Degree project

UNIVERSITY OF SKÖVDE

A Data-Driven Approach for Detecting and Understanding Abnormal Process Behaviour in Industrial Ring Rolling

Master degree Project in Informatics with a specialization in Data Science

Second Cycle 15 credits

Spring Term 2026

Student: Njoud Alshedah

Supervisor: Yacine Atif

Examiner: Juhee Bae

Abstract

This study investigates data-driven approaches to abnormal process behaviour detection in an industrial ring-rolling process. A threshold-based classification rule and three supervised machine learning models were evaluated using high-frequency process measurements and product-quality assessments collected across six predefined production phases.

Exploratory analysis was conducted to compare descriptive signal patterns between normal and abnormal observations across the phases. Based on these comparisons, Phase 3 was selected for closer examination and threshold-based analysis; however, no formal statistical comparison established it as the most informative phase. Logistic Regression, Random Forest, and XGBoost were evaluated separately using H_pressure, H_speed, and H_output as input features, without production phase as an explicit model input. Random Forest achieved the highest overall accuracy among the initial evaluations, whereas XGBoost achieved higher abnormal-class recall. Applying a threshold of 0.3 to the XGBoost test-set predictions resulted in recall of 0.99, precision of 0.17, and accuracy of 0.38 for abnormal-behaviour detection, demonstrating high sensitivity at the cost of a high false-positive rate.

The study therefore presents a phase-informed analytical approach rather than a phase-aware monitoring framework. The threshold-based component provides a directly understandable rule, while no dedicated explainability analysis was conducted for the supervised models.

Keywords: Abnormal Behaviour Detection, Machine Learning, Industrial Process Data, Exploratory Statistical Analysis, Phase-Informed Analysis, Industrial Process Monitoring, Ring-Rolling Process.

Table of contents

Abstract 0
1.Introduction 1
1.1 Problem definition and Research Question 2
1.2 Research Objectives 2
1.3 Contributions of the Study 3
2 Background and Related Work 3
2.1 Industrial Process and Monitoring Context 4
2.2 Industrial Anomaly Detection 5
2.3 Machine Learning for Industrial Monitoring 6
2.4 Context-Specific Research Motivation 7
3.Methodology and Implementation 7
3.1 Research Method 7
3.2 Alternative Research Method 8
3.3 Dataset Description 9
3.4 Feature Construction 9
3.5 Production Phase Analysis 10
3.6 Threshold-Based Baseline Classifier 10
3.7 Supervised Machine Learning Models 11
3.7.1 Logistic Regression (Baseline Model) 12
3.7.2 Random Forest (Tree-Based Model) 12
3.7.3 XGBoost (Boosted Model) 12
3.7.4 Phase-Informed Classification Approach 13
3.8 Experimental Considerations 13
4.Experimental Analysis and Results 14
4.1 Exploratory Statistical Analysis 14
4.1.1 Phase-Based Analysis Results 14
4.1.2 Feature Analysis Results 15
4.1.3 Threshold Derivation 17
4.1.4 Pressure-Based Threshold Classification Results 18
4.1.5 multi-signal Threshold-Based Classification Results 18
4.2 Machine Learning Results 19

4.2.1 Logistic Regression 19
4.2.2 Random Forest 19
4.2.3 XGBoost Model 20
4.3 Comparative Model Evaluation 20
4.3.1 ROC Curve Analysis 20
4.3.2 Precision–Recall Analysis 21
4.3.3 Exploratory XGBoost Threshold Analysis 22
4.3.4 Scope of the Recorded XGBoost Output 22
4.3.5 Comparison Between Threshold-Based Classification and Machine Learning Models 23
5.Discussion 23
5.1 Interpretation of Findings 24
5.2. Ethical, Societal, and Sustainability Aspects 25
5.3 Study Limitations and Generalisability 27
6.Conclusion and Future Work 28
6.1 Conclusion 28
6.2 Future Work 29

1.Introduction

Industrial manufacturing systems generate large volumes of process data through sensors and monitoring systems, providing valuable opportunities to improve process monitoring, abnormal behaviour detection, and operational decision-making (Braei & Wagner, 2020). Analysing these data enables manufacturers to better understand process behaviour, identify deviations from normal operation, and support more reliable production systems. As manufacturing processes become increasingly automated and data-rich, data-driven monitoring has become an important component of modern industrial operations.

A major challenge in industrial process monitoring is that process variables are highly interconnected. Small deviations in one variable may influence other process variables and ultimately affect the quality of the final product. In addition, industrial processes often exhibit complex and non-linear relationships, making abnormal behaviour difficult to detect using conventional monitoring techniques alone (Kang & Lee, 2024; Qin, 2012).

To address these challenges, a variety of monitoring approaches have been proposed. Traditional statistical methods and threshold-based monitoring approaches provide interpretable results and remain widely adopted in industrial environments. However, they often have limited capability to capture complex relationships among multiple process variables. Consequently, supervised machine learning techniques have received increasing attention because they can learn patterns directly from industrial process data and improve abnormal behaviour detection performance (Pang et al., 2021; Yan et al., 2023).

Although machine learning has demonstrated promising results for industrial monitoring, exploratory statistical analysis remains an important part of the Data Science workflow to help identify informative process characteristics, examine differences between production stages, and inform feature selection before predictive modelling. Combining exploratory statistical analysis for process understanding with supervised machine learning can therefore improve both model interpretability and decision support in industrial applications (Chalapathy & Chawla, 2019; Pang et al., 2021).

This thesis investigates abnormal process behaviour in a ring rolling process at Ovako AB, where horizontal pressure, rolling speed, and control output are continuously monitored throughout production. The manufacturing process consists of six production phases, each representing a distinct stage of the rolling operation. Examining process behaviour across these production phases provides valuable insight into how descriptive signal patterns differ between normal and abnormal observations and supports the selection of features for subsequent predictive modelling.Within this industrial context, the research adopts a comparative empirical evaluation design to assess different approaches for abnormal behaviour detection. Exploratory statistical analysis is first used to investigate process behaviour and identify informative production phases and process variables. The resulting findings inform the subsequent predictive modelling stage, where a threshold-based classification approach is compared with supervised machine learning models. Through this framework, the study aims to support interpretable and data-driven industrial process monitoring for abnormal behaviour detection.

1.1 Problem definition and Research Question

The ring-rolling process is a dynamic industrial manufacturing system in which hydraulic pressure, rolling speed, and control output vary continuously throughout production. These process signals are associated with the deformation behaviour of the ring and the resulting product geometry. Their interdependencies make it challenging to characterise how the observed signal patterns relate to abnormal process behaviour.

A key challenge in industrial process monitoring is that the final production outcome alone does not describe how process-signal patterns vary across different parts of the production process. Although classification models can distinguish between normal and abnormal observations, their predictions do not by themselves explain the phase-level patterns observed in the data or the contribution of individual input variables. This motivates a combined analysis in which exploratory comparisons across predefined production phases are used to contextualise a subsequent evaluation of supervised classification models.

This study therefore aims to examine descriptive differences between normal and abnormal observations across predefined production phases and to evaluate the use of selected process signals for abnormal-behaviour classification. Three process signals are investigated: horizontal pressure, which represents the deformation forces applied during ring rolling, together with rolling speed and control output, which represent control-system responses throughout production. The exploratory phase-level comparisons are used to inform the focus of the subsequent threshold-based analysis and the selection of input variables for the supervised models.

The study is driven by the following main research question:

Main RQ:

How can a phase-informed data-driven analysis approach support abnormal process behaviour detection in an industrial ring-rolling process?

To investigate the main research question, three supporting research questions are formulated:

RQ1:

What descriptive differences between normal and abnormal observations are identified across the predefined production phases, and how do these findings inform the selection of a phase for closer analysis?

RQ2:
Which process signals are most strongly associated with abnormal process behaviour within the identified production phase?

RQ3:
How effectively can a threshold-based classification approach and supervised machine learning models detect abnormal process behaviour using the identified process signals?

1.2 Research Objectives

The objective of this thesis is to apply a phase-informed data-driven approach to the analysis and classification of abnormal process behaviour in an industrial ring-rolling process. To achieve this objective, the study first examines descriptive differences between normal and abnormal

observations across predefined production phases and uses these exploratory findings to select a phase for closer analysis. It then investigates class-specific patterns in three process signals horizontal pressure, rolling speed, and control output and uses these signals in the subsequent classification analysis. A threshold-based classification rule and three supervised machine learning models are evaluated using precision, recall, F1-score, accuracy, and ROC-AUC where applicable. Particular attention is given to abnormal-class recall while the accompanying precision and accuracy are considered when interpreting the cost of prioritising sensitivity. The study thereby provides a context-specific empirical evaluation of phase-level exploratory analysis and abnormal-behaviour classification rather than proposing a new phase-aware monitoring framework.

1.3 Contributions of the Study

This study contributes a context-specific empirical analysis of abnormal process behaviour in an industrial ring-rolling process. Exploratory comparisons across predefined production phases were used to examine descriptive differences between normal and abnormal observations and to inform the selection of Phase 3 for closer threshold-based analysis. The study also examined three process signals H_pressure, H_speed, and H_outputand subsequently used them as input features in the supervised classification models. In the fitted Random Forest model, H_pressure received the highest model-derived feature-importance value among these three variables; however, this result is interpreted as specific to the evaluated dataset and model configuration.

