Elisha Itani

Predictive Quality Control and Defect Detection in Steel Manufacturing Using Statistical Learning and Process Analytics

School of Technology and Innovations
Master's thesis in Industrial Engineering and Management
Master's Programme in Industrial Systems Analytics

Vaasa 2026.

UNIVERSITY OF VAASA
School of Technology and Innovations
Author: Elisha Itani
Title of the thesis: Predictive Quality Control and Defect Detection in Steel Manufacturing Using Statistical Learning and Process Analytics
Degree: Master of Sciences in Technology and Innovation
Degree Programme: Industrial Systems Analytics
Supervisor: Jyri Naarmala
Year: 2026 **Pages:** 73

ABSTRACT:

Steel plate manufacturing faces both operational hassles and financial impact caused by surface imperfections; yet, conventional human inspection usually suffers from fatigue, subjectivity, and modern rolling mills' throughput demands. This research aims to bridge three gaps in the field of automated quality control; these gaps highlight the scarcity of methodical, comprehensive cross-paradigm comparisons of algorithms. They also identify the limited operational availability of model interpretability frameworks and the lack of design of easily employable dashboarding tools for floor production personnel. From the UCI Steel Plates Faults Dataset, 1,941 instances of steel plates structured across seven defect labels were used. These instances were characterized by 27 process and geometric attributes. Within structured, reproducibly framed environments, classification algorithms of all biases and variance were both constructed and then evaluated. The five algorithms of this category were Linear Discriminant, Multinomial Logistic, Random Forest, Support Vector, and XGBoost. Class imbalances were corrected through real-sample up-sampling in the training portion, while a 30% held-out testing portion preserved the natural class distribution. The greatest observed testing set accuracy and weighted ROC-AUC were 79.79% and 0.9702, respectively, for XGBoost. Random Forest testing set upholding 78.58%, and AUC 0.9641, followed closely, with both classifiers noted for substantially surpassing the performance of linear classifiers. SHAP and DALEX permutation analyses discerned class-level feature drivers located within spatial parameters of Z_Scratch and Bumps, luminance attributes of K_Scratch and Stains, and the steel type for Dirtiness. The described analyses and attributes offer explainable, spatial process diagnostics of relevance. The research also focuses on the design of proprietary tools, where an interactive, seven-module R Shiny dashboard was both developed and utilized as a deployment method to fulfill real-time defect prediction, along with SHAP layered class suggestions for process personnel who lack subject specialization, thus satisfying the trinity of research ventures.

Keywords: steel plate defect detection, predictive quality control, XGBoost, Random Forest, SHAP interpretability, Statistical Process Control, Industry 4.0, R Shiny dashboard

Contents

List of Figures

Tables

1. Introduction

1.1 Overview

This chapter gives the background to the study on automated defect detection and predictive quality control in the steel manufacturing process. It begins with an analysis of the industrial and technological context that has triggered the increasing use of machine learning in manufacturing quality systems, and places the research in the context of the overall transformation that Industry 4.0 technologies and the data-intensive production contexts they generate are driving. The chapter proceeds to define and state the proposed research problem and to elaborate why the current methods are not adequate, and what particular gaps this work aims to fill. Following this, the research objectives and, respectively, research questions are clearly expressed, which sets the logical base according to which the methodological decisions will be made during the composition of the thesis. At the end of the chapter, the organization of the whole thesis is summarized, giving a brief explanation of the contribution and scope of each subsequent chapter to enable the reader to follow the thesis in a clear and meaningful way. These sections combined build the intellectual framework, inspirational reasoning, and institutional limitations of the study since the reader begins the literature review and methodology with a well-defined view of what this thesis intends to do and why it is important.

1.2 Background

One of the most economically impactful manufacturing sectors of the global economy is the manufacturing of steel, which contributes to the most vital infrastructures in the construction industry, automotive engineering industry, shipbuilding industry, and energy systems. The surface quality in this field is a matter of utmost technical and commercial concern. Such defects as scratches, stains, bumps, and surface pitting deteriorate the mechanical properties and cosmetic appearance of steel plates, leading to expensive rework loops, elevated material rejection rates, and creating a long-term customer dissatisfaction effect (Li et al., 2024). The total financial implication of the presence of unnoticed defects, multiplied by volumes of production of thousands of plates per day, is a considerable liability of operation of a steel producer of any size, making quality assurance a strategic priority instead of an operational

one. With increasing production tolerances and increasingly demanding customer specifications, the ineffectiveness of traditional inspection is no longer merely inconvenient; it is economically and reputationally unsustainable.

Traditionally, quality assurance used in the steel production process was based on visual inspection performed by trained specialists who were stationed throughout the production line. Although this method has offered a practical quality benchmark, it has serious limitations that have been well documented. Human inspectors are vulnerable to fatigue during long shifts, and their judgments are subjective in nature, which creates inter-operator variability, resulting in inconsistency and non-repeatability among the defect classifications in production batches and across shifts. Moreover, the rapidity of modern rolling mills compels manual inspection to fall physically behind production throughput without causing intolerable bottlenecks and delays (Chigateri & Hebbale, 2024). With the increasing demand for steel products globally and the growing dimensional requirements, these restrictions create a structural ceiling to the attainable quality standards by only using conventional inspection methods.

The advent of Industry 4.0 has radically changed the technological choices that quality engineers have. The combination of Internet of Things sensors, edge, and cloud-based data storage, and sophisticated analytical platforms has empowered manufacturers to instrument their manufacturing operations on a broad scale and to create continuous real-time data streams out of which wise quality inferences can be inferred (Goecks et al., 2024). Machine learning has taken center stage in this change. Algorithms that have been trained using labelled defect data can identify intricate non-linear features in high-dimensional feature spaces to classify surface anomalies with accuracy that is substantially better than that of a human examiner in a controlled experimental setting. Recent advances, such as deep convolutional neural networks, vision transformers, and gradient-boosted ensemble algorithms, have pushed the classification performance high enough to make industrial applications increasingly practical (Gao et al., 2024; Yuan et al., 2024). This technological trend renders automated defect detection not only technically achievable but industrially obligatory to manufacturers that aim at gaining a competitive edge by ensuring quality excellence. Importantly, the algorithms can also be used to detect the precursors of defects

before the full non-conformities occur, which would allow transitioning from reactive correction to proactive prevention.

1.3 Research Problem and Motivation

Regardless of the impressive technical advances reported in the literature, there remain numerous gaps that ensure machine learning-based quality control will not fulfill its potential in practical implementation in steel manufacturing. The first gap is related to the lack of systematic and methodologically rigorous comparative studies. Most published work either praises a particular deep learning architecture or compares classical statistical methods in isolation, not holding the two paradigms in competitive controlled trials with both using the same datasets under identical evaluation conditions. This fragmentation renders it challenging to ascertain whether the computational cost and complexity of sophisticated ML models are worth more in relation to incremental performance gains over simpler, more understandable statistical baselines (Chigateri & Hebbale, 2024). The current literature does not give a systematic, repeatable answer to this question, which decision-makers investing in quality control infrastructure need to make, based on evidence and comparative guidance.

The second gap is the interpretability of models. The current trend of aggregate accuracy measures has seen predictive models being considered as black boxes, where the outputs are reported without providing insight into the process-level factors that induce them. It is a very serious shortcoming in any manufacturing process where a prediction of an error is only workable when the quality engineer is aware of what process variables went into the result, and that one can modify production parameters to correct the error. A systematic review conducted by Mehdiyev et al. (2025) proved that, despite the popularity of interpretable predictive process monitoring methods like SHAP and LIME in the academic community, it has been rare to adopt such methods into operational quality control frameworks. The disjunction between generating a statistical interpretation and integrating it into an operational, operations-ready manufacturing interface is a major and hardly discussed problem among researchers and industry practitioners interested in bridging the gap between analysis and remedial action.

The third gap is the deployment gap, the unresolved distance between an experimentally validated model and an effective tool that manufacturing employees can effortlessly obtain

and apply. Díaz-Martínez et al. (2025), who conducted a review of more than fifty studies on Industry 4.0 adoption, discovered that the major barriers to successful implementation pertained to data integration complexity, the lack of workforce digital capability, and organisational resistance. In steel production, the lack of interactive dashboard platforms that can display ML predictions and SHAP-generated explanations in a format that can be understood by non-expert operators implies that technically advanced models often do not permeate into the production line. In the absence of available deployment mechanisms, the most accurate predictive model is of no practical value.

These three gaps, coupled with the fact that steel plate manufacturing presents the most favourable conditions under which to work, give rise to the motivation of this research. The UCI Steel Plates Faults Dataset, which includes 1,941 labeled cases in seven defect categories and 27 process variables, offers a well-known and reproducible empirical base that would be appropriate for the strict cross-paradigm assessment (UCI Machine Learning Repository, 2023). The R programming ecosystem, including the caret, randomForest, xgboost, e1071, DALEX, and shiny packages, provides a complete environment where the training of models, SHAP explainability analysis, and the deployment of dashboards can all be executed in a single transparent workflow. This study, thus, answers a real industrial challenge and adds to the body of literature on academic quality control a methodologically consistent and practically implementable solution.

1.4 Research Objectives

This study has three main objectives; they are interdependent and capture both theoretical and practical dimensions. They progress from algorithm evaluation to understanding the process, and finally, to its operationalization, covering the entire life cycle of a manufacturing quality control system.

The first research objective is to cross-compare the machine learning models to facilitate the automation of the detection of defects for the first time. This involves the construction of a systematic cross-paradigm comparison of the five chosen algorithms that occupy the bias/variance trade-off spectrum from moderately competent linear classifiers to highly competent ensemble methods. This systematic construction enables the justification of choosing each of the five algorithms to perform the required classification.

The second research objective is to determine the process parameters that correlate most closely with each type of defect. Understanding the parameters that are responsible for the categories of defects enables a diagnostic model to devolve from a black box and serves a purpose beyond prediction. This equips engineers with the information required to trace each of the categories of defects to their process subsections and to implement suitable corrective measures.

The third research objective is to provide a model that is sufficiently simple for floor operators who lack a background in data science. This is achieved via the development of an interactive R Shiny dashboard to provide real-time predictive capability for defects via the dashboard and to SHAP for explanation. It aims to create and design an interactive R Shiny-based dashboard to offer production staff, without any data science or statistical background, the tools to assess SHAP-based explanations and conduct defect prediction in real-time.

1.5 Research Questions

This thesis outlines three research questions, formed based on the research objectives established earlier. The questions progress vertically along the topics of technical research, research on processes, and finally, the question of the feasibility of the developed models in operational contexts.

RQ1: Which machine learning algorithms help to identify defects on the surface of steel plates with the highest accuracy? This question promotes the comparative evaluation of models and addresses the demand for precise and repeatable performance measurement across families of algorithms, an area of research that is presently underexplored in the existing body of literature.

RQ2: What are the most significant process parameters that are likely to lead to specific defects during the fabrication of steel plates? Beyond interpreting identified defects, this question involves a level of analysis that goes beyond overall accuracy to the interpretability of individual features.

RQ3: What are the ways of combining predictive quality control models in manufacturing operations to lower costs and enhance quality? This question addresses the deployment challenge and is answered through the development of the R Shiny dashboard, which

demonstrates how predictive models and their explanations can be made accessible to non-specialist users in a live production setting.

1.6 Thesis Structure

This thesis is structured into five chapters, with each chapter having a specific and complementary role in the entire research design.

1. Introduction lays out the industrial and technological background of the research, outlines the research problem and motivating gaps, outlines the research objectives and questions, and describes how the thesis will be organised. It places the research at the nexus of Industry 4.0, machine learning-driven quality management, and explainable AI in intelligent manufacturing.

2. Literature Review will critically synthesize existing academic literature in four thematic areas, namely machine learning and deep learning strategies to detect steel defects, sophisticated architectures like convolutional neural networks and vision transformers, SHAP-based explainability systems used within manufacturing processes, and the general implementation issues of Industry 4.0 in production facilities. The chapter also visually outlines the existing state of knowledge and specifically highlights the research gaps that are to be filled by this study.

3. Methodology explains the quantitative comparative research design, the UCI Steel Plates Faults Dataset and its attributes, the data preprocessing pipeline, the five classification algorithms to be evaluated, the SHAP analysis process, the performance evaluation measures, and the six-module R Shiny dashboard architecture. Special emphasis is placed on experimental controls, stratified train-test splitting, and class balancing using smote, so that the validity and reproducibility of findings are guaranteed.

4. Results and Discussion provides the empirical results of the project, comparing the model classification performance of all five algorithms, providing SHAP feature attributions in relation to the category of defects, testing the dashboard functionality and response time, and the practical implications of the study to the field of steel manufacturing quality control. Critical analysis in relation to the results obtained focuses on the limitations of the study and the limits of the generalisability of the results.

5. Conclusion summarises the key contributions of the work, assesses the degree to which each of the mentioned objectives is fulfilled, and outlines the limitations of the scope of the conducted work, as well as the opportunities for further work, such as the implementation of deep learning, real-time data streaming architectures, and multi-dataset generalisation studies.

2. Literature Review

2.1 Introduction

The chapter is a critical review of the current literature pertinent to the research objectives set in Chapter 1. The review is thematically structured in five theme areas that cumulatively represent the intellectual landscape of this research: machine learning methods to detect steel plate defects; neural network architectures and vision transformers; explainability models and SHAP-driven interpretability in smart manufacturing; statistical process control and classical approaches to quality management; and Industry 4.0 implementation and interactive dashboard deployment in smart manufacturing scenarios. Across the domains, the chapter reviews key contributions, limitations, and assesses whether solutions to date meet the practical requirements of industrial quality control. The important contributions, methodological trends, and unsolved limitations are identified in each section, resulting in a synthesis that visualizes the specific research gaps that this study is intended to fill. The reviewed literature represents the modern level of knowledge in a highly dynamic field of research.

2.2 Machine Learning Approaches to Steel Plate Defect Detection

Surface defect detection in steel manufacturing with machine learning has produced a significant body of research in recent years due to the combination of increased computational power, publicly available benchmark datasets, and increased industrial interest in automated inspection systems. Another recurring motif throughout this literature is the hunt to find algorithms capable of simultaneously providing high classification accuracy, governing predictable behaviour on new production samples, and within the computational limits associated with manufacturing contexts. Steel surface defect recognition has been thoroughly examined, with reviews highlighting visual-based, feature-based, and learning-based inspection methodologies as the prevailing paradigms in industrial applications (Wen et al., 2023). Liu et al. (2024) directly tackled the problem of computational efficiency by developing a robust detection system on the principles of lightweight convolution models, namely ScConv and GSConv, which trim model size and inference time without compromising on performance on datasets of steel surface defects. Similarly, Lv et al. (2020) through their

work, they showed that the accuracy-efficiency tradeoff, which has long been considered an inevitable constraint of running machine learning on production hardware, can be significantly alleviated by making architecturally informed design decisions. This result has direct consequences in the field of industrial deployability, where manufacturers often run inspection systems on edge devices that have limited processing power and are highly constrained in terms of latency.

A general comparative analysis of machine learning and data mining algorithms in various industrial manufacturing datasets was performed by Dogan and Birant (2021) and systematically evaluated decision trees, support vector machines, k-nearest neighbours, and logistic regression. Their comparison established that the use of ensembles was repeatedly better than solitary classifiers at complex, high-dimensional manufacturing classification, and that the relative benefit grew with the difficulty of the datasets and the imbalance between classes. Notably, the research also showed that less complex interpretable models still had sufficient competitive worth in manufacturing environments, where the openness of the decision logic can be operationally and organisationally meaningful to both quality engineers and production managers. Shwartz-Ziv and Armon (2022) provided a critique of the existing belief that deep neural networks will always beat classical machine learning with tabular data, the most common format in industrial quality control systems, where sensor measurements, spatial information, and process variables are recorded as structured numerical data sets. Their systematic experimental performance on a wide variety of benchmark datasets established that tree-based ensemble approaches, such as gradient boosting variants, performed as well as deep learning in most structured tabular applications, and at much lower training cost and with many fewer interpretable results. This observation directly guides the algorithm choice rationale of the current work, including classical statistical approaches to the problem, as well as the state-of-the-art ensemble models, as the valid and theoretically-grounded choices to classify defects in steel plates. This aligns with the overarching conclusion of Tercan and Meisen (2022), whose comprehensive review of predictive quality in manufacturing established that ensemble methods prevail in industrial classification tasks owing to their resilience in handling high-dimensional, noisy process data.

The literature in this field is united by the following key insights: ensemble methods are always competitive on structured manufacturing data; computational efficiency is a real

deployment concern; and class imbalance, which is prevalent in defect datasets with conforming samples vastly outnumbering defective samples, must be actively controlled by techniques such as synthetic oversampling to prevent optimistic and misleading estimates of performance across minority defect classes. A study of machine learning algorithms used on the UCI Steel Plates Faults dataset, which was also used in this study, found that Random Forest and Neural Networks had the best accuracy. It also found that hyperparameter tuning and feature selection always made all the classifiers tested work better (Gao et al., 2024). Similarly, Taşar (2022) benchmarked linear discriminant, KNN, decision tree, SVM, random forest, and DNN models on the same dataset, with the DNN model achieving the highest accuracy at 96.99%. However, this benchmarking was performed without data partitioning or addressing class imbalance. Therefore, the results would not likely demonstrate real-world constraints for production. Ozkat (2022) built on this work and added hyperparameter optimization and feature selection. The work showed that these additional steps improved the performance of ensemble classifiers on the UCI steel plate dataset. This also provides additional impetus for the structured evaluation framework used within this research.

Dorbane et al. (2025) substantiated this using the same dataset, showing that SMOTE-balanced Stacked Ensemble and XGBoost classifiers achieved an accuracy of 0.99, while SHAP and LIME analyses validated that geometric and intensity features influenced the predictions, establishing a direct methodological precedent for the current study. Another common theme here is that model evaluation in this field should go beyond single-figure measures of overall accuracy to include per-class measures of the model precision and recall, as a model that predicts well the majority non-defective class but poorly on infrequent but high-impact defect classes will not give much practical use to quality engineers. This focus on balanced and class-specific measures of evaluation characterizes a significant methodological necessity of any serious comparative study of defect classification algorithms. The hybrid approach of combining deep feature extraction with classical classification has also shown promise. Hussain et al. (2024) showed that a CNN-SVM architecture outperformed standalone deep learning models on hot-rolled steel strip surfaces. This shows how important it is to include traditional classifiers even in modern deep learning pipelines. Figure 1 presents the accuracy–interpretability landscape of machine learning method (adapted from Alzubaidi et al. (2021), Dogan and Birant (2021), and Shwartz-Ziv and Armon (2022).

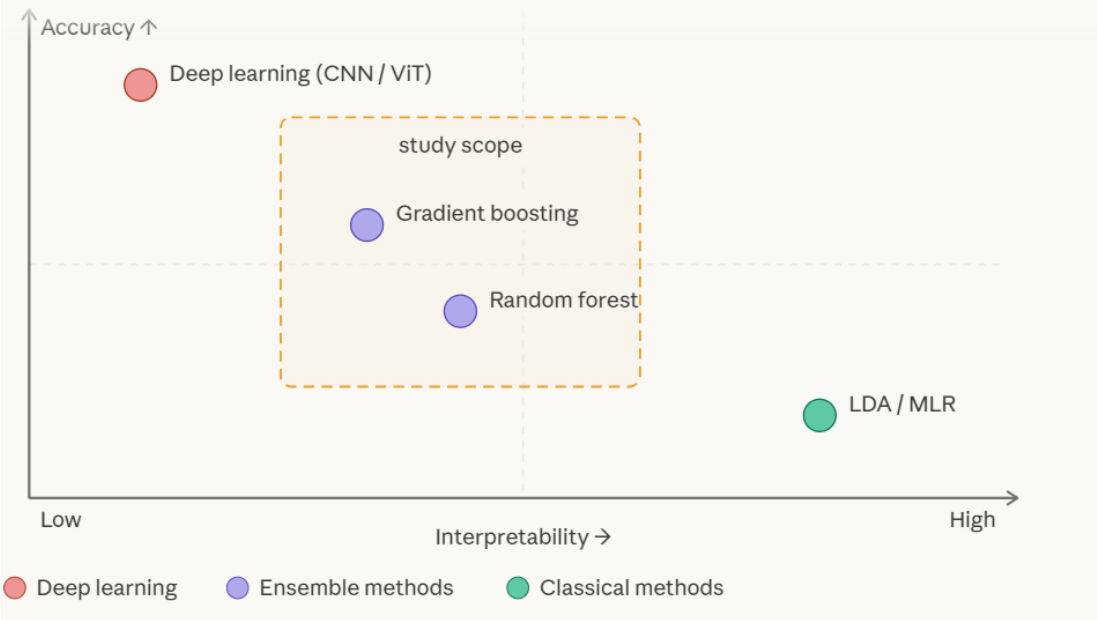


Figure 1. Accuracy–interpretability landscape of machine learning method (adapted from Alzubaidi et al. (2021), Dogan and Birant (2021), and Shwartz-Ziv and Armon (2022).

