SLOVAK UNIVERSITY OF TECHNOLOGY
IN BRATISLAVA

FACULTY OF CHEMICAL AND FOOD TECHNOLOGY

Reg. No.: FCHPT-FCHPT-16584-112282

Anomaly Detection in Industrial Data

MASTER THESIS

2026 **Bc. Patrícia Arbetová**

SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA

FACULTY OF CHEMICAL AND FOOD TECHNOLOGY

Reg. No.: FCHPT-FCHPT-16584-112282

Anomaly Detection in Industrial Data

MASTER THESIS

Study programme:	Automation and Information Engineering in Chemistry and Food Industry
Study field:	Cybernetics
Training workspace:	Institute of Information Engineering, Automation, and Mathematics
Thesis supervisor:	doc. Ing. Radoslav Paulen, PhD.
Consultants:	Ing. Rastislav Fáber, Ing. Karol Ľubušký

2026 **Bc. Patrícia Arbetová**

Slovak University of Technology in Bratislava
Department of Information Engineering and Process Control

Faculty of Chemical and Food Technology
Academic year: 2025/2026
Reg. No.: FCHPT-16584-112282

STU FCHPT

MASTER THESIS TOPIC

Student: **Bc. Patrícia Arbetová**
Student's ID: 112282
Study programme: Automation and Information Engineering in Chemistry and Food Industry
Study field: Cybernetics
Thesis supervisor: doc. Ing. Radoslav Paulen, PhD.
Head of department: prof. Ing. Miroslav Fikar, DrSc.
Workplace: Department of Information Engineering and Process Control

Topic: **Anomaly Detection in Industrial Data**

Language of thesis: English

Specification of Assignment:

The main objective of this work is the design and comparison of methods for detecting anomalies and fault conditions in industrial time series. Early and accurate detection of unexpected behavior (e.g., sensor failures, process deviations) is crucial to ensure production stability and safety. The thesis will analyze historical operational data from a real industrial process. Several approaches based on statistical methods and machine learning will be investigated and compared in order to identify the most suitable algorithm for the given application.

Tasks:
Processing and visualization of industrial time-series data
Application of selected anomaly detection algorithms (e.g., clustering-based methods, autoencoders)
Design of an early warning system for abnormal process behavior
Evaluation of detection performance and sensitivity to false alarms

Selected bibliography:

1. FÁBER, Rastislav; MOJTO, Martin; ĽUBUŠKÝ, Karol and PAULEN, Radoslav. From Data to Alarms: Data-driven Anomaly Detection Techniques in Industrial Settings. In: *Proceedings of the 34th European Symposium on Computer Aided Process Engineering /15th International Symposium on Process Systems Engineering.* Amsterdam, Holandsko: Elsevier, 2024, p. 2857–2862. ISBN 9780443288241.
2. Foorthuis, R. On the nature and types of anomalies: a review of deviations in data. International Journal of Data Science and Analytics, 12, 2021.
3. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. ACM Computing Surveys, 41:1–58, 2009.

Declaration of Honour

I declare that the submitted master thesis was completed on my own, in cooperation with my supervisor and consultants, with the help of professional literature and other information sources, which are cited in my thesis in the reference section.

During the preparation of this thesis, I used ChatGPT as an auxiliary tool for consultation, especially for code writing, text formulation, and language corrections. All generated outputs were reviewed, verified, and adapted by me, and I take full responsibility for the final content of the thesis.

As the author of my diploma thesis, I declare that I did not break any third party copyrights.

Acknowledgment

I would like to express my sincere gratitude to my supervisor, doc. Ing. Radoslav Paulen, PhD., for his professional guidance, patience, valuable feedback, and continuous support throughout the preparation of this thesis. His advice and constructive comments helped me to better understand the topic and improve the quality of this work.

My sincere thanks also go to Ing. Rastislav Fáber and Ing. Karol Ľubušký for their helpful consultations, practical advice, and willingness to share their expertise. Their support was especially valuable for understanding the industrial data used in this work and for connecting the theoretical methods with practical process knowledge.

Finally, I would like to thank my family and friends for their encouragement and support during my studies, especially my best friend, who had to survive my moods during the writing of this thesis. I would also like to thank all sources of caffeine that helped me get through the final days of writing this thesis.

Abstract

Industrial processes produce large amounts of time-series data from online sensors and laboratory analyses. Online measurements are available frequently, but they can be noisy, inaccurate, or affected by sensor faults, while laboratory measurements are more reliable but collected much less often. This difference makes anomaly detection difficult, especially when labels of abnormal behavior are not available. This thesis focuses on anomaly detection in industrial data using measurements with different fidelity levels, where low-fidelity data come from online sensors and high-fidelity data come from laboratory analyses. A multi-fidelity model is used to combine both data sources and to create a model-derived reference signal, which is then used to define pseudo Ground Truth anomalies. Several anomaly detection methods were studied and compared using a confusion matrix. We used different feature selection and reduction approaches to create various input data subsets. To improve model performance, the Nelder–Mead optimization algorithm was used to tune the parameters of each method based on confusion-matrix metrics computed against the pseudo Ground Truth. Three main approaches were used: simple output signal processing methods, prediction methods, and a clustering method. We evaluated simple output signal processing methods to determine whether they can be used for anomaly detection, and we also tested an unsupervised clustering method to study whether ground truth labels are necessary. The results show that simple output signal processing methods are not sufficient but can be used in combination with other methods. On the other hand, the best results were achieved with prediction methods especially in combination with one of the output signal processing methods. The clustering method showed better performance than the output signal processing methods, but it was not able to match the performance of the best model.

Key words: anomaly detection, pseudo ground truth, optimization, confusion matrix, sensors

Abstrakt

Priemyselné procesy produkujú veľké množstvo časových dát z online senzorov a laboratórnych analýz. Online merania sú dostupné často, môžu však obsahovať šum, nepresnosti alebo byť ovplyvnené poruchami senzorov, zatiaľ čo laboratórne merania sú spoľahlivejšie, ale získavajú sa výrazne menej často. Tento rozdiel komplikuje detekciu anomálií, najmä v prípadoch, keď nie sú dostupné označenia abnormálneho správania procesu. Táto diplomová práca sa zaoberá detekciou anomálií v priemyselných dátach s využitím meraní s rôznymi úrovňami presnosti, pričom dáta s nízkou presnosťou pochádzajú z online senzorov a dáta s vysokou presnosťou z laboratórnych analýz. Na prepojenie oboch zdrojov dát je použitý multifidelitný model, ktorý vytvára modelom odvodený referenčný signál. Tento signál je následne použitý na definovanie pseudo-referenčných označení anomálií. V práci bolo analyzovaných a porovnaných viacero metód detekcie anomálií pomocou matice zámeny. Na vytvorenie rôznych vstupných dátových podmnožín boli použité viaceré prístupy výberu a redukcie príznakov. Na zlepšenie výkonnosti modelov bol použitý Nelderov-Meadov optimalizačný algoritmus, ktorým boli ladené parametre jednotlivých metód na základe metrík matice zámeny vypočítaných vzhľadom na pseudo-referenčné označenia anomálií. Použité boli tri hlavné prístupy: jednoduché metódy úpravy výstupného signálu, predikčné metódy a zhlukovacia metóda. Jednoduché výstupové metódy boli hodnotené s cieľom zistiť, či môžu byť použité na detekciu anomálií. Zároveň bola testovaná aj zhlukovacia metóda bez učiteľa s cieľom preskúmať, či sú pri detekcii anomálií potrebné referenčné označenia. Výsledky ukázali, že jednoduché metódy úpravy výstupného signálu nie sú postačujúce, ale môžu byť použité v kombinácii s inými metódami. Najlepšie výsledky boli dosiahnuté pomocou predikčných metód, najmä v kombinácii s jednou z metód úpravy výstupného signálu. Zhlukovacia metóda dosiahla lepšie výsledky ako výstupové metódy, nedokázala však dosiahnuť výkonnosť najlepšieho modelu.

Kľúčové slová: detekcia anomálií, pseudo referenčné označenie, optimalizácia, matica zámeny, senzory

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Modern industrial factories generate a large amount of data from sensors and instruments [8]. There is a clear difference between sensor and laboratory measurements. Laboratory analyses are generally considered more accurate and trustworthy and are used as a reference for evaluating process quality. They represent high-fidelity (HF) data, which means that they are accurate but less frequent. On the other hand, sensor measurements provide low-fidelity (LF) data that are more frequent but less reliable. Because of their continuous operation in harsh industrial conditions, LF measurements are more likely to contain abnormal behavior. Such deviations are referred to as anomalies and can make the interpretation of process data unreliable. In industrial monitoring [18], undetected anomalies may lead to incorrect conclusions about the process state, delayed operator response, false alarms, or missed fault conditions. As a result, they can negatively affect product quality, process stability, economic performance, and safety. Therefore, detecting anomalies in LF sensor data is an important task for reliable process monitoring and early warning system design.

Several approaches have been proposed for anomaly detection in industrial and time-series data. These methods usually detect anomalies by monitoring deviations of a measured signal from its expected behavior. Recent studies have investigated different machine learning approaches for anomaly detection. Traditional signal-processing and output-based methods are often used because they are simple, computationally efficient, and easy to interpret. In industrial process monitoring, such methods are commonly used to characterize normal process variation and detect abnormal changes in measured variables [17]. Prediction based

approaches have also been widely studied, where the expected behavior of the system is learned from data and anomalies are detected as deviations from this behavior. Zhang et al. (2022) applied Decision Tree and Random Forest models for anomaly detection in financial data, where observations were classified as normal or abnormal [28]. Mingrui et al. (2023) studied anomaly detection in multivariate time-series data using a neural-

network-based model, which was designed to detect deviations from expected temporal patterns [10]. Clustering-based and density-based approaches have also been applied to anomaly detection, where anomalous observations are usually identified as points located in sparse regions or outside the main data structure. Mehmood et al. (2025) used a hybrid method combining Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Isolation Forest (iForest) for outlier detection in process data [12]. Although these studies show that different statistical and machine learning methods can be useful for detecting abnormal behavior in complex datasets, many existing works focus on a single method, a single class of methods, or one hybrid approach. This makes it difficult to compare different anomaly detection strategies under the same industrial case study, pseudo ground truth definition, and evaluation framework.

In contrast, this work provides a comprehensive comparison of several anomaly detection approaches. The compared methods include simple output-based approaches, such as the Exponentially Weighted Moving Average (EWMA) filter and the difference method, prediction-based approaches, such as Ordinary Least Squares (OLS) regression and Gaussian Process Regression (GPR), and the cluster-based method Density-Based Spatial Clustering of Applications with Noise (DBSCAN). Besides the comparison of individual detection methods, different input data feature selection and reduction techniques, including Principal Component Analysis (PCA) and Partial Least Squares (PLS), are also evaluated.

We use the Nelder-Mead algorithm to optimize selected method parameters according to performance metrics from the confusion matrix. This allows the methods to be compared not only in terms of detected anomalies, but also with respect to false alarms and missed fault conditions. Since manually labeled anomaly data are not available, this work uses the prediction of a multi-fidelity (MF) model, created by Fáber et al. (2025), as a reference ground truth for anomaly detection [6].

1.1 Motivation

Industrial processes are very complex systems in which even small deviations from normal operating conditions can gradually develop into significant operational problems. Therefore, it is often difficult to determine whether the process is operating normally or whether an abnormal situation is developing. Anomaly detection is important because undetected abnormal behavior can lead to reduced product quality, increased material or energy losses, and potential safety risks. For this reason, engineers need reliable and interpretable methods that can detect abnormal process behavior at an early stage.

This work is motivated by the need to understand which anomaly detection methods are practical and effective in real industrial processes. The goal is to compare simple and easily interpretable methods with more advanced data-driven approaches and to identify the most suitable strategy for the considered industrial application. Such a comparison can support better decision-making in process monitoring and contribute to earlier detection of abnormal process behavior.

1.2 Problem Statement

In this thesis we are working with time series data. The industrial process is monitored using a set of sensors, producing multivariate input measurements. These input measurements are represented as

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{L,1} & x_{L,2} & \cdots & x_{L,n} \end{bmatrix} \in \mathbb{R}^{L \times n},$$

where each row corresponds to a time stamp, L, and each column corresponds to a specific sensor signal, n.

The corresponding output measurements are collected as a single-column time-series

$$\mathbf{y}^{LF} = \begin{bmatrix} y_{1,1} \\ y_{2,1} \\ \vdots \\ y_{L,1} \end{bmatrix} \in \mathbb{R}^{L \times 1},$$

which aligns temporally with the rows of $\mathbf{X}$ and represents the process output.

The HF laboratory measurements of the output variable are available at a limited number of time stamps, M, and represented as

$$\mathbf{y}^{HF} = \begin{bmatrix} y_{1,1} \\ y_{2,1} \\ \vdots \\ y_{M,1} \end{bmatrix} \in \mathbb{R}^{M \times 1}.$$

The true process output $\mathbf{y}^{*LF}$ represents the ideal, noise-free behavior of the LF data

and $\mathbf{y}^{*HF}$ represents the ideal behavior of the HF data. In practice, it cannot be observed directly, since all real measurements include some form of error:

$$\mathbf{y}^{LF} = \mathbf{y}^{*LF} + \varepsilon_{LF}, \quad \mathbf{y}^{HF} = \mathbf{y}^{*HF} + \varepsilon_{HF},$$

where ε_{LF} and ε_{HF} denote errors, with $\varepsilon_{HF} \ll \varepsilon_{LF}$. Therefore, actual observations always reflect the sum of the true process and the measurement noise, rather than $\mathbf{y}^*$ itself.

1.3 Objectives and Contributions

The main objective of this thesis is to design, implement, and compare selected anomaly detection methods for industrial time-series data. The work focuses on historical operational data from a real industrial process, where both LF sensor measurements and HF laboratory measurements are available. Following the assignment of the thesis, the first objective is the processing and visualization of industrial time-series data. This includes preprocessing of sensor measurements, removal of unreliable sensors, handling of missing values, and creation of input data subsets based on different sensor selection strategies. The second objective is the application of selected anomaly detection algorithms. In this work, output-based methods, prediction-based methods, and a clustering-based method are implemented and compared using a confusion matrix.

