Automobile Engine Fault Diagnosis Using Machine Learning Method

21900416 AN GYEON HEAL
21900727 JIN GA RAM

School of Mechanical & Control Engineering

10.01.2024

Industrial AI & Automation
Prof. Young-Keun Kim

CONTENTS

Engine Fault Diagnosis

BACKGROUND

Engine Fault Diagnosis

Necessity

- Vehicle engine malfunctions increase emissions of harmful pollutants like CO, HC, and CO2, contributing to environmental problems.
- Engine malfunctions lead to high repair costs, unexpected downtime, and potential financial and safety risks.
- Engines may experience mechanical or electronic failures (e.g., sensor issues, pressure problems, injector defects), complicating the repair process.

Technical Challenges

- Traditional Engine Fault Diagnosis Methods Rely on Manual Inspections and Specialized Tools (e.g., fuel pressure measurements, OBD-2 scanners).
- It’s Labor-Intensive, Expensive, and Require Expert Knowledge, It’s Vulnerable to Human Error.

Solution

- By Applying Machine Learning, Variables from Engine Testing Experiments can be used for Classification to Identify Types of Engine Faults.
- Classification will Help Quickly Diagnose and Resolve Engine Faults, Enabling Fast and Efficient Repairs.

PROBLEM STATEMENT

Engine Fault Diagnosis

Objective

By **Developing** a Model that Outperforms those Implemented in Existing Journals,

Accurately **Identify** types of Engine faults, Providing Greater **Speed** and **Precision** in Engine Repairs.

Detailed Objectives

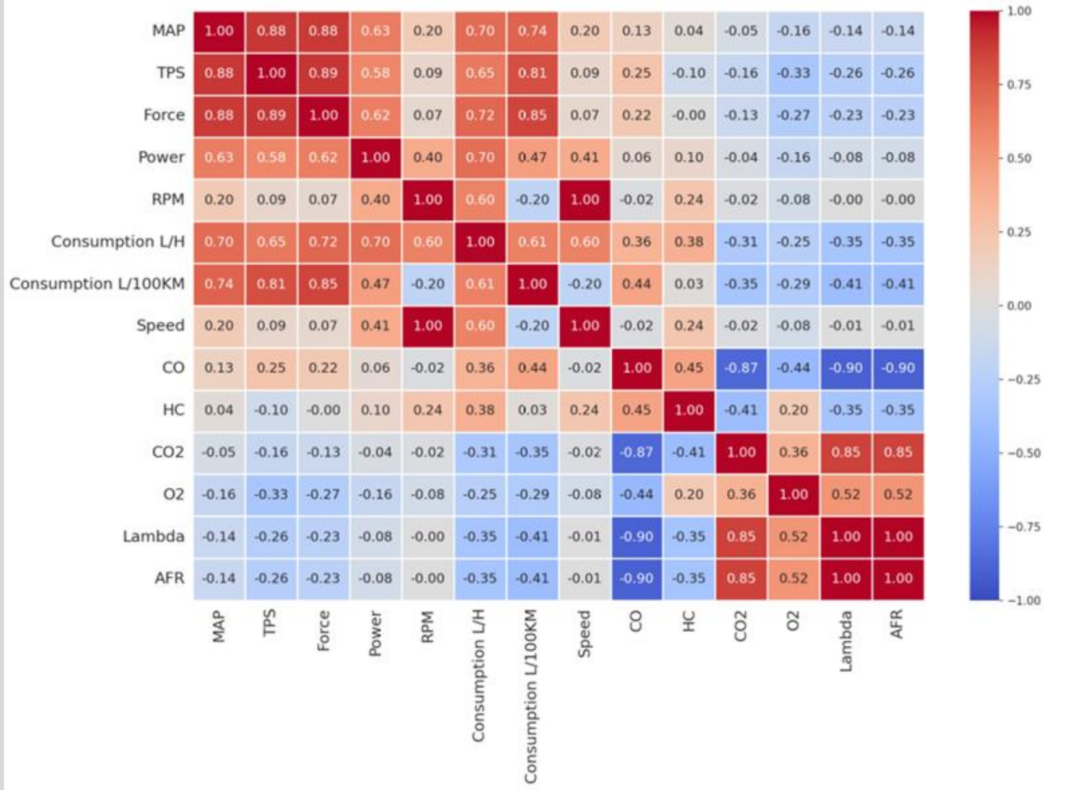
1. Classify Datasets by Machine Learning Model Using "EngineFaultDB" Datasets.
2. Understand and Analysis Datasets through Feature Analysis.
3. Improve Model Performance through Feature Extraction and Feature Reduction/Selection
4. Compare the Model Performance with Baseline Journal.