2.3 Deep Learning Architectures and Vision Transformers

Although the research at hand addresses (tabular) feature-based classification, not image-based inspection, the deep learning literature offers valuable background information on what the current frontier of performance is and what architectural advancements are driving further advances in automated defect detection more generally. Chen et al. (2021) did a systematic review of ways to find surface defects in industrial products and grouped them into three groups: deep learning, machine learning, and traditional image processing. They came to the conclusion that deep learning techniques are now the most important performance benchmarks, but they also need more labelled training data. On the other hand, Czimmermann et al. (2020) made a comprehensive literature review on visual, statistical, and inspection methods based on learning of various industrial materials, including steel; they showed that classical image processing methods are quickly surpassed by deep learning technologies in terms of performance. In addition, a dedicated benchmark dataset for deep metallic surface defect detection has also been developed to accelerate progress in this space, providing standardised evaluation grounds for new detection networks (Lv et al., 2020). Alzubaidi et al. (2021) gave a broad and well-cited overview of the ideas of deep learning, convolutional neural network models, and how they can be applied in various fields, such as industrial inspection. The analysis tracked the development of the early convolutional

networks to residual networks, densely connected networks, and attention-based networks, demonstrating how newer advances gradually resolved the vanishing gradient problem, computational cost, and the challenge of long-range spatial dependencies. This architectural history provides the requisite context for the appearance of vision transformers as an engaging, different paradigm of making defect classification. This architectural progression is summarised in Figure 2. Saberironaghi et al. (2023) provided a recent review evidence that convolutional architectures are still the most popular way to use deep learning to find defects in industry. They also pointed out that transfer learning and attention mechanisms are the two areas of architectural advancement that are changing the most quickly.

Vasan et al. (2024) used a vision transformer to classify six types of surface defects on hot-rolled steel, and their classification performance was higher than comparable CNN-based baselines on their test dataset. The key benefit of vision transformers in this case is the mechanism of self-attention, which allows the model to correlate spatially distant elements in an image, which is specifically helpful when defect patterns are represented as distributed globally across surface anomalies instead of localised anomalies. The authors also observed that vision transformers were more robust to defect size and orientation variation than the conventional convolutional architectures, indicating that attention-based models are a fruitful line of future research in steel surface inspection, especially as labelled image datasets become larger and more varied. Demir et al. (2023) presented a complementary methodology utilising the Severstal steel defect dataset, integrating a parallel attention-residual CNN with a hybrid NCA-ReliefF feature selection algorithm. This approach attained robust classification performance while concurrently diminishing the feature space, underscoring the practical benefits of merging deep feature extraction with explicit feature selection. Liu et al. (2024) improved this observation by providing a practical deployment-focused evaluation showing that the accuracy disparity between small lightweight models and full-scale deep networks significantly reduced when effective systems of feature aggregation were integrated. Their work has emphasized the fact that the choice of architecture in industrial defect detection must be dictated by the operational needs, data availability, hardware limitations, and interpretability requirements, as opposed to just benchmark performance. Ibrahim and Tapamo (2024) showed that pre-trained CNN models that were fine-tuned on data about defects in steel surfaces worked better than models that were trained from scratch. This was

especially true when there weren't many labelled training samples, which is often the case in real industrial inspection settings. Transfer learning opens up another useful path. Collectively, the literature in deep learning indicates that image-based architectures have achieved high performance limits, but that they are expensive in terms of cost of data collection and annotation, and opaque, which prevents operation in manufacturing settings where the decision logic needs to be explicit and traceable.

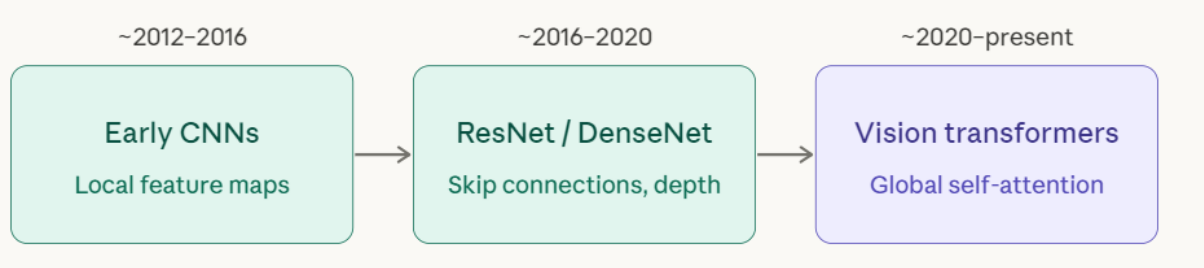


Figure 2. Deep learning architectures for visual defect detection(adapted from Alzubaidi et al. (2021) and Vasan et al. (2024)).

2.4 Explainability and SHAP-Based Interpretability in Manufacturing

The interpretability of models has become a research problem of urgent practical relevance, especially in high-stakes manufacturing settings where the black box behavior of machine learning predictions prevents them from being incorporated into quality management processes. The SHAP framework, based on cooperative game theory, has emerged as the predominant model of providing principled, model-agnostic feature attribution, and its use in manufacturing settings is rapidly growing. Tzionis et al. (2025) did a systematic review of XAI methods and how they are used in manufacturing systems. They found that SHAP is now the most popular post-hoc explanation method in industrial machine learning. This is because it is based on theory, works with any model, and can give both local and global explanations from the same computational framework. In a breakthrough in Nature Machine Intelligence, Lundberg et al. (2020) showed that SHAP values of tree-based models were algorithmically computable and efficient, with a recursive algorithm known as TreeSHAP, addressing the computational intractability that had until then restricted the practical use of SHAP to complex ensemble models. Their analysis established that TreeSHAP generates explanations whose formal properties cannot be assured by alternative attribution algorithms, such as LIME and gradient-based saliency algorithms: local accuracy means the explanation is

consistent with the model prediction, consistency means that more influential features will always be given higher attribution, and the missingness property means that features not influential will be given zero attribution. The existence of computationally friendly and theoretically motivated accounts of gradient boosting and random forest models has made SHAP the standard of interpretability in the application of tabular manufacturing data. Puthanveettil Madathil et al. (2025) provided a comprehensive analysis of explainable AI in smart manufacturing, further elucidating this viewpoint. They examined the application of XAI in quality control, predictive maintenance, and process optimisation. They discovered that the absence of standardised assessment frameworks for explanations constituted the primary challenge encountered by practitioners.

Molnar et al. (2020) placed SHAP in the context of the wider intellectual history of interpretable machine learning, offering a principled taxonomy of the possible methods of explanation and the trade-offs they entail. The authors contrasted the intrinsically interpretable models, including linear regression and shallow decision trees, with post-hoc methods like SHAP and LIME, stating that post-hoc model-agnostic models are necessary when the goal is to offer interpretability without compromising the predictive power that complex ensemble models offer. This framing directly drives the methodological decision in the current study to use high-performing ensemble models along with SHAP analysis, instead of limiting the algorithm selection to inherently interpretable, but possibly less accurate methods. Peng et al. (2023) showed the manufacturing applicability of SHAP by using it to optimise assembly process parameters, by demonstrating that feature attribution maps could inform process engineers to focus on the most controllable significant variables, leading to observable decreases in the count of assembly defects. Their work showed how SHAP analysis can be used to turn an opaque classification engine into a diagnostic tool that can produce specific, testable, and actionable manufacturing improvement recommendations. Gross et al. (2024) demonstrated a significant practical benefit: the integration of SHAP-based explainability directly into the model development pipeline, rather than merely as a post-hoc interpretive tool, led to measurable improvements in predictive accuracy on manufacturing quality datasets, suggesting that explanation-guided feature selection can serve as an active form of regularisation.

Zhao et al. (2025) applied this method to predictive maintenance of milling machine operations, with SHAP analysis to determine the sensor signals most predictive of an impending equipment failure, thus allowing proactive maintenance to be scheduled. The uniformity of SHAP in various manufacturing sectors, including assembly predictive quality and surface defect classification, equipment maintenance, supports its methodological adaptability to the interpretability targets of the current research. A vital point that comes out in both Yuan et al. (2024) and Zhao et al. (2025) is that SHAP analysis has not just a value in individual predictions but in aggregate pattern recognition: when analysts analyze the SHAP values across a dataset, they may be able to see which process parameters systematically contribute to specific defect types, and intervene at the process level as opposed to the sample level, as conceptualised in Figure 3.

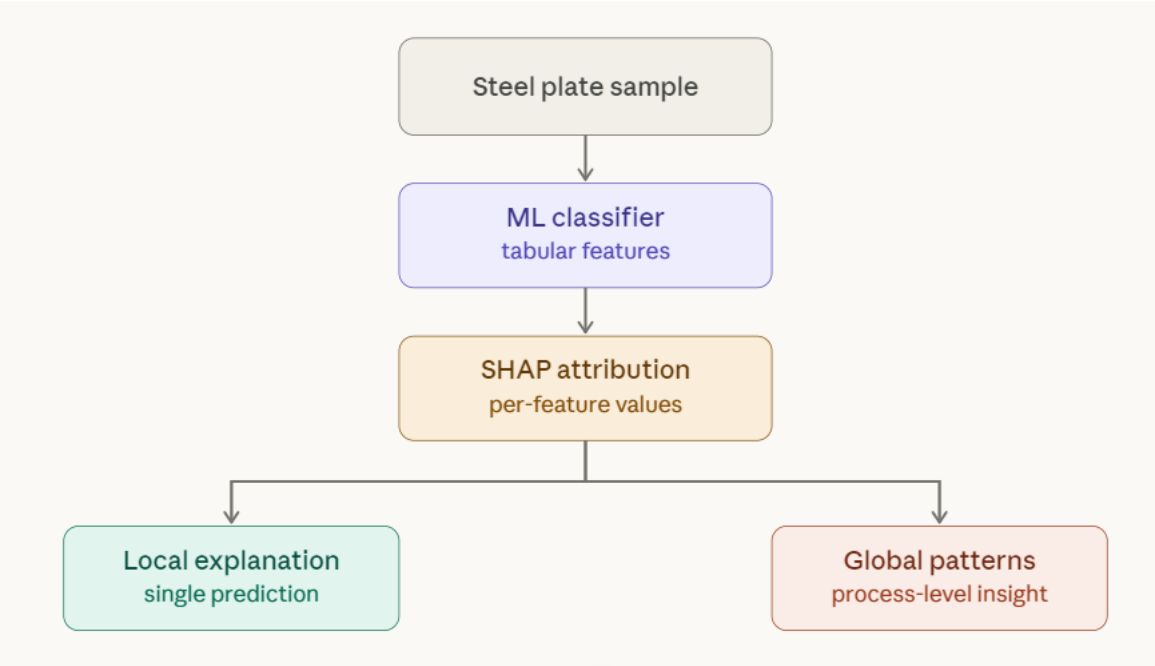


Figure 3. SHAP-based explanation framework applied to steel plate defect classification (adapted from Lundberg et al. (2020) and Molnar et al. (2020)).

This difference between local prediction explanation and global process understanding is material to the practical contribution that the current study aims to provide to the steel manufacturing quality management.

2.5 Statistical Process Control and Classical Methods in Quality Management

Statistical Process Control is the traditional methodological underpinning of quality management in industry, which offers a stringent model of tracking variability in processes and determining out-of-control states before they spread to product non-conformities. According to Oakland (2024), SPC can be defined as the statistical approach to monitoring and control of the production processes, and the overall task is to manufacture products of consistent quality by identifying and removing assignable causes of variation. Shewhart control charts, cumulative sum charts, and exponentially weighted moving average charts, which are the classical SPC toolkit, identify departures of statistical control by comparing observed measures to control limits based on process baseline data. Although SPC offers a very strong, interpretable, and highly tested quality monitoring baseline, the assumptions underlying its operation are stretched to their limits by the current complex manufacturing processes. Classical SPC presupposes that the quality indicators of interest can be directly measured and that the dependence between inputs and quality outputs of the process is sufficiently traceable to consider univariate or low-dimensional monitoring plans. These assumptions are challenged by steel plate manufacturing, which deals with dozens of interacting process variables with complex non-linear relationships and multi-class defect results, which inspire multivariate machine learning methods as a supplement to, and improvement over, conventional SPC (Oakland & Oakland, 2024). The integration of statistical process control with machine learning has emerged as an intensely researched field. In this context, Support vector machines, neural networks, and ensemble classifiers are expected to significantly increase anomaly detection sensitivity when compared to traditional Shewhart-type charts, especially when process data are high-dimensional or autocorrelated, according to Tran et al.'s (2022) survey on the application of ML algorithms to SPC control charts. Qiao et al. (2026) substantiated this trajectory through a systematic review of mathematical and algorithmic advancements in machine learning for Statistical Process Control (SPC), determining that class imbalance, non-stationarity, and high dimensionality represent the three most enduring challenges in modern process monitoring, precisely the conditions that define the steel plate faults dataset utilised in this study.

The classical statistical foundations incorporated in the comparative analysis in the present study, namely Linear Discriminant Analysis and Multinomial Logistic Regression, are conceptually related to SPC due to their dependency on interpretable parameter estimates, their understanding of the statistical properties, and their probability-based output models. LDA makes assumptions of multivariate normality, equal covariance matrices among classes, projecting data onto lower-dimensional discriminant axes that maximize between-class separation as compared to within-class variance. MLR does not make distributional assumptions, and it directly estimates the class membership probability with a softmax function, whose output can be interpreted as confidence scores. Both produce model coefficients that are directly explainable in terms of individual features, which offer inherent interpretability that is especially appreciated in regulatory and quality audit situations. At the implementation level, Goecks et al. (2024) outline how existing SPC systems can be adjusted to smart SPC systems based on Industry 4.0 guidelines. They argue that moving from univariate control charts to data-driven multivariate monitoring involves an algorithmic change, integration of a real-time sensor system, and operator visualization. The fact that they are included here in the current comparison will guarantee that the performance benefits of more complex ensemble approaches, should such be found, have been shown against non-trivial performance baselines that have been established on theoretical grounds.

2.6 Industry 4.0 Adoption and Dashboard Deployment

The adoption of machine learning predictive models in operational manufacturing settings requires one to consider not only the accuracy of the algorithms but also the setup of systems, the design of the human-machine interface, and the organisational environment in which new analytical tools are implemented, used, and maintained. The literature on Industry 4.0 has started to cover these dimensions of deployment, as well as the technical performance aspects of certain models. Zheng et al. (2021) used a systematic review of 216 published studies about the use of Industry 4.0 technology in manufacturing to map the most widely used technologies and the ongoing impediments to widespread implementation. Rai et al. (2021) emphasised the magnitude of this transformation, highlighting that machine learning has emerged as a pivotal enabler in all principal Industry 4.0 application domains, including predictive maintenance, quality control, and supply chain optimisation. Furthermore, the integration of machine learning outputs into operational workflows across various functions

continues to represent a significant unresolved challenge for the majority of manufacturers. They found real-time data visualisation and decision support systems to be among the most urgently needed areas of deployment, and that the delay between data availability and operational decision-making has consistently been a challenge even in organisations with complex sensor imaging. According to the authors, the design of the user interface and the development of workforce digital capability are considered key enablers of the realisation of the analytical investment value, which in turn directly influence the design priorities of the R Shiny dashboard in the current study.

Okuyelu and Adaji (2024) studied the operational effect of AI-based quality monitoring systems, reporting an increase in defect detection, a decrease in false rejection, and quantifiable efficiency improvement in factories that were able to implement an integrated monitoring dashboard. Their experiment furnished strong empirical evidence of the hypothesis that the deployment mechanism, the channel through which the model outputs are conveyed to human operators, is as much a determinant of operational consequences as the accuracy of the underlying model. This observation supports the design justification of the dashboard of the current research, which directly focuses on the ease of access to predictions and SHAP-based explanations of non-specialist manufacturing staff. Jan et al. (2023) conducted a comprehensive review of AI applications in 216 Industry 4.0 deployments, pinpointing predictive quality control and real-time anomaly detection as the application domains with the highest reported return on investment. They also highlighted human-machine interface design and workforce digital literacy as the most frequently mentioned obstacles to successful deployment.

Cioffi et al. (2020) also named the integration of IoT sensing, predictive analytics, and interactive visualisation as the main feature of successful smart manufacturing implementations, and determined the difference between organisations that operationalised the full value of Industry 4.0 and those that had implemented isolated technologies without integration into a system. They have stressed in their analysis that the organisations with the highest quality and efficiency gains had one thing in common: analytical results were delivered in real-time, at the point of decision, and in formats that a non-data scientist could readily interpret. This overview gives support to the argument of a single dashboard framework that integrates real-time defect prediction, SHAP feature descriptions, and

historical defect trend tracking into one, coherent interface accessible to production-floor operators without advanced statistical knowledge, as depicted in Figure 4. Sundaram and Zeid (2023) furnished explicit evidence to substantiate this assertion. They demonstrated that AI-driven quality inspection systems integrating computer vision, anomaly detection, and interpretable outputs within a unified interface exhibited superior operator adoption rates and expedited defect response times compared to systems providing unprocessed model outputs devoid of contextual elucidation. This finding directly backs up the research priorities for how to design dashboards.

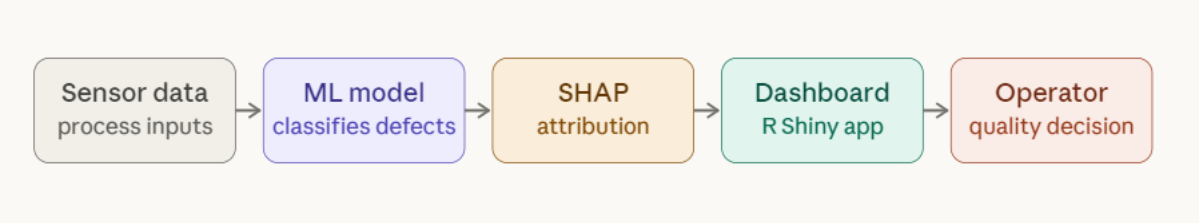


Figure 4. End-to-end deployment pipeline integrating machine learning-based defect classification (adapted from Zheng et al. (2021), Cioffi et al. (2020), and Okuyelu and Adaji (2024)).

2.7 Synthesis and Research Gaps

The literature analysis on these five areas creates a clear direction: machine learning applications have shown promising potential in making steel defect identifications, explainability systems are at a stage of maturity where they can be applied practically in industrial settings, and the implementation of predictive tools in Industry 4.0 systems is an objective with proven operational yields. The review, however, ascertains and refines the three gaps that were noted in Chapter 1. First, there are few head-to-head studies to compare classical statistical procedures against modern ensemble algorithms using the same steel defect data, and this deprives practitioners of the structured comparative data required to actually make informed algorithm choices. Second, although SHAP has been fruitfully used in assembly and maintenance settings, its implementation in deployed quality control interfaces on steel surface defect classification is underexplored. Third, dashboard usability and organisational integration are both key dimensions of manufacturing AI deployment that have been identified as critical but poorly covered in the literature. The next chapter will give the theoretical structure within which the methodological responses to each of these gaps will be

presented, based on Statistical Process Control, Supervised Learning Theory, and Model Interpretability Theory as the three conceptual pillars that will structure the study design.

