

ENCIRCLE

Enabling Circular Value Chains via Production Digitization and Human Empowerment

Deliverable D3.1 Integrated IIoT and soft sensors network for sustainable manufacturing and traceability

Project Month: M16

Type: Report

Dissemination level: Public

Rev.	Prepared by	Controlled by	Approved by	Date
2.0	Mario Suppa (RINA), D. Oikonomou (CERTH), D. Kotzaitis (CERTH), L. Leontaris (CERTH), C. Gkogkos (CERTH), A. Mitsiaki (CERTH), K. Roumeliotis (CERTH), D. Skoulis (CERTH), P. Charalampous (CERTH), M. Chatzakis (CERTH), A. Tzelepakis (CERTH), E. Theodoridis (CERTH)	Alice Reina (RINA), N. Dimitriou (CERTH) (WP3 Leader)	N. Dimitriou (CERTH)(Coordinator)	30/01/2026

TABLE OF CONTENTS

	Page
HISTORY OF CHANGES	3
LIST OF TABLES	3
LIST OF FIGURES	4
ABBREVIATIONS AND ACRONYMS	8
SUMMARY	10
1 INTRODUCTION	11
2 STATE OF THE ART	12
2.1 INDUSTRIAL INTERNET OF THINGS (IIOT) FOR MANUFACTURING	12
2.2 SOFT SENSORS FOR REAL-TIME ESTIMATION AND FORECASTING OF ENVIRONMENTAL IMPACT IN MANUFACTURING	15
2.2.1 Real-time image segmentation for Quality Control based on Deep Learning	15
2.2.2 Illumination Control for Enhanced Image Quality in Vision-Based Industrial Inspection	19
2.2.3 Predictive Analytics for Aluminium Melting Processes	22
2.3 DIFFUSION-BASED IMAGE GENERATION FOR INDUSTRIAL FURNACE MONITORING	23
2.4 AUDIO-BASED DEFECT DETECTION	25
3 SENSORS NETWORK	28
3.1 ALUMIL PILOT SITE	28
3.1.1 Edge Computing and Data Flow	29
3.1.2 Scrap Sorting	30
3.1.3 Scrap Loading	32
3.1.4 Process Monitoring	38
3.2 ZINQ PILOT SITE	41
3.2.1 Edge Computing and Data Flow	42
3.2.2 Pre-Treatment and Process Monitoring	42
3.2.3 Hot-Dip Galvanization and Post-Treatment	44
4 SOFT SENSORS	48
4.1 ALUMIL SOFT SENSORS	48
4.1.1 YOLO-based Visual Soft Sensors for Process Monitoring	48
4.1.2 Image Quality Stabilization via Adaptive Illumination and Exposure Control	53
4.1.3 Heat demand prediction based on scrap images	67
4.1.4 Dross prediction based on scrap images	77
4.1.5 Diffusion-based image generation of internal furnace dross formation	83
4.2 ARCEL SOFT SENSORS	84
4.2.1 Defect Description Prediction	85
4.2.2 Audio-based Defect Detection	94
4.3 ZINQ SOFT SENSORS	100
4.3.1 LLM Enhanced Industrial Defect Classification and Reasoning	101
4.3.2 Zero-Shot Anomaly Detection via LMM-Guided Inverse Reasoning	105
5 IIOT AND DATA GATHERING	108
5.1 IMPLEMENTATION IN PILOT 1 – ALUMIL	113
5.2 IMPLEMENTATION IN PILOT 2– ARCELIK	117
5.3 IMPLEMENTATION IN PILOT 3 – ZINQ	118
6 CONCLUSIONS	120

HISTORY OF CHANGES

VERSION	PUBLICATION DATE	CHANGE
1.0	16/12/2025	Initial RINA contribution
1.1	16/01/2026	CERTH contribution
2.0	26/01/2026	Internal review version
2.1	27/01/2026	Reviewer 1 & 2 comments (WLI, F-IEM)
2.2	28/01/2026	Revisions after internal review
2.3	29/01/2026	Quality review check (RINA)
2.4	30/01/2026	Final version for submission

LIST OF TABLES

Table 1: TP-LINK Archer MR600 Specifications.	29
Table 2: NVIDIA Jetson AGX Orin Specifications.	29
Table 3: HP OMEN Specifications.	30
Table 4: FLIR Blackfly S Specifications.	31
Table 5: Fujinon Lens CF12ZA-1S Specifications.	31
Table 6: NVIDIA Jetson AGX Xavier Specifications.	32
Table 7: Illumination Specifications	32
Table 8: Baumer VCXG.2-241C Specifications.	34
Table 9: Elgato HD60X Capture Card Specifications.	40
Table 10: Dell Latitude 5440 Specifications.	42
Table 11: GV-5290FA-C-HQ Specifications.	43
Table 12: Fujinon Lens CF8ZA-1S Specifications.	43
Table 13: Thermal camera configuration.	44
Table 14: FLIR A50 95° Thermal Camera Specifications.	46
Table 15: Multi-Illumination Dataset Statistics	59
Table 16: Summarized comparative results.	64
Table 17: Experimental results of the heat demand soft sensor on the test dataset	75
Table 18: Experimental Evaluation of Vision-Based Dross Prediction Models	82
Table 19: ARCEL production dataset rows and columns.	86
Table 20: Experimental results.	92
Table 21: Dataset statistics.	95
Table 22: Classification evaluation metrics.	99

LIST OF FIGURES

Figure 1: Evolution of industrial systems towards Industry 4.0, highlighting the increasing role of connectivity, cyber-physical systems and Industrial Internet of Things technologies (adapted from Afrin et al. [2]).	12
Figure 2: IIoT network automation architecture with standard services, such as generic communication, management and interoperability functions (image from [3]).	13
Figure 3: Hierarchical development architecture of the IIoT intelligence (image from [4]).	14
Figure 4: Conceptual research model linking Green IIoT adoption with circular economy practices and environmental performance (image from [5]).	14
Figure 5: Schematic representation of the YOLOv8 (image from [6])	16
Figure 6: YOLO-v8 architecture (a) Backbone (b) Neck (c) Head (image from [37])	17
Figure 7: SAM2 architecture. For each frame, the segmentation is predicted using the current input and information stored from previous frames. Frames are processed one at a time by the image encoder and combined with past object memories through memory attention. The mask decoder then produces the segmentation for the current frame, and a memory encoder updates the stored representations for future frames. (image from [64])	18
Figure 8: System Overview of the Automated Exposure Control Framework (image from [40])	19
Figure 9: Comparison between conventional built-in auto-exposure and the DRL-based framework. The DRL-AE agent achieves rapid convergence under lighting changes and significantly enhances object detection and feature extraction performance. (image from [32])	20
Figure 10: Background on DINOv2 training data construction (image from [69])	22
Figure 11: CLIP trains an image encoder and a text encoder to match paired image–text examples via a contrastive objective. At inference, class names or descriptions are embedded as text prototypes and compared to an image embedding to enable zero-shot classification. (image from REF _Ref6314 \n \n h [14][14])	23
Figure 12: LDM architecture. (image from [88])	24
Figure 13: A text prompt is encoded and passed to the SDXL base model, which generates an initial latent representation. The latent is then optionally forwarded to a refiner model that performs the final denoising steps to produce the final image. (image from [66])	24
Figure 14: High-level overview of FLUX.1. (image from [90])	25
Figure 15: SOM-LSTM defect detection pipeline, which generates different FFT frequencies, it embeds them into two dimensional space and then it utilizes an LSTM for classification. (Image from [78])	26
Figure 16: Causal Audio Transformer (CAT) and proposed acoustic attention module (Image from [80]).	27
Figure 17: Conformer encoder architecture (Image from [81])	27
Figure 18: ALUMIL pilot site sensor overview.	28
Figure 19: Local Network.	29
Figure 20: TP-LINK Archer MR600.	29
Figure 21 : NVIDIA Jetson AGX Orin.	30
Figure 22: FLIR Blackfly S.	31
Figure 23: Fujinon Lens CF12ZA-1S.	31
Figure 24: NVIDIA Jetson AGX Xavier.	32
Figure 25: Camera capture above the scrap sorting conveyor belt.	32
Figure 26: LED Floodlight 500W Cold White 5000K	33
Figure 27: Scrap loading camera with casing and illumination device.	33
Figure 28: Baumer VCXG.2-241C.	34
Figure 29: 3D snapshots from camera casing design.	34
Figure 30: Initial non-optimal scrap loading captures.	35
Figure 31: Initial non-optimal scrap loading captures.	36
Figure 32: Weight sensor output on LED display.	37
Figure 33: Production data for oven loads.	37

PROJECT 101178230 – ENCIRCLE

Figure 34: Production data for castings.	38
Figure 35: SCADA system for furnace condition management.	38
Figure 36: Initial approach recording the screen with a camera.	39
Figure 37: ALUM SCADA server signal mirroring via capture card and OCR.	39
Figure 38: The capture card and the cabling used for monitor mirroring.	39
Figure 39: Cabling to mirror the ALUM SCADA data server to the computational edge node.	40
Figure 40: OCR captured data from the SCADA software.	40
Figure 41: In furnace image captures.	41
Figure 42: ZINQ Pilot site sensor overview.	42
Figure 43: GV-5290FA-C-HQ machine vision camera.	43
Figure 44: Fujinon Lens CF8ZA-1S.	43
Figure 45: Hot dip galvanization.	44
Figure 46: Projected camera installation.	44
Figure 47: Hot- Dip Galvanization parts.	45
Figure 48: FLIR A50 95° Thermal Camera.	46
Figure 49: Documentation of which bath is used for how long for each order.	47
Figure 50: Thickness measurement	47
Figure 51: Production parameters.	47
Figure 52: Visualization of the automated semantic annotation process within the LabelMe interface. The system converts YOLO-based detections into editable polygon labels, where distinct colors correspond to specific scrap categories (e.g., circles, alumil_return)	50
Figure 53: Statistical analysis of the training dataset, illustrating class distribution and bounding box spatial properties.	51
Figure 54: Precision-Confidence curve for the YOLO-based scrap detection model. High precision levels are maintained across most categories, ensuring reliable identification of scrap types before furnace entry.	52
Figure 55: F1-Confidence curve illustrating the optimal operational point.	53
Figure 56: The camera is mounted above the scrap wagon near the factory door, where direct sunlight enters the scene (a), causing fixed exposure settings to produce underexposed images at night (b) and overexposed images during daylight (c).	54
Figure 57: Camera exposure is adjusted based on image brightness measurements. At each iteration, the exposure time is increased or decreased in discrete steps until the measured brightness falls within the predefined acceptable range. The top example shows a saturated image corrected by reducing the exposure over four steps, while the bottom example shows an underexposed image corrected by increasing the exposure to reach an adequate brightness level.	56
Figure 58: In (a), different types of defects and surface textures manifest differently as illumination conditions change. In the left and middle columns, the low illumination level (L0) appears optimal for inspecting the surface defects, while in the right column, higher illumination (L6) enhances the defect's prominence.	58
Figure 59: For each sample, illustrated at the top, 8 original images were captured under 8 different illumination levels. At the bottom, each original image was split into patches, capturing defects across the 8 levels (columns). The defect types include Deformation (Df), Scratch (Sc), and Spot (Sp).	58
Figure 60: DRL-based Multi-Illumination Optimization.	61
Figure 61: DRL agent training and inference framework for illumination control. During training the segmentation performance is used to reward the agent. During inference the agent determines the final decision after T interactions with the environment.	62
Figure 62: Comparison results (harmonic IoU) between the proposed DRL approach and the baseline models.	64
Figure 63: Iterative multi-illumination optimization starting from the initial states L7 (top 4 rows) and L0 (bottom 4 rows).	66
Figure 64: State-transition heatmaps of agent decisions. In (a), (b), (e) and (f) two test samples at $t=0$ and $t=10$ are illustrated, with the respective heatmaps in (c), (d), (g) and (h).	67

PROJECT 101178230 – ENCIRCLE

Figure 65: Selected images for loading alumil_2025-05-02_18-27-34 containing [alumil_bars, alumil_returns, alumil_returns_painted].	69
Figure 66: Sensor timeseries for loading alumil_2025-05-02_18-27-34 with the variables corresponding to the cold furnace: CC_ROOF_TEMPERATURE, CC_HEAT_DEMAND, and CC_CHAMBER_PRESSURE	70
Figure 67: Characteristic internal furnace camera frames selected for loading alumil_2025-05-02_18-27-34.	70
Figure 68: Dross detected timeseries for loading alumil_2025-05-02_18-27-34.	71
Figure 69: Proposed model architecture for heat demand prediction. (i) Pre-processing pipeline including masking based on scrap segmentation masks and patching, (ii) per-patch embeddings generated from backbone visual model, (iii) aggregation and regression head for final heat demand prediction.	71
Figure 70: Examples of cropped images focused on individual scrap categories for loading alumil_2025-05-02_18-27-34.	72
Figure 71: Representative samples from the processed dataset, illustrating the geometric transformations (rotation and cropping) applied to scrap images and the selected frames for dross detection.	79
Figure 72: DINOv2-based Embedding Pipeline for Dross Percentage Estimation	80
Figure 73: Training Loss Evolution (MSE) across 100 Epochs for Different Input Configurations.	82
Figure 74: RMSE across 100 Epochs for Different Input Configurations.	83
Figure 75: Qualitative comparison between real furnace dross images (top), images generated by fine-tuned Stable Diffusion 1.5 (middle), and images generated by fine-tuned Stable Diffusion XL (bottom). SDXL samples exhibit higher perceptual similarity to real images, with more coherent texture and improved sharpness.	84
Figure 76: Raw production data.	85
Figure 77: Input (white) and output (grey) columns of production data.	86
Figure 78: Fault types over production time.	87
Figure 79: Fault type distribution.	87
Figure 80: Fault description distribution.	88
Figure 81: Correlation heatmaps: (a) Fault type and Model. (b) Fault type and Days. (c) Fault type and Category. (d) Fault description and Model. (e) Fault description and Days. (f) Fault description and Category.	89
Figure 82: Time series analysis between aggregate categories and production date.	89
Figure 83: Bar chart plot of aggregated fault descriptions and models.	90
Figure 84: Bar chart plot of aggregated fault descriptions and fault type.	90
Figure 85: Bar chart plot of aggregated fault descriptions and category4 column.	90
Figure 86: High level diagram for the experiment.	91
Figure 87: Line plot of precision per defect description for all models.	93
Figure 88: Line plot of recall per defect description for all models.	93
Figure 89: Line plot of f1-score1 per defect description for all models.	94
Figure 90: ARCEL audio capturing.	95
Figure 91: ToyADMOS Conveyor dataset spreadsheet.	96
Figure 92: Conformer based audio classification architecture.	97
Figure 93: 2D and 3D PCA of conformer embeddings per fault type.	98
Figure 94: Classification training and validation loss behaviour.	99
Figure 95: Per-class performance metrics.	100
Figure 96: Proposed pipeline for explainable/interpretable defect classification. In the first stage we use Anomaly-OV to extract descriptions of the type and possible cause of defect. In the second stage we encode the human-understandable text output using the CLIP model in order to use the representations for training a classification head. In the third stage we train a classification head for classifying images between normal and defective, using as input the representations produced in the second stage.	103
Figure 97: Using Anomaly-OV for extracting corrective actions.	103
Figure 98: Evaluation of the proposed model in terms of classification accuracy.	104

PROJECT 101178230 – ENCIRCLE

Figure 99: Qualitative Results. Type and Cause of Defect Responses	104
Figure 100: Qualitative Results. Corrective Actions Responses	105
Figure 101: AUROC results for MVTEC and VISA datasets.	107
Figure 102: Comparison with SOTA.	107
Figure 103: Encircle dedicated library and used driver in middleware	108
Figure 104: ENCIRCLE dedicated library and used driver in middleware	109
Figure 105: ENCIRCLE solution for monitoring docker deploy	110
Figure 106: SmartUnifier use cases.	111
Figure 107: Integration with equipment.	112
Figure 108: SmartUnifier scalability.	112
Figure 109: Alumil sensor data	113
Figure 110: Dross estimations data	114
Figure 111: Data attribute from REST service setup	114
Figure 112: Alumil scrap inspections with polygons	115
Figure 113: Image with polygons configuration	115
Figure 114: Scrap inspection history with Zip download feature (images and data) for each extracted result	116
Figure 115: Dross estimations history with CSV download feature (images and data)for each extracted result	116
Figure 116: Alumil sensor data plot	117
Figure 117: ARCEL pilot implementation plan.	118
Figure 118: ZINQ pilot implementation plan.	119

ABBREVIATIONS AND ACRONYMS

AAS	Asset Administration Shell
AST	Audio Spectrogram Transformer
AUROC	Area Under the Receiver Operating Characteristic Curve
CNN	Convolutional Neural Network
CRF	Camera Response Function
CSV	Comma-Separated Values
CUDA	Compute Unified Device Architecture
DDPG	Deep Deterministic Policy Gradient
DINOv2	Self-Supervised Vision Transformer Model (v2)
DL	Deep Learning
DRL	Deep Reinforcement Learning
DSS	Decision Support System
FFT	Fast Fourier Transform
FOV	Field of View
FPS	Frames Per Second
GAN	Generative Adversarial Network
GAE	Generalized Advantage Estimation
GIIoT	Green Industrial Internet of Things
GPU	Graphics Processing Unit
HC	Hot Chamber
IIoT	Industrial Internet of Things
IoT	Internet of Things
IR	Infrared
JSON	JavaScript Object Notation
KPI	Key Performance Indicator
LDM	Latent Diffusion Model
LLM	Large Language Model
LMM	Large Multimodal Model
LoRA	Low-Rank Adaptation
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAP	Mean Average Precision
MFCC	Mel-Frequency Cepstral Coefficients
MQTT	Message Queuing Telemetry Transport
MSE	Mean Squared Error
NeRF	Neural Radiance Fields
OCR	Optical Character Recognition
PCA	Principal Component Analysis
PLC	Programmable Logic Controller
PPO	Proximal Policy Optimization
PVS	Promptable Visual Segmentation
RA	Reinforcement Learning Agent
REST	Representational State Transfer
RLU	Remaining Useful Life
RMSE	Root Mean Squared Error

PROJECT 101178230 – ENCIRCLE

RNN	Recurrent Neural Network
ROI	Region of Interest
RUL	Remaining Useful Life
SAM	Segment Anything Model
SAM2	Segment Anything Model 2
SCADA	Supervisory Control and Data Acquisition
SD	Stable Diffusion
SDXL	Stable Diffusion XL
SDEdit	Stochastic Differential Editing
SGD	Stochastic Gradient Descent
SOM	Self-Organizing Map
SOTA	State of the Art
STFT	Short-Time Fourier Transform
TOPS	Trillion Operations Per Second
TPU	Tensor Processing Unit
U-Net	Convolutional Neural Network Architecture
VAE	Variational Autoencoder
ViT	Vision Transformer
VLM	Vision-Language Model
VMD	Vision Model Descriptor
ZSAD	Zero-Shot Anomaly Detection

SUMMARY

This deliverable presents the design and implementation of an integrated Industrial Internet of Things (IIoT) and soft sensor infrastructure developed within Work Package 3 of the ENCIRCLE project. The overall objective is to support sustainable, circular manufacturing by enabling non-intrusive data acquisition, real-time monitoring, predictive analytics, and intelligent decision support across heterogeneous industrial environments.

The work covers the deployment of sensor networks and edge computing architectures at three industrial pilot sites (ALUMIL, ARCELMIK, and ZINQ). These infrastructures integrate legacy equipment with modern IIoT devices using non-invasive strategies such as machine vision, thermal sensing, audio acquisition, and Optical Character Recognition (OCR), ensuring interoperability without disrupting existing production processes. Data acquisition and routing are handled through MQTT-based middleware and standardized REST interfaces, enabling reliable data flow toward higher-level analytics and visualization platforms.

Building on this sensing layer, the deliverable introduces a comprehensive set of AI-based soft sensors that transform heterogeneous multi-modal data into actionable process intelligence. Vision-based soft sensors estimate scrap composition, furnace heat demand, and dross formation, while adaptive illumination and exposure control mechanisms improve image stability in challenging industrial conditions. Additional soft sensors address audio-based defect detection, text-based defect description prediction, and explainable industrial anomaly detection, leveraging deep learning, foundation models, and large multimodal language models. Particular emphasis is placed on handling data scarcity through self-supervised learning, zero-shot methods, and human-in-the-loop strategies.

The proposed solutions demonstrate the feasibility of real-time, explainable, and scalable intelligence for industrial operations, supporting early estimation of energy consumption, material losses, and quality deviations. These capabilities directly contribute to improved resource efficiency, reduced waste, and enhanced traceability. Finally, the integration with ENCIRCLE data platforms enables seamless use of soft sensor outputs by downstream components such as Digital Twins, environmental impact analytics, and green control loops. Overall, this deliverable establishes a robust technical foundation for AI-driven, circular manufacturing aligned with ENCIRCLE's sustainability objectives.

1 INTRODUCTION

Work Package 3 delivers an integrated digital ecosystem supporting sustainable manufacturing by combining IIoT, AI, Digital Twins and advanced human-in-the-loop control. The set of tasks enables monitoring, prediction, simulation and optimization of production processes, with a strong focus on environmental impact reduction, traceability and transparency.

In particular, an IIoT infrastructure with edge computing is implemented to interconnect sensors, actuators and industrial controllers, ensuring interoperability, cybersecurity and compliance with the AAS standard. This infrastructure enables near real-time remote monitoring and control of production assets, while supporting data collection and recordkeeping for product traceability across the entire lifecycle.

Building on this foundation, intelligent soft sensors are developed for real-time estimation and forecasting of environmental KPIs (e.g. energy consumption, emissions and waste), leveraging AI and statistical models capable of exploiting multi-sensor correlations and temporal patterns. Soft sensors also include models for Remaining Useful Life and Remaining Value estimation of assets and products, supporting maintenance planning, lifecycle extension and circular economy strategies.

In parallel, high-fidelity Digital Twins are created on-the-fly using neural-implicit techniques (e.g. NeRF), by combining video-based multimodal data, technical drawings and information provided by the soft sensors. The resulting Digital Twins are AAS-compliant, include logistics aspects, and enable simulation, training and advanced analysis, while significantly reducing time, cost and human effort through automated workflows and expert-guided active learning approaches.

Finally, Digital Twins feed AI-based green control loops, including reinforcement learning methods, to optimize production processes in an eco-friendly manner. These control loops continuously analyse real-time IIoT data to detect anomalies and suboptimal operating conditions, triggering process reconfigurations at machine and production-line level. Decision-making is supported by dedicated decision support interfaces and human-in-the-loop mechanisms, ensuring alignment between AI recommendations, production objectives and long-term sustainability goals.

Overall, WP3 provides an end-to-end solution for green manufacturing, integrating monitoring, forecasting, simulation and intelligent control, with a strong emphasis on sustainability, traceability and human-machine collaboration.

2 STATE OF THE ART

2.1 INDUSTRIAL INTERNET OF THINGS (IIOT) FOR MANUFACTURING

The Industrial Internet of Things (IIoT) refers to the application of IoT technologies in industrial environments, enabling the integration of sensors, actuators and industrial equipment to support data-driven manufacturing operations [1]. IIoT systems are designed to operate in heterogeneous industrial environments, where diverse production assets and control systems coexist and need to be interconnected. Through the deployment of sensing devices and communication technologies, IIoT infrastructures enable real-time data acquisition and continuous monitoring of manufacturing processes. Commonly, IIoT solutions adopt layered architectures, comprising sensing and actuation layers, communication infrastructures and application-level platforms for data processing. Compared to consumer-oriented IoT applications, IIoT deployments impose stricter requirements in terms of reliability, latency and security, reflecting their use in mission-critical industrial operations. In this context, IIoT primarily acts as an enabling infrastructure, providing the foundational connectivity and data acquisition capabilities upon which higher-level analytics, assessment, and optimization are built.

As illustrated in Figure 1, the transition towards Industry 4.0 is characterized by the increasing integration of connectivity and data-driven technologies within industrial systems. The systematic review by Afrin et al. [2] provides a comprehensive overview of Industrial Internet of Things (IIoT) implementations across a wide range of industrial sectors, highlighting how IIoT infrastructures are deployed to interconnect sensors, machines and industrial systems for monitoring and control purposes. The implementations analysed in this study demonstrate that IIoT enables continuous data acquisition from heterogeneous production assets, supporting real-time visibility of industrial processes and improved operational efficiency. However, the same review also identifies several challenges that hinder large-scale deployment, including interoperability issues among heterogeneous devices, data management complexity, cybersecurity concerns and limitations related to scalability and reliability. To address these challenges, Afrin et al. report the adoption of standardized communication protocols, the use of IoT gateways and middleware platforms, and the integration of edge computing to reduce latency and communication overhead.

