNet), an architecture designed to capture spatial and temporal correlations for urban traffic states of road segments prediction. In a similar direction, (WANG et al., 2022) discussed fusion strategies for fixed and mobile data sources (connected vehicles) for real-time traffic state estimation, employing techniques such as Online Model Parameter Estimation (OMPE) and reporting case studies on Next Generation SIMulation (NGSIM).

Additionally, (ZENG; MIWA; MORIKAWA, 2016) developed a vehicle-dynamics-based model on probe vehicle data to predict carbon monoxide (CO) emissions and support eco-efficient routing. From the perspective of transport demand and temporal forecasting, (VISHNU et al., 2023) predicted electric vehicle freight demand using LSTM, achieving superior results compared to conventional models (MAE $\approx$ 4 kW; RMSE $\approx$ 5.9 kW). Other works, such as those by (KAYES; AL., 2019; MA et al., 2019; RAVINDRA et al., 2019), analyze various aspects of emissions and air quality, including relationships with meteorological conditions and the temporal-spatial variation of pollutants.

These studies highlight the growing application of deep learning models, particularly LSTM and its variants, in forecasting time series related to urban mobility and air quality. However, a clear research gap remains in the use of real onboard emissions data from heavy-duty fleets in continuous operation, particularly when combined with meteorological information. While most previous works focus on controlled environments, isolated engines, or data from fixed monitoring stations, this study is pioneering in leveraging continuously collected fleet data in Brazil, enriched through spatio-temporal kriging of meteorological variables, and benchmarking multiple multi-horizon forecasting strategies within a unified framework.

3
Theoretical Basis

This chapter presents the theoretical and conceptual foundations necessary to understand the phenomena, technologies, and approaches analyzed throughout this thesis. It first discusses the formation and impact of NO_x, key elements in the environmental problem under investigation. It then addresses technologies for emission measurements, particularly on board NO_x sensors, and how they are implemented in urban environments. Finally, it presents predictive modeling and machine learning techniques used to quantify and estimate pollutant emissions, providing a solid background for the subsequent experimental and analytical stages.

3.1
What is NO_x?

The term **NO_x** refers to a group of nitrogen oxides, primarily nitric oxide (NO) and nitrogen dioxide (NO_2), produced predominantly during high temperature combustion processes, especially in internal combustion engines. While NO is the principal compound emitted directly from vehicle exhaust, it quickly oxidizes in the atmosphere to form NO_2, a more reactive and harmful pollutant. Both gases act as precursors in the formation of tropospheric ozone and secondary fine particulate matter ($PM_{2.5}$), posing serious risks to human health and contributing to environmental degradation (ORGANIZATION, 2021).

The emission of atmospheric pollutants by motor vehicles remains one of the foremost environmental and public health challenges in modern urban areas. Among the most critical pollutants is NO_x, especially from diesel powered vehicles. Numerous studies, including those by the World Health Organization (WHO) (ORGANIZATION, 2021), (III; DOCKERY, 2006), and (LELIEVELD et al., 2015), have established a strong association between chronic exposure to NO_x and increased incidence of respiratory and cardiovascular diseases, as well as elevated premature mortality.

Given its significance, NO_x emission control is subject to stringent regulations in many countries. In Brazil, environmental frameworks such as Conselho Nacional do Meio Ambiente (CONAMA) resolutions and the Programa de controle de emissões veiculares (PROCONVE) define emission thresholds and encourage the adoption of cleaner vehicular technologies (Conselho Nacional do Meio Ambiente, 2018).

3.2 Measurement Technologies and the Role of NO_x Sensors

Traditional laboratory based measurement techniques, such as chassis dynamometers and controlled test chambers, while highly accurate, often fail to represent actual traffic conditions. In this context, on board NO_x sensors have emerged as a cost-effective and scalable solution for real-time, infield emissions monitoring.

Figure 3.1 illustrates a typical installation layout in vehicles equipped with a Selective Catalytic Reduction (SCR) after treatment system. In this configuration, one NO_x sensor is positioned upstream of the SCR catalyst to measure engine out emissions, while a second sensor is located downstream to monitor the NO_x concentration after the catalyst. This dual sensor arrangement enables precise calculation of SCR efficiency and supports the detection of system malfunctions.

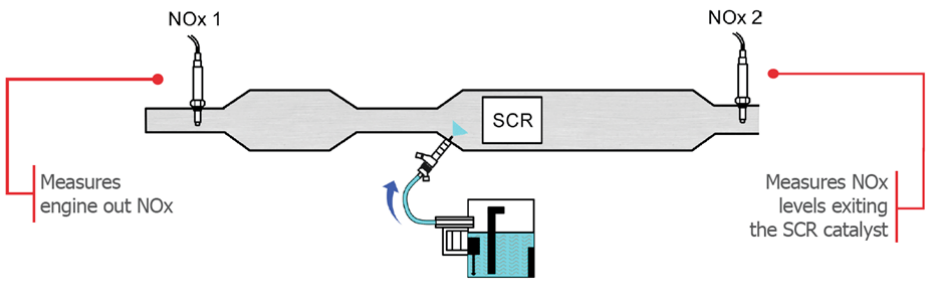

Figure 3.1: Typical arrangement of NO_x sensors in a Selective Catalytic Reduction (SCR) system (Standard Brand, 2025).

A widely used example is the *UniNOx* sensor by Continental, which employs two electrochemical cells: the **Nernst cell**, estimating oxygen concentration through a voltage differential across platinum electrodes embedded in stabilized zirconia, and the **pumping cell**, maintaining a controlled oxygen level in a sealed chamber by modulating an applied current, thereby inferring NO_x concentrations. The device operates near 800°C with 125 or 250 V supply, providing fast response and stable measurements. It covers 0–1,500 ppm with a precision of ± 10 ppm. The expected lifetime is 36,000 h and was used during the sample period for less than 1,000 h. Communication is digital via CAN (J1939), enabling seamless integration with telemetry systems and cloud-based databases (SILVA; CRUZ; NUNES, 2024). Figure 3.2 shows the UniNOx sensor.

Figure 3.2: UniNOx sensor (Continental Aftermarket, 2025).

The deployment of NO_x sensors in densely populated regions provides essential data for environmental diagnostics and the formulation of evidence based public policies. In areas like the Metropolitan Region of Rio de Janeiro, characterized by high traffic density and fleet heterogeneity, such technology enables high-resolution analysis of vehicular pollution across various segments, time intervals, routes, and meteorological conditions.

Incorporating these sensor based datasets into predictive modeling workflows facilitates scenario simulations, hotspot identification, and policy impact assessments, thereby supporting sustainable mobility strategies and safeguarding public health.

3.3 Classical Machine Learning (ML)

In this field, the algorithm makes use of features from a phenomenon of interest to learn and adapt through experience in order to improve the evaluation of a metric, rather than being explicitly told the output (PANDEY et al., 2020). These features are a useful representation of the underlying information to be extracted. For this reason, it is assumed that the training data was constructed through a feature engineering process. The learning process uses features (attributes) either pre-selected by a person or extracted through representation learning techniques. When properly learned, they can be used as inputs to a supervised model, seeking the optimal data representation (BENGIO; COURVILLE; VINCENT, 2013). A large amount of training data is used to enable the model to understand patterns. Therefore,

the choice of attributes/features is of essential importance in generating appropriate learning models (BEZERRA, 2016).

3.3.1 Linear Regression (LR)

Linear regression is a statistical technique that models the relationship between a dependent variable y and one or more independent variables $x_1, x_2, \ldots, x_p$, assuming that this relationship is approximately linear (JAMES et al., 2023). The general form of the model can be written as:

$$\hat{y} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j \tag{3-1}$$

3.3.2 Decision Tree

A decision tree is a supervised machine learning method based on the divide and conquer strategy, where the problem is recursively partitioned into smaller subproblems until reaching a terminal condition (FACELI et al., 2021). Structurally, a decision tree consists of internal nodes representing tests on attributes, branches corresponding to outcomes of these tests, and leaf nodes that assign a decision or prediction, which may be discrete (classification) or continuous (regression) (RUSSELL; NORVIG; DAVIS, 2010). The algorithm seeks to maximize the separation between classes by selecting at each split the attribute that yields the highest discriminatory power commonly measured by criteria such as information gain, gain ratio, or Gini index. Decision trees are interpretable, handle both numerical and categorical data, and can model non linear relationships without requiring prior assumptions about data distribution. However, they are prone to overfitting, especially when grown deep without pruning (JAMES et al., 2023).

3.3.3 Random Forest (RF)

Random Forest is an ensemble learning method that mitigates the overfitting tendency of decision trees, while preserving their low bias, by allowing trees to grow deep, using bootstrap samples, constructing a large number of them and aggregating their predictions (SILVA et al., 2020). Each tree is trained on a bootstrap sample of the training data, and at each split, only a random subset of predictors is considered, a technique known as the feature bagging approach. This dual randomization (sampling of instances and features) reduces correlation among trees, improving generalization performance.

Classification is achieved via majority voting among trees, while regression predictions are obtained by averaging outputs (JAMES et al., 2023). Random Forests are robust to noise, handle high-dimensional datasets effectively, and can provide internal estimates of generalization error through out-of-bag (OOB) error estimation. Furthermore, they offer measures of variable importance, which are valuable for feature selection in exploratory data analysis (FACELI et al., 2021).

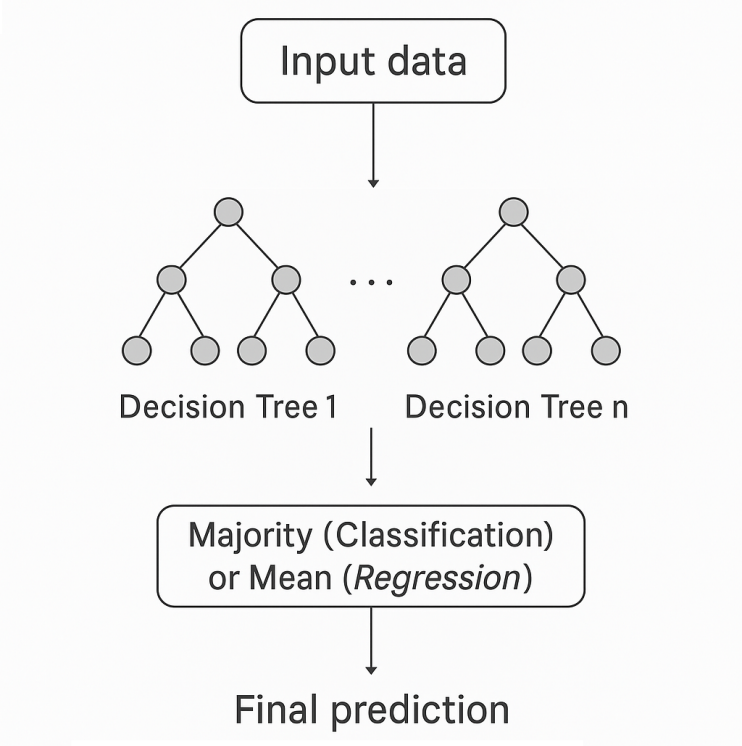

Figure 3.3: Random Forest Diagram

3.3.4 Extreme Gradient Boosting (XGBoost)

XGBoost is a scalable and efficient implementation of the gradient boosting framework, introduced by (CHEN; GUESTRIN, 2016), designed to optimize both computational speed and model performance. Gradient boosting builds an ensemble of weak learners, typically shallow decision trees, in a sequential manner, where each new tree is trained to correct the residual errors of the previous ensemble. As a consequence, boosting primarily reduces *bias*, while *variance* can increase but is usually controlled via appropriate regularization and hyperparameter choices (JAMES et al., 2023).

XGBoost extends this concept with several engineering and algorithmic enhancements: (i) explicit L_1 and L_2 regularization terms in the objective function to control model complexity and prevent overfitting. (ii) sparsity-

aware algorithms to handle missing values and sparse datasets efficiently. (iii) weighted quantile sketch for approximate tree learning on large datasets. (iv) parallelized tree construction for high scalability (JAMES et al., 2023). These innovations make XGBoost particularly effective for large scale predictive modeling tasks with tabular data.

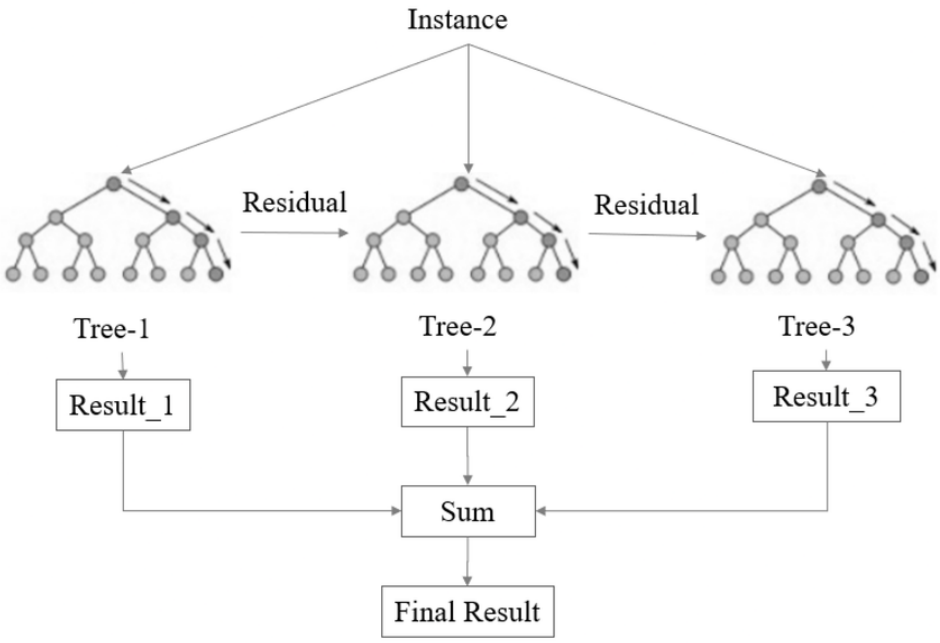

Figure 3.4: Simplified diagram of the XGBoost learning process. (WANG; CHAKRABORTY; CHAKRABORTY, 2020)

3.4 Deep Learning

Deep learning is a subfield of artificial intelligence (AI) that uses machine learning models composed of one or more processing layers arranged hierarchically, where higher-level features are formed by the combination of lower-level features. This allows the model to learn data representations with multiple levels of abstraction (LECUN; BENGIO; HINTON, 2015). Deep learning can discover complex structures in large datasets using backpropagation algorithms to indicate how the machine should change its internal parameters, which are used to compute representations in each layer based on the representation from the previous layer.

Its architecture is composed of multiple layers in which most (or all) components can learn. These models aim to simulate the human brain, where each neuron has different weights and, when activated (i.e., when the feature it computes is found), it becomes input to the next layer, which may or may not use that information. Many neurons compute nonlinear input-output

mappings. Each one transforms its input to increase both the selectivity and invariance of the representation. With a sufficient number of nonlinear layers, the model can implement highly complex functions over its inputs, which are simultaneously sensitive to details. Thus, each region of neurons, each layer in a neural network, combines patterns detected by the immediately preceding region to form more complex features (BEZERRA, 2016).

3.4.1 LSTM

Long Short-Term Memory (LSTM) networks are a specialized form of recurrent neural networks designed to overcome the limitations of traditional RNNs in learning long-term dependencies (GREFF et al., 2017). The core innovation of LSTMs lies in their memory cell architecture, which maintains information over extended sequences through a set of multiplicative gates (LINDEMANN et al., 2021).

The input gate regulates the flow of new information into the cell state, determining which parts of the input are relevant to store. The forget gate controls which information in the cell state should be discarded, allowing the network to remove irrelevant or outdated data. The output gate manages the exposure of the cell state to the next layers, deciding what information is passed forward in the sequence. Together, these gates enable precise control over the storage, retention, and retrieval of information, mitigating issues such as vanishing or exploding gradients common in standard RNNs (GOODFELLOW; BENGIO; COURVILLE, 2016).

The figure 3.5 shows the schematic diagram of an LSTM model. This gated mechanism allows LSTMs to effectively capture patterns in sequential data where dependencies can span large temporal distances. As a result, LSTMs have become widely used in domains such as natural language processing, speech recognition, and time series forecasting, particularly in cases where the influence of earlier observations is critical to predicting future outcomes (LECUN; BENGIO; HINTON, 2015). Their ability to balance short-term adaptability with long-term memory retention makes them a strong candidate for modeling complex temporal dynamics in real-world datasets (LINDEMANN et al., 2021).