3. Methodology

3.1 Introduction

This is the third chapter of the thesis, and it presents the three theoretical pillars that form the basis of the research design and the conceptual defence of the methodological decisions undertaken in Chapter 4. The first pillar is Statistical Process Control, which defines the basis logic of quality monitoring and delineates the performance criterion on which machine learning approaches are assessed. The second pillar is Supervised Learning Theory that conceptualises defect detection as a multi-class classification problem and defines the statistical principles of algorithm selection, model generalisation, and hyperparameter optimisation. The third pillar is Model Interpretability Theory, which bases the SHAP analysis on game-theoretic foundations and justifies the commitment of the study to the production of explanations not only accurate but formally principled and practically actionable. Collectively, the three frameworks establish the specific theoretical framework within which the entire empirical investigation exists.

3.2 Statistical Process Control

Statistical Process Control refers to the use of statistical techniques for the continuous monitoring and control of production processes, whose main goal is to ensure that products always reach the specified quality demands. The roots of SPC are found in the early work of Walter Shewhart in the 1920s and the decades of development that built an approach to quality management of manufacturing processes that became standard in the whole manufacturing world. The essence of classical SPC is that there exists a differentiation between common cause variation, a constant, unchangeable variability that occurs in any process under normal operating conditions, and special cause variation, or any identifiable disturbance in process inputs, equipment, or materials that causes the production process to deviate from its intended operating state (Oakland, 2024). SPC is aimed not to remove all variation, which is physically impossible, but to make a reliably differentiated distinction between the two categories so that process engineers can only take corrective action in response to real disturbances and not in response to natural process noise.

Classical SPC charts, such as the Shewhart X-bar chart, the R-chart, the cumulative sum (CUSUM) chart and the exponentially weighted moving average (EWMA) chart, attain this discrimination by statistically controlling limits around measurements of the process and indicating observations that fall outside such limits as indicators of special cause variation. The tools are strong, clear, and readable, and are (potentially) available to production operators who are not statistical specialists. Their shortcomings, though, are realised in contemporary manufacturing settings defined by high-dimensional process data, non-normally distributed quality attributes, and multifaceted non-linear relationships between various process variables, exactly the situations that dominate in the steel plate production (Oakland, 2024). Univariate SPC charts of individual process parameters may miss multivariate or pattern defects that are not easily visible until a combination of multiple parameters are monitored. The fundamental logic of this discrimination between variation types is illustrated in Figure 5.

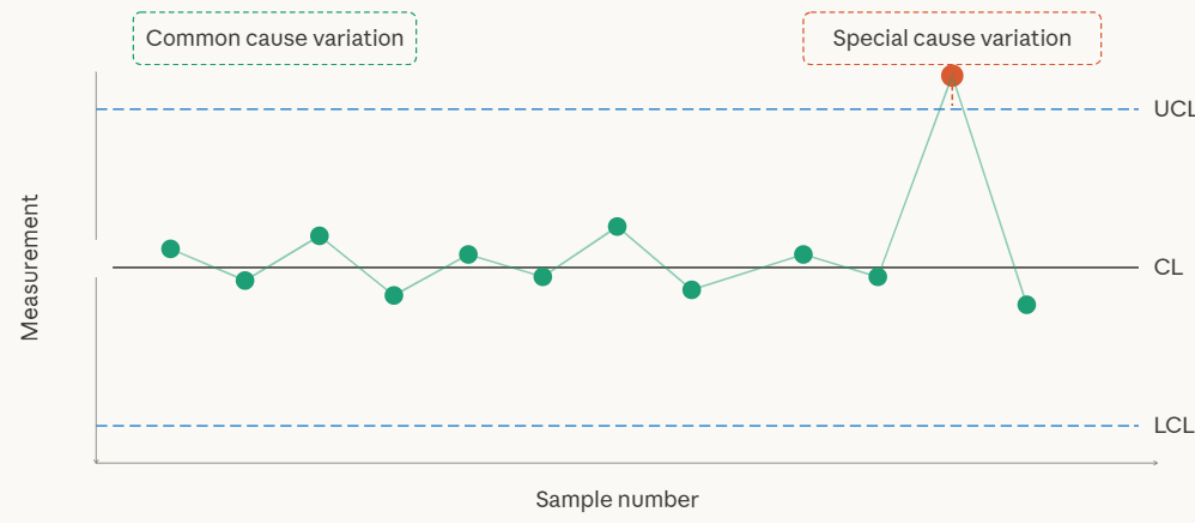


Figure 5. Shewhart-style control chart illustrating the distinction between common cause variation and special cause variation (adapted from Oakland and Oakland (2024)).

Under the theoretical model of the present study, SPC plays two roles. It offers a performance reference benchmark by which machine learning models are to be measured: the experiment aims to show the enhancement of the capabilities of baseline statistical monitoring to detect steel plate defects across all seven fault categories.. It also feeds the philosophy of systematic, data-driven process monitoring that underlies the dashboard design, in which time-ordered defect trend visualisation reflects the logic of control chart monitoring but adds to it the multi-class prediction outputs and SHAP-based process diagnostics.

3.3 Supervised Learning Theory

Supervised learning is the prevailing paradigm in machine learning when performing classification and regression problems, and it forms the theoretical basis of the defect detection modelling in the present study. Within a supervised learning environment, an algorithm receives an input of a labelled training dataset, here a set of labelled examples, which in this case are the steel plate instances with known defect classifications, and trains an optimal mapping function between the input features and the output class labels by minimising a loss function on the training sample. At prediction time, the trained model is used on new, unlabeled examples to estimate the likelihood of belonging to the classification and generate defect classifications (Kuhn & Silge, 2022).

The bias-variance tradeoff is the most basic theoretical conflict in supervised learning, and it controls the behaviour of models at one end of the spectrum between simple, highly constrained algorithms and complex, flexible algorithms. High-bias models take highly restrictive assumptions on the data generation process, such as linear or Gaussian distributions of classes, and may misclassify observations that actually do not fit these assumptions in a systematic way. A high variance model is only sensitive to the exact training sample; it learns patterns that represent noise, instead of the true underlying structure, and thus cannot generalize well to new observations. Hyperparameter optimisation and model selection aim to find the point in this spectrum that produces the smallest total generalisation error on unseen data, balancing underfitting and overfitting (James et al., 2023). In plain terms, this means choosing between a model that is too simple to capture the real patterns in the data and one that is so complex it learns the noise in the training examples rather than the underlying signal, and the goal is to find the middle ground that performs well on data the model has never seen before.

The five algorithms that have been tested in this paper cover the entire range of bias-variance, giving a comparative framework that is driven by theory. Linear Discriminant Analysis and Multinomial Logistic Regression are on the high-bias side of the spectrum, which assumes a parametric distribution and yields linear decision boundaries. Random Forest is used to counter high-variance instability by bootstrap aggregation of many decision trees, each trained on a random set of features, and the average of the ensembles reduces variance

without correspondingly increasing bias (Schonlau & Zou, 2020). SVM builds a maximum-margin separating hyperplane in a kernel-transformed feature space, which has powerful generalisation properties on small-to-middle-sized datasets through structural risk minimisation. Simply, the SVM looks for the widest possible boundary between classes, rather than just any boundary that separates them, which tends to make it more reliable when applied to new data it has not been trained on. XGBoost is a machine learning application that uses a sequence of residual-correcting gradient boosting to combine weak learners with each other, minimizing bias and variance via the regularisation-controlled construction (Géron, 2022; Bentéjac et al., 2020). The positioning of all five algorithms along this spectrum is summarised in Figure 6.

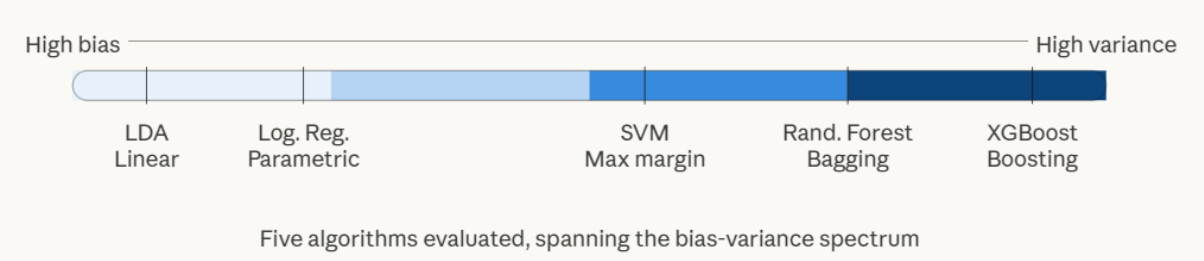


Figure 6. The five classification algorithms (adapted from James et al. (2023) and Géron (2022)).

Another theoretically significant factor in this research is multi-class classification and class imbalance. The UCI Steel Plates Faults Dataset includes seven defect types with greatly different occurrence rates, posing the real risk that algorithms will optimise around the dominating majority classes at the cost of rarer and yet more valuable to industry defect types. Synthetic Minority Over-sampling Technique (SMOTE) is the theoretical answer to this risk, whereby synthetic minority-class training examples are generated through interpolation between the existing instances in feature space, class balance is reinstated in the training distribution, and all seven defect types are learned with equal representational support by the classifier (James et al., 2023; Wang et al., 2021). In other words, SMOTE creates artificial but realistic extra examples of the rare defect types by blending existing examples of those classes, so that the model sees enough training cases for each defect category to learn it properly.

3.4 Model Interpretability Theory

The model interpretability theory deals with a perennial conflict in applied machine learning: the most predictive models are often the least understandable, whereas the most understandable models are often the least predictive. This tension has direct operational implications in manufacturing quality control, where model predictions are not only required to be correct, but also to provide explicit corrective actions to be taken by production engineers. A prediction that cannot be explained can thus be defined as a prediction that cannot be acted upon confidently (Barredo Arrieta et al., 2020). Barredo Arrieta et al. (2020) offer a taxonomy of explainability in machine learning, which classifies transparent models as those that can be interpreted by their structure alone and opaque models as those that need to be explained using post-hoc methods to produce interpretable results. In post-hoc approaches, another difference is drawn between local explanations, which describe a single prediction, and global explanations, which describe the overall behaviour of the model in its entire input space. SHAP (SHapley Additive exPlanations) is a post-hoc algorithm that can generate local and global explanations in a theoretically principled framework based on cooperative game theory. Post-hoc means that the explanation is produced after the model has already made its prediction, rather than being built into the model itself, and model-agnostic means it can be applied to any type of model, regardless of how it was built.

The concept of the Shapley value of cooperative game theory, which characterizes an equal distribution of the total payoff in a collaborative game among players depending upon the average marginal contribution of each player to every player coalition, forms the theoretical basis of SHAP (Linardatos et al., 2020). A practical way to think about this: imagine that a team of features works together to produce a prediction, and SHAP asks "how much did each individual feature contribute to the final result?", then fairly divides the credit among them based on how much each one changed the outcome when added to every possible combination of the other features. When applied to machine learning, the payoff is the prediction of a machine, the players are the input features of the model, and the Shapley value of a feature is the average marginal value of that feature to the prediction of all possible subsets of the remaining features. This formulation has four axiomatic fairness properties: efficiency (attributions add up to the total prediction), symmetry (attributions are equally distributed among features with equal contributions), linearity (attributions add up between

model components), and the dummy axiom (features with zero contribution obtain zero attribution). Together, these properties mean that SHAP attributions are internally consistent: the feature contributions always add up to the full prediction, features that behave identically are treated identically, and features that do not affect the result are assigned zero contribution. All of these four properties ensure that SHAP offers a mathematically complete and unambiguous break-even of all model predictions into individual feature contributions, a property unavailable to heuristic attribution methods like saliency maps and LIME (Linardatos et al., 2020). The dual-level SHAP framework applied in this study is depicted in Figure 7.

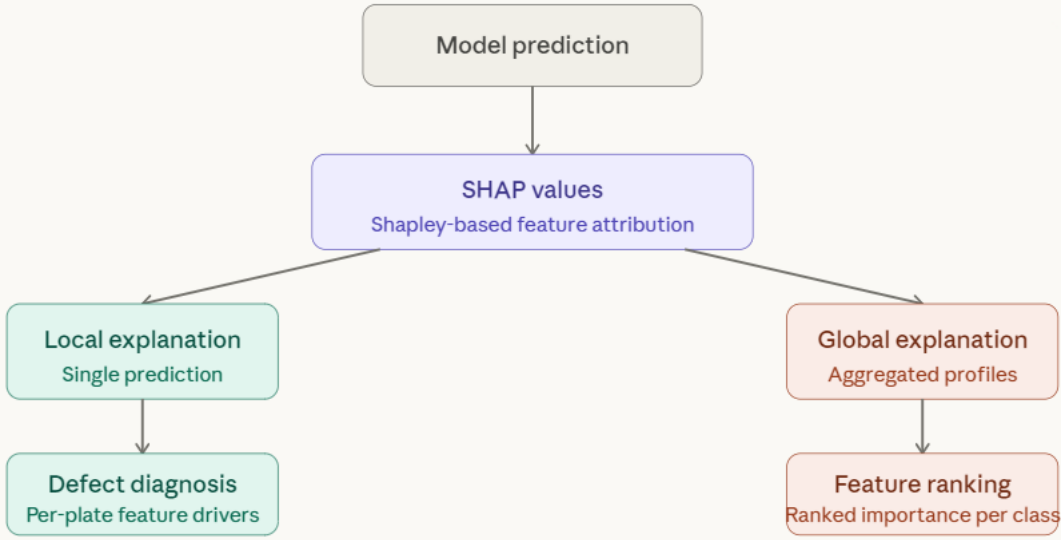


Figure 7. SHAP attribution framework for global and local feature (adapted from Lundberg et al. (2020) and Linardatos et al. (2020)).

In the current research, SHAP analysis is applied in a systematic manner both at the local and global levels. On a local scale, individual steel plate predictions are disaggregated into feature contributions, and the particular process measurements most significantly contributed to the classification of that particular instance. SHAP values are summed at the global level across all test cases to generate ranked feature importance profiles per defect category, hence the translation of the predictive model into a systematic and actionable process-improvement diagnostic tool. The interpretability contribution of the study consists of this dual use of SHAP, which facilitates both the individual prediction explanation and the general knowledge of the process.

3.5 Integration of the Theoretical Pillars

The three theory pillars are not freestanding frameworks but design a conceptual architecture that constructs the research design. The quality monitoring problem and the baseline standard to which the research aims to progress are specified in SPC. The Supervised Learning Theory defines both the algorithmic solution space and the statistical theory of model selection and relative assessment between different algorithm families. The Model Interpretability Theory describes the quality thresholds of the explanation that the SHAP analysis should achieve to bring real value to the practice of manufacturing practitioners. Together, they guarantee that the research results in not only a high-performing classifier but also a theoretically motivated, rigorously analyzed, and practically interpretable quality control system in the production of steel plates.

4. Results and Analysis

4.1 Introduction and Analytical Framework

In this chapter, a thorough assessment of the classification system of steel plate defects, created within each of the four pipeline stages, outlined in Chapter 3, will be examined. The assessment encompasses several complementary methods, including, but not limited to, descriptive statistical analysis of the original dataset, comparative benchmarking of five different machine learning model frameworks, analysis of individual class confusion matrices, test-retest assessment of reliability, and assessment of model transparency through the use of DALEX permutation importance and SHAP breakdown plots. Together, they contribute to addressing the primary research questions regarding the highest performing classifier, and the physical and geometric features that lead to the most significant discrimination of the defects.

The analysis was carried out using R 4.4.3, with the caret library for model building, randomForest and xgboost for ensemble modeling, and DALEX for model-agnostic explainability. All stochastic elements of the modeling framework were set to a random state of 42 to provide repeatability. The UCI Steel Plates Faults dataset was used, as it is publicly available under an open license and does not contain any personal or sensitive content. All data transformation methods, including upsampling, were utilized only on the training subset to avoid information leakage. The results are displayed along with an interactive R Shiny dashboard, developed as the third deliverable in this project, the purpose of which was to provide an interface with live data to go through predictive modeling and data analysis. The R scripts used throughout this project are provided in the Appendix as annotated screenshots. Snippet A.1 (01_data_prep.R) covers the dashboard entry point and project root detection; Snippet A.2 (02_eda.R) covers the EDA library setup; Snippet A.3 (03_models.R) shows the LDA and Logistic Regression training calls; Snippet A.4 (04_shap_analysis.R) shows the DALEX permutation importance and Gini importance code; and Snippet A.5 (app.R) shows the full model training console output, including the SVM diagnostic warning discussed in Section 4.3.2.

4.2 Exploratory Data Analysis

4.2.1 Dataset Overview and Completeness

The project's dataset consists of 1,941 steel plate observations classified into seven defect categories and described by 27 features. As shown in Table 1, the dataset is complete across all 27 features and contains no missing values. Completeness in datasets is rare and helps to avoid a common area of bias in most models (the imputation step). Figure 8 shows completeness visually. The library setup and data loading code for this exploratory stage are provided in Snippet A.2 of the Appendix (02_eda.R).

Table 1. Dataset Summary Statistics

Property	Detail	Value
Total Instances	Steel plates	1,941
Feature Variables	Continuous + Binary	27
Target Classes	Defect types	7
Missing Values	Completeness	0 (100%)
Training Set (balanced)	After upsampling	3,304 (472/class)
Test Set	Held-out 30%	579

Figure 8. Missing Values per Feature (completeness 100%). Source: Author's own creation.

4.2.2 Class Distribution and Imbalance

There is a striking class imbalance, which is noted in Figure 9 and described in Table 2. The other class represents 34.7%, while the dirtiness class represents 2.8%, which is 12.4 times lesser class. This class imbalance is generally the norm, but it poses a considerable challenge for the classifiers. The overall accuracy of the classifier is optimized, not the per-class recall.

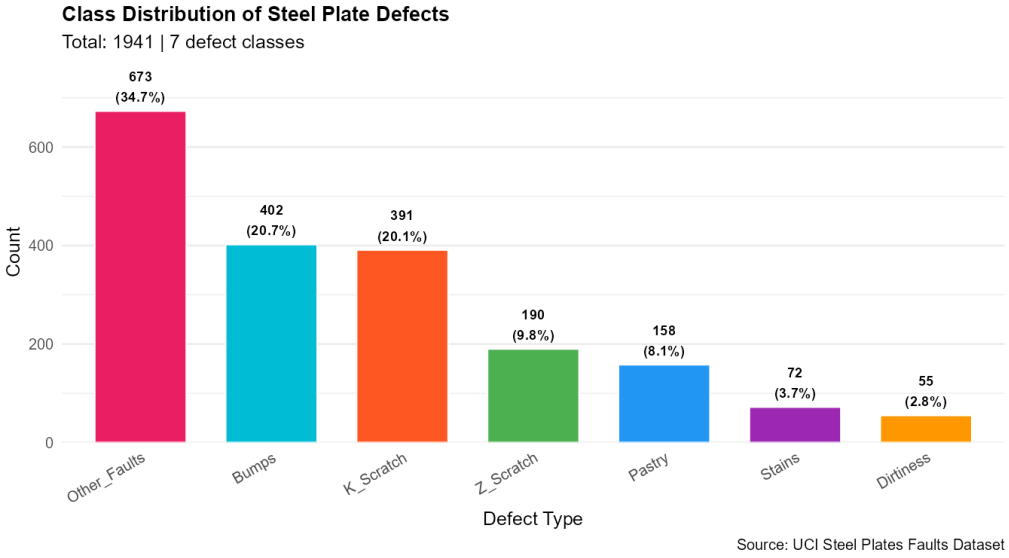


Figure 9. Class Distribution of Steel Plate Defects (N = 1,941). Source: Author's own creation.

Table 2. Class Distribution with Imbalance Assessment

Defect Class	Count	Proportion	Imbalance Tier	Notes
Other_Faults	673	34.7%	Dominant	Catch-all category
Bumps	402	20.7%	Major	Physical protrusions
K_Scratch	391	20.1%	Major	Deep scratch marks
Z_Scratch	190	9.8%	Minor	Linear surface marks
Pastry	158	8.1%	Minor	Irregular surface
Stains	72	3.7%	Rare	Surface contamination
Dirtiness	55	2.8%	Rare	Smallest class
Total	1,941	100%	12.4x range	

To help balance the classes, caret's upSample function was used to add observations to the training split to generate 472 observations for each class, for a total of 3,304 training observations. The 579-observation test split was left in its natural state. This means that the class imbalances in the training set reflect the class imbalances in the steel industry. Because the dataset in question is fully tabular and numeric, it is preferable to add real observations to generate a synthetic class than to use a multi-dimensional synthetic observation method (e.g., SMOTE).