BASELINE SURVEY-JOURNAL

Engine Fault Diagnosis

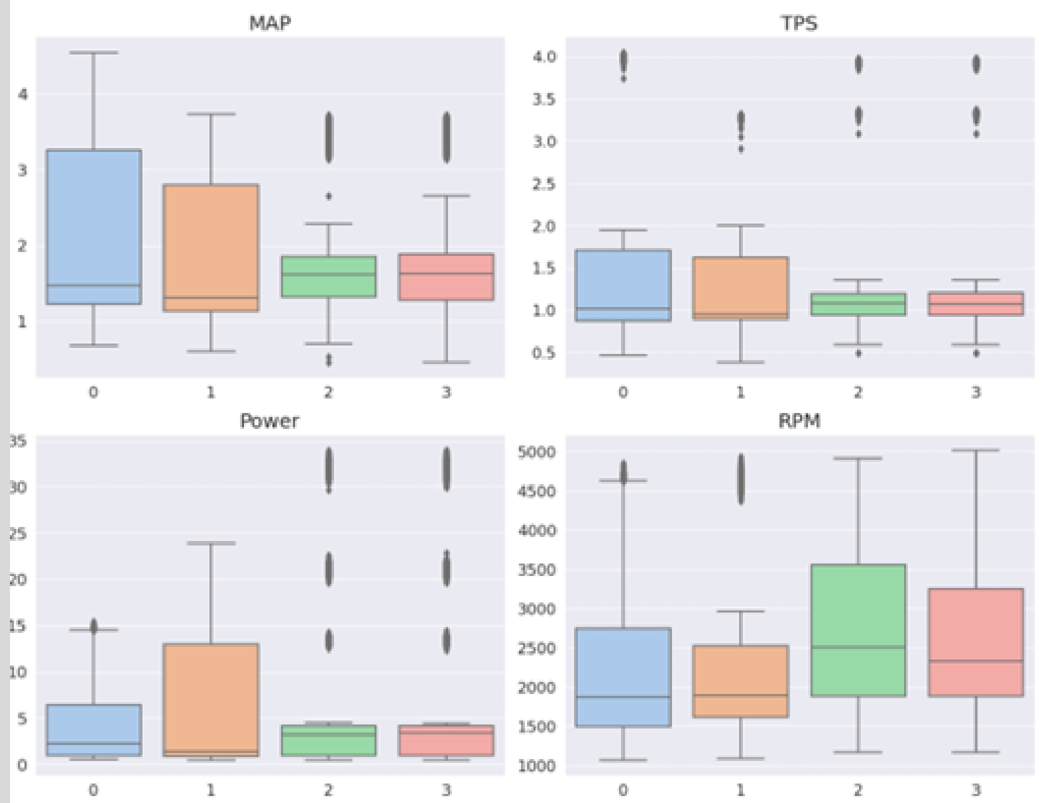
Journal (IEEE): LINK

1. Correlation Analysis



- **Correlation :**
  How variables influence each other
- **Multicollinearity :**
  Reducing highly correlated variables

2. Box Plot Analysis



- **Data Distribution Summary**
- **Identifying Outlier**
- **Skewness Check**
- **Range and Variability Check**

3. Data Preprocessing & Classification

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

- Each Datasets has a different range → **Scaling is necessary.**
- **Min-Max Scaler**

- Logistic Regression
- Decision Tree
- Random Forest
- SVM
- KNN
- Naive Bayes
- Feed-Forward Neural Network

4. Performance Confusion Metrics

- **Accuracy**

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

- **Precision**

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- **Recall**

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- **F1-Score**

$$\text{F1} - \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Classifier	Accuracy	Precision	Recall	F1-score
LR	0.576	0.574	0.576	0.574
DT	0.750	0.750	0.750	0.750
RF	0.748	0.748	0.748	0.748
SVC	0.747	**0.768**	0.747	0.715
KNN	**0.751**	0.751	**0.751**	**0.751**
NB	0.394	0.370	0.394	0.353
Neural Net.	0.749	0.748	0.749	0.748

<Performance – Testsets>

Performance of KNN = 0.751

[1] Vergara, M., Ramos, L., Rivera-Campoverde, N. D., & Rivas-Echeverría, F. (2023). Enginefaultdb: a novel dataset for automotive engine fault classification and baseline results. *IEEE Access*, *11*, 126155-126171.