The main contribution of this work is a systematic comparison of several anomaly detection strategies, pseudo-ground-truth definition, and evaluation framework. The results provide insight into which methods are suitable for detecting anomalies in industrial time-series data and how feature selection, dimension reduction, and parameter optimization influence the final detection performance.

Chapter 2

Anomalies

Anomalies are patterns in a dataset that are unusual in some way and do not follow general trends [7]. Figure 2.1 shows an example of data points in a two-dimensional space. Two large data clusters, N_1 and N_2, represent areas without anomalies based on trends. Clusters A_1, A_2, and A_3 are located far from these regions and are relatively small, indicating that they are probably groups of anomalies [4]. A threshold is usually chosen so that points lying far from normal regions are classified as anomalies. The choice of this threshold depends on the specific task and the probability of encountering such points in normal operation [20].

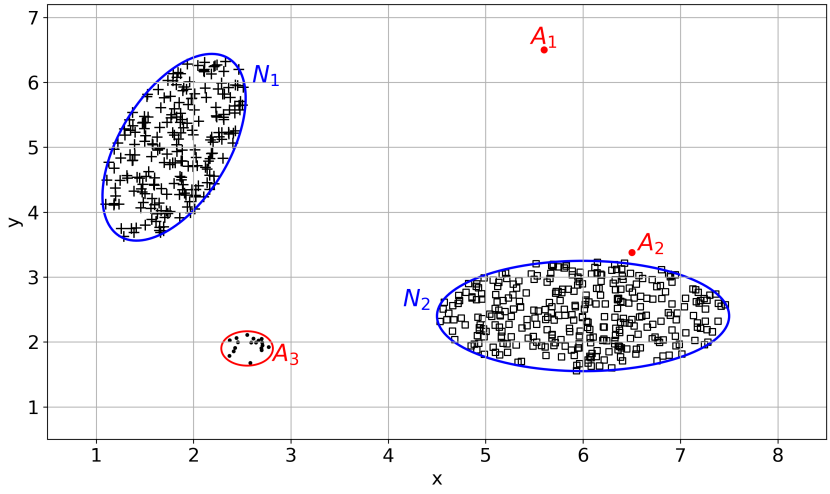

Figure 2.1: Example of normal and anomalous clusters in a two-dimensional space

2.1 Types of Anomalies

Anomalies are divided into three main classes: point anomalies, contextual anomalies and collective anomalies. Each type represents a different kind of deviation from expected patterns. Figure 2.2 shows examples of these classes highlighted with red color.

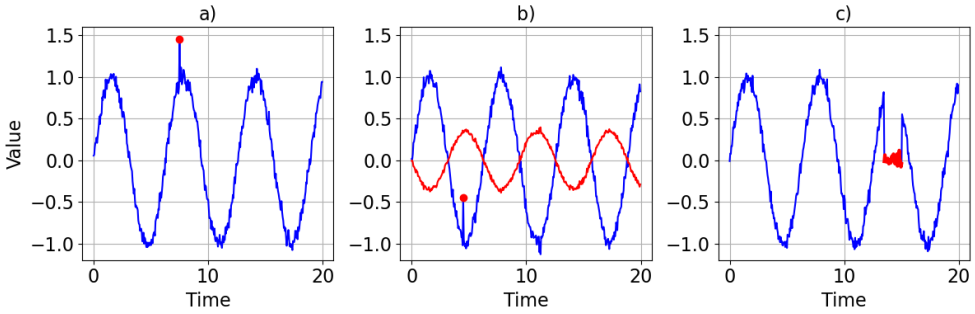

Figure 2.2: Examples of anomaly types: a) Point anomaly, b) Contextual anomaly, c) Collective anomaly

Point anomalies stand out from the rest of the data measurements and are the simplest to detect. Contextual anomalies are data instances whose abnormality depends on the context in which they occur, while they may be considered normal outside that context. They can be a single point anomaly or a group anomaly. An example of contextual point anomaly and group anomaly is shown in the Figure 2.2b). Sets of related or dependent data instances that deviate from the behavior of the entire dataset are known as collective anomalies. In this case individual points may not be anomalies by themselves, but when combined with other data points, they can be identified as a group of anomalies.

2.2 Anomaly Detection

Features of Anomaly Detection

A good anomaly detection system should provide features as self-service, holistic, explainable and extensive. These features are important when choosing or designing anomaly detection methods for industrial data.

When we want our system to be extensive it means that the used method should handle a large range of data and perform rather well with uncommon, near-constant data, as well as other challenging data conditions. The method must be flexible and able to handle these situations. Another important feature is self-service, in reality engineers and operators use these systems and some of them may not be experts in machine learning. For this reason, anomaly detection should be easy enough so it is understandable for everyone. The system also should be holistic, that means able to analyze the whole data flow. It should detect anomalies from high to low levels, no matter how near or how far are they from the normal data. The final feature is explainability of detection system. Most, if not all, of the users do not see the logic behind the system, so it is important to understand why a point is marked as an anomaly. This helps to build trust in the system and allows better decision-making in real applications [2].

Anomaly Detection Approaches

There are three main types of anomaly detection approaches based on available labeled anomaly points [20]: Supervised, Unsupervised and Semi-Supervised. The approach where training data are fully labeled is known as Supervised. In this setting, labels indicate whether a point is normal or anomalous. The model learns to distinguish between normal data and anomalies. However, because anomalous data are often rare and not fully representative, this approach is less common in practical applications. On the other hand, Unsupervised anomaly detection is an approach with only unlabeled data available for training a model. Last approach, Semi-Supervised anomaly detection, combines both unlabeled and labeled data. Usually, most of the data are unlabeled, and only a few points are. Labeling can be expensive and often requires domain experts, such as engineers.

Chapter 3

Methodology

3.1 Data Preprocessing and Feature Engineering

Before anomaly detection, data preparation is an important step. The available data are usually divided into separate subsets used for model development, parameter tuning, and final evaluation. The training set (TR) is used to fit the model and to estimate model parameters. During this stage, the model learns relationships and patterns present in the data. In machine learning applications, a validation set can be used during model development to compare different model settings and tune hyperparameters. The validation set helps to evaluate the model on data that were not directly used for fitting, while still allowing adjustments of the model before the final evaluation. This reduces the risk of selecting a model that performs well only on the training data. The test set (TS) is kept separate from the training and tuning procedure and is used only to evaluate the final performance of the selected model on unseen data [3, 14]. This separation provides a more reliable estimate of how well the model generalizes to new measurements and helps to reduce the risk of overfitting.

There are several ways to split a dataset: random sampling, stratified dataset, and cross-validation splitting. The choice of the splitting method depends on the type of data, the size of the dataset, the data quality, and the goal of the model.

Random splitting, in this method, the data are randomly divided into training, validation, and testing sets according to a selected ratio. This method is simple and often works well when the data are balanced and all parts of the dataset have similar properties.

Stratified splitting method is useful when the dataset is imbalanced, which means that some classes have many samples and other classes have only a few samples. Stratified splitting keeps the same class proportions in the training, validation, and testing sets. This helps the model learn from all classes and reduces bias in the evaluation.

Cross-validation is another commonly used approach. In this method, the dataset is divided into several smaller parts. The model is trained and evaluated several times, and each part is used as a validation set once. This gives a more reliable estimate of model performance [3].

Data cleaning is an important part of data preparation. It is used to identify and remove incorrect, incomplete, or unreliable measurements from the dataset [24]. In industrial time-series data, this step can include the removal of non-functional sensors, sensors with sudden value fluctuations, and sensors with long linear or constant segments. Such measurements may not represent the real process behavior and can negatively affect the performance of anomaly detection methods.

Standardization is also essential in preprocessing. It transform data to measurements with zero mean and unit variance [23]. This is done by subtracting the mean value of the feature and dividing by its standard deviation [14]:

$$z_i = \frac{g_i - \mu}{\sigma},$$

where g_i is the original value, μ is the mean value of the feature, σ is the standard deviation of the feature, and z_i is the standardized value. This approach is well suited for many machine learning models [23], especially linear and non-linear models that are sensitive to feature scaling.

3.1.1 Sensor Selection

Correlation-Based Selection

One part of data cleaning is also a removing duplicate signals, and correlation can help with this [24]. Correlation is a statistical measure that shows how strongly two variables are connected. A correlation value close to 1 means that these variables are close to each other, while a value close to 0 means the opposite. Correlation values can also be negative, down to -1. That means the variables are strongly related but in an opposite way, which means they behave similarly but in a mirrored manner. We can express this using the correlation matrix:

$$\Lambda = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}$$

where each ρ_{ij} is a correlation coefficient between two sensors. This method is simple and fast, which is important for industrial data with many sensors. It also helps to find redundant sensors. If two sensors measure similar things, they will have a high correlation in absolute value, which means that one of them can be removed without losing information. By keeping only the sensors with high correlation to the output and removing redundant ones anomaly detection can become faster and more accurate without losing important information [16].

Domain Knowledge

Domain knowledge plays an important role in sensor selection. It helps to identify which sensors are relevant for the given industrial process. Experts, such as process engineers, understand how the system works and can decide which variables have real influence on the output and remove sensors that are not important, even if they appear statistically relevant. This helps to guide the feature selection process and reduce the search space for potentially relevant input variables. However, the final effect on model simplicity and performance depends on the specific dataset and application [1].

3.1.2 Dimension Reduction

Principal Component Analysis (PCA) and Partial Least Squares (PLS) are commonly used methods for dimensionality reduction and feature extraction in machine learning.

PCA is an unsupervised dimensionality reduction method that transforms the original input variables into a new lower-dimensional representation. The new variables, called principal components, are linear combinations of the original variables. They are constructed so that the first principal component captures the largest possible variance in the data, while the following components capture the remaining variance under the condition that they are mutually uncorrelated. In this way, PCA can reduce the number of variables and help to simplify the input space while preserving a large part of the information contained in the original dataset. This can be useful for reducing computational complexity, improving visualization, and limiting problems related to high dimensionality. However, PCA is based only on the structure and variance of the input data. Therefore, it does not directly consider the relationship between the input variables and the output variable, and components with high variance are not always the most important for prediction or classification tasks [23].

PLS is a method that is similar to PCA, but it also considers the relationship between input variables and the output variable, so that means this method is supervised. PLS finds latent variables that maximize the covariance between the input data and the

output. This makes PLS more suitable for predictive modeling tasks, where the goal is to explain or predict the output variable [19].

Both methods can improve the performance of anomaly detection systems by reducing dimensionality and highlighting important patterns in the data [14].

3.2 Methods for Anomaly Detection

3.2.1 Output Signal Processing Methods

This section describes methods that improve anomaly detection by adjusting and analyzing the output signals from industrial systems. These techniques process the raw data to reduce noise and highlight changes in the signal. To detect anomalies, a threshold is created around this signal and any point out of this threshold is labeled as anomaly.

Exponentially Weighted Moving Average (EWMA) filter

EWMA [15], is a commonly used method for smoothing time series data and detecting trends in signals. It computes a weighted average of past observations, where the weights decrease exponentially over time. This means that more recent observations have a higher influence on the result than older ones.

Let x_t represent the input signal at time t and y_t the filtered output. The EWMA can be defined recursively as

$$y^{EWMA} = \alpha x_t + (1 - \alpha) y_{t-1},$$

where $\alpha \in (0, 1]$ is the smoothing parameter. The value of α controls how fast the influence of past observations decreases. A larger value of α gives more weight to recent data and makes the filter more sensitive to changes. On the other hand, a smaller value of α results in stronger smoothing and slower response to changes in the signal. This way the method can reduce noise while still preserving important trends in the data. In industrial applications, EWMA is useful for filtering noisy sensor signals and improving anomaly detection. By smoothing the signal, it becomes easier to identify significant deviations that may indicate abnormal behavior in the system.

Difference Method

The difference method is a simple technique used to highlight changes in time series data. It is based on computing the difference between consecutive values of the signal.

Let x_t represent the input signal at time t. The first difference is defined as

$$y^{\text{diff}} = x_t - x_{t-1},$$

which can be interpreted as a discrete approximation of the first-order derivative. If the signal is constant or changes slowly, the difference values will be close to zero. On the other hand, sudden changes in the signal will result in large difference values, and are, most of the time, marked as anomalies. The difference method is useful in anomaly detection because anomalies often appear as sudden jumps or drops in the signal. By transforming the data using differences, these changes become more visible and easier to detect.

3.2.2 Prediction Based Methods

Prediction-based methods use past observations to estimate future values. These methods build models that capture patterns in historical data and then use them to predict outputs for new inputs. Two common approaches are Ordinary Least Squares regression and Gaussian Process Regression. After predicting the signal behavior a threshold is set around it and any point from the original measurement out of this boundary is labeled as anomaly. These two methods were selected because they represent two different types of regression models (linear and non-linear).

Ordinary Least Squares (OLS) Regression

OLS regression is a statistical method used to model the relationship between input variables and an output variable. The goal is to find a linear function that best fits the observed data.

The model can be expressed as

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i,$$

where $\beta_0, \beta_1, \ldots, \beta_p$ are model parameters and ε_i represents the error term.

OLS estimates the parameters by minimizing the sum of squared differences between the observed values and the predicted values. This can be written as

$$\min \sum_{i=1}^{N} (y_i - \hat{y}_i)^2.$$

By solving this optimization problem, we obtain the estimated parameters

$$\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p,$$

which define the best fitting linear model.

The predicted output is then given as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}.$$

This method is widely used because it provides a simple way to model relationships between variables and approximate real data using a linear function [11].

Gaussian Process Regression (GPR)

GPR is a machine learning method used for regression tasks. Instead of defining a fixed model structure, GPR defines a distribution over possible functions, which allows the model to flexibly adapt to complex data patterns. A Gaussian process is defined as a collection of random variables, where any finite subset has a joint Gaussian distribution. It can be written as

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')),$$

where $m(x)$ is the mean function and $k(x, x')$ is the covariance function (kernel). The kernel defines the similarity between data points and plays a key role in shaping the predicted function [25].