The study further provides a comparative evaluation of a threshold-based classification rule and three supervised machine learning models: Logistic Regression, Random Forest, and XGBoost. The evaluation reports precision, recall, F1-score, accuracy, and ROC-AUC where applicable and demonstrates the trade-off between abnormal-class sensitivity and false-positive outcomes. The threshold-based component provides a directly understandable decision rule, whereas no dedicated explainability analysis was conducted for the supervised models. Accordingly, the contribution is presented as a phase-informed analytical approach and comparative empirical evaluation rather than as a new phase-aware or operationally validated monitoring framework.

2 Background and Related Work

The ring rolling process is a thermo-mechanical forming process in which a heated steel workpiece is gradually formed into a seamless ring through the coordinated movement of multiple machine components. Horizontal movement controls the ring diameter, while vertical movement controls the ring height. Additional components, including support arms and positioning systems, contribute to process stability, and the entire system must operate in a coordinated manner to ensure dimensional accuracy. As illustrated in Figure 2.1, the process is governed by five coordinated mechanical movements: horizontal movement, vertical movement, support arm out, support arm in, and tower movement.

2.1 Industrial Process and Monitoring Context

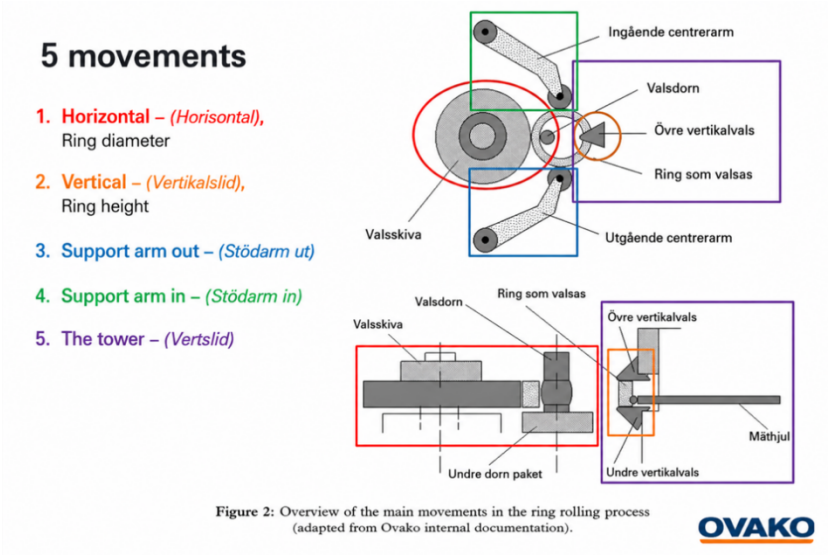

Figure 2.1: Main movements in the ring rolling process.
Source: Adapted from Ovako internal documentation.

From a data perspective, the process is monitored using several process signals that describe the operating conditions of the system throughout production. The main signals analysed in this study are horizontal pressure, rolling speed, and control output. Horizontal pressure represents the applied pressure on the ring and is calculated as the difference between the positive and negative pressure readings. This signal directly reflects the physical deformation occurring during the ring rolling process.

In contrast, the output and speed signals are more closely related to the behaviour of the control system. The output signal reflects how the system responds to process changes or deviations, while rolling speed is continuously adjusted to maintain process stability. This distinction separates signals that primarily represent the physical deformation process, such as pressure, from signals that primarily represent control-system responses, such as output and speed.

Understanding this distinction is important for abnormal process behaviour analysis because deviations in pressure are more likely to indicate process instability, whereas variations in speed and output may reflect the system's response to changing operating conditions rather than the underlying cause of abnormal behaviour. For analytical purposes, the production process is divided into six production phases, each representing a distinct stage of the ring rolling operation. The characteristics of these production phases are summarised in Table 2.1

Table 2.1: Description of the production phases in the ring rolling process.

Phase	Description	Process Role
Phase 1	Initialization stage	System startup and initial stabilization
Phase 2	Adjustment stage	Process alignment and operational regulation
Phase 3	Main deformation stage	Critical deformation stage where instability becomes more visible
Phase 4	Intermediate stabilization	System balancing during continued deformation
Phase 5	Final deformation adjustments	Final shaping and process correction
Phase 6	Process completion stage	Ending phase and stabilization before process completion

Understanding the differences between these production phases is essential for identifying where abnormal process behaviour becomes most distinguishable during production. Phase 3 represents the primary deformation stage, during which the process is most susceptible to instability, according to the company's operational documentation.

For industrial process monitoring, it is important to distinguish between the underlying physical process and the corresponding control-system response. For example, a change in the output signal does not necessarily indicate that the output itself is the source of the abnormal behaviour. Instead, it often reflects the system's response to a deviation that occurred earlier in the manufacturing process. Making this distinction supports a more accurate interpretation of abnormal process behaviour by helping identify the underlying source of instability rather than its downstream effects.

2.2 Industrial Anomaly Detection

Threshold-based and rule-based approaches are commonly described in the industrial monitoring literature because they provide transparent and interpretable decision rules. These approaches are typically based on expert-defined thresholds and process knowledge and are

widely used as baseline techniques for abnormal behaviour detection (Qin, 2012; Chandola et al., 2009). However, industrial systems are often complex, and the strong relationships among process variables can make it difficult to clearly distinguish between normal and abnormal behaviour (Qin, 2012; Chandola et al., 2009).

Traditional industrial monitoring approaches have primarily relied on statistical measures such as the mean, variance, and standard deviation. These techniques are straightforward to implement and effective for describing overall process behaviour; however, they do not fully capture the complex and nonlinear relationships among process variables (Qin, 2012). This challenge becomes even greater in manufacturing systems where multiple process signals interact throughout production (Gupta et al., 2014).

Modern industrial processes generate multiple process variables simultaneously, creating multivariate monitoring challenges. Unlike univariate monitoring, where each process signal is analysed independently, multivariate monitoring considers the relationships and dependencies among multiple variables. This is particularly important in industrial environments, where abnormal process behaviour may not be evident from a single signal but instead becomes distinguishable through the combined behaviour of multiple process variables. Consequently, multivariate monitoring has become an important component of industrial anomaly detection.

2.3 Machine Learning for Industrial Monitoring

Machine learning and deep learning algorithms have been increasingly applied to address the complexity of industrial anomaly detection (Yan et al., 2023; Pang et al., 2021). These techniques can identify complex patterns in multivariate industrial process data and detect abnormal behaviour more effectively than traditional statistical approaches. However, many machine learning algorithms have limited interpretability because they do not explicitly explain how their predictions are generated (Pang et al., 2021; Chalapathy & Chawla, 2019). This limitation is particularly important in industrial environments, where operators require both accurate predictions and interpretable results to support operational decision-making.

For this reason, recent research in industrial monitoring has increasingly focused on combining predictive performance with process understanding. Studies have highlighted the importance of analysing process behaviour and relationships among process signals to improve the interpretation of monitoring results and support more reliable industrial decision-making (Qin, 2012; Blázquez-García et al., 2021).

Feature importance analysis is widely used in industrial machine learning applications to identify which process variables contribute most to predictive performance. Understanding feature importance improves model interpretability by helping engineers relate model predictions to the underlying physical process. Machine learning algorithms such as Logistic Regression, Random Forest, and XGBoost have demonstrated strong predictive performance in industrial fault detection and process monitoring applications (Breiman, 2001; Chen & Guestrin, 2016; Pang et al., 2021). Consequently, recent research increasingly combines interpretable threshold-based approaches with supervised machine learning techniques to balance transparency and predictive capability in industrial monitoring systems.

Previous studies have applied machine learning techniques in radial–axial ring rolling processes for quality prediction and form-error classification. Fahle et al. (2021a) evaluated several machine learning models using industrial process data to classify ring ovality and assess their

suitability for quality prediction. Similarly, Fahle et al. (2021b) investigated preprocessing strategies for industrial process data collected from a ring rolling machine and evaluated their influence on machine-learning-based form-error classification.

In contrast, the present study focuses on abnormal process behaviour detection rather than quality prediction. The present study applies exploratory phase analysis, a threshold-based classification rule, and supervised machine-learning models as related but distinct analytical components. Their combination is described as a phase-informed analytical approach rather than a new or interpretable monitoring framework.

2.4 Context-Specific Research Motivation

Previous studies have applied machine-learning methods to ring-rolling quality prediction and form-error classification (Fahle et al., 2021a, 2021b). However, combining exploratory analysis, phase comparison, and supervised machine learning does not in itself constitute a general methodological research gap. Instead, the present study addresses a context-specific analytical problem by examining how these existing techniques can be applied to explore abnormal process behaviour in the investigated industrial ring-rolling case.

Accordingly, the contribution of this study is empirical and context-specific. It documents phase-level signal patterns and the outputs and limitations of the implemented classification approaches rather than proposing a new methodological framework.

3.Methodology and Implementation

3.1 Research Method

This study adopts an industrial case-study approach with descriptive and exploratory modelling components. Historical production data from an industrial ring-rolling process were analysed without manipulating the production conditions. According to Wohlin et al. (2012), a case study is appropriate for investigating a phenomenon within its real-world context using existing data without direct manipulation of experimental variables.