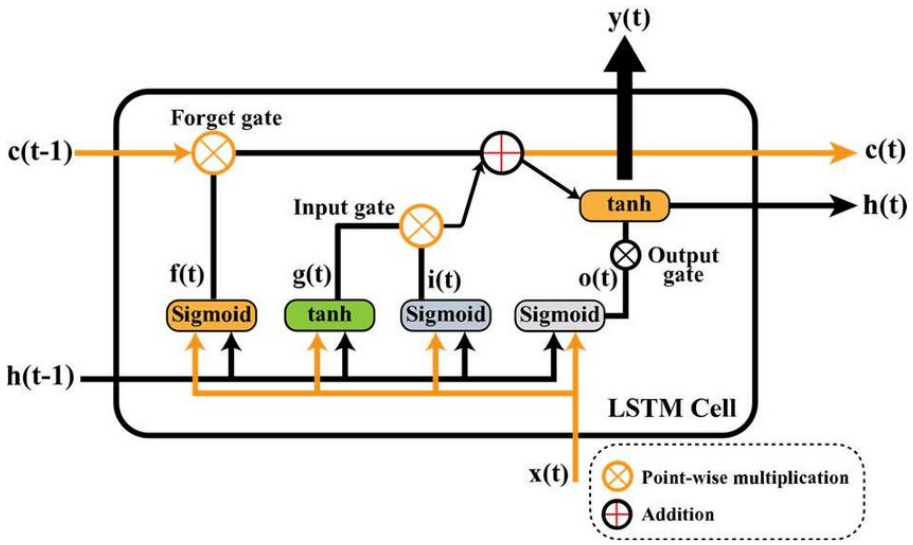

Figure 3.5: Schematic diagram of LSTM model (TAMAL et al., 2022)

3.5 Evaluation Metrics

Evaluating the performance of predictive models is essential to ensure their generalization capability. As presented in (KUHN; JOHNSON, 2013), the employed metrics quantify the difference between predicted and observed values, focusing on measures based on absolute and squared error.

3.5.1 Coefficient of Determination (R^2)

The coefficient of determination quantifies the proportion of the total variance in that is explained by the model. It is defined as:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{3-2}$$

According to Kuhn e Johnson (2013), although R^2 is widely used, it can be misleading in extrapolation scenarios or when the model suffers from overfitting.

3.5.2 Root Mean Squared Error (RMSE)

The RMSE penalizes large individual errors more heavily, making it appropriate when the goal is to minimize substantial deviations (KUHN; JOHNSON, 2013):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2} \tag{3-3}$$

3.5.3
Mean Absolute Error (MAE)

The MAE calculates the mean of the absolute errors between predictions and observed values (FERREIRA, 2020):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3-4}$$

3.5.4
Mean Absolute Percentage Error (MAPE)

The MAPE, similar to the MAE, expresses the average difference between actual and predicted values as a percentage (FERREIRA, 2020). Despite being intuitive and unitless, it has caveats: it is undefined at $y = 0$ and penalizes underestimates when y is small, which can induce bias. Here, a small ϵ mitigates division by zero, and MAPE is interpreted alongside MAE/RMSE. Its formula is given by:

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i + \epsilon}\right| \tag{3-5}$$

4
Methods

The implemented system employs multiple architectures to forecast NOx emissions over time windows. The approaches include traditional statistical ML models as well as a RNN based on LSTM enabling comparisons between different modeling strategies. This chapter presents the framework 4.1, the data sources 4.2, the preprocessing steps 4.3, the models adopted 4.4, and the configuration of the experiments 4.5.

4.1
Framework Overview

To structure the methodology and highlight the novelty of this work, we propose a unified framework for NOx forecasting (Figure 4.1). The framework is composed of five main stages:

1. **Raw Data Collection:** Integration of heterogeneous sources, including onboard emission sensors, vehicle specifications, trip records, meteorological stations, and geographical data.

2. **Pre-processing:** Data cleaning, validation, and temporal/spatial interpolation of meteorological variables (using linear/cubic interpolation and kriging).

3. **Trip Segmentation:** Division of trips into homogeneous sub-segments based on continuity rules, ensuring temporal consistency for forecasting.

4. **Data Modeling:** Construction of feature sets (F1F4) combining operational, meteorological, and contextual variables.

5. **Model Training and Forecasting:** Comparative evaluation of regression models (Linear Regression, Random Forest, XGBoost, and LSTM) under multiple forecasting strategies (single-step, recursive, multi-output, and direct).

This high-level pipeline emphasizes the methodological contribution of the study: combining continuous real-world vehicle data with interpolated meteorological features and systematically benchmarking forecasting strategies rarely explored in the literature.

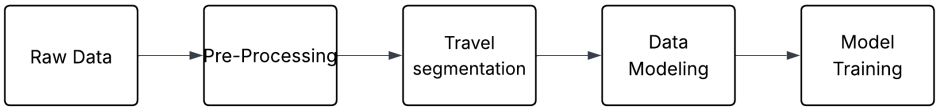

Figure 4.1: Proposed framework for NOx emission forecasting.

4.2 Data Sources

The data used in this study were collected from multiple reliable sources, including onboard vehicle emission measurements, meteorological observations, and geographical elevation data. The integration of these sources enables a contextualized and detailed modeling of NOx emissions from heavy-duty vehicles. The data sources are described below:

- **Trip Data:** records of trip start and end times, distances traveled, timestamps, and aggregated information about transported cargo.
- **Emission Data:** continuous time series of NOx emissions and other operational parameters obtained from onboard sensors.
- **Vehicle Specifications:** information such as type, model, engine power, year of manufacture, and vehicle category.
- **Refueling Records:** logs of refueling operations performed during trips.
- **Meteorological Data:** collected from automatic weather stations located in the state of Rio de Janeiro, provided by the Brazilian National Institute of Meteorology (INMET), with hourly sampling.

Table 4.1 presents the main attributes extracted from the sources listed above, encompassing temporal, geographic, meteorological, operational, and contextual information that form the foundation for predictive modeling.

Table 4.1: Attributes of the data used in NOx emission forecasting

Source / Category	**Attribute(s)**	**Type**	**Description**
Emission (onboard)	timestamp	Datetime	Timestamp
	order	Integer	Data index
	SensorHours	Float	number running hours
	NOx, NOx_max, NOx_min, NOx_dp	Float	NOx levels
	samples	Integer	Sample count per reading
	O_2	Float	Oxygen concentration
	position	String	latitude and longitude
Vehicle Specifications	engine_power	Float	Engine power
	vehicle_model	String	Vehicle model
	vehicle_brand	String	Vehicle manufacturer
	vehicle_weight	Float	vehicle Weight
	tank_capacity	Float	tank size
	vehicle_classification_euro, vehicle_type	String	Environmental classification
	vehicle_type	String	operational classification
	vehicle_pto_use	Boolean	PTO use
	vehicle_gas_reduction	Boolean	gas reduction status
	vehicle_year	Integer	Year of manufacture
	plate, vehicle_card	String	License plate and ID card
	fleet_number	Integer	Fleet number
	fleet_type	String	Fleet category
Trip Data	tripNumber	Integer	Unique trip ID
	travel_trip, tripVehicleNumber	String	Vehicle trip association
	startedAt, finishedAt, actualTripStartDate, actualTripEndDate	Datetime	Start and end trip timestamp
	stopNumber	Integer	Stop sequence identifier
	stopName, stopType	String	Stop label and classification
	stopOdometer	Float	Odometer
	stopVehicleWeightBefore, stopVehicleWeightAfter	Float	weight metrics per stop
	tripInitialWeightOfVehicle, tripFinalWeightOfVehicle	Float	Weight at trip start and end
	locationCode, vacuumSpec	String	Trip location and vacuum specs
Meteorology (INMET)	Total Precipitation	Float	Hourly Cumulative rainfall
	Atmospheric Pressure	Float	pressure levels
	Air Temperature, Dew Point	Float	Temperature and dew point
	Max/Min Temperature (prev. hr)	Float	Past hour temperature extremes
	Relative Humidity (avg/max/min)	Float	Humidity range statistics
	Wind (direction, gust, speed)	Float	Wind speed, gusts, and direction
	Global Radiation	Float	Incoming solar radiation
Geography (AMB-DATA)	elevation, delta_elevation	Float	Altitude and elevation variation

4.3 Preprocessing

The preprocessing phase ensures data quality for future analyses by removing inconsistent or irrelevant records. This step is critical in environmental studies that deal with large volumes of data, as emphasized by (HAN; KAMBER; PEI, 2012).

4.3.1 Validation of NOx Emissions

Based on established vehicle emission standards, a valid range for NOx levels was defined, along with validation rules involving oxygen concentration and geographic boundaries. Records that did not meet these conditions were removed:

- Minimum NOx level: 100 ppm.
- Maximum NOx level: 1500 ppm.
- Atmospheric O_2 concentration must not exceed 21%.
- Higher levels of O_2 typically correspond to lower NOx emissions, and vice versa.
- Latitude must be between -26.0 and -19.0, and longitude between -46.0 and -39.0.

4.3.2 Temporal Interpolation of Meteorological Data

To ensure that the meteorological data have appropriate temporal resolution for time series modeling, a temporal interpolation process was applied to each weather station.

Initially, meteorological data were collected from annual CSV files (2022, 2023, 2024), containing hourly information from various INMET automatic weather stations located in the state of Rio de Janeiro. Each annual dataset was linked to a complementary file containing the geographic coordinates (latitude, longitude, and altitude) of each station.

After loading the data, column names were standardized to ensure consistency during processing. Subsequently, the datasets from all meteorological stations were concatenated into a unified structure. Irrelevant attributes such as global radiation, station foundation date, and administrative metadata were removed, preserving only the core meteorological variables: temperature, humidity, wind, atmospheric pressure, and precipitation. The following preprocessing steps were then applied:

1. Samples were grouped by hour using the `data_hora` field, rounded to the nearest hour.

2. Numerical variables were then interpolated to fill hourly gaps. Linear interpolation was used for most variables, while cubic interpolation was employed specifically for precipitation due to its more abrupt variability.

3. After hourly interpolation, the data were upsampled to a one-minute resolution.

4. A second interpolation step was performed at the minute level, using the same interpolation methods for each variable.

5. Any remaining missing values were filled using forward and backward propagation of the nearest valid observations.

6. Station specific constant columns, such as latitude, longitude, altitude, and station name, were preserved unchanged.

The final result was a set of minute resolution meteorological data for all stations. These data were saved in `Parquet` format to enable efficient querying and merging with other temporally indexed data sources, such as vehicle emission data.

This process is crucial for enabling temporal fusion between onboard truck sensor data (which has high granularity) and meteorological data, which originally have hourly resolution and may contain temporal gaps.

4.3.3 Spatial Interpolation of Meteorological Data (Kriging)

Following the temporal interpolation of meteorological data, spatial interpolation was conducted to estimate weather variables (precipitation, temperature, wind, and others) at the precise positions of the trucks over time.

For this purpose, the Kriging method with an exponential model was employed. Kriging is a well established geostatistical technique that estimates values at unsampled points by leveraging spatial correlations from neighboring sampled data (KRIGE, 1951; MATHERON, 1962; CRESSIE, 1993).

The interpolation process involved the following steps:

1. For each NOx emission sample recorded by a truck, all meteorological stations that had measurements at the corresponding minute were identified.

2. The geodesic distance between the truck's location and each available meteorological station was computed. Only stations within a 50 km radius were retained.

3. Using the coordinates and corresponding values of the filtered stations, the Kriging method was applied to estimate the meteorological variables at the truck's exact position.

4. This procedure was repeated for each meteorological variable of interest: precipitation, atmospheric pressure, air temperature, dew point temperature, and relative humidity.

In addition, truck refueling data were incorporated through an asynchronous join based on timestamps using the `merge_asof` method. This enabled the association of each emission record with the most recent available refueling information prior to the event.

Finally, the complete dataset comprising NOx emission records, spatially interpolated meteorological variables, and refueling data was saved in `Parquet` format.

This procedure ensures that NOx emission forecasting models account for climate data that are both spatially aligned and temporally synchronized with the actual vehicle positions, thereby enhancing the contextual accuracy and predictive performance of the models.

4.3.4 Trip Segmentation

To enhance the granularity of the analysis and enable more efficient modeling, complete trips (`tripNumber`) were divided into smaller segments based on the temporal sequence of the data.

The segmentation process was conducted through the following steps:

1. Time differences between consecutive readings within each trip were calculated. When the time gap exceeded 1 minute, it was considered a discontinuity in the trip, indicating the beginning of a new segment.

2. A sequential identifier, `mini_tripNumber_id`, was assigned to mark each new segment within the same trip.

3. Segments with insufficient data points were filtered. Only segments containing at least 12 records (12 minutes) were retained for analysis, ensuring a minimal temporal base for predictive modeling.

4. Finally, the resulting DataFrame was reordered and reindexed, yielding a dataset composed of multiple segments per original trip, with an average segment size of approximately `50.6` readings.

This procedure enabled structured handling of data discontinuities, while also reducing the impact of interruptions, long stops, or sensor failures by segmenting the data into temporally cohesive and homogeneous blocks.

4.4 Regression Models

For the NOx emission forecasting task, a combination of statistical, ensemble, and deep neural network models was adopted to capture different aspects of meteorological and environmental data.

LR was included as a baseline model due to its simplicity and interpretability. It is often used as a reference in regression tasks (JAMES et al., 2023).

RF was selected for its robustness to correlated variables, its ability to capture nonlinear interactions, and its good performance even without intensive hyperparameter tuning. It has been successfully applied in environmental and geoscientific contexts (SILVA et al., 2020).

XGBoost was incorporated for its high effectiveness on tabular data and its recognition as one of the most accurate boosting algorithms for structured problems, including recent studies on pollutant emission forecasting (CHEN; GUESTRIN, 2016; WEN; LU; JHANG, 2021; PARK et al., 2023).

Finally, the **LSTM** network was employed for its ability to model long-term temporal dependencies, a critical feature in meteorological time series, and has been recommended in several studies for emission and environmental variable forecasting (HOCHREITER; SCHMIDHUBER, 1997; GREFF et al., 2017; LINDEMANN et al., 2021).

This model combination aims to comparatively evaluate linear, tree-based, and deep learning approaches, providing a comprehensive analysis of the forecasting problem.

4.4.1 Evaluation Metrics

Rigorous evaluation of model performance is essential in environmental applications such as vehicle emission forecasting, due to the direct impact on public policy formulation and mitigation strategies. As highlighted by (KUHN; JOHNSON, 2013), the use of multiple metrics allows for capturing

different aspects of model performance, including average accuracy, robustness to outliers, and explanatory power.

In this work, widely adopted metrics in the regression literature were used to quantify the discrepancy between predicted and observed values. Four metrics were employed to assess the predictive quality of the models: **Coefficient of Determination (R^2)**, **Root Mean Squared Error (RMSE)**, **Mean Absolute Error (MAE)**, and **Mean Absolute Percentage Error (MAPE)**.

The choice of these metrics is aligned with best practices in predictive modeling applied to environmental and time series data, as discussed in (KUHN; JOHNSON, 2013) and (FERREIRA, 2020). Their combined use ensures a comprehensive assessment of model performance, addressing accuracy, robustness, and practical relevance.

4.5 Experiment Configuration

This section outlines the structure of the experiments conducted, focusing on data splitting and training strategies. The motivations for the implemented models are detailed in Subsection 4.4, while the performance metrics are described in Subsection 4.4.1.

4.5.1 Dataset Splitting

All numerical variables were normalized using the *MinMaxScaler*. A 5-fold cross-validation strategy ($K = 5$) was adopted, ensuring independence across vehicles and trips. In this strategy, each fold splits entire trips into 75% for training, 15% for validation, and 10% for testing. No portion of a trip used for training was included in the validation or testing sets. Importantly, the splits were *not chronological*: trips were randomly assigned to folds, so validation/test trips may occur earlier or later in time than training trips.

For each trip, after some testing, sliding windows of size $w = 10$ were created. For the finals RF and LR models, training and validation sets were combined for final model fitting. The input data format was adjusted according to the model type: flattened feature vectors for classical machine learning models and temporal sequences for the LSTM neural network.

4.5.2 Multi-Horizon Forecasting Strategies

Beyond single-step NOx emission forecasting, this study explores multi-horizon forecasting strategies, i.e., estimating multiple future time steps of the target variable. These approaches are essential for real-world applications such as route planning, adaptive engine control, or forecasting environmental impact over longer windows.

Three distinct methods were implemented and compared:

1. **Recursive Forecasting:** A single model is trained to predict the next NOx value ($t + 1$). The predicted value is then fed back as input to generate the subsequent prediction, iteratively continuing until the desired forecast horizon is reached. This approach is simple and efficient but may suffer from error accumulation over time (GOODFELLOW; BENGIO; COURVILLE, 2016).

2. **Multi-output Forecasting:** The model is trained to predict an entire sequence of future NOx values in a single step, producing a vector output. This approach enables the model to capture dependencies between future steps and maintain consistency across the predicted horizon, as discussed in (LI; ZHANG; DONG, 2023).