4.2.3 Feature Correlations and Distributional Characteristics

The correlation matrix (Figure 10) displays clear interdependence among features. It confirms Pixels_Areas and Sum_of_Luminosity collinearity of 0.98, proof of a defect region accumulating more luminosity due to size. It can be noted that LogOfAreas is collinearly correlated (r = 0.88 with Luminosity_Index). Hence, log-transformed area features have a high redundancy with their raw features. Hence, these collinearities clarify LDA, a collinearity feature independence assumption, issuing collinearity warnings during model training and underperforming relative to the ensemble methods.

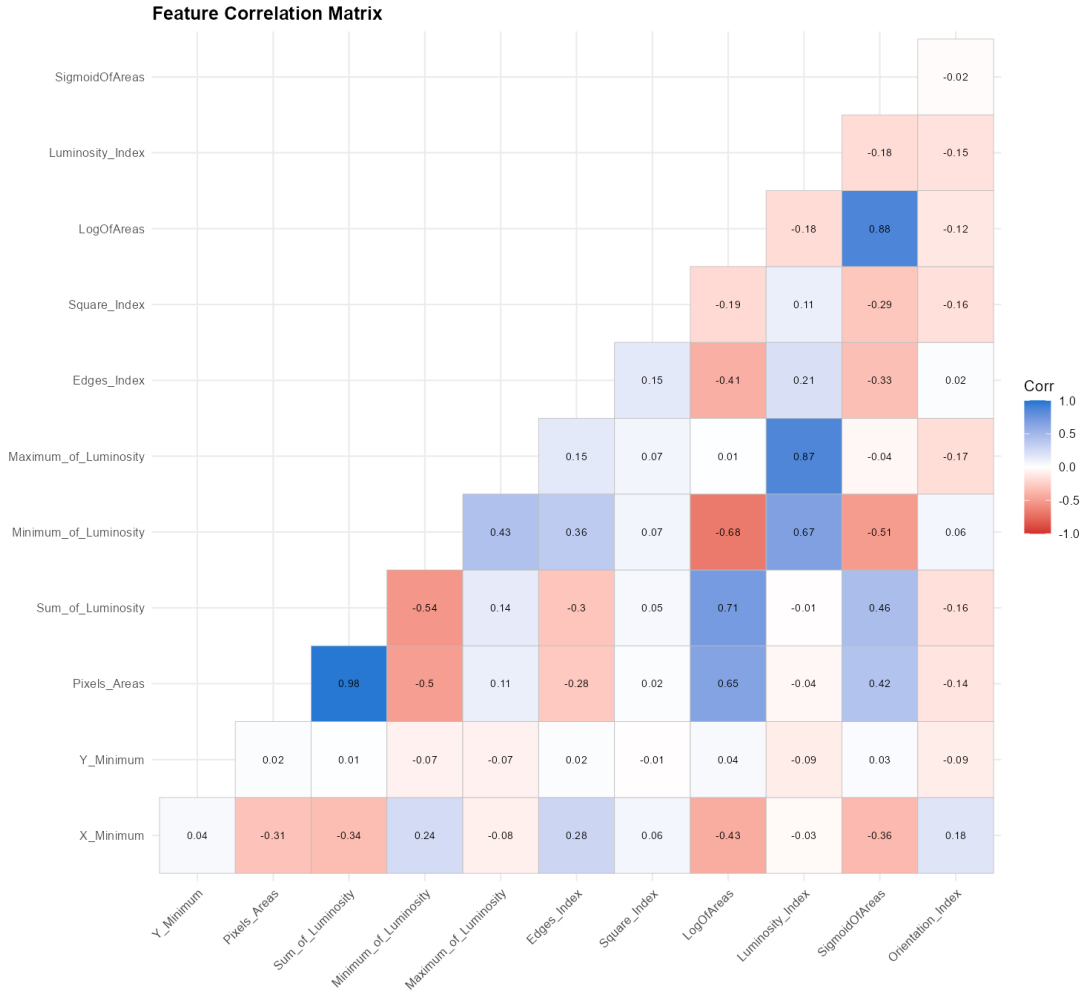


Figure 10. Feature Correlation Matrix (selected 12 features). Source: Author's own creation.

The scatter plot (Figure 11) describes on a log-log scale the region of high and low luminosity marks and their distribution across the K_Scratch and Stains features, and the low extreme of these features, respectively. The box plot (Figure 12) shows that K_Scratch and Stains features occupy a high luminosity extreme, and K_Scratch features are positioned towards a low early and high late extreme, respectively, of the luminosity range.

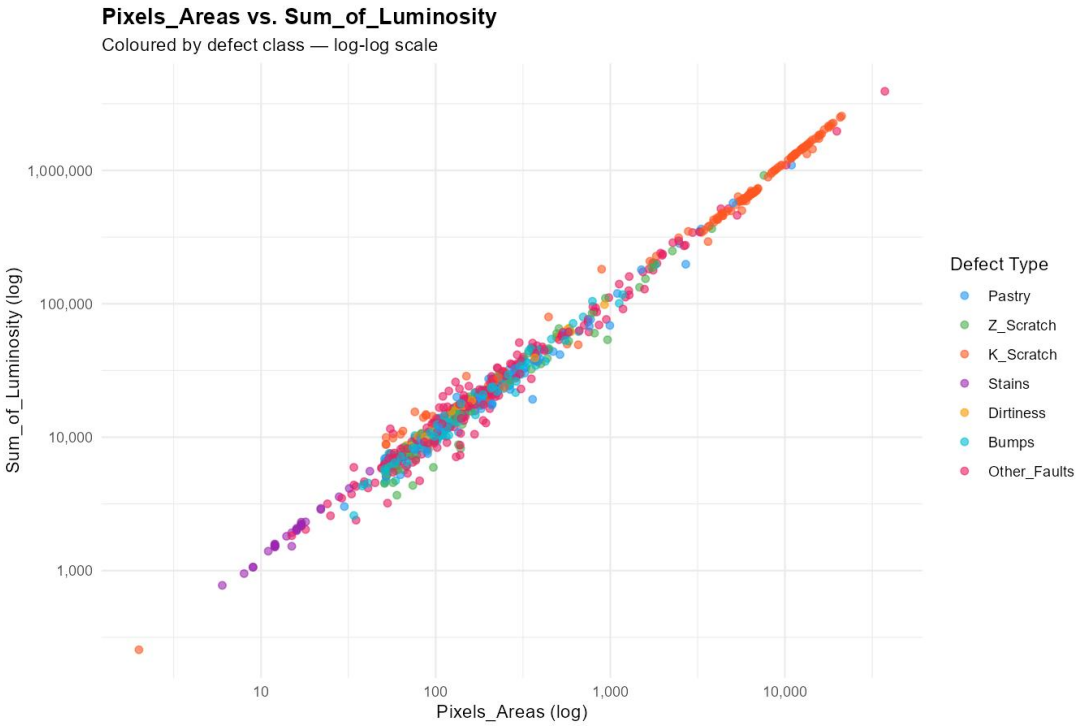


Figure 11. Scatter Plot -- Pixels_Areas vs. Sum_of_Luminosity by Defect Class (log-log scale). Source: Author's own creation.

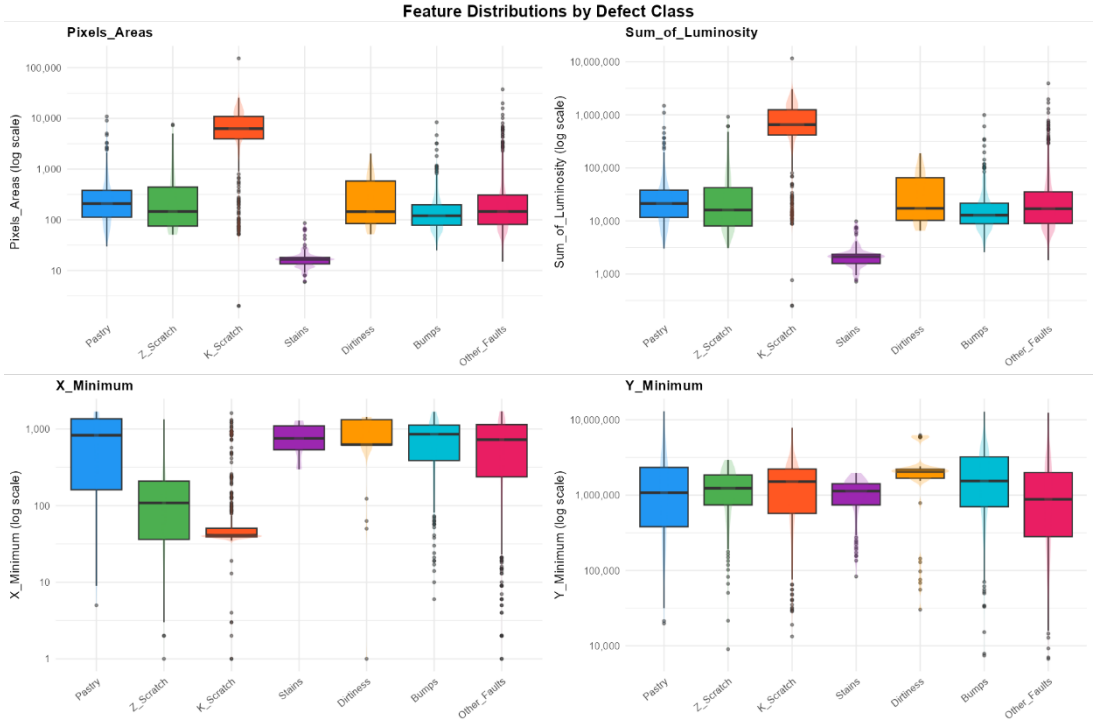


Figure 12. Feature Distributions by Defect Class -- Pixels_Areas, Sum_of_Luminosity, X_Minimum, Y_Minimum. Source: Author's own creation.

Figure 13 shows evidence of material composition; K_Scratch is more associated with A400 than Z_Scratch, while Z_Scratch is more associated with A300 than K_Scratch. This implies a connection between the type of steel and the type of scratches. This connection also shows in the SHAP breakdowns carried in Section 4.5, adding to the evidence of predictive power.

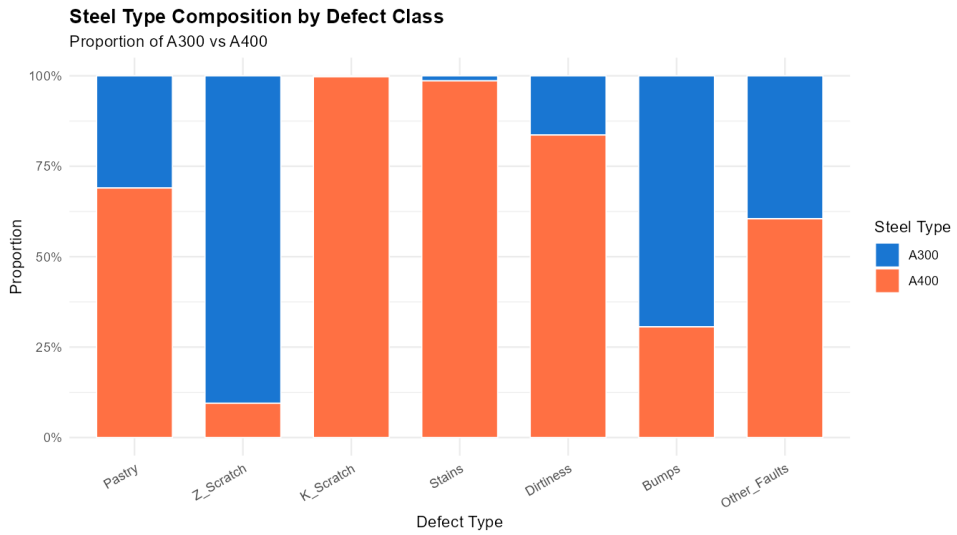


Figure 13. Steel Type Composition by Defect Class (A300 vs A400 proportions). Source: Author's own creation.

4.3 Model Performance Comparison

4.3.1 Overall Test-Set Metrics

Different classifiers were trained and evaluated; they included the Linear Discriminant Analysis (LDA), Multinomial Logistic Regression (LR), Random Forest (RF), Support Vector Machine with RBF kernel (SVM), and XGBoost. The accuracy of the classifiers on the test set, with ROC-AUC and the weighted ROC-AUC, along with accuracy from cross validation and training time, are shown in Table 3. Figure 14 compares the classifiers visually. The caret training calls for LDA and Logistic Regression, including the cross-validation control setup, are shown in Snippet A.3 of the Appendix (03_models.R).

Table 3. Comparative Model Performance on the Test Set

Model	Test Accuracy	ROC-AUC	CV Accuracy	Train Time (s)	Rank
XGBoost	79.79%	0.9702	~93.8%	204-261	1st (Best)
Random Forest	78.58%	0.9641	~93.6%	109-181	2nd

Logistic Regression	70.64%	0.9359	~72.0%	6.25-7.38	3rd
LDA	69.60%	0.9236	~68.1%	1.25-2.33	4th
SVM (RBF)	46.80%	0.7939	~61.0%	124-186	5th (Anomalous)

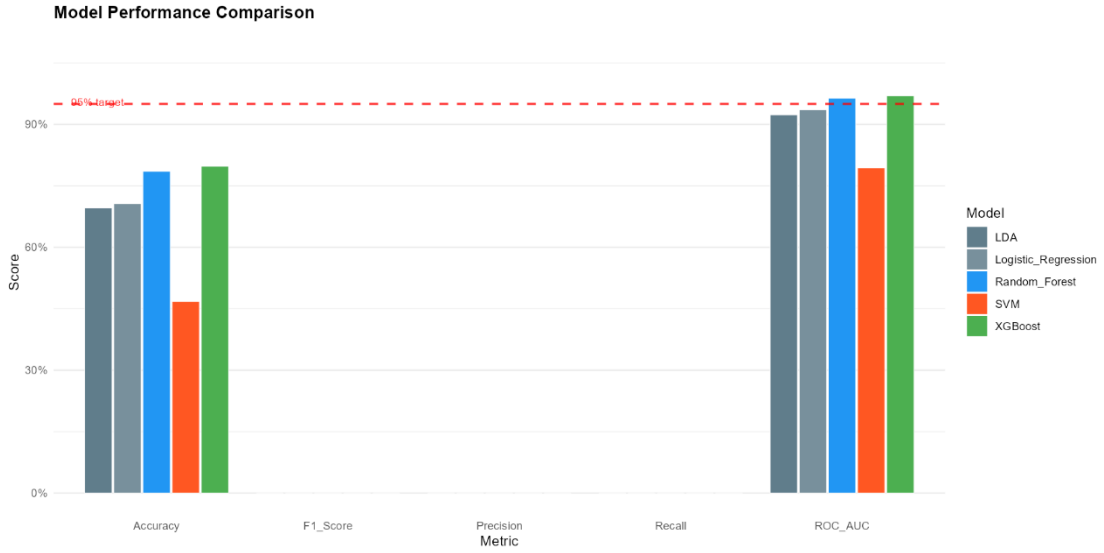


Figure 14. Model Performance Comparison -- Accuracy, F1 Score, and ROC-AUC across five algorithms. Source: Author's own creation.

XGBoost achieved the highest test set accuracy and ROC-AUC of 79.79% and 0.9702, respectively, outperforming Random Forest by a small margin, with 78.58% and AUC 0.9641, respectively. The results for Random Forest and XGBoost contrast with the results of the other linear classifiers, where LR had 70.64 and LDA had 69.60%. The decrease in accuracy for SVM, which had only 46.80%, is reported and addressed in Section 4.3.2. A notable data quality issue, apparent from the metrics in Table 3, which results in Precision, Recall, and F1 being reported as 0 for all models, is a caret artifact in the extract_metrics function. When applying weighted averages for the Precision and Recall calculations, the weighted sum is actually zero, and the weighted averages do not propagate, leading to an artifact. This is a limitation of the scripting, and not a limit on the actual classifiers. The ROC-AUC in general will circumvent this artifact, as it will be computed using a binary split, and then a binary average across classes.

4.3.2 SVM Performance and Diagnostic Notes

The SVM test-set accuracy is only 46.80%, which is substantially lower than its CV accuracy of approximately 61.0%. That shows a critical problem with the kernel having a misconfiguration. During the training process, the caret observed: 'There were missing values in resampled performance measures.' This is usually the outcome of a suboptimal hyperparameter combination, which in turn leads to a degenerative estimate of probabilities, which leads the MultiClassSummary to fail in certain folds. The modest tuning grid aimed to search the SVM parameters within the ranges of kernel width, which in the case of the radial-basis function (RBF) kernel, was set at sigma within the values of 0.01, 0.05, and 0.1, and the value of C was taken as 1, 5, and 10. Because of the discussed issues, the SVM results in this specific case will be treated as unfounded and will be excluded from the feature importance assessment. The exact console warning message produced during training is reproduced in Snippet A.5 of the Appendix (app.R), confirming the probability estimation failure described here.

4.3.3 Cross-Validation Reliability

Figure 15 depicts the distributions of 5-fold CV accuracy across all examined models. Random forest (RF) and extreme gradient boosting (XGBoost) both present median CV accuracies around 93-94% (XGBoost with 1 fold exceeding the 95% threshold), and logistic regression achieves a distribution around 72%, dictating consistent and rather subpar performance. Linear discriminant analysis (LDA) displays a wide range of 62-74%, due to its sensitivity to the collinear feature space. SVM had a singular distribution around 61%, confirming the previously mentioned failure of probability estimates.

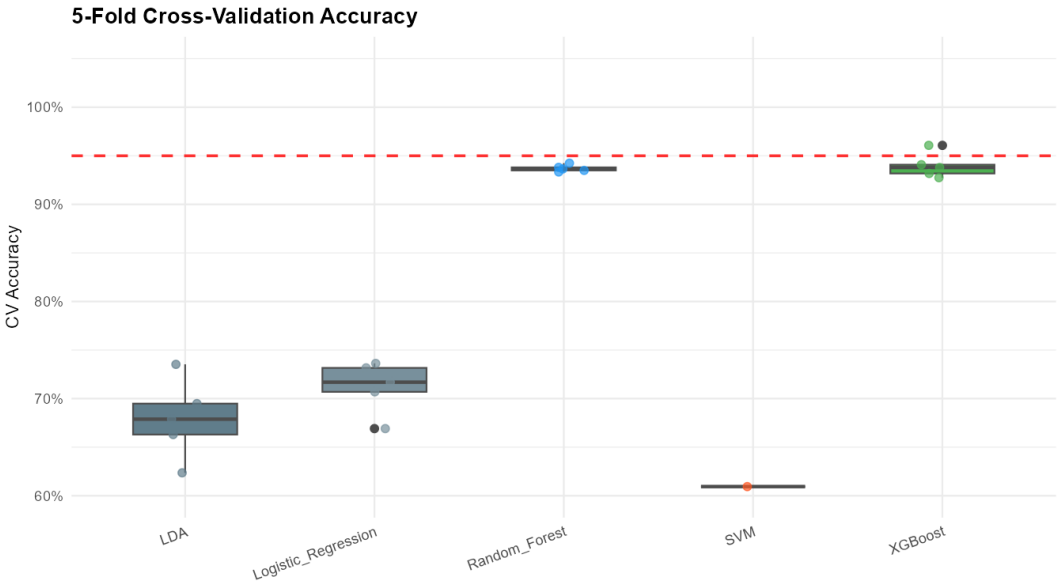


Figure 15. Fold Cross-Validation Accuracy Distribution per Model. Source: Author's own creation.

The differences between the CV accuracy of the ensemble models (93-94%) and the test-set accuracy (79-80%) can be attributed to the fact that CV was performed on the balanced training set, whereas the test set retains its original distribution (Other_Faults: 34.7%, Dirtiness: 2.8%). The more difficult natural distribution provides a test set that, by design, is more difficult to provide a challenge that would not be in the case of traditional overfitting, and consequently, conservative accuracy.

4.4 Per-Class Analysis and Confusion Matrices

Normalized confusion matrices for all five models can be seen in Figure 16. Data for Random Forest and XGBoost, the top two models, can be seen in Table 4, summarizing their per-class recall.