The analysis documents a threshold-based classification approach and three supervised machine-learning implementations: Logistic Regression, Random Forest, and XGBoost. These methods represent different approaches to abnormal process behaviour classification. However, because the evaluation was based on a single train–test split without independent validation, cross-validation, or systematic hyperparameter tuning, the results should be interpreted as a descriptive comparison of the implemented model configurations rather than as a definitive comparison of the algorithms.

Although the process measurements are time-stamped and collected sequentially, the supervised-learning component does not formulate the task as time-series forecasting or sequence classification. Each retained sensor observation is treated as an independent supervised-learning instance, and temporal dependencies between consecutive measurements are not explicitly modelled. Time-series studies are therefore used only to describe the industrial data context, sequential measurements, and potential temporal dependence, rather than to justify the implemented supervised-learning methodology.

Following the exploratory statistical analysis, Phase 3 was selected for closer examination

because descriptive differences were observed between normal and abnormal observations in the investigated process signals. Phase 3 was used for the phase-specific exploratory and threshold-based analyses. The supervised machine-learning models, however, were evaluated separately using observation-level data and were not restricted to Phase 3. Accordingly, the phase-level analysis provides exploratory context rather than constituting a phase-aware predictive framework.

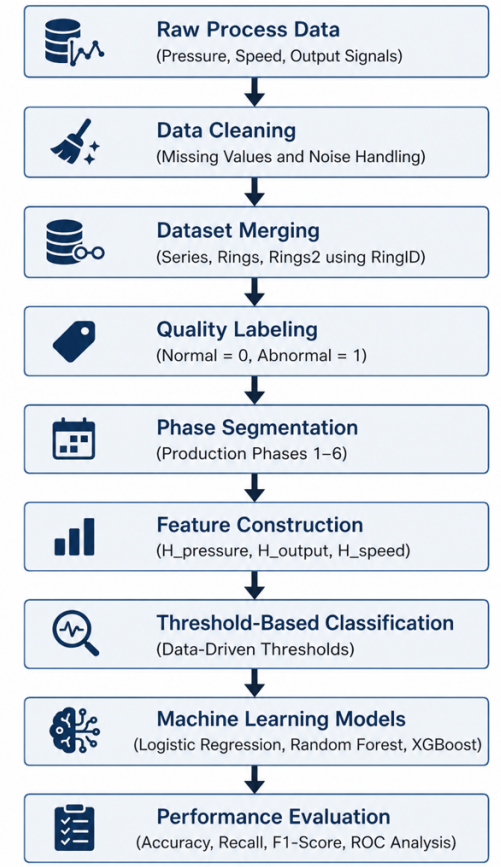

Figure 3.1: Data preprocessing and analysis workflow.

Figure 3.1 summarises the main stages of the methodology, including data preparation, production phase analysis, feature construction, threshold-based classification, supervised machine learning classification, and comparative model evaluation.

3.2 Alternative Research Method

An experimental research method was considered but was not suitable for the objectives of this study. Experimental studies typically require direct manipulation of input variables and controlled observation of their effects on outcome variables (Wohlin et al., 2012). Such an approach was not feasible as the researcher did not have access to the industrial system for conducting controlled experiments or modifying process parameters during production.

Instead, the study relies on historical production data collected during normal industrial operation. Since the available data represent previously completed production runs, it was not possible to alter process conditions or observe their effects under controlled settings. Consequently, an industrial case study was considered more appropriate because it allows abnormal process behaviour to be investigated within its natural production environment,

without altering the conditions under which the data were originally collected.

3.3 Dataset Description

The analysis utilises multiple interrelated datasets generated from an industrial ring rolling process, capturing both process measurements and product quality outcomes.

The primary dataset, **Series**, contains high-frequency multivariate industrial process measurements, where each record corresponds to a sensor measurement collected during production. The dataset includes process variables such as horizontal pressure, rolling speed, control output, positional values, and production phase labels. Each record is linked to an individual production run through the **RingID** and **TimePosix** attributes, enabling process measurements to be associated with specific production runs and production phases. Two additional production-level datasets, **Rings** and **Rings2**, provide supplementary quality information and process metadata for each production run. All datasets were integrated using the shared **RingID** attribute, allowing process measurements to be linked with the corresponding product quality outcomes.

The normal and abnormal class labels used in this study were provided by the industrial company based on ovality assessment. Production runs were classified as acceptable or unacceptable according to the company's quality criteria, resulting in a binary quality label. **Label 1** represents normal production runs, whereas **Label 2** represents abnormal production runs. The combined dataset contains **336,099** normal observations and **356,952** abnormal observations, indicating a near-balanced class distribution at the observation level.

The abnormal behaviour detection problem is formulated as a supervised binary classification task. The prediction target is the binary quality label derived from the company-provided ovality assessment. The input variables used by the classification models are the three constructed features **H_pressure**, **H_speed**, and **H_output**, which are described in Section 3.4.

Table 3.1 summarises the main datasets used in this study and their corresponding contents.

Dataset	Description
Series	High-frequency industrial process dataset including pressure, speed, output, phase indicators, and timestamps
Rings	Production-level dataset including ovality, conicity, and process metadata
Rings2	Additional aggregated dataset with supplementary quality and process-related information

3.4 Feature Construction

Features were constructed from process signals recorded during the industrial ring rolling

process. The selected signals were identified by the industrial company as relevant to process performance and form the input variables used in the classification models.

The **H_pressure** feature represents the horizontal pressure applied during rolling and is calculated as the difference between the positive and negative pressure readings. This feature directly reflects the physical deformation forces acting on the ring during production.

Two deviation-based features were constructed to quantify the difference between the actual and target operating conditions. **H_dev** represents the horizontal deviation and is calculated as the difference between the actual horizontal position and its corresponding set point. Similarly, **V_dev** represents the vertical deviation and is calculated as the difference between the actual vertical position and its corresponding set point. These features quantify the extent to which the process deviates from its intended operating conditions.

H_speed represents the horizontal rolling speed, whereas **H_output** represents the control output signal. Both features describe the response of the control system during production and were included as input variables for the classification models.

All constructed features were retained at the observation level. Each record corresponds to an individual sensor measurement associated with a specific production run and production phase, preserving the original measurement granularity of the dataset.

3.5 Production Phase Analysis

The dataset included a production-phase indicator that assigned each observation to one of six predefined production phases. As part of the exploratory statistical analysis, descriptive comparisons were conducted between normal and abnormal observations across these phases to examine how the investigated process signals varied between the two classes.

The phase-based analysis compared groups of observations according to their predefined production phases. It did not model temporal relationships between observations within or across phases. Phase 3 exhibited observable descriptive differences between normal and abnormal observations in the investigated signals and was therefore selected for closer exploratory and threshold-based analysis. However, no formal statistical comparison, standardised effect-size analysis, or phase-specific predictive evaluation was conducted.

Accordingly, the selection of Phase 3 should be understood as an exploratory and context-specific decision rather than evidence that it was the most informative or predictively important production phase. Phase 3 was used in the phase-specific exploratory and threshold-based analyses, whereas the supervised machine-learning models were evaluated separately using observation-level data and were not restricted to Phase 3.

3.6 Threshold-Based Baseline Classifier

A threshold-based classification approach was implemented as an interpretable baseline for abnormal process behaviour detection. Unlike supervised machine learning models, the proposed approach classifies observations using statistically derived decision thresholds, providing a transparent and easily interpretable decision boundary that can be directly related to the observed process behaviour.

The primary decision threshold was defined using the H_pressure feature. Rather than selecting

an arbitrary cut-off value, the threshold was calculated as the mean of the 25th percentile of the normal-class distribution and the 75th percentile of the abnormal-class distribution. This strategy positions the decision boundary between the two class distributions while reducing the influence of outliers, thereby providing a threshold that reflects the empirical characteristics of the industrial process.

A secondary decision criterion was introduced using H_output, which represents the control-system response during production. This additional criterion complements the primary pressure-based threshold by incorporating information about the system response to process variability.

The threshold-based classification rule was defined as follows. An observation was classified as abnormal if it belonged to **Phase 3** and satisfied at least one of the following conditions:

- **H_pressure** is below the calculated threshold value.
- **H_output** is below its median value.

Based on the exploratory analysis described in Section 3.5, Phase 3 was selected for the threshold-based classification analysis because observable descriptive differences between normal and abnormal observations were present in the investigated process signals. This selection was exploratory and was not based on a formal statistical comparison of predictive importance across the six production phases. The phase condition was applied only to the threshold-based rule; the supervised machine learning models were evaluated separately using observation-level data without production phase as an input feature.

The threshold-based rule was evaluated using a confusion matrix and classification metrics including precision, recall, F1-score, and accuracy. Because the decision threshold was derived from class-specific Phase 3 statistics and evaluated using the same available dataset, its results should be interpreted as an exploratory, directly understandable rule-based reference rather than as an independent estimate of generalisation performance.

3.7 Supervised Machine Learning Models

Supervised machine learning models were applied alongside the threshold-based classification approach to identify abnormal process behaviour. Unlike the threshold-based classification approach, which relies on statistically derived decision thresholds, supervised learning algorithms learn decision boundaries directly from labelled data and can therefore capture more complex relationships among process variables.