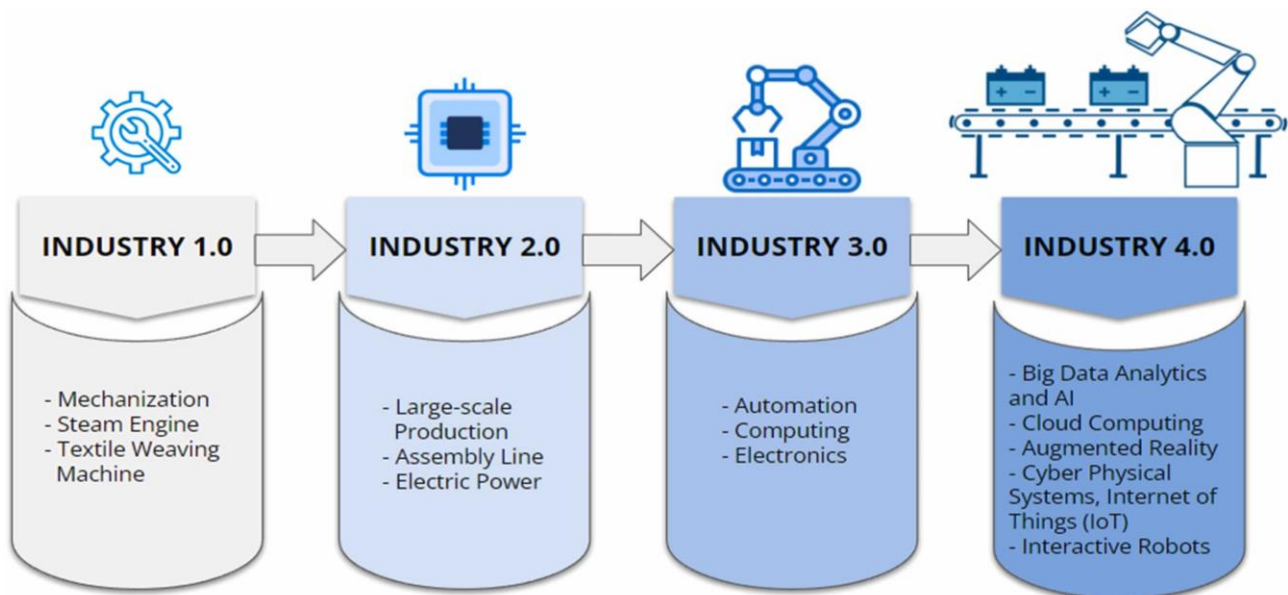


Figure 1: Evolution of industrial systems towards Industry 4.0, highlighting the increasing role of connectivity, cyber-physical systems and Industrial Internet of Things technologies (adapted from Afrin et al. [2]).

The survey by Chi et al. [3] on IIoT network automation highlights the increasing complexity of industrial communication infrastructures and the need for intelligent mechanisms to manage heterogeneous devices, protocols and data flows. Network automation is identified as a key enabler for scalable and reliable IIoT deployments, supporting dynamic configuration, monitoring and optimization of industrial networks. The literature emphasizes that IIoT environments typically comprise heterogeneous sensing devices, industrial controllers and legacy systems interconnected through multiple communication technologies, which poses significant challenges in terms of interoperability, latency and reliability. To address these challenges, automated network management approaches based on software-defined networking, programmable gateways

and middleware layers have been proposed, enabling flexible data routing and resource allocation. Overall, network automation is positioned as a critical layer between the physical IIoT infrastructure and higher-level industrial platforms, ensuring that data generated by sensors and production assets can be delivered reliably and in near real time to downstream monitoring, assessment and control components. Figure 2, illustrates a layered IIoT communication architecture comprising device, network and application layers. In this work, we focus on the device and network layers, where heterogeneous sensors and actuators are integrated via gateways and middleware. Together, these layers ensure dependable data capture and transport, while providing protocol mediation and interoperability that support scalable delivery to monitoring and control services.

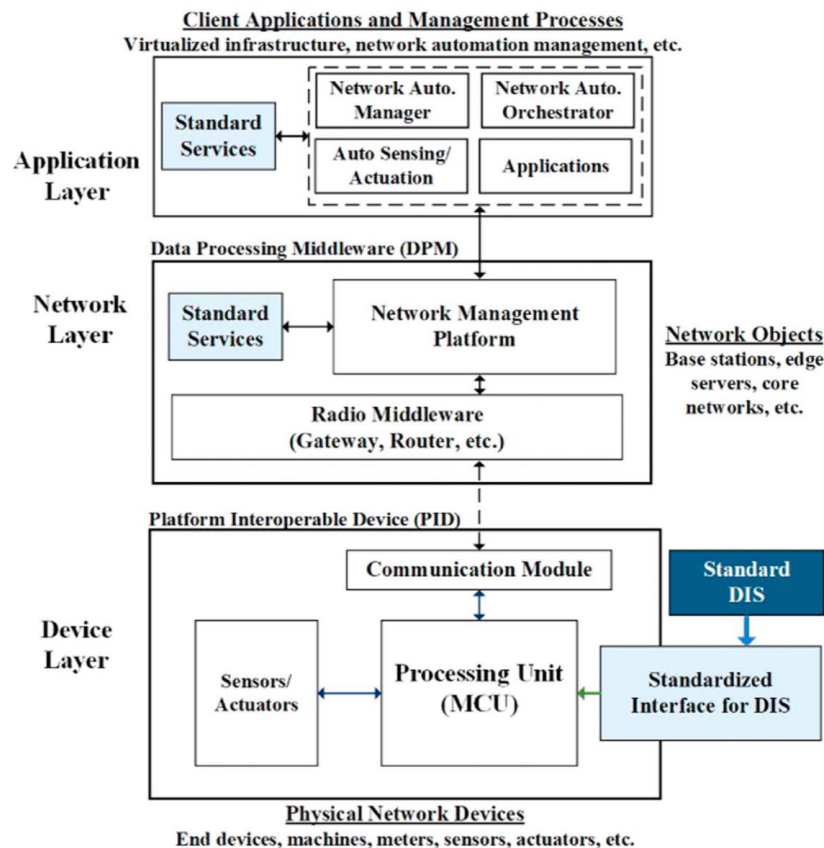


Figure 2: IIoT network automation architecture with standard services, such as generic communication, management and interoperability functions (image from [3]).

The literature review by Hu et al. [4] on IIoT intelligence emphasizes that, beyond connectivity and data acquisition, intelligence mechanisms built on top of IIoT infrastructures play a key role in supporting smart manufacturing. The review indicates that IIoT intelligence focuses on transforming raw data collected from sensors and industrial systems into actionable insights for monitoring, assessment and decision support. Such intelligence layers typically leverage data analytics, estimation and reasoning mechanisms to improve production efficiency, quality and operational reliability. Importantly, IIoT intelligence is not treated as an isolated component, but as a functional layer tightly coupled with the underlying IIoT platforms and communication infrastructures.

The study further highlights that the effectiveness of intelligent manufacturing solutions depends on reliable data availability, interoperability and near real-time processing capabilities provided by the IIoT network. In the context of the present work, emphasis is placed on the intelligence layer built on top of IIoT infrastructures, which enables data-driven assessment and decision support rather than application-specific implementations (Figure 4). Overall, IIoT intelligence is positioned as a key enabler for data-driven manufacturing, forming the basis for advanced assessment, optimization and control functionalities.

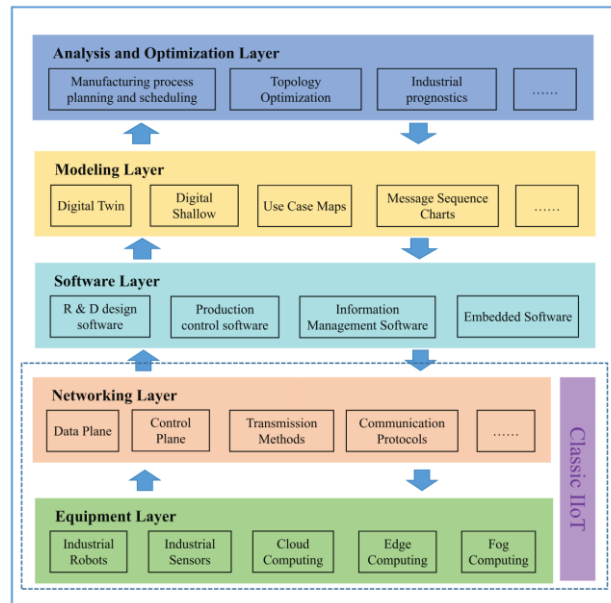


Figure 3: Hierarchical development architecture of the IIoT intelligence (image from [4]).

The study by Xu et al. [5] investigates IIoT adoption from a higher-level organizational perspective, focusing on how Green IIoT (GIIoT) contributes to environmental performance through dynamic capabilities and circular economy practices. As illustrated in Figure 4, GIIoT adoption does not directly lead to improved environmental outcomes, but instead operates through mediating mechanisms such as sensing, learning, seizing and reconfiguring capabilities, as well as circular economy practices. This perspective highlights that IIoT infrastructures should not be viewed solely as data collection and automation layers, but as enablers of organizational learning and adaptive decision-making. Such a view aligns with higher-level analytics, where sensor data and analytics outputs are transformed into strategic insights that support sustainability-oriented decision processes. In contrast to purely technical architectures, this model emphasizes the importance of integrating IIoT-driven analytics with organizational capabilities, enabling firms to translate real-time data into environmental impact reduction and resource efficiency improvements.

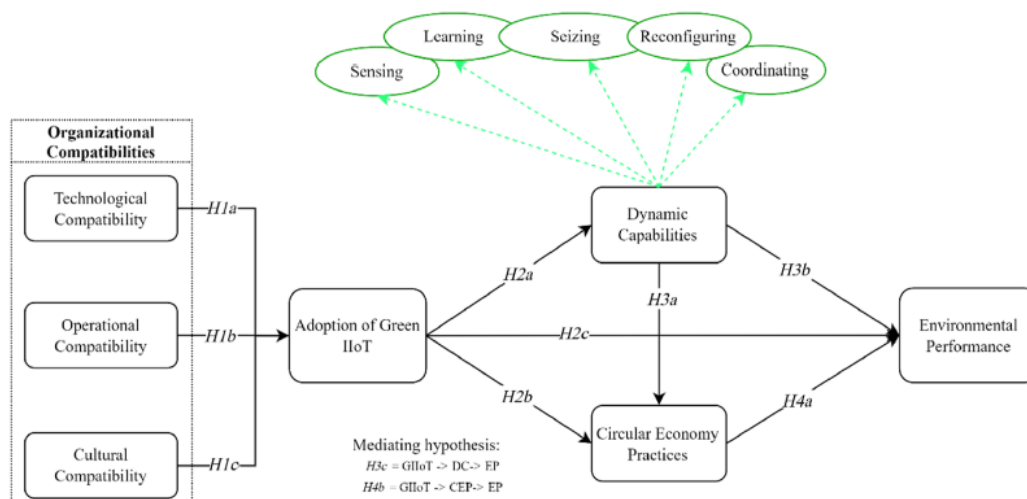


Figure 4: Conceptual research model linking Green IIoT adoption with circular economy practices and environmental performance (image from [5]).

While layered IIoT architectures provide a general blueprint for connectivity, recent literature stresses that “best-effort” networking is often insufficient for manufacturing control loops and safety-relevant coordination. Zhang et al. [92] provide a comprehensive survey of Time-Sensitive Networking (TSN) for industrial automation, positioning TSN as a key pathway towards deterministic, converged industrial Ethernet that can simultaneously support heterogeneous traffic classes (e.g., periodic control, alarms, monitoring streams) with bounded latency and jitter. The survey synthesizes how TSN mechanisms, spanning precise time synchronization, scheduled traffic shaping, reliability support and resource management, address stringent industrial requirements and enable tighter IT/OT convergence. Importantly, this line of work reinforces that the “network layer” in IIoT is not merely a transport substrate, but a critical determinant of whether real-time industrial data can be delivered with the guarantees required by downstream monitoring, assessment and closed-loop control.

Beyond deterministic delivery, manufacturing IIoT deployments must address interoperability across vendors and “brown-field” assets. Trifonov and Heffernan [93] argue that the combination of OPC UA (semantic, service-oriented industrial data exchange) with TSN (real-time Ethernet behaviour) provides a vendor-independent successor pathway for factory networks, explicitly targeting cross-vendor interoperability and converged traffic support. Their work describes an implementation-oriented view of an OPC UA-TSN factory network (with emphasis on the field layer) and reports timing measurements indicating that OPC UA-TSN can meet basic critical timing needs expected of fieldbus-style communication. This complements broad IIoT surveys by grounding interoperability in concrete protocol stacks and migration considerations, which is particularly relevant when IIoT data acquisition must integrate heterogeneous sensors, PLCs and legacy equipment without sacrificing timing guarantees.

Finally, large-scale IIoT connectivity expands the attack surface across sensing devices, gateways, networks and platforms. Alotaibi [94] provides a recent, structured survey of IIoT security requirements, attacks and AI/edge-enabled mitigation approaches, organizing threats by IIoT layers and emphasizing that edge computing can reduce latency and bandwidth pressure while also enabling localized security functions (e.g., detection/response closer to assets). The survey highlights a key deployment trade-off: pushing capabilities to the edge can strengthen real-time security monitoring and resilience, but it also increases operational complexity (e.g., distributed patching and maintenance). For manufacturing IIoT, these findings reinforce that reliability/latency requirements must be addressed jointly with security and lifecycle management, since the integrity and availability of sensor data directly condition the trustworthiness of higher-level analytics and impact assessment functions.

2.2 SOFT SENSORS FOR REAL-TIME ESTIMATION AND FORECASTING OF ENVIRONMENTAL IMPACT IN MANUFACTURING

2.2.1 Real-time image segmentation for Quality Control based on Deep Learning

Recent work in the literature investigates the use of vision-based deep learning architectures (Figure 5), such as YOLOv8, for object detection and process monitoring in industrial environments with complex backgrounds and challenging visual conditions, as presented in Schneidereit et al. [6][6]. As illustrated through the architectural diagrams and heatmap-based visualizations included in the paper, the use of multi-scale feature extraction allows the model to reliably interpret complex industrial images, even in the presence of visually similar materials and reflective surfaces. In addition, explainability mechanisms show which parts of the image the model relies on when making its predictions, enabling indirect monitoring of the process and increasing confidence in the resulting vision-based estimations. These characteristics match the core idea of soft sensing, where visual information is used to derive meaningful process indicators instead of direct physical measurements.

Based on the same methodological principles, YOLOv8 [91] was employed in the Alumil industrial use case, where images of scrap wagons were acquired prior to furnace charging and automatically annotated to identify and classify the different scrap categories contained within the wagon. Furthermore, a vision-based soft sensing approach was applied for dross estimation during the melting process, using images captured by an internal furnace camera to annotate and estimate the surface coverage of dross inside the furnace. Further details regarding the implementation, data processing pipeline, and experimental results of these applications are provided in section 4. This demonstrates the applicability of state-of-the-art vision-based object detection approaches, as discussed in the literature, to real industrial soft sensing tasks involving scrap characterization and in-furnace process monitoring.

Related by the comparative study of Haque et al. [35], which evaluates its performance against the traditional Mask R-CNN framework in quality assurance tasks. Their research highlights that YOLOv8-seg provides a significantly more efficient balance between inference speed and detection precision, achieving a high mean Average Precision of 0.939 while requiring substantially lower computational time compared to Mask R-CNN. However, while the effectiveness of such architectures has been widely demonstrated across various industrial domain, including leather defect segmentation, metallic plate inspection, corrosion detection in industrial imagery, and even human posture analysis. Most existing literature focuses on standardized manufacturing environments where products exhibit uniform features and controlled orientations. In contrast, the secondary aluminium industry presents unique challenges that have not been widely addressed; specifically, the visual appearance of scrap categories can vary drastically from batch to batch, and the material is typically accumulated and piled within wagons, leading to complex occlusions and overlapping objects. Our implementation addresses these challenges by developing a vision-based soft sensor capable of identifying diverse scrap categories within each individual loading sequence, moving beyond simple defect detection toward robust material characterization in unpredictable and unstructured industrial environments.

The selection of the YOLOv8 architecture is further substantiated by comparative performance evaluations in industrial contexts. According to Yao et al. [36], who conducted a benchmark study on surface defect detection in metallic plates, YOLOv8 consistently outperforms its predecessors, such as YOLOv5 and YOLOv7, in both detection accuracy and processing speed. Their experimental results demonstrate that YOLOv8 achieves superior mean Average Precision while maintaining an exceptionally fast inference time of 2.3 ms. This performance boost is primarily attributed to the integration of the C2f module, which replaces older bottleneck structures to enhance feature extraction and allow the model to identify small-scale defects more effectively. By achieving state-of-the-art results on metallic surfaces with similar reflective properties to industrial materials, this framework provides a solid technical foundation for implementing high-speed, vision-based sensing in metal production lines.

The applicability of the YOLOv8 architecture in the recycling sector is further demonstrated by the work of Anwar et al. [37], who developed a real-time waste detection and classification model for automated waste management. Their study addresses the inherent inefficiencies of manual sorting by utilizing the YOLOv8 deep learning architecture (illustrated in Figure 6), which is optimized for high-speed feature extraction and precise material categorization. Their study addresses the inherent inefficiencies and safety concerns of manual sorting processes by utilizing a deep learning framework capable of accurately categorizing municipal waste into five distinct classes: paper, plastic, metal, glass, and organic materials. By training the model on a diverse dataset of recyclable items, the researchers achieved a mean Average Precision of 0.82, highlighting the architecture's effectiveness in identifying various material types under real-time operational conditions. This implementation underscores the potential of YOLO-based soft sensors to enhance industrial sustainability by enabling high-speed, automated material recovery within a circular economy framework. The real-time processing capabilities of YOLOv8 allow for the immediate identification and sorting of recyclables on moving conveyor systems, significantly improving efficiency compared to traditional human-led methods. Such advancements in waste classification provide a critical scientific reference for the Alumi use case, as they validate the use of state-of-the-art object detection to manage and categorize raw materials in demanding industrial environments.

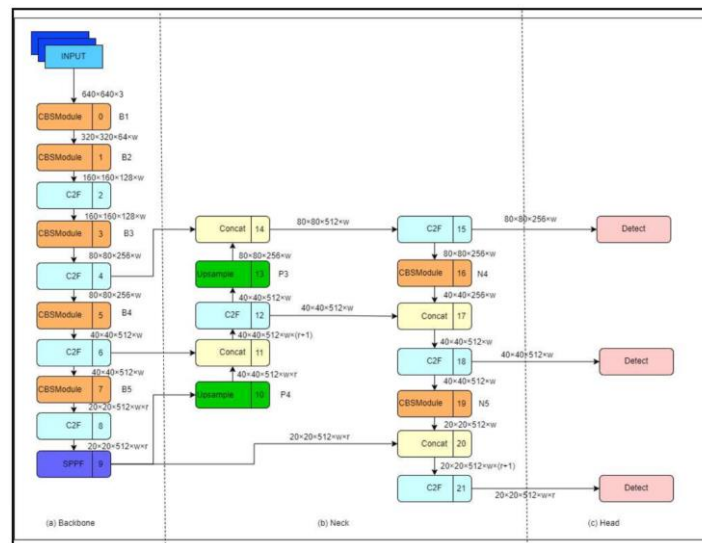


Figure 6: YOLO-v8 architecture (a) Backbone (b) Neck (c) Head (image from [37])

Segment Anything Model 2 (SAM 2) represents the current state of the art in foundation models for visual segmentation across both images and videos. It is a major evolution of the original Segment Anything Model (SAM), extending the concept of promptable, zero-shot segmentation from static images to full spatio-temporal video understanding [64]. SAM 2 is designed to solve the Promptable Visual Segmentation (PVS) task, which generalises image segmentation to video. In this setting, a user (or an upstream system such as an object detector) provides prompts such as points, bounding boxes, or masks on any frame, and the model predicts consistent segmentation masks (masklets) for the target object across all video frames. This formulation allows SAM 2 to handle complex visual phenomena such as occlusions, object motion, appearance changes, and re-identification over time.

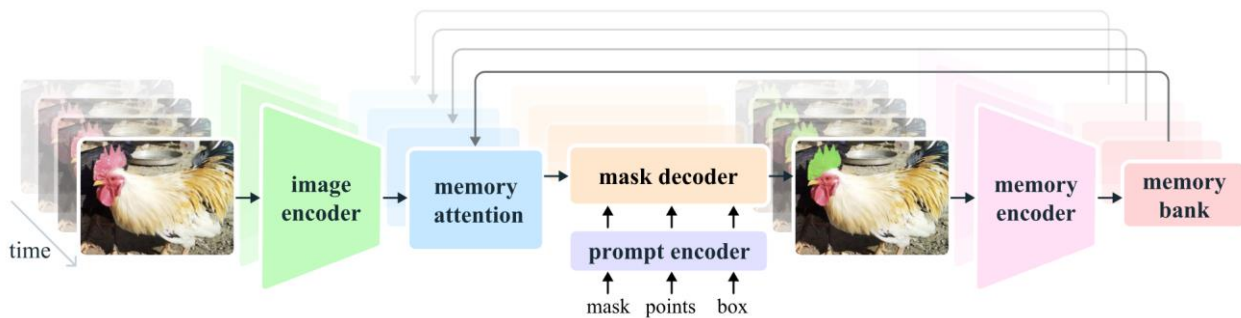


Figure 7: SAM2 architecture. For each frame, the segmentation is predicted using the current input and information stored from previous frames. Frames are processed one at a time by the image encoder and combined with past object memories through memory attention. The mask decoder then produces the segmentation for the current frame, and a memory encoder updates the stored representations for future frames. (image from[64])

Architecturally, as visualized in Figure 7, SAM 2 extends the original SAM design by introducing a memory-augmented transformer that enables temporal reasoning. Each video frame is first encoded by an image encoder into feature embeddings. These embeddings are then conditioned on a memory bank that stores spatial features and object pointers from previous frames. A memory attention module cross-attends to these stored representations, allowing the model to track objects over time and recover from occlusions. A lightweight prompt-aware mask decoder then produces segmentation masks for the current frame. A key innovation in SAM 2 is the streaming memory design, which allows it to process arbitrarily long videos in real time. Rather than recomputing all frames jointly, the model consumes frames sequentially, storing only a fixed number of recent memory representations. This enables efficient long-term tracking without sacrificing segmentation quality. In addition, SAM 2 includes an occlusion prediction head, allowing it to explicitly model whether an object is currently visible, which is critical in industrial environments where materials may be temporarily hidden by other objects or machinery.

SAM 2 is trained on the SA-V dataset, a very large video segmentation dataset collected using a human-in-the-loop data engine. The dataset contains over 35.5 million segmentation masks across 50,900 videos, more than an order of magnitude larger than any previous video segmentation dataset. Importantly, SA-V is not limited to predefined object categories: it includes arbitrary objects, parts, and materials, which makes SAM 2 particularly suitable for open-world industrial perception tasks such as scrap detection, slag and dross segmentation, and material flow monitoring. In benchmarks, SAM 2 significantly outperforms previous video segmentation systems, while requiring three times fewer user interactions. It also improves over the original SAM in image segmentation accuracy while being up to six times faster, due to its more efficient hierarchical image encoder and optimized attention mechanisms.

For ENCIRCLE's use cases, SAM 2 is particularly relevant because it provides high-quality, category-agnostic segmentation of complex industrial scenes. This enables robust extraction of scrap regions from cluttered wagon images and furnace camera streams, even when scrap piles overlap, deform, or become partially occluded. The use of SAM 2 therefore provides a state-of-the-art foundation for generating the precise spatial masks required for reliable downstream embedding extraction and energy-demand prediction.

A major challenge in modern industrial quality control is the lack of extensive labeled datasets for every possible defect. To address this, recent research has focused on Few-Shot Defect Segmentation (FDS), which aims to train models using only a handful of examples. A pivotal study by Liu et al. [87] provides a comprehensive analysis of this domain. The authors introduced a new, realistic benchmark for general industrial scenarios and compared traditional 'meta-learning' algorithms against modern Vision Foundation Models (VFMs). Their detailed experiments revealed a crucial finding: while traditional methods struggle to adapt to complex and diverse industrial textures, VFMs demonstrate superior generalization capabilities. This is due to the robust, pre-trained feature space of Foundation Models, which allows them to understand and

segment novel defects without requiring massive retraining. Building on these insights, the ENCIRCLE project adopts a VFM-based strategy for the specific challenge of dross estimation in the Alumil pilot. Since dross forms irregular, amorphous patterns on the molten surface a 'non-trivial' texture that is difficult to label in large quantities standard segmentation methods often fail. By leveraging the strong baseline established by Liu et al., our soft sensor utilizes the deep visual representations of a Foundation Model to accurately segment dross regions. This ensures high-precision process monitoring and efficient skimming, even under conditions of extreme data scarcity.

2.2.2 Illumination Control for Enhanced Image Quality in Vision-Based Industrial Inspection

Image quality is a fundamental prerequisite for reliable vision-based quality control and high-precision annotation in industrial environments. In practice, image degradation arises both from improper camera exposure settings and from suboptimal illumination configurations, particularly under dynamic and heterogeneous lighting conditions. As a result, prior work has addressed image quality improvement through two complementary directions: adaptive exposure control, which dynamically regulates camera parameters to stabilize brightness over time, and illumination control, which optimizes the type, intensity, and spatial distribution of light to enhance contrast and suppress reflections. Together, these approaches form the basis for robust visual data acquisition in industrial inspection systems, where consistent image quality is critical for the performance of downstream learning-based analysis.