3. **Direct Forecasting:** Separate models are trained to predict each future time step independently ($t + 1$, $t + 2$, ..., $t + h$), without relying on recursive inputs. This method avoids error propagation but entails higher computational costs, as it requires fitting multiple independent models (CHEVILLON, 2007).

5
Results

This chapter presents and analyzes the results obtained from preprocessing to the predictive modeling of NOx emissions, based on operational and environmental data collected from heavy-duty vehicles. The experiments are organized into four major parts: preprocessing, feature selection, spatial analysis of emissions, and model evaluation under different forecasting strategies. The main findings are detailed below.

5.1
Preprocessing

This section outlines the data filtering and transformation procedures employed to ensure data quality. Four major steps were undertaken: NOx emission validation, temporal interpolation of meteorological data, spatial interpolation via kriging, and segmentation of trips into smaller segments.

The process began with 1,852,606 raw readings collected from onboard NOx emission sensors. After applying validation criteria, 137,831 inconsistent readings were removed, resulting in a total of 1,714,775 valid NOx measurements with 12 columns (vehicle_number, vehicle_name timestamp, order, Sensor_Hours, NOx, NOx_max, NOx_min, NOx_dp, O2, longitude, latitude).

Subsequently, multiple data sources were merged, including emission records, trip data, vehicle specifications, and refueling logs. Vehicle data were joined using the vehicle ID, while trip and refueling data were aligned using both vehicle ID and timestamp to ensure accurate temporal matching. This integration expanded the dataset from 12 to 49 columns but reduced the number of rows to 543,733, as emission records without corresponding trip information were discarded.

Next, hourly meteorological data from 28 INMET weather stations across the state of Rio de Janeiro were processed. Time series spanning 2022 to 2024 were concatenated and interpolated to fill missing hourly entries using linear interpolation. Minute-level interpolation followed to match the granularity of emission data, resulting in a meteorological dataset of 37,137,506 rows and 23 columns.

With temporally aligned meteorological data, spatial interpolation was performed using kriging. For each emission record, weather stations within a 50 km radius were used to estimate local environmental conditions. The

variables subjected to spatial interpolation included precipitation, atmospheric pressure, air temperature, dew point temperature, and relative humidity. This step increased the number of columns from 49 to 54 and the number of rows stayed the same.

Finally, trips were segmented into smaller, coherent segments based on temporal continuity. Segments with fewer than 12 records were discarded. Irrelevant or duplicate columns were removed using objective criteria: (i) very high proportions of missing values (no usable coverage across trips), (ii) constant columns (single value), and (iii) Administrative or redundant fields that do not provide domain information beyond other maintained variables. The final preprocessed dataset consisted of 326,805 records and 27 columns:

`timestamp`, `vehicle_number`, `vehicle_model`, `vehicle_type`, `vehicle_year`, `dia_da_semana`, `periodo_dia`, `latitude`, `longitude`, `distancia_metros`, `delta_elevation`, `velocidade_media_kmh_ajustada`, `flag_velocidade`, `aceleracao_media_kmh2_ajustada`, `NOx`, `tripNumber`, `status`, `peso_atual`, `pressao_atmosferica`, `temperatura_orvalho`, `temperatura_ar`, `umidade_relativa`, `consumption`, `km_driven`, `odometer`, `mini_tripNumber_id`, `chuva`.

These features include temporal data (e.g., timestamp, day of the week), geolocation (latitude and longitude), meteorological conditions, vehicle characteristics (e.g., average speed, vehicle type), and trip metadata (e.g., trip ID, segment ID).

This preprocessing ensures that trips are isolated and well-defined for each vehicle, preventing data leakage across trips during model training. Details of the data splitting strategy are provided in Subsection 4.5.1.

5.2 Feature Selection

To determine the most relevant features for the predictive modeling of NOx emissions, a statistical correlation analysis with the target variable was conducted, complemented by feature importance rankings obtained from the *Random Forest* and *XGBoost* algorithms.

Tables 5.1 and 5.2 present the Pearson and Spearman correlation coefficients between each attribute and NOx emissions. The variable `velocidade_media_kmh_ajustada` (adjusted average speed) exhibited the strongest correlations in both analyses.

This finding aligns with the results of (SILVA et al., 2025), who demonstrated, based on real-world data collected from continuously operating urban

buses, that average speed is one of the key determinants of NOx emission variability.

Table 5.1: Pearson correlation coefficients with NOx emissions

Variable	**Pearson Coefficient**
adjusted_average_speed_kmh	0.466
distance_meters	0.257
odometer	0.202
vehicle_model	0.122
consumption	0.112
adjusted_average_acceleration_kmh2	0.076
longitude	0.057
latitude	0.049
air_temperature	0.034

Table 5.2: Spearman correlation coefficients with NOx emissions

Variable	**Spearman Coefficient**
adjusted_average_speed_kmh	0.502
distance_meters	0.399
odometer	0.102
consumption	0.092
latitude	0.084
adjusted_average_acceleration_kmh2	0.076
air_temperature	0.062
longitude	0.045

Furthermore, Figures 5.1 and 5.2 show the feature importance rankings generated by the Random Forest and XGBoost models, respectively. It can be observed that variables related to vehicle dynamics (`adjusted_average_speed_kmh`, `distance_meters`, `odometer`) were consistently identified as the most relevant for predicting NOx emissions.

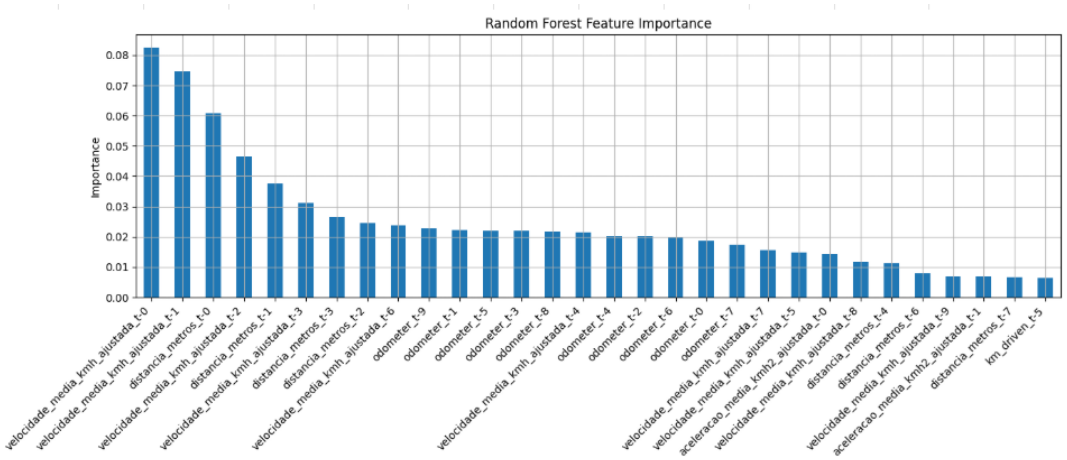

Figure 5.1: Feature importance according to the Random Forest model

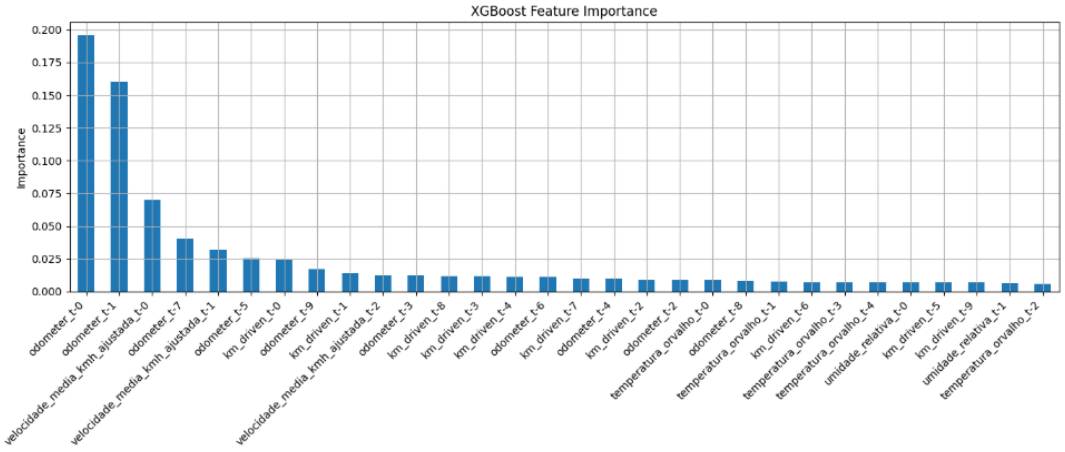

Figure 5.2: Feature importance according to the XGBoost model

Based on the combined analyses, feature selection was organized into four distinct experimental configurations (F1 to F4), as described below:

- **F1**: Includes spatial variables (`latitude`, `longitude`), vehicle dynamics (`average_speed`, `average_acceleration`, `distance_meters`), historical usage (`km_driven`, `odometer`), and the previous value of (`NOx`).
- **F2**: A subset of F1, excluding spatial variables. This configuration focuses solely on operational data from the vehicle.
- **F3**: An expansion of F1 that incorporates interpolated meteorological variables (`rain`, `air_temperature`, `dew_temperature`, `relative_humidity`).
- **F4**: A simplified version of F3, excluding latitude and longitude.

These configurations enable the evaluation of the relative contribution of different groups of features spatial, climatic, and operational, to the prediction of vehicle emissions, while consistently including the previous NOx value to reflect the temporal nature of the data.

5.3 Descriptive Analysis of the Dataset

The preprocessing pipeline began with 1,852,606 raw emission records from embedded vehicle sensors. After merging meteorological, operational, and trip data, and performing temporal and spatial interpolations, the final dataset comprised 326,805 refined records and 27 attributes. This dataset was then passed through the feature selection process, resulting in configurations F1 to F4 with 12 variables in its largest version. The following section present the descriptive statistics for the most relevant variables used in the modeling process.

5.3.1 Main Variables

Table 5.3 summarizes the key statistical characteristics of the main variables used in this study, including operational, emission, and meteorological data.

Table 5.3: Statistical summary of main variables

Variable	Mean	Std. Dev.	Min	Q1	Median	Q3	Max
Avg. Speed (km/h)	23.97	27.24	0.00	0.41	9.83	48.86	120.88
NOx (ppm)	389.47	217.19	100.00	231.00	322.00	485.00	1496.00
Kilometers Driven (km)	270.47	179.81	7.00	136.00	225.00	382.00	1358.00
Pressure (hPa)	1001.78	23.51	869.97	996.33	1010.56	1015.62	1033.34
Air Temp. (řC)	23.03	4.43	8.17	20.00	23.03	26.03	38.08
Humidity (%)	78.10	15.54	16.43	67.26	80.44	91.43	102.80
Rain (binary)	0.57	0.50	0.00	0.00	1.00	1.00	1.00

The statistical summary (Table 5.3) reveals key characteristics of the dataset across operational, emission, and environmental dimensions. Operational variables such as average speed and kilometers driven exhibit substantial dispersion (Std. Dev. = 27.24 and 179.81, respectively), reflecting heterogeneity in vehicle activity and route profiles. NOx emissions show a wide range (100-1496 ppm), with a mean of (389.47 ppm), indicating frequent peaks likely linked to traffic congestion, acceleration events, or topographical variation.

Meteorological variables also display marked variability: air temperature ranges from 8.17řC to 38.08řC, consistent with the climatic diversity encountered along the fleet's routes. The binary rain variable has a mean of 0.57, suggesting that precipitation was present in more than half the recorded trips.

These variations across domains highlight the complexity of the emission process and reinforce the importance of incorporating both operational and environmental variables in predictive modeling.

5.4 Spatial Visualization of Emissions Along a Route

Figure 5.3 illustrates the spatial distribution of NOx emissions along a real route traveled by one of the fleet vehicles. Emissions are displayed on a map using a color gradient.

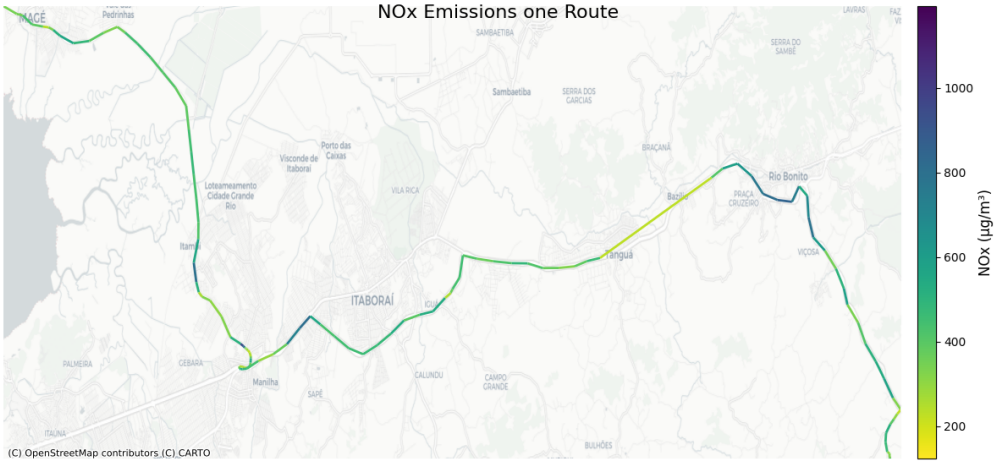

Figure 5.3: NOx emissions along a route ($\mu g/m^3$)

It is evident that the highest emission levels (represented in purple/dark blue) tend to occur in urban segments or areas with heavy traffic, such as around Itaboraí and Rio Bonito. This pattern is consistent with the findings of (ZENG; MIWA; MORIKAWA, 2016), who emphasized the role of acceleration and average speed in low flow zones as critical factors influencing NOx emissions.

5.5 Model Hyperparameters

The hyperparameters used for each predictive model are detailed below:

- **Random Forest (RF):** `n_estimators = 150`, `max_depth = 10`, `min_samples_split = 5`, `min_samples_leaf = 10`, `max_features = "sqrt"`.
- **XGBoost:** `n_estimators = 300`, `max_depth = 10`, `learning_rate = 0.05`, `subsample = 0.8`, `colsample_bytree = 0.8`, `early_stopping_rounds = 10`.
- **LSTM:** Architecture: `LSTM(64)` → `LSTM(32)` → `Dense(`n`_horizon)`. Training strategies included `ReduceLROnPlateau` and `ModelCheckpoint`.

5.6 Comparative Analysis of Models and Feature Configurations

This section presents a comparative performance analysis of four predictive models: Linear Regression, Random Forest, XGBoost, and LSTM. Evaluated across two vehicle types (Small Truck and Medium Truck) and four distinct feature configurations (F1 to F4). The models were assessed under four different forecasting scenarios:

1. **Single-step forecasting**: the model predicts only the next value of the time series at each iteration;

2. **Recursive forecasting by trip**: the predicted value is recursively fed back as input for the subsequent prediction, simulating a continuous forecasting cycle over the entire trajectory;

3. **Direct forecasting**: a separate model is trained for each forecasting horizon ($n = 10$);

4. **Multi-output forecasting**: a single model is trained to simultaneously predict the next $n = 20$ values in the time series.

This multi-scenario approach allows for the evaluation of not only the relative importance of features and the robustness of the models, but also the impact of different temporal forecasting strategies on predictive performance.

5.6.1 Single-Step Forecasting

Tables 5.4 and 5.5 summarize the results obtained for the two vehicle types considered: Small Truck and Medium Truck. Under each of the four feature configurations (F1 to F4).