The choice of the kernel function significantly affects the performance of the model. One commonly used kernel is the **Rational Quadratic (RQ) kernel**, which can be interpreted as a scale mixture of Radial Basis Function kernels with different length scales. It is defined as

$$k(x_i, x_j) = \left(1 + \frac{d(x_i, x_j)^2}{2\lambda l^2}\right)^{-\lambda},$$

where l is the length scale parameter and λ controls the relative weighting of different scales [14]. Another important kernel is the **White Noise kernel**, which models the noise of the signal as independently and identically normally-distributed. It is defined as

$$k(x_i, x_j) = \sigma_n^2 \delta_{ij},$$

where σ_n^2 represents the variance of the noise and δ_{ij} is the Kronecker delta defined as [14]

$$\delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

Together, these kernels have good potential to model both smooth variations at different length scales and random measurement noise, which makes them suitable for noisy industrial time-series data.

3.2.3 Cluster Based Methods

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

DBSCAN is a clustering algorithm that groups data points based on their density. The main idea is that clusters are formed in areas where many data points are close to each other, while points that are far from others are considered noise.

The algorithm is controlled by two main parameters. The first parameter defines the radius of a neighborhood around each point (`eps`). The second parameter defines the minimum number of points that must be present in this neighborhood for the point to be considered part of a cluster (`min_samples`). If a point has enough neighboring points within the given radius, it is included in a cluster. If this condition is not fulfilled, the point is treated as noise. Clusters are then created by connecting points that satisfy this density condition. This approach allows DBSCAN to identify clusters of different shapes and sizes, while also separating noise from the main data structure [9].

3.3 Evaluation and Optimization

3.3.1 Evaluation Metrics

Confusion matrix is a table that helps understand how well a classification model works. Confusion matrix organizes predictions into four categories. True Positives (TP) are correct positive predictions, True Negatives (TN) are correct negative predictions, False Positives (FP) are incorrect positive predictions, and False Negatives (FN) are incorrect negative predictions. These four values create a 2×2 table (Figure 3.1) that shows exactly what model predicted correctly.

From the confusion matrix, you can calculate several important metrics to measure model performance.

Accuracy is the simplest metric used to evaluate a classification model. It represents the percentage of correct predictions made by your model.

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

Precision tells us how many of your positive predictions are actually correct. High

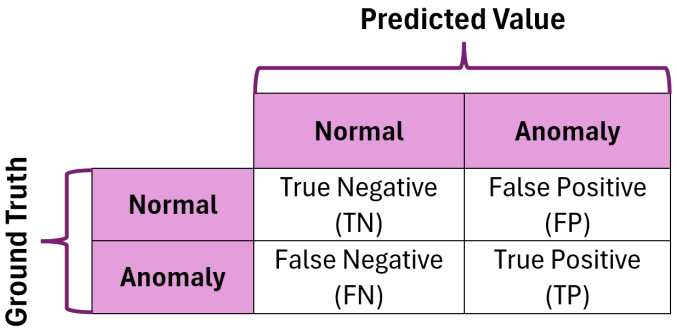

Figure 3.1: Confusion matrix

precision means that model identify small number of FP and correctly detect anomalies.

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

Sensitivity tells us how many actual positive cases you found. High sensitivity is preferred because it means that the model is able to detect most of the actual anomalies and minimizes the number of missed anomalous events. The formula for sensitivity based on a confusion matrix is:

$$Sensitivity = \frac{TP}{TP + FN}$$

Precision and sensitivity often have an opposite relationship. When we try to improve one of them, the other one can become worse. It is important to find a balance between precision and sensitivity. For this reason, the *F1-Score* is used. The F1-score has values between 0 and 1. A higher value means better performance. It combines precision and sensitivity into one value and is useful when we need to evaluate both measures together [26]. The F1-score is the harmonic mean of precision and sensitivity and is calculated as:

$$F1\text{-}Score = \frac{2 \cdot (Precision \cdot Sensitivity)}{Precision + Sensitivity} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

Specificity tells how many actual negative cases were correctly identified by the model. High specificity is preferred because it means that the model produces a small number of false alarms. It is defined as the ratio of true negative predictions to all actual negative samples and is calculated as [5]:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

The Receiver Operating Characteristic (ROC) curve is used to evaluate the performance of a classification model or a diagnostic test. The ROC curve represents the relationship between sensitivity and specificity for different threshold values. If a threshold value for the predictive variable is chosen below the lowest observed value, the point $(0, 0)$ is generated in the ROC space. As the threshold value increases, additional points are generated and connected to form the ROC curve. A threshold value greater than the highest observed value produces the point $(1, 1)$. The diagonal line connecting the points $(0, 0)$ and $(1, 1)$ represents predictions equivalent to random guessing. Points located above this diagonal line indicate better predictive performance. In general, the further the ROC curve is above the diagonal line, the better the predictive ability of the model.

The ROC curve also demonstrates the trade-off between sensitivity and specificity. These two quantities are inversely related, meaning that increasing sensitivity usually leads to lower specificity, while increasing specificity leads to lower sensitivity. A commonly used quantitative measure derived from the ROC curve is the Area Under the Curve (AUC), also referred to as the c-statistic. The AUC evaluates the ability of the model to discriminate between classes. Higher AUC values indicate better classification performance. The interpretation of AUC values can be summarized as follows [27, 13]:

- $\text{AUC} \leq 0.6$ – Fail,
- $0.6 < \text{AUC} \leq 0.7$ – Poor,
- $0.7 < \text{AUC} \leq 0.8$ – Fair,
- $0.8 < \text{AUC} \leq 0.9$ – Good,
- $0.9 < \text{AUC} \leq 1.0$ – Excellent.

3.3.2 Optimization

The Nelder-Mead Algorithm [22], is a method used for finding the minimum of a function. It is a direct search method, which means that it does not require derivatives of the function. This makes it useful for problems where the function is not smooth or is difficult to differentiate. The algorithm works with a geometric object called a simplex. In a two-dimensional space, the simplex is a triangle, and in higher dimensions it has more points. Each point of the simplex represents one possible solution. At the beginning, an initial simplex is created. Then, the algorithm evaluates the function at

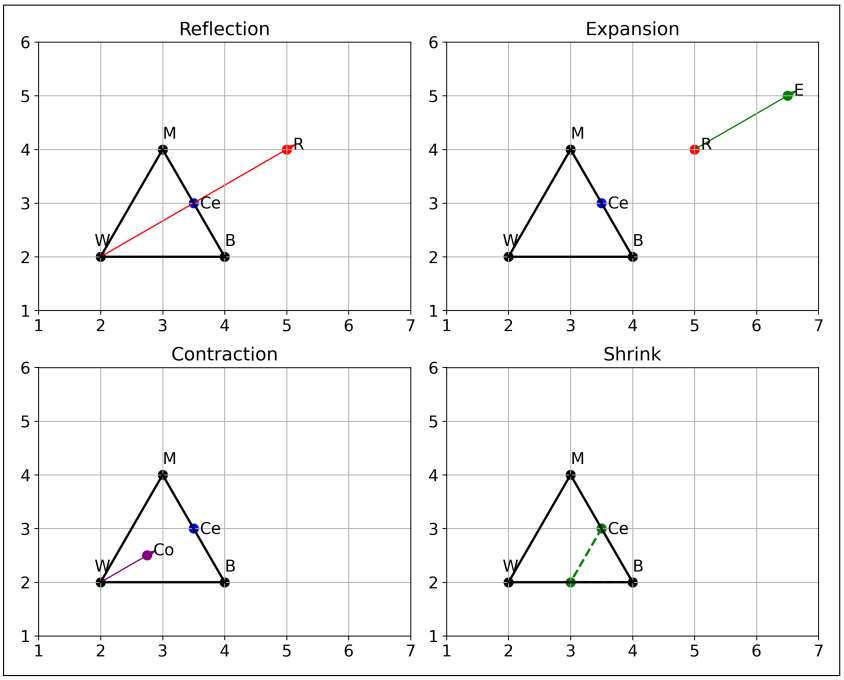

Figure 3.2: Example of Reflection, Expansion, Contraction and Shrink

each point of the simplex. Based on these values, the simplex is updated using several operations, such as reflection, expansion, contraction, and shrinking.

Reflection moves the worst point of the simplex to the opposite side of the simplex, usually towards a better region of the search space. Expansion continues in the same direction as reflection and tries to find an even better point if the reflection improves the solution. Contraction moves the worst point closer to the better points if reflection does not give a good result. Shrink reduces the size of the whole simplex by moving all points closer to the best point, which helps the algorithm focus on a smaller region.

Figure 3.2 shows examples of reflection, expansion, contraction and shrink. For better understanding, the letter B denotes the best point, M denotes the middle point, and W denotes the worst point of the simplex. The centroid is denoted by Ce and is calculated from the B and M points. The reflection point is denoted by R, the expansion point by E, and the contraction point by Co. These points help to illustrate how the simplex moves during the optimization process. The algorithm repeats this process until a stopping condition is satisfied, for example when the change in function values becomes

very small. The Nelder-Mead method is simple to implement and does not require gradient information. However, it may converge slowly and does not always guarantee finding the global minimum [22].

Chapter 4

Implementation

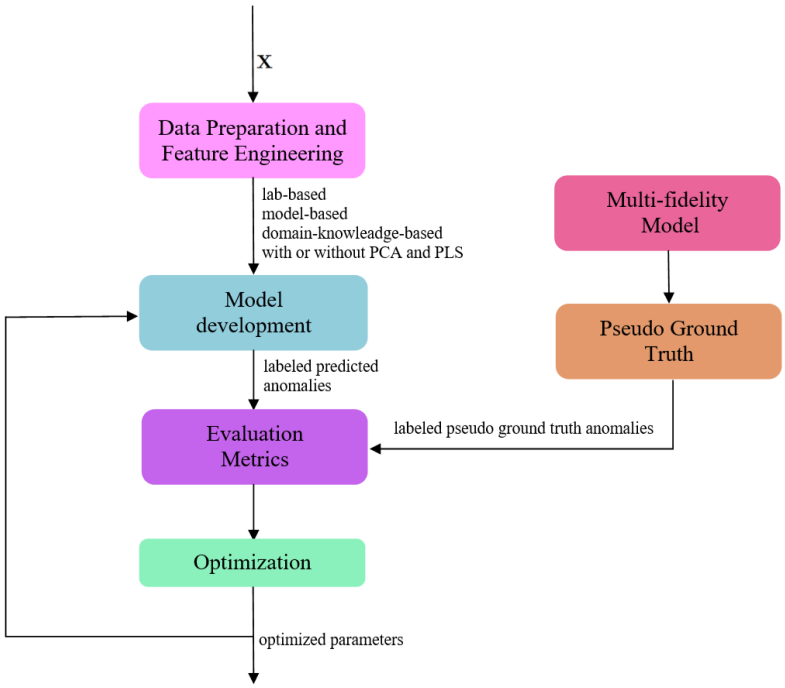

Figure 4.1: Implementation workflow of the proposed anomaly detection framework

Figure 4.1 presents the overall workflow of the implemented anomaly detection framework. We started with data preparation and feature engineering, where we transformed the input signal using different feature extraction approaches. Then, we used the prepared data for model development and compared the detected anomalies with the labeled pseudo ground truth anomalies. Based on this comparison, we evaluated the

model using selected evaluation metrics. Finally, we optimized the model parameters according to the obtained results.

4.1 Case Study

Alkylation is a chemical process in which isoparaffins react with olefins to produce a high-octane product called alkylate. It is added to gasoline to improve its quality. Catalysts such as sulfuric acid or hydrofluoric acid are used in this process. Based on this we divide alkylation in two categories, Hydrofluoric acid alkylation and Sulfuric acid alkylation

Hydrofluoric acid alkylation process

In this process, isobutane reacts with olefins in the presence of hydrofluoric acid, which acts as a catalyst. The feed for the process consists of isobutane and olefins. Before entering the reactor, the feed is treated to remove sulfur compounds and other impurities. After this pretreatment, the feed is mixed in a specific ratio and sent to the reactor. The reactor operates at relatively low temperatures and moderate pressure, which helps control the reaction. Inside the reactor, olefins react with isobutane and are almost completely converted into alkylate, which is the main product. The output from the reactor contains alkylate, excess isobutane, light hydrocarbons such as propane and n-butane, and the acid catalyst. This mixture is then sent to a settler, where the hydrofluoric acid separates from the hydrocarbons because of its higher density. The acid is collected and recycled back to the reactor. The remaining hydrocarbons are further processed in separation units, such as fractionators, where individual components are separated. Some of the separated streams are recycled back into the process, for example isobutane, while others are removed as final products, such as propane and alkylate.

Sulfuric acid alkylation process

It is similar to hydrofluoric acid alkylation, as it also uses isobutane and olefins to produce alkylate. The main difference is that sulfuric acid is used as the catalyst instead of hydrofluoric acid. In this process, the reaction usually takes place at lower temperatures compared to hydrofluoric acid alkylation. For this reason, additional cooling is required to maintain stable operating conditions. The process often uses refrigeration systems to control the temperature in the reactor. The feed preparation and reaction steps are similar, but the separation and handling of the acid are different. Sulfuric acid is heavier and more difficult to recycle, so the process requires more

complex handling and regeneration of the acid. Hydrofluoric acid alkylation typically achieves higher yield and does not require expensive refrigeration. However, due to safety concerns related to hydrofluoric acid, sulfuric acid alkylation is often preferred in practice. It provides a good balance between safety, operability, product quality, and yield [21].

Available Measurements

In this work, anonymized data from Slovnaft, a.s. related to the sulfuric acid alkylation process are used as a case study.

Figure 4.2 shows the input matrix $\mathbf{X}$ consisting of multiple sensor signals. The dataset contains measurements from $n = 731$ sensors over $L = 10\,000$ time stamps. The signals represent different physical quantities, such as concentration, temperature, and flowrate. They are sampled every 11 minutes and show different levels of variability, including noise and occasional faults. Some of the sensors are not fully functional and contain missing values (NaN).

Figure 4.3 presents $\mathbf{y}_{LF}$ together with $\mathbf{y}_{HF}$. Both represent the concentration of recycled isobutane. The LF signal contains $L = 10\,000$ time stamps and is sampled every 11 minutes. In contrast, $\mathbf{y}_{HF}$ is available only once per month, resulting in a total of 11 measurements. It can be observed that the online signal does not always follow the laboratory values closely. In some parts of the signal, there are visible deviations between the two measurements. Since laboratory data are considered more reliable, these differences indicate potential anomalies in the sensor measurements.

Based on this observation, anomaly detection methods are applied to identify abnormal behavior in the signal.