Afifi et al. [31] address exposure correction as a fundamental image quality problem, targeting both underexposed and overexposed images that suffer from reduced contrast and loss of detail. Their work formulates exposure correction as a coarse-to-fine learning problem combining color and detail enhancement, supported by a large-scale dataset covering a wide range of exposure conditions, and demonstrates improved performance particularly for overexposed images. Wang et al. [39] proposed an automated exposure control framework, as illustrated in Figure 8, that utilizes a 'coarse-to-fine' strategy to rapidly adjust the camera's exposure time to an optimal level in real-time. This optimization is driven by a specialized image quality metric that weights gradient magnitudes according to photometric sensitivity, ensuring that the selected exposure maximizes structural information rather than merely balancing average brightness. Crucially, this approach incorporates a photometric calibration module to model the Camera Response Function (CRF) and vignetting effects. By leveraging this calibration, the system applies photometric compensation to eliminate brightness inconsistencies caused by exposure shifts and vignetting, ensuring robust and reliable visual data for downstream computer vision algorithms despite ambient fluctuations.

To establish a baseline of stability against dynamic illumination changes, we adopt the automated exposure control framework proposed by Wang et al. [39]. Their methodology, illustrated in Fig. 2, utilizes a 'coarse-to-fine' strategy to rapidly adjust the camera's exposure time to an optimal level in real-time. This optimization is driven by a specialized image quality metric that weights gradient magnitudes according to photometric sensitivity, ensuring that the selected exposure maximizes structural information rather than merely balancing average brightness. Crucially, this approach incorporates a photometric calibration module to model the Camera Response Function (CRF) and vignetting effects. By leveraging this calibration, the system applies photometric compensation to eliminate brightness inconsistencies caused by exposure shifts and vignetting, ensuring robust and reliable visual data for downstream computer vision algorithms despite ambient fluctuations.

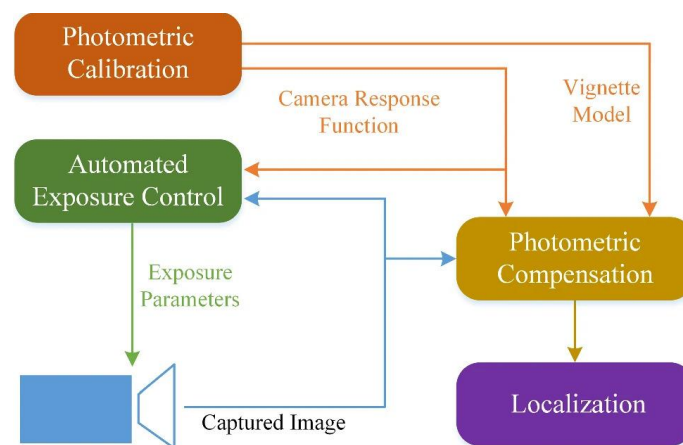


Figure 8: System Overview of the Automated Exposure Control Framework (image from 40)]

PROJECT 101178230 – ENCIRCLE

To address the issue of specular reflections characteristic of aluminium surfaces, an adaptive exposure control algorithm is implemented based on the principles established by Pan et al. [38]. The mechanism operates on a feedback loop comprising three stages:

1. **Histogram Analysis:** Real-time evaluation of the current frame's pixel intensity distribution to detect overexposed (saturated) regions.
2. **Gradient-based Optimization:** Calculation of an image quality metric that prioritizes edge visibility and texture details over mere brightness.
3. **Iterative Adjustment:** Automatic recalibration of the camera's shutter speed (exposure time) to maximize information entropy, ensuring that distinct scrap categories remain identifiable even under fluctuating industrial lighting conditions.

A recent advancement in this field is the work by Lee et al. [32], which introduces a Deep Reinforcement Learning (DRL) approach designed for industrial cameras. Unlike older methods, this system can adjust camera settings like exposure time and gain in just a few milliseconds. It is built to handle sudden lighting changes, ensuring the camera always captures clear images. For our task, this technology is highly beneficial because it ensures that the YOLO-based detection and the polygonal annotations are performed on high-quality data, as illustrated in Figure 9. By reducing motion blur and glare in real-time, it helps us achieve more stable and accurate scrap identification, even when factory lighting is unpredictable.

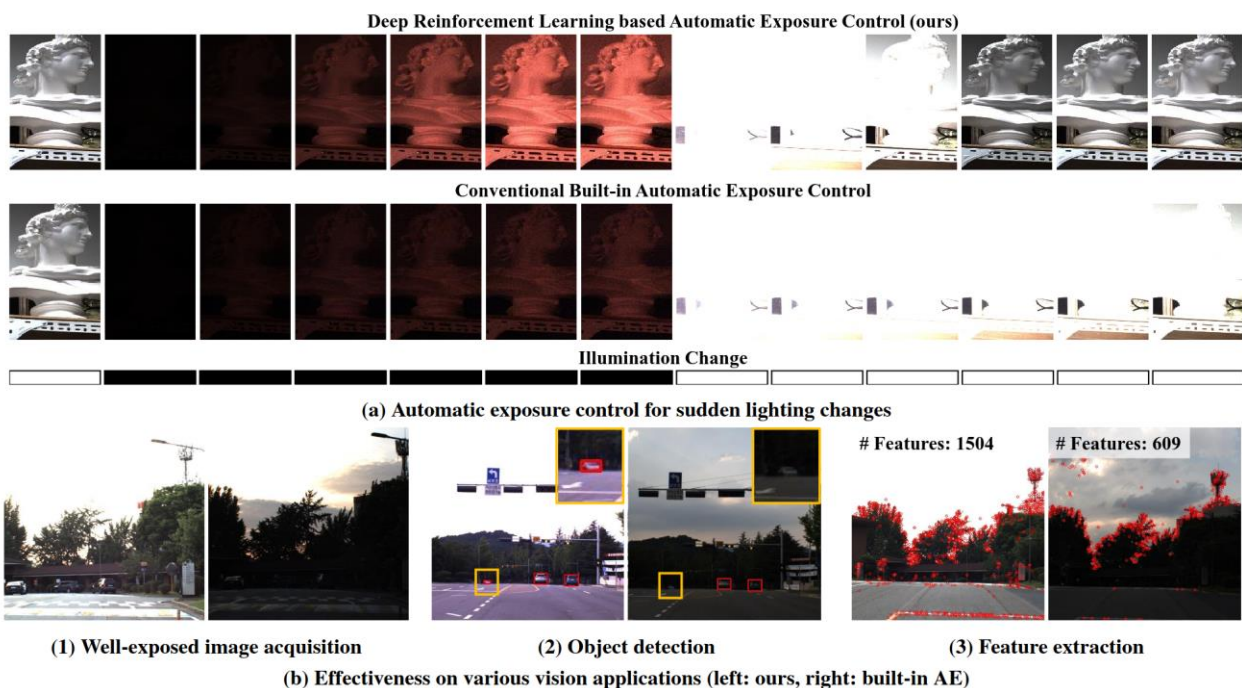


Figure 9: Comparison between conventional built-in auto-exposure and the DRL-based framework . The DRL-AE agent achieves rapid convergence under lighting changes and significantly enhances object detection and feature extraction performance. (image from [32])

While exposure control addresses temporal variations in brightness, it represents only one component of image quality optimization in vision-based quality control systems. In industrial inspection scenarios, particularly those involving metallic materials, illumination design and control play a decisive role in determining image contrast, reflection behaviour, and feature visibility. Reflective metals, including aluminium-based materials commonly encountered in recycling streams, exhibit strong specular responses that amplify sensitivity to illumination configuration. As a result, inspection performance is heavily influenced by illumination type, intensity, spatial distribution, and interaction with uncontrolled ambient light, motivating systematic investigation of illumination control strategies as a core element of reliable automated quality assessment.

The automation of quality control has been significantly advanced through developments in Deep Learning (DL), machine vision sensors, and illumination systems. In vision-based inspection and measurement setups, illumination constitutes a key factor that directly affects image quality and, consequently, the reliability of inspection outcomes. The selection or

optimization of illumination for machine vision applications typically relies on manual adjustments and on-site measurements to achieve stable and acceptable performance. In demanding industrial environments involving metallic materials, inspection systems are required to detect surface defects across varying geometries, scales, and reflective properties, rendering illumination control particularly critical. Reflective metals commonly encountered in recycling streams, including aluminium-based materials, exhibit strong specular behaviour that makes surface inspection highly sensitive to illumination conditions. This sensitivity complicates manual inspection, increasing both time requirements and dependence on experienced personnel. In such settings, illumination solutions must account for uncontrolled ambient lighting, which introduces shadows, specular highlights, and sensor saturation. They must further consider an a priori analysis of illumination technique, intensity, and spatial coverage based on material properties. As a result, illumination strategy selection becomes highly application-specific and often necessitates extensive market evaluation and pre-deployment testing, increasing system complexity, cost, and deployment time.

Deep Learning has substantially improved traditional machine learning approaches for automated quality inspection. Surface inspection tasks present challenges including heterogeneous defect characteristics, domain-specific constraints, limited labelled data, and real-time processing requirements. DL addresses these challenges [8][8] through its capacity to efficiently analyse high-dimensional and imbalanced datasets, to mitigate data scarcity via transfer learning [40] few-shot learning [41], and generative modeling approaches [42], and to support real-time inference and decision-making [43].

Reinforcement Learning (RL) is a computational paradigm for optimizing sequential decision-making processes under the Markov property, where future states depend solely on the current state and selected action. By leveraging DL for representation learning from high-dimensional sensory inputs, Deep Reinforcement Learning (DRL) has emerged as an active research area [44], with demonstrated success in robotics, smart manufacturing, game playing, and autonomous driving. Within industrial manufacturing contexts, DRL-based solutions for quality inspection remain comparatively underexplored relative to applications in scheduling, robotics, and process control [45],[46]. Moreover, RL-based methods that directly optimize physical illumination conditions during image acquisition, rather than relying on post-acquisition enhancement, are still largely absent from existing practice.

Existing studies and reported limitations can be broadly grouped into two thematic areas that are central to vision-based quality inspection in industrial environments. The first concerns Deep Learning (DL) approaches for surface inspection, with a focus on metallic materials, such as the ones in ALUMIL and ZINQ cases. The second concerns the use of Reinforcement Learning (RL) in smart manufacturing, particularly at the quality control stage.

Surface Inspection with Deep Learning

Deep Learning has been established as a dominant paradigm for surface inspection [40], owing to its effectiveness in image classification, object detection, and image segmentation tasks. In surface inspection, defect segmentation refers to pixel-level classification, where each pixel is assigned to a defect class or background. The effective extraction of multi-scale semantic features and the modeling of global contextual dependencies have been identified as key challenges in image segmentation. Architectural strategies such as skip connections [47], average pooling and atrous convolutions [48], as well as attention mechanisms [49], have been proposed to address these challenges. In [50], RSADUNet, a U-Net-based architecture employing multi-branch training and dense skip connections, is introduced for detecting small-scale defects on metal surfaces. Leveraging feature pyramid representations, Defect-Aux-Net is presented in [51], as a multi-task architecture augmented with Squeeze-and-Excitation and Spatial Attention mechanisms, jointly exploiting information from segmentation masks, object-level detection, and image-level classification. In a further study [52], the SPEED network is proposed as a real-time segmentation framework for metal surface defects, combining a Semantic Prior branch with an Extremely Efficient Dilated branch to address intra-class variability and inter-class similarity.

Despite notable progress in defect segmentation architectures, illumination conditions during image acquisition remain insufficiently addressed. Uncontrolled illumination on metallic surfaces can introduce reflections whose spatial patterns closely resemble actual defects, leading to ambiguous visual cues. Such effects can induce misleading activations in neural networks and result in systematic false positive detections [53]. To mitigate these issues, [54] introduces an image quality assessment metric to guide illumination adjustment and training data selection, demonstrating improved defect recognition on metal surfaces. However, the approach is limited to specific defect categories and relies on strong assumptions regarding background–foreground contrast characteristics. In [55], a deflectometry-based illumination strategy is proposed within the DeepInspection framework for specular surface inspection in automotive applications. While effective under controlled conditions, the complexity of the required setup and calibration, along with limited adaptability to variations in geometry and surface appearance, constrain its applicability in heterogeneous industrial environments such as metal recycling.

Reinforcement Learning in Smart Manufacturing

A recent review on Deep Reinforcement Learning in smart manufacturing [45] indicates that the majority of reported applications focus on scheduling, process control, and robotics, with comparatively few contributions addressing quality inspection. In [56], an RL-based path planning strategy employing a Deep Attention Actor-Critic architecture with CNN and Long Short-Term Memory (LSTM) components is proposed to enhance inspection performance in wafer probing. In [57], a

Model-driven Adaptive Proximal Policy Optimization (MAPPO) method is introduced for bolt posture adjustment, incorporating adaptive reward mechanisms and a probability tree to improve convergence and accuracy relative to standard PPO [58]. Illumination-related optimization is explored in [59], where a Deep Deterministic Policy Gradient (DDPG) framework is used to jointly adjust illumination and camera parameters for improved defect detection in CMOS and CCD glass lid inspection. In another approach [60], a robot-assisted inspection system is proposed that sequentially acquires images around an object, using Iterative Learning Control and Reward-Weighted Policy Learning to optimize motion planning based on objective image quality criteria.

While these RL-based methods address selected aspects of quality control, illumination control for real-world metallic surface inspection remains comparatively underexplored. This gap is particularly evident in scenarios involving heterogeneous materials, variable geometries, and reflective properties, such as those encountered in aluminium recycling. Experimental investigations building on existing DRL-based visual systems [61], will be presented in 4.1.2. These investigations aim to examine illumination as a controllable component of the sensing process, linking adaptive illumination selection with DL-based inspection to address the variability and complexity of industrial metal inspection environments.

2.2.3 Predictive Analytics for Aluminium Melting Processes

The DINOv2 framework is a self-supervised visual representation learning method based on a Vision Transformer (ViT) architecture and a student–teacher training scheme. The model is trained without explicit labels and learns to produce semantically meaningful visual representations by enforcing consistency between different augmented views of the same image. As a result, DINOv2 provides strong, task-agnostic image embeddings that can be directly used for downstream tasks without fine-tuning the backbone network. The pretrained model outputs high-dimensional global image embeddings, typically derived from the ViT class token, which capture rich semantic information about the input image. Owing to its foundation-model nature, DINOv2 has demonstrated strong generalization across a wide range of visual domains, making it well suited for feature extraction in application-specific pipelines [69].

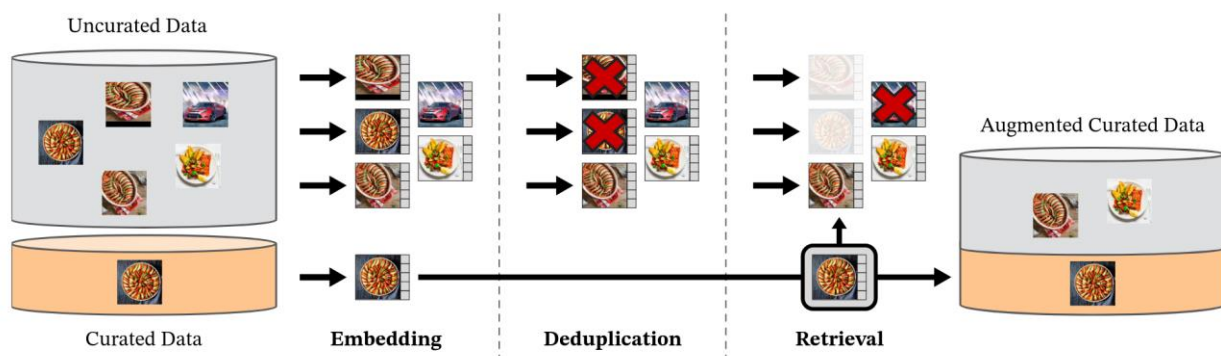


Figure 10: Background on DINOv2 training data construction (image from [69])

A significant contribution to the field of industrial safety and computer vision is the DINOSAFE methodology (2024) [72], which addresses the challenges of detecting hazardous situations in production facilities. To overcome the limitations of traditional deep learning—specifically the need for massive labeled datasets and high computational power—DINOSAFE proposes a hybrid approach combining dense optical flow with the DINOv2 foundation model. While optical flow is utilized to estimate the motion of objects in video streams, DINOv2 provides high-dimensional, self-supervised visual representations that capture complex object properties without requiring manual annotation. By training lightweight linear classifiers on these features, the system can identify and categorize risky scenarios, such as pedestrian-forklift interactions or falling loads, in real-time. Experimental results on real-world industrial datasets, including a publicly released set of 6,500 images, demonstrate that this self-supervised feature extraction significantly reduces the manual labeling effort while maintaining high detection accuracy and computational efficiency. Inspired by recent methodologies like DINOSAFE, which combine DINOv2 with motion analysis to reduce labeling efforts, the ENCIRCLE project utilizes DINOv2's self-supervised features to categorize scrap material.

Despite the success of frameworks like DINOv2 in industrial safety, a significant research gap remains regarding the use of self-supervised embeddings for predicting continuous process parameters in heavy industry. Most existing literature focuses on trivial classification tasks or well-defined object detection. The ENCIRCLE project addresses these limitations by introducing a novel vision-based soft sensor. While previous works use DINOv2 for identification, we leverage its task-agnostic embeddings to predict dross formation values and material composition in non-trivial environments. A major challenge we address is that prediction becomes significantly more complex when different categories of scrap are piled on

top of each other, leading to high levels of occlusion. Our approach is specifically designed to handle these accumulated materials and the extreme specular reflections of molten aluminium. The ENCIRCLE framework connects vision technology with industrial control by turning visual data into process insights, leading to improved energy efficiency and material recovery.

2.3 DIFFUSION-BASED IMAGE GENERATION FOR INDUSTRIAL FURNACE MONITORING

CLIP (Contrastive Language–Image Pre-training) is a dual-encoder model that learns a shared embedding space for images and text [14][14]. Instead of predicting a fixed set of labels with a classifier head, it trains an image encoder and a text encoder together so that the embedding of an image is close to the embedding of its matching caption and far from the embeddings of non-matching captions (and symmetrically, each caption is closest to its paired image). This contrastive pairing objective is what makes CLIP broadly transferable: once trained, the model can compare any image to any text string by measuring similarity in that shared space. At inference time, this enables “zero-shot” recognition by turning class names (or richer class descriptions) into text embeddings and treating those embeddings as an implicit linear classifier, with the predicted class being the one whose text embedding is most similar to the image embedding.

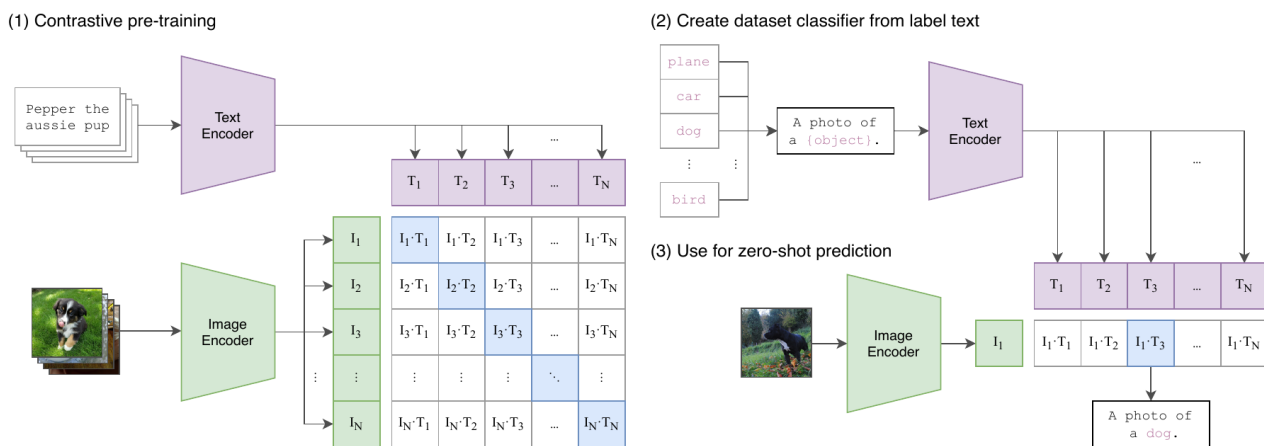


Figure 11: CLIP trains an image encoder and a text encoder to match paired image–text examples via a contrastive objective. At inference, class names or descriptions are embedded as text prototypes and compared to an image embedding to enable zero-shot classification. (image from REF_Ref6314 \n \n [14][14])

OpenCLIP is an open, community-driven implementation and scaling effort around the same core idea, jointly training an image encoder and a text encoder with a contrastive image–text pairing objective, but with an emphasis on reproducible training recipes, a wide selection of pretrained checkpoints, and flexible architectural choices (multiple ViT sizes, variants, and training datasets) [65]. In diffusion systems, including SDXL, CLIP-style text encoders matter because they provide the conditioning representation that steers generation: the prompt is encoded into a dense semantic representation that the denoiser can attend to.

Stable Diffusion 1.5 (SD 1.5) is among the most widely adopted open text-to-image generative models and remains a strong baseline for latent diffusion–based generation. SD 1.5 follows the Latent Diffusion Model (LDM) paradigm [88]: rather than performing denoising in pixel space, images are first compressed with a variational autoencoder (VAE) into a lower-dimensional latent representation, and diffusion is learned in that latent space, as illustrated in Figure 12. This design reduces compute and memory requirements substantially while preserving perceptual quality, making both fine-tuning and inference practical on moderate GPU hardware.

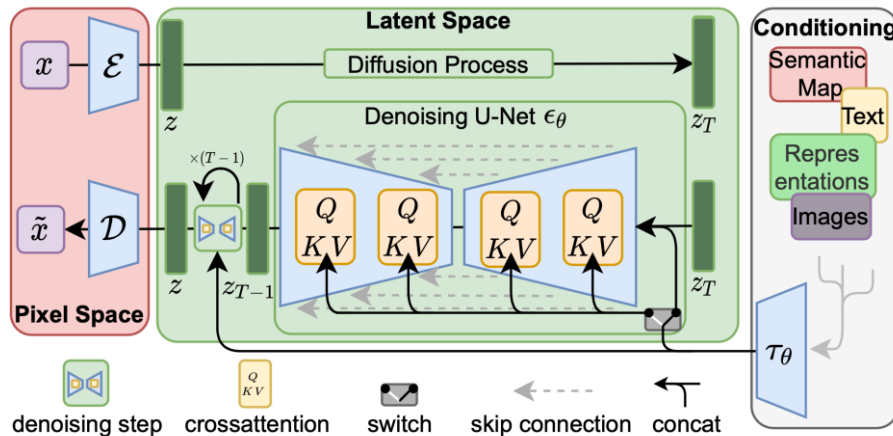


Figure 12: LDM architecture. (image from [88])

Text conditioning in SD 1.5 is typically provided by a frozen CLIP ViT-L/14 text encoder [14][14], where token embeddings condition the denoising UNet through cross-attention layers. During sampling, SD 1.5 commonly uses classifier-free guidance (CFG) [89] to trade off prompt adherence and diversity; correspondingly, training includes dropping text conditioning with some probability so that guided and unguided predictions can be combined at inference time. SD 1.5 is also particularly valuable in applied settings because it has a mature ecosystem (Diffusers pipelines, LoRA tooling, ControlNet extensions) and enables efficient domain adaptation via parameter-efficient fine-tuning methods such as LoRA, making it a natural predecessor to SDXL in modern diffusion pipelines.

Stable Diffusion XL (SDXL) is a Latent Diffusion Model (LDM) that improves over earlier Stable Diffusion (SD 1 and SD 2) versions primarily through (i) a substantially larger denoising backbone and (ii) stronger, richer text conditioning Figure 13. Architecturally, SDXL scales the UNet by redistributing transformer computation within the UNet and expanding the cross-attention context. A key design choice is the use of two text encoders, CLIP ViT-L and OpenCLIP ViT-G [65], whose penultimate embeddings are concatenated for conditioning. SDXL also conditions on a pooled embedding derived from OpenCLIP [65], in addition to standard cross-attention conditioning.

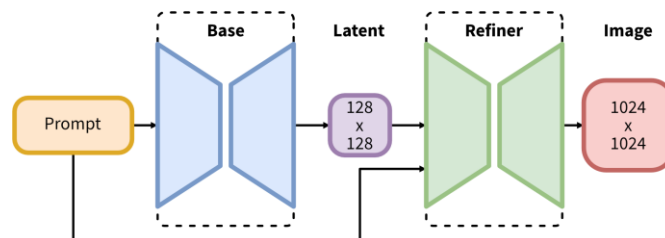


Figure 13: A text prompt is encoded and passed to the SDXL base model, which generates an initial latent representation. The latent is then optionally forwarded to a refiner model that performs the final denoising steps to produce the final image. (image from [66])

A common way to use SDXL is as an “ensemble-of-experts” two-model pipeline: (1) the base model generates (noisy) latents, and (2) a dedicated refiner model performs the final denoising steps, specializing in improving local detail and perceptual quality. Importantly, the base model can also be used standalone (without the refiner) when you prefer a simpler or faster pipeline. An alternative SDXL workflow is a two-stage generation strategy that leverages a high-resolution step via SDEdit [67]. In this setup, the base model first generates latents at the desired output size. Then, a second specialized high-resolution process applies SDEdit-style latent refinement using the same prompt, effectively “editing”/denoising latents toward a sharper result. This option is typically slower than using the refiner directly because it requires more denoising function evaluations, but it can provide additional control and quality gains depending on the use case.