Table 5.4: Model performance for Small Truck under configurations F1-F4

Configuration	**Model**	R^2	RMSE	MAE	MAPE
F1	Linear Regression	91.96	75.36	48.34	11.76
	Random Forest	84.19	105.68	68.05	16.49
	XGBoost	86.91	96.16	58.85	13.34
	LSTM	**93.29**	**68.84**	**41.69**	**10.28**
F2	Linear Regression	92.01	75.14	48.20	11.72
	Random Forest	85.06	102.72	66.00	15.97
	XGBoost	88.13	91.55	56.36	12.74
	LSTM	**92.99**	**70.38**	**45.42**	**11.50**
F3	Linear Regression	91.92	75.54	48.62	11.89
	Random Forest	83.07	109.35	71.38	17.47
	XGBoost	86.43	97.90	60.12	13.67
	LSTM	**92.45**	**73.01**	**46.33**	**11.97**
F4	Linear Regression	91.96	75.34	48.49	11.86
	Random Forest	84.34	105.18	68.17	16.64
	XGBoost	88.71	89.28	55.47	12.69
	LSTM	**92.49**	**72.82**	**42.06**	**9.77**

Table 5.5: Model performance for Medium Truck under configurations F1-F4

Configuration	**Model**	R^2	RMSE	MAE	MAPE
F1	Linear Regression	**90.00**	**54.32**	36.80	11.11
	Random Forest	80.10	76.62	46.97	13.61
	XGBoost	85.41	65.62	38.46	10.90
	LSTM	87.20	61.46	**31.46**	**8.29**
F2	Linear Regression	**90.07**	**54.14**	36.63	11.07
	Random Forest	81.80	73.27	44.21	12.82
	XGBoost	86.52	63.07	36.54	10.33
	LSTM	87.63	60.41	**31.48**	**8.51**
F3	Linear Regression	**89.94**	**54.48**	36.95	11.14
	Random Forest	77.40	81.65	51.00	14.68
	XGBoost	85.08	66.33	38.97	10.91
	LSTM	89.14	56.59	**32.46**	**9.51**
F4	Linear Regression	90.00	54.31	36.81	11.10
	Random Forest	80.12	76.60	46.63	13.41
	XGBoost	86.31	63.55	36.88	**10.37**
	LSTM	**91.14**	**51.14**	**32.90**	10.51

For the **Small Truck**, the LSTM achieved the best results, achieving the best overall results across all metrics. With errors dropping as low as 9.77% in F4 and 10.28% in F1. Its R^2 values exceeding 92% in all cases. LR was the second most consistent model, maintaining $R^2 \approx 92\%$ and MAPE around 11-12%, while XGBoost improved accuracy when additional features were included (e.g., F4). Random Forest, on the other hand, showed the weakest performance, with higher error rates across all settings.

For the **Medium Truck**, LR demonstrated remarkable stability, achieving the highest R^2 in F1-F3 (90%) and maintaining MAPE near 11%. However, the LSTM once again provided the lowest MAE and MAPE values, reaching as little as 8.29% in F1 and 8.51% in F2. Its best overall performance occurred in F4, where it achieved the highest R^2 (91.14) along with the lowest errors. XGBoost ranked just below these two, showing robust results when meteorological variables were included, while Random Forest consistently lagged in accuracy.

In summary, **LSTM was the dominant model for both vehicle types**, consistently achieving the lowest MAE and MAPE across all configurations, while also delivering high R^2 values. Linear Regression proved to be a stable and competitive baseline, particularly for the Medium Truck, whereas XGBoost showed incremental benefits with richer feature sets. Random Forest was the least effective across both datasets.

5.6.2 Recursive Forecasting by Trip

In the recursive forecasting scenario, each prediction is fed back as input into the model, simulating a continuous forecasting process along the trajectory. Tables 5.6 and 5.7 summarize the performance of the models for Small and Medium Trucks under the different feature configurations (F1-F4).

Table 5.6: Performance of models for Small Truck - Recursive forecast per trip (F1-F4)

Configuration	**Model**	R^2	RMSE	MAE	MAPE
F1	Linear Regression	**88.55**	**86.41**	**55.60**	**14.40**
	Random Forest	80.22	113.57	74.78	19.84
	XGBoost	83.70	103.10	64.39	15.87
	LSTM	64.19	152.81	108.22	30.91
F2	Linear Regression	88.75	85.66	54.75	14.07
	Random Forest	81.07	111.10	73.15	19.26
	XGBoost	84.89	99.27	61.37	14.93
	LSTM	**89.25**	**83.73**	**48.06**	**12.57**
F3	Linear Regression	**88.51**	**86.55**	**55.94**	**14.57**
	Random Forest	78.33	118.88	79.71	21.47
	XGBoost	83.32	104.28	65.81	16.76
	LSTM	66.44	147.94	101.06	25.90
F4	Linear Regression	88.66	85.98	55.26	14.26
	Random Forest	80.06	114.04	75.74	20.00
	XGBoost	84.79	99.59	62.95	15.87
	LSTM	**90.58**	**78.39**	**43.92**	**11.29**

Table 5.7: Performance of models for Medium Truck - Recursive forecast per trip (F1-F4)

Configuration	**Model**	R^2	RMSE	MAE	MAPE
F1	Linear Regression	**84.62**	**67.36**	42.30	12.55
	Random Forest	73.02	89.23	54.35	15.62
	XGBoost	80.78	75.31	**42.08**	**11.71**
	LSTM	1.59	170.41	120.67	37.29
F2	Linear Regression	**84.70**	**67.20**	42.12	12.50
	Random Forest	75.17	85.59	51.17	14.72
	XGBoost	81.66	73.56	40.64	11.28
	LSTM	84.18	68.32	**33.53**	**8.89**
F3	Linear Regression	**84.55**	**67.51**	**42.48**	12.58
	Random Forest	69.94	94.18	58.67	16.75
	XGBoost	80.02	76.78	43.40	**11.89**
	LSTM	n/a	n/a	n/a	n/a
F4	Linear Regression	84.62	67.36	42.30	12.53
	Random Forest	73.19	88.95	53.63	15.24
	XGBoost	81.31	74.26	**41.23**	**11.37**
	LSTM	**85.62**	**65.15**	42.37	14.54

For the **Small Truck**, LR achieved consistently high performance across all configurations, with R^2 values above 88% and MAPE around 14-15%. However, the **LSTM** demonstrated superior performance in F2 and F4, where it achieved the lowest errors, with R^2 values up to 90.58% and MAPE dropping to 11.29%. On the other hand, LSTM performance deteriorated sharply in F1 and F3, where the inclusion of spatial features led to reduced accuracy ($R^2 < 67\%$, MAPE $> 25\%$). XGBoost and Random Forest presented stable but lower accuracy than Linear Regression, with R^2 values between 78-85% and MAPE between 15-21%.

For the **Medium Truck**, LR was again the most stable model, delivering consistent performance across F1-F4 with $R^2 \approx 84.5\%$ and MAPE around 12-13%. XGBoost performed competitively, especially in error-based metrics, where it achieved the lowest MAE and MAPE in F1 and F3 (e.g., MAE = 42.08, MAPE = 11.71 in F1). The **LSTM**, showed high sensitivity to feature configurations: in F2 it reached a MAPE of just 8.89%, the best result among all models and configurations. Nevertheless, it was highly unstable in F1 and did not learn in F3. These outcomes highlight the vulnerability of LSTM to spatial feature sets in recursive mode.

Overall, recursive forecasting per trip shows that while classical models such as Linear Regression and XGBoost provide stable and reliable predictions, LSTM has the potential for the best accuracy but is highly dependent on

feature configuration. In particular, F2 and F4 configurations proved to be the most favorable for LSTM, enabling significantly better error reductions compared to the other approaches.

Figures 5.4 and 5.5 illustrate the predictive performance of the LSTM model along full trajectories under the favorable F4 configuration, showing strong alignment between predicted and observed values.

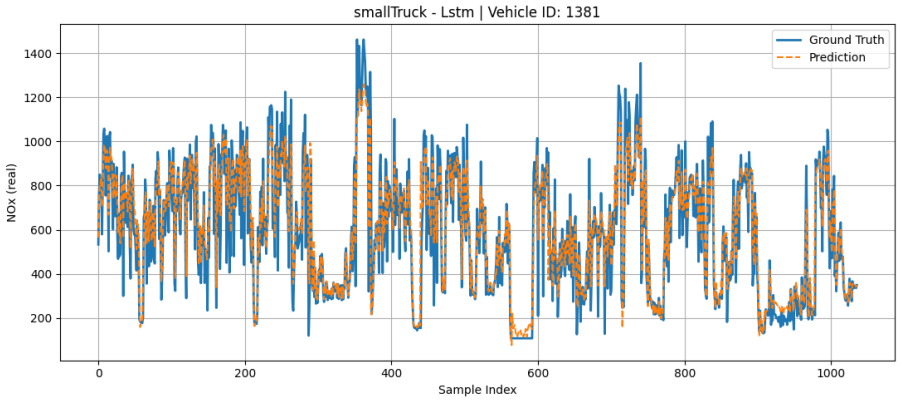

Figure 5.4: Recursive forecast for Small Truck using LSTM - Example from vehicle 1381 (F4)

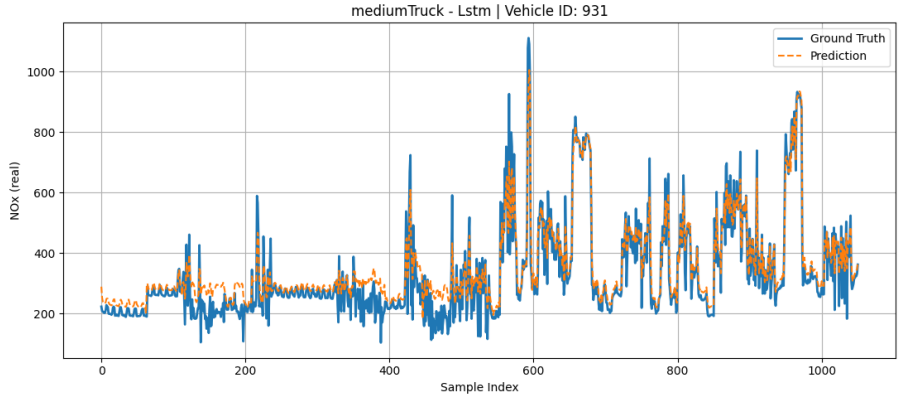

Figure 5.5: Recursive forecast for Medium Truck using LSTM - Example from vehicle 931 (F4)

5.6.3 Direct Forecasting ($n = 10$)

In this scenario, independent models were trained for each forecasting horizon ($n = 10$), aiming to evaluate step by step performance. Despite being conceptually promising, this approach yielded the worst results among all conducted experiments.

Tables 5.8 and 5.9 present the performance for both Small and Medium Trucks under the four feature configurations (**F1-F4**). The overall results

reveal that R^2 values remained substantially lower than those obtained with either the recursive or multi-output approaches, indicating limited ability to capture temporal dependencies when training separate models for each horizon.

Table 5.8: Performance of models for Small Truck - Direct Forecasting (F1-F4, $n = 10$)

Configuration	**Model**	R^2	RMSE	MAE	MAPE
F1	Linear Regression	61.15	166.45	112.28	27.86
	Random Forest	63.84	160.60	110.85	28.20
	XGBoost	**65.07**	**157.83**	106.91	27.06
	LSTM	60.70	167.42	**106.71**	**25.79**
F2	Linear Regression	61.06	166.65	112.40	27.89
	Random Forest	**61.64**	**165.40**	113.73	28.60
	XGBoost	61.56	165.59	112.84	28.05
	LSTM	59.55	169.84	**108.80**	**26.04**
F3	Linear Regression	62.27	163.04	108.64	28.94
	Random Forest	60.18	167.48	113.34	30.65
	XGBoost	**62.62**	**162.26**	108.17	28.82
	LSTM	61.20	165.34	**104.44**	**25.96**
F4	Linear Regression	61.30	166.14	112.25	28.02
	Random Forest	63.38	161.60	112.03	28.70
	XGBoost	**64.87**	**158.29**	108.81	28.02
	LSTM	59.36	170.24	**110.49**	**26.27**

Table 5.9: Performance of models for Medium Truck - Direct Forecasting (F1-F4, $n = 10$)

Configuration	**Model**	R^2	RMSE	MAE	MAPE
F1	Linear Regression	48.42	123.87	86.35	25.96
	Random Forest	48.55	123.72	86.53	25.68
	XGBoost	**51.61**	**119.98**	82.26	24.26
	LSTM	46.45	126.22	**78.80**	**21.90**
F2	Linear Regression	48.52	123.75	86.09	25.88
	Random Forest	48.91	123.27	86.07	25.58
	XGBoost	**49.96**	**122.01**	84.34	24.86
	LSTM	46.61	126.03	**78.89**	**22.37**
F3	Linear Regression	48.48	123.80	86.17	25.85
	Random Forest	46.60	126.03	88.39	26.12
	XGBoost	**50.50**	**121.35**	83.94	24.70
	LSTM	46.13	126.59	**81.65**	**23.28**
F4	Linear Regression	48.60	123.66	85.98	25.79
	Random Forest	47.42	125.06	87.69	25.97
	XGBoost	**49.32**	**122.79**	85.26	25.14
	LSTM	45.77	127.01	**81.01**	**23.36**

For the **Small Truck**, XGBoost provided the best R^2 across most configurations (up to 65.07 in F1 and 62.62 in F3), while Random Forest and Linear Regression trailed slightly behind. However, the **LSTM** consistently outperformed in terms of error metrics, achieving the lowest MAE and MAPE values across all configurations (e.g., MAPE $\approx$ 25-26%). This suggests that while tree-based models explained more variance, LSTM was more effective at minimizing absolute and relative prediction errors.

For the **Medium Truck**, results followed a similar pattern. XGBoost again achieved the highest R^2 values across settings (peaking at 51.61 in F1 and 50.50 in F3), whereas the **LSTM** achieved the lowest error values.

In summary, the direct forecasting modeling approach highlighted a trade-off between variance explained and error minimization. XGBoost generally delivered the best R^2 and RMSE, but the LSTM consistently achieved the best MAE and MAPE across all scenarios, reinforcing its strength in error reduction. The overall lower performance compared to recursive and multi-output strategies underscores the limitations of training fully independent models for each forecasting step.

Figure 5.6 shows the full trajectory for the small Truck in the F3 configuration. Despite its simplicity and modularity, the limited performance of this approach highlights the challenges in capturing temporal dependencies when training a separate model for each forecasting step.

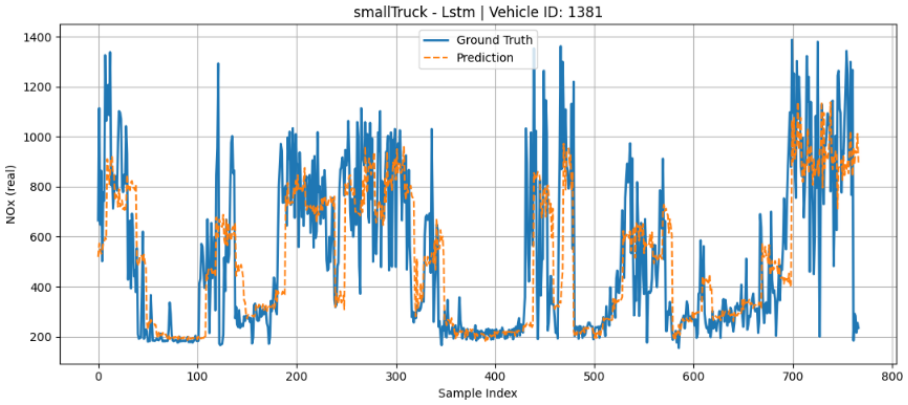

Figure 5.6: Direct forecasting ($n = 10$) - LSTM performance for Small Truck

5.6.4 Multi-output Forecasting ($n = 20$)

In this approach, a single model was trained to generate multiple outputs simultaneously ($n = 20$), producing the entire future sequence in one forward pass.

Overall, the multi-output models achieved competitive results, often surpassing the independent models in accuracy, although they still underperformed compared to the recursive forecasts. Tables 5.10 and 5.11 summarize the performance across all configurations (**F1-F4**).

Table 5.10: Performance of models for Small Truck - Multi-output forecasting (F1-F4, $n = 20$)

Configuration	Model	R^2	RMSE	MAE	MAPE
F1	Linear Regression	69.77	123.99	88.02	21.13
	Random Forest	74.59	113.70	81.97	20.59
	XGBoost	**76.76**	**108.72**	**77.15**	19.49
	LSTM	68.07	127.44	85.18	**18.34**
F2	Linear Regression	68.89	125.80	89.50	21.51
	Random Forest	**70.80**	**121.86**	87.53	21.38
	XGBoost	70.79	121.88	86.20	20.75
	LSTM	70.32	122.87	**79.45**	**17.21**
F3	Linear Regression	70.11	123.30	87.84	21.33
	Random Forest	74.56	113.74	82.52	21.11
	XGBoost	**74.65**	**113.55**	82.15	21.38
	LSTM	73.27	116.60	**75.73**	**16.84**
F4	Linear Regression	69.30	124.95	89.11	21.68
	Random Forest	**72.13**	**119.06**	86.11	21.65
	XGBoost	69.57	124.40	90.23	22.50
	LSTM	71.75	119.87	**77.32**	**16.99**

Table 5.11: Performance of models for Medium Truck - Multi-output forecasting (F1-F4, $n = 20$)

Configuration	**Model**	R^2	RMSE	MAE	MAPE
F1	Linear Regression	45.98	90.06	66.94	20.74
	Random Forest	53.88	83.22	63.13	19.72
	XGBoost	54.62	82.55	62.17	19.66
	LSTM	**55.40**	**81.83**	**59.05**	**17.38**
F2	Linear Regression	45.33	90.61	67.51	20.90
	Random Forest	**52.58**	**84.39**	64.79	20.04
	XGBoost	50.33	86.36	65.96	20.69
	LSTM	51.68	85.19	**60.67**	**17.52**
F3	Linear Regression	46.07	89.99	66.93	20.62
	Random Forest	53.19	83.85	63.84	19.95
	XGBoost	**55.22**	**82.01**	62.35	19.83
	LSTM	51.30	85.51	**61.59**	**18.03**
F4	Linear Regression	45.55	90.43	67.40	20.75
	Random Forest	**52.24**	**84.69**	64.88	20.20
	XGBoost	51.24	85.57	65.91	20.61
	LSTM	49.46	87.12	**61.98**	**17.53**

For the **Small Truck**, XGBoost delivered the highest R^2 values (up to 76.76 in F1), with Random Forest providing stable performance across most settings. However, the **LSTM** consistently achieved the lowest error rates, standing out for both MAE and MAPE. In particular, in configurations F2-F4, its MAPE values dropped below 18%, making it more accurate in relative terms than the tree-based models.