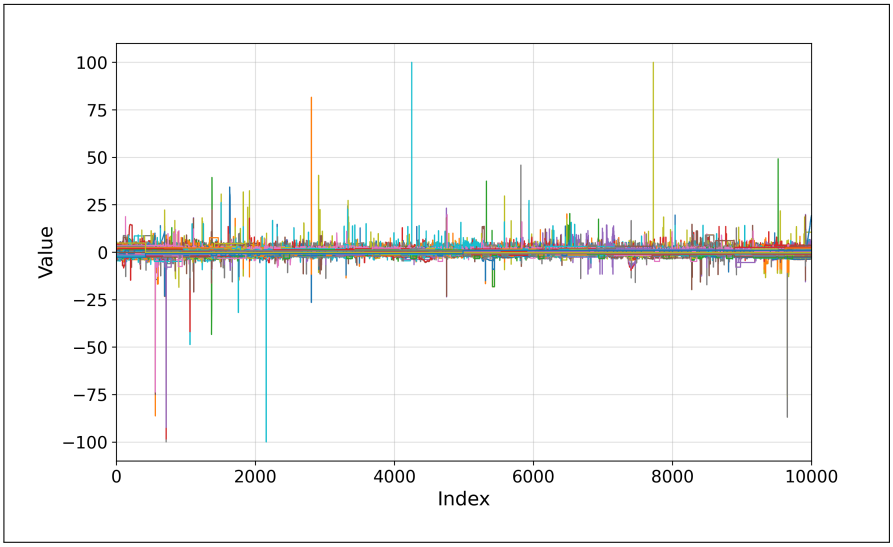

Figure 4.2: Online sensor measurements used as input data.

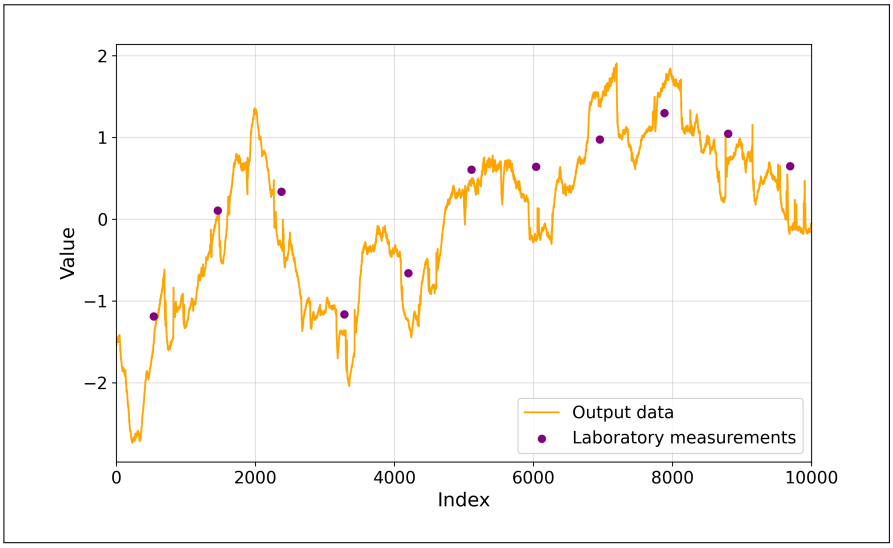

Figure 4.3: Output data and laboratory measurements representing the concentration value of recycled isobutane.

4.2 Multi-fidelity Model (MF)

The MF model used in this work was proposed by Fáber et al. (2025) and was created based on the idea of correcting the difference between $\mathbf{y}_{LF}$ and $\mathbf{y}_{HF}$, closely described in Section 4.1. First, feature selection methods such as Principal Component Regression, Partial Least Squares, LASSO, and Stepwise Regression are used to select the most important input variables. These variables are then used to build a model that describes the process behavior. To capture time dependencies in the data, time-lagged variables are included in the model. This is important because the process contains recycle streams, which introduce delays and dynamic behavior.

A Gaussian Process model is then used to model the deviation between the dynamic LF model prediction and the available $\mathbf{y}_{HF}$. This deviation represents a correction term that is added to the dynamic LF prediction. The final MF model therefore does not directly correct the original online signal, but corrects the prediction of the dynamic LF model obtained using SIPPY. In this way, the MF model combines the temporal behavior captured by the dynamic LF model with the accuracy information provided by the $\mathbf{y}_{HF}$ [6].

Pseudo Ground Truth (pGT)

Because true verified anomaly labels are not available, our evaluation uses MF-derived pGT labels. Therefore, the reported metrics quantify agreement with the MF reference criterion, not necessarily agreement with real process faults. Anomalies are identified by comparing the original online measurements shown in Figure 4.3 with the MF reference signal. A threshold is defined around this reference signal, and all points outside this threshold are marked as anomalies. This process is illustrated in Figure 4.4. The threshold is derived empirically from the available HF laboratory samples. First, the absolute differences between the LF online measurements and the corresponding HF laboratory values are calculated at the laboratory sampling points. Their average represents the typical mismatch between the online analyzer and the laboratory reference. Since the MF signal is used as a corrected estimate of the expected process behavior, one half of this average mismatch is selected as a tolerance band for Pseudo Ground Truth labeling. The threshold can be expressed as

$$\tau_{pGT} = \frac{1}{2} \cdot \frac{1}{M} \sum_{k=1}^{M} \left| \mathbf{y}^{LF} - \mathbf{y}^{HF} \right|,$$

where M is the number of laboratory samples. The threshold value used in this work is 0.436.

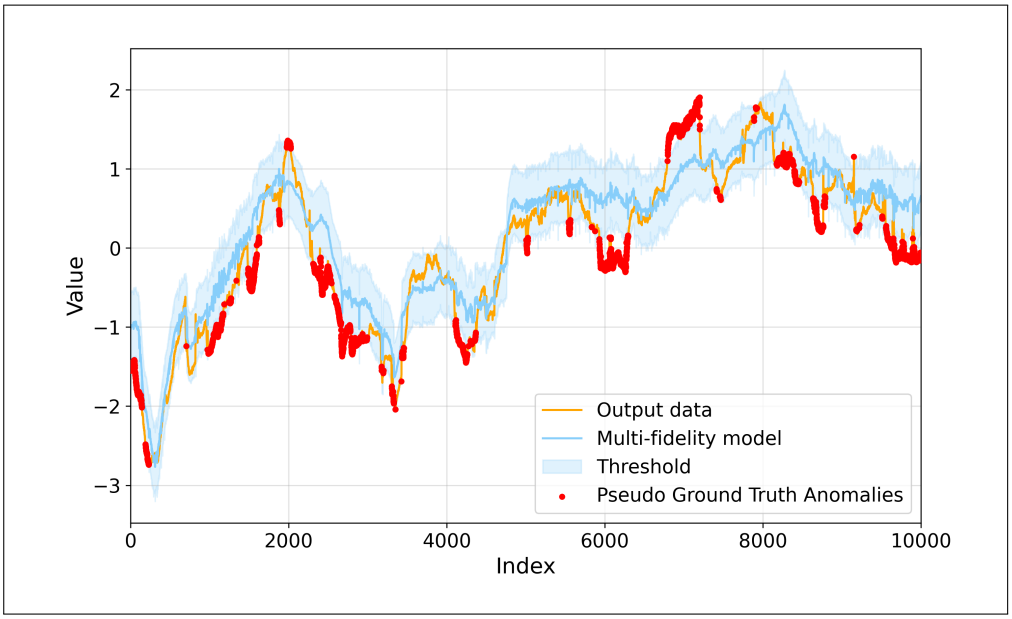

Figure 4.4: Pseudo Ground Truth anomalies based on the Multi-fidelity reference signal with a threshold.

This generated pGT is later used to evaluate the performance of anomaly detection methods using the confusion matrix, closely described in Section 3.3.1.

4.3 Data Preparation and Feature Engineering

Figure 4.5 shows the workflow used in this section. The original input data matrix **X** is first preprocessed to remove unreliable sensors and handle missing values. After preprocessing, feature selection is performed using three different approaches: lab-based, model-based, and domain-knowledge-based selection. The selected input subsets are either used directly for model development or further transformed using dimensionality reduction methods, namely PCA and PLS. In this way, several input configurations are created and later used for anomaly detection model development.

At the beginning, all available measurements are divided into TR and TS sets. Since the data represent a time series, the split is performed chronologically rather than randomly. The first 70 % of the data are used as the TR set, while the remaining 30% are used as

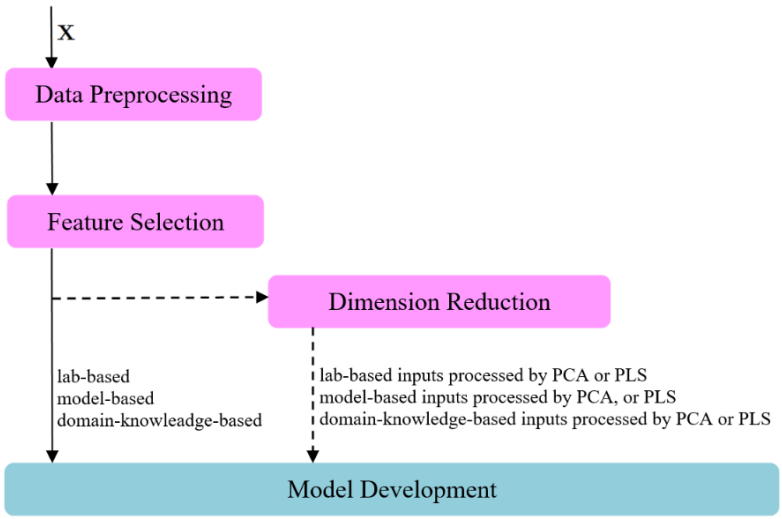

Figure 4.5: Workflow of data preparation and feature engineering used before model development.

the TS set. The result of this split is TR set with size 7000×731 and TS 3000×731. We did not use separate validation set in this work because of the limited length of the available dataset and the small number of HF laboratory measurements. Creating an additional validation subset would further reduce the amount of data available for model training and final testing. If a larger dataset or a longer measurement period were available, a separate validation set could be used. All preprocessing parameters, including standardization, feature-selection criteria, PCA/PLS transformations, model parameters, and detection thresholds, were determined using only the training set. We transformed testing set using the parameters fitted on the training set.

Data Preprocessing

Sensors that contain more than 80 % of missing values (NaN) are removed from the dataset. For the remaining sensors, missing values are replaced using interpolation based on neighboring values. Some sensors show almost linear behavior over time. These sensors are identified using linear regression, and sensors with a coefficient of determination R^2 higher than 0.9 are removed, since they do not provide useful dynamic information. In addition, sensors that do not change their value for a long period of time (more than 500 consecutive samples) are also removed. Sensors with sudden large spikes greater than 5 units are considered unreliable and are removed as

well.

Feature Selection

In this work, three different approaches are used to create subsets of input variables:

- The first approach we decided to use is called *lab-based*. In this method, the correlation between input sensors, obtained from the data preprocessing, and laboratory measurements is evaluated. Since laboratory measurements are available only at specific time points, the corresponding sensor values are selected in a small window of ±5 indices around each laboratory measurement. These values are averaged and then compared with the laboratory data. Sensors with correlation higher than 80 % are selected. To avoid redundancy, we evaluated correlation between the selected sensors, and sensors with mutual correlation higher than 95 % were removed. With this approach we selected five important sensors.

- The second approach is called *model-based*. In this case, the correlation between input sensors, from data preprocessing, and the MF model output described in Section 4.1 is calculated. Sensors with correlation higher than 80 % are selected. Similarly to the previous approach, highly correlated sensors are removed by applying a threshold of 95 % for mutual correlation. In total we obtained six sensors.

- The third approach is based on domain knowledge and is referred to as the *domain knowledge* approach. In this case, sensors are selected based on expert knowledge about the industrial process, as described in Section 3.1.1. Using knowledge from Section 3.1.1, we separated 6 important sensors. These sensors are chosen directly based on process understanding and are not limited to the preprocessing-based selection.

Although the three approaches use different selection criteria, they all identified one common sensor, which suggests that this sensor is relevant from both data-driven and process-based perspectives. The same selected sensors are then applied to the testing dataset to ensure consistency between TR and TS data. We standardized these three approaches before they are used in the anomaly detection models, as is described in Section 3.1.

Dimension Reduction

After feature selection, dimensionality reduction methods can be applied to further reduce the complexity of the input set while preserving information about how the input variables describe the output signal. In this work, PCA and PLS are used. The goal of these methods is to reduce the number of input variables while keeping the most important information. To determine how many components should be used, the relationship between the input data and the output signal is analyzed. For both PCA and PLS, the coefficient of determination R^2 is calculated for an increasing number of components. This shows how well the selected components explain the variability of the output signal. At the beginning, adding more components significantly improved the model. However, around the third point, the improvement becomes very small. The optimal number of components is selected at the point where the curve starts to flatten. This means that adding more components does not bring significant improvement, and the main information is already captured by the selected components.

4.4 Model Development and Optimization

4.4.1 Output Signal Processing Methods

EWMA-Based Anomaly Detection

The EWMA method was implemented as an output-based anomaly detection approach. This method was applied directly to the output signal $\mathbf{y}_{LF}(t)$ in order to smooth the data and reduce noise. This method is described in Subsection 3.2.1. We calculated EWMA first on TR dataset where α was set up and then used also on TS dataset After computing the EWMA signal, we performed anomaly detection by comparing the original signal with the smoothed signal. Upper and lower bounds were optimized and defined as:

$$y^{EWMA} \pm \tau^{EWMA}$$

If the original signal exceeded these bounds, the point was classified as an anomaly.

Difference Method

We applied this method directly to the output signal $\mathbf{y}_{LF}(t)$ by computing the difference between consecutive measurements in order to highlight abrupt changes in the signal. Since the differencing operation computes the difference between consecutive samples, the resulting signal contains one fewer data point than the original signal. Therefore, the pGT labels were adjusted accordingly by removing the first element to ensure proper alignment with the differenced signal. After computing the differenced signal, anomaly detection was performed by evaluating its magnitude. Upper and lower

bounds were defined as:

$$y^{\text{diff}} \pm \tau^{\text{diff}}$$

4.4.2 Prediction Based Methods

OLS-Based Anomaly Detection

The OLS method was implemented as a model-based anomaly detection approach. Several input data configurations were considered for the OLS model. In total, we evaluated nine different variants of the input matrix after feature selection and dimension reduction. These variants were created from three types of input selection approaches: lab-based, model-based, and domain-knowledge-based inputs. Each of these input sets was also evaluated in its original form and after dimensionality reduction using PCA and PLS. Therefore, the following input configurations were tested:

- lab-based inputs,
- lab-based inputs processed by PCA,
- lab-based inputs processed by PLS,
- model-based inputs,
- model-based inputs processed by PCA,
- model-based inputs processed by PLS,
- domain-knowledge-based inputs,
- domain-knowledge-based inputs processed by PCA,
- domain-knowledge-based inputs processed by PLS.