BASELINE SURVEY-DATASETS

Engine Fault Diagnosis

Datasets Name

- EngineFaultDB (Supervised Learning)

Achieved by

- Mary Vergara, Diego Rivera, Francklin Rivas-Echeverría

Test Engine

- C14NE Spark Ignition Engine

Specification	Detail
Maximum power	83.7 HP @ 6000 RPM
Torque	113.56 N.m @ 3000 RPM
Displacement	1388 cc
Injection system	Multipoint
Fuel consumption	6.8 l/100 km
Valve configuration	SOHC

<Test Engine Configuration>

Engine Data Collection Method

<Engine Data Collection / Gas Analyzer Device>

- Gas Analyzer
- USB 6008 Data Acquisition Card (DAQ)

Datasets (Github) : LINK

INPUT

1. Manifold Absolute Pressure (MAP)
   - Pressure inside manifold [kPa]
2. Throttle Position Sensor
   - Position of Throttle: about fuel injection, ignition time, etc. [%]
3. Force
   - Engine torque/rotational force [N]
4. Power
   - Energy transferred in engine [kW]
5. Revolutions Per Minute (RPM)
   - The times crankshaft rotates per minute
6. Fuel consumption L/H
   - Engine's fuel consumption rate
7. Fuel consumption L/100KM
   - Engine's fuel efficiency by distance
8. Speed
   - Vehicle's travel speed [km/h]
9. Carbon monoxide (CO)
   - CO concentration in the exhaust gases [%]
10. Hydrocarbons (HC)
    - Hydrocarbons concentration [%]
11. Carbon dioxide (CO2)
    - CO2 concentration : combustion efficiency [%]
12. Oxygen (O2)
    - Oxygen concentration : insights about combustion [%]
13. Lambda
    - Air-fuel equivalence ratio
14. Air-Fuel Ration (AFR)
    - Ratio of the air fuel in the combustion chambers

OUTPUT

Fault type 0: Normal (16,000 entries)

Fault type 1: Rich mixture - High Pressure, Incorrect Sensor, etc. (10,988 entries)

Fault type 2: Lean mixture – Low Pressure, Incorrect Sensor, etc. (15,000 entries)

Fault type 3: Low Voltage – Worn Spark, Defective Coil, etc. (14,001 entries)

[2] Thomas, L. (2024). EngineFaultDB Dataset. GitHub. https://github.com/Leo-Thomas/EngineFaultDB

PROJECT PROGRESS SCHEME

Engine Fault Diagnosis

Method

Data Analysis → Preprocessing → Feature Extraction → Feature Reduction/Selection → Classification → Evaluation

Data Analysis
① Plot Data to know datasets Tendency

Preprocessing
① Remove Outlier data
② Correlation Check

Feature Extraction
① Extract Time-domain
② Extract Frequency-domain
③ Analyze Kurtogram / STFT

Feature Reduction/Selection
① PCA
② Manual / Automatic Selection

Classification
① Logistic Regression
② Decision Trees
③ Random Forests
④ SVM
⑤ KNN, etc.

Evaluation
① Accuracy
② Precision
③ Recall
④ F1-Score

Expected Outcome

Extracting/Reducing/ Selecting Features →

① Derive Meaningful Information
② Migrates Overfitting
③ Enhance Computational Efficiency

GOAL

Higher Performance than Baseline Journal (≥ 0.751)

Schedule

	1st Week	2nd Week	3rd Week
Data Analysis	■		
Preprocessing	■		
Feature Extraction	■		
Feature Reduction/Selection		■	
Classification		■	
Evaluation			■
Writing Report			■

Role Distribution

AN GYEON HEAL
- Analyze about Baseline Journal
- Search Engine Structure & Vulnerable Part or Process
- Research about Diagnosis Process

JIN GA RAM
- Search Applicable Way or Field
- Search Additional Datasets
- Research about Additional ML Method that can Improve Model Performance