For the **Medium Truck**, the results showed greater variability. Once again, **LSTM** dominated the error-based metrics, consistently obtaining the best MAE and MAPE values across all configurations (e.g., MAPE $\approx$ 17-18% in F1-F4), even though its R^2 scores lagged behind XGBoost and Random Forest. XGBoost achieved the highest R^2 (55.22 in F3), while Random Forest performed reliably in F2 and F4.

In summary, the multi-output setup allowed LSTM to excel in error reduction, making it effective at minimizing relative deviations, while XGBoost often led in explained variance (R^2). As with the independent models, this approach was unable to outperform the recursive forecasting strategy, which remained superior in both stability and predictive accuracy.

5.7 Discussion

The results presented in this chapter highlight the feasibility of predictive modeling of NOx emissions in heavy-duty vehicles through the integration of meteorological data. Across all experiments, the findings confirm that feature enrichment play a decisive role in improving predictive accuracy.

The correlation and feature importance analyses revealed that variables associated with vehicle dynamics, particularly `velocidade_media_kmh_ajustada` (adjusted average speed), are strongly related to NOx emission levels. Furthermore, the spatial visualization of estimated emissions reinforces the link between high-traffic zones and elevated NOx levels, emphasizing the need for localized and context-aware strategies to mitigate environmental impacts.

When comparing feature configurations, it became evident that the inclusion of meteorological variables (F3 and F4) did not consistently enhance predictive performance compared to the baseline configurations (F1 and F2). In several cases, models trained on vehicle-only features (F1, F2) achieved comparable or even superior results. This indicates that while environmental factors such as air temperature, relative humidity, and rainfall may contribute marginal improvements under specific conditions, their predictive value is secondary to core vehicle dynamics.

Across forecasting strategies, the LSTM model consistently provided the best error reduction, particularly in single-step and recursive forecasting scenarios, where it achieved MAPE values below 10% for certain configurations. Its performance was highly sensitive to feature configurations, showing instability when spatial features were introduced in recursive mode.

Linear Regression proved to be a surprisingly competitive baseline, especially for Medium Trucks, where it often outperformed more complex models (Random Forest and XGBoost) in terms of R^2 and stability. This suggests that the optimal model choice depends on both the vehicle type and the temporal dynamics of the data series.

Regarding multi-horizon strategies, the multi-output LSTM approach ($n = 20$) provided the best trade-off between efficiency and accuracy for long-term predictions. Although not always surpassing recursive forecasts in precision, it enabled simultaneous estimation of multiple future steps, making it a practical choice for real-time deployment. In contrast, the independent multi-model strategy ($n = 10$) consistently yielded the weakest results, highlighting the limitations of ignoring temporal dependencies across horizons.

Notably, the minimum MAE value of 31.46 ppm achieved by the LSTM,

though higher than those reported in more controlled studies such as (BANASIEWICZ et al., 2023) (MAE $\approx$ 13.7 ppm for NOx) and (SHAH et al., 2023) (MAE $\approx$ 0.58 for CO_2), reflects the complexity of the real-world, dynamic urban environment studied here. Unlike underground mining operations (BANASIEWICZ et al., 2023) or large-scale passenger car datasets in Europe (SHAH et al., 2023), this work focuses on heavy-duty vehicles, which operate under heterogeneous traffic conditions, varying topography, and frequent acceleration/deceleration cycles. These operational challenges naturally introduce greater variability in emission patterns, explaining the comparatively higher prediction errors observed.

In summary, the findings show that vehicle operation variables remain the primary drivers of predictive performance, while meteorological data can provide marginal benefits in specific scenarios but do not consistently improve accuracy. The robustness of LSTM, particularly in autoregressive and multi-output settings, makes it a strong candidate for deployment in real-world fleet monitoring systems. At the same time, the stability of simpler models such as Linear Regression suggests that hybrid strategies, leveraging both advanced deep learning and interpretable baselines, may be valuable for practical, cost-efficient emission forecasting.

6
Conclusion

This study presented a comprehensive framework for the predictive modeling of NOx emissions in heavy-duty vehicles, based on real-world operational and meteorological data. By testing four feature configurations (F1-F4) and multiple modeling strategies, including single-step, recursive, direct, and multi-output forecasts, systematically assessed the capacity of different machine learning models to capture the temporal and spatial dynamics of vehicular emissions.

The results highlight that vehicle operation variables, especially `velocidade_media_kmh_ajustada` (adjusted average speed), remain the dominant predictors of NOx emissions, consistently aligning with prior research (SILVA et al., 2025). The LSTM model proved to be the most effective in reducing error metrics, frequently achieving the lowest MAE and MAPE across forecasting modes, and reaching values below 10% in the most favorable configurations. However, LSTM also exhibited high sensitivity to feature selection, with performance deteriorating sharply when spatial variables were included in recursive mode. This underlines its dual nature: capable of superior accuracy under optimized conditions, but unstable under less favorable configurations.

Regarding feature configurations, the inclusion of meteorological variables (F3, F4) did not consistently improve predictive performance compared to the baseline (F1, F2). While in some cases they provided marginal benefits, in others they introduced instability. This demonstrates that meteorological information, though contextually important, has limited predictive power compared to direct vehicle operation data.

Among the long-horizon forecasting techniques, the one-shot multi-output approach emerged as the most promising, offering a favorable balance between performance and extended forecasting capability ($n = 20$). Furthermore, it is noteworthy that, even with a limited set of variables directly obtained from sensors, the models achieved competitive results. The combination of spatial interpolation techniques (kriging) for meteorological variables and derived calculations, such as estimated speed and acceleration, proved sufficient to build effective representations of vehicular emission behavior.

A key contribution of this work is its application to heavy-duty fleets in continuous operation in Brazil, a context characterized by heterogeneous traffic conditions, variable topography, and diverse climatic patterns. Unlike previous

studies conducted in controlled environments or with passenger car datasets, this case study demonstrates the feasibility of deploying advanced forecasting methods in highly dynamic and operationally challenging settings. This novelty strengthens both the scientific and practical contributions of the framework: advancing methodological research while directly supporting transport companies and environmental policymakers in emission-aware decision making.

When compared to previous studies, such as (BANASIEWICZ et al., 2023) and (SHAH et al., 2023), the results obtained here were competitive, reinforcing the robustness of the applied methods even in the context of continuous real-world operation with commercial vehicles.

Limitations and Future Work

Despite promising results, several limitations should be acknowledged. First, the emission sensor provides latitude and longitude, enabling the calculation of variables like speed and acceleration. However, the data lacked engine-specific measurements (e.g., temperature, torque, fuel flow), which could enrich model inputs. Second, meteorological variables were derived from interpolations of fixed monitoring stations, often located far from vehicle routes and at lower temporal frequencies, potentially obscuring localized weather effects. Lastly, the NOx sensor itself records emissions at one-minute intervals, which limits the ability to capture transient peaks and short-term fluctuations in driving dynamics and emissions.

In light of these limitations, the following directions are proposed for future work:

- Incorporating new onboard sensors to capture engine-specific and localized climatic data (e.g., internal temperature, pressure, wind direction/speed), thereby improving data quality and coverage;
- Higher-frequency emission sensing, at sub-minute or per-second intervals, to better reflect short-term variations in vehicle operation and emissions;
- Evaluating the use of Transformer-based architectures, which offer greater capacity for modeling complex temporal dependencies and multi-scale inputs;
- Extending the study to other vehicle classes such as tractor-trailers or urban buses, increasing practical applicability;
- Testing the model across different regions and vehicle types to assess its generalization capabilities.

This work contributes to the literature by combining real-world vehicle data, deep learning methods, and environmental features in a unified framework. While the study highlights the limitations of meteorological inputs and the instability of complex models like LSTM under certain conditions, it also demonstrates the feasibility of achieving high predictive accuracy in realistic urban contexts. Offering a promising foundation for the development of intelligent emission management systems in the transportation sector.

7
Bibliography

AGENCY, E. E. **Air quality in Europe-2022 report**. [S.l.], 2022. Disponível em: https://www.eea.europa.eu/publications/air-quality-in-europe-2022.

AGENCY, U. E. P. **Policy Assessment for the Review of the Ozone National Ambient Air Quality Standards**. [S.l.: s.n.], 2014.

BANASIEWICZ, A. et al. Forecasting of nox emissions of diesel lhd vehicles in underground mines-an ann-based regression approach. **Applied Sciences**, v. 13, n. 17, 2023. ISSN 2076-3417. Disponível em: https://www.mdpi.com/2076-3417/13/17/9965.

BENGIO, Y.; COURVILLE, A.; VINCENT, P. Representation learning: A review and new perspectives. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 35, n. 8, p. 1798–1828, 2013.

BEZERRA, E. Introdução à aprendizagem profunda. In: _____. [S.l.: s.n.], 2016. p. 57–86.

BRUNEKREEF, B.; HOLGATE, S. T. Air pollution and health. **The Lancet**, Elsevier, v. 360, n. 9341, p. 1233–1242, 2002.

CHAVES, M. G. D. et al. Particulate matter forecast and prediction in curitiba using machine learning. **Frontiers in Big Data**, Volume 7 - 2024, 2024. ISSN 2624-909X. Disponível em: https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2024.1412837.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: Association for Computing Machinery, 2016. (KDD '16), p. 785–794. ISBN 9781450342322. Disponível em: https://doi.org/10.1145/2939672.2939785.

CHEVILLON, G. Direct multi-step estimation and forecasting. **Journal of Economic Surveys**, v. 21, n. 4, p. 746–785, 2007. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-6419.2007.00518.x.

Conselho Nacional do Meio Ambiente. **Resolução CONAMA nž 490, de 16 de novembro de 2018**. 2018. https://www.in.gov.br/materia/-/asset_publisher/Kujrw0TZC2Mb/content/id/51829611. Estabelece limites máximos de emissão de poluentes para veículos automotores pesados novos.

Continental Aftermarket. **UniNOx Sensor**. 2025. https://www.continental-aftermarket.com/en-en/products/small-production-runs-for-special-purpose-vehicles/sensors-switches/uninox-sensor. Accessed: 13 Aug. 2025.

CRESSIE, N. A. C. **Statistics for Spatial Data**. Revised. New York: Wiley, 1993. (Wiley Series in Probability and Mathematical Statistics). ISBN 0-471-00255-0.

FACELI, K. et al. **Inteligência artificial: uma abordagem de aprendizado de máquina**. [S.l.]: LTC, 2021.

FENG, S. et al. A macro-micro spatio-temporal neural network for traffic prediction. **Transportation Research Part C: Emerging Technologies**, v. 156, p. 104331, 2023. ISSN 0968-090X. Disponível em: https://www.sciencedirect.com/science/article/pii/S0968090X23003200.

FERNáNDEZ-AVILéS, G.; MATTERA, R.; SCEPI, G. Factor-augmented autoregressive neural network to forecast nox in the city of madrid. **Socio-Economic Planning Sciences**, v. 95, p. 101958, 2024. ISSN 0038-0121. Disponível em: https://www.sciencedirect.com/science/article/pii/S0038012124001575.

FERREIRA, F. **Prevendo números: Entendendo métricas de regressão**. 2020. Acessado em 3 de julho de 2025. Disponível em: https://medium.com/data-hackers/prevendo-n%C3%BAmeros-entendendo-m%C3%A9tricas-de-regress%C3%A3o-35545e011e70.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [S.l.]: MIT Press, 2016. http://www.deeplearningbook.org.

GREFF, K. et al. Lstm: A search space odyssey. **IEEE Transactions on Neural Networks and Learning Systems**, v. 28, n. 10, p. 2222–2232, 2017.

HAN, J.; KAMBER, M.; PEI, J. 3 - data preprocessing. In: HAN, J.; KAMBER, M.; PEI, J. (Ed.). **Data Mining: Concepts and Techniques (Third Edition)**. Third edition. Boston: Morgan Kaufmann, 2012, (The Morgan Kaufmann Series in Data Management Systems). p. 83–124. ISBN 978-0-12-381479-1. Disponível em: https://www.sciencedirect.com/science/article/pii/B9780123814791000034.

HE, T.-L. et al. Deep learning to evaluate us nox emissions using surface ozone predictions. **Journal of Geophysical Research: Atmospheres**, v. 127, n. 4, p. e2021JD035597, 2022. E2021JD035597 2021JD035597. Disponível em: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021JD035597.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural Computation**, v. 9, n. 8, p. 1735–1780, Nov 1997. ISSN 0899-7667.

III, C. A. P.; DOCKERY, D. W. Health effects of fine particulate air pollution: Lines that connect. **Journal of the Air & Waste Management Association**, Taylor & Francis, v. 56, n. 6, p. 709–742, 2006. Disponível em: https://doi.org/10.1080/10473289.2006.10464485.

JAMES, G. et al. **An Introduction to Statistical Learning: with Applications in R**. 2nd. ed. Springer, 2023. ISBN 978-1-0716-1418-1. Disponível em: https://www.statlearning.com/.

JIANG, X.-Q.; MEI, X.-D.; FENG, D. Air pollution and chronic airway diseases: what should people know and do? **Journal of Thoracic Disease**, v. 8, n. 1, 2015. ISSN 2077-6624. Disponível em: https://jtd.amegroups.org/article/view/5951.

KAYES, I.; AL. et. The relationships between meteorological conditions and air quality in urban areas: A case study in dhaka, bangladesh. **Atmospheric Research**, Elsevier, v. 218, p. 190–202, 2019.

KRIGE, D. **A Statistical Approach to Some Mine Valuations and Allied Problems at the Witwatersrand**. Tese (Doutorado), 01 1951.

KUHN, M.; JOHNSON, K. **Applied Predictive Modeling**. Springer, 2013. ISBN 978-1-4614-6849-3. Disponível em: https://doi.org/10.1007/978-1-4614-6849-3.

KUMARI, S.; SINGH, S. K. Machine learning-based time series models for effective co_2 emission prediction in india. **Environmental Science and Pollution Research**, v. 30, n. 55, p. 116601–116616, November 2023. ISSN 1614-7499. Disponível em: https://doi.org/10.1007/s11356-022-21723-8.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, v. 521, p. 436–44, 05 2015.

LELIEVELD, J. et al. The contribution of outdoor air pollution sources to premature mortality on a global scale. **Nature**, Nature Publishing Group, v. 525, n. 7569, p. 367–371, 2015.

LI, Z.; ZHANG, X.; DONG, Z. Tsf-transformer: a time series forecasting model for exhaust gas emission using transformer. **Applied Intelligence**, Springer, v. 53, n. 13, p. 17211–17225, 2023. ISSN 1573-7497. Disponível em: https://doi.org/10.1007/s10489-022-04326-1.

LINDEMANN, B. et al. A survey on long short-term memory networks for time series prediction. **Procedia CIRP**, v. 99, p. 650–655, 2021. ISSN 2212-8271. 14th CIRP Conference on Intelligent Computation in Manufacturing Engineering, 15-17 July 2020. Disponível em: https://www.sciencedirect.com/science/article/pii/S2212827121003796.

MA, T. et al. Air pollution characteristics and their relationship with emissions and meteorology in the yangtze river delta region during 2014-2016. **Journal of Environmental Sciences**, v. 83, p. 8–20, 2019. ISSN 1001-0742. Disponível em: https://www.sciencedirect.com/science/article/pii/S1001074218329954.

MATHERON, G. **Traité de géostatistique appliquée**. [S.l.]: Éditions Technip, 1962. (Memoires, v. 1).

ORGANIZATION, W. H. **WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide**. World Health Organization, 2021. Disponível em: https://www.who.int/publications/i/item/9789240034228.

PANDEY, Y. et al. Machine learning in the oil and gas industry, including geosciences, reservoir engineering, and production engineering with python. 01 2020.