Two different output signals were considered as target variables for the OLS model. The first one was the LF output signal and the second one was the HF output signal. The LF-based OLS model was trained using the LF output signal $\mathbf{y}_{LF}$ and the HF-based OLS model, was trained using the available laboratory measurements, $\mathbf{y}_{HF}$.

The OLS model was fitted on the TR dataset according to the least-squares principle described in Section 3.2.2. After prediction, anomaly detection was performed by

comparing the measured output signal with the predicted output signal. We optimized and implemented bounds as:

$$\hat{y}_i^{OLS} \pm \tau_i^{OLS}$$

where $\hat{y}_i^{OLS}$ represents the output predicted by the OLS model for every combination.

GPR-Based Anomaly Detection

GPR was used to model the relationship between the nine input preprocessed variables and the output signal. For this method, we used the same input configurations and target signals as we described in the OLS-based method. Therefore, nine different input variants and two types of output signals (LF output signal and HF laboratory measurements) were considered. We created kernel function consisting of a RQ kernel and a White Noise kernel.

The model was trained on the TR dataset and subsequently used to predict the output signal for the TS dataset. The predicted output can be expressed as:

$$\hat{y}_i^{GPR} = f(\mathbf{x}_i)$$

where $f(\mathbf{x}_i)$ represents the nonlinear function estimated by the Gaussian Process model. After prediction, anomaly detection was performed using optimized threshold (τ_i^{GPR}), in the same way as before.

4.4.3 Cluster Based Methods

DBSCAN-Based Anomaly Detection

In this method we decided to use only original input configurations (lab-based, model-based and domain-knowledge-based inputs), without PCA or PLS transformations. Unlike OLS and GPR, DBSCAN does not predict the output signal. Instead, it identifies anomalies directly from the structure of the input data based on point density. Data points located in low-density regions are considered potential anomalies. We set `eps` and `min_samples` parameters based on the TR dataset and subsequently applied them to the TS dataset. In DBSCAN, data points labeled as noise are assigned the label -1, and these points were classified as anomalies. Additionally, since this method creates clusters, we decided to apply the condition that if one of the created clusters contains more than 51 % of pGT anomalies, the entire cluster is considered anomalous.

4.4.4 Combination of Methods

In addition to the individual methods described above, a combination of methods was also investigated in order to improve anomaly detection performance.

Predictive Models with EWMA

In this work, predictive models combined with the EWMA smoothing technique were also examined. The goal was to observe how smoothing of the selected predicted signal influences the resulting anomaly detection performance. Only selected prediction-based models were considered in this analysis based on their previous results. First, the output signal was predicted using the selected model. Then, the EWMA method was applied to the predicted signal to obtain a smoother version:

$$\hat{y}_j^{EWMA} = \text{EWMA}(\hat{y}_j^{OLS})$$

where $\hat{y}_j^{OLS}$ represent the selected predicted outputs.

After smoothing, anomaly detection was performed by comparing the measured output signal with the smoothed predicted signal using threshold ($\tau_j^{OLS-EWMA}$) in the same way as in the previous methods. The resulting prediction curves were subsequently compared with the original prediction-based methods using ROC curves and AUC values.

DBSCAN with Differenced Signal

In this approach, the DBSCAN method was applied not to the original input data, but to the differenced signal. Specifically, the input to the DBSCAN algorithm consisted of the first-order differences of the selected input variables, which emphasize abrupt changes in the signal and suppress slow trends. For this combined method, three input configurations were evaluated: lab-based, model-based, and domain knowledge-based inputs.

The differencing operation was applied to each input variable as:

$$x^{\text{diff}} = x(t) - x(t-1)$$

The DBSCAN algorithm was then applied to the transformed input space. By operating on the differenced data, the method focuses on detecting anomalies characterized by sudden changes rather than absolute deviations.

4.4.5 Parameter Optimization

For all methods, the model parameters were optimized using the TR dataset. The optimization was based on evaluation metrics derived from the confusion matrix. Two optimization criteria were used, the F1-score and a weighted combination of sensitivity and specificity ($0.5 \cdot \text{sensitivity} + 0.5 \cdot \text{specificity}$). Different parameters were optimized depending on the method. For the EWMA-based method, we decided to optimized both the smoothing parameter α and the detection threshold, τ. For the DBSCAN method, the parameters `eps` and `min_samples` were optimized. For all remaining methods, only the detection threshold was optimized. After optimization on the TR dataset, the selected parameters were applied to the TS dataset without further adjustment.

Chapter 5

Results and Discussion

This chapter presents the results obtained for the investigated anomaly detection methods. First, the optimized parameters are summarized for all considered methods and input configurations. Afterwards, the detection performance of the individual approaches is evaluated and discussed.

5.1 Optimized Metrics

5.1.1 Output Signal Processing Methods

For the EWMA-based method, the optimized parameters were determined using the criterion based on a combination of sensitivity and specificity. The smoothing parameter α was set to 0.003, and the detection threshold τ^{EWMA} was set to 0.407.

For the difference method, only the detection threshold τ^{diff} was optimized. Based on the same optimization criterion, the threshold was set to 0.110. These values were then used for anomaly detection in both the TR and TS datasets.

5.1.2 Prediction Based Methods

Tables 5.1 and 5.2 present the optimized thresholds for all considered input configurations, including lab-based, model-based, and domain knowledge-based inputs, as well as their PCA and PLS transformations. The results are shown separately for the LF and HF target signals and for both regression models (OLS and GPR). From these tables, it can be observed that the threshold values vary depending on the selected input configuration, model type, and target signal. In general, the thresholds for the HF signal are higher than those for the LF signal. This difference can be related to the nature of the target signals. The LF signal represents the standard output data,

which may not fully reflect the real system behavior. As a result, the prediction models can more easily fit this signal, leading to smaller differences between predicted and measured values and, consequently, lower threshold values. In contrast, the HF signal is based on laboratory measurements, which better represent the true system behavior. Therefore, larger differences between predicted and measured values can occur, which leads to higher optimized threshold values.

Table 5.1: Optimized thresholds using sensitivity and specificity

Model	Data	**Lab based**			**Model based**			**Domain knowledge**		
		Orig.	PCA	PLS	Orig.	PCA	PLS	Orig.	PCA	PLS
OLS	LF	0.289	0.304	0.360	0.293	0.360	0.300	0.122	0.266	0.313
	HF	0.558	0.360	0.495	0.461	0.470	0.495	0.325	0.270	0.178
GPR	LF	0.169	0.289	0.286	0.120	0.304	0.299	0.300	0.300	0.405
	HF	0.464	0.445	0.461	0.469	0.495	0.470	0.361	0.439	0.385

Table 5.2: Optimized thresholds using F1-score

Model	Data	**Lab based**			**Model based**			**Domain knowledge**		
		Orig.	PCA	PLS	Orig.	PCA	PLS	Orig.	PCA	PLS
OLS	LF	0.274	0.304	0.320	0.293	0.360	0.300	0.111	0.152	0.116
	HF	0.580	0.364	0.528	0.520	0.495	0.495	0.227	0.261	0.176
GPR	LF	0.005	0.131	0.124	0.032	0.296	0.300	0.000	0.006	0.000
	HF	0.508	0.445	0.461	0.469	0.512	0.555	0.371	0.468	0.385

5.1.3 Cluster Based Methods

The optimized DBSCAN parameters for all configurations are shown in Tables 5.3-5.4. The parameters were optimized using the two different criteria we mentioned in Section 4.4.5. These tables show how the parameters *eps* and *min_samples* change depending on the selected input group (lab-based, model-based, and domain knowledge-based). The results indicate that both parameters vary across configurations, with *eps* values remaining within a relatively narrow range, while *min_samples* shows more noticeable variation. Table 5.3 corresponds to optimization based on sensitivity and specificity and Table 5.4, uses the F1-score as the optimization criterion.

Table 5.3: Optimized DBSCAN parameters using sensitivity and specificity

Parameter	Lab based	Model based	Domain knowledge
eps	0.151	0.188	0.191
min_samples	10	6	7

Table 5.4: Optimized DBSCAN parameters using F1-score

Parameter	Lab based	Model based	Domain knowledge
eps	0.150	0.173	0.155
min_samples	10	8	10

5.1.4 Combination of Methods

Predictive Models with EWMA

In addition to the complete set of configurations, we selected several prediction models that achieved the best anomaly detection performance for further analysis. The analysis focused on the OLS model with the HF target signal using the following input configurations: lab-based PLS, model-based PCA and PLS, and domain-knowledge-based PCA. For these models, the EWMA method was additionally applied to the predicted output signal in order to smooth the prediction curves. The corresponding optimized thresholds for the combined prediction-based and EWMA-smoothed methods $\tau_j^{OLS-EWMA}$ are shown in Tables 5.5 and 5.6.

Table 5.5: Optimized thresholds for the best prediction-based methods combined with EWMA using sensitivity and specificity

Model	Data	Lab(PLS)	Model (PCA)	Model (PLS)	Domain (PCA)
OLS	HF	0.501	0.469	0.465	0.271

Table 5.6: Optimized thresholds for the best prediction-based methods combined with EWMA using the F1-score

Model	Data	Lab (PLS)	Model (PCA)	Model (PLS)	Domain (PCA)
OLS	HF	0.503	0.468	0.510	0.270

DBSCAN with Differenced Signal

Tables 5.7 and 5.8 present the results for DBSCAN when applied to differenced input data. Compared to the original input data, the values of *eps* are generally higher, indicating that larger distances between data points are allowed when clustering is performed on differenced signals.

Table 5.7: Optimized DBSCAN parameters using sensitivity and specificity (differenced inputs)

Parameter	Lab based	Model based	Domain knowledge
eps	0.490	0.560	0.503
min_samples	11	3	12

Table 5.8: Optimized DBSCAN parameters using F1-score (differenced inputs)

Parameter	Lab based	Model based	Domain knowledge
eps	0.317	0.451	0.101
min_samples	24	18	24

The increase in both *eps* and *min_samples* suggests that the transformed input data are more spread out in the feature space, while still requiring a stronger density condition to form clusters. A larger *eps* is needed to connect neighboring samples, whereas a larger *min_samples* prevents small random groups of points from being classified as dense regions. This indicates that the differenced inputs change the distance structure of the data and require a more conservative clustering criterion.

5.2 Model Development

5.2.1 Output Based Methods

Figure 5.1 shows the development of the signal in time after applying the EWMA method. The resulting F1-score obtained using optimization based on the combination of sensitivity and specificity was 0.414, while the value of the function $0.5 \cdot \text{Sensitivity} + 0.5 \cdot \text{Specificity}$ reached 0.588. These results indicate that the EWMA method was able to detect some anomalies in TS dataset, but the overall performance was moderate. Therefore, this method could be more suitable when combined with another more advanced anomaly detection method.

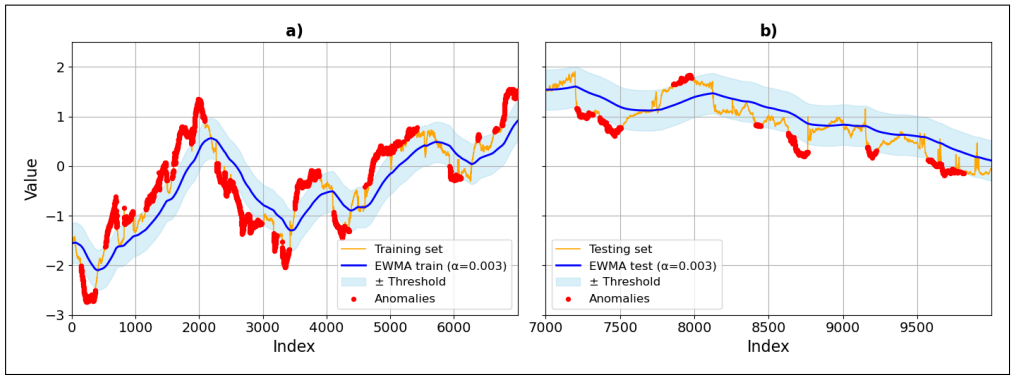

Figure 5.1: Anomalies based on EWMA a) training set, b) test set

Figure 5.2 shows the anomalies detected using the difference method. In this case, the resulting F1-score was very low, only 0.020, while the value of the function $0.5 \cdot \text{Sensitivity} + 0.5 \cdot \text{Specificity}$ was 0.496. These results indicate that the method detected only a very small number of anomalies and therefore did not provide satisfactory anomaly detection performance.