FLUX.1 represents a newer generation of transformer-based latent generative models, replacing the classic UNet denoiser with a large Diffusion Transformer (DiT-style) backbone trained using rectified flow / flow matching objectives [90]. FLUX.1 is a rectified flow transformer trained in the latent space of an autoencoder, following the same high-level efficiency motivation as latent diffusion, but modernizing the denoiser architecture toward transformer blocks with improved scaling behaviour. The model uses a combination of double-stream and single-stream blocks (as illustrated in Figure 14), where

image and text tokens are processed separately early on (double-stream), then concatenated into a unified sequence processed by deeper transformer layers (single-stream), before decoding back to pixels via the latent autoencoder.

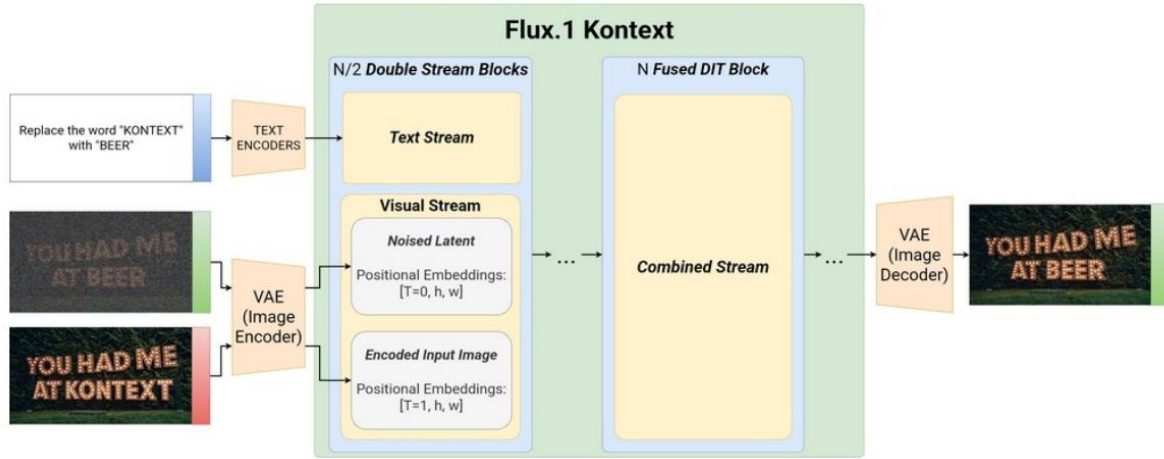


Figure 14: High-level overview of FLUX.1. (image from [90])

FLUX.1 comprises a unified generation-and-editing system by conditioning on both text instructions and optional context images, enabling (i) standard text-to-image generation and (ii) instruction-based editing under the same architecture. A key design element is token-sequence concatenation, where context-image latent tokens are appended alongside target tokens, allowing flexible resolutions/aspect ratios and supporting iterative workflows without specialized editing heads. The paper emphasizes improved multi-turn stability (less “character drift”) and interactive runtimes (reported as a few seconds at 1024×1024), positioning FLUX.1 as a practical SOTA direction for controllable, iterative visual generation.

Low Rank Adaptation (LoRA) is a parameter-efficient fine-tuning method designed to adapt large pretrained models without updating (and without storing optimizer state for) the full weight set [68]. The core assumption is that task adaptation does not require a full-rank update to dense weight matrices. Instead, for a pretrained weight matrix $W_0 \in \mathbb{R}^{d \times k}$, LoRA constrains the update ΔW to be low-rank by parameterizing it as $\Delta W = BA$, with $B \in \mathbb{R}^{d \times r}$, $A \in \mathbb{R}^{r \times k}$, where $r \ll \min(d, k)$. The forward pass becomes:

$$h = W_0 x + \Delta W x = W_0 x + B A x$$

while W_0 remains frozen and only A, B are trained. This yields a large reduction in trainable parameters and optimizer memory while maintaining strong adaptation capacity.

Two practical properties make LoRA especially attractive for “foundation-model adaptation” workflows. First, LoRA modules can be swapped per task while keeping a single shared frozen base model, which reduces storage and task-switching overhead. Second, LoRA can be deployed with no additional inference latency by merging the low-rank update into the base weight (i.e., using $W = W_0 + BA$ at inference time). Although introduced and evaluated primarily in the context of Transformer adaptation (often focusing on attention projection matrices), the same mechanism transfers directly to diffusion pipelines in practice (e.g., adapting SDXL’s attention blocks) because the dominant trainable components are still large linear layers where low-rank updates are effective.

2.4 AUDIO-BASED DEFECT DETECTION

Acoustic anomaly detection in industrial environments typically relies on Deep Neural Network architectures such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs/LSTMs). These approaches frequently employ time-frequency representations, like Fast Fourier Transform (FFT) and Spectrograms [73], to bridge the gap between raw audio and visual classifiers.

CNN-based audio classification

Convolutional Neural Networks (CNNs) excel at capturing local spectral features (e.g., sharp defective sounds), a capability effectively demonstrated by Zhu et al. [74] in their approach, in which raw acoustic signals converted into Short-Time Fourier Transform (STFT) spectrograms and depth-wise separable convolutions were applied for defect detection. This

PROJECT 101178230 – ENCIRCLE

architecture allows for the efficient extraction of high-frequency transient defect signatures on edge devices with significantly reduced parameters. In a similar approach [75], the WAV formatted files using the Mel-Frequency Cepstral Coefficients (MFCC) and then a CNN approach was used for downstream classification. However, standard CNNs often struggle to model long-range temporal dependencies efficiently, limiting their ability to characterize complex, time-evolving machinery sounds without the aid of sequential processing mechanisms.

RNN-based audio classification

RNNs/LSTMs, on the other hand, excel at handling time-series data by maintaining a memory of past inputs, a capability that has been effectively adapted for specific industrial machinery. For instance, a hybrid LSTM-Autoencoder architecture that fuses one-dimensional raw audio signals with Mel-spectrogram features to identify anomalies in compressor systems [77]. Similarly, other approaches have integrated Self-Organizing Maps (SOM) with LSTMs, utilizing sound conversion mechanisms to enhance fault diagnosis efficiency in industrial motors [78]. However, despite these algorithmic advancements, these architectures inherently suffer from sequential processing problems such as limited scalability and temporal attention bottlenecks, compared to modern attention-based models.

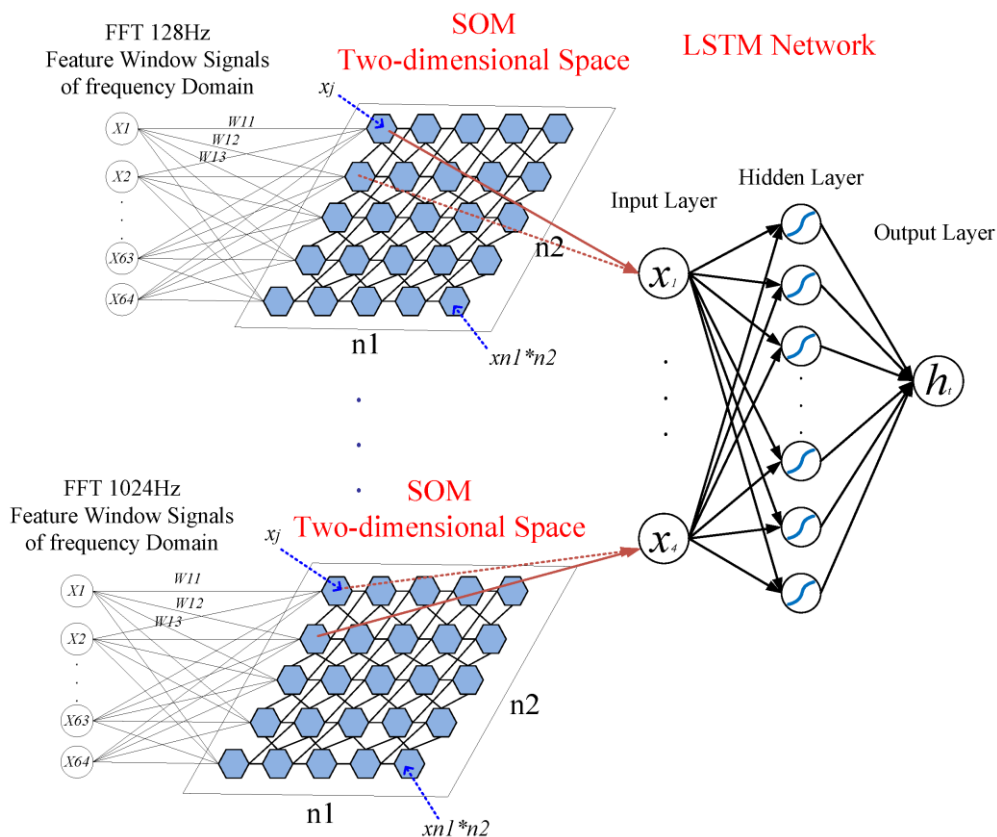


Figure 15: SOM-LSTM defect detection pipeline, which generates different FFT frequencies, it embeds them into two dimensional space and then it utilizes an LSTM for classification. (Image from [78])

Transformer-based audio classification

These problems were partially mitigated by the introduction of Transformer architecture [76]. Approaches like the Audio Spectrogram Transformer (AST) [79] treat the audio Mel-spectrogram as a vision task by segmenting the time-frequency map into a sequence of overlapping patches. To compensate for the lack of locality constraints and the scarcity of labeled industrial data, these architectures utilize learnable positional embeddings to retain temporal order and leverage transfer learning from large-scale image datasets. While CAT [80], introduces a Causal Audio Transformer (CAT) consisting of a Multi-Resolution Multi- Feature (MRMF) feature extraction with an acoustic attention block for more optimized audio modeling, instead of using a simple Visual Transformer. Even though pure Transformer models offer a powerful global context solution, they often lack the inductive bias required to extract fine-grained local patterns from audio spectrograms without massive datasets, which are rarely available in specific industrial pilot sites.

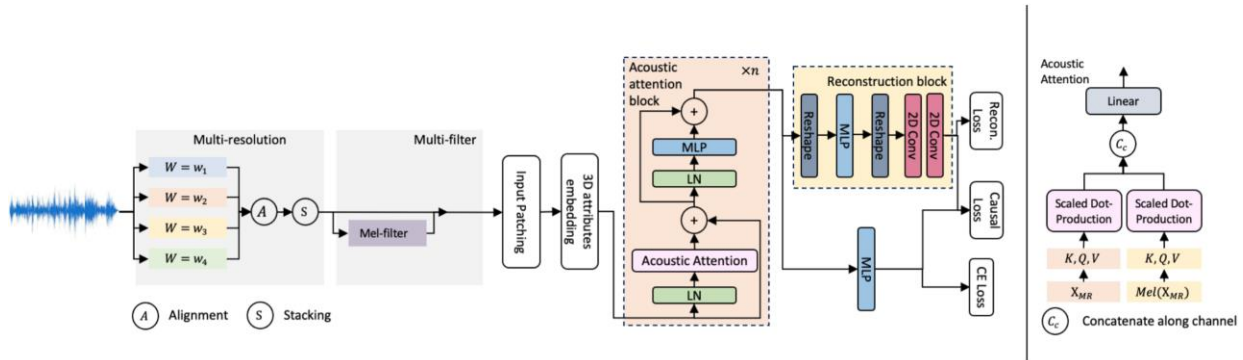


Figure 16: Causal Audio Transformer (CAT) and proposed acoustic attention module (Image from [80]).

To bridge these gaps, we implemented a Conformer-based audio classification pipeline, an architecture that is the state-of-the-art for audio recognition [81][82] tasks fusing the best properties of CNNs and Transformers. By integrating Convolution-augmented Transformers, our soft sensor simultaneously captures local acoustic features and global temporal patterns. This capability allows the model to detect subtle defect signatures that may be distributed across the audio clip, outperforming traditional CNN or LSTM baselines. However, despite its proven employment in audio event recognition, a significant gap exists in the literature regarding its application to industrial environments, where it has not typically been utilized for defect detection tasks.

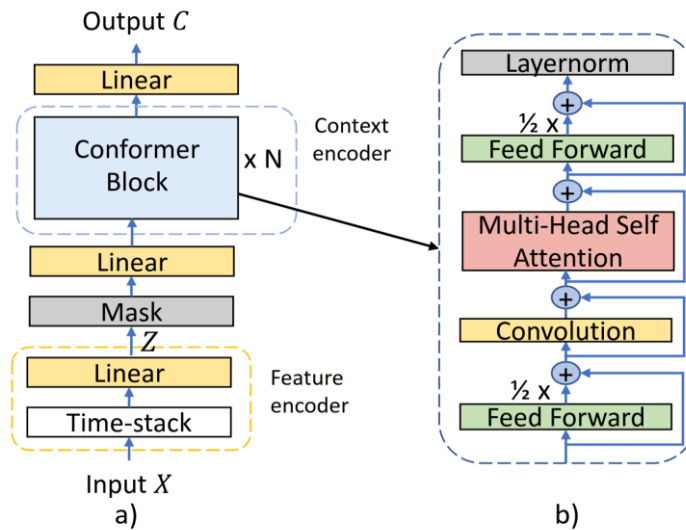


Figure 17: Conformer encoder architecture (Image from [81])

3 SENSORS NETWORK

The Sensors Network constitutes the physical data acquisition layer of the ENCIRCLE architecture. Its primary goal is to digitize shopfloor operations by connecting legacy technology and the project's high-level analytics engines. This network is designed as an intelligent edge ecosystem capable of ingesting heterogeneous data streams. Sensors can register telemetry and high-bandwidth image streams, which are then standardized in order to be used from the Soft Sensors for predictions, and other downstream ENCIRCLE modules.

3.1 ALUMIL PILOT SITE

This section details the sensor network architecture deployed for the ALUMIL (ALUM) pilot site. The objective of this layer is to establish a robust data acquisition infrastructure that bridges the gap between legacy sensors and the project's IIoT requirements. The network comprises a mix of existing industrial detectors, newly installed IIoT devices, and established camera hardware designed to support sustainable manufacturing and traceability.

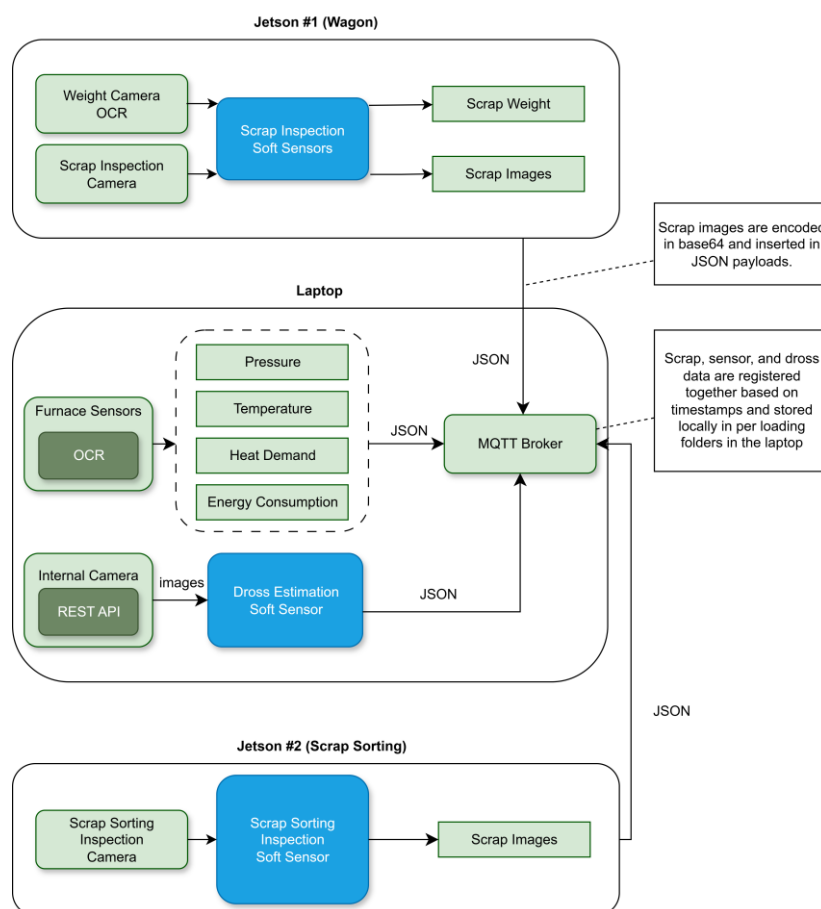


Figure 18: ALUMIL pilot site sensor overview.

The ALUM sensor network is engineered to capture critical process parameters across two primary production stages: the raw material intake (Scrap Sorting and Scrap Loading) and the scrap melting process. To achieve the granularity required, the physical infrastructure has been upgraded to support high-frequency data transmission and optical sensors. This modernization relies on a multi-tiered edge computing architecture, integrating NVIDIA Jetson nodes for local image processing with a central Local Aggregation Server that unifies diverse data streams via MQTT. All this infrastructure is combined within a local private network that utilizes a TP-LINK Archer MR600 router for Wi-Fi and ethernet connectivity (Figure 19). Furthermore, to overcome the connectivity limitations of legacy equipment, the network employs a non-invasive digitization strategy, by utilizing machine vision and Optical Character Recognition (OCR) to extract real-time telemetry from existing screens and scales without disrupting critical control loops.

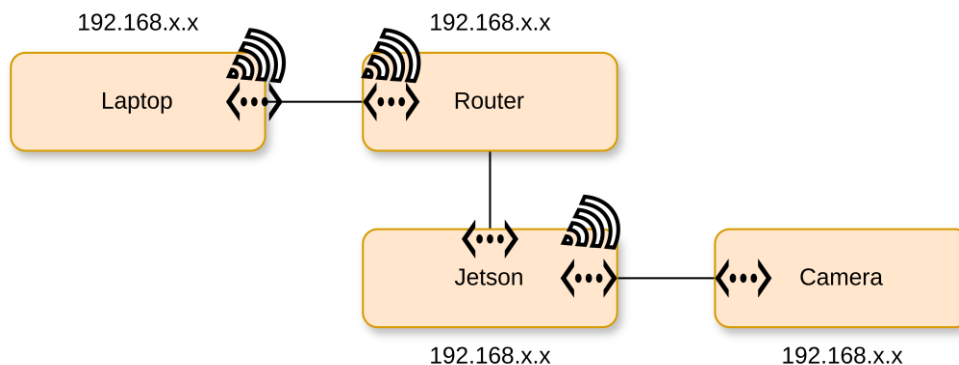


Figure 19: Local Network.

Table 1: TP-LINK Archer MR600 Specifications.

Protocols	IPv4/IPv6
Wireless Standards	IEEE 802.11a/n/ac 5 GHz // IEEE 802.11b/g/n 2.4 GHz
Frequency	Dual Band (2.4 & 5GHz)



Figure 20: TP-LINK Archer MR600.

3.1.1 Edge Computing and Data Flow

To efficiently manage the heterogeneous data streams generated by the shopfloor sensor network a multi-level architecture has been deployed. This infrastructure is designed to process data as close to the source as possible in order to minimize network latency and reduce bandwidth congestion. The architecture is organized into a two-level hierarchy: distributed Vision Edge Nodes responsible for real-time inference and signal digitization, and a central Local Aggregation Server that acts as the site's primary gateway, handling protocol unification, data storing, and synchronization.

As Vision Edge Nodes we integrate multiple NVIDIA Jetson AGX Orin instances, connected to each machine vision camera. These units handle direct connection to the machine vision cameras and perform the initial heavy lifting, such as image buffering. This ensures that only extracted data (values) or pre-processed images are sent upstream, reducing network latency.

Table 2: NVIDIA Jetson AGX Orin Specifications.

Size	110mm x 110mm x 71.65mm
Energy Consumption	15-60W
GPU	2048-core NVIDIA Ampere architecture GPU with 64 Tensor Cores

PROJECT 101178230 – ENCIRCLE

Computational Capacity	275 TOPS
CPU	12-core Arm® Cortex®-A78AE v8.2 64-bit CPU 3MB L2 + 6MB L3
CSI Camera	16-lane MIPI CSI-2 connector
Storage	64GB eMMC 5.1



Figure 21 : NVIDIA Jetson AGX Orin.

A dedicated on-premises computational edge node (Table 3), functions as the central Local Aggregation Server. This HP OMEN laptop acts as the primary gateway between the shopfloor edge nodes and the remote ENCIRCLE server infrastructure. Beyond simple connectivity, it hosts the custom middleware solution described in Figure 20, designed to unify the diverse data protocols present on the shop floor. Crucially, this local server operates as the site's MQTT broker, managing the message exchange between devices. It is also responsible for the temporary storage of data to buffer streams during network interruptions, the scheduling of data transmission to other components in order to optimize bandwidth usage, and on-the-fly pre-processing to ensure data quality before upload.

Table 3: HP OMEN Specifications.

CPU	Intel Core i7-1075H 2.6 GHz
Est. Energy Consumption	100 W
GPU	NVIDIA GeForce RTX 3060 Mobile / Max-Q
Storage Memory	500GB NVMe M.2 SSD
RAM	16 GB
Connectivity	Ethernet, Wi-Fi

The middleware, is constructed on a modular .NET backend that functions as a robust protocol abstraction layer, designed to unify the diverse connectivity landscape of the shop floor. It aggregates data via MQTT, which is primarily utilized by the Jetson nodes and newer IIoT devices. Acting as a comprehensive bridge, the middleware collects both the processed OCR data from the edge nodes and raw telemetry, subsequently forwarding this consolidated information to upper-level systems, such as the Environmental Impact Analytics Engine, via a standardized REST API or secure MQTT forwarding. This architecture ensures that the Sensors Network presents a single data interface to the rest of the ENCIRCLE platform, effectively abstracting local hardware complexities.

3.1.2 Scrap Sorting

The sorting phase represents the first critical control point for ensuring the purity of the recycled aluminium inputs. Traditionally, this stage relies heavily on manual inspection, which leaves the process vulnerable to undetected contaminants that can degrade the final alloy quality. To transition towards a fully traceable and automated quality control system, the ENCIRCLE sensor network introduces a continuous visual inspection layer. By deploying high-resolution machine vision

PROJECT 101178230 – ENCIRCLE

directly over the sorting conveyor belt, the system moves from reactive sampling to proactive, real-time characterization of the entire material stream, enabling the early detection of impurities before they reach the furnace.

A high-resolution FLIR machine vision camera (Figure 22 and Table 4) is mounted directly over the scrap sorting conveyor belt, along with a Fujinon CF12ZA-1S lens (Figure 23). Equipped with special illumination to handle ambient lighting changes, this sensor captures high-frequency images of the moving scrap stream, serving as the data source for contamination detection models. The camera is then connected with the NVIDIA Jetson Xavier (Figure 24 and Table 5) in order to upload and pre-process the scrap sorting photos, before the soft-sensor will predict the impurities.

Table 4: FLIR Blackfly S Specifications.

Sensor	Sony IMX304 (Color)
Sensor Type	1.1" CMOS
Shutter Type	Global Shutter
Pixel Size	3.45x3.45 μm
Resolution	4096 x 3000 px
FPS	10
Data Interface	Gigabit Ethernet, Transfer rate 1000 Mbits/sec
Power Supply	12 V 1.5 A, 6-pin Hirose or Power over Ethernet (PoE)



Figure 22: FLIR Blackfly S.

Table 5: Fujinon Lens CF12ZA-1S Specifications.

Focal Length	12 mm
Iris	F1.8-F16
Iris Type	Manual
Pixel Pitch	2.74 μm
Min. Working Distance	100 mm
Wavelength Range	Visible



Figure 23: Fujinon Lens CF12ZA-1S.

Table 6: NVIDIA Jetson AGX Xavier Specifications.

Size	7 mm x 100 mm
Energy Consumption	10-30W
GPU	512-Core Volta GPU with Tensor Cores
Computational Capacity	32 TOPS
CPU	8-Core ARM v8.2 64-Bit CPU, 8 MB L2 + 4 MB L3
CSI Camera	16-lane MIPI CSI-2 connector
Storage	32GB eMMC 5.1


Figure 24: NVIDIA Jetson AGX Xavier.

Figure 25: Camera capture above the scrap sorting conveyor belt.

3.1.3 Scrap Loading

The precise description of scrap inputs is critical for aluminium recycling. Operators typically load the furnace with heterogeneous batches containing a mix of "clean" (pre-consumer) and "contaminated" (post-consumer) scrap, resulting in an approximate, non-uniform weight ratio in each loading wagon. To accurately map this variability, a Baumer machine vision camera (Figure 28 and Table 8), along with a designated lens (Figure 23 and Table 5), casing and floodlight (Figure 26 and Table 7), was deployed above the scrap loading wagon area (Figure 27). In order to safely attach the camera above the wagon (Figure 27), we've designed and 3D-printed (Figure 29) a casing that encloses and protects the camera from harsh industrial conditions.