PARK, J. J. et al. Development of a light and accurate nox prediction model for diesel engines using machine learning and xai methods. **International Journal of**

Automotive Technology, Springer, v. 24, n. 2, p. 559–571, 2023. ISSN 1976-3832. Disponível em: https://doi.org/10.1007/s12239-023-0047-0.

QING, L.; HONGLI, H. Prediction model of the nox emissions based on long short-time memory neural network. In: **2020 Chinese Automation Congress (CAC)**. [S.l.: s.n.], 2020. p. 2795–2797.

RAVINDRA, K. et al. Real-time monitoring of air pollutants in seven cities of north india during crop residue burning and their relationship with meteorology and transboundary movement of air. **Science of The Total Environment**, v. 690, p. 717–729, 2019. ISSN 0048-9697. Disponível em: https://www.sciencedirect.com/science/article/pii/S0048969719327895.

RUSSELL, S.; NORVIG, P.; DAVIS, E. **Artificial Intelligence: A Modern Approach**. Prentice Hall, 2010. (Prentice Hall series in artificial intelligence). ISBN 9780136042594. Disponível em: https://books.google.com.br/books?id=8jZBksh-bUMC.

SHAH, S. et al. A comparative study ofămachine learning andădeep learning techniques forăprediction of$\check{a}co_2$emission inăcars. In: SHAKYA, S.; BALAS, V. E.; HAOXIANG, W. (Ed.). **Proceedings of Third International Conference on Sustainable Expert Systems**. Singapore: Springer Nature Singapore, 2023. p. 749–758. ISBN 978-981-19-7874-6.

SILVA, A. A. et al. Petrofacies classification using machine learning algorithms. **GEOPHYSICS**, v. 85, n. 4, p. WA101–WA113, 2020. Disponível em: https://doi.org/10.1190/geo2019-0439.1.

SILVA, L.; CRUZ, A.; NUNES, A. Real-time monitoring of nitrogen oxides emission factors using sensors in the exhaust pipes of heavy vehicles in the metropolitan region of rio de janeiro. **Journal of the Brazilian Chemical Society**, v. 35, 02 2024.

SILVA, L. F. M. da et al. Real-time analysis of nox emissions in heavy-duty diesel vehicles: Impact of speed and variations across vehicle groups. **Atmospheric Pollution Research**, v. 16, n. 9, p. 102572, 2025. ISSN 1309-1042. Disponível em: https://www.sciencedirect.com/science/article/pii/S1309104225001746.

Standard Brand. **Diesel Nitrogen Oxide (NOx) Sensors**. 2025. Accessed: 12 Aug. 2025. Disponível em: https://www.napaechlin.com/en/products/diesel/diesel-components/diesel-nitrogen-oxide-nox-sensors.

SUN, H. et al. Development of an lstm-ccf-ma model for predicting nox emission and exhaust temperature of a diesel engine. **International Journal of Automotive Technology**, v. 26, n. 2, p. 437–450, 2025. ISSN 1976-3832. Disponível em: https://doi.org/10.1007/s12239-024-00152-1.

TAMAL, M. et al. Forecasting of solar photovoltaic output energy using lstm machine learning algorithm. In: . [S.l.: s.n.], 2022. p. 1–6.

THALLER, E. I. et al. Moderate increases in ambient pm2.5 and ozone are associated with lung function decreases in beach lifeguards. **Journal of Occupational and Environmental Medicine**, v. 50, n. 2, 2008. ISSN 1076-2752.

TOSCANO, D.; MURENA, F. Atmospheric ship emissions in ports: A review. correlation with data of ship traffic. **Atmospheric Environment: X**, v. 4, p. 100050, 2019. ISSN 2590-1621. Disponível em: https://www.sciencedirect.com/science/article/pii/S259016211930053X.

VISHNU, G. et al. Short-term forecasting of electric vehicle load using time series, machine learning, and deep learning techniques. **World Electric Vehicle Journal**, v. 14, n. 9, 2023. ISSN 2032-6653. Disponível em: https://www.mdpi.com/2032-6653/14/9/266.

WANG, W.; CHAKRABORTY, G.; CHAKRABORTY, B. Predicting the risk of chronic kidney disease (ckd) using machine learning algorithm. **Applied Sciences**, v. 11, p. 202, 12 2020.

WANG, Y.; YU, Y.; LI, J. Predicting the transient nox emissions of the diesel vehicle based on lstm neural networks. In: **2020 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS)**. [S.l.: s.n.], 2020. p. 261–264.

WANG, Y. et al. Real-time joint traffic state and model parameter estimation on freeways with fixed sensors and connected vehicles: State-of-the-art overview, methods, and case studies. **Transportation Research Part C: Emerging Technologies**, v. 134, p. 103444, 2022. ISSN 0968-090X. Disponível em: https://www.sciencedirect.com/science/article/pii/S0968090X21004332.

WEN, H.-T.; LU, J.-H.; JHANG, D.-S. Features importance analysis of diesel vehicles nox and co2 emission predictions in real road driving based on gradient boosting regression model. **International Journal of Environmental Research and Public Health**, v. 18, n. 24, 2021. ISSN 1660-4601. Disponível em: https://www.mdpi.com/1660-4601/18/24/13044.

YU, J. et al. Deep learning models for pv power forecasting: Review. **Energies**, v. 17, n. 16, 2024. ISSN 1996-1073. Disponível em: https://www.mdpi.com/1996-1073/17/16/3973.

ZENG, W.; MIWA, T.; MORIKAWA, T. Prediction of vehicle co2 emission and its application to eco-routing navigation. **Transportation Research Part C: Emerging Technologies**, v. 68, p. 194–214, 2016. ISSN 0968-090X. Disponível em: https://www.sciencedirect.com/science/article/pii/S0968090X16300110.

ZHANG, Z. et al. Probe data-driven travel time forecasting for urban expressways by matching similar spatiotemporal traffic patterns. **Transportation Research Part C: Emerging Technologies**, v. 85, p. 476–493, 2017. ISSN 0968-090X. Disponível em: https://www.sciencedirect.com/science/article/pii/S0968090X17302826.

A
Forecasting Results for Small Trucks

A.1
F1

Recursive Forecasting

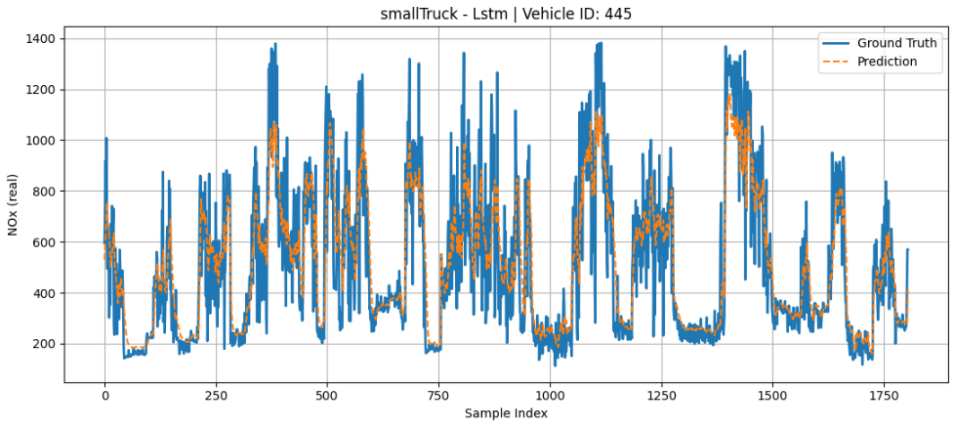

Figure A.1: Recursive forecast for Small Truck using LSTM - Example from vehicle 445 (F1)

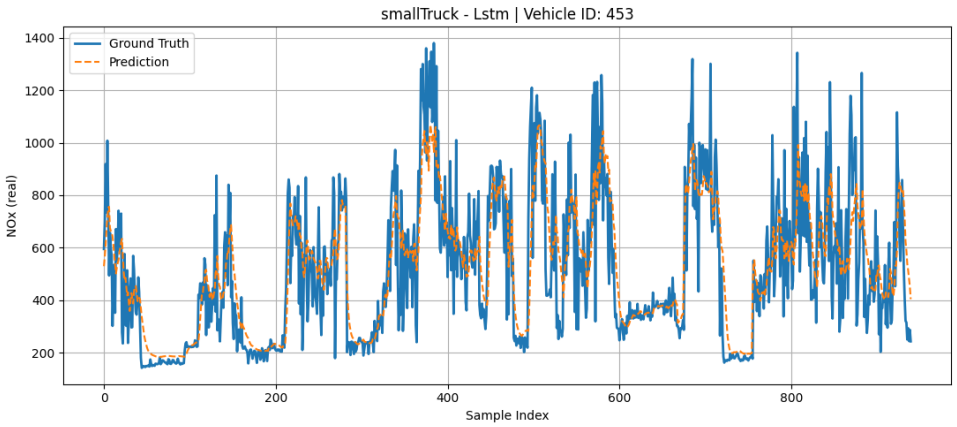

Figure A.2: Recursive forecast for Small Truck using LSTM - Example from vehicle 453 (F1)

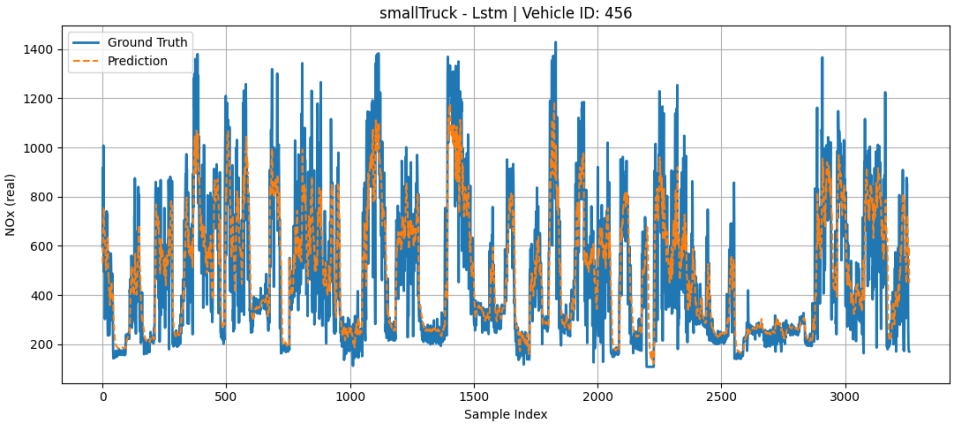

Figure A.3: Recursive forecast for Small Truck using LSTM - Example from vehicle 456 (F1)

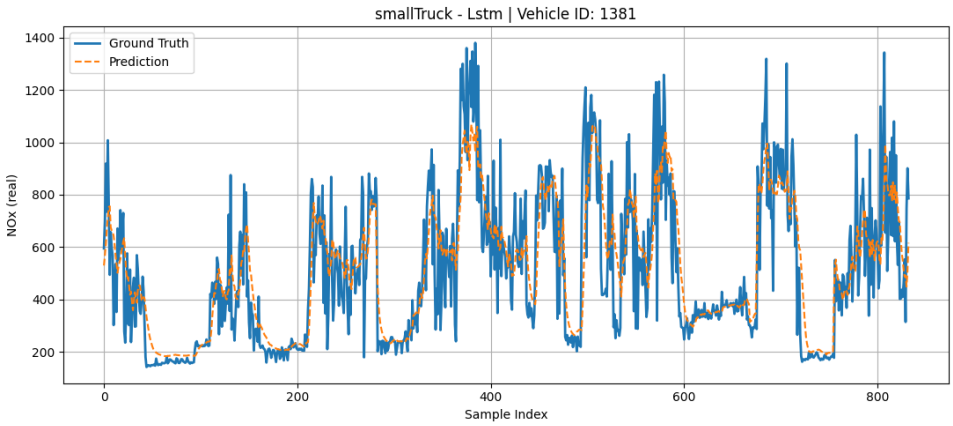

Figure A.4: Recursive forecast for Small Truck using LSTM - Example from vehicle 1381 (F1)

A.2
F3

Recursive Forecasting

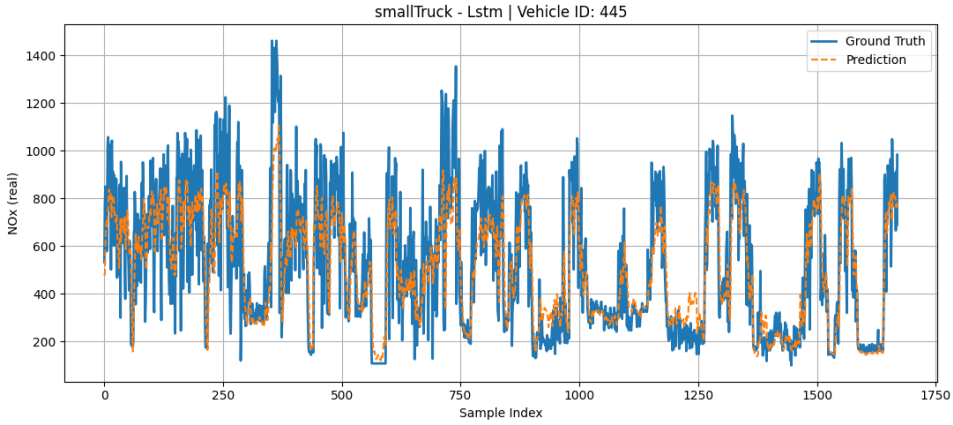

Figure A.5: Recursive forecast for Small Truck using LSTM - Example from vehicle 445 (F3)

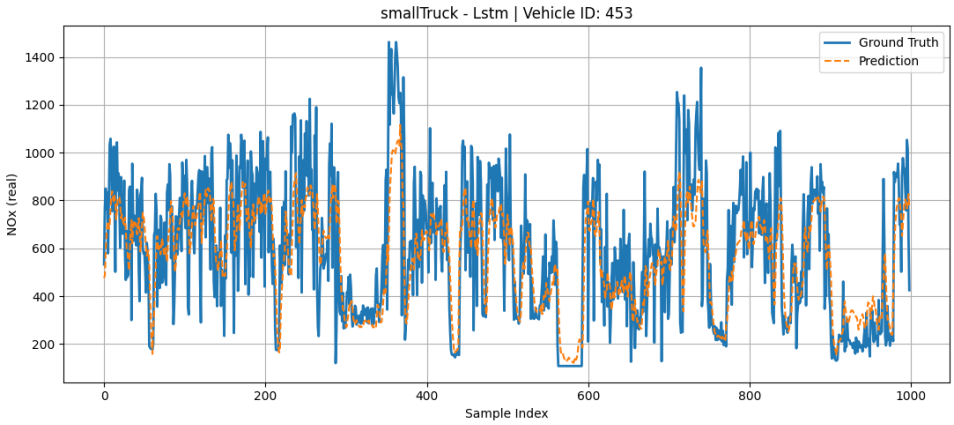

Figure A.6: Recursive forecast for Small Truck using LSTM - Example from vehicle 453 (F3)

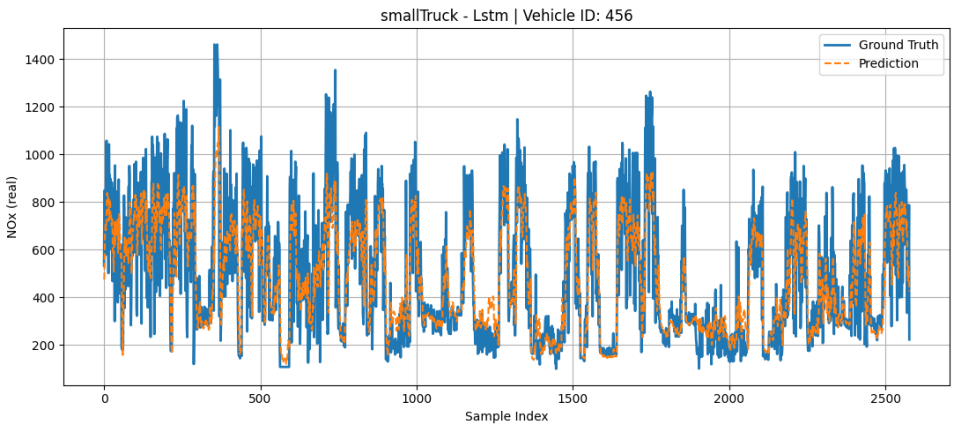

Figure A.7: Recursive forecast for Small Truck using LSTM - Example from vehicle 456 (F3)