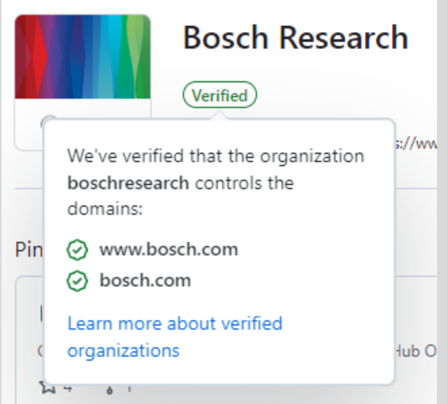
ADDITIONAL RESEARCH

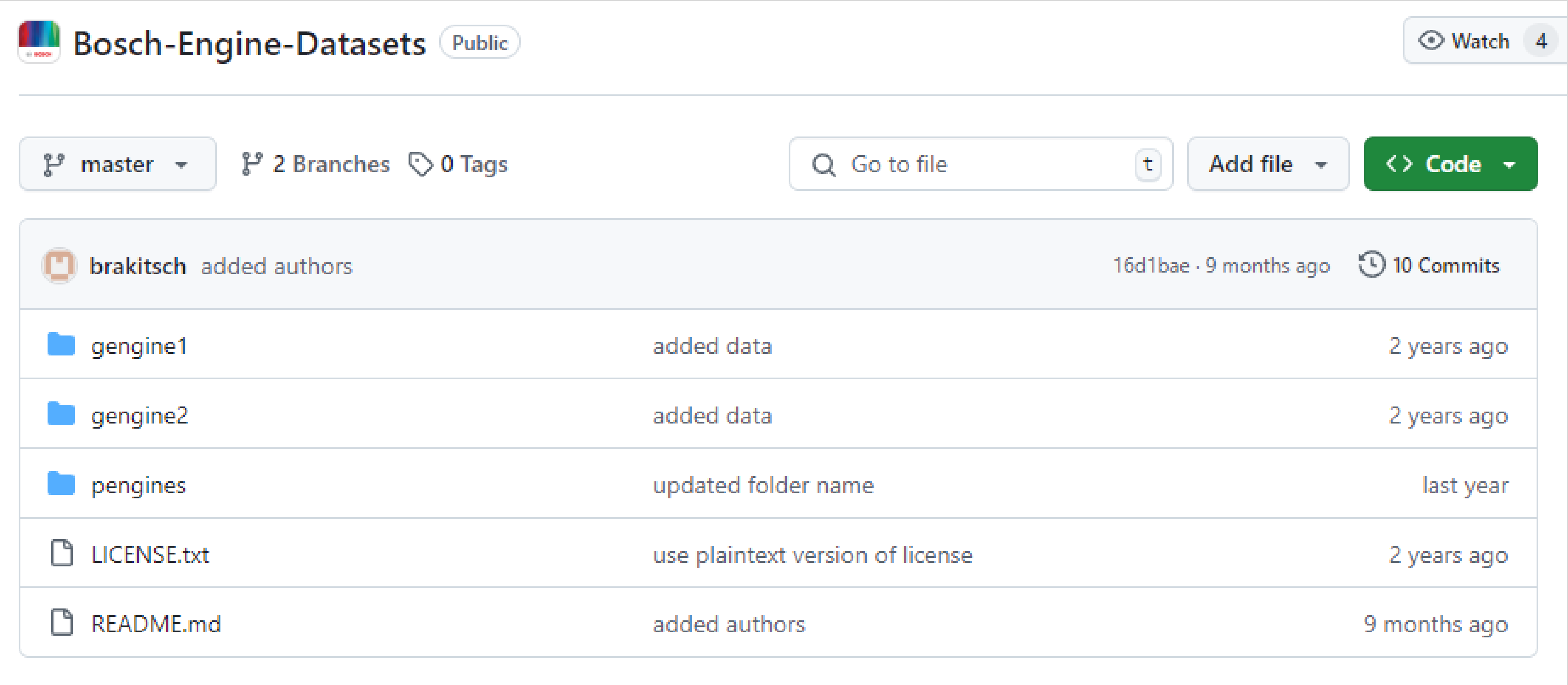
Engine Fault Diagnosis

Further Study

Apply the Model to the Open BOSCH Engine Dataset Available on BOSCH's Official GitHub.