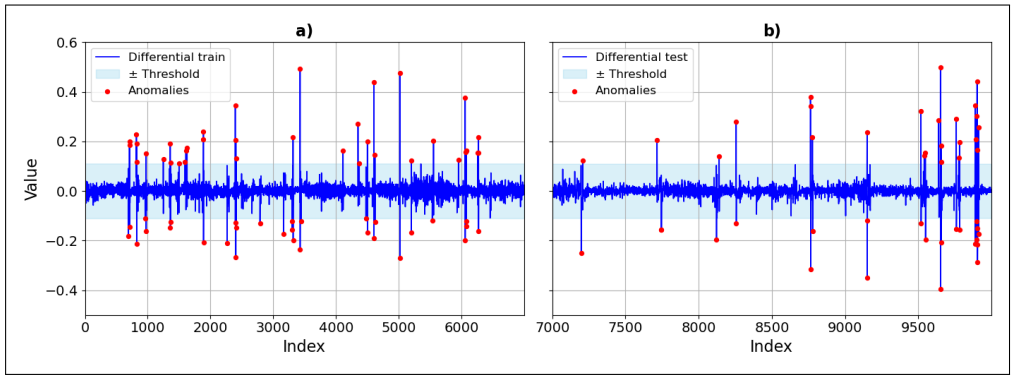

Figure 5.2: Anomalies based on difference method a) training set, b) test set

5.2.2 Prediction Based Methods

Since a large number of results were obtained for the prediction-based methods, only one selected example is shown in Figure 5.3. The figure illustrates the time evolution of the prediction-based anomaly detection method divided into the training set and the test set. The plot shows the measured signal, the predicted signal, the threshold

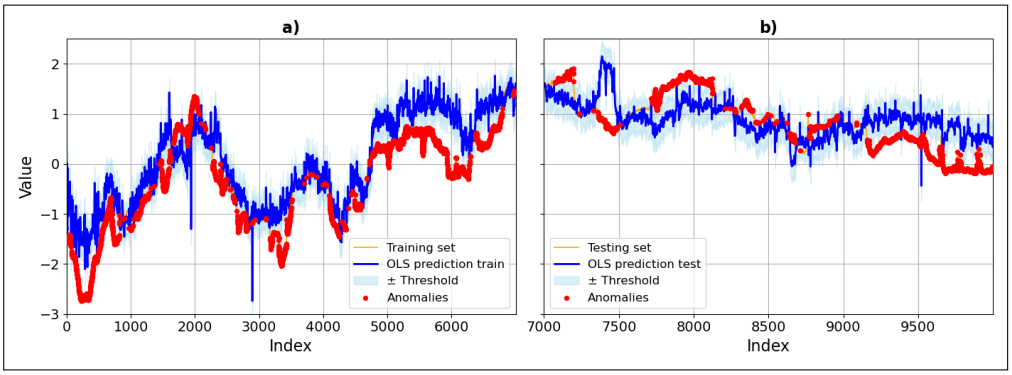

Figure 5.3: Anomalies based on one of prediction methods a) training set, b) test set

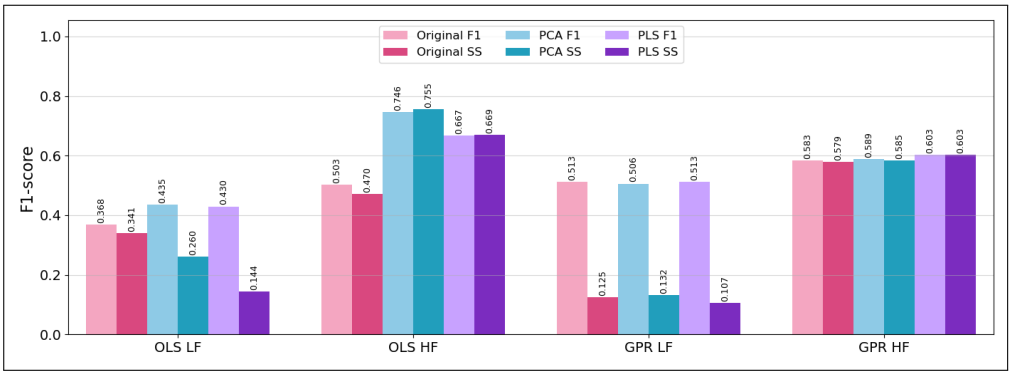

Figure 5.4: Comparison of F1-score values for prediction-based methods using domain-knowledge-based input data.

region, and the detected anomalies.

Figures 5.4-5.9 present the TS set results for OLS and GPR models trained with LF and HF target signals. Different input transformations are compared, including the original input data, PCA, and PLS. The labels F1 and SS indicate the optimization criterion used for threshold selection, where F1 corresponds to optimization based on the F1-score and SS corresponds to optimization based on the combination of sensitivity and specificity.

Figures 5.4-5.6 present results of F1-score. It can be observed that the best performance is generally achieved when PCA or PLS transformed inputs are used. In particular, the OLS HF models show the highest F1-score values across most configurations.

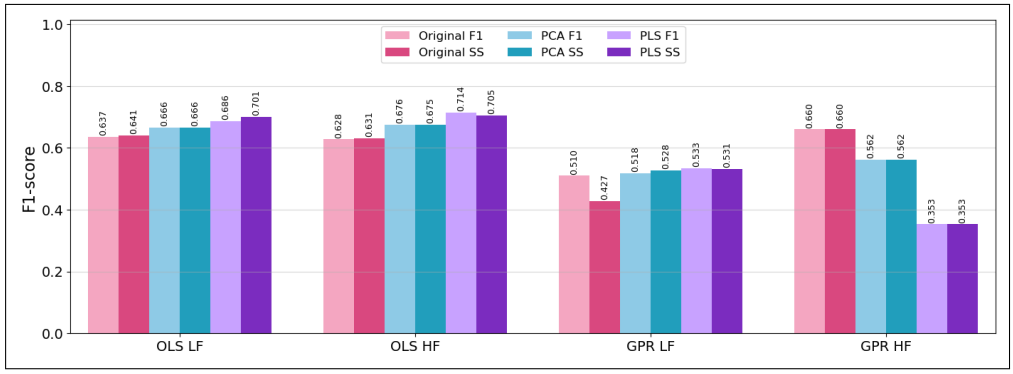

Figure 5.5: Comparison of F1-score values for prediction-based methods using lab-based input data.

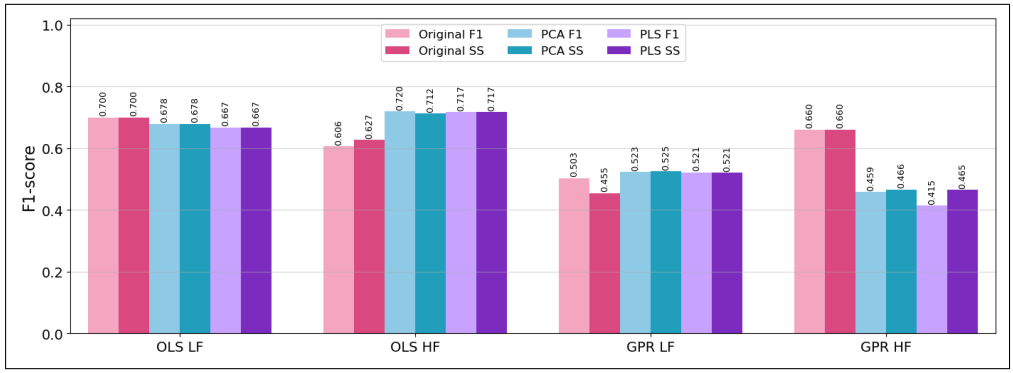

Figure 5.6: Comparison of F1-score values for prediction-based methods using model-based input data.

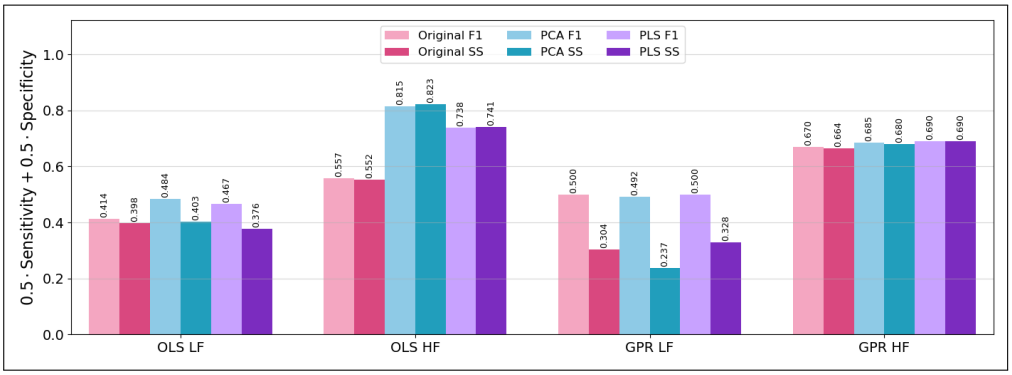

Figure 5.7: Comparison of 0.5·sensitivity+0.5·specificity values for prediction-based methods using domain-knowledge-based input data.

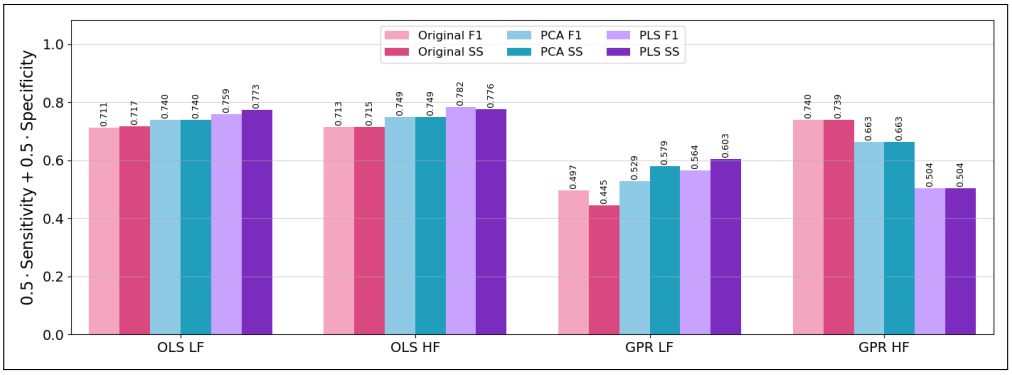

Figure 5.8: Comparison of 0.5·sensitivity+0.5·specificity values for prediction-based methods using lab-based input data.

This suggests that the laboratory-based target signal combined with dimensionality reduction methods can improve anomaly detection performance. In contrast, the LF target signal often leads to lower F1-score values, especially for the GPR model.

Figures 5.7-5.9 present the values of the combined metric 0.5·sensitivity+0.5·specificity for the prediction-based methods. In general, the OLS HF models achieved better performance than the LF variants, while the use of PCA and PLS transformations often improved the final results compared to the original input data.

Overall, the best performing configurations were obtained for the domain-knowledge-based PCA input, the Model-based PCA and PLS inputs, and the Lab-based PLS

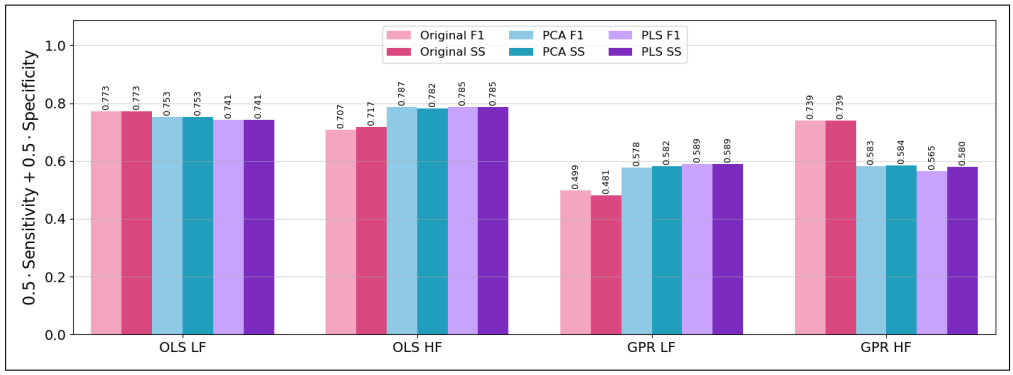

Figure 5.9: Comparison of 0.5·sensitivity+0.5·specificity values for prediction-based methods using model-based input data.

input configuration. These combinations achieved the highest values for both the F1-score and the combined sensitivity and specificity metric.

5.2.3 Cluster Based Methods

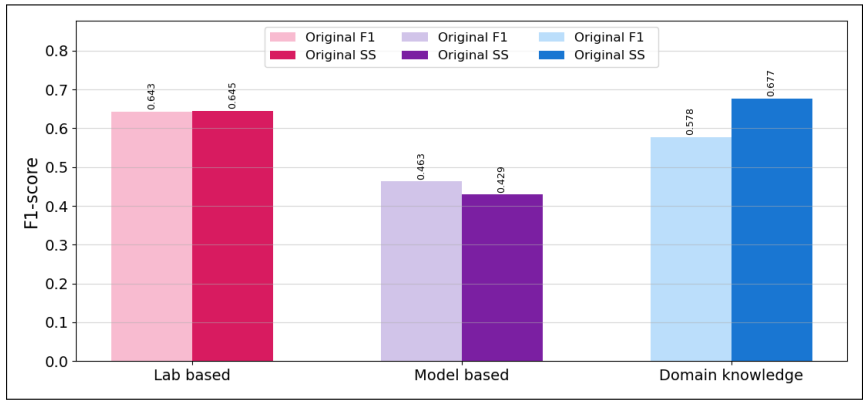

Figure 5.10: Comparison of F1-score values for the DBSCAN-based anomaly detection method using lab-based, model-based, and domain-knowledge-based input configurations.

Figures 5.10 and 5.11 present results of F1-score and combination of sensitivity and specificity for the DBSCAN method with three main input configurations. It is noticeable that DBSCAN-based anomaly detection method strongly depends on

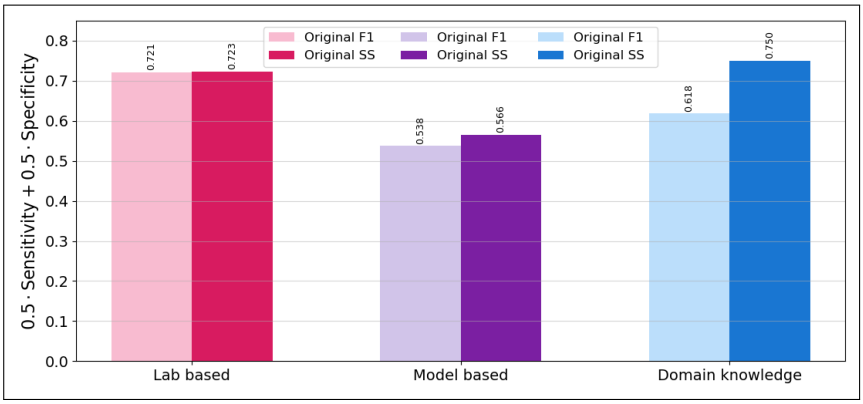

Figure 5.11: Comparison of 0.5·Sensitivity+0.5·Specificity values for the DBSCAN-based anomaly detection method using lab-based, model-based, and domain-knowledge-based input configurations.

the selected input configuration and optimization criterion. Among the evaluated input configurations, the model-based inputs achieved the lowest performance for both the F1-score and the combined sensitivity and specificity metric. The highest performance was obtained for the domain-knowledge-based configuration optimized using the combination of sensitivity and specificity.