Table 7: Illumination Specifications

Type	LED Floodlight 500W Cold White 5000K
-------------	---

PROJECT 101178230 – ENCIRCLE

Rated Power	500W
Supply Voltage	100-240V
Color Temperature	5000K
Lumen Output	60000lm
LED Type	SMD3030
Ingress Protection Rating	IP65



Figure 26: LED Floodlight 500W Cold White 5000K

Following the initial installation and testing phase, several operational challenges related to image exposure were identified, primarily due to fluctuating ambient lighting throughout the day, which resulted in inconsistent image quality. To mitigate this issue, a dedicated illumination device (Figure 26) was integrated into the infrastructure to standardize lighting conditions. Additionally, after identifying multiple non-optimal images in the dataset (Figure 30), including blurry images, overly dark or bright exposures, high-contrast scenes, and frames failing to capture the entire wagon, the camera's physical position was readjusted to optimize the viewing angle and improve overall image consistency. Complementing these hardware adjustments, a soft sensor was developed to dynamically regulate the camera's brightness and contrast based on ambient lighting conditions, enabling the generation of higher-quality datasets for training the soft sensor models

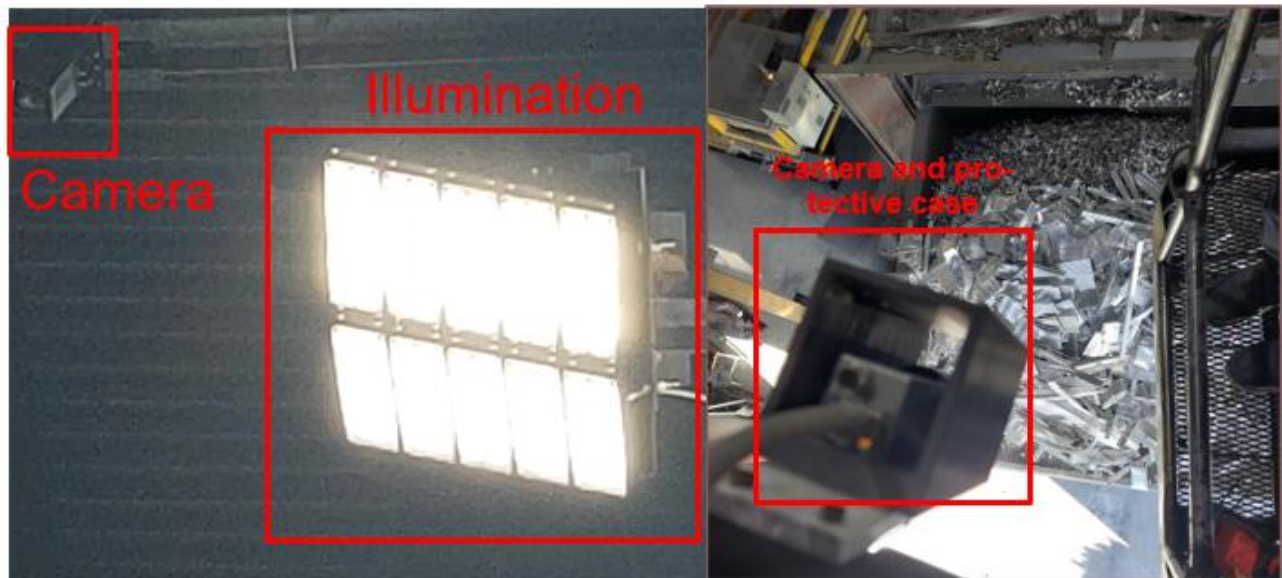
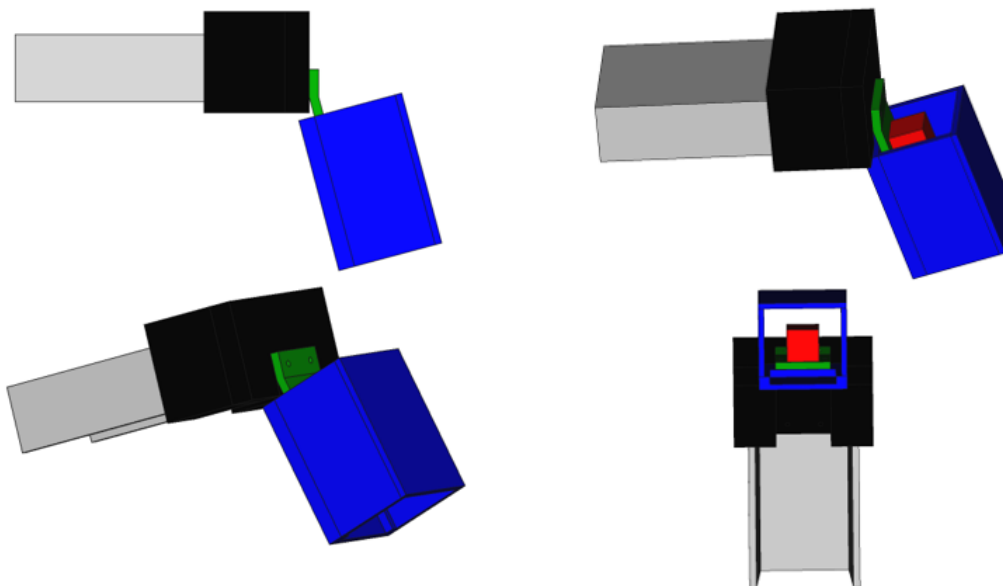


Figure 27: Scrap loading camera with casing and illumination device.

Table 8: Baumer VCXG.2-241C Specifications.

Sensor	Sony IMX540 Gen4 (Color)
Sensor Type	1.2" CMOS
Shutter Type	Global Shutter
Pixel Size	2.74x2.74 μm
Resolution	5312 $\times$ 4592 px
FPS	5
Data Interface	Gigabit Ethernet, Transfer rate 1000 Mbits/sec
Power Supply	M8/8 pins or Power over Ethernet (PoE)


Figure 28: Baumer VCXG.2-241C.

Figure 29: 3D snapshots from camera casing design.

Furthermore, the camera's physical position was readjusted after identifying several non-optimal images in the dataset (Figure 30) such as blurry, overly dark or bright, high-contrast, or images that fail to capture the entire wagon. These adjustments were necessary to improve the viewing angle and overall image consistency. In addition, a soft sensor was developed to dynamically adapt the camera's brightness and contrast based on ambient lighting conditions, thereby enabling the generation of higher-quality datasets for training the soft sensor models.



Figure 30: Initial non-optimal scrap loading captures.

As a result of these corrective actions, the previously identified imaging issues were effectively resolved. The roof-mounted wagon camera now consistently captures the scrap loading process throughout production, providing a stable field of view of the entire wagon (Figure 31). The improved camera placement, combined with the adaptive brightness and contrast control, ensures reliable visual coverage under varying operational and lighting conditions. This enables the continuous acquisition of high-quality image data that accurately represents the loading process, in order to support robust model training and soft sensor deployment.



Figure 31: Initial non-optimal scrap loading captures.

Parallel to this visual monitoring, the system addresses the challenge of recording the mass of these batches. The facility utilizes industrial scales for the wagons, but these legacy devices feature standalone digital displays with no network connectivity. To digitize this without invasive hardware modifications, a secondary machine vision camera has been installed to face the scale's digital readout directly. This sensor captures the numerical display (see Figure 25) in real-time, effectively bypassing the need for a physical interface with the scale's internal controller. Complementing this, the system also integrates historical production data concerning the weight distribution and the total gross weight of past scrap loadings (Figure 33), as well as the chemical composition of aluminium for each casting (Figure 34). This legacy dataset provides a baseline for the expected mass ranges per scrap type, which is then utilized to validate the live camera captures and refine the accuracy of the soft sensor models.



Figure 32: Weight sensor output on LED display.

 Alumil ΧΥΤΗΡΙΟ			
	Type	Weight	Datetime
A/A	ΠΟΡΟΣ	ΚΙΛΑ	ΗΜ. ΦΟΡΤΩΣΗΣ
1	T-BARS	5200	01/05/2025 12:41
2	Δέματα	1180	01/05/2025 12:41
3	Επιστροφές Alumil Αβαφο	1850	01/05/2025 12:51
4	SCRAP Διαλογής	1500	01/05/2025 12:51
5	Scrap Loading		Weight
6	SCRAP Διαλογής	1430	01/05/2025 13:22
7	Επιστροφές Alumil Αβαφο	1200	01/05/2025 14:05
8	SCRAP Διαλογής	1800	01/05/2025 14:05
10	T-BARS	2076	01/05/2025 14:07
11	Πλάκες METAM	1125	01/05/2025 14:07
9	SCRAP Διαλογής	1600	01/05/2025 14:35
12	Επιστροφές Alumil Αβαφο	1200	01/05/2025 14:35
13	Επιστροφές Alumil Αβαφο	1300	01/05/2025 15:07
14	SCRAP Διαλογής	1600	01/05/2025 15:07
15	Επιστροφές Alumil Αβαφο	1500	01/05/2025 15:35
16	SCRAP Διαλογής	1500	01/05/2025 15:35

Figure 33: Production data for oven loads.

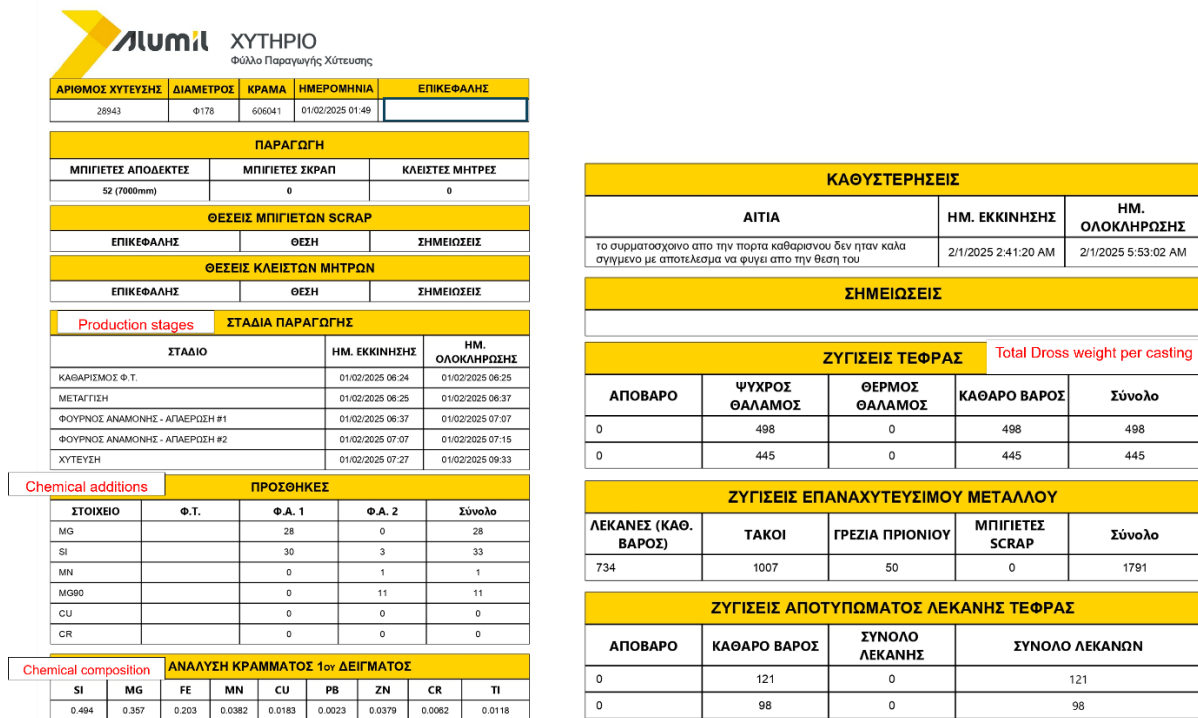


Figure 34: Production data for castings.

By combining the overhead visual feed with the digitized weight signal the system's soft sensors can classify the composition of each scrap batch with high fidelity, and by correlating the visual evidence of the scrap type (clean vs. contaminated) with the precise weight logged for that specific timestamp, the system can then predict the specific mass contribution of each scrap category within a single load.

3.1.4 Process Monitoring

The monitoring of the melting furnace is critical for optimizing energy consumption and ensuring high quality aluminium output. The furnace's conditions are managed via legacy industrial sensors integrated with Programmable Logic Controllers (PLCs) and visualized through a SCADA system (see Figure 35). Accessing these data streams directly via the PLC backplane is often prohibitively costly, requiring specialized intervention and requiring production pauses to safely integrate new network bridges. To avoid these operational disruptions and the risks associated with altering critical control loops, the Sensors Network employs a non-invasive strategy.

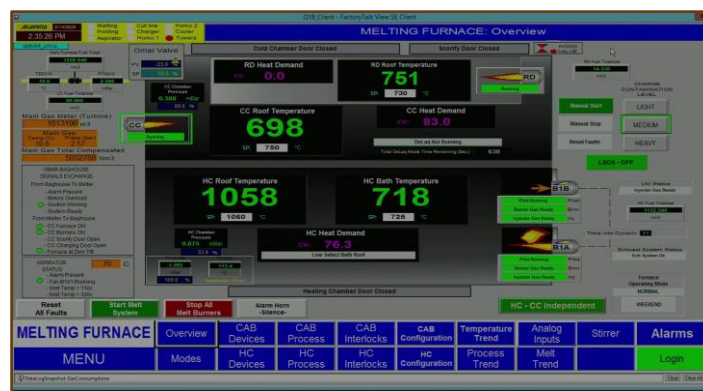


Figure 35: SCADA system for furnace condition management.



Figure 36: Initial approach recording the screen with a camera.

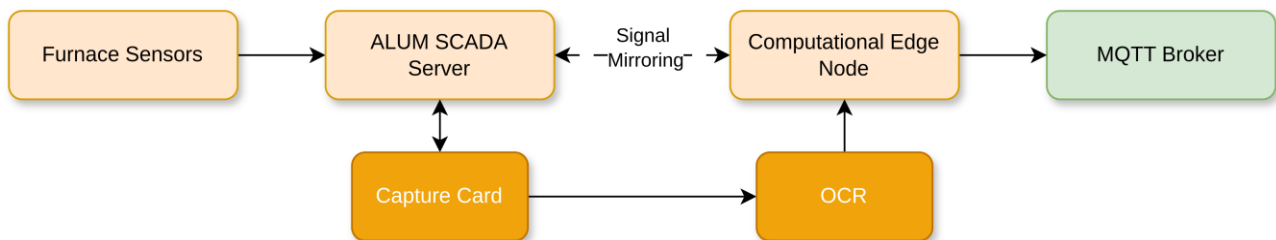


Figure 37: ALUM SCADA server signal mirroring via capture card and OCR.



Figure 38: The capture card and the cabling used for monitor mirroring.

PROJECT 101178230 – ENCIRCLE

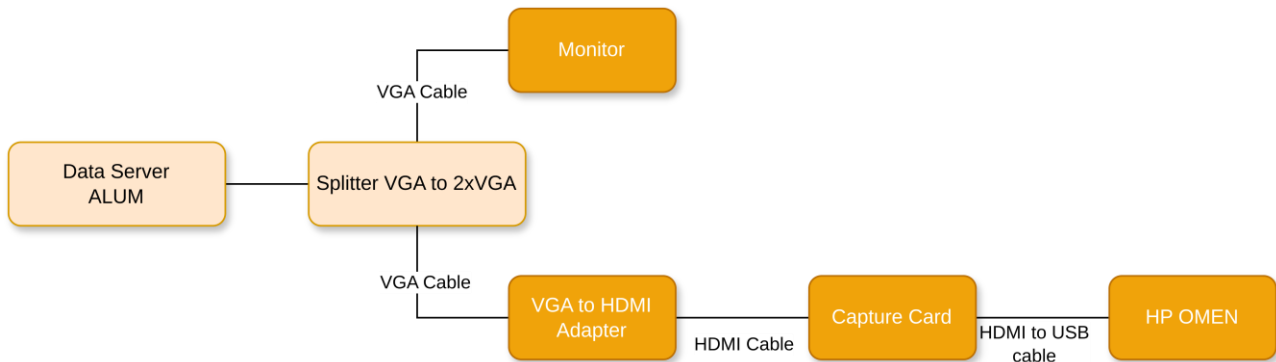


Figure 39: Cabling to mirror the ALUM SCADA data server to the computational edge node.

These video frames are processed locally by the edge node, which extract the numerical values and convert them into structured data (Figure 40). This allows the system's soft sensors to predict the furnace's energy profile without ever touching the sensitive network.

Table 9: Elgato HD60X Capture Card Specifications.

Connectivity	USB 3.0
Supported Resolutions	1080p / 1080i / 720p / 480p / 1440p
Video Input/Output	HDMI 2.0
Audio Input/Output	HDMI

timestamp	loading	star	loading	en	CC	CHAMBER	PRE	CC	ROOF	TEM	CC	HEAT	DE	DOOR	STATU	RD	ROOF	TE	HC	ROOF	TE	HC	BATH	T	HC	HEAT	D	HC	CHA		
2025-11-14	2025-11-13	2	2025-11-13	2			0.151			676				14	Not Closed			714		1019			692			83.8		0.12			
2025-11-14	2025-11-13	2	2025-11-13	2			0.198			676				15	Closed			714		1019			692			84		0.116			
2025-11-14	2025-11-13	2	2025-11-13	2			0.191			676				15	Closed			714		1019			692			83.9		0.139			
2025-11-14	2025-11-13	2	2025-11-13	2			0.148			676				15	Closed			714		1019			692			83.9		0.144			
2025-11-14	2025-11-13	2	2025-11-13	2			0.147			676				15	Closed			714		1019			692			84		0.14			
2025-11-14	2025-11-13	2	2025-11-13	2			0.148			676				15	Cold Chamber			714		1019			692			84		0.139			
2025-11-14	2025-11-13	2	2025-11-13	2			0.152			676				15	Closed			714		1019			692			84		0.147			
2025-11-14	2025-11-13	2	2025-11-13	2			0.154			676				15	Closed			714		1019			692			84		0.145			
2025-11-14	2025-11-13	2	2025-11-13	2			0.157			676				15	Closed			714		1019			692			84		0.147			
2025-11-14	2025-11-13	2	2025-11-13	2			0.158			676				15	Closed			714		1019			692			84		0.15			
2025-11-14	2025-11-13	2	2025-11-13	2			0.159			676				15	Closed			714		1019			692			84.1		0.146			
2025-11-14	2025-11-13	2	2025-11-13	2			0.16			676				15	Closed			714		1019			692			84		0.147			
2025-11-14	2025-11-13	2	2025-11-13	2			0.162			676				15	Closed			714		1019			692			84.1		0.146			
2025-11-14	2025-11-13	2	2025-11-13	2			0.162			676				15	Closed			714		1019			692			84.1		0.147			
2025-11-14	2025-11-13	2	2025-11-13	2			0.164			676				15	Closed			714		1019			693			84.1		0.151			
2025-11-14	2025-11-13	2	2025-11-13	2			0.164			676				15	Closed			714		1019			692			84.1		0.155			
2025-11-14	2025-11-13	2	2025-11-13	2			0.164			676				15	Closed			713		1019			692			84		0.159			
2025-11-14	2025-11-13	2	2025-11-13	2			0.164			676				15	Closed			714		1019			692			84.1		0.159			
2025-11-14	2025-11-13	2	2025-11-13	2			0.165			676				15	Closed			713		1019			692			84.1		-0.002			

2025-11-14	2025-11-13	2	2025-11-13	2	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed
2025-11-14	2025-11-13	2	2025-11-13	2	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed
2025-11-14	2025-11-13	2	2025-11-13	2	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed
2025-11-14	2025-11-13	2	2025-11-13	2	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed
2025-11-14	2025-11-13	2	2025-11-13	2	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed
2025-11-14	2025-11-13	2	2025-11-13	2	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed
2025-11-14	2025-11-13	2	2025-11-13	2	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed
2025-11-14	2025-11-13	2	2025-11-13	2	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed
2025-11-14	2025-11-13	2	2025-11-13	2	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed
2025-11-14	2025-11-13	2	2025-11-13	2	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed
2025-11-14	2025-11-13	2	2025-11-13	2	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed
2025-11-14	2025-11-13	2	2025-11-13	2	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed
2025-11-14	2025-11-13	2	2025-11-13	2	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed
2025-11-14	2025-11-13	2	2025-11-13	2	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed
2025-11-14	2025-11-13	2	2025-11-13	2	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed
2025-11-14	2025-11-13	2	2025-11-13	2	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed
2025-11-14	2025-11-13	2	2025-11-13	2	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed
2025-11-14	2025-11-13	2	2025-11-13	2	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed
2025-11-14	2025-11-13	2	2025-11-13	2	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed
2025-11-14	2025-11-13	2	2025-11-13	2	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed
2025-11-14	2025-11-13	2	2025-11-13	2	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed
2025-11-14	2025-11-13	2	2025-11-13	2	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed
2025-11-14	2025-11-13	2	2025-11-13	2	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed
2025-11-14	2025-11-13	2	2025-11-13	2	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed
2025-11-14	2025-11-13	2	2025-11-13	2	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed
2025-11-14	2025-11-13	2	2025-11-13	2	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed
2025-11-14	2025-11-13	2	2025-11-13	2	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed
2025-11-14	2025-11-13	2	2025-11-13	2	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed
2025-11-14	2025-11-13	2	2025-11-13	2	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed
2025-11-14	2025-11-13	2	2025-11-13	2	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1496	Closed	734	0.016	1					

Figure 40: OCR captured data from the SCADA software.

PROJECT 101178230 – ENCIRCLE

Complementing the external telemetry, the physical state of the scrap during the recycling is monitored via a specialized in-furnace camera. This thermal-resistant optical sensor withstands the extreme environment of the melting chamber. It provides a direct visual feed of the scrap pyrolysis procedure, allowing the system to observe the phase changes of the material and the combustion of organic contaminants (Figure 41). Furthermore, this sensor is focal for monitoring dross accumulation on the melt surface, as the visual data acquired here serves as the ground truth for the soft sensors that estimate dross coverage percentages, a key metric for yield efficiency.

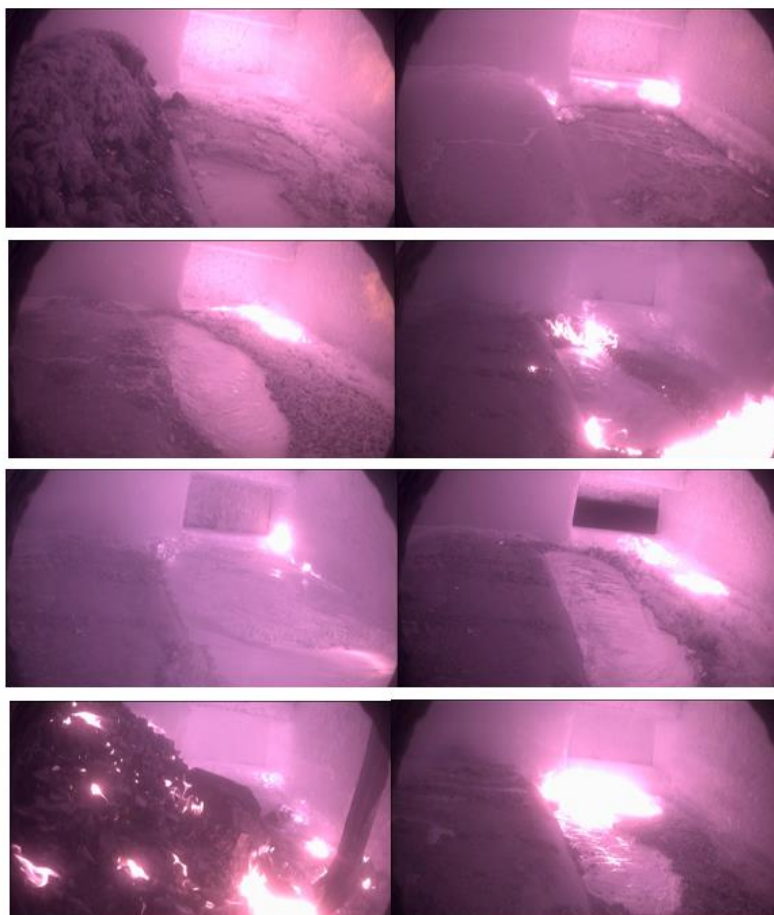


Figure 41: In furnace image captures.

The integration of these disparate data sources (scalar and time-series values from the OCR system and unstructured image data from the thermal camera) is managed through a timestamp-based temporal synchronization mechanism. The Local Aggregation Server ensures that every frame captured by the in-furnace camera is precisely aligned with the corresponding readings extracted from the SCADA screen. This alignment is crucial for the downstream soft sensors, allowing them to correlate specific operational conditions (e.g. a spike in heat demand) with physical phenomena observed inside the furnace.

3.2 ZINQ PILOT SITE

This section details the sensor network architecture deployed for the ZINQ pilot site. The objective of this layer is to establish a robust data acquisition infrastructure that bridges the gap between legacy sensors and the project's IIoT requirements. The network comprises a mix of existing industrial sensors, and newly installed thermal and machine vision camera hardware.