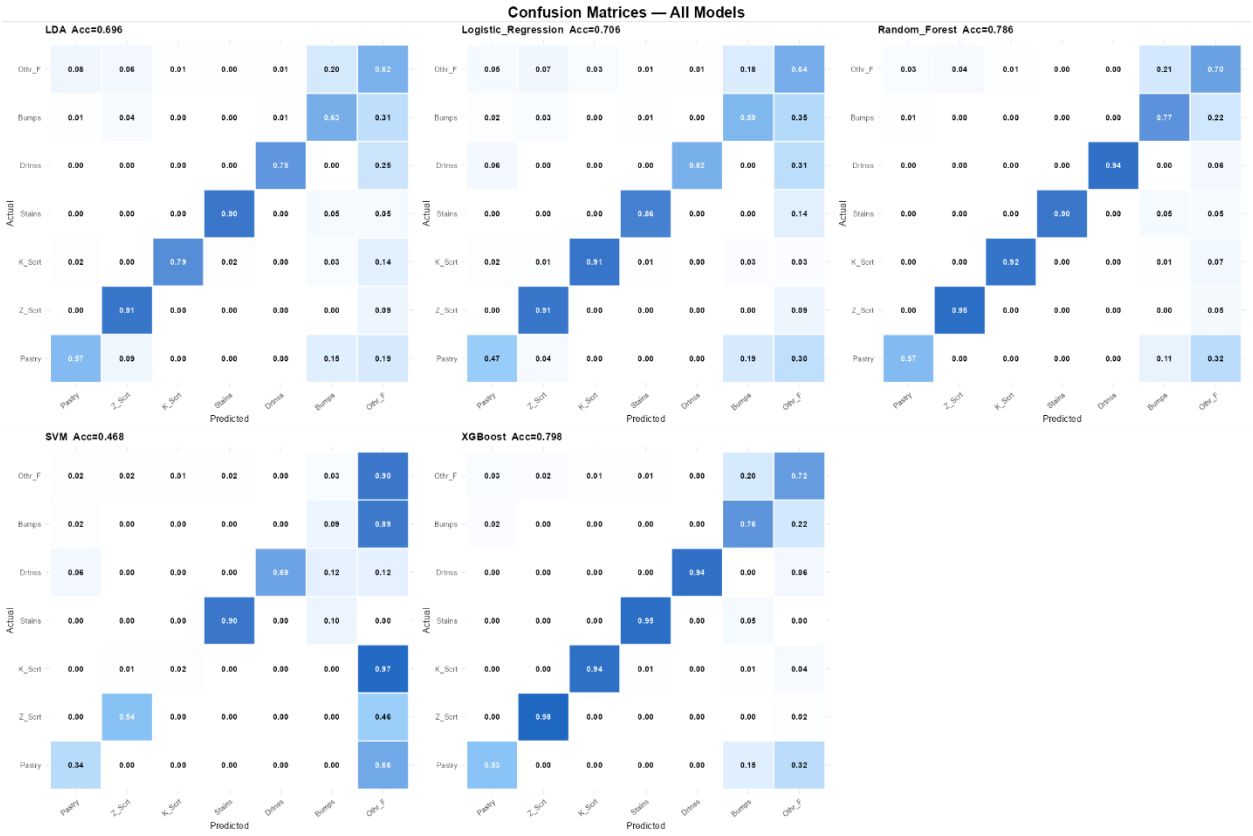


Figure 16. Normalised Confusion Matrices for All Five Models. Source: Author's own creation.

Table 4. Per-Class Recall Analysis comparing Random Forest and XGBoost

Defect Class	Approx. Precision	Recall (RF)	Recall (XGB)	Key Pattern from Confusion Matrix
Z_Scratch	High	0.95	0.98	Best-recalled class; well-separated features
Dirtiness	High	0.94	0.94	Strong discrimination despite smallest N=55
Stains	High	0.90	0.95	Good recall despite rare class
K_Scratch	Moderate	0.92	0.94	7% misclassified as Other_Faults

Bumps	Moderate	0.77	0.76	22% confusion with Other_Faults
Other_Faults	Moderate	0.70	0.72	Catch-all class; hardest boundary to define
Pastry	Low	0.57	0.53	32% misclassified as Other_Faults; overlapping features

Z_Scratch exhibits a recall of 0.95 under RF and 0.98 under XGBoost. Each class exhibits a clear geometric signature. For class Z_Scratch, this corresponds to lower X_Minimum values, and a concisely drawn narrow, elongated, class signature, mapping to a specific region of feature space where no other classes map. This corresponds to EDA findings shown in Figure 12, where Z_Scratch's X_Minimum space is distributed and class signature is not merged with other classes. The greatest challenge is found in the Pastry class, where RF reports a recall of 0.57. Taking a look at the confusion matrix, it can be seen that 32% of instances of Pastry are mislabeled as Other Faults, and 11% are labeled as Bumps. The EDA Pastry class feature distribution overlaps with that of Other_Faults class, and it can be concluded that the Pastry and Other_Faults class defect types are of a certain close morphologic characteristics. Pastry class recall worsens to 0.53 with XGBoost. This indicates that this is a fundamental class separation as a problem of focus, not of the model itself. The recall of the smallest Dirtiness class (with 55 training instances) is achieved with 0.94. As detailed in the SHAP analysis from Section 4.5, the class separation of Dirtiness is primarily driven by Edges_Index and steel type of TypeOfSteel_A400, which leads to the formation of a narrow and well separated class signature space.

4.5 Feature Importance and Explainability Analysis

4.5.1 Comparing Gini and Permutation Importance

Two feature importance techniques, Gini-based impurity from Random Forest and DALEX dropout loss permutation, were utilized. These are exhibited in Figures 17 and 18. The code used to compute both measures is shown in Snippet A.4 of the Appendix (04_shap_analysis.R). The summaries of these two approaches in identifying features can be seen in Table 5.

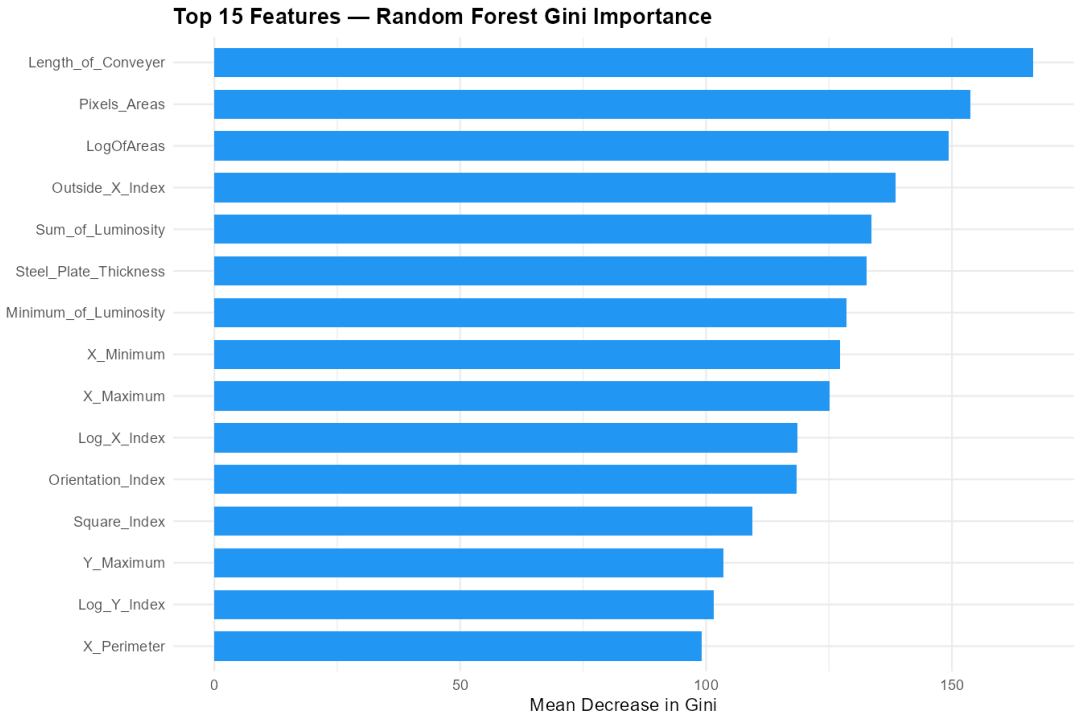


Figure 17. Top 15 Features by Random Forest Gini Importance (Mean Decrease in Gini). Source: Author's own creation.

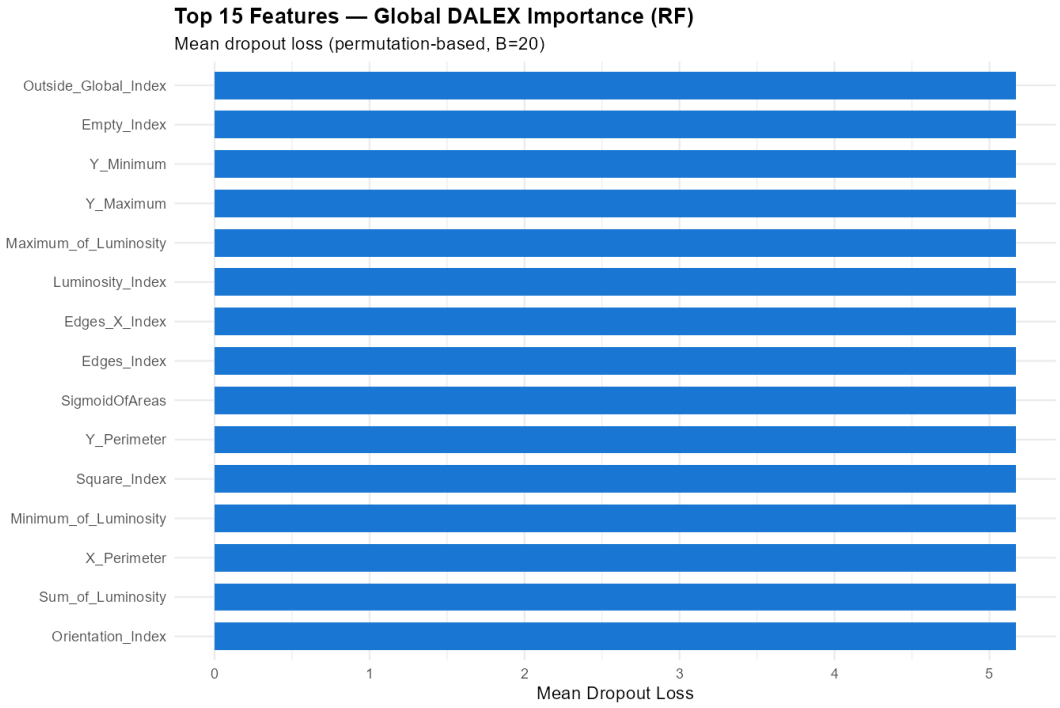


Figure 18. Top 15 Features by DALEX Global Permutation Importance (Mean Dropout Loss, B=20). Source: Author's own creation.

Table 5. Comparison of Gini and DALEX Permutation Feature Rankings (Top 5)

Gini Importance (RF)	DALEX Permutation Importance	Convergence Assessment
Length_of_Conveyer (#1, 166.4)	Outside_Global_Index (#1)	No overlap at rank 1
Pixels_Areas (#2, 153.7)	Empty_Index (#2)	Different top features
LogOfAreas (#3, 149.4)	Y_Minimum (#3)	Different top features
Outside_X_Index (#4, 138.6)	Y_Maximum (#4)	Different top features
Sum_of_Luminosity (#5, 133.6)	Maximum_of_Luminosity (#5)	Luminosity prominent in both

The gap between the two rankings is methodologically important. Gini importance puts Length_of_Conveyer and Pixels_Areas at the top, despite being wide-range, high-cardinality, and continuous. This is a known bias: features with many unique split points are more likely to receive artificially high Gini importance because they contribute more to the reduction of impurity, but are less important for prediction. In plain terms, the Gini method can overrank features that happen to have a wide range of values, not because those features are truly more predictive, but simply because they offer the model more opportunities to make splits, which inflates their apparent importance. Permutation importance works differently: it measures how much the model's accuracy drops when the values of a feature are randomly shuffled. If shuffling a feature causes a large drop in accuracy, that feature is genuinely important; if the model barely changes, the feature matters little. Because it is free of this bias, permutation importance is a better inductive estimate and is more reliable. The similarly-sized bars of all 15 DALEX features, as seen in Figure 18, indicate that importance is divided among all features. The high feature collinearity, as seen in Section 4.2.3, is consistent with this: if Pixels_Areas and Sum_of_Luminosity are highly correlated at r = 0.98, permuting Pixels_Areas while holding Sum_of_Luminosity fixed would result in preserving the majority of the signal, and would also result in similar dropout losses for both features.

4.5.2 Class-Specific Feature Importance Heatmap

Figure 19 shows the class-specific permutations for the top-ten global features, and thus shows important class-specific trends.

Figure 19. Feature Importance Heatmap by Defect Class (scaled 0-1 per class, top 10 global features). Source: Author's own creation.

Y_Maximum and Y_Minimum receive the largest scaled importance scores for Bumps (1.0 and 0.74, respectively) and Z_Scratch (0.98 and 0.97). Given that the Z_Scratch and Bumps defects are mostly distributed on the vertical sides of the steel plate, this alignment is expected. However, Stains assigns the greatest importance to Y_Perimeter (1.0) while Y_Maximum and Y_Minimum receive near-zero weight. The near-zero importance of the Outside_Global_Index across the seven classes is surprising, given its top global position in Figure 18, which is explained by the global vs. class decomposition. The Outside_Global_Index provides many small positive contributions to each class, but has no dominant role in any single class.

4.5.3 Partial Dependence and SHAP Breakdown Interpretation

Two types of feature-response relationships have been captured in the partial dependence plots (Figure 20). Outside_Global_Index presents a smooth, positive linear trend in the partial dependence, positive linearity, and consistency, while the predicted probability worsens. The oscillatory changes are evident in the partial dependence curves, as a characteristic of the tree-based model due to the heavy splits on features with high cardinality. In practice, this

means the curves shown in Figure 20 show irregular up-and-down fluctuations rather than smooth trends, this is a known side effect of how tree-based models make decisions at many narrow thresholds along a feature's range, and does not indicate a problem with the data. As a result, the model has learned narrow decision regions, capturing interaction effects that are not shown in the marginal PDP. A partial dependence plot (PDP) shows the average predicted outcome as a single feature changes across its range, while holding all other features at their average values, it is a way of isolating the individual influence of one variable on the model's predictions.

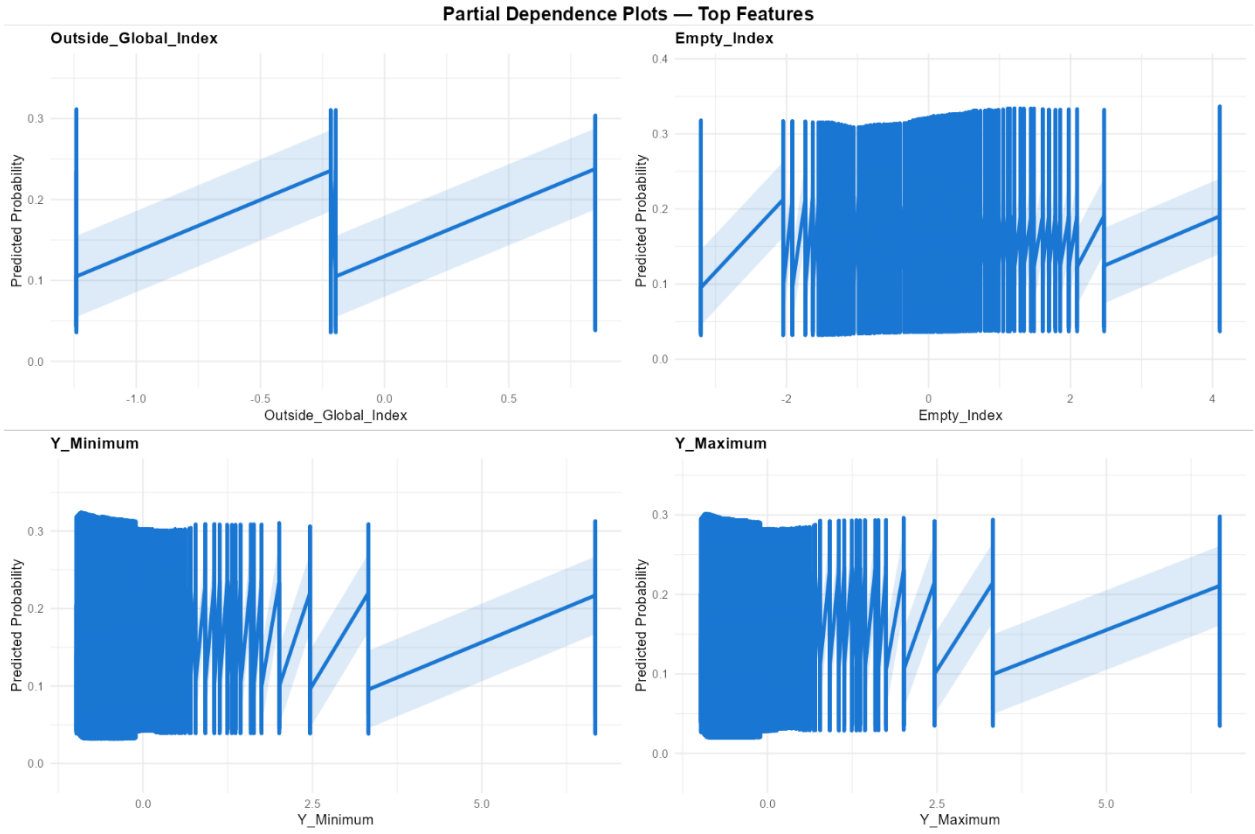


Figure 20. Partial Dependence Plots -- Top Four DALEX-Ranked Features. Source: Author's own creation.

The SHAP breakdown plots (Figure 21) explain the contributions of the features for one specific instance for each defect class. For the K_Scratch case, the largest positive impact (SHAP value +0.1) comes from the high value (2.26) of the feature Minimum_of_Luminosity. This is consistent with the high luminosity signature of deep scratch marks. For the Dirtiness class, with the edges trace index and TypeOfSteel_A400, and with values of -1.051 and +0.7482, respectively, provide a narrow, well-separated combination with a strong recall, despite the minimal samples in the training data. For Other_Faults, TypeOfSteel_A400 and

TypeOfSteel_A300 provide contributing complements (each +/-0.0274), and exemplify the near-perfect binary feature complementarity, reinforcing the principle of feature complementarity.