The models were trained using three input features: horizontal pressure (H_pressure), horizontal control output (H_output), and horizontal rolling speed (H_speed). The features H_dev and V_dev were excluded because they primarily represent positional control accuracy rather than the core process behaviour and demonstrated limited contribution during the exploratory statistical analysis. The three retained features were selected based on their ability to distinguish between normal and abnormal observations, as identified through the exploratory statistical analysis and production phase analysis.

The target variable was a binary quality label derived from the company-provided ovality assessment, distinguishing normal and abnormal production runs. The dataset was divided into training and test sets using an 80/20 random split at the observation level, with a fixed random seed of 42 to ensure reproducibility. Chronological order was not preserved during this

partitioning. The split was not grouped by production run (RingID); therefore, the procedure did not explicitly prevent observations belonging to the same production run from being allocated to both sets. This represents a potential source of dependence between the training and test observations and should be considered when interpreting the reported generalisation performance. A group-based split by RingID would provide a more rigorous evaluation of performance on unseen production runs.

Class weighting was applied during model training to account for the slight imbalance between normal and abnormal observations.

Three supervised classification algorithms were evaluated: Logistic Regression, Random Forest, and XGBoost. These algorithms were selected to represent different levels of model complexity and interpretability, enabling a systematic comparison of abnormal process behaviour detection performance.

Model performance was evaluated using confusion matrices together with precision, recall, F1-score, accuracy, and ROC-AUC where applicable. No composite performance measure or single criterion representing an overall "balance" was defined. Abnormal-class recall was given particular attention because it measures the proportion of abnormal observations identified, while precision, F1-score, and accuracy were reported to show the classification outcomes associated with prioritising sensitivity. Receiver Operating Characteristic (ROC) and Precision–Recall analyses were also conducted to examine model performance across classification thresholds. An exploratory threshold analysis was performed for XGBoost using a classification threshold of 0.3. Finally, Random Forest feature-importance values were examined to describe the model-derived relative importance of the three input variables within that fitted model.

3.7.1 Logistic Regression (Baseline Model)

Logistic Regression was selected as the baseline supervised classification model because of its simplicity and interpretability. The model estimates the probability of abnormal process behaviour as a linear function of the input features, producing a directly interpretable decision boundary. Class weighting was applied to improve sensitivity towards abnormal observations during model training.

3.7.2 Random Forest (Tree-Based Model)

A Random Forest classifier was included to investigate whether a tree-based ensemble approach could improve upon the baseline. The model constructs multiple decision trees from randomised subsets of the training data and aggregates their predictions, enabling it to capture non-linear relationships between features. Class weighting was applied consistent the Logistic Regression model.

3.7.3 XGBoost (Boosted Model)

An XGBoost classifier was employed to further evaluate predictive performance. The model builds an ensemble of decision trees sequentially, with each tree trained to correct the residual errors of its predecessor, resulting in a powerful classifier that is well suited to structured tabular data. To address class imbalance, the scale_pos_weight parameter was used to assign greater weight to abnormal observations during training. A probability threshold of 0.3 was examined as an exploratory choice to investigate its effect on abnormal-class recall. It was not

selected through a systematic optimisation procedure or an independent validation set.

3.7.4 Phase-Informed Classification Approach

The exploratory phase analysis and the supervised classification analysis were conducted as related but distinct components of the study. Phase 3 was selected for the threshold-based analysis based on the descriptive patterns observed between normal and abnormal observations. In contrast, the supervised machine learning models were trained and evaluated using observation-level data with H_pressure, H_speed, and H_output as input features. Production phase was not included as an input feature, and the supervised models were not restricted to Phase 3.

Accordingly, the study does not implement a phase-aware predictive model or an integrated monitoring framework. Instead, it applies a phase-informed analytical approach in which exploratory phase-level comparisons provide context for the subsequent threshold-based and supervised classification analyses. The predictive results should therefore be attributed to the evaluated model configurations rather than to explicit incorporation of production-phase information.

3.8 Experimental Considerations

The dataset was partitioned using a single 80/20 random train–test split at the observation level, with a fixed random seed of 42 to make the split reproducible. Chronological order was not preserved, and the split was not grouped by production run (RingID). Consequently, observations belonging to the same production run could be allocated to both the training and test sets. Because multiple sensor observations from the same production run may be related, this creates a potential source of data leakage and dependence between the two partitions. The reported metrics should therefore not be interpreted as reliable estimates of performance on completely unseen production runs or future observations. A more rigorous evaluation would use a group-based split by RingID and, where prediction of future production is intended, preserve chronological order. No feature scaling was applied, and the three input features were supplied to the models in their original measurement units.

For the XGBoost model, a classification threshold of 0.3 was applied to the predicted probabilities for the test set instead of the default threshold of 0.5. No separate validation set was created, and the threshold was not selected through a systematic optimisation procedure. Because the test set was used both to examine the threshold and to report the resulting performance metrics, the threshold analysis does not provide an independent estimate of generalisation performance. The value of 0.3 should therefore be described as an exploratory threshold rather than an optimal or independently validated operating point.

K-fold cross-validation was not employed. The classifiers were evaluated using the same single train–test split. Consequently, the reported results provide only one performance estimate for each implemented model configuration and do not demonstrate performance stability across alternative data partitions. Future work should use grouped cross-validation based on RingID to prevent observations from the same production run from appearing in different folds.

No systematic hyperparameter optimisation, such as grid search, randomised search, or Bayesian optimisation, was performed. The models were evaluated using fixed configurations specified directly in the implementation. Therefore, the comparison should be interpreted as a

descriptive evaluation of the implemented model configurations rather than as a comparison of the optimal achievable performance of the algorithms.

4.Experimental Analysis and Results

4.1 Exploratory Statistical Analysis

An exploratory statistical analysis was conducted to examine differences between normal and abnormal observations across the six production phases. The analysis focused on three process signals: H_pressure, H_output, and H_speed, with the objective of examining descriptive differences between normal and abnormal observations across the production phases between the two classes. Across all production phases, H_pressure exhibited the largest separation between normal and abnormal observations. In Phase 3, the mean pressure for normal production runs was approximately 49.66, compared with 46.47 for abnormal production runs, indicating a clear distinction between the two classes. The distributions of H_output and H_speed also differed between the two classes; however, their separation was less pronounced than that observed for H_pressure.

Based on the exploratory phase-level comparisons, Phase 3 was selected for subsequent analysis because it exhibited observable descriptive differences between normal and abnormal observations in the investigated process signals. The Phase 3 analysis examined class-specific patterns in H_pressure, H_speed, and H_output. These descriptive findings informed the subsequent threshold-based classification analysis described in Section 3.6 and the selection of input variables for the supervised machine learning models. They should not be interpreted as statistical evidence that Phase 3 or any individual signal was the most predictively informative.

4.1.1 Phase-Based Analysis Results

The dataset was analysed across all six production phases to examine signal behaviour and identify the production phase that provided the strongest separation between normal and abnormal observations.

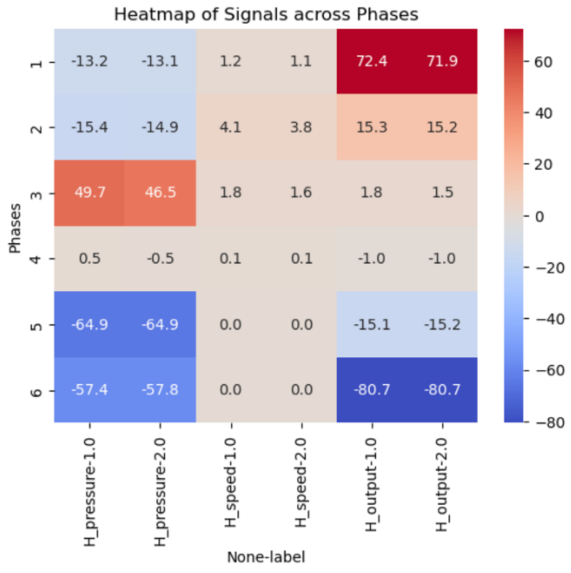

Figure 4.1: Heatmap of Signal Values Across Production Phases

As shown in Figure 4.1, the phase-level mean values of the investigated process signals differed between normal and abnormal observations. In Phase 3, the mean values for normal observations were higher than those for abnormal observations for H_pressure (49.66 compared with 46.48), H_speed (1.83 compared with 1.62), and H_output (1.80 compared with 1.46). These results describe differences between the class-specific means but do not establish that the individual observations were consistently separated. Furthermore, because the three signals are expressed on different measurement scales, the absolute magnitudes of their mean differences cannot be used directly to determine which signal provides the strongest discriminatory or predictive information.

To examine Phase 3 in greater descriptive detail, the phase was divided into three analytical segments: Initial, Main, and Final. This subdivision was used solely for the exploratory analysis of signal characteristics within Phase 3. The segments were not used in the train–test partitioning or included as input features in the supervised machine learning models.