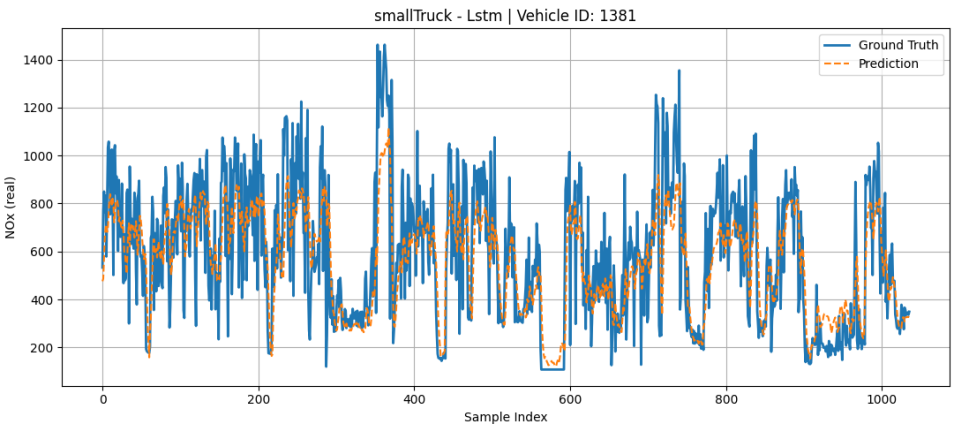

Figure A.8: Recursive forecast for Small Truck using LSTM - Example from vehicle 1381 (F3)

Direct Forecasting (n=10)

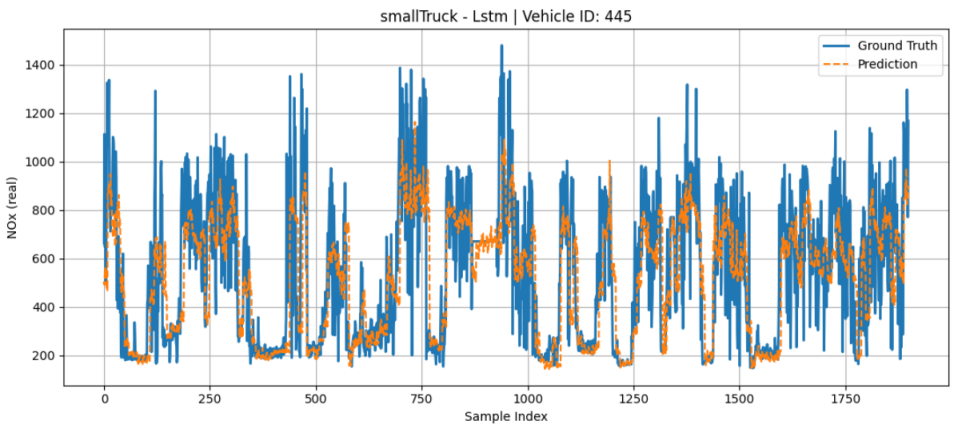

Figure A.9: Direct Forecasting for Small Truck using LSTM - Example from vehicle 445 (F3)

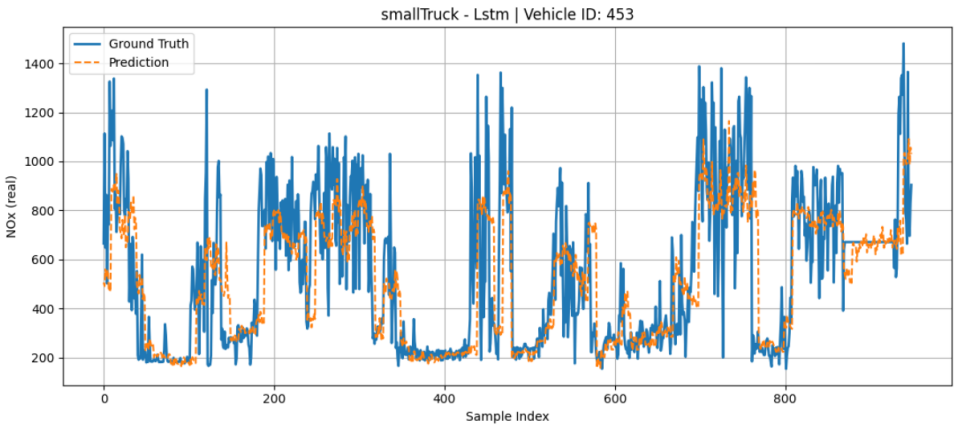

Figure A.10: Direct Forecasting for Small Truck using LSTM - Example from vehicle 453 (F3)

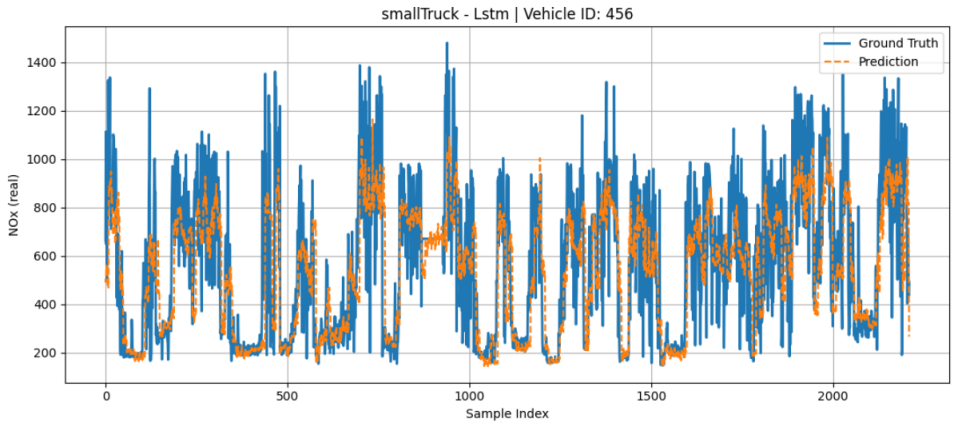

Figure A.11: Direct Forecasting for Small Truck using LSTM - Example from vehicle 456 (F3)

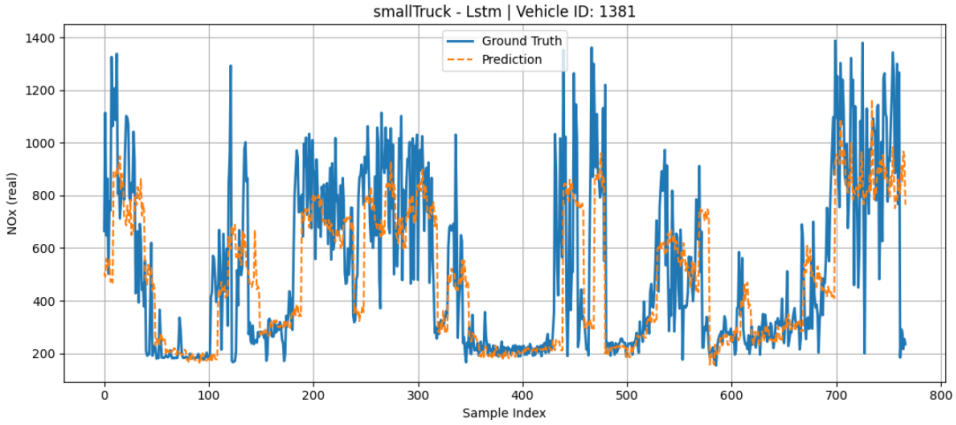

Figure A.12: Direct Forecasting for Small Truck using LSTM - Example from vehicle 1381 (F3)

Multiple Outputs (n=20)

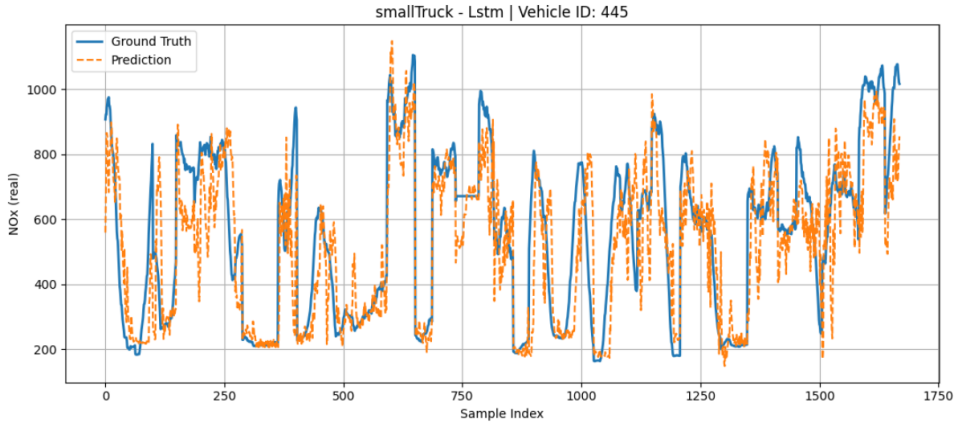

Figure A.13: Multiple outputs forecast for Small Truck using LSTM - Example from vehicle 445 (F3)

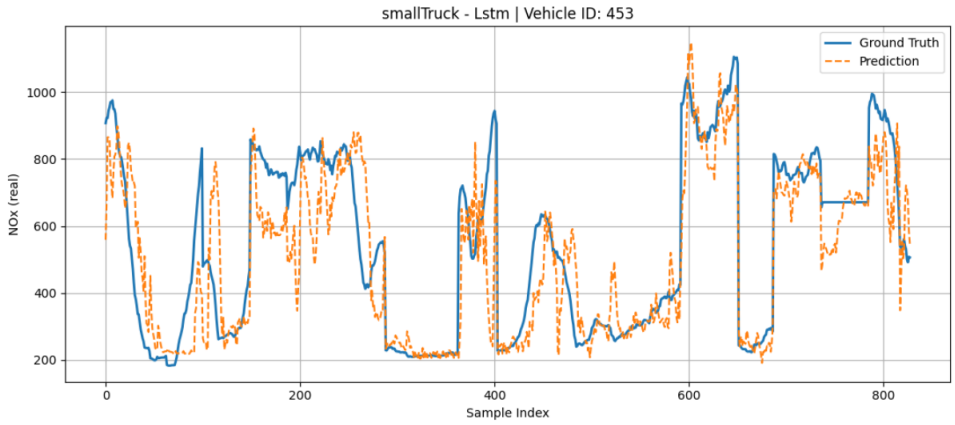

Figure A.14: Multiple outputs forecast for Small Truck using LSTM - Example from vehicle 453 (F3)

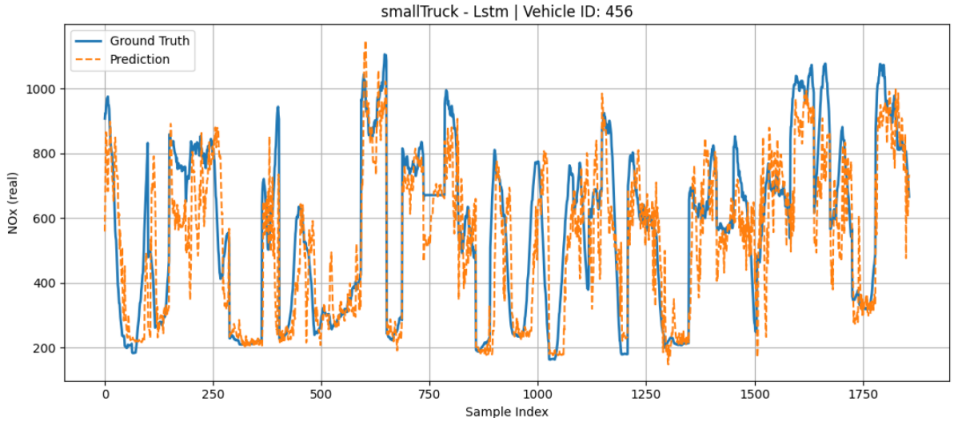

Figure A.15: Multiple outputs forecast for Small Truck using LSTM - Example from vehicle 456 (F3)

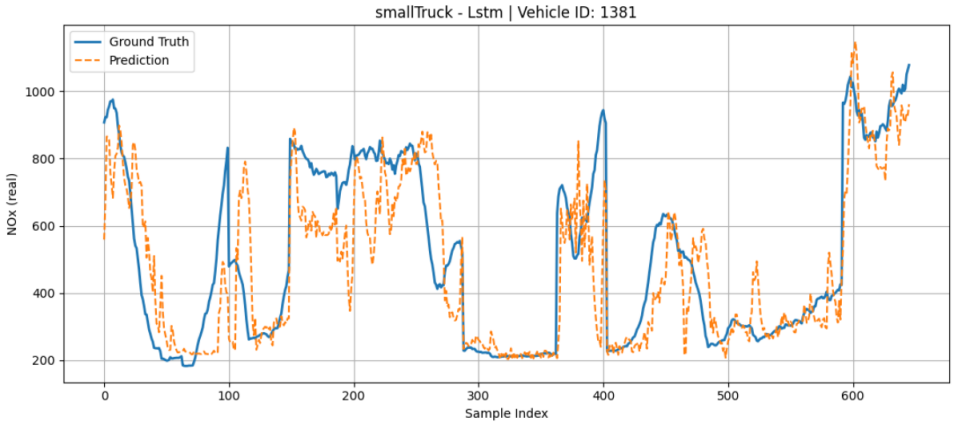

Figure A.16: Multiple outputs forecast for Small Truck using LSTM - Example from vehicle 1381 (F3)

A.3
F4

Recursive Forecasting

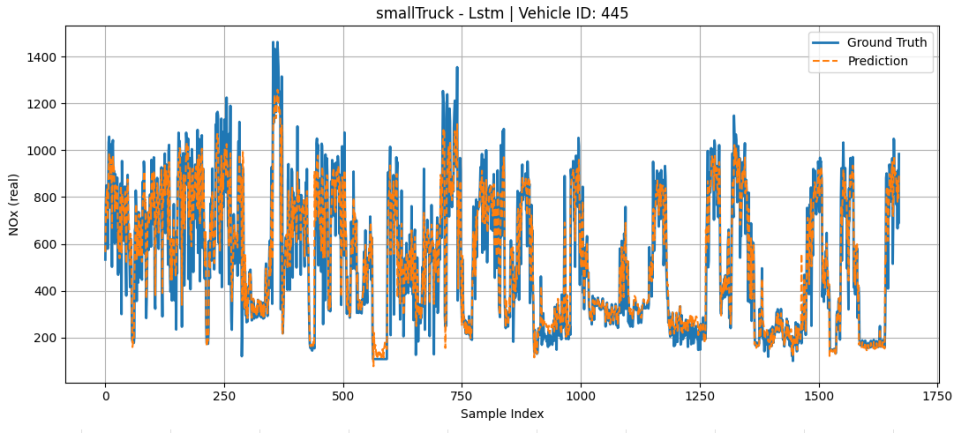

Figure A.17: Recursive forecast for Small Truck using LSTM - Example from vehicle 445 (F4)

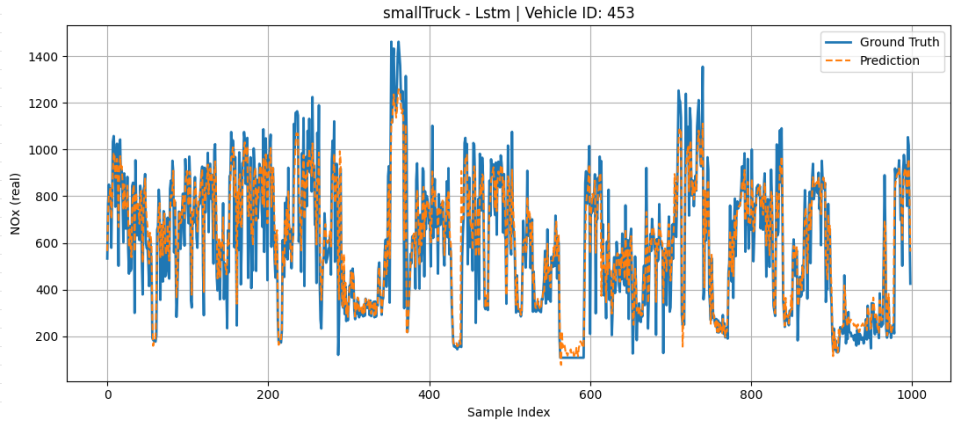

Figure A.18: Recursive forecast for Small Truck using LSTM - Example from vehicle 453 (F4)