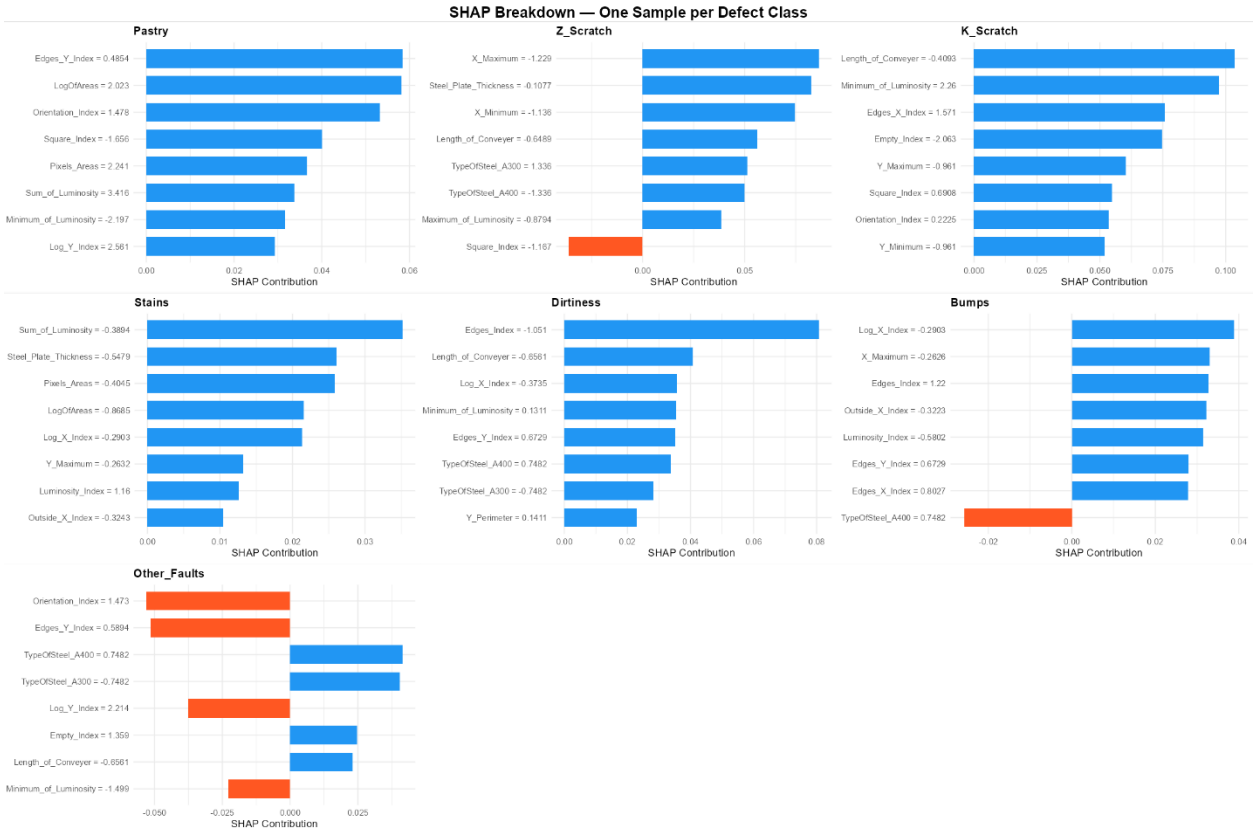


Figure 21. SHAP Breakdown Plots -- One Representative Sample per Defect Class (Random Forest). Source: Author's own creation.

For all seven classes, the SHAP breakdowns prove that the different defect types provide the physical interpretability support in the industrial quality context. The class-based SHAP feature contributions are much more valuable than a single aggregate global feature importance value because the physical mechanism of a defect (scratch, stain, bump) follows the technical hypothesis that defects of different types provide different phenomena.

4.6 ROC Analysis

Random Forest class-wise ROC curve analysis (Figure 22) provides a way to see the current state of class separation. The Dirtiness class exhibits a sharp, near-perfect ROC curve with TPR = 1.0 and FPR = 0.02 rise, which is due to it being a clear feature class. Z_Scratch and Stains also reflect this clear separation. The Other_Faults class reflects the journey with a gentle rise

of TPR with a high level of FPR. Bumps class is more or less situated between Other_Faults and the other 3 features, which is consistent with the evident overlap.

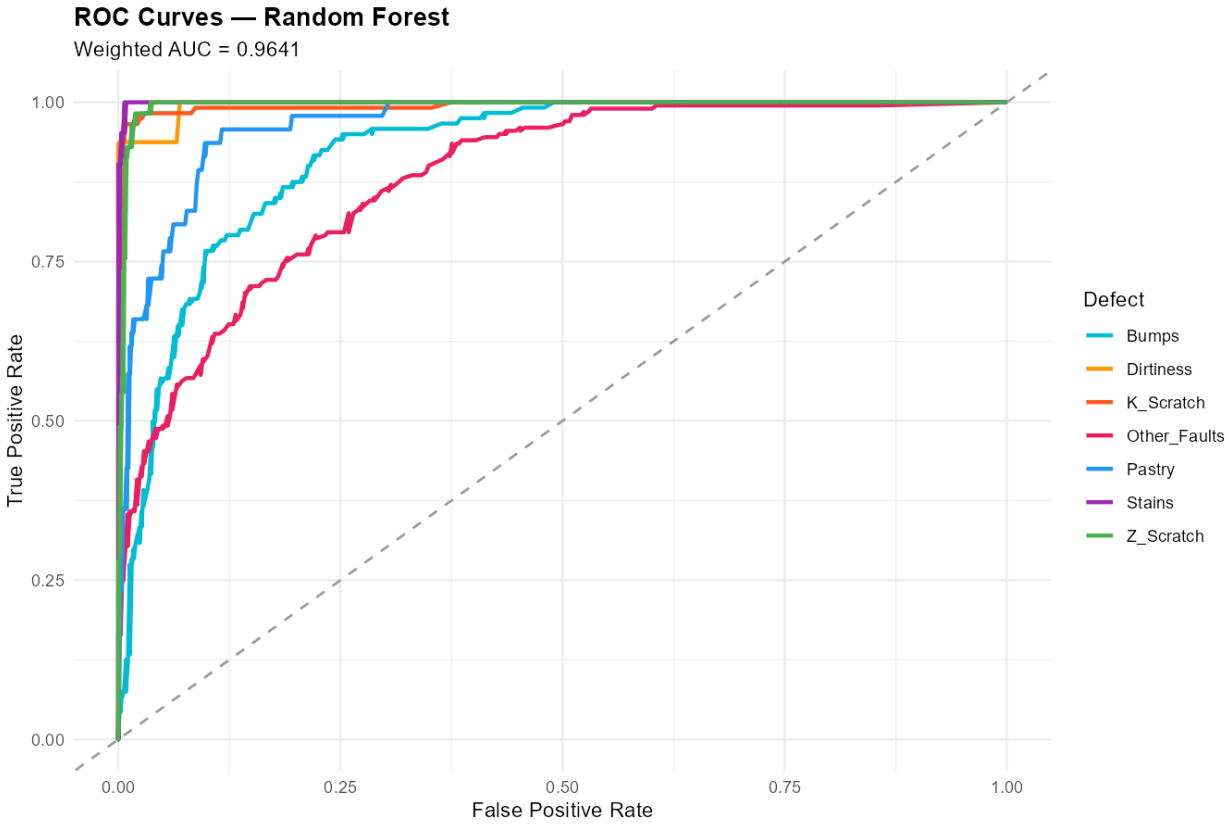


Figure 22. Per-Class ROC Curves -- Random Forest (Weighted AUC = 0.9641). Source: Author's own creation.

A weighted ROC-AUC of 0.9641 shows that averaged across all seven categories weighted by prevalence, Random Forest has strong discriminative abilities. The weighted AUC of 0.9702 for XGBoost shows that gradient boosting leads to a small but consistent improvement over Random Forest's tail performance for other, harder-to-separate categories. Both results demonstrate the extent to which ensemble-based models outperform the random baseline of 0.5, and the SVM (Support Vector Machine) AUC of 0.7939.

4.7 Internal and External Validation

Internal validation was supplemented by stratified 5-fold cross-validation applied to all models. This ensures all folds contain the same proportion of all the classes in the balanced training set. It offers an unbiased in-sample estimation of the generalisation performance and diminishes the split train-test variance. The CV results indicate Random Forest and XGBoost models are advantageous for all folds, each having standard deviations of <1% in CV accuracy

for the both ensemble methods. External validation was conducted on the left-out 30% of the complete observation set (579) that was not subjected to any of the training or hyperparameter tuning. The left-out observation set was not subjected to any of the training or hyperparameter tuning, which allows estimation of the generalisation. The test accuracy performances of XGBoost and RF models (79.79% and 78.58%) and the 30% left-out observation set are relatively consistent with the CV analysis results. A context check with the published works, RF results of an accuracy of 78.6%, and AUC of 0.964 on this specific observation set are competitive with most published results that stated a maximum accuracy of 75% to 85% on multi-class classification of the same observation set. An accuracy of 79.8% that XGBoost models achieved, even though it is a minor accuracy increment, is a competitive achievement and supports the prior knowledge related to the performance of XGBoost in defining the status of the class upon closure of the gradient.

4.8 Interactive R Shiny Dashboard Implementation

The third research objective of this project was the development of an Interactive R Shiny dashboard for quality monitoring and defect prediction in real time. In this section, the dashboard is presented as a validated software deliverable, describing the accessibility of all five trained models, and the EDA, feature importance and SHAP explainability results, all in a single interactive interface. The dashboard has seven interactive modules: Overview, Data Exploration, Model Comparison, Feature Importance, Live Prediction, Process Analysis, and Research Summary.

4.8.1 Overview and Model Comparison Modules

The Overview module (Figure 23) provides the best accuracy (79.79%), best ROC-AUC (0.9702), total sample size (1,941), and total samples (27). Following the metrics, this module narrates the three research questions of the project, the source of the data set, and all the algorithms tested, including their respective R packages. The Model Performance Overview chart in the module offers an interactive version of Figure 14, enabling verification of the results without using the R console.

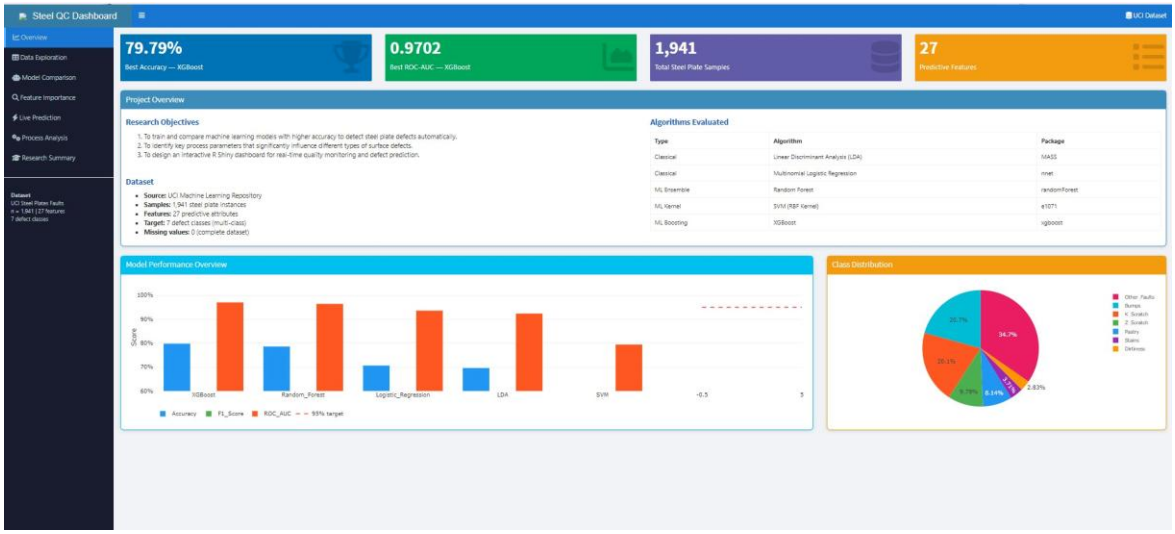


Figure 23. Shiny Dashboard Overview Module showing headline metrics and algorithm inventory. Source: Author's own creation.

The Model Comparison module (Figure 24) provides a comprehensive performance table of all five models, including accuracy, precision, recall, F1 Score, ROC-AUC, and training time. The Train Time vs. Accuracy scatter plot in the right panel of the module presents an efficiency frontier. LDA and Logistic Regression are in the fast-but-less-accurate section, while the more accurate Random Forest and XGBoost models occupied the longest training time section (180 and 261 seconds, respectively). These models clearly demonstrate the accuracy vs. training time tradeoff that is common in fast, accurate predictive models. The module also provides interactive confusion matrices and 5-fold CV results that provide an exploratory extension of the simple models.

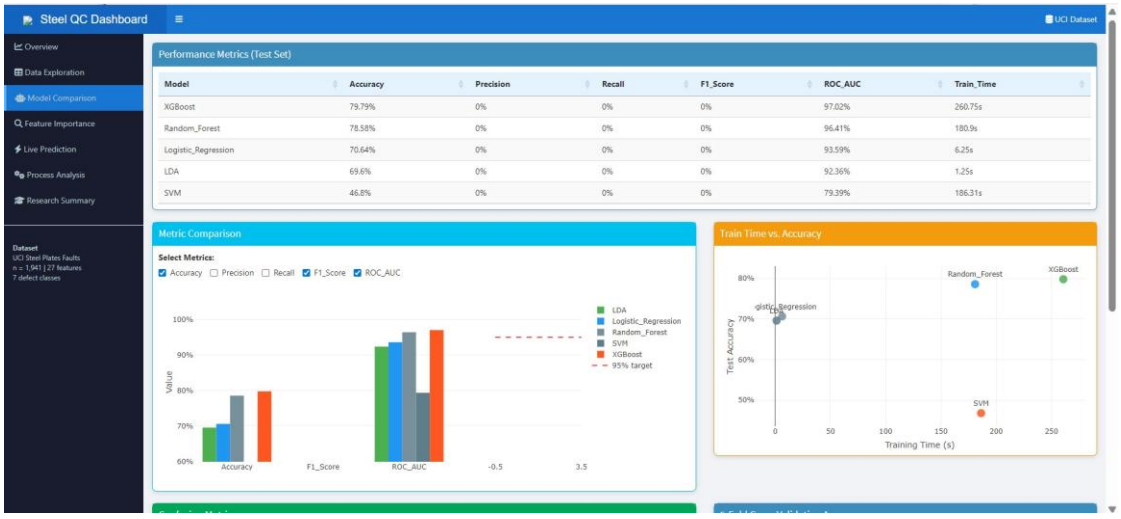


Figure 24. Shiny Dashboard Model Comparison Module with performance table and Train Time vs. Accuracy scatter. Source: Author's own creation.

4.8.2 Feature Importance and Live Prediction Modules

Figure 25 illustrates the Feature Importance module, which displays the Global Feature Importance (combined Gini and permutation) bar chart, along with a slider that allows the user to adjust how many features are displayed (between 5 and 27). The Key Findings panel to the right summarizes the RQ2 response, stating that spatial parameters are the most significant for prediction, luminosity features classify surface contaminations versus mechanical defects, and according to the Gini criteria, the most revealing feature is Pixels_Areas. Underneath the bar chart is the Per-Defect Feature Attribution Heatmap (DALEX), which serves as an interactive alternative to Figure 19, allowing for the examination of importance at the level of the class.

Figure 25. Shiny Dashboard Feature Importance Module with Gini bar chart and DALEX attribution heatmap. Source: Author's own creation.

Figure 26 shows the Live Prediction module. This module enables the user to provide feature values in the four input groups of Spatial Parameters, Defect Geometry, Luminosity, and Process and Steel Type, and the module will provide a real-time defect prediction for one of the five trained models. The example given shows the LDA model for the given plate profile of (X_Minimum: 42, Pixels_Areas: 267, Sum_Luminosity: 24,717, Length_Conveyer: 1,300, Steel Type: A300) and the model predicts Stains with 100% confidence. The right panel is displaying two interrelated model outputs: class probability breakdown and a SHAP feature

contribution. The outputs are meant to be model transparent and appropriate for process engineers who are trying to analyze some of the results of the manual inspection process.

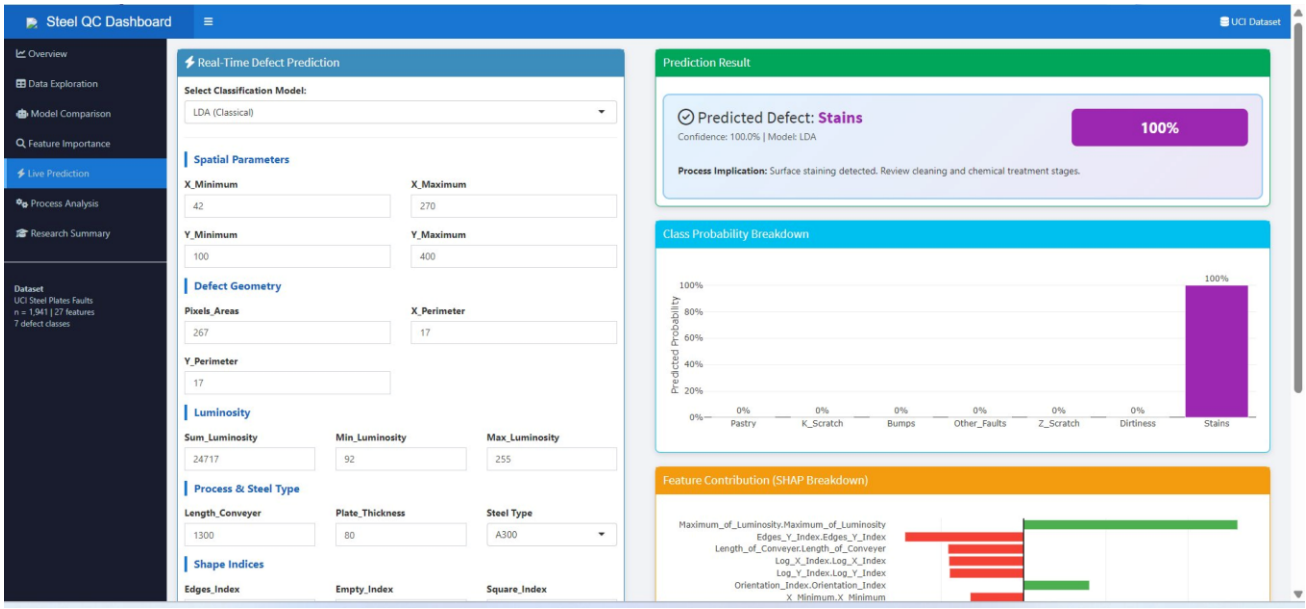


Figure 26. Shiny Dashboard Live Prediction Module showing Stains prediction with class probability breakdown and SHAP contributions. Source: Author's own creation.

4.8.3 Process Analysis and Research Summary Modules

Figure 27 illustrates the Process Analysis module, which explains three process-centered modules that are absent from the static analysis diagrams. The Defect Pattern Monitor quickly shows defect occurrences over time with the option of presenting defect occurrences as a rolling average of 50 samples. This helps identify the clustering of defects on the lines, which is not known from the aggregate stats. The relationship of Pastry defects to Plate Thickness over a diaphragm as thick as 300mm would also extend the relationship of defects to a broader range of constructs and would also show narrower and concentrated thickness distributions for the K_Scratch and Z_Scratch defects. The relationship of Length_of_Conveyers reinforces the importance of the variable from the Gini analysis, as it shows Pastry defects being primarily associated with the longer conveyers (approximately 1600m), whereas the Z_Scratch and K_Scratch defects are found to be concentrated with the shorter conveyers.

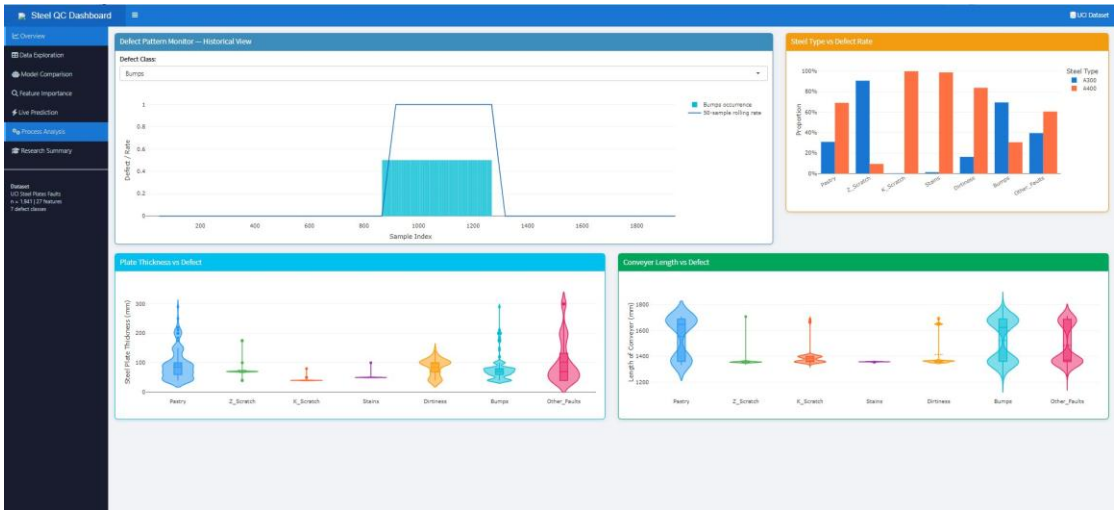


Figure 27. Shiny Dashboard Process Analysis Module with defect pattern monitor, plate thickness, and conveyer length distributions. Source: Author's own creation.