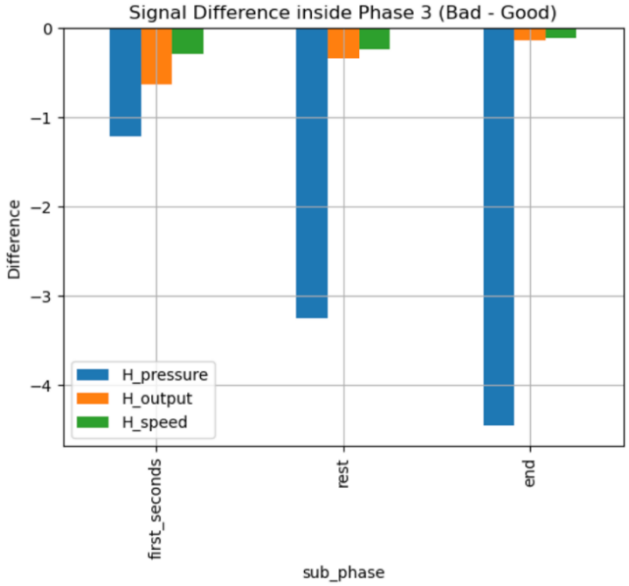

Figure 4.2: Signal differences across Phase 3 segments.

Figure 4.2 illustrates the differences between normal and abnormal production runs across the three analytical segments of Phase 3.

The largest separation is observed in the Final segment, particularly for H_pressure, indicating that this signal provides the strongest discriminatory information within Phase 3. The differences observed across these segments represent descriptive comparisons and should not be interpreted as part of a temporal progression model of abnormal behaviour within Phase 3.

Overall, Phase 3 exhibited observable descriptive differences between normal and abnormal observations. However, no formal statistical or predictive comparison established it as the most informative phase.

4.1.2 Feature Analysis Results

Table 4.1: Mean signal differences between abnormal and normal observations in Phase 3.

Signal	Difference (Abnormal − Normal)
H_pressure	-3.18
H_speed	-0.21
H_output	-0.34

Table 4.1 summarises the mean differences between normal and abnormal production runs during Phase 3. All three process signals exhibited lower mean values for abnormal observations than for normal observations. Among the investigated signals, H_pressure showed the largest absolute difference (−3.18), followed by H_output (−0.34) and H_speed (−0.21). These findings indicate that H_pressure provided the strongest discriminatory information and therefore supported its selection as the primary feature for the subsequent threshold-based classification approach and supervised machine learning models.

Table 4.2: Mean pressure-component values across production phases.

Phase	Label	H_p_plus	H_p_minus
1	1	22.73	35.94
2	1	39.32	54.77
3	1	104.68	55.01
4	1	101.20	100.71
5	1	90.84	155.79
6	1	50.41	107.83
1	2	22.79	35.91
2	2	37.77	52.62
3	2	103.23	56.75
4	2	98.53	99.01
5	2	86.87	151.79
6	2	50.56	108.34

Table 4.2 presents the mean values of the positive (H_p_plus) and negative (H_p_minus) pressure components across the six production phases for both normal and abnormal production runs. Phase 3 recorded the highest H_p_plus values for both classes, whereas H_p_minus remained comparatively lower. Consequently, the calculated H_pressure (H_p_plus – H_p_minus) reached its highest values during Phase 3, with normal production runs exhibiting a higher net pressure (49.67) than abnormal production runs (46.48). This finding further supports the selection of H_pressure as the primary feature used in the threshold-based classification approach and the subsequent supervised machine learning models.

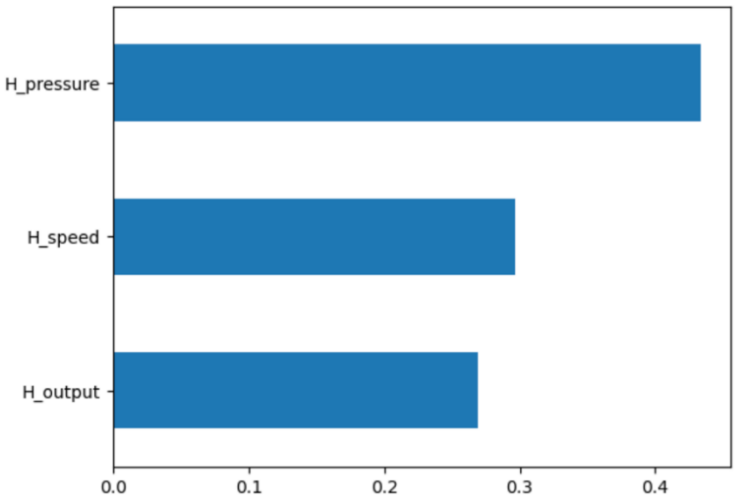

Figure 4.3: Feature Importance Based on Random Forest

As shown in Figure 4.3, H_pressure received the highest feature importance score (≈ 0.44), followed by H_speed (≈ 0.30) and H_output (≈ 0.25). This ranking is consistent with the findings of the exploratory statistical analysis and the threshold-based classification approach, both of which showed that H_pressure had the highest model-derived importance within the fitted Random Forest configuration between normal and abnormal observations. The consistency of these findings across the exploratory statistical analysis, feature importance analysis, and threshold-based classification approach further supports the selection of H_pressure as the primary feature for abnormal process behaviour detection.

4.1.3 Threshold Derivation

The threshold was not used to generate the class labels. The labels were provided by the industrial company based on ovality assessment. The threshold was introduced solely as an interpretable decision boundary within the threshold-based classification approach and served as an analytical reference for comparing the threshold-based classification approach with the supervised machine learning models.

Table 4.3: Quantile values for threshold derivation (Phase 3)

Class	**Quantile**	**Value**
Normal (Class 0)	25%	36.205
Abnormal (Class 1)	75%	62.391

The threshold was calculated as the midpoint between these two values:

Threshold ≈ 49.3

This value represents a data-driven boundary positioned between the two class distributions. It should be noted that this threshold is specific to the current dataset and operating conditions; applying the model to a different product type or production environment would require re-evaluation of the threshold value.

4.1.4 Pressure-Based Threshold Classification Results

Table4.4: Confusion Matrix for the Pressure-Based Threshold Classification Approach

	Predicted 0	Predicted 1
Actual 1 (Normal)	268632	67467
Actual 2 (Abnormal)	272206	84746

Applying the derived threshold of 49.3 to H_pressure within Phase 3, the threshold-based classification approach achieved an abnormal-class recall of 0.24, a precision of 0.56, and an overall accuracy of 0.51. Although the approach correctly identified a proportion of abnormal observations, most abnormal observations were classified as normal (272,206 false negatives), indicating that a single pressure-based threshold provides only partial separation between the two classes. These findings suggest that H_pressure alone is insufficient to capture the full complexity of abnormal process behaviour and therefore motivate the use of supervised machine learning models capable of learning more complex relationships among process variables.

4.1.5 multi-signal Threshold-Based Classification Results

Table 4.5 :Confusion Matrix (Pressure + Output + Speed)

	Predicted 0	**Predicted 1**
Actual 1 (Normal)	281451	54648
Actual 2 (Abnormal)	286833	70119

Extending the threshold-based classification approach to incorporate H_output and H_speed alongside H_pressure did not improve classification performance under the initial multi-signal configuration (Table 4.5). The resulting abnormal-class recall (0.20) was lower than that achieved by the pressure-based threshold classification approach, indicating that simply combining additional process signals without an appropriate decision structure does not necessarily improve abnormal behaviour detection.

Table 4.6: Confusion Matrix for the Enhanced Multi-Signal Threshold-Based Classification Approach)

	Predicted 0	Predicted 1
Actual 1 (Normal)	238763	97336
Actual 2 (Abnormal)	240541	116411

The enhanced multi-signal threshold-based classification approach (Table 4.6) achieved an abnormal-class recall of 0.33 and an F1-score of 0.41, representing an improvement over both the pressure-based threshold classification approach and the initial multi-signal configuration. This improvement was accompanied by a moderate increase in false positives, illustrating the trade-off between detection sensitivity and false-alarm rate.

These findings demonstrate that the effectiveness of a threshold-based classification approach depends primarily on the design of the decision rules rather than simply increasing the number of input signals. Although the enhanced approach improved performance, the remaining classification errors indicate that manually defined thresholds are limited in their ability to capture the complex relationships among industrial process variables. This finding provides the motivation for evaluating supervised machine learning models in the following sections.

4.2 Machine Learning Results

Supervised machine learning models were applied to address the limitations of the threshold-based classification approach, using H_pressure, H_output, and H_speed as input features.

4.2.1 Logistic Regression

Logistic Regression was evaluated as the baseline supervised classification model, assuming a linear relationship between the input features and the target variable. The model achieved an overall accuracy of approximately 0.50, indicating performance close to random classification. These results suggest that the relationship between the selected process signals and abnormal process behaviour cannot be adequately represented by a linear decision boundary, motivating the use of more flexible nonlinear machine learning models.

4.2.2 Random Forest

A Random Forest classifier was applied to capture non-linear relationships between features and the target variable. The model achieved an overall accuracy of 0.80 in the implemented evaluation.

Table 4.7: Random Forest Classification Performance

Metric	Class 0	Class 1
Precision	0.89	0.25
Recall	0.88	0.25

Metric	Class 0	Class 1
F1-score	0.88	0.25
Accuracy	0.80	—

Despite the reported overall accuracy, abnormal-class recall remained relatively low (0.25), indicating limited effectiveness in detecting abnormal observations.