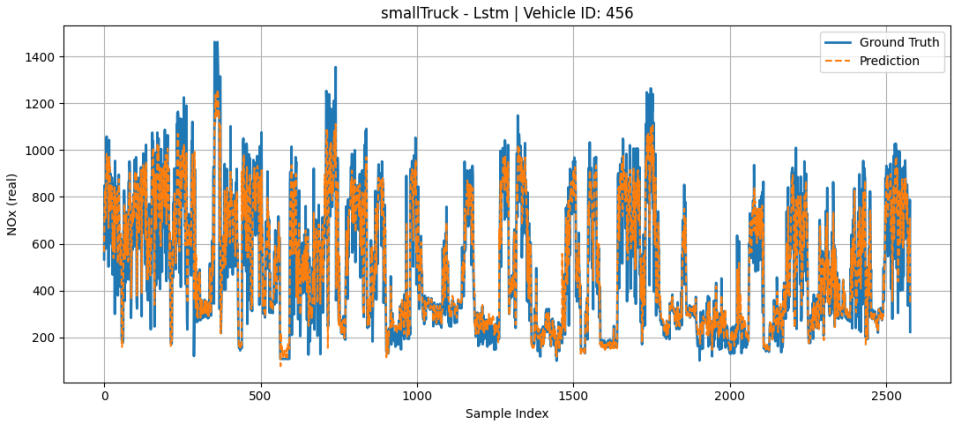

Figure A.19: Recursive forecast for Small Truck using LSTM - Example from vehicle 456 (F4)

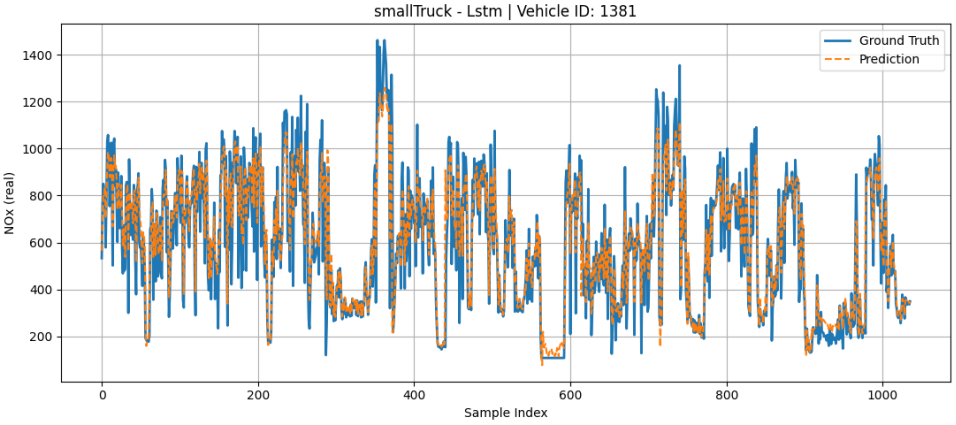

Figure A.20: Recursive forecast for Small Truck using LSTM - Example from vehicle 1381 (F4)

Multiple Outputs (n=20)

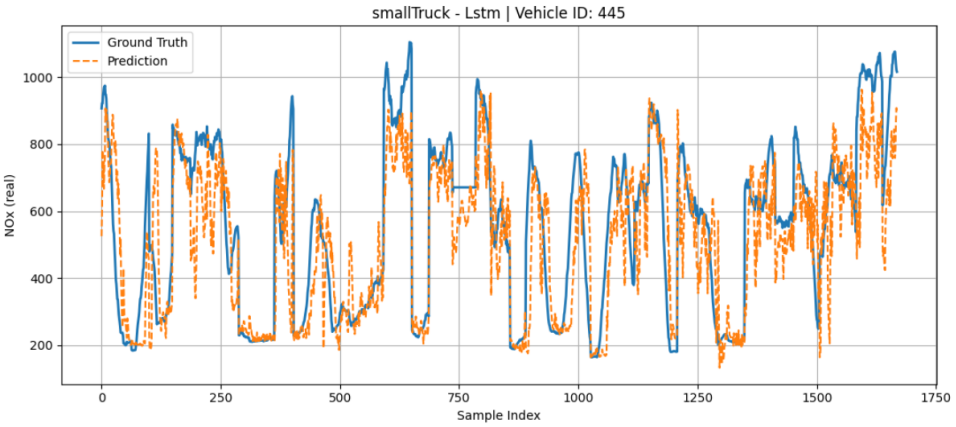

Figure A.21: Multiple outputs forecast for Small Truck using LSTM - Example from vehicle 445 (F4)

B
Forecasting Results for Medium Trucks

B.1
F1

Direct Forecastinging (n=10)

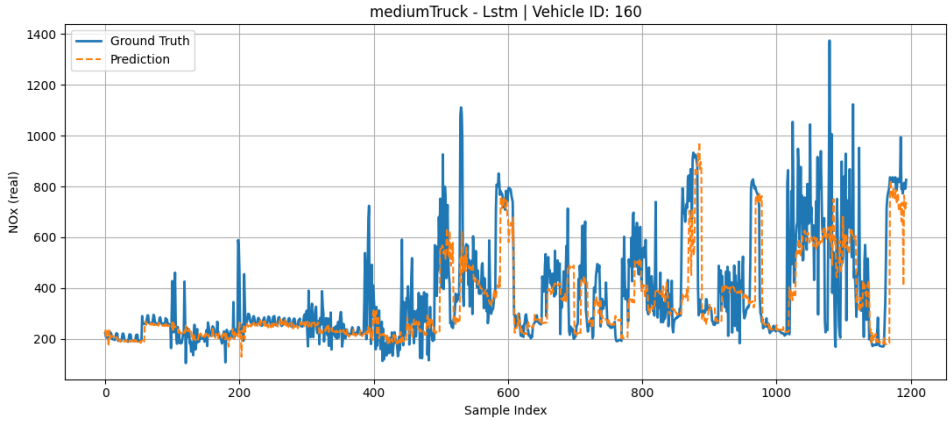

Figure B.1: Direct Forecasting for Medium Truck using LSTM - Example from vehicle 160 (F1)

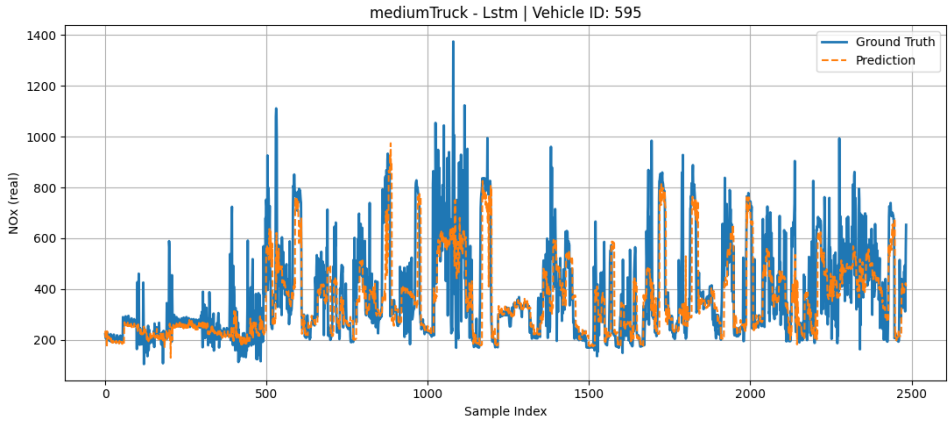

Figure B.2: Direct Forecasting for Medium Truck using LSTM - Example from vehicle 595 (F1)

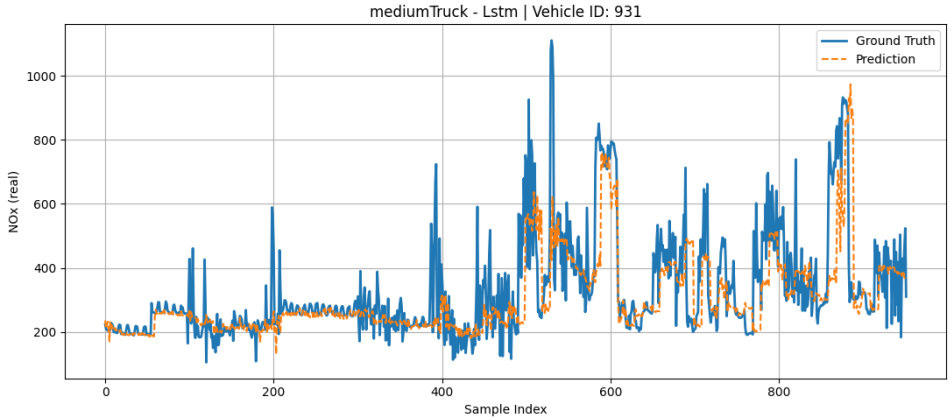

Figure B.3: Direct Forecasting for Medium Truck using LSTM - Example from vehicle 931 (F1)

Multiple Outputs (n=20)

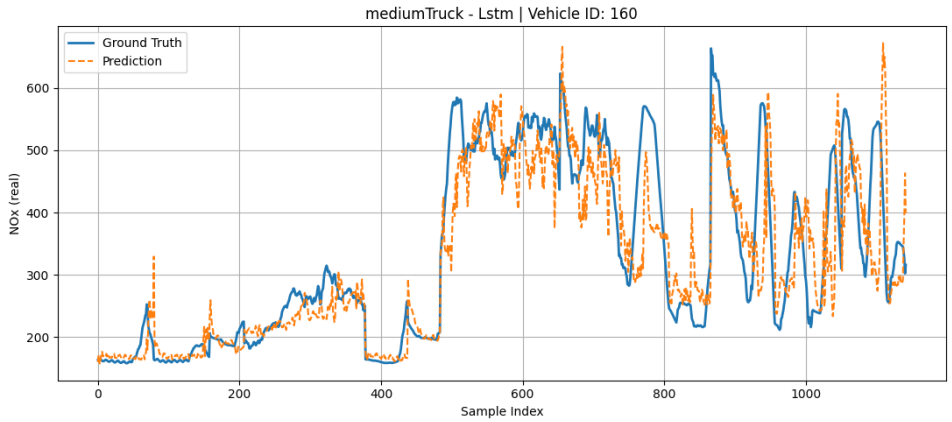

Figure B.4: Multiple outputs forecast for Medium Truck using LSTM - Example from vehicle 160 (F1)

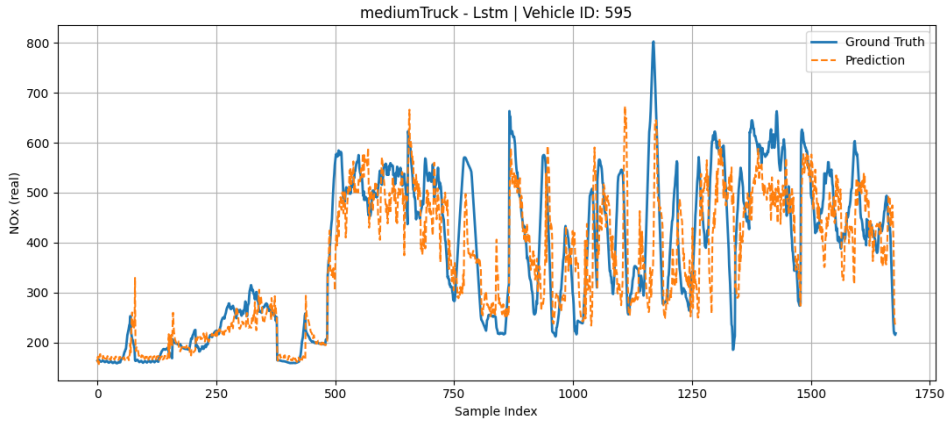

Figure B.5: Multiple outputs forecast for Medium Truck using LSTM - Example from vehicle 595 (F1)

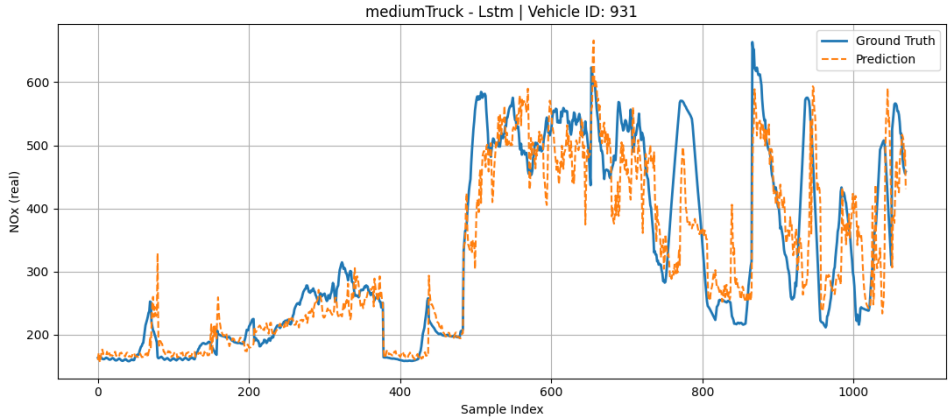

Figure B.6: Multiple outputs forecast for Medium Truck using LSTM - Example from vehicle 931 (F1)

B.2
F3

Direct Forecastinging (n=10)

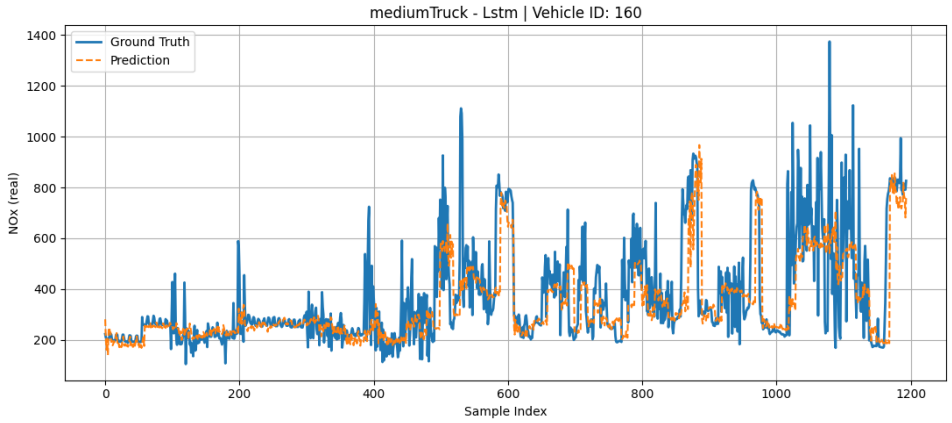

Figure B.7: Direct Forecasting for Medium Truck using LSTM - Example from vehicle 160 (F3)

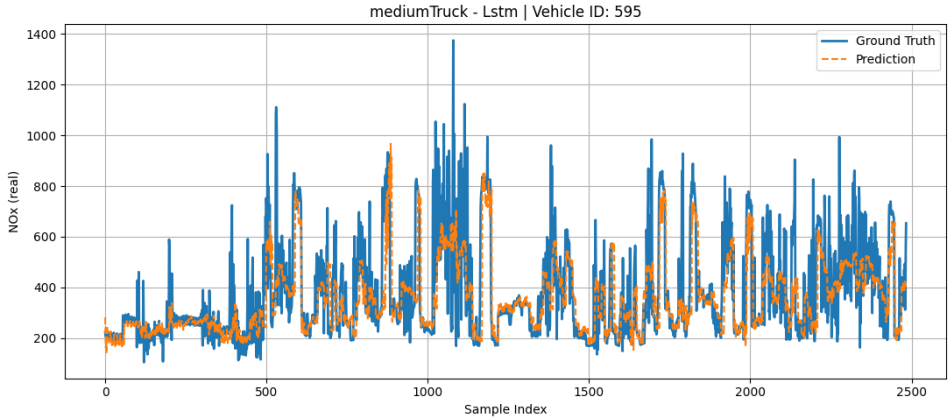

Figure B.8: Direct Forecasting for Medium Truck using LSTM - Example from vehicle 595 (F3)

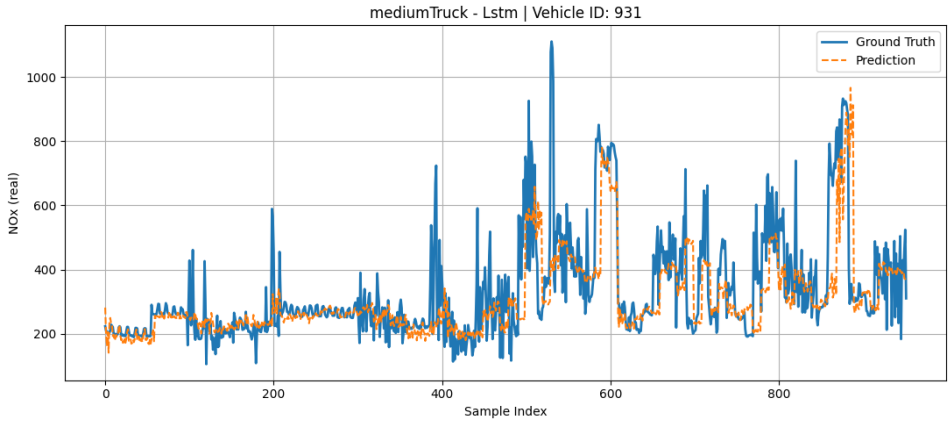

Figure B.9: Direct Forecasting for Medium Truck using LSTM - Example from vehicle 931 (F3)