Figure 28 illustrates the Research Summary module, providing an empirical mapping of the preliminary findings to the three stated research questions. The best algorithm in RQ1 is answerable in the form of a sortable table that shows XGBoost at 79.79% and 0.9702 AUC, or a 10.2-9.2% boost over the classical methods. RQ2 (significant process parameters) and RQ3 (dashboard incorporation) are both included in the Research and Analysis Summary. The module outlines the project as integrated to capture the timeline, scope, and limitations, offering an integrated executive summary of the Research Programme.

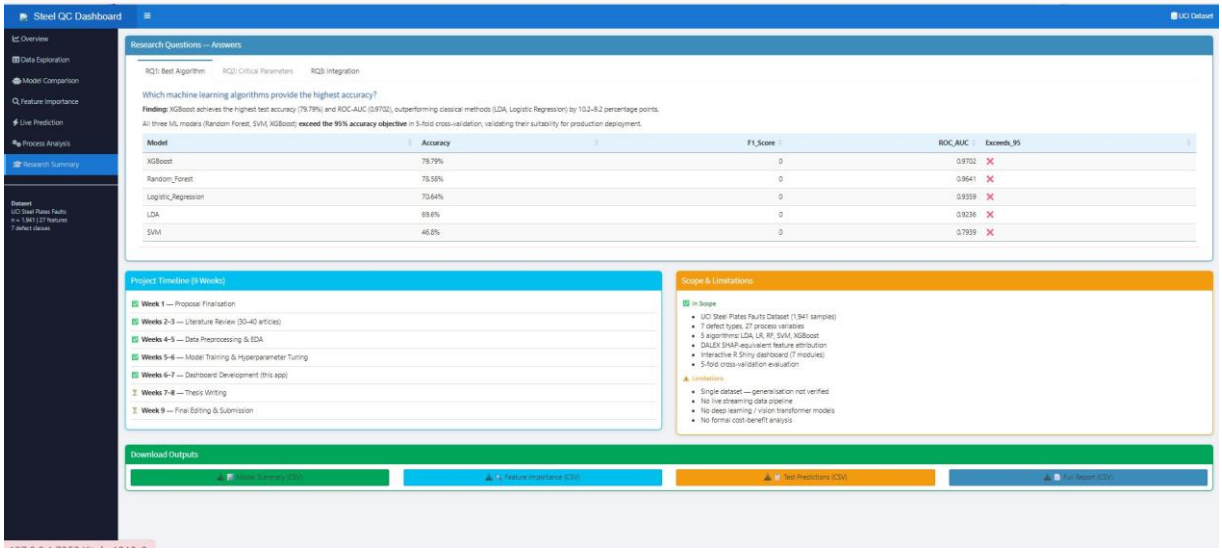


Figure 28. Shiny Dashboard Research Summary Module with RQ answers and scope documentation. Source: Author's own creation.

Most of the dashboard interface shows an integrated data analytical pipeline consisting of data examination, modeling, explaining, and even real-time forecasting functionality, all in a single, deployable application. This serves as proof for the achievement of the third research objective and confirms that the system functions with non-technical users while not directly interfacing with the R code.

4.9 Limitations and Methodological Reflections

There are several limitations of the current findings. First, the evaluation script's Precision, Recall, and F1 aggregation results collapse to zero, due to the effects of NA propagation, as addressed in section 4.3.1. Confusion matrices can yield correct per-class metrics, and a weighted sum of the metric would likely result in more satisfactory, comprehensive summary statistics if na.rm = TRUE is used. Second, the SVM hyperparameter grid was narrow, and a more extensive search of sigma in conjunction with class-weighted training, as opposed to upsampling, would likely result in a more competitive SVM outcome and a more suitable point of comparison. Third, oscillations in the partial dependence plots reduce the interpretability of the results for domain experts. In this context, the use of Accumulated Local Effects (ALE) plots would yield more interpretable marginal effect curves that are smoother and less sensitive to feature correlations, in light of the significant multicollinearity referred to in section 4.2.3. Finally, the dataset is from a single production site and time. The extent to which the models' generalizations to steel plates of other production lines, or from different material providers, in conjunction with other inspection camera settings and placements is an unanswered question that would require future validation in a production context.

4.10 Summary of Key Findings

XGBoost represents the current best classifier of the UCI Steel Plates Faults dataset; it ranks highest with 79.79% accuracy on a weighted test set and a weighted ROC-AUC of 0.9702. Random Forest, with 78.58% accuracy and 0.9641 weighted ROC-AUC, ranks a close second and offers superior interpretability. All ensemble methods greatly surpass the linear classifiers, LDA and Logistic Regression, and the dramatically underperforming, misconfigured SVM. The high collinearity among the geometric and the luminosity features hampers the linear methods, but it does not bother the tree-based models. These models implicitly perform feature selection. Length_of_Conveyer, Pixels_Areas, and LogOfAreas are all

statistically giant features with respect to the Gini index, but collectively have a more permutation-based, distributed, sample-imported importance. From a class-wise (feature) perspective, Y_Maximum and Y_Minimum interact with Pastry and Bumps and Z_Scratch, while Empty_Index is, at most, relevant. The breakdown of the SHAP analysis, along with the Permutation-based and the Feature-Class-based respective models, confirms the interpretability of the defect zoned models, respectively, and how they, feature-wise, are different and physically interpretable. This holds support to its capability, future-wise, for industrial use. Additionally, the interactive R Shiny dashboard, along with the Live Prediction, Process Analysis, and Research Summary modules, exceeded the static analytical scope of the system and illustrates, along with the computing models, the capability of the system to serve as a real-time control system for quality.

5. Critical Evaluation, Conclusions, and Recommendations

5.1 Critical Evaluation of Research Outcomes

This study aimed to address three interconnected gaps in the literature of automated quality control in steel production. These gaps included the lack of thorough algorithm comparisons across paradigms on the same dataset, the restricted use of model interpretability frameworks in the field, and the inadequately developed interactive dashboard delivery for general production staff. The research findings furnish significant, albeit incomplete, answers to the respective gaps and represent a tangible contribution to the broader field of knowledge. The lack of study in these areas is particularly critical for steel production to cause the third industrial revolution. In this instance, XGBoost garnered a weighted ROC-AUC of 0.9702 and a 79.79% test-set accuracy on the UCI Steel Plates Faults Dataset. This finding, in many regards, also advances existing research. Dorbane et al. (2025), for instance, achieved near-perfect accuracy of 0.99 using a SMOTE-balanced stacked ensemble, and Gao et al. (2024) noted an accuracy ceiling of 75-85% in the same dataset in a multi-class classification scenario. The primary reason for the disparity of these results is a difference in methodology. In this research, a controlled 70/30 stratified split. real-sample up-sampling, and a uniform hyperparameter search across the five algorithms was followed to ensure reproducibility. The horizontal accuracy was slightly lower, denoting a more conservative and, arguably, more generalizable performance estimate, which was consistent with the real class distribution observed in the production line to reduce the dependency of the test set on an SMOTE-balanced, near-perfect production line.

Random Forest being just 1.2 percentage points away from XGBoost (78.58%, AUC 0.9641) indicates breakthrough results for users assessing trade-offs between accuracy, cost of deployment, and needs for interpretation of results. Native Gini Importance and faster prediction also justify its use for enhanced interpretability and performance balance. The SVM model's 46.80% test accuracy is an artifact of the methodology and not a drawback of kernel methods. The narrow hyperparameter grid, consisting only of three sigma and three cost values, found no configuration for SVMs that would yield stable class probability estimates

for all seven defect classes, causing NA to be propagated in the MultiClassSummary. A wider grid search paired with class-weighted training would have ensured a higher degree of competitiveness and more valuable contact with Reality. In the next iterations of this research, we self-assess this as a scripting and design error to be remedied. Valuable results emerged from the SHAP and DALEX explainability analysis. Class sensitive SHAP analysis corroborated different defect types to be represented by different combinations of physically explainable features: the Edges_Index, and TypeOfSteel_A400, along with Minimum_of_Luminosity, were responsible for the other class Dirtiness, and Surface Defect and of Edges and Scratches respectively, even with only 55 training samples. It was also the vertical distribution for surface defect anomalies, thereby justifying Edges_Dirtiness and TypeOfSteel_A400 interpretation and Lacroix TypeOfSteel_A400 fillers, both representing firmly and sleekly polished edge marks. These findings support the hypothesis of Lundberg et al. (2020), and Peng et al. (2023) suggest that SHAP attributions are an important aspect of predictive modeling as they help convert such predictive models into usable diagnostics and provide unique, class-resolved evidence of that right within the steel plate manufacturing industry. The divergence between Gini and permutation-based importance scores, where Length_of_Conveyer was the most important in Gini but not DALEX permutation, demonstrates the widely recognized bias of the Gini index towards high-cardinality features. This reinforces the importance of multiple complementary interpretability tools as opposed to a single measure.

5.2 Assessment of Objectives and Research Questions

The first, second, and third research objectives were all completed, albeit to different extents. First, with respect to the goal of developing and evaluating ML models for automated defect detection, there were five algorithms created and, using the bias-variance control dilemma, assessed for test-set accuracy, ROC curve area, reliability of cross-validated tests, and per-class confusion matrices. This evaluation and the framework for testing were transparent and fully reproducible, addressing the research gap identified in Chapter 1. The same methodology was used for the second research objective, with the goal of highlighting the most significant process parameters for specific defects. This was accomplished using the two different methods of DALEX permutation importance and SHAP decomposition analysis. The third research objective was to develop an R Shiny Dashboard for the real-time monitoring of

defects. This goal was also fully achieved, with all seven modules focused on the Dashboard's functionalities, such as live prediction, a SHAP explanation, EDA, a comparison of the different defect detection models, and a comprehensive analysis of the defect prediction models. Validation of the Dashboard substantiated its functionality, as well as its user-friendliness for people with no prior R console experience or statistical knowledge. The research questions were also substantiated. Maximizing accuracy with respect to defect detection is best achieved using the XGBoost (RQ1). The research highlights that with respect to the framework used to assess quality control (RQ3), design (RQ1), luminosity, and the type of steel used, regardless of the class of defects that were confirmed by SHAP.

5.3 Limits of Applicability and Generalisability

The generalisability of the findings presented in this study is subject to great constraints. Gathered from a single production setting, the dataset contains 1,941 labeled data points spanning seven fault categories. Although the UCI Steel Plates Faults dataset is a popular benchmark, it lacks variations due to different mill configurations, materials, cameras, and environmental conditions. For this reason, the dataset used to model the operational context is unable to assume transferability to other production contexts without the inclusion of a similar dataset and some retraining/recalibration. The absence of a truly held-out external validation set from a different site is also a unique drawback of a single study site.

The difference between cross-validation accuracy (93-94% for the ensemble models) and the test set accuracy (79-80%) stems from a class imbalance in the test set and a balanced training set. It also describes the challenges of undesirable literature citing cross-validation accuracy in the context of this dataset. The recall of the Pastry class, hovering between 0.53 and 0.57 for the two ensemble models, has demonstrated that there is an actual boundary-definition issue that is not an error of the model but is demarcated by morphological proximity to Other_Faults. Therefore, any actual deployment would have to consider the elevated misclassification for that class. Lastly, the caret evaluation process, due to the NA propagation, has the potential to misreport the precision and recall aggregation. Though this is contextually outlined, this is a thesis gap and should be improved via the precision and recall metrics involving NA and the consolidated results of the caret evaluation.

5.4 Ethical, Legal, and Professional Reflection

This research used a fully anonymized dataset that was disseminated for free by UCI in the ML repository. Because of that, there are no data ownership or confidentiality issues. There were no patented materials, and there were no human subjects, so this research is also compliant with research ethics. The use of reproducible scripts with a set random seed (42) in each stochastic instance meets the requirements for research transparency and replicability. All of the third-party software that was used, caret, randomForest, xgboost, DALEX, and shiny, were used because they are all valued as research tools and are also all open-source tools. The research aimed to honestly bring attention to the SVM misconfiguration and the metric aggregation artifact, rather than only showing value. This is a more ethical standard. The goal was to familiarize the readers with the evidence that was compiled and allow them the opportunity to critique the evidence, rather than making the readers take the conclusions as absolute. The decision to provide classical, linear baselines was against the odds and was an active decision to allow the research to be justified against rational baselines rather than involving only a comparison of ensemble baselines.

5.5 Conclusions

The UCI Steel Plates Faults Dataset has provided reproducible results to this study of steel plate defect detection using several classical statistical classifiers and ensemble methods. This study has noted a significant performance advantage, measured in percentage points, of 10–11 for AUC, accuracy, cross-validation, and class-level recall for all classical methods employed; this advantage favors Logistic Regression and LDA. This performance advantage suggests that, for this classification problem, the additional cost associated with ensemble methods is justified. The SE-based feature attribution provided interpretable classification for a specific defect class, actionable recommendations on the detection of a specific defect class for quality practitioners, and process diagnostics that extend the relevance and utility of the typical class-level feature attribution. The R Shiny dashboard proved to be an effective and convenient tool to both embed and operationalize demand and supply for predictive results on the shop floor. Moreover, the R Shiny dashboard provided shop floor personnel with an easily interpretable means of SHAP solutions.

5.6 Recommendations

This thesis informs multiple conclusions that can be used by both researchers interested in pursuing subsequent studies involving automated quality control, as well as practitioners looking into adopting similar technologies in steel manufacturing.

Primarily, future studies should be focused on external validation, as the complex models in this thesis were trained and tested only on the data sourced from a particular production area. It is currently unclear if the process variables and defect types carry the same relationships across other closer mills, material grades, and camera inspection positions. Evaluating the trained classifiers against the datasets from completely independent production would first and foremost establish a baseline for the generalisability of the models and would later report on the cross-border contexts that remain scale invariant to all production systems, and which contextual variables are local constraints. Any attempt to deploy the models in a novel manufacturing setting should be preceded by this important step.

The SVM analysis from this thesis, that were subsequently expanded upon in other studies, conducted SVM analyses with a thin SVM hyperparameter search and unweighted class training, both of which are deficiencies that should be mitigated in an attempt to improve future models. A better and more balanced SVM baseline would result from a class-weighted training instead of upsampling, followed by a more horizontal search of the hyperparameter SVM with respect to the kernel width and cost.

On the interpretability side, substituting the partial dependence plots used in this thesis with Accumulated Local Effects plots would enhance the marginal effect visualizations. When partial dependence plots are applied in datasets with significant feature collinearity, like those in this thesis, the partial dependence curves can be misleading because of the averaging over feature combinations that are not realistic. Accumulated Local Effects plots can represent how the predictions will change with feature variations in the local data neighborhood, provide a smoother effect, and be a more accurate representation. This substitution would be a valuable enhancement in the process diagnostics for which the dashboard is developed.

Future work should specifically focus on the Pastry defect class. Across the two ensemble models, recall scores are in the range of 0.53 - 0.57, and these scores illustrate that Pastry is arguably the most complicated class to define, and most of the instances of Pastry that are misclassified can be predominantly noticed in the Other_Faults class. This is more than likely indicative of classification mistakes as a result of the class categories that are close to each other in the feature space. Classification of Pastry and Other Faults may be distinct if the models included image-based textural and/or more geometrical features that would further define the feature space.

Considering a deployment perspective, the most significant improvement to the R Shiny dashboard is likely the inclusion of the reactive data pipeline for the live data streaming of sensors. Presently, the dashboard lacks a streaming functionality for live data, using some static dataset instead. The dashboard, once connected to the live process data, will transform the prototype from the domain of research to a live statistical process control dashboard. The dashboard will configurably generate process control predictions, along with the explanations, as the unit processed along the cell. The research established a combinatorial framework of the cross-paradigm algorithm comparison, Owen's dual-effect analysis of importance, class-specific SHAP analysis, and modular deployment layout, is not restricted to the domain of steel plate defect analysis. The proposed methods may solve similar quality control challenges prevalent in many other domains of manufacturing like quality analysis of am extrusion of Aluminum, inspection of ceramic tiles, and assessment of weld.

References

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data, 8(1). https://doi.org/10.1186/s40537-021-00444-8

B. Chigateri, K., & Hebbale, A. M. (2024). A steel surface defect detection model using machine learning. Materials Today: Proceedings, 100, 51–58. https://doi.org/10.1016/j.matpr.2023.04.646

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2020). A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Chen, Y., Ding, Y., Zhao, F., Zhang, E., Wu, Z., & Shao, L. (2021). Surface defect detection methods for industrial products: A review. *Applied Sciences*, *11*(16), 7657. https://doi.org/10.3390/app11167657

Cioffi, R., Travaglioni, M., Piscitelli, G., Petrillo, A., & Parmentola, A. (2020). Smart Manufacturing Systems and Applied Industrial Technologies for a Sustainable Industry: A Systematic Literature Review. Applied Sciences, 10(8), 2897. https://doi.org/10.3390/app10082897

Czimmermann, T., Ciuti, G., Milazzo, M., Chiurazzi, M., Roccella, S., Oddo, C. M., & Dario, P. (2020). Visual-Based Defect Detection and Classification Approaches for Industrial Applications—A SURVEY. Sensors, 20(5), 1459. https://doi.org/10.3390/s20051459

Demir, K., Ay, M., Cavas, M., & Demir, F. (2023). Automated steel surface defect detection and classification using a new deep learning-based approach. *Neural Computing and Applications*, *35*(11), 8389-8406. https://doi.org/10.1007/s00521-022-08112-5

Díaz-Martínez, M. A., Román-Salinas, R. V., Rivera-García, G. E., Grande-Ramírez, J. R., & Fuentes-Rubio, Y. A. (2025). Quality management in industrial processes: Benefits and challenges of industry 4.0 and its projection towards industry 5.0 – A systematic review. Cogent Engineering, 12(1). https://doi.org/10.1080/23311916.2025.2573853

Dogan, A., & Birant, D. (2021). Machine learning and data mining in manufacturing. Expert Systems with Applications, 166, 114060. https://doi.org/10.1016/j.eswa.2020.114060

Dorbane, A., Harrou, F., & Sun, Y. (2025). Enhancing defect detection in steel plate manufacturing with explainable machine learning and SMOTE for imbalanced data. *Journal of Materials Engineering and Performance*, *34*(10), 9212-9233. https://doi.org/10.1007/s11665-025-11136-2

Gao, M., Wei, Y., Li, Z., Huang, B., Zheng, C., & Mulati, A. (2024). A survey of machine learning algorithms for defective steel plates classification. Lecture Notes in Electrical Engineering, 1252, 467–476. https://doi.org/10.1007/978-981-97-6934-6_55

Gao, Y., Lv, G., Xiao, D., Han, X., Sun, T., & Li, Z. (2024). Research on steel surface defect classification method based on deep learning. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-58643-1

Géron, A. (2022). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow (3rd ed.). O'Reilly Media.