4.2.3 XGBoost Model

To further explore model performance, XGBoost was applied using a boosting-based approach. The model achieved an accuracy of approximately **0.68**, indicating moderate performance.

Table 4.8: XGBoost Classification Performance

Metric	Class 0	Class 1
Precision	0.94	0.25
Recall	0.68	0.72
F1-score	0.79	0.37
Accuracy	0.68	—

The XGBoost model achieved higher abnormal-class recall (0.72) than the Random Forest model, while its overall accuracy (0.68) was lower. These results reflect different performance priorities rather than establishing the overall superiority of either model. In this study, particular attention was given to abnormal-class recall because it measures the proportion of abnormal observations identified. Precision, F1-score, and accuracy were considered alongside recall to describe the associated false-positive cost and overall classification performance. Under this emphasis on abnormal-class recall, XGBoost was carried forward to the subsequent exploratory threshold analysis; however, it should not be interpreted as providing the most effective overall balance across all evaluation metrics.

4.3 Comparative Model Evaluation

4.3.1 ROC Curve Analysis

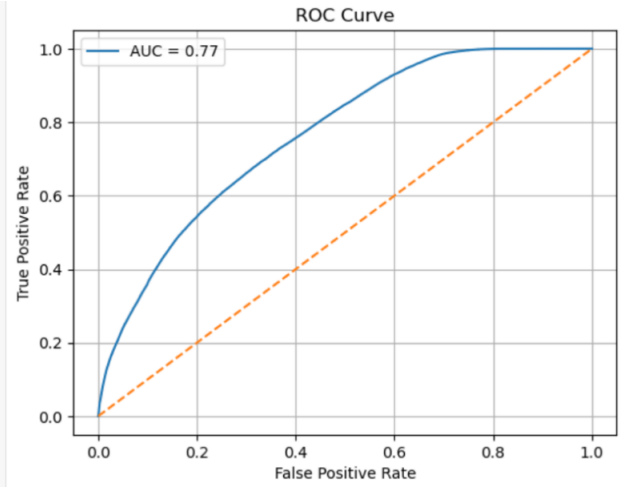

Figure 4.4: ROC curve for the XGBoost model.

Figure 4.4 presents the Receiver Operating Characteristic (ROC) curve for the XGBoost model. The model achieved an Area Under the Curve (AUC) of 0.77, indicating a moderate-to-good ability to distinguish between normal and abnormal observations across different classification thresholds.

ROC analysis was performed for the XGBoost model to examine its ability to distinguish between normal and abnormal observations across different classification thresholds. The ROC analysis provides a threshold-independent assessment of ranking performance and complements the classification metrics reported for the evaluated model configuration.

4.3.2 Precision–Recall Analysis

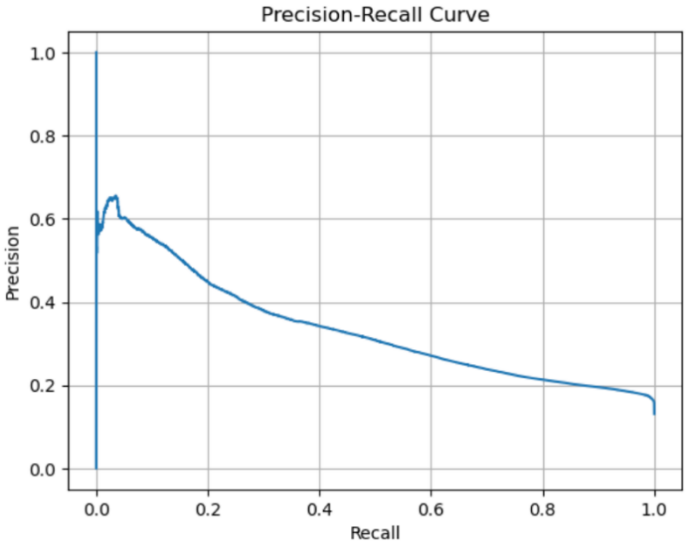

Figure 4.5 Precision–Recall Curve for the XGBoost model.

As shown in Figure 4.5, the curve illustrates the trade-off between precision and recall for the abnormal class across different decision thresholds. The Precision–Recall analysis complements the ROC analysis by showing how increases in abnormal-class sensitivity are associated with changes in precision. These results are therefore interpreted as a trade-off between detecting abnormal observations and limiting false-positive classifications, rather than as evidence that XGBoost is the most effective model overall.

4.3.3 Exploratory XGBoost Threshold Analysis

The default classification threshold of 0.5 was used in the initial XGBoost evaluation. In a subsequent exploratory analysis, a threshold of 0.3 was applied to the predicted probabilities for the test set to examine its effect on abnormal-class recall and the other performance metrics. The threshold was not selected through a systematic optimisation procedure or using an independent validation set. Therefore, it should not be interpreted as an optimal or independently validated operating threshold.

Table 4.9: XGBoost Performance with Adjusted Threshold

Metric	Class 0	Class 1
Precision	0.99	0.17
Recall	0.28	0.99
F1-score	0.44	0.29
Accuracy	0.38	—

Applying the classification threshold of 0.3 resulted in an abnormal-class recall of 0.99, indicating that almost all abnormal observations in the test set were identified. However, this high sensitivity was accompanied by low abnormal-class precision (0.17) and low overall accuracy (0.38), indicating a high false-positive rate. The threshold should therefore not be interpreted as providing balanced or strong overall predictive performance.

Because the threshold of 0.3 was applied to the test set rather than selected using a separate validation set, the procedure does not constitute independent model calibration. The reported metrics may have been influenced by the use of the test data during the exploratory threshold analysis and should therefore be interpreted cautiously. In future work, the classification threshold should be selected on a separate validation set according to an explicit performance criterion, while the test set should remain untouched until the final evaluation.

4.3.4 Scope of the Recorded XGBoost Output

Building on the results presented in Section 4.3.3, the XGBoost test-set predictions were evaluated at a classification threshold of 0.3. At this threshold, the model achieved abnormal-class recall of 0.99, abnormal-class precision of 0.17, and overall accuracy of 0.38. The ROC-AUC value of 0.77 summarises ranking performance across classification thresholds and should therefore be interpreted separately from the results obtained at the specific threshold of 0.3. The high recall at this threshold was accompanied by low precision and accuracy, indicating a high false-positive rate rather than strong overall predictive performance.

The supervised XGBoost model was evaluated using observation-level data with H_pressure, H_speed, and H_output as input features. It was not restricted to Phase 3, and production phase was not included as an input feature. Phase 3 was instead used separately in the exploratory and

threshold-based analyses. Consequently, the implemented procedure does not constitute a final phase-aware decision model or an integration of phase-based filtering with XGBoost. The relationship between these components should be described as phase-informed rather than phase-aware.

The threshold-based rule and the supervised machine learning models provide different analytical perspectives. The threshold-based component offers a directly understandable decision rule based on observed process values, whereas the supervised models were assessed primarily through their classification metrics. No dedicated explainability analysis was conducted for the supervised models; therefore, their results should not be interpreted as evidence of model interpretability or as a validated operational decision-support framework.

4.3.5 Comparison Between Threshold-Based Classification and Machine Learning Models

The implemented supervised machine-learning models produced different metric trade-offs. Logistic Regression achieved a reported accuracy of approximately 0.50 under the implemented evaluation. Random Forest achieved the highest reported accuracy of 0.80, but its abnormal-class recall was only 0.25. XGBoost, at the exploratory threshold of 0.3, achieved abnormal-class recall of 0.99, but this was accompanied by low abnormal-class precision of 0.17 and overall accuracy of 0.38. Because no overall model-selection criterion was defined, these results do not establish a best-performing model or an appropriate balance across recall, precision, and accuracy.

The threshold-based rule and the supervised machine-learning models provide different analytical perspectives. The threshold-based approach provides a directly understandable decision boundary based on the observed process values, while the supervised models can represent more complex relationships among the selected input variables. However, because the evaluation was based on a single observation-level split without independent validation, the reported results do not demonstrate that the supervised models provide greater generalisable predictive performance. Moreover, no dedicated explainability analysis was conducted for the supervised models; therefore, they should not be considered inherently interpretable.

Overall, the threshold-based and supervised machine learning analyses provide complementary perspectives on abnormal process behaviour detection. The threshold-based approach provides a directly understandable decision rule, whereas the supervised models were evaluated using predictive performance metrics. These components are therefore considered part of a phase-informed analytical approach rather than an integrated phase-aware monitoring framework.

5.Discussion

The first research question examined the descriptive differences observed between normal and abnormal observations across the predefined production phases. Based on the exploratory summaries of the investigated process signals, Phase 3 was selected for closer examination and subsequent threshold-based classification analysis. Within Phase 3, class-specific differences were observed in H_pressure, H_speed, and H_output. However, the analysis did not include a formal statistical comparison of the phases or a phase-specific evaluation of predictive importance. The findings therefore support the use of Phase 3 as the focus of the phase-specific exploratory and threshold-based analyses in this study but do not establish that it is the most informative production phase.