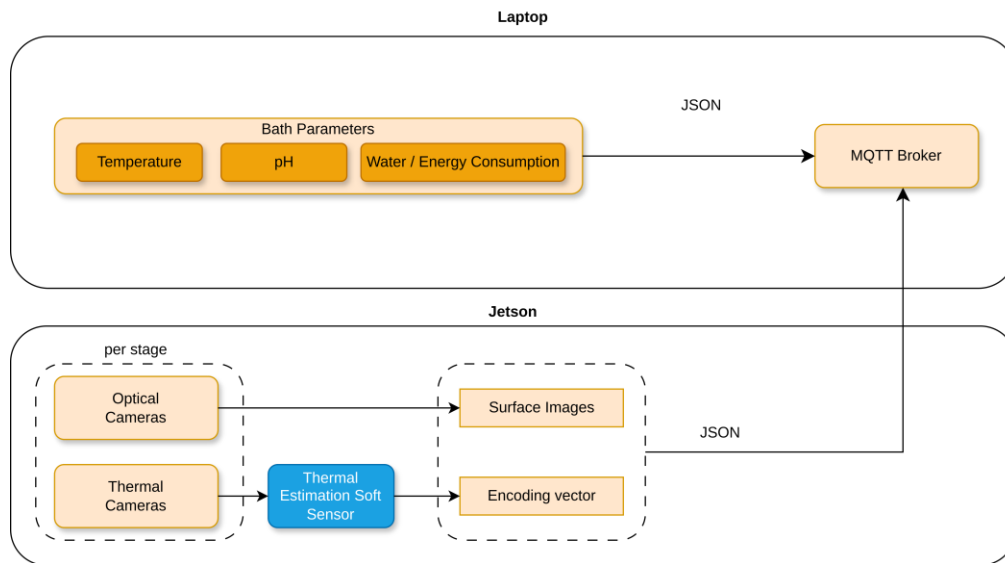


Figure 42: ZINQ Pilot site sensor overview.

3.2.1 Edge Computing and Data Flow

To manage the computational load of processing high-resolution visual and thermal streams, the ZINQ site utilizes a consolidated edge computing architecture. A powerful central edge node, a Dell Latitude 5440 (Table 10), handles data aggregation and local processing. Equipped with an Intel Core i5 processor and 16GB of RAM, this device functions as the site's Local Aggregation Server. It interfaces directly with legacy sensors and connects to the Jetson device, via Gigabit Ethernet (GigE), that interfaces directly with both the thermal and the machine vision cameras.

Table 10: Dell Latitude 5440 Specifications.

CPU	Core i5 1335U-3.4GHz
Storage Memory	SSD 512GB
RAM	16 GB
Connectivity	Ethernet, Wi-Fi

This Local Aggregator Server is responsible for protocol abstraction, via the integrated Middleware, and data transmission. The middleware and the installed MQTT broker, unify data streams from the legacy sensors (via PLC/Modbus bridges) and the modern optical sensors (Machine Vision and Thermal). Additionally, the thermal time-series data is synchronized with the visual snapshots of the material, creating "process batch" record. This consolidated data is then forwarded to downstream ENCIRCLE modules via a standardized REST API or secure MQTT forwarding.

3.2.2 Pre-Treatment and Process Monitoring

The pre-treatment phase relies on a tightly integrated multi-modal sensor array to enable fine-grained monitoring and control of material handling, bath chemical state, and thermal operating conditions, all of which are critical for achieving consistent surface activation and ensuring the robustness, adhesion, and uniformity of the downstream galvanic coating. To track the flow of production, the facility leverages legacy sensors integrated into the overhead cranes, which provide the essential temporal data needed to correlate the physical movement and dipping times of material batches with specific process steps.

Complementing this tracking, a FLIR A50 Thermal Camera (Figure 48 and Table 14) has been deployed to monitor the thermal stability of the baths. This camera sensor captures real-time thermal gradients to detect anomalies such as uneven heating. Furthermore, to guarantee optimal surface preparation before the galvanizing kettle, a machine vision camera (Figure 43 and Table 11) equipped with a specialized lens (Figure 44 and Table 12) captures high-resolution imagery of

PROJECT 101178230 – ENCIRCLE

the incoming materials, serving as the ground truth for identifying surface impurities like rust residues that could compromise coating quality. Data collection from the thermal and machine vision cameras will be completed by M18, with progress updates systematically documented and reported in Deliverable D3.2.



Figure 43: GV-5290FA-C-HQ machine vision camera.

Table 11: GV-5290FA-C-HQ Specifications.

Sensor	Sony IMX267 (Color)
Sensor Type	1" CMOS
Shutter Type	Global Shutter
Pixel Size	3.45 x 3.45 μm
Resolution	4104 x 2174 px
FPS	13
Data Interface	Gigabit Ethernet, Transfer rate 1000 Mbits/sec
Power Supply	M8/8 pins or Power over Ethernet (PoE)



Figure 44: Fujinon Lens CF8ZA-1S.

Table 12: Fujinon Lens CF8ZA-1S Specifications.

Focal Length	8 mm
Iris	F1.8-F16
Iris Type	Manual
Pixel Pitch	2.74 μm
Min. Working Distance	100 mm

Wavelength Range

Visible

3.2.3 Hot-Dip Galvanization and Post-Treatment

The core of the ZINQ process involves the immersion of steel parts into the molten zinc kettle, where the steel, after undergoing pre-treatment through degreasing, pickling, and fluxing, is dipped into zinc heated to approximately 450 °C, allowing a reaction to occur between iron and zinc that forms a series of tightly bonded alloy layers and a durable outer zinc coating, which together provide long-term corrosion protection, mechanical robustness, and uniform coverage even on complex geometries characteristic of the hot-dip galvanization process (Figure 45). Precise thermal management is significant, as temperature variations directly correlate with coating thickness and zinc consumption efficiency. To capture this dynamic thermal profile, the FLIR A50 Thermal Camera (Figure 48 and Table 14) remains active throughout the dipping cycle. It logs the bath's surface temperature distribution, identifying potential cold or overheating spots. This thermal data serves as the primary input for the soft sensors developed, which are designed to predict the final coating characteristics and zinc usage based on the specific thermal history of each dip. The data collection from these sensors is ongoing and will be completed by M18, and the corresponding progress and preliminary results will be documented and reported in Deliverable D3.2.

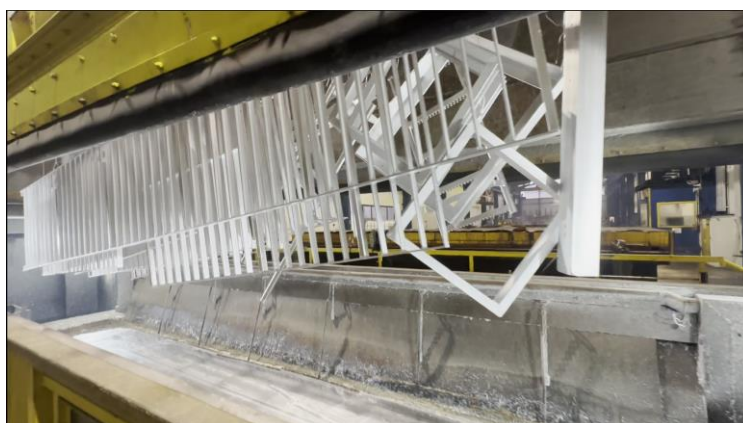


Figure 45: Hot dip galvanization.



Figure 46: Projected camera installation.

Table 13: Thermal camera configuration.

Object distance

4-5 meters

PROJECT 101178230 – ENCIRCLE

Horizontal FOV	95°
Sensor width	10.9
Effective focal length	6-7.5 mm
Horizontal pixel count	640 px

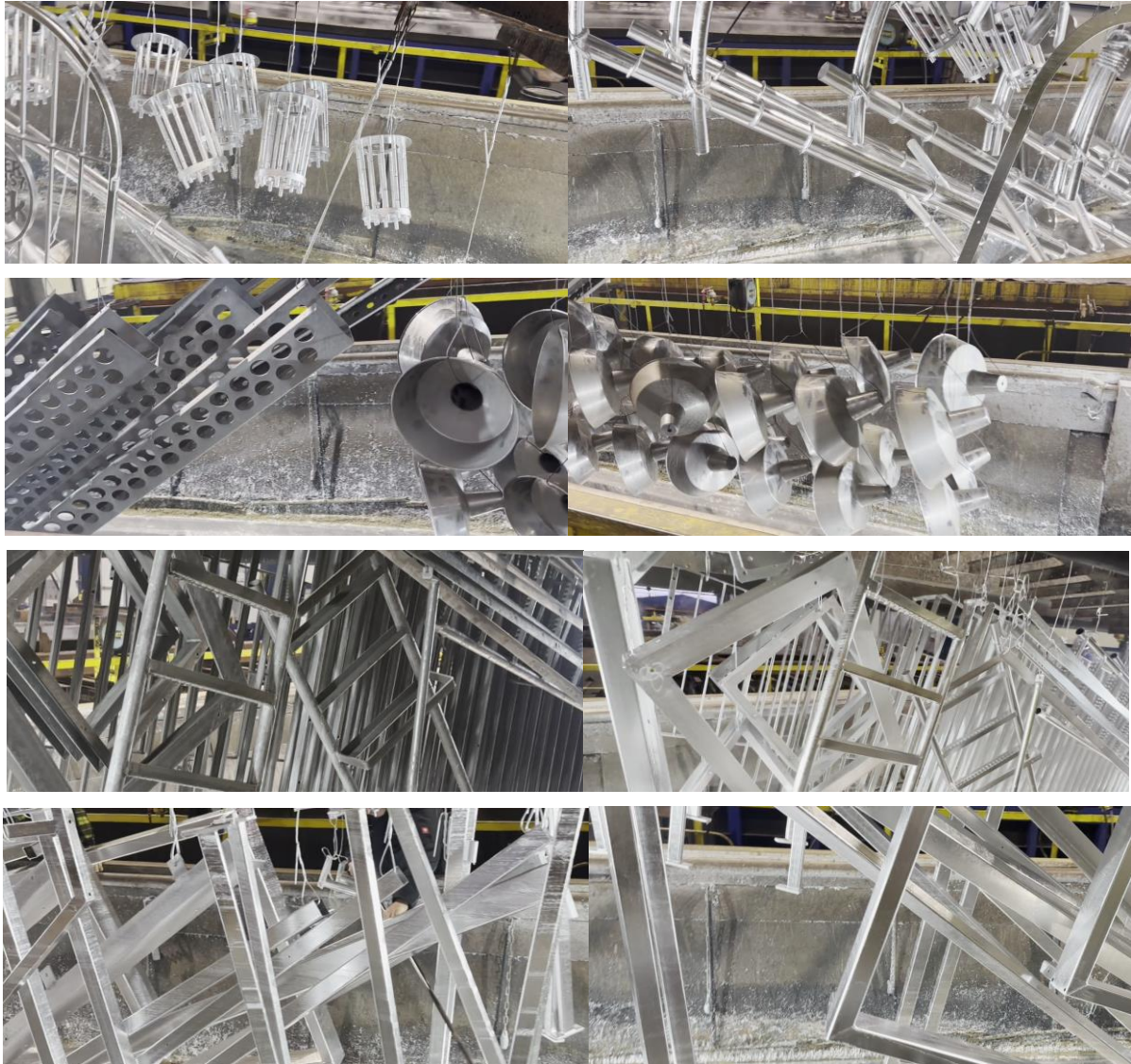


Figure 47: Hot- Dip Galvanization parts.

To ensure reliable thermal camera captures of the bath during the hot-dip galvanization cycle, careful attention was given to the positioning and optical configuration of the thermal imaging system. The camera is projected to be installed at an elevated and fixed vantage point overlooking the galvanization bath and the parts (Figure 47), while minimizing occlusions from structural elements and moving equipment. The camera-to-object distance was constrained by plant layout and safety requirements and was therefore set within the range of 4–5 m from the bath surface and the final camera position was determined after experimental trials evaluating the field-of-view (FoV) and optimal bath coverage.

PROJECT 101178230 – ENCIRCLE

To accurately determine this object distance and validate the camera geometry, four to five fixed reference objects with known dimensions and spatial locations (e.g., bath edges, beams) were used during casting procedures. These reference items enabled effective object distance and field-of-view verification directly within the operational environment. Based on the measured object distance and the required coverage area, the effective focal length was calculated, and an appropriate camera was selected to balance spatial resolution, field of view, and thermal sensitivity.



Figure 48: FLIR A50 95° Thermal Camera.

Following the galvanization process, the sensor network shifts focus to surface quality verification during the post-treatment phase. To automate this critical inspection step, the facility utilizes the GV-5290FA-C-HQ machine vision camera setup (Figure 43 and Table 11). This sensor captures high-resolution imagery of the finished galvanized surfaces. These images function as the digital ground truth for the project's automated defect detection models, which are trained to identify specific galvanization faults.

Table 14: FLIR A50 95° Thermal Camera Specifications.

Accuracy	+/- 2 degrees Celsius or better
FOV	95° × 74°
IR Resolution	640 × 480 px
Operating Temperature	-20 to 50 °C
Temperature measuring range	-20 to 1000 °C
Data Interface	Gigabit Ethernet, Transfer rate 1000 Mbits/sec
Power Supply	Power over Ethernet (PoE)

Complementing this real-time acquisition, the system also integrates a comprehensive set of historical production data derived from both digital sources within the facility. This legacy dataset is composed data that documents station/bath durations (Figure 49) for past orders, alongside manual quality control logs specifically detailing coating thickness measurements and documents that contain daily production parameters such as flux bath temperature and PH. By consolidating these records, the system establishes a statistical baseline for normal process variability, which is used to validate the live camera readings and refine the accuracy of the soft sensors.

PROJECT 101178230 – ENCIRCLE

Auftrag 2030065929								
#	POS	Arbeitsplatz	ID	Ereignis Start	Ereignis Stop	Schicht	Traverse	Bemerkung
2	30	Schwarzlager Feld 1	BUFFMIXED	19.02.2025 09:48:41		FS		
3	30	Andrahtstation 4	LOADING	20.02.2025 06:00:00	20.02.2025 07:41:05	FS	6-0	01:41:05
5	30	Entfettung	CLEANING	20.02.2025 07:41:08	20.02.2025 07:57:49	FS	6-0	00:16:41
6	30	Beize 2	PICKLING	20.02.2025 07:57:49	20.02.2025 08:29:18	FS	6-0	00:31:29
7	30	Beize 1	PICKLING	20.02.2025 08:29:18	20.02.2025 10:40:43	FS	6-0	02:11:25
8	30	Spüle	RINSING	20.02.2025 10:40:43	20.02.2025 10:42:45	FS	6-0	00:02:02
9	30	Flussmittelbad	FLUXING	20.02.2025 10:42:45	20.02.2025 10:44:47	FS	6-0	00:02:02
16	30	Trockenofen	DRYING	20.02.2025 10:45:08	20.02.2025 11:02:10	FS	6-0	00:17:02
17	30	Kessel	ZINCBATH	20.02.2025 11:02:10	20.02.2025 11:13:24	FS	6-0	452°
18	30	Quertransport	BUFFWHITE	20.02.2025 11:13:24	20.02.2025 11:20:17	FS	6-0	00:06:53
19	30	Abdrahtstation 1	UNLOADING	20.02.2025 11:20:17	20.02.2025 12:20:22	FS	6-0	01:00:05
20	30	Waage	WEIGHING	20.02.2025 12:20:22		MS	6-0	

Figure 49: Documentation of which bath is used for how long for each order.

1	2	3	4	5
Anker Schroeder ASDO GmbH	Anker Schroeder ASDO GmbH	Anker Schroeder ASDO GmbH	Anker Schroeder ASDO GmbH	Anker Schroeder ASDO GmbH
55	55	55	55	55
70	70	70	70	70
2039192291	2039192291	2039192291	2039192291	2039192291
Schwere Profil	Schwere Profil	Schwere Profil	Schwere Profil	Schwere Profil

Thickness measurements

Messung	X1	X2	X3	X4	X5	GemX	
Messung 1.1	387	449	396	385	384	400,2	349.3
Messung 1.2	342	376	355	370	362	361	
Messung 1.3	277	293	291	293	279	286,6	
Messung 2.1	248	313	269	334	256	284	281.5
Messung 2.2	257	256	254	256	296	263,8	
Messung 2.3	296	306	284	301	296	296,6	
Messung 3.1	287	304	296	299	298	296,8	

Figure 50: Thickness measurement

ZINQ®											
QMS FORMBLÄTTER											
GFW097											
Tägliche Produktionsdaten											
Datum											
Tag der Woche	Montag		Dienstag		Mittwoch		Donnerstag		Freitag		
Tonnage in KG	-		-		-		-		-		
Produktionsdaten											
Tauchungen	0		0		0		0		0		
SCB	0		0		0		0		0		
Stahlbau	0		0		0		0		0		
Entfetter	E1	E2	E1	E2	E1	E2	E1	E2	E1	E2	
Temp. Entfettung (20-24°C) / PH	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	
Temp. H ² O Becken (30-35°C)	18,0		18,0		18,0		18,0		18,0		
Leratens 1019	10	0	10	0	10	0	10	0	10	0	
Fluxbad											
Laufzeit Flux-Press in h	15,00		15,00		15,00		15,00		15,00		
Rechnerischer Verbrauch Flux	-		-		-		-		-		G
Wöchentliche Zugabe Flux in ltr.											
Wöchentliche Zugabe Zinkchlorid in ltr.											
Temp. Flux (50-60°C) / PH- Wert	48,0	3,89	45,0	3,85	44,0	3,96	40,0	4,02	40,0	4,03	
Salmiak 18% in ltr. (06:30 Uhr ablesen)	40		40		40				40		
Wasserstoffperoxid- Zugabe in ltr.											
20% H ₂ O ₂ + 80% H ₂ O	30		30		30		30		30		
Polymer in ltr. (06:30 Uhr ablesen)	60		60		60		60		60		

Figure 51: Production parameters.

4 SOFT SENSORS

This section presents the development and implementation of Soft Sensors within the ENCIRCLE framework, building upon the IIoT infrastructure and physical sensor networks described in the sections 3 and 4.3. In line with the ENCIRCLE's objective of enabling circular and sustainable manufacturing, these models are employed to predict critical process, quality, and environmental indicators that cannot be directly or continuously measured through hardware sensors alone. Leveraging advanced AI algorithms and data-driven modelling, these soft sensors transform heterogeneous multi-modal data, such as images, audio signals, and production parameters, into actionable insights supporting real-time monitoring, prediction, and decision-making. The soft sensing solutions described in this section constitute a core intelligence layer of WP3, feeding higher-level modules like Digital Twins, environmental impact analytics engine, and optimization loops, while ensuring adaptability, and reduced instrumentation costs across.

4.1 ALUMIL SOFT SENSORS

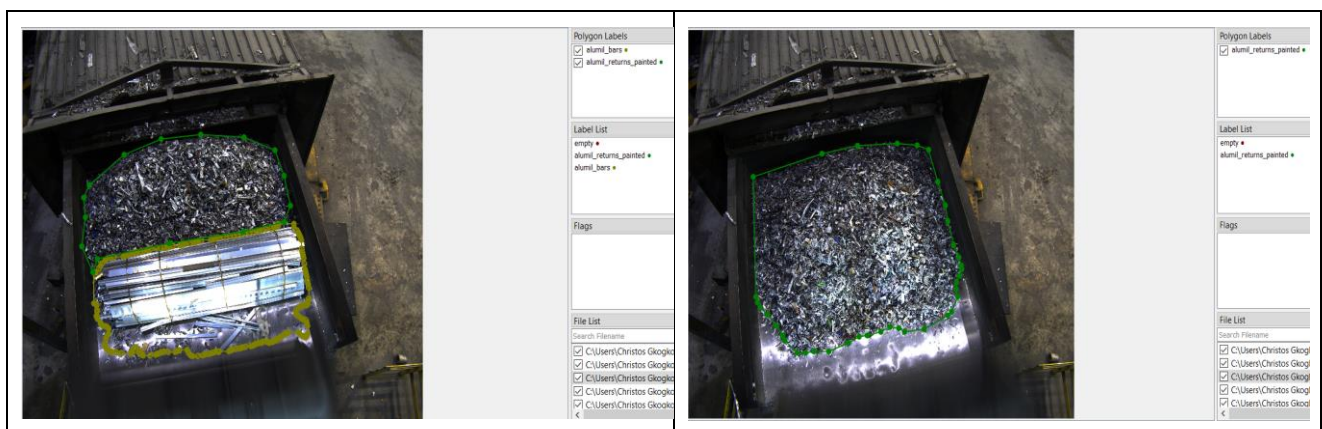
This subsection focuses on the Soft Sensor solutions developed and validated at the ALUMIL pilot site, addressing key challenges in aluminium recycling processes. The proposed models exploit vision-based and data-driven approaches to estimate critical process variables such as scrap composition, heat demand, and dross formation, which directly influence energy consumption, material efficiency, and environmental performance. By integrating deep learning models with real-time image acquisition and edge computing solutions, ENCIRCLE enables continuous process monitoring of the melting process, while supporting predictive control optimization.

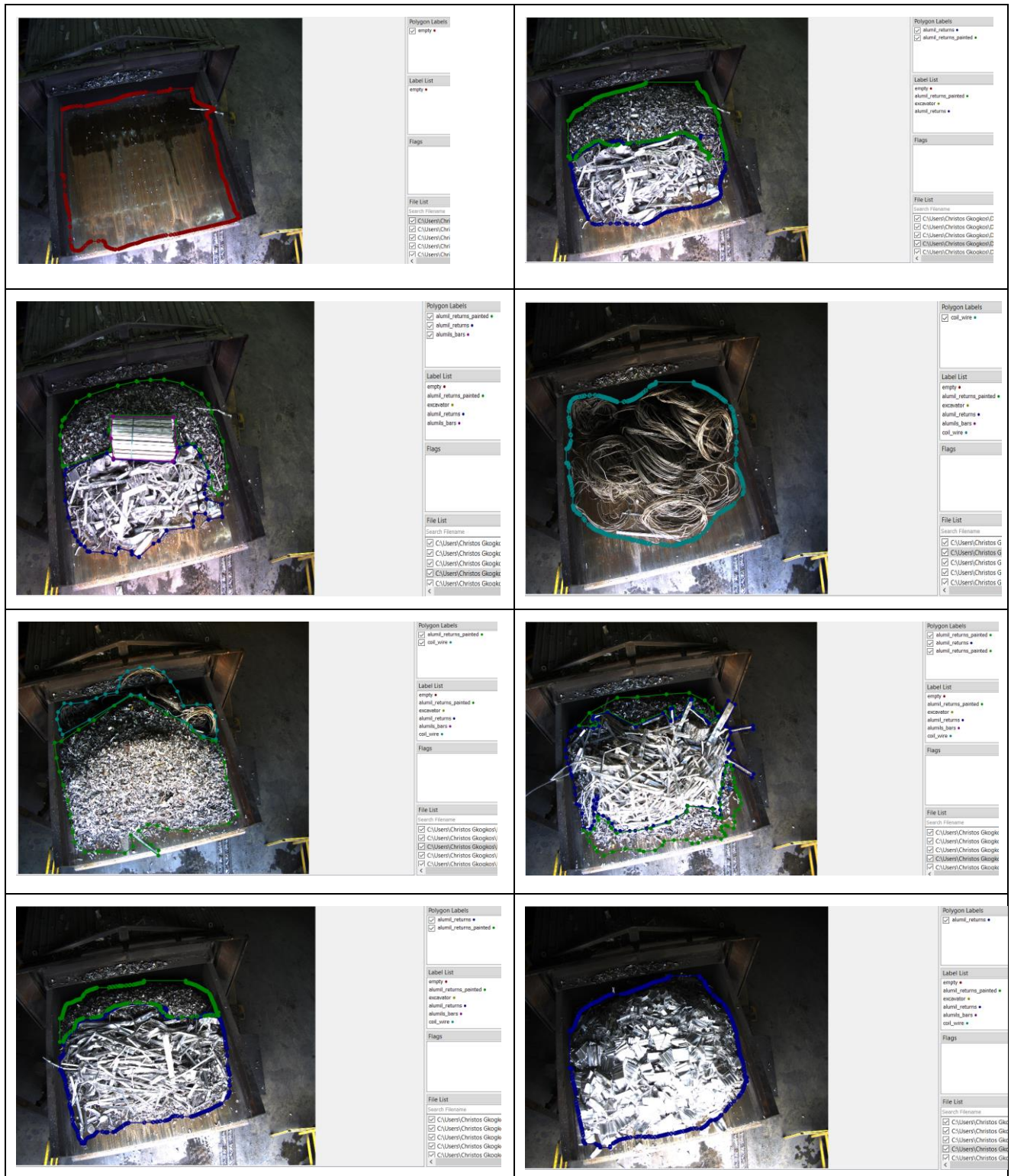
4.1.1 YOLO-based Visual Soft Sensors for Process Monitoring

Within the developed system, a YOLO-based computer vision pipeline was implemented as the core component of a visual soft sensing framework for process monitoring. The objective was not limited to object detection, but to determine the loading status of the process by analyzing how scrap appears and disappears over time in images captured before furnace entry.