5.2.4 Combination of Methods

Predictive Models with EWMA

As shown in Figure 5.3, the predicted signals still contain a relatively high amount of noise. Therefore, for the four best performing prediction-based models, an additional EWMA filter was applied to the prediction output in order to smooth the signal and potentially improve the anomaly detection performance. Figure 5.12 compares the ROC curves obtained without the EWMA filter and with the EWMA filter applied to the output signal. Using the ROC curves and the AUC metric, it is possible to evaluate whether the application of the EWMA filter improved the anomaly detection performance. The comparison indicates that the ROC curves after EWMA application are generally shifted closer to the upper left corner of the graph, which corresponds to better classification performance. This improvement can also be observed from the AUC metric values. Based on both the ROC curves and the AUC values, it can be concluded that the application of the EWMA filter improved the final results for all selected models.

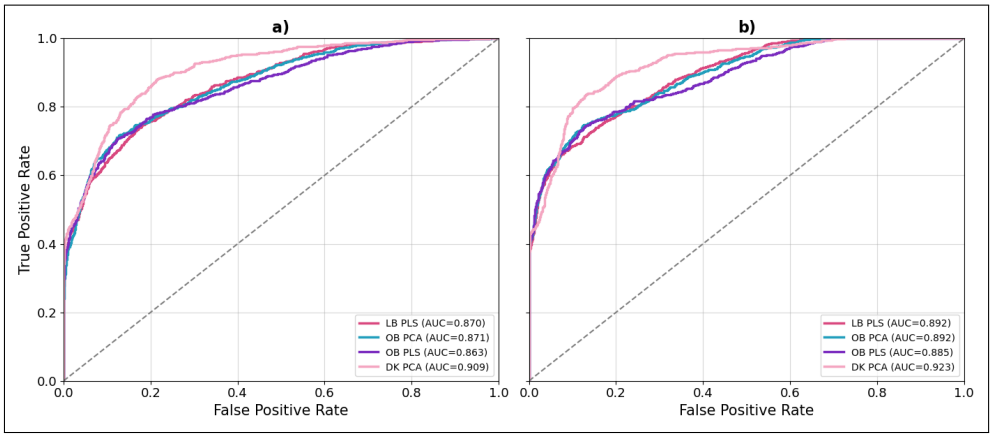

Figure 5.12: Comparison of ROC curves for the best prediction-based models: a) without EWMA filtering, b) with EWMA filtering applied to the output signal.

DBSCAN with Differenced Signal

The DBSCAN method was also evaluated using the differenced input signals instead of the original input data. The resulting F1-score values and the combined metric $0.5 \cdot \text{Sensitivity} + 0.5 \cdot \text{Specificity}$ are shown in Figures 5.13 and 5.14. As expected, the use of differenced input signals did not lead to an improvement in the anomaly detection performance. The difference method is mainly suitable for detecting point anomalies caused by sudden changes in the signal, but it is not able to effectively detect groups of anomalies or slowly developing anomalous behavior.

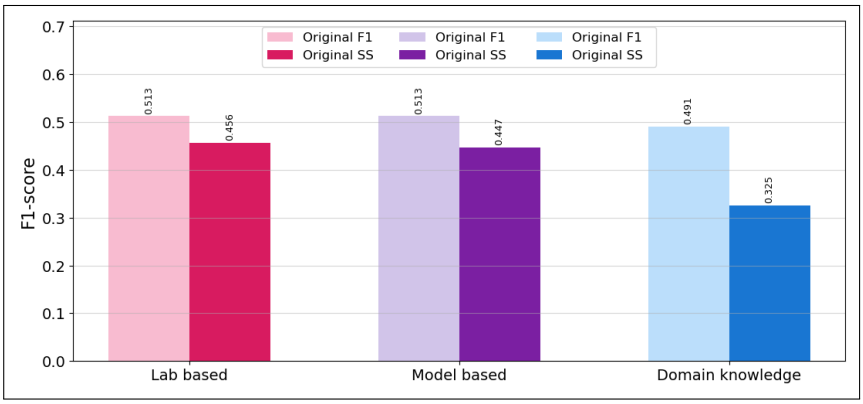

Figure 5.13: Comparison of F1-score values for the DBSCAN-based anomaly detection method combined with the differenced signal using lab-based, model-based, and domain-knowledge-based input configurations.

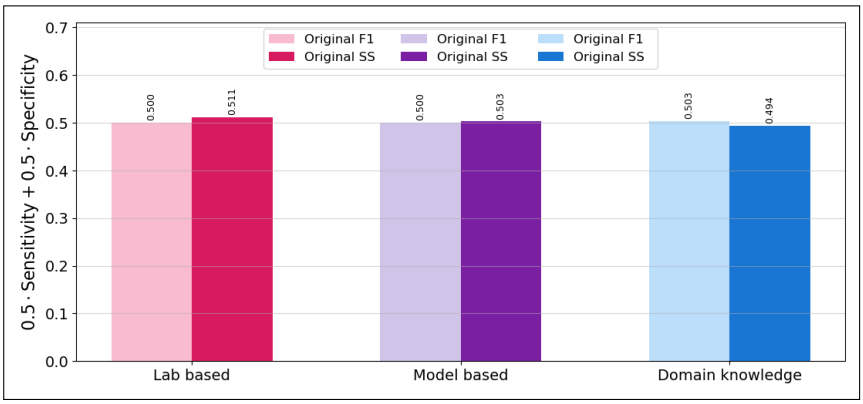

Figure 5.14: Comparison of 0.5·Sensitivity+0.5·Specificity values for the DBSCAN-based anomaly detection method combined with the differenced signal using lab-based, model-based, and domain-knowledge-based input configurations.

Chapter 6

Conclusions and Future Work

In this thesis, several anomaly detection methods for industrial time-series data were studied, implemented, and compared. The work focused on the analysis of data from a sulfuric acid alkylation process, where online sensor measurements and laboratory measurements with different fidelity levels were available. Since real anomaly labels were not provided, a multi-fidelity model was used to create a reference signal and generate pGT labels for further evaluation. The main focus of this work was the comparison of simple statistical methods with more advanced machine learning approaches for anomaly detection in industrial data. Different preprocessing techniques, sensor selection strategies, and dimensionality reduction methods were applied in order to improve model performance and reduce the influence of noisy industrial measurements. The work also examined how different types of input data and optimization criteria affect the final anomaly detection results.

6.1 Summary of Findings

The first investigated approaches were the output-based methods, namely the EWMA filter and the difference method. The EWMA method was able to identify several anomalous regions, but the obtained results indicate that it alone is not sufficient for reliable anomaly detection in complex industrial data. The difference method achieved the weakest results among all investigated methods, but that was expected since this method is mainly suitable for detecting point anomalies. Although the method was able to highlight abrupt changes in the signal, only a small number of anomalies were successfully detected.

Among all investigated approaches, prediction-based methods achieved the best overall performance. In general, the OLS regression models provided more stable and reliable results than the GPR models. The best overall performances were achieved by the

OLS HF models. For OLS HF prediction models, implementation of PCA and PLS generally improved the obtained results. On the other hand, prediction of LF signals showed several limitations. Large fluctuations of F1-score values could be observed for both OLS LF and GPR LF models across different input configurations. The most significant effect appeared for the domain knowledge-based GPR LF models. Since the selected domain-knowledge sensors describe the LF signal very accurately, the Gaussian Process model was able to fit the training data too closely, which likely caused overfitting. As a result, the optimized thresholds obtained during training were often extremely small. This behavior negatively affected the confusion matrix metrics on the testing dataset and reduced the generalization ability of the model.

Since it was noticeable that the predicted signals still contained a relatively large amount of noise, we decided to apply an EWMA smoothing filter to the outputs of selected prediction-based models. Based on the ROC analysis, the smoothing of predicted outputs using EWMA had a positive effect on anomaly detection performance. These results suggest that EWMA can be successfully used as a post-processing technique for prediction-based anomaly detection methods.

The last applied method was DBSCAN, which achieved better performance than the simple output-based methods, especially compared to the difference method. However, DBSCAN was not able to outperform the best prediction-based OLS models. One possible explanation is that DBSCAN is fundamentally an unsupervised method. Even though parameter optimization using evaluation metrics was applied, the optimization process was not sufficient to overcome the limitations of an unsupervised clustering approach compared to supervised prediction-based methods trained directly on reference outputs. We also decided to investigate an additional combination of DBSCAN and differenced inputs. However, this approach did not provide satisfactory results. The obtained behavior was expected because the difference method is more suitable for highlighting abrupt changes in the signal than for creating a stable feature space for clustering-based methods.

Overall, the obtained results confirmed that combining industrial process knowledge, dimensionality reduction techniques, and supervised prediction-based methods can significantly improve anomaly detection performance in industrial time-series data.

6.2 Future Research Directions

In the future, we would like to compare the performance of the domain knowledge-based sensor selection approach with additional feature selection methods, such as Least

Absolute Shrinkage and Selection Operator (LASSO).

Another interesting direction for future work would be to investigate additional anomaly detection methods, such as Isolation Forest and Random Forest algorithms. We would also like to investigate the implementation of neural networks and deep learning methods. We could train a neural network to learn the normal behavior of industrial signals and automatically identify situations where the signal behavior significantly changes.

Appendix A

Resumé

V súčasnosti priemyselné procesy [8] produkujú veľké množstvo časových dát z online senzorov, analyzátorov a laboratórnych meraní. Tieto dáta sú dôležité pre monitorovanie procesu, hodnotenie jeho stability a včasné odhalenie neštandardného správania. V praxi však merania často obsahujú šum, chýbajúce hodnoty, nepresnosti alebo poruchy meracích zariadení. Takéto problémy môžu spôsobiť nesprávnu interpretáciu stavu procesu a následne viesť k zníženiu kvality produktu alebo bezpečnostným rizikám. Z tohto dôvodu je detekcia anomálií dôležitou súčasťou moderného priemyselného monitorovania. Anomálie [7] predstavujú vzory alebo pozorovania v dátach, ktoré sú určitým spôsobom neobvyklé a nezhodujú sa so všeobecným správaním dát.

Metódy detekcie anomálií možno rozdeliť podľa dostupnosti označených dát na metódy s učiteľom, bez učiteľa a s čiastočným dohľadom. Pri metódach s učiteľom sú dostupné označenia normálnych a anomálnych bodov, podľa ktorých sa model učí rozlišovať medzi triedami. V priemyselných aplikáciách však takéto označenia často chýbajú, pretože anomálie sú zriedkavé a ich označovanie si vyžaduje odborné znalosti. Metódy bez učiteľa pracujú iba s neoznačenými dátami a snažia sa nájsť odlišné správanie priamo zo štruktúry dát. Metódy s čiastočným dohľadom kombinujú malé množstvo označených dát s väčším množstvom neoznačených dát [20].

Pred samotnou detekciou anomálií je potrebné dáta vhodne rozdeliť na trénovaciu, validačnú a testovaciu sadu. Trénovacia množina sa používa na nastavenie modelov a odhad ich parametrov. Validačná množina slúži počas vývoja modelu na výber vhodných hyperparametrov, porovnanie rôznych nastavení modelu a zníženie rizika preučenia. Testovacia množina sa používa až na finálne overenie výkonnosti na dátach, ktoré model počas nastavovania nevidel [3, 14]. Ďalšou časťou prípravy dát je čistenie. Tento krok slúži na odstránenie nesprávnych, neúplných alebo nespoľahlivých meraní [24]. Dôležitou súčasťou predspracovania dát je aj štandardizácia. Pri štandardizácii sa každá premenná transformuje tak, aby mala nulovú strednú hodnotu a jednotkovú smerodajnú odchýlku [23, 14]. Tento krok je dôležitý najmä pri metódach, ktoré sú

citlivé na mierku vstupných premenných, napríklad pri regresných alebo zhlukovacích metódach.

Pri práci s veľkým počtom senzorov je dôležitý aj výber vstupných premenných. Jedným z jednoduchých prístupov je korelačný výber príznakov [16]. Korelácia vyjadruje mieru vzťahu medzi dvoma premennými a môže pomôcť identifikovať senzory, ktoré sú silne spojené s výstupnou veličinou. Zároveň môže pomôcť odhaliť redundantné senzory, ktoré nesú veľmi podobnú informáciu. Okrem štatistických metód môže byť pri výbere senzorov dôležitá aj doménová znalosť. Odborníci na proces vedia určiť, ktoré premenné majú fyzikálny alebo technologický význam [1]. Po výbere príznakov je možné použiť metódy redukcie rozmernosti. PCA je metóda bez učiteľa, ktorá transformuje pôvodné premenné na nové nekorelované komponenty tak, aby prvé komponenty zachytili čo najväčšiu variabilitu dát [23]. PLS je podobná metóda, ale na rozdiel od PCA zohľadňuje aj vzťah medzi vstupnými premennými a výstupnou veličinou. Preto je vhodná najmä pre predikčné úlohy, kde je cieľom nájsť také latentné premenné, ktoré čo najlepšie opisujú výstupný signál [19].

V práci boli porovnávané tri hlavné skupiny metód detekcie anomálií. Prvú skupinu tvorili metódy založené na spracovaní výstupného signálu. Patrí sem exponenciálne vážený kĺzavý priemer (EWMA) a diferenčná metóda. EWMA je metóda vyhladzovania časových radov, ktorá počíta vážený priemer aktuálnej a predchádzajúcich hodnôt, pričom novšie hodnoty majú väčší vplyv ako staršie hodnoty [15]. Diferenčná metóda je založená na výpočte rozdielu medzi dvoma po sebe idúcimi hodnotami signálu. Je vhodná najmä na zvýraznenie náhlych zmien alebo skokov v dátach.

Druhú skupinu tvorili predikčné metódy. Ich základnou myšlienkou je vytvoriť model očakávaného správania výstupného signálu a následne označiť ako anomálie tie body, kde sa skutočné meranie výrazne líši od predikcie. V práci boli použité dva regresné prístupy: metóda obyčajných najmenších štvorcov (OLS) a Gaussovská procesná regresia (GPR). OLS regresia hľadá lineárny vzťah medzi vstupnými premennými a výstupom minimalizáciou súčtu štvorcov rozdielov medzi pozorovanými a predikovanými hodnotami [11]. GPR je flexibilnejšia nelineárna metóda, ktorá modeluje rozdelenie nad možnými funkciami a používa jadrové funkcie na opis podobnosti medzi dátovými bodmi [25].