A possible process-related interpretation of the observed Phase 3 patterns concerns the operating conditions associated with this stage of the ring-rolling process. During Phase 3, deformation-related activity and the interaction among hydraulic pressure, rolling speed, and control-system adjustments may make differences between normal and abnormal observations more visible in the investigated data. However, the present study did not test the underlying physical mechanisms or establish a causal relationship between these operating conditions and the observed class-specific differences. This interpretation should therefore be treated as context-specific and exploratory. More generally, previous industrial process-monitoring research emphasises that the relevance of process variables may depend on the operational context in which they are measured rather than remaining constant throughout an entire production process (Qin, 2012).

The additional analytical segmentation of Phase 3 provided a more detailed descriptive comparison of the process signals within the selected production stage. This segmentation was conducted solely as part of the exploratory analysis and was not incorporated into the supervised machine learning models. Phase 3 was used to focus the threshold-based analysis, whereas Logistic Regression, Random Forest, and XGBoost were evaluated separately using observation-level data with H_pressure, H_speed, and H_output as input features and without production phase as an explicit model input. Consequently, the phase-specific findings provide descriptive context for the classification analysis but should not be interpreted as evidence that restricting the supervised models to Phase 3 improves predictive performance.

5.1 Interpretation of Findings

The methodological choices adopted in this study were guided by the objective of maintaining consistency between the analytical methodology, the characteristics of the industrial process data, and the physical behaviour of the ring rolling process.

Rather than representing a unified phase-aware monitoring framework, the exploratory statistical analysis, production phase analysis, threshold-based classification, and supervised machine learning analyses were used as related but methodologically distinct components of a phase-informed analytical approach. The phase-level analysis provided descriptive context, while the classification approaches were evaluated separately according to their respective performance characteristics.

The decision to combine a threshold-based classification approach with supervised machine learning was motivated by the complementary strengths of the two approaches. The threshold-based classification approach provides a transparent and interpretable decision boundary that can be directly related to observed process behaviour, making it suitable for industrial environments where decision transparency is essential. In contrast, supervised machine learning models are capable of learning more complex relationships among process variables that cannot be represented by manually defined thresholds alone. Consequently, relying exclusively on either approach would have provided only a partial understanding of abnormal process behaviour. The threshold-based classification approach offers interpretability, whereas supervised machine learning can represent more complex relationships, although the present evaluation does not establish greater generalisable predictive performance.

Another important methodological decision was to retain the process measurements at the observation level throughout the analysis. This decision preserved the original measurement

granularity of the industrial process data and allowed the exploratory statistical analysis to identify the descriptive differences across the production stages. between normal and abnormal observations. Maintaining the original observations also ensured that the threshold-based classification approach and the supervised machine learning models were developed using the same underlying process information, thereby providing a consistent basis for comparing the different abnormal process behaviour detection approaches.

The selection of **Phase 3** for the predictive modelling stage was supported by quantitative evidence obtained during the exploratory statistical analysis rather than by industrial knowledge alone. Although the primary deformation stage is recognised by the industrial company as an operationally important part of the manufacturing process, the statistical analysis independently demonstrated that Phase 3 exhibited observable descriptive differences between normal and abnormal observations and was therefore selected for closer exploratory analysis. However, no formal statistical comparison of phase-specific predictive importance was conducted; therefore, Phase 3 cannot be considered the most informative or predictively strongest phase based on the present analysis.

The selection of the supervised machine learning models also followed a structured rationale. **Logistic Regression** was evaluated as a linear baseline to determine whether abnormal process behaviour could be represented using a simple linear decision boundary. **Random Forest** was selected because of its established performance on structured industrial datasets and its ability to provide feature importance measures that support model interpretation (Breiman, 2001). **XGBoost** was included as a gradient boosting model capable of learning more complex nonlinear relationships while allowing exploratory probability-threshold analysis to examine abnormal-class detection (Chen & Guestrin, 2016). Evaluating these three models in combination provided a descriptive comparison of the implemented configurations of different supervised learning strategies, consistent with the methodological design of this industrial case study.

5.2. Ethical, Societal, and Sustainability Aspects

Although this study focuses on industrial process monitoring and does not involve personal, medical, or other sensitive personal data, several ethical, societal, and sustainability considerations remain important when developing and deploying industrial decision-support systems.

Data Confidentiality and Industrial Privacy

The dataset analysed in this study was obtained from a real industrial manufacturing environment. Such datasets may contain commercially sensitive information relating to manufacturing processes, production performance, and operational practices. To protect industrial confidentiality, production identifiers were removed before the analysis, and no commercially sensitive information is disclosed in this thesis. Furthermore, all results are presented in an aggregated form to ensure that the collaborating industrial company cannot be identified through individual production records.

Responsible Use of Machine Learning for Decision Support

The phase-informed analytical approach presented in this study is intended to support industrial decision-making rather than replace human expertise.

The experimental results demonstrated that increasing abnormal-class recall is accompanied by

an increase in false-positive classifications, highlighting the practical trade-off between detection sensitivity and operational efficiency. Consequently, predictions generated by the supervised machine learning models should be interpreted as decision-support information that complements operator experience and engineering judgement rather than as fully autonomous decisions.

An important characteristic of the phase-informed analytical approach is that threshold-based classification and supervised machine learning are presented as distinct but complementary analyses. The threshold-based classification approach provides a directly understandable decision boundary based on the observed process values, whereas the supervised machine-learning models can represent more complex relationships among the selected input variables. However, the present study did not demonstrate that combining these components improves predictive performance, facilitates model validation, or increases operational trust. Furthermore, no dedicated explainability analysis was conducted for the supervised models; therefore, they should not be considered inherently interpretable.

Sustainability and Resource Efficiency

Abnormal process behaviour may lead to material waste, unnecessary energy consumption, increased machine wear, production downtime, and additional manufacturing costs. Earlier detection of abnormal process behaviour therefore has the potential to improve resource utilisation, reduce manufacturing waste, and increase production efficiency. These outcomes contribute to more sustainable industrial manufacturing and are consistent with the objectives of the **United Nations Sustainable Development Goals (SDGs)**, particularly **SDG 9 (Industry, Innovation and Infrastructure)** and **SDG 12 (Responsible Consumption and Production).**

Industry 4.0 and Human Oversight

The increasing adoption of Industry 4.0 technologies has accelerated the use of automated monitoring and intelligent decision-support systems within industrial manufacturing. Despite these technological advances, effective industrial monitoring continues to require human oversight. Automated models should therefore assist rather than replace engineering judgement, particularly when production decisions involve operational uncertainty. By combining a directly understandable threshold-based classification rule with supervised machine learning analyses, the phase-informed analytical approach provides complementary information for industrial decision support. However, no dedicated explainability analysis was conducted for the supervised models, and therefore they should not be considered inherently interpretable or explainable.

Long-Term Reliability and Generalisability

The analytical approach was examined using historical production data collected under a single industrial operating configuration. Industrial manufacturing environments continuously evolve because of machine ageing, maintenance activities, operational adjustments, and changes in product specifications. Consequently, the statistical characteristics of the process data may also change over time. To maintain reliable predictive performance, both the threshold-based classification approach and the supervised machine learning models should therefore be periodically monitored, validated, and recalibrated before deployment under new operating conditions.

The practical relevance of this study lies in its use of data from an industrial ring-rolling process and its focus on detecting abnormal process behaviour. The analysis combines exploratory

comparisons across predefined production phases, a threshold-based classification rule, and a comparative evaluation of supervised machine learning models. The threshold-based component provides a directly understandable decision rule, whereas the supervised models were evaluated primarily in terms of predictive performance. Because no dedicated explainability analysis or operational deployment study was conducted, the supervised models are not claimed to provide an interpretable or validated decision-support system. The findings instead provide an empirical basis for further evaluation and development in the investigated industrial context.

5.3 Study Limitations and Generalisability

Internal Validity

Several factors may affect the internal validity of this study. First, the class labels used for training and evaluation were derived from the company's ovality assessment, which reflects actual quality judgements. However, these labels are based on expert interpretation and may be subject to human error, inconsistency between assessors, or shifts in quality standards over time. Different labelling criteria could produce different label assignments, which would directly affect both model training and the interpretation of results. This constitutes a fundamental threat to construct validity the extent to which the labels accurately represent the theoretical construct of abnormal process behaviour.

Second, the pressure threshold used in the rule-based model was derived statistically from the dataset using fixed quantile boundaries (25th and 75th percentiles). The choice of these boundaries involves methodological subjectivity; alternative quantile selections would yield different threshold values and potentially different classification outcomes. The threshold should therefore be understood as one defensible choice rather than a uniquely optimal solution.

Third, model performance was evaluated using a fixed 80/20 train–test split, which provides a single performance estimate dependent on which observations were assigned to each partition. The use of stratified k-fold cross-validation would have provided more stable and generalisable performance estimates by averaging results across multiple train–test configurations. Its absence limits confidence in the reported metrics, particularly for the minority class where performance variability across folds tends to be higher. This choice was made to maintain consistency with the evaluation approach used in the directly comparable study by Fahle et al. (2021a), enabling a more meaningful comparison with previously reported results. Future studies should adopt cross-validation to produce more robust estimates, particularly when datasets are of moderate size.

Fourth, the supervised models were applied to observation-level features without explicit modelling of sequential dependencies between time steps. Although the production phase analysis distinguishes observations according to their production stage, ordered temporal patterns within phases were not explicitly captured or modelled. Sequential modelling approaches, such as recurrent neural networks or time-series-specific classifiers, may better represent the dynamic progression of abnormal behaviour in future work to capture temporal dependencies that could provide additional predictive information for abnormal process behaviour detection.