4.1.1.1 Problem Formulation

Currently, in the secondary aluminium production process, the lack of semantic data regarding raw materials creates a significant blind spot. Traditional sensors measure weight but fail to identify material types or operational context. Therefore, the problem addressed in this section is the development of a visual soft sensing framework that transforms discrete image captures into actionable process intelligence. A critical constraint in developing such a framework is the scalability of dataset creation. The system is required to identify 10 distinct scrap categories, specifically alumil bars, alumil dust, alumil returns, painted returns, circles, cleanser, coil wire, OMAN, and cubes as visually depicted in Figure 52. However, manually annotating the thousands of images required to train deep learning models for these diverse categories constitutes a prohibitive operational bottleneck. Consequently, the problem is formulated to necessitate an automated annotation pipeline. The primary requirement is that for each acquired image, the model must output bounding boxes and class labels that are automatically generated in LabelMe-compatible JSON format. This functionality is essential to significantly reduce annotation effort, enabling the rapid construction of large-scale, consistent datasets while ensuring high-precision detection of the scrap types entering the furnace.





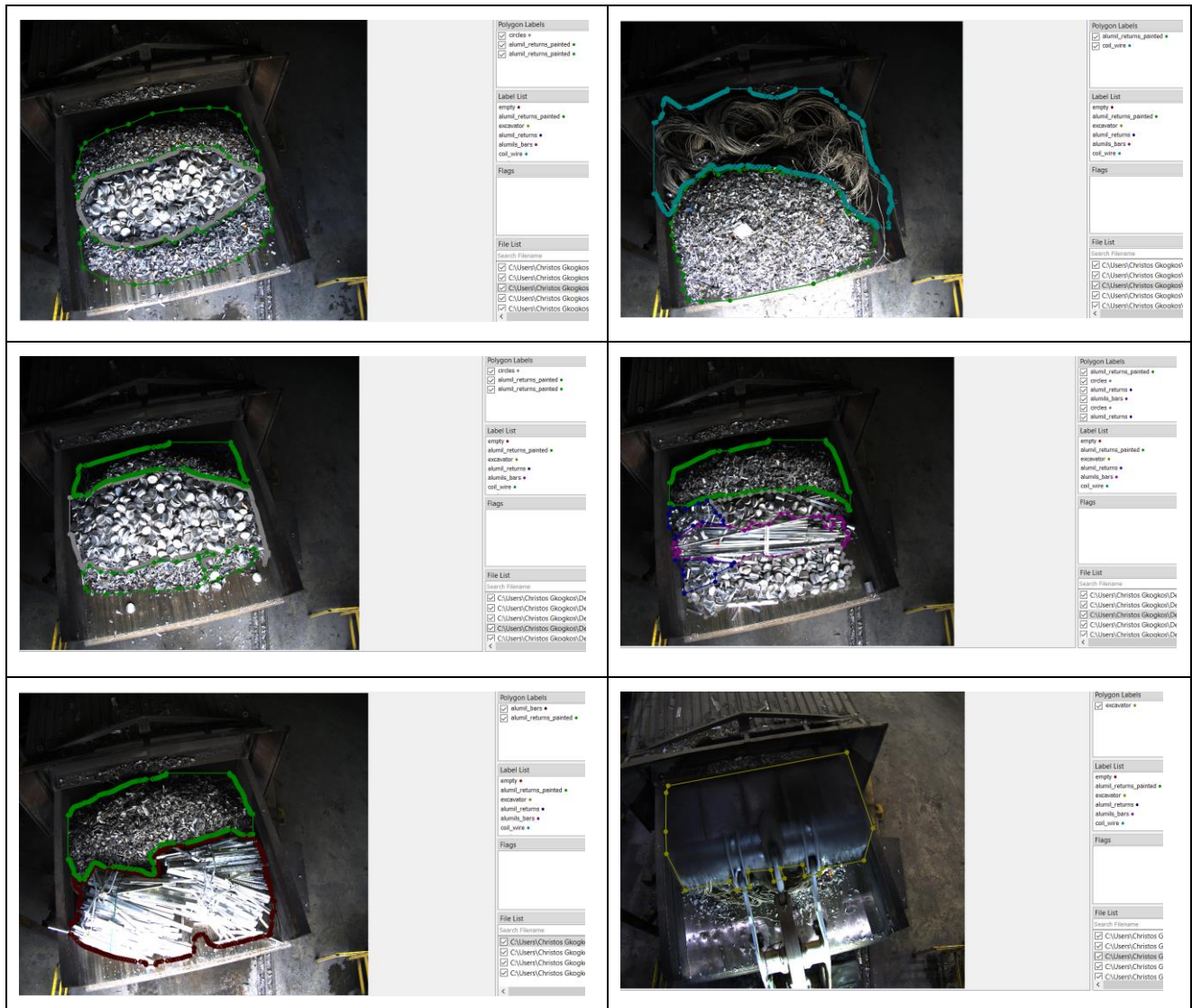


Figure 52: Visualization of the automated semantic annotation process within the LabelMe interface. The system converts YOLO-based detections into editable polygon labels, where distinct colors correspond to specific scrap categories (e.g., circles, alumi_return)

4.1.1.2 Dataset

The dataset utilized for training and validating visual soft sensors comprises a total of 3,756 images, split into 3,004 training images and 752 validation images. The dataset includes annotations for instance segmentation across multiple classes related to the process monitoring (e.g., alumi_returns, cubes, coil_wire, etc.).

As illustrated in the dataset distribution analysis, the classes exhibit significant imbalance (see Figure 53). The categories alumi returns painted, alumi returns, and cubes represent the majority of instances, exceeding 1,000 samples each, whereas classes such as OMAN, column, and alumi dust are underrepresented with fewer than 100 instances. The spatial analysis of the bounding boxes indicates that while objects are generally centered, there is a substantial variance in object dimensions (width vs. height), reflecting the diverse geometry of the monitored items.

For the task of scrap-region detection and segmentation, the YOLOv8-Seg architecture was employed with the segmentation head enabled. The input images were resized to a resolution of 640 x 640 pixels (image size) to balance inference speed and detection accuracy. The training process was configured for a maximum of 500 epochs, utilizing a batch size

PROJECT 101178230 – ENCIRCLE

of 8 to accommodate the memory constraints of the 12GB VRAM GPUs. To ensure optimal convergence and prevent overfitting, an early stopping mechanism was implemented with a patience of 100 epochs.

The optimization strategy relied on the auto setting (typically defaulting to Stochastic Gradient Descent - SGD), initialized with a learning rate of 0.01, a momentum factor of 0.937, and a weight decay of 0.0005 for regularization. The training schedule included a linear warmup period of 3.0 epochs. Finally, to enhance the model's robustness against variations, Mosaic augmentation was applied with a probability of 1.0.

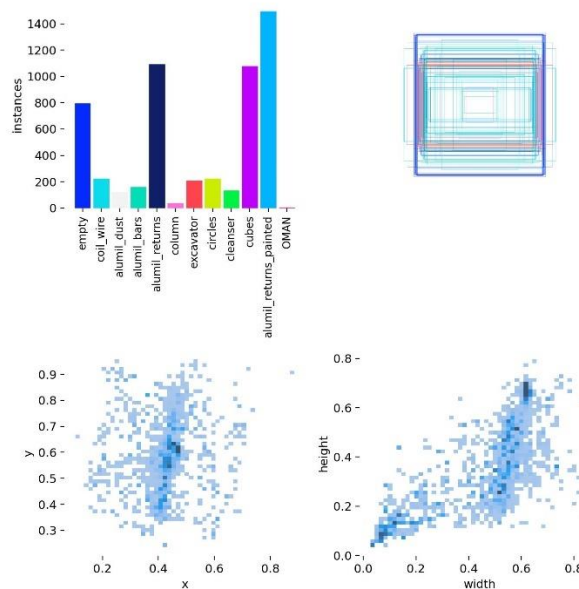


Figure 53: Statistical analysis of the training dataset, illustrating class distribution and bounding box spatial properties.

4.1.1.3 Architecture

The proposed visual soft sensor is built upon a modular architecture that integrates deep learning for perception with deterministic logic for process state inference. The system is composed of three primary subsystems: the Model-Assisted Annotation Pipeline, the Real-time Perception Engine, and the event inference module.

1. Model-Assisted Annotation Pipeline. To address the scalability of dataset creation, an automated workflow was established:

- Pre-annotation: A baseline YOLO model generates initial bounding box predictions for the 10 target scrap categories.
- Serialization: These detections are automatically exported into LabelMe-compatible JSON files, embedding class labels and spatial coordinates.
- Expert Verification: Domain experts review and refine these pre-labels using the LabelMe interface. This "Human-in-the-Loop" approach drastically reduces annotation time while ensuring high-quality ground truth data for the final model training.

2. Real-time Perception Engine The core of the online detection system is built upon the YOLOv8 object detection network. The fundamental architecture of this model, comprising the Backbone, Neck, and Head modules, has been detailed in the State-of-the-Art section and is illustrated in Figure 6. For the specific requirements of the ENCIRCLE soft sensor, this architecture was utilized to process discrete image snapshots from the scrap wagon. The model's prediction head was adapted to classify the 10 specific scrap categories defined in the problem formulation. For every processed frame, the engine outputs a list of detected objects, their class probabilities, and bounding box coordinates, effectively transforming raw pixel data into semantic process information.

4.1.1.4 Experimental Setup

All experiments were conducted on an on premise high-performance server designed to support intensive deep learning workloads. The workstation is equipped with an Intel Core i9-13900K CPU and 32 GB of system RAM. To accelerate the

PROJECT 101178230 – ENCIRCLE

training process, the system utilizes two NVIDIA GeForce RTX 3080 Ti GPUs (12 GB VRAM each). The training workflow was optimized to maximize throughput by configuring distinct training tasks per GPU, enabling parallel execution of experiments. The operating system used was Ubuntu 22.04.5 LTS (x86_64). The experimental pipeline was implemented in Python 3.10, leveraging a CUDA-enabled deep learning stack configured as follows:

- Deep Learning Framework: PyTorch 2.4.1
- GPU Drivers: NVIDIA Driver 550.144.03
- CUDA Runtime: 12.4

4.1.1.5 Experimental Results

The performance of the YOLO-based detection model was evaluated using confidence-dependent precision and F1-score analysis. High precision levels were observed across most confidence thresholds, while the optimal operational point was identified at a confidence value of approximately 0.83, yielding an overall F1-score close to 0.96. These evaluation metrics are illustrated in the Precision-Confidence and F1-Confidence curves shown in Figure 54 and Figure 55, respectively. Dataset analysis revealed class imbalance and significant variation in object size and position; nevertheless, the model maintained stable detection performance across all scrap categories.

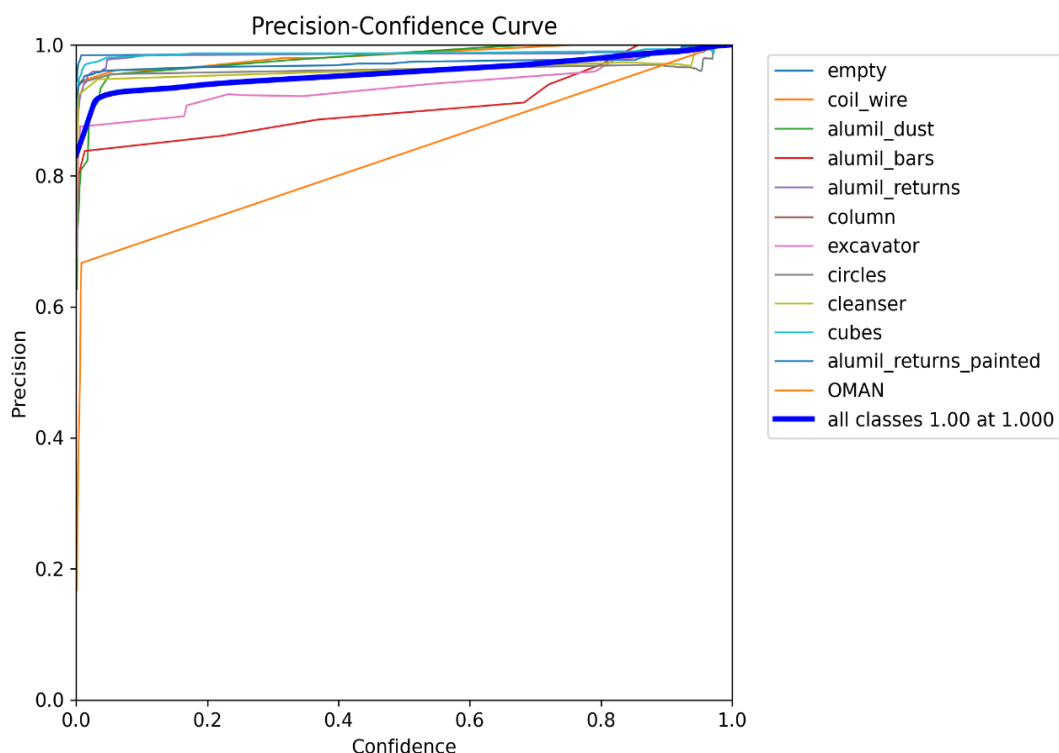


Figure 54: Precision-Confidence curve for the YOLO-based scrap detection model. High precision levels are maintained across most categories, ensuring reliable identification of scrap types before furnace entry.

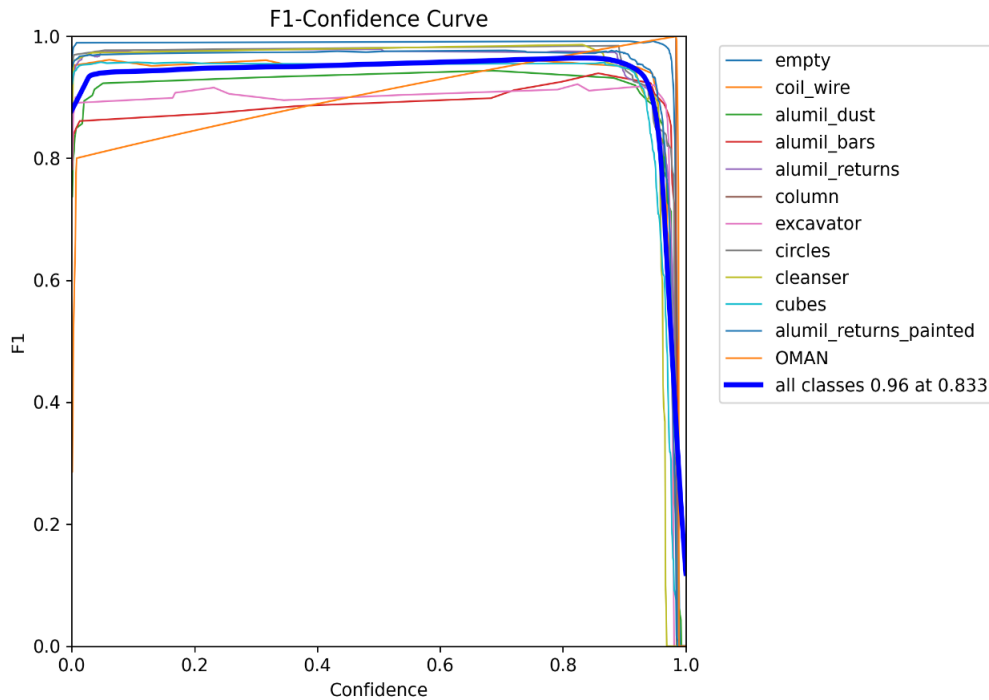


Figure 55: F1-Confidence curve illustrating the optimal operational point.

4.1.1.6 Conclusion

The YOLO-based visual soft sensor demonstrated strong capability in monitoring aluminium recycling processes under visually complex conditions. Experimental results show consistently high detection accuracy across scrap categories, with **precision and F1-score values exceeding 0.95** at the selected operational threshold. The use of annotations further improved material boundary prediction, enabling more accurate downstream material quantification. Deployed at the edge nodes, the model achieves **real-time inference** while maintaining robustness against occlusions and appearance variability. Overall, this soft sensor provides a reliable foundation for automated real-time process monitoring and material traceability in the ALUMIL pilot.

4.1.2 Image Quality Stabilization via Adaptive Illumination and Exposure Control

4.1.2.1 Exposure Control Using Sliding-Window Image Statistics

To cope with strong illumination variability caused by daylight ingress and artificial lighting conditions, an image brightness monitoring mechanism was integrated into the system as a baseline exposure control strategy during aluminium scrap acquisition. The approach is rule-based and is intended to stabilize image quality under changing illumination conditions, while also serving as a reference point for subsequent learning-based optimization methods, which will be deployed for the Digital Twin developments in T3.3.

As illustrated in Figure 56 (a), the camera is installed above the wagon where the scrap is loaded and in close proximity to the factory's opening door, resulting in direct sunlight entering the scene. Consequently, using a fixed exposure setting leads to underexposed images during night-time operation and overexposed images during daylight conditions, as depicted in Figure 56 (b) and (c), respectively.

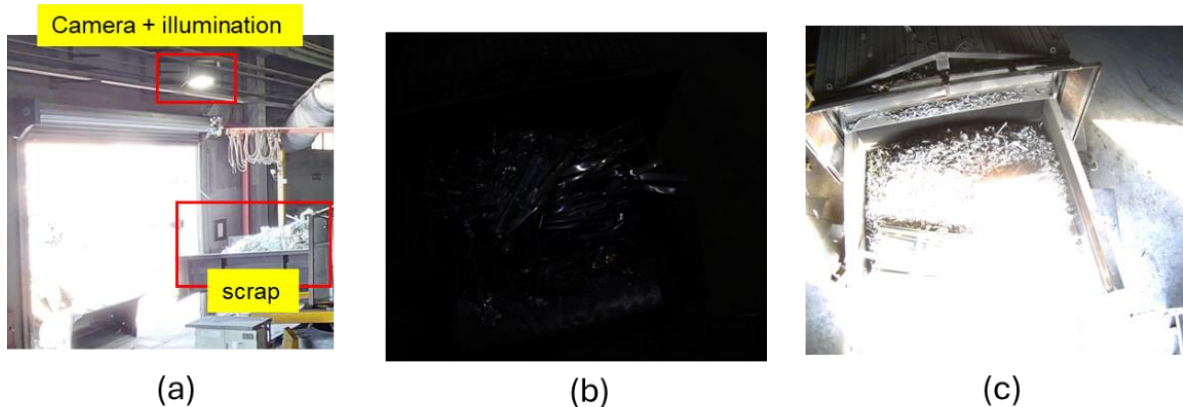


Figure 56: The camera is mounted above the scrap wagon near the factory door, where direct sunlight enters the scene (a), causing fixed exposure settings to produce underexposed images at night (b) and overexposed images during daylight (c).

A brightness-related metric is computed for each captured frame and acts as a soft sensor of image quality. The metric is evaluated over a sliding window of recently acquired frames to mitigate the effect of transient lighting fluctuations. In practice, the method monitors the fraction of saturated pixels, indicating overexposure, and the fraction of dark pixels, indicating underexposure, to detect deviations from a predefined operating range. When such deviations are observed, the camera exposure time is adjusted in discrete steps to return the image stream to a nominal regime suitable for downstream segmentation and visual analysis.

In addition, images acquired under different exposure settings are stored to support offline analysis and to enable the development and validation of more advanced illumination and exposure control strategies.

Inputs

- Sliding window size: X frames

Number of most recent frames over which brightness statistics are aggregated. The sliding window smooths transient illumination effects (e.g. brief reflections or glare) and ensures that exposure adjustments respond to persistent lighting changes rather than single-frame outliers.

- High-intensity threshold: T_{high} (e.g. 200)

Pixel-level intensity threshold ($[0-255]$) above which pixels are considered overexposed or near saturation. It defines which pixels contribute to the saturation metric and is determined by the camera bit depth and acceptable highlight clipping. In the current setup, a value of 200 was selected empirically through trial acquisitions and visual inspection of the resulting images.

- Low-intensity threshold: T_{low} (e.g. 50)

Pixel-level intensity threshold ($[0-255]$) below which pixels are considered underexposed. It defines which pixels contribute to the darkness metric and captures loss of detail in shadowed regions. In the current setup, a value of 50 was selected empirically through trial acquisitions and visual inspection of the resulting images.

- Saturation threshold : S_{max}

Maximum acceptable fraction of high-intensity pixels within an image (or sliding window). When exceeded, it indicates excessive overexposure at the image level and triggers a reduction in camera exposure. A value of 20% was selected empirically through trial acquisitions and visual inspection of the resulting images.

- Darkness threshold: D_{max}

Maximum acceptable fraction of low-intensity pixels within an image (or sliding window). When exceeded, it indicates underexposure and triggers an increase in camera exposure. A value of 30% was selected empirically through trial acquisitions and visual inspection of the resulting images.

- Exposure step size: ΔE (microseconds)

PROJECT 101178230 – ENCIRCLE

Discrete increment (in microseconds) by which the camera exposure time is increased or decreased. It controls the responsiveness of the adjustment, with larger steps enabling faster correction at the risk of oscillation. In the current setup, a value of 1000 μ s was selected empirically, as it provided a suitable trade-off between rapid convergence and stable exposure regulation.

- Exposure limits: E_{min}, E_{max}

Lower and upper bounds on allowable exposure time imposed by camera hardware constraints and motion-blur considerations. These limits prevent invalid or unsafe exposure settings during operation. In the current setup, $E_{min}=100\mu$ s, $E_{max}=60000\mu$ s.

Output

The output of the procedure is the updated camera exposure time $E_t \in [E_{min}, E_{max}]$ which is applied in the subsequent image acquisition cycle in order to maintain image brightness within the predefined operating range under varying illumination conditions. Below are the basic algorithmic steps that we followed for adjusting the exposure based on the image properties.

Let $I_t \in \{0, \dots, 255\}^{H \times W}$ denote the image acquired at time step t using exposure E_{t-1} . Let X denote the sliding window size, T_{high} and T_{low} the intensity thresholds for saturation and darkness, respectively. Let $[S_{min}, S_{max}]$ denote the acceptable saturation range, D_{max} the acceptable darkness bound, ΔE the exposure adjustment step, and E_{min}, E_{max} the minimum and maximum allowable exposure values.

Step 1: Per-frame image statistics

The saturation fraction for frame t is defined as

$$S_t = \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W 1(I_t(i, j) > T_{high}),$$

where $1(\cdot)$ denotes the indicator function.

The darkness fraction for frame t is defined as

$$D_t = \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W 1(I_t(i, j) < T_{low}).$$

Step 2: Sliding-window aggregation

Maintain sliding windows of the most recent X values:

$$\{S_{t-X+1}, \dots, S_t\}, \{D_{t-X+1}, \dots, D_t\}.$$

Compute the window-averaged statistics as

$$\dot{S}_t = \frac{1}{X} \sum_{k=t-X+1}^t S_k, \dot{D}_t = \frac{1}{X} \sum_{k=t-X+1}^t D_k.$$

Step 3: Exposure update rule

The exposure for the next acquisition step is updated according to

$$E_t = \begin{cases} \max(E_{min}, E_{t-1} - \Delta E), & \text{if } \dot{S}_t > S_{max}, \\ \min(E_{max}, E_{t-1} + \Delta E), & \text{if } \dot{D}_t > D_{max}, \\ E_{t-1}, & \text{otherwise.} \end{cases}$$

Step 4: Exposure application

PROJECT 101178230 – ENCIRCLE

The updated exposure value E_t is applied for the acquisition of the next frame I_{t+1} .

Two examples are illustrated in Figure 57. At the top, an example is shown where an initially saturated image is progressively improved by reducing the exposure over four adjustment steps. At the bottom, an underexposed image is correspondingly corrected by increasing the exposure until adequate brightness is achieved.

This baseline method provides a simple and computationally lightweight mechanism for stabilising exposure under varying illumination conditions and demonstrates the feasibility of online exposure adjustment using recent image statistics. However, it exhibits several important limitations that constrain its robustness and generality.

First, the approach is inherently scene-dependent. Changes in scrap composition, surface reflectance, or the presence of highly reflective objects can alter pixel intensity distributions independently of illumination, leading to misleading exposure adjustments. Second, the method is reactive rather than anticipatory, relying solely on recent frames and offering limited responsiveness to abrupt illumination changes such as direct sunlight exposure. Third, the use of fixed thresholds and discrete exposure steps can lead to oscillatory behaviour near decision boundaries, particularly in heterogeneous scenes. Finally, the metric is global and does not account for spatial structure, semantic relevance, or task-specific image quality criteria.

To overcome these limitations, subsequent experiments will investigate reinforcement learning–based exposure and illumination control strategies. In this formulation, exposure adjustment will be treated as a sequential decision-making problem, where actions are optimised with respect to an objective reward function that incorporates segmentation performance, image quality indicators, and stability constraints. This enables adaptive control policies that respond to complex scene dynamics and optimise illumination in a task-aware manner, moving beyond fixed heuristics toward data-driven illumination optimisation.