Datasets Information





Datasets

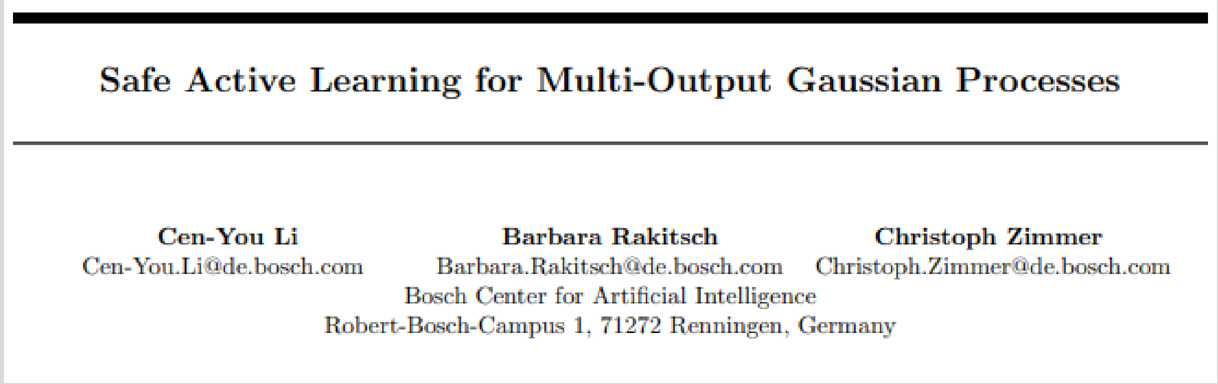
#1. gengine 1
- Speed, Load, Lambda, Ignition Angle, Fuel cutoff, CO, CO2, HC, NOx, O2, Temperature(Manifold), Temperature (Catalyst)

#2. gengine 2
- Speed, Load, Lambda, Ignition Angle, HC, NOx, O2, Temperature(Manifold), Temperature (Catalyst)

#3. pengines
- engine speed, engine load, intake valve opening, air fuel ratio, specific fuel consumption, temperature exhaust manifold, temperature (Catalyst), cylinder pressure, HC, NOx

[4] Bosch Research. (n.d.). Bosch Engine Datasets. GitHub. https://github.com/boschresearch/Bosch-Engine-Datasets

Related Survey/Research



Safe Active Learning for Multi-Output Gaussian Processes

Cen-You Li, Cen-You.Li@de.bosch.com
Barbara Rakitsch, Barbara.Rakitsch@de.bosch.com
Christoph Zimmer, Christoph.Zimmer@de.bosch.com
Bosch Center for Artificial Intelligence
Robert-Bosch-Campus 1, 71272 Renningen, Germany

- ❖ Uses "Multiple Output Gaussian Process" (Safe Active Learning)
- ❖ RMSE Performance Comparison: Active Learning MOGP (≤ 0.4)
- ❖ Datasets : BOSCH-Engine-Datasets
- ❖ Supervised Training : Labeling Output (HC, NOx, O2, etc.)

Plan

- ❖ Train ML Model which Labeled by (HC, NOx, O2, etc.) and Compare Performance

OR

- ❖ Diagnose Engine Fault Using BOSCH-Engine-Datasets

[3] Li, C. Y., Rakitsch, B., & Zimmer, C. (2022, May). Safe active learning for multi-output gaussian processes. In International Conference on Artificial Intelligence and Statistics (pp. 4512-4551). PMLR.

REFERNCES

Engine Fault Diagnosis

Baseline Survey – Journal & Datasets

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 10 October 2023, accepted 2 November 2023, date of publication 8 November 2023, date of current version 14 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3331316

RESEARCH ARTICLE

EngineFaultDB: A Novel Dataset for Automotive Engine Fault Classification and Baseline Results

MARY VERGARA[1], LEO RAMOS[2,3], (Student Member, IEEE), NÉSTOR DIEGO RIVERA-CAMPOVERDE[4], (Member, IEEE), AND FRANCKLIN RIVAS-ECHEVERRÍA[3,5], (Senior Member, IEEE)