Tretiu skupinu tvorila zhlukovacia metóda DBSCAN [9], teda hustotne založené zhlukovanie aplikácií s prítomnosťou šumu. Ide o algoritmus, ktorý vytvára zhluky v oblastiach s vysokou hustotou bodov a body v oblastiach s nízkou hustotou označuje ako šum. Algoritmus je riadený dvoma hlavnými parametrami: `eps`, ktorý určuje polomer okolia bodu, a `min_samples`, ktorý určuje minimálny počet bodov potrebných

na vytvorenie hustej oblasti.

Výkonnosť jednotlivých metód bola hodnotená pomocou matice zámeny. Tá rozdeľuje výsledky klasifikácie na správne detegované anomálie, správne detegované normálne body, falošné alarmy a vynechané anomálie. Z týchto hodnôt boli vypočítané metriky ako presnosť, citlivosť, špecificita a F1-skóre [26, 5]. Okrem toho boli pre vybrané modely použité krivky operačnej charakteristiky prijímača (ROC krivky) a plocha pod ROC krivkou (AUC), ktoré umožňujú hodnotiť klasifikačnú schopnosť modelu pri rôznych prahových hodnotách [27, 13].

Parametre jednotlivých metód boli optimalizované pomocou Nelderovho-Meadovho algoritmu. Ide o priamu optimalizačnú metódu, ktorá nevyžaduje výpočet derivácií a pracuje so simplexom [22].

Praktická časť práce bola realizovaná na anonymizovaných dátach zo spoločnosti Slovnaft, a.s., ktoré pochádzajú z procesu alkylácie kyselinou sírovou. V práci boli použité dáta z 731 senzorov počas 10 000 časových krokov. Online merania boli dostupné každých 11 minút. Výstupný signál predstavoval koncentráciu recyklovaného izobutánu. Laboratórne merania tej istej veličiny boli dostupné iba raz mesačne, spolu v počte 11 meraní. Keďže laboratórne hodnoty sú považované za spoľahlivejšie, rozdiely medzi online signálom a laboratórnymi meraniami naznačovali možné anomálie v online meraní.

Keďže skutočné overené označenia anomálií neboli dostupné, v práci boli použité pseudo-referenčné označenia anomálií odvodené z modelu s viacerými úrovňami presnosti. Tento model bol navrhnutý v práci Fáber et al. a slúži na kombináciu online dát s nižšou presnosťou a laboratórnych dát s vyššou presnosťou [6]. Výsledkom je korigovaný referenčný signál, ktorý bol v tejto práci použitý ako základ pre vytvorenie pseudo-referenčných anomálií, ktoré boli vytvorené porovnaním pôvodného online výstupného signálu s referenčným signálom modelu s viacerými úrovňami presnosti.

Dostupné dáta boli rozdelené chronologicky na trénovaciu a testovaciu množinu v pomere 70 % ku 30 %. Trénovacia množina mala rozmery 7000×731 a testovacia množina mala rozmery 3000×731. Samostatná validačná množina nebola použitá z dôvodu obmedzenej dĺžky dostupných dát a malého počtu laboratórnych meraní. Všetky parametre predspracovania, kritériá výberu senzorov, transformácie PCA a PLS, parametre modelov a detekčné prahy boli určené iba na trénovacej množine. Testovacia množina bola následne transformovaná pomocou parametrov nastavených na trénovacej množine.

V rámci čistenia dát boli odstránené senzory s viac ako 80 % chýbajúcich hodnôt. Chýbajúce hodnoty v zostávajúcich signáloch boli doplnené interpoláciou. Ďalej boli odstránené senzory s takmer lineárnym priebehom, senzory s dlhodobo konštantnou hodnotou a senzory s výraznými náhlymi skokmi. Následne boli vytvorené tri skupiny vstupných premenných. Prvá skupina bola založená na korelácii medzi senzormi a laboratórnymi meraniami. Druhá skupina bola založená na korelácii medzi senzormi a výstupom modelu s viacerými úrovňami presnosti. Tretia skupina bola vytvorená na základe doménových znalostí o procese. Tieto skupiny senzorov boli následne použité buď priamo, alebo boli transformované pomocou PCA a PLS.

EWMA metóda bola v implementácii aplikovaná priamo na online výstupný signál. Najprv bol na trénovacej množine vypočítaný vyhladený signál a boli optimalizované parameter vyhladzovania a detekčný prah. Následne bola rovnaká metóda použitá aj na testovacej množine. Anomálie boli detegované porovnaním pôvodného signálu s vyhladeným signálom. Diferenčná metóda bola aplikovaná takisto priamo na výstupný signál, pričom sa počítali rozdiely medzi po sebe idúcimi meraniami. Keďže týmto výpočtom vznikne signál kratší o jeden bod, pseudo-referenčné označenia boli upravené odstránením prvého prvku, aby zostalo zachované správne časové zarovnanie.

Pre predikčné metódy bolo testovaných deväť vstupných konfigurácií. Tieto konfigurácie vznikli kombináciou troch prístupov výberu senzorov, teda lab-based, model-based a domain-knowledge-based výberu, s tromi reprezentáciami vstupov: pôvodné vstupy, vstupy transformované pomocou PCA a vstupy transformované pomocou PLS. Modely OLS a GPR boli trénované buď na LF online výstupnom signáli, alebo na HF laboratórnych meraniach. Po natrénovaní modelu boli anomálie detegované na základe rozdielu medzi meraným výstupom a predikovaným výstupom. Ak tento rozdiel prekročil optimalizovaný prah, bod bol označený ako anomália.

Pri DBSCANe boli použité iba pôvodné vstupné konfigurácie bez PCA a PLS transformácií. Na rozdiel od OLS a GPR DBSCAN nepredikuje výstupný signál, ale hľadá anomálie priamo v štruktúre vstupných dát na základe hustoty bodov. Body označené algoritmom ako šum boli považované za anomálie. Okrem toho bola použitá dodatočná podmienka, podľa ktorej bol celý zhluk označený ako anomálny, ak obsahoval viac ako 51 % pseudo-referenčných anomálií.

Okrem samostatných metód boli skúmané aj kombinácie metód. Prvou kombináciou boli predikčné modely s EWMA vyhladzovaním. Táto analýza bola vykonaná iba pre vybrané predikčné modely s najlepšou výkonnosťou. Cieľom bolo zistiť, či vyhladenie predikcie zlepší separáciu medzi normálnym a anomálnym správaním. Druhou kombináciou bol DBSCAN aplikovaný na diferencované vstupné signály.

Optimalizácia parametrov bola realizovaná na trénovacej množine. Pre EWMA metódu boli optimalizované parameter vyhladzovania a detekčný prah. Pre DBSCAN boli optimalizované parametre `eps` a `min_samples`. Pre ostatné metódy bol optimalizovaný iba detekčný prah. Po optimalizácii boli vybrané parametre aplikované na testovaciu množinu bez ďalšieho upravovania.

Výsledky ukázali, že jednoduché metódy založené na spracovaní výstupného signálu majú obmedzenú výkonnosť. EWMA metóda dokázala zachytiť časť anomálnych oblastí, ale ako samostatná metóda nebola dostatočne spoľahlivá. Diferenčná metóda dosiahla veľmi nízke F1-skóre 0.020 a kombinovanú metriku 0.496. Tento výsledok bol očakávaný, pretože diferenčná metóda je vhodná najmä na detekciu náhlych bodových zmien, ale nedokáže dobre zachytiť dlhšie alebo postupne sa vyvíjajúce anomálie.

Najlepšie výsledky dosiahli predikčné metódy, najmä modely OLS trénované na laboratórnych HF dátach. Vo všeobecnosti sa ukázalo, že HF cieľový signál poskytuje vhodnejší referenčný základ pre detekciu anomálií ako LF online signál. Pri LF modeloch, najmä pri GPR, sa v niektorých konfiguráciách objavili veľmi malé optimalizované prahy, čo naznačuje, že model dokázal trénovacie dáta prispôsobiť príliš presne. To mohlo viesť k horšej schopnosti zovšeobecnenia na testovacej množine. Použitie PCA a PLS často zlepšilo výsledky predikčných modelov, čo potvrdzuje význam redukcie rozmernosti pri práci s veľkým počtom senzorov. Najlepšie konfigurácie boli dosiahnuté pri vstupoch založených na doménových znalostiach transformovaných pomocou PCA, pri modelovo založených vstupoch transformovaných pomocou PCA a PLS a pri laboratórne založených vstupoch transformovaných pomocou PLS.

Keďže predikované signály stále obsahovali krátkodobý šum, pre štyri najlepšie predikčné modely bol dodatočne aplikovaný EWMA filter na predikovaný výstup. Porovnanie ROC kriviek pred a po vyhladení ukázalo, že EWMA filter zlepšil klasifikačnú schopnosť vybraných modelov. ROC krivky sa po aplikácii EWMA posunuli bližšie k ľavému hornému rohu grafu a zlepšili sa aj hodnoty AUC. To naznačuje, že EWMA je vhodná najmä ako doplnková postprocesná metóda pre predikčné prístupy.

DBSCAN dosiahol lepšie výsledky ako jednoduché výstupné metódy, najmä v porovnaní s diferenčnou metódou, ale nedokázal prekonať najlepšie predikčné modely. Výkonnosť DBSCANu výrazne závisela od výberu vstupných premenných a optimalizačného kritéria. Ako dodatočný experiment bol DBSCAN aplikovaný aj na diferencované vstupné signály. Tento prístup však neviedol k zlepšeniu. Zvýšenie parametrov `eps` a `min_samples` naznačilo, že diferencované vstupy sú v priestore príznakov viac rozptýlené a na vytvorenie zhlukov vyžadujú väčšie okolie aj prísnejšiu hustotnú podmienku. Diferencovanie teda síce zvýrazňuje náhle lokálne zmeny, ale narúša

štruktúru dát potrebnú na stabilné zhlukovanie.

Na základe dosiahnutých výsledkov možno konštatovať, že pre analyzovaný priemyselný proces sú najvhodnejšie predikčné metódy, najmä OLS modely trénované na HF laboratórnych meraniach v kombinácii s vhodným výberom vstupov a redukciou rozmernosti. Jednoduché metódy založené na výstupnom signáli môžu byť užitočné ako pomocné alebo doplnkové nástroje, ale samostatne neposkytujú dostatočnú výkonnosť.

Bibliography

[1] A. B. Bada, A. B. Garko, D. Gabi, and M. S. Argungu. The role of domain knowledge in feature selection for machine learning, 2025.

[2] S. Bauskar. Enhancing system observability with machine learning techniques for anomaly detection, 2024.

[3] N. Buhl. Training, validation, test split for machine learning datasets, 2024.

[4] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41:1–58, 2009.

[5] A. de Giorgio, G. Cola, and L. Wang. Systematic review of class imbalance problems in manufacturing. *Journal of Manufacturing Systems*, 71:620–644, 2023.

[6] R. Fáber, M. Vaccari, R. Bacci di Capaci, K. Ľubušký, G. Pannocchia, and R. Paulen. Soft-sensor-enhanced monitoring of an alkylation unit via multi-fidelity model correction. 2025.

[7] R. Foorthuis. On the nature and types of anomalies: A review of deviations in data. *International Journal of Data Science and Analytics*, 12, 2021.

[8] X. Kong, X. Jiang, B. Zhang, J. Yuan, and Z. Ge. Latent variable models in the era of industrial big data: Extension and beyond. *Annual Reviews in Control*, 54:167–199, 2022.

[9] J. Liu, H. Qin, Z. Liu, S. Wang, Q. Zhang, and Z. He. A density-based spatial clustering of application with noise algorithm and its empirical research. *Highlights in Science, Engineering and Technology*, 7:174–179, 2022.

[10] M. Ma, L. Han, and C. Zhou. Btad: A binary transformer deep neural network model for anomaly detection in multivariate time series data. *Advanced Engineering Informatics*, 56:101949, 2023.

[11] M. Majka. Ordinary least squares, 2024.

[12] Z. Mehmood and Z. Wang. Hybrid iforest-dbscan for anomaly detection and wind power curve modelling. *Expert Systems with Applications*, 289:128381, 2025.

[13] F. S. Nahm. Receiver operating characteristic curve: Overview and practical use for clinicians. *Korean Journal of Anesthesiology*, 75(1):25–36, 2022.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[15] M. Perry. The exponentially weighted moving average. In *Wiley Encyclopedia of Operations Research and Management Science*. 2010.

[16] T. Pham-Gia and V. Choulakian. Distribution of the sample correlation matrix and applications. *Open Journal of Statistics*, pages 330–344, 2014.

[17] s. Joe Qin. Survey on data-driven industrial process monitoring and diagnosis. *Annual Reviews in Control*, 36:220–234, 12 2012.

[18] M. S. Reis and G. Gins. Industrial process monitoring in the big data/industry 4.0 era: From detection, to diagnosis, to prognosis. *Processes*, 5(3), 2017.

[19] R. Rosipal and N. Krämer. Overview and recent advances in partial least squares. In *Subspace, Latent Structure and Feature Selection*, pages 34–51. 2006.

[20] L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, and G. Montavon. A unifying review of deep and shallow anomaly detection. 2021.

[21] H. B. Salah, P. Nancarrow, and A. Al-Othman. Ionic liquid-assisted refinery processes – a review and industrial perspective. *Fuel*, 302:121195, 2021.

[22] S. Singer and J. Nelder. Nelder-mead algorithm. *Scholarpedia*, 4(7):2928, 2009. Revision #91557.

[23] A. Subasi. Data preprocessing. In *Practical Machine Learning for Data Analysis Using Python*, pages 27–89. 2020.

[24] Tableau. Guide to data cleaning: Definition, benefits, components, and how to clean your data.

[25] J. Wang. An intuitive tutorial to gaussian processes regression, 2020.

[26] D. Wei. Essential math for machine learning: Confusion matrix, accuracy, precision, recall, f1-score, 2024.

[27] S. Yang and G. Berdine. The receiver operating characteristic curve. *The Southwest Respiratory and Critical Care Chronicles*, 5:34, 2017.

[28] Q. Zhang. Financial data anomaly detection method based on decision tree and random forest algorithm. *Journal of Mathematics*, 2022(1):9135117, 2022.