External Validity

The study was conducted using data from a single industrial ring rolling environment provided by one company. The findings therefore reflect the specific operating conditions, product configurations, and quality-control procedures of that environment. Whilst the signals examined (pressure, speed, and control output) are common to radial-axial ring rolling systems more broadly, the specific thresholds, phase boundaries, and feature importance rankings identified here may not transfer directly to systems operating under different parameters or producing different ring geometries. Whether the framework would perform equivalently in another production facility, with different machine configurations or label definitions, remains an open empirical question. Replication across multiple industrial environments would be required before generalisation could be responsibly claimed.

Validity of Drawn Conclusions

The conclusions concerning H_pressure and Phase 3 are limited to the data, variables, and analytical procedures used in this study. The phase-level analysis was exploratory and descriptive; therefore, it does not establish the statistical significance or predictive superiority of Phase 3 over the other production phases. Among the three input variables included in the fitted Random Forest model, H_pressure received the highest model-derived feature-importance value. However, this result is specific to that model configuration and dataset and does not establish that H_pressure is universally the most informative process signal. Additional variables, such as torque, positional measurements, or thermal data, could produce different results. Furthermore, the observation-level train–test split and the use of test-set predictions in the exploratory threshold analysis limit the strength of conclusions concerning generalisation. The findings should therefore be interpreted as dataset- and model-specific rather than as universally generalisable conclusions.

6.Conclusion and Future Work

6.1 Conclusion

This thesis investigated a phase-informed data-driven approach to abnormal process behaviour detection using historical production data from an industrial ring-rolling process. Within this context, a comparative empirical evaluation was conducted to assess a threshold-based classification approach and supervised machine learning models. Exploratory analysis of process signals across predefined production phases was used to examine class-specific patterns and inform the subsequent analysis.

The exploratory analysis showed that the observed descriptive patterns between normal and abnormal observations varied across the six production phases. Based on these exploratory summaries, Phase 3 was selected for closer examination because observable class-specific differences were present in the investigated process signals. However, no formal statistical comparison of predictive importance across the phases was conducted. Phase 3 should therefore be interpreted as the focus of the exploratory and threshold-based analyses rather than as the statistically established most informative production phase. The supervised machine learning models were evaluated separately using observation-level data and did not explicitly include production phase as an input feature.

The supervised models used H_pressure, H_speed, and H_output as input features. In the fitted Random Forest model, H_pressure received the highest model-derived feature-importance value

among these three variables. This result is specific to the evaluated dataset and model configuration and does not establish that H_pressure is universally the most informative or discriminative process signal. Because no feature-ablation analysis was conducted, the individual or complementary contribution of each variable to classification performance was not independently established.

The comparative evaluation demonstrated differences among the investigated classification approaches. The threshold-based approach provided a directly understandable decision rule, while the supervised models represented alternative methods for modelling relationships among the selected process variables. At a classification threshold of 0.3, XGBoost achieved an abnormal-class recall of 0.99, but this high sensitivity was accompanied by low abnormal-class precision of 0.17 and low overall accuracy of 0.38, indicating a high false-positive rate. Because this threshold was applied to the test set as part of an exploratory threshold analysis, these results should not be interpreted as an independently validated optimal operating point. No single model was shown to be superior across all evaluation criteria; the preferred configuration depends on whether the monitoring priority is abnormal-class sensitivity, precision, or overall accuracy.

The principal contribution of this study is a context-specific, phase-informed analytical approach that combines exploratory phase-level comparisons with threshold-based and supervised classification analyses. The phase analysis provides descriptive context rather than being explicitly incorporated into the supervised predictive models. The evaluated methods provide complementary analytical perspectives, and their performance should be interpreted according to the specific evaluation criteria rather than as evidence of a unified phase-aware monitoring framework.

This study has several limitations. The analysis was conducted using historical production data collected from a single industrial environment, and only three process signals were included in the evaluation. Consequently, the findings should be interpreted within the context of the investigated manufacturing process. Nevertheless, the phase-informed analytical approach presented in this study provides a reproducible basis for future studies involving additional industrial datasets, process variables, and manufacturing environments.

Overall, the study provides a context-specific and phase-informed exploratory analysis of abnormal process behaviour in the investigated industrial setting. It does not establish a most informative production phase, a best performing model, an interpretable XGBoost system, or a validated phase-aware monitoring framework. Nevertheless, it provides a transparent account of the observed phase level patterns, the metric trade-offs produced by the implemented classification approaches, and the methodological limitations that should guide future research.

6.2 Future Work

Several directions for future research emerge from the findings and limitations of this study and provide opportunities to further develop and validate the phase-informed analytical approach. Process measurements such as torque, current, temperature, and positional variables may provide complementary information for abnormal process behaviour detection. Expanding the feature set may improve classification performance and provide a more comprehensive understanding of the physical mechanisms associated with abnormal process behaviour.

Second, the phase-informed analytical approach should be evaluated using production data

collected from multiple industrial environments, different machine configurations, and a wider range of product types. Such validation would enable the robustness and generalisability of the approach to be assessed under varying manufacturing conditions and determine whether the observed phase-level patterns and feature rankings remain consistent across different operating environments.

Third, future work could investigate adaptive threshold calibration methods for the threshold-based classification approach. The threshold used in this study was derived from the current dataset and operating conditions. Developing adaptive threshold selection strategies that automatically adjust to changing production conditions could improve the long-term reliability of the threshold-based classification approach in industrial applications.

Fourth, the analytical approach was evaluated using historical production data. Future work could investigate how the evaluated analytical components may be adapted and validated for real-time industrial environments. Such an implementation would require continuous data acquisition, online model inference, and periodic model validation to ensure reliable performance under changing production conditions.

Fifth, future research could investigate additional supervised machine learning models and advanced model interpretation techniques to further improve predictive performance and model transparency. Methods such as **SHAP (SHapley Additive exPlanations)** and **LIME (Local Interpretable Model-agnostic Explanations)** may provide more detailed explanations of individual prediction outcomes and complement the global feature importance analysis presented in this study (Guidotti et al., 2018).

Finally, future work could explore how the phase-informed analytical approach may be adapted and validated for use within industrial decision-support systems. Combining abnormal process behaviour detection with process visualisation, operator guidance, and interactive monitoring interfaces could support the practical evaluation of the approach in industrial settings.

References

Blázquez-García, A., Conde, A., Mori, U., & Lozano, J. A. (2021). A review on outlier/anomaly detection in time series data. *ACM Computing Surveys*, *54*(3), 1–33. https://doi.org/10.1145/3444690

Braei, M., & Wagner, S. (2020). *Anomaly detection in univariate time-series: A survey on the state-of-the-art* (arXiv:2004.00433). arXiv. https://arxiv.org/abs/2004.00433

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chalapathy, R., & Chawla, S. (2019). *Deep learning for anomaly detection: A survey* (arXiv:1901.03407). arXiv. https://arxiv.org/abs/1901.03407

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, *41*(3), 1–58. https://doi.org/10.1145/1541880.1541882

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Fahle, S., Glaser, T., & Kuhlenkötter, B. (2021a). Investigation of machine learning models for a time series classification task in radial–axial ring rolling. In G. Daehn, J. Cao, B. Kinsey, T. Altan, A. E. Tekkaya, & Y. Wang (Eds.), *Forming the future: Proceedings of the 13th International Conference on the Technology of Plasticity* (pp. 589–600). Springer. https://doi.org/10.1007/978-3-030-75381-8_48

Fahle, S., Kneißler, A., Glaser, T., & Kuhlenkötter, B. (2021b). Research on preprocessing methods for time series classification using machine learning models in the domain of radial-axial ring rolling. In *Lightweight design in product and production development* (pp. 525–534). Springer Vieweg. https://doi.org/10.1007/978-3-662-62138-7_49

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 1–42. https://doi.org/10.1145/3236009

Gupta, M., Gao, J., Aggarwal, C. C., & Han, J. (2014). Outlier detection for temporal data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, *26*(9), 2250–2267. https://doi.org/10.1109/TKDE.2013.184

Kang, S., & Lee, J. H. (2024). Data-driven process monitoring and fault diagnosis: A comprehensive survey. *Processes*, *12*(2), Article 251. https://doi.org/10.3390/pr12020251

Pang, G., Shen, C., Cao, L., & Van Den Hengel, A. (2021). Deep learning for anomaly detection: A review. *ACM Computing Surveys*, *54*(2), 1–38. https://doi.org/10.1145/3439950

Qin, S. J. (2012). Survey on data-driven industrial process monitoring and diagnosis. *Annual Reviews in Control*, *36*(2), 220–234. https://doi.org/10.1016/j.arcontrol.2012.09.004

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M., Regnell, B., & Wesslén, A. (2012). *Experimentation in software engineering*. Springer. https://doi.org/10.1007/978-3-642-29044-

2

Yan, P., Abdulkadir, A., Luley, P. P., Rosenthal, M., Schatte, G. A., Grewe, B. F., & Stadelmann, T. (2023). *A comprehensive survey of deep transfer learning for anomaly detection in industrial time series* (arXiv:2307.05638). arXiv. https://arxiv.org/abs/2307.05638