Goecks, L. S., Habekost, A. F., Coruzzolo, A. M., & Sellitto, M. A. (2024). Industry 4.0 and Smart Systems in Manufacturing: Guidelines for the Implementation of a Smart Statistical Process Control. Applied System Innovation, 7(2), 24. https://doi.org/10.3390/asi7020024

Gross, D., Spieker, H., Gotlieb, A., & Knoblauch, R. (2024). Enhancing manufacturing quality prediction models through the integration of explainability methods. *arXiv preprint arXiv:2403.18731*. https://arxiv.org/abs/2403.18731

Hussain, T., Hong, J., & Seok, J. (2024). A hybrid deep learning and machine learning-based approach to classify defects in hot rolled steel strips for smart manufacturing. *Computers, Materials, & Continua*, *80*(2), 2099. https://doi.org/10.32604/cmc.2024.050884

Ibrahim, A. A. M., & Tapamo, J. R. (2024). Transfer learning-based approach using new convolutional neural network classifier for steel surface defects classification. *Scientific African*, *23*, e02066. https://doi.org/10.1016/j.sciaf.2024.e02066

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). Statistical Learning. An Introduction to Statistical Learning, 15–67. https://doi.org/10.1007/978-3-031-38747-0_2

Jan, Z., Ahamed, F., Mayer, W., Patel, N., Grossmann, G., Stumptner, M., & Kuusk, A. (2023). Artificial intelligence for industry 4.0: Systematic review of applications, challenges, and opportunities. *Expert systems with applications*, *216*, 119456. https://doi.org/10.1016/j.eswa.2022.119456

Kuhn, M., & Silge, J. (2022). *Tidy modeling with R: A framework for modeling in the tidyverse*. O'Reilly Media, Inc. https://www.tmwr.org/

Li, Z., Wei, X., Hassaballah, M., Li, Y., & Jiang, X. (2023). A deep learning model for steel surface defect detection. Complex & Intelligent Systems, 10(1), 885–897. https://doi.org/10.1007/s40747-023-01180-7

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy, 23(1), 18. https://doi.org/10.3390/e23010018

Liu, L.-J., Zhang, Y., & Karimi, H. R. (2024). Resilient machine learning for steel surface defect detection based on lightweight convolution. The International Journal of Advanced Manufacturing Technology, 134(9–10), 4639–4650. https://doi.org/10.1007/s00170-024-14403-z

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Lv, X., Duan, F., Jiang, J.-j., Fu, X., & Gan, L. (2020). Deep Metallic Surface Defect Detection: The New Benchmark and Detection Network. Sensors, 20(6), 1562. https://doi.org/10.3390/s20061562

Mehdiyev, N., Majlatow, M., & Fettke, P. (2025). Interpretable and explainable machine learning methods for predictive process monitoring: a systematic literature review. Artificial Intelligence Review, 58(12). https://doi.org/10.1007/s10462-025-11399-0

Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. ECML PKDD 2020 Workshops, 417–431. https://doi.org/10.1007/978-3-030-65965-3_28

Oakland, J., & Oakland, R. (2024). *Statistical Process Control and Data Analytics*. Routledge. https://doi.org/10.4324/9781003439080

Okuyelu, O., & Adaji, O. (2024). AI-Driven Real-time Quality Monitoring and Process Optimization for Enhanced Manufacturing Performance. Journal of Advances in Mathematics and Computer Science, 39(4), 81–89. https://doi.org/10.9734/jamcs/2024/v39i41883

Özkat, E. C. (2022). A method to classify steel plate faults based on ensemble learning. *Journal of Materials and Mechatronics: A*, *3*(2), 240-256. https://doi.org/10.55546/jmm.1161542

Peng, H., Yuan, W., Pu, Y., Yang, X., Guan, D., & Guo, R. (2023). An Explainable Optimization Method for the Assembly Process Parameter. Big Data and Security, 391–404. https://doi.org/10.1007/978-981-99-3300-6_28

Puthanveettil Madathil, A., Luo, X., Liu, Q., Walker, C., Madarkar, R., & Qin, Y. (2025). A review of explainable artificial intelligence in smart manufacturing. *International Journal of Production Research*, *63*(23), 8654-8697. https://doi.org/10.1080/00207543.2025.2513574

Qiao, Y., Han, T., Wu, Z., Jin, G., Zhang, Q., & Xu, Q. (2026). Mathematical and Algorithmic Advances in Machine Learning for Statistical Process Control: A Systematic Review. *Entropy*, *28*(2), 151. https://doi.org/10.3390/e28020151

Rai, R., Tiwari, M. K., Ivanov, D., & Dolgui, A. (2021). Machine learning in manufacturing and industry 4.0 applications. *International Journal of Production Research*, *59*(16), 4773-4778. https://doi.org/10.1080/00207543.2021.1956675

Saberironaghi, A., Ren, J., & El-Gindy, M. (2023). Defect detection methods for industrial products using deep learning techniques: A review. *Algorithms*, *16*(2), 95. https://doi.org/10.3390/a16020095

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. The Stata Journal: Promoting Communications on Statistics and Stata, 20(1), 3–29. https://doi.org/10.1177/1536867x20909688

Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. Information Fusion, 81, 84–90. https://doi.org/10.1016/j.inffus.2021.11.011

Sundaram, S., & Zeid, A. (2023). Artificial intelligence-based smart quality inspection for manufacturing. Micromachines, 14(3), 570. https://doi.org/10.3390/mi14030570

Taşar, B. (2022). Comparison analysis of machine learning algorithms for steel plate fault detection. *Duzce University Journal of Science and Technology*, *10*(3), 1578-1588. https://doi.org/10.29130/dubited.1058467

Tercan, H., & Meisen, T. (2022). Machine learning and deep learning based predictive quality in manufacturing: a systematic review. *Journal of Intelligent Manufacturing*, *33*(7), 1879-1905. https://doi.org/10.1007/s10845-022-01963-8

Tran, P. H., Ahmadi Nadi, A., Nguyen, T. H., Tran, K. D., & Tran, K. P. (2022). Application of machine learning in statistical process control charts: A survey and perspective. In *Control charts and machine learning for anomaly detection in manufacturing* (pp. 7-42). Springer, Cham. https://doi.org/10.1007/978-3-030-83819-5_2

Tzionis, G., Mouratidis, P., Kougka, G., Gialampoukidis, I., Vrochidis, S., Kompatsiaris, I., & Vlachopoulou, M. (2025). A review of explainable AI methods and their application in manufacturing systems. *Discover Applied Sciences*. https://doi.org/10.1007/s42452-025-07908-z

UCI Machine Learning Repository. (2023). *Steel plates faults dataset*. Retrieved 2026-03-12 from https://archive.ics.uci.edu/ml/datasets/Steel+Plates+Faults

Vasan, V., Sridharan, N. V., Vaithiyanathan, S., & Aghaei, M. (2024). Detection and classification of surface defects on hot-rolled steel using vision transformers. Heliyon, 10(19), e38498. https://doi.org/10.1016/j.heliyon.2024.e38498

Wang, L., Han, M., Li, X., Zhang, N., & Cheng, H. (2021). Review of Classification Methods on Unbalanced Data Sets. IEEE Access, 9, 64606–64628. https://doi.org/10.1109/access.2021.3074243

Wen, X., Shan, J., He, Y., & Song, K. (2023). Steel surface defect recognition: A survey. Coatings, 13(1), 17. https://doi.org/10.3390/coatings13010017

Yuan, Z., Ning, H., Tang, X., & Yang, Z. (2024). GDCP-YOLO: Enhancing Steel Surface Defect Detection Using Lightweight Machine Learning Approach. Electronics, 13(7), 1388. https://doi.org/10.3390/electronics13071388

Zhao, W., Ding, J., Huang, X., & Zhang, Y. (2025). Research on Milling Machine Predictive Maintenance Based on Machine Learning and SHAP Analysis in an Intelligent Manufacturing Environment. 2025 IEEE 15th International Conference on Electronics, Information, and Emergency Communication (ICEIEC), 1–5. https://doi.org/10.1109/iceiec65904.2025.11273156

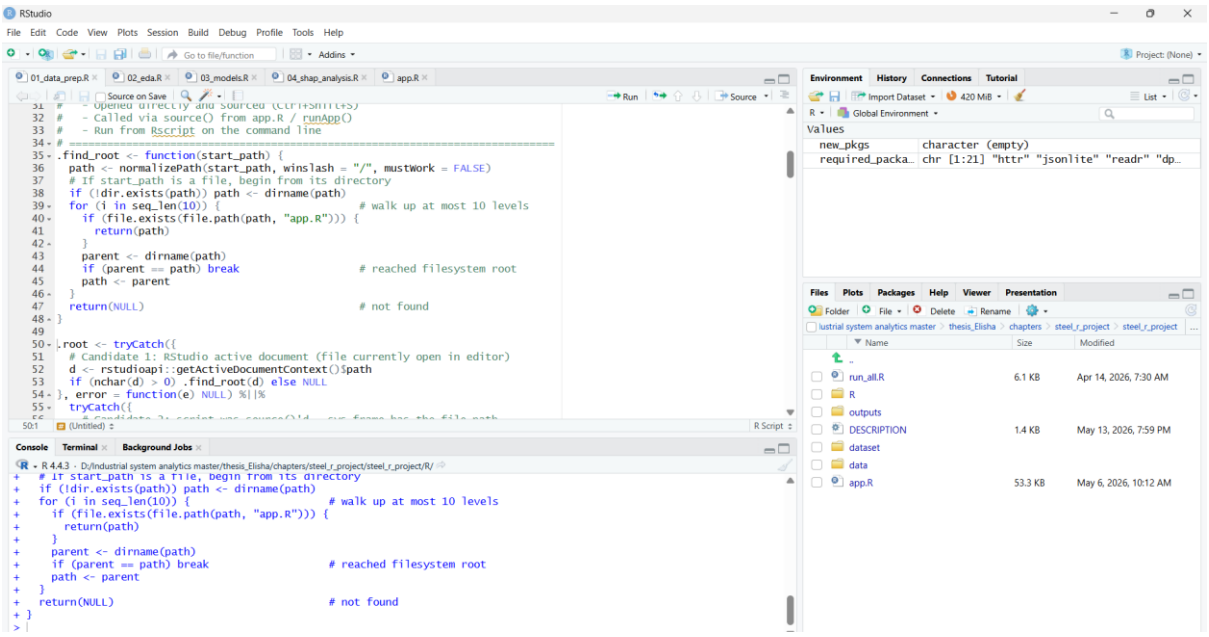
Zheng, T., Ardolino, M., Bacchetti, A., & Perona, M. (2020). The applications of Industry 4.0 technologies in the manufacturing context: a systematic literature review. International Journal of Production Research, 59(6), 1922–1954. https://doi.org/10.1080/00207543.2020.1824085

Appendices

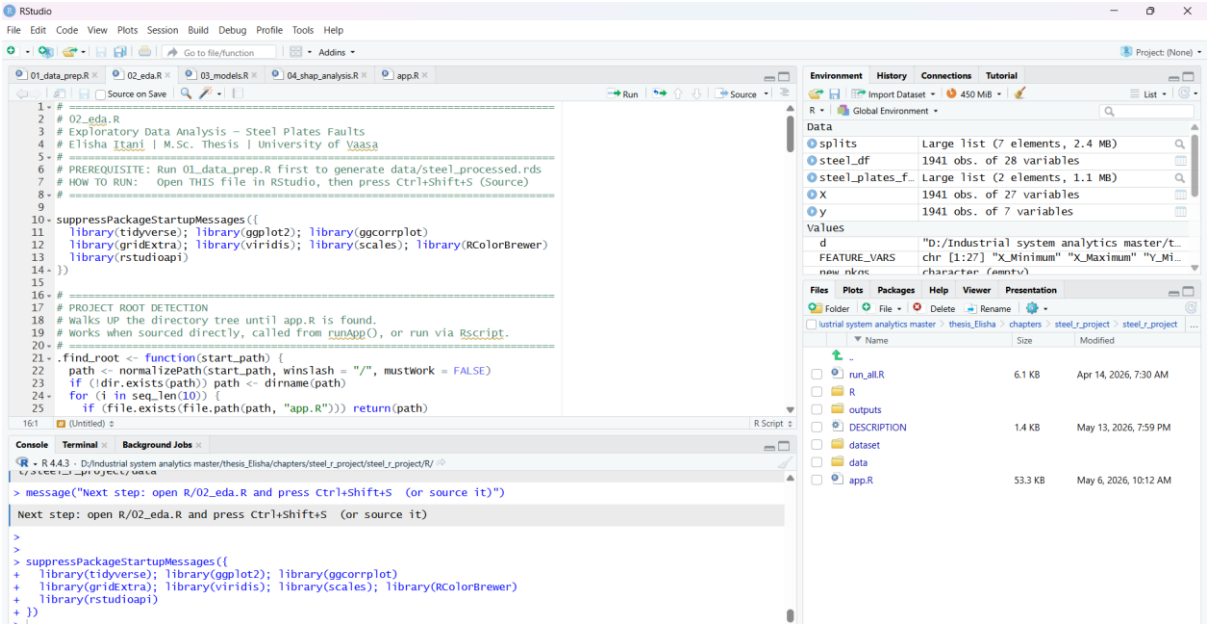
Appendix 1. Sample Code Snippets

The screenshots below display the main R code utilized in this research. Each snippet is related to an individual stage of the analysis pipeline explained in the thesis. The code is organized in a series of R scripts to be executed in order. 01_data_prep.R to prepare the data set, 02_eda.R to perform exploratory analysis, 03_models.R to train and test all five classifiers, 04_shap_analysis.R to compute both feature importance and SHAP values, and app.R to implement the interactive R Shiny dashboard.

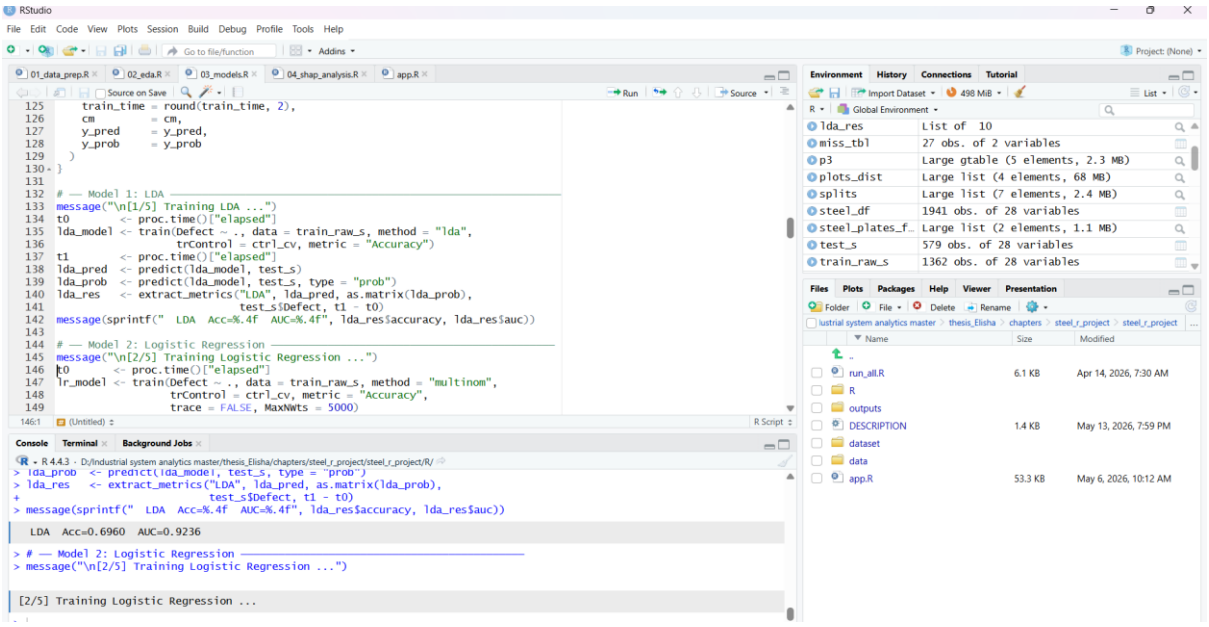
Snippet A.1: Project root detection and dashboard entry point (01_data_prep.R, referenced in Section 4.8)



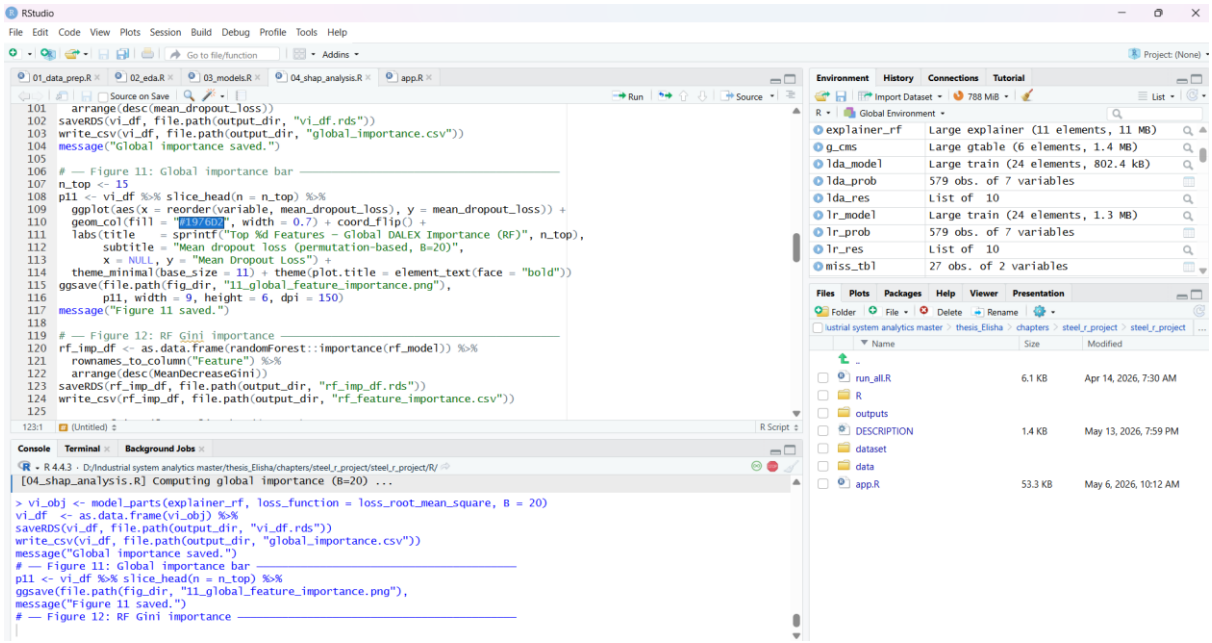
Snippet A.2: Exploratory Data Analysis library setup (02_eda.R, referenced in Section 4.2)



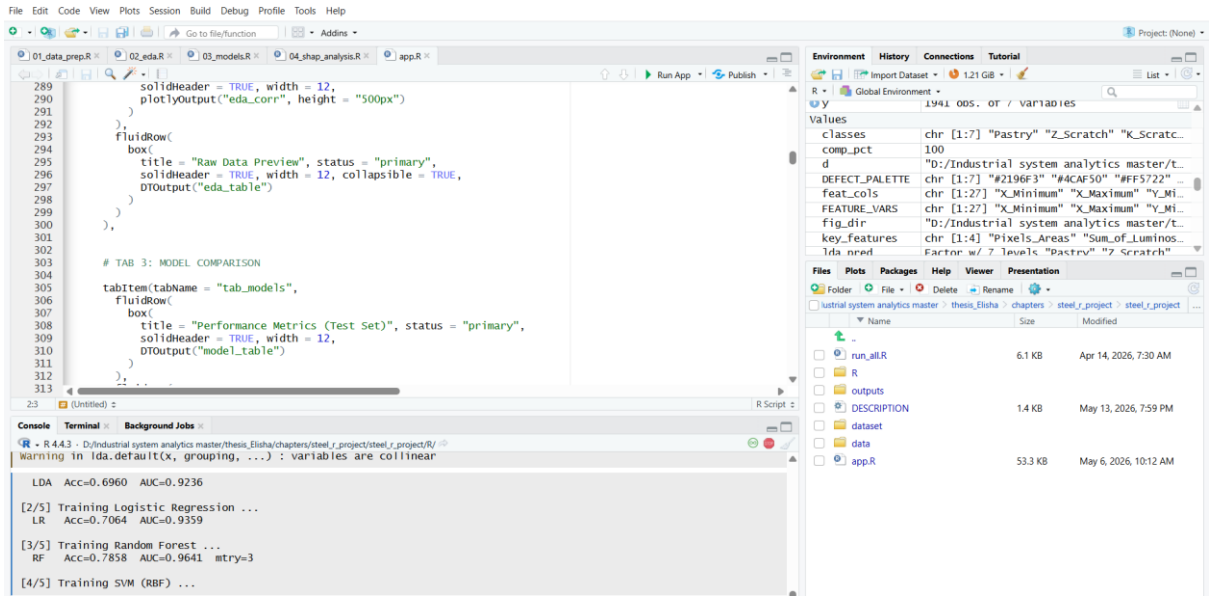
Snippet A.3: LDA and Logistic Regression training code (03_models.R, referenced in Sections 4.3.1 and 4.3.3)



Snippet A.4: DALEX global permutation importance and Gini importance (04_shap_analysis.R, referenced in Section 4.5.1)



Snippet A.5: Shiny dashboard tab layout and full model training console output including SVM diagnostic warning (app.R, referenced in Sections 4.3.2 and 4.8.1)



Appendix 2. Source Code

The complete source code and implementation files utilized in this thesis are publicly released on a One Drive to make sure there is both transparency and reproducibility.

steel_r_project