[1]Higher School of Engineering, Science and Technology, Valencian International University, 46002 Valencia, Spain
[2]Computer Vision Center, Universitat Autónoma de Barcelona, Bellaterra, 08193 Barcelona, Spain
[3]Kauel Inc., Houston, TX 77027, USA
[4]Grupo de Investigación en Ingeniería del Transporte (GIIT), Universidad Politécnica Salesiana, Cuenca 010105, Ecuador
[5]Escuela de Ingeniería, Pontificia Universidad Católica del Ecuador Sede Ibarra, Ibarra 100112, Ecuador

Corresponding authors: Mary Vergara (maryjosefina.vergara@professoruniversidadviu.com) and Leo Ramos (leo.ramos@kauel.com)

This work was supported by the Grupo de Investigación en Ingeniería del Transporte (GIIT) at the Universidad Politécnica Salesiana, Ecuador.

[1] Vergara, M., Ramos, L., Rivera-Campoverde, N. D., & Rivas-Echeverría, F. (2023). Enginefaultdb: a novel dataset for automotive engine fault classification and baseline results. *IEEE Access*, *11*, 126155-126171.

[2] Thomas, L. (2024). EngineFaultDB Dataset. GitHub. https://github.com/Leo-Thomas/EngineFaultDB

Additional Research

Safe Active Learning for Multi-Output Gaussian Processes

Cen-You Li Cen-You.Li@de.bosch.com
Barbara Rakitsch Barbara.Rakitsch@de.bosch.com
Christoph Zimmer Christoph.Zimmer@de.bosch.com
Bosch Center for Artificial Intelligence
Robert-Bosch-Campus 1, 71272 Renningen, Germany

Abstract

Multi-output regression problems are commonly encountered in science and engineering. In particular, multi-output Gaussian processes have been emerged as a promising tool for modeling these complex systems since they can exploit the inherent correlations and provide reliable uncertainty estimates. In many applications, however, acquiring the data is expensive and safety concerns might arise (e.g. robotics, engineering). We propose a safe active learning approach for multi-output Gaussian process regression. This approach queries the most informative data or output taking the relatedness between the regressors and safety constraints into account. We prove the effectiveness of our approach by providing theoretical analysis and by demonstrating empirical results on simulated datasets and on a real-world engineering dataset. On all datasets, our approach shows improved convergence compared to its competitors.

1 Introduction

Active learning (AL) selects the most informative data sequentially according to previous measurements and an acquisition function (Krause et al., 2008; Houlsby et al., 2011; Zhang et al., 2016). The objective is to optimize a model without labeling unnecessary data. The problem setup is closely related to Bayesian op- of a machine are not supposed to crash any objects. A system should avoid generating high pressure, high temperature, or explosion. Safe learning addresses this by incorporating and learning safety constraints (Sui et al., 2015). Schreiter et al. (2015) and Zimmer et al. (2018) combine safety considerations with AL so that the data selection is done only in the determined safe domain.

These works, however, rarely considered multi-output (MO) regression problems, despite them commonly encountered in science, engineering and medicine (Xu et al., 2019; Zhang and Yang, 2021; Liu et al., 2018). In such problems, it is possible to consider individual tasks or outputs independently, but the plausibly shared mechanisms are ignored, and the performances or data efficiency might be deteriorated. Zhang et al. (2016) dealt with AL on MO models but focused on efficient computation of AL with large datasets and safe exploration was not addressed.

We consider safe AL for MO regression models that exploit the correlations. In particular, we focus on problems in which different output components may not be synchronously observed (e.g. due to different measuring cost or difficulty). MO Gaussian processes (GPs) are natural candidates for these problems (Bonilla et al., 2008; Álvarez and Lawrence, 2011; Álvarez et al., 2012; van der Wilk et al., 2020), due to their capability of capturing the correlations among different outputs and of quantifying the uncertainty.

In our work, we consider as main model the Linear Model of Coregionalization (LMC, Journel and Huijbregts (1976)), in which each output is modeled as a weighted sum of shared latent functions. Each

[3] Li, C. Y., Rakitsch, B., & Zimmer, C. (2022, May). Safe active learning for multi-output gaussian processes. In International Conference on Artificial Intelligence and Statistics (pp. 4512-4551). PMLR.

[4] Bosch Research. (n.d.). Bosch Engine Datasets. GitHub. https://github.com/boschresearch/Bosch-Engine-Datasets