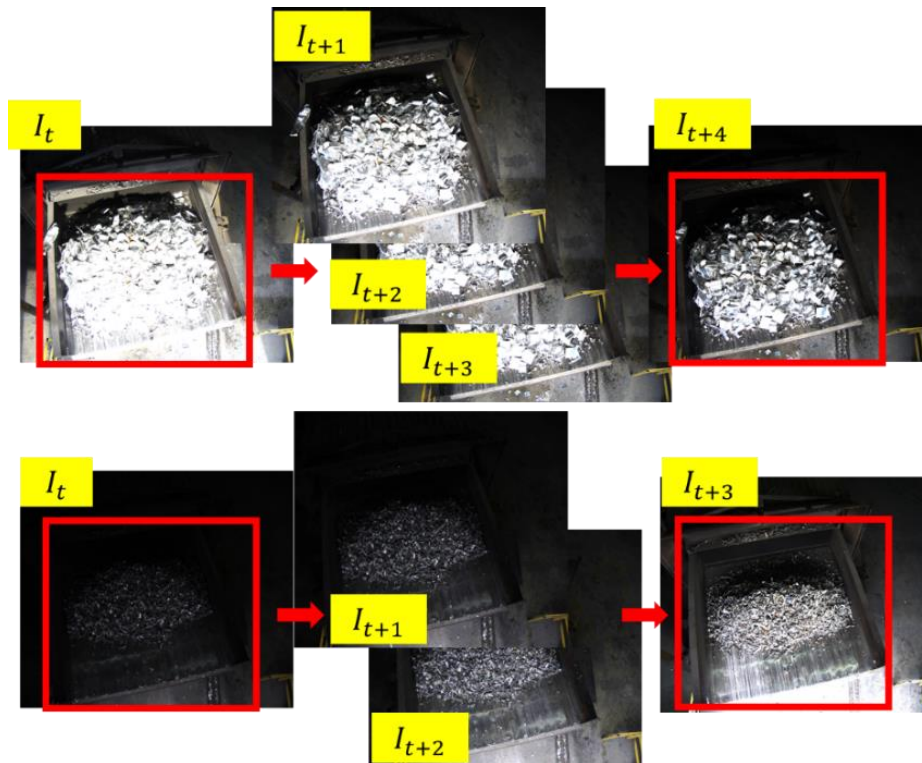


Figure 57: Camera exposure is adjusted based on image brightness measurements. At each iteration, the exposure time is increased or decreased in discrete steps until the measured brightness falls within the predefined acceptable range. The top example shows a saturated image corrected by reducing the exposure over four steps, while the bottom example shows an underexposed image corrected by increasing the exposure to reach an adequate brightness level.

4.1.2.2 Illumination Control for Enhanced Image Quality in Vision-Based Industrial Inspection

While the baseline exposure control mechanism demonstrates that online adjustment of camera parameters can stabilise image brightness under variable lighting, its limitations highlight the need for more task-aware control strategies. In particular, improving image quality in isolation is insufficient when the downstream objective is reliable visual analysis, as different illumination conditions can favour or suppress features relevant to segmentation. For this reason, subsequent experiments shift the focus from exposure-only regulation to illumination control informed by segmentation performance, enabling a systematic investigation of how illumination directly impacts algorithmic performance rather than brightness alone.

Vision-based segmentation on metallic surfaces is strongly influenced by illumination conditions, as illumination directly determines image contrast, reflection behaviour, and the visibility of surface features exploited by learning-based models. To investigate these effects in a systematic and data-driven manner, experimental studies were conducted on reflective steel surfaces, which exhibit optical properties comparable to aluminium in terms of specularly and sensitivity to illumination. The experiments leverage an existing large-scale, controlled multi-illumination dataset, enabling robust analysis of illumination–image interactions while production data from aluminium recycling are being collected within WP3.

In this context, segmentation is employed as a generic analytical tool for quantifying how variations in external illumination affect image quality and the prominence and separability of visually relevant surface characteristics, rather than as a task tied to a specific defect taxonomy or manufacturing process. The experimental results show that changes in illumination intensity and configuration can substantially alter the visual representation of surface features, even when the underlying material state remains unchanged. Some surface cues become more distinguishable under lower illumination, while others require higher intensity to emerge from background reflections or sensor noise. Conversely, excessive illumination can lead to saturation and loss of discriminative detail, degrading segmentation performance. These findings confirm that no single illumination configuration is universally optimal for reflective metallic surfaces.

The insights gained from these experiments provide a methodological foundation for illumination-aware inspection in aluminium recycling scenarios, where the primary objective is the visual differentiation of scrap categories and contaminants rather than defect localization. By developing and validating illumination-adaptive strategies on steel data with sufficient scale and variability, the proposed methodologies can be transferred to aluminium inspection setups, supporting robust image quality control under the heterogeneous and reflective conditions encountered in real recycling environments.

4.1.2.3 Dataset

To study illumination effects under controlled conditions, representative surface regions with varying textures were captured at multiple illumination levels. The same regions were imaged under low and higher illumination intensities to enable direct comparison of image contrast and saturation effects across illumination settings.

Experiments were conducted using a controlled inspection setup designed to minimize uncontrolled ambient illumination and ensure repeatable acquisition conditions. The setup consists of two LED (Light-Emitting Diode) illumination units, two HikRobot MV-CS200-10GC machine vision cameras with a resolution of 20 Mpixels (5472×3648), and a mounting fixture for the samples. Illumination is controlled by 0–10V dimming devices, allowing reproducible intensity adjustments. To simplify the data collection process, we discretized the dimmer signal of each illumination unit into steps corresponding to approximately 12.5% of the maximum output. From the resulting combinations, we selected eight illumination configurations spanning low to high intensity, discarding redundant settings that produced saturated or visually indistinguishable images. We denote these configurations as illumination levels $L_0, L_1, \dots, L_7$, which include extreme low and high cases to test the system's adaptability under diverse illumination conditions and abrupt changes.

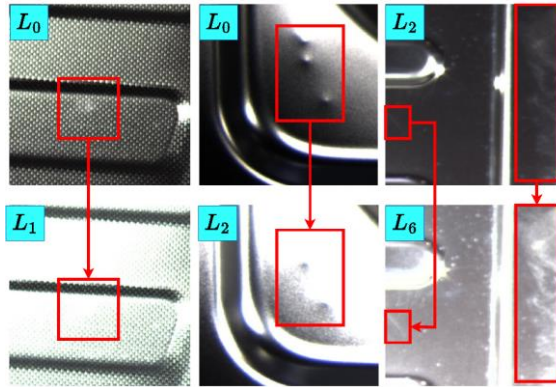


Figure 58: In (a), different types of defects and surface textures manifest differently as illumination conditions change. In the left and middle columns, the low illumination level (L_0) appears optimal for inspecting the surface defects, while in the right column, higher illumination (L_6) enhances the defect's prominence.

An existing dataset comprising stainless steel samples of varying sizes, geometries, and surface textures was used for experimentation. The dataset originates from an industrial production context and includes a range of naturally occurring surface irregularities as well as mechanically induced patterns, enabling systematic examination of how illumination conditions affect the visibility and segmentation of surface features on reflective materials.

In this context, surface irregularities are not treated as application-specific defects, but rather as representative visual structures used to study the impact of illumination on image quality and segmentation behaviour. The insights gained from these experiments are directly relevant to aluminium recycling scenarios, where the inspection objective shifts from defect localization to the visual identification of contaminants and scrap categories, such as plastics, coatings, or residual paint. By analysing how illumination control influences segmentation performance on reflective steel surfaces, the experiments provide transferable understanding for improving image-based segmentation under the heterogeneous and reflective conditions encountered in aluminium scrap inspection.

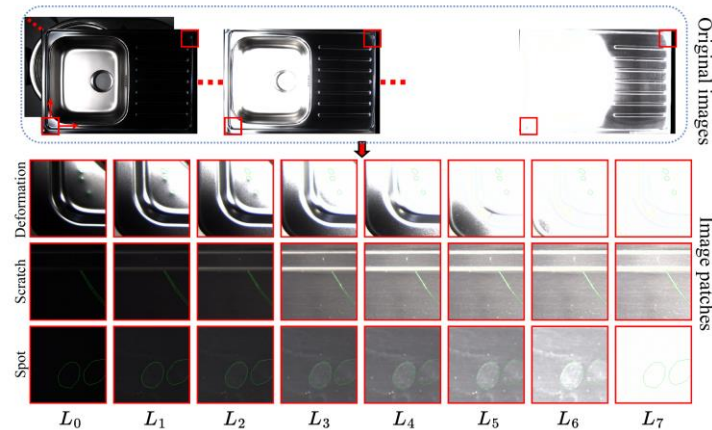


Figure 59: For each sample, illustrated at the top, 8 original images were captured under 8 different illumination levels. At the bottom, each original image was split into patches, capturing defects across the 8 levels (columns). The defect types include Deformation (Df), Scratch (Sc), and Spot (Sp).

Each sample was captured under 8 illumination levels, $L_0, L_1, \dots, L_7$. As depicted in Figure 59 on the top, increasing illumination does not result in uniform intensity distribution across the surface, leaving certain regions less illuminated or over-exposed compared to others. To address this, our approach segments the surface into regions and then determines the optimal illumination level for each. We annotated the dataset for scratch, deformation and spots defects, by drawing precise

polygons around the defective regions, using the LabelMe tool, as depicted in Figure 59 on the bottom rows, outlined with green.

The original images of 5472×3648 resolution were split into overlapping patches of size 512×512 . We used an overlap of 0.85, to generate multiple illumination angles for each image patch. To minimize class imbalance due to large background areas, non-defective patches were excluded. We split the resulting dataset D of 12200 images to 6455/1644/4101 train/validation/test ($D_{train}/D_{val}/D_{test}$). The dataset statistics are presented in Table 15.

Table 15: Multi-Illumination Dataset Statistics

Statistic	Value
Images per Level	12200
Total Images	97600
Image Size	512×512 px
Class Distribution	Bg 96.266%, Sc 1.780%
	Df 0.068%, Sp 1.885%
Train / Val / Test	6455 / 1644 / 4101

4.1.2.4 Architecture

We first present the defect segmentation approach adopted to establish a performance baseline. Subsequently, illumination control is formulated as a reinforcement learning (RL) problem, and the proposed deep reinforcement learning (DRL) framework is introduced

Defect Segmentation

Our study primarily focuses on optimizing illumination to improve segmentation accuracy. Let $F(I; \theta)$ denote the segmentation network, which maps an input image I of size $H \times W \times 3$ to per-pixel class scores y of size $H \times W \times N$. Pixel-wise class probabilities are obtained via a softmax operation, and the predicted segmentation map is given by $S(x, y) = \arg \max_n y(x, y, n)$, with $n = 0, \dots, N - 1$

In this study, DeepLabv3 [48] and MA-Net [49] were adopted to implement the segmentation function $F(\cdot)$. The initial objective was to analyse the impact of varying illumination conditions on segmentation accuracy, establishing a baseline aligned with current industrial practices that deploy high-performing networks under fixed illumination settings. To this end, eight different models were trained for each network, namely $M_{L_0}, M_{L_1}, \dots, M_{L_7}$ where each model was trained using images acquired under the corresponding illumination level. In addition, one model, M_{L_c} , was trained using images from all illumination levels combined.

Since the development of a state-of-the-art segmentation architecture is not within the scope of this work, two well-established architectures were selected to represent core mechanisms commonly reported in the semantic segmentation literature. Specifically, the DeepLab family of architectures, widely used as baselines, introduced the Atrous Spatial Pyramid Pooling (ASPP) module, which applies atrous convolutions to efficiently capture multi-scale contextual information by enlarging the receptive field using different atrous rates. The ASPP module is further enhanced with global average pooling to incorporate global context information. MA-Net follows an encoder–decoder architecture inspired by U-Net, while integrating self-attention mechanisms, namely the Position-wise Attention Block (PAB) for modeling spatial dependencies and the Multi-scale Fusion Attention Block (MFAB) for capturing channel-wise dependencies. Representative encoder backbones were used, namely ResNet-50 for DeepLabv3 and EfficientNet-B5 for MA-Net. Both encoders are computationally efficient, which is a key requirement in industrial inspection scenarios. Heavier architectures did not demonstrate substantially improved performance, rendering the trade-off between accuracy and computational cost impractical.

RL problem formulation

The illumination control problem is formulated as a Markov Decision Process (MDP) defined by the tuple (S, A, P, r, γ) , where S and A denote the state and action spaces, respectively. P denotes the state transition probability function, r is the reward function, and γ in the range $[0, 1]$ is the discount factor.

Illumination is treated as a discrete control variable with eight predefined levels L_0 to L_7 , derived from the acquisition setup. The action set is defined as $A = \{-1, 0, 1\}$, where -1 decreases the illumination level by one step, 0 maintains the current level, and $+1$ increases the illumination level by one step. Let s_t in the range 0 to 7 denote the current illumination index. State transitions are defined as $s_{t+1} = \min(7, \max(0, s_t + a_t))$, where L_{s_t} corresponds to the applied physical illumination

PROJECT 101178230 – ENCIRCLE

level. This discrete formulation reflects the structure of the collected dataset, while the proposed framework remains compatible with continuous or digitally addressable illumination devices, as demonstrated in [61].

To define the reward function, we first introduce the evaluation metric $E^F(S, \hat{S}_t)$, where S and $\hat{S}_t$ denote the predicted and ground-truth segmentation maps, respectively. This metric quantifies segmentation performance under a given illumination state. For each sample, E_j^F is computed for all illumination levels $j = 0, \dots, k - 1$, and the maximum achievable performance is defined as:

$$E_{max} = \max_j E_j^F$$

This value corresponds to the best achievable segmentation performance for the given sample. At time step t , the reward signal encourages the agent to select the illumination level whose segmentation performance matches the optimal value:

$$r_t = \begin{cases} +1, & \text{if } E_t^F = E_{max}, \\ -1, & \text{otherwise.} \end{cases}$$

The objective is to learn a policy that maximizes the expected discounted return:

$$J(\pi_\theta) = E_{\pi_\theta} \left[\sum_{t=0}^T \gamma^t \cdot r_t \right]$$

A neural network is used to learn the policy $\pi_\theta(a_t | s_t)$, where θ denotes the trainable parameters. The network receives an observation s_t at time step t and selects an action according to:

$$a_t \sim \pi(\cdot | s_t)$$

DRL Framework

Proximal Policy Optimization (PPO) [58] is used to optimize the policy parameters by maximizing the expected return. Given an advantage estimate $\hat{A}_t$ (computed with GAE [62]), PPO defines the probability ratio

$$r_t(\theta) = \frac{\pi_\theta(a_t | s_t)}{\pi_{\theta_{old}}(a_t | s_t)}.$$

To prevent excessively large policy updates, PPO maximizes a clipped surrogate objective:

$$L^{clip}(\theta) = \hat{E}_t \left[\min(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon) \hat{A}_t) \right].$$

The method follows an actor-critic formulation, where the critic predicts the state value $V_\theta(s_t)$. The full optimization combines the policy objective, a value regression term, and an entropy regularizer:

$$L(\theta) = \hat{E}_t \left[L^{clip}(\theta) - c_1 L_t^{VF}(\theta) + c_2 H[\pi_\theta](s_t) \right],$$

where the value-function loss is

$$L_t^{VF}(\theta) = \left(V_\theta(s_t) - V_t^{\text{target}} \right)^2,$$

and the policy entropy is

$$H[\pi_\theta](s_t) = - \sum_a \pi_\theta(a | s_t) \log \pi_\theta(a | s_t),$$

which regularizes the policy update, ensuring a balance between exploitation and exploration.

The coefficients c_1 and c_2 weight the value and entropy terms respectively. We use the standard PPO hyperparameters: $\epsilon=0.1$, $c_1=0.5$, and $c_2=0.01$. The basic algorithmic steps of the proposed DRL-based multi-illumination optimization are shown in the algorithmic steps presented in Figure 60.

Algorithm 1 DRL-based Multi-Illumination Optimization

```

TRAIN_SEG( $\mathcal{D}_{train}$ ):
  Train segmentation models  $F_j$  for all illumination levels  $L_j$ .
TRAIN_PPO( $\mathcal{D}_{val}$ ):
  Precompute best illumination per sample:
  for each  $(I, \hat{S}) \in \mathcal{D}_{val}$ :
     $E_j^F \leftarrow E^F(F_j(I), \hat{S})$ 
     $E_{max}(I) \leftarrow \max_j E_j^F$ 
  for iteration = 1, ...,  $M$ :
    Collect rollout of length  $T_{batch}$ :
    for  $t = 0, \dots, T_{batch} - 1$ :
       $a_t \sim \pi_{\theta_{old}}(\cdot | s_t)$ 
       $s_{t+1} \leftarrow \min(7, \max(0, s_t + a_t))$ 
       $r_t \leftarrow \begin{cases} +1 & \text{if } E_{s_{t+1}}^F(I) = E_{max}(I) \\ -1 & \text{otherwise} \end{cases}$ 
    Compute returns  $R_t$  and advantages  $\hat{A}_t$  (GAE)
    Update policy and value function using PPO
     $\theta_{old} \leftarrow \theta$ 
INFER( $I$ ) for each  $I \in \mathcal{D}_{test}$ :
   $s_0 \leftarrow$  initial illumination state
  for  $t = 0, \dots, T_{infer} - 1$ :
     $a_t \sim \pi_{\theta}(\cdot | s_t)$ 
     $s_{t+1} \leftarrow \min(7, \max(0, s_t + a_t))$ 
  return  $F_{s_{T_{infer}}}(I)$ 

```

Figure 60: DRL-based Multi-Illumination Optimization.

The PPO algorithm was employed to train a DRL agent, based on the framework illustrated in Figure 61. The proposed DRL framework is built upon the typical RL training scheme, where the environment is represented by the set of images captured under different illumination levels $L_0, L_1, \dots, L_7$. The DRL agent is implemented based on an Actor-Critic network. At each step, it selects an action that adjusts the illumination level and the environment returns a new observation corresponding to the resulting illumination level.

As illustrated in the training block in Figure 61, outlined with dots pattern, the new observation is processed by a neural network $F(\cdot)$ which interprets the image data into a reward signal r_t that is provided as feedback to the agent. This interpretation is achieved using the evaluation metric E^F which compares the predictions against the ground truth. In our case, the comparison employs the widely recognized metric, mean Intersection over Union (mIoU). This approach ensures that the agent receives rewards for generating outputs that result in optimal illumination levels, where the network $F(\cdot)$ achieves its highest segmentation performance.

During inference, each image is presented to the agent with an initial illumination state s_t . The agent iteratively selects illumination adjustments over T_{infer} steps, as illustrated in the inference block in Figure 61, outlined with crossed-line pattern.

In our case, we fixed the inference horizon to $T_{infer}=10$, to ensure the agent can explore all eight illumination levels and maintain the decision, including extreme cases, e.g. $L_0 \leftrightarrow L_7$. This final state $s_{T_{infer}}$ is then passed to the segmentation model which generates the defect segmentation map. The performance is compared against baseline models operating under fixed illumination conditions, as well combination strategies which aggregate segmentation outputs across all illumination levels.

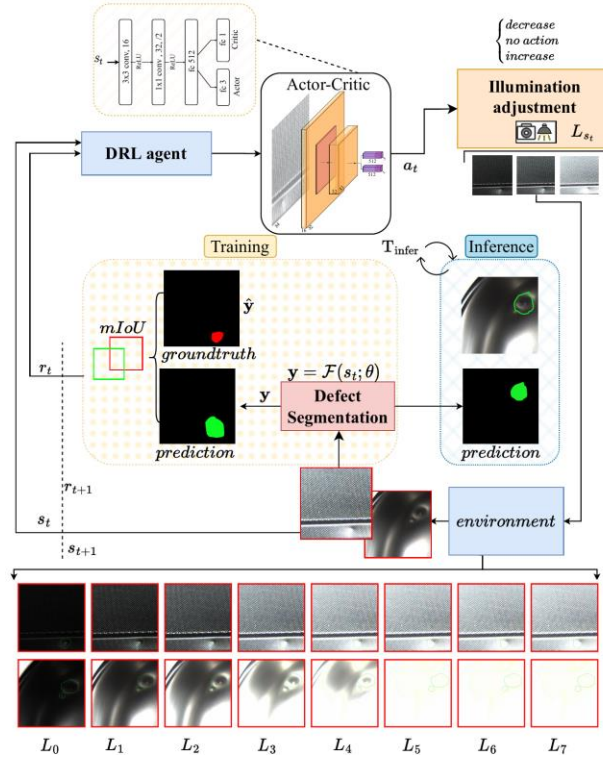


Figure 61: DRL agent training and inference framework for illumination control. During training the segmentation performance is used to reward the agent. During inference the agent determines the final decision after T interactions with the environment.

The architecture of the Actor-Critic network is depicted in Figure 61, outlined with diagonal line pattern. We adopted a simple CNN network structure since more complex architectures did not yield significant performance gain. As illustrated, the two networks, the actor and the critic, share the same parameters of the two convolutional layers and the first fully connected layer. The last fully connected layer of critic outputs the value function estimate, while for the actor, the network's outputs are passed through a SoftMax function creating a categorical distribution to compute the action probabilities. The agent then samples an action from the distribution, $a \sim \text{Categorical}(p_1, p_2, p_3)$, where p_i is the probability of selecting action a_i .

4.1.2.5 Experimental Setup

Our approach was implemented using PyTorch 2.4 library and Python 3.9 environment on a system with Nvidia GeForce RTX 3080Ti GPU 16GB, i9-12900H CPU and 32GB RAM. For the defect segmentation, we trained both Deeplabv3 and MA-Net with Adam optimizer on D_{train} , with the input images resized to 256×256 . The initial learning rate was set empirically to 0.001 with a step decay factor of 0.8 every 7 epochs. Early stopping was employed if segmentation accuracy on the validation set did not improve for 10 consecutive epochs. The model with the best performance on D_{val} was selected to generate the reward signals during the DRL training. The same model was evaluated as the baseline on D_{test} .

For the DRL training, the agents were trained for $M=3 \times 10^6$ iterations on D_{val} and policy rollout batch of $T_{batch}=2000$ using the Adam optimizer with initial learning rate set to 0.00025 and linear decay. Training proceeds in an episodic manner, where each episode involves a different image from the D_{val} , with the agent exploring illumination transitions for each image. Episode length was restricted to 10 steps, providing a bounded rollout horizon and ensure PPO rollouts cycle through the validation set, rather than accumulating long trajectories on a single image. Early stopping was applied when the performance did not improve for 1×10^5 iterations. Input images were resized to 84×84 , the default resolution for the

PROJECT 101178230 – ENCIRCLE

OpenAI Gym framework [63], since increasing the resolution did not increase performance. The PPO model with the highest performance on D_{val} was selected for inference on D_{test} . The total training time, including segmentation training for all illumination levels and the DRL optimization phase, was 482 hours.

4.1.2.6 Experimental Results

To evaluate the DRL agent's performance in improving baseline, we compared it against two approaches, fixed illumination and voting over predictions. In the fixed illumination approach each of the trained models M_{L_i} is evaluated in each of the illumination levels L_i during inference. The voting approach serves as a baseline combination strategy to test whether aggregating predictions from all illumination levels can improve performance. Within this approach, we examined two strategies, standard majority voting L_{MV} , which uses standard majority voting from the predictions across all illumination levels and foreground-voting, L_{FV} which prioritizes foreground predictions.

For L_{FV} , let $S_{ij}(x, y)$ be the predicted class at pixel (x, y) of model M_{L_i} , under illumination L_j , and $P_{ij}(x, y, c)$ its softmax score. Defective vote counts are $n_c^{(i)}(x, y) = \sum_{j=0}^7 1(S_{ij}(x, y) = c), c > 0$.

$$S_{FV}^{(i)}(x, y) = \begin{cases} \underset{c > 0}{\operatorname{argmax}} n_c^{(i)}(x, y), & (i) \\ \underset{j}{\operatorname{argmax}} P_{ij}(x, y, S_{ij}(x, y)), & (ii) \\ 0, & (iii) \end{cases}$$

where (i) is used when a unique defective class achieves the highest vote count, (ii) when multiple defective classes tie and the maximization over j is restricted to illumination levels whose predicted label belongs to the tied defective classes (iii) when all predictions correspond to background.

For L_{MV} , we use the same vote counts $n_c^{(i)}(x, y) = \sum_{j=0}^7 1(S_{ij}(x, y) = c)$ computed over all classes c , and define

$$S_{MV}^{(i)}(x, y) = \underset{c}{\operatorname{argmax}} n_c^{(i)}(x, y).$$

with ties broken by random selection among the tied classes.

Additionally, our approach is compared with Double Deep Q-Network (DDQN) approach (Hasselt et al.), which has been widely used in the literature as an improvement over DQN, mitigating the problem of overestimating the Q-values. The results are presented in Table 2, where bold text indicates the best performance, and underlined text highlights the next three best performances. We use the harmonic mean of IoU (hIoU) (Xian et al.) as the main evaluation metric, but we also report mIoU, to align with standard practices. The hIoU metric for C number of classes is defined as

$$hIoU = C / \sum_{c=1}^C \frac{1}{IoU_c}$$

In the segmentation task with the fixed illumination setting, MA-Net generally outperforms DeepLabv3, achieving an improvement of 6.3%. Specifically, the best performing model M_{L_1} trained with DeepLabv3, achieves 26.0% in hIoU under L_1 level, while with MA-Net, it achieves 32.3% under level L_2 . For the models $M_{L_3} \dots M_{L_C}$, it is evident that training becomes inefficient when images captured under levels higher than L_3 are used, as saturated images dominate the dataset, hindering the learning process.

Among the voting approaches, L_{MV} performs poorly across all models because it aggregates both foreground and background predictions. Since background predictions dominate across illumination levels, the few illumination settings that correctly detect defects are outvoted. Conversely, L_{FV} , which restricts voting to foreground predictions, achieves more stable results across all models.

Arch.		Illum. Level	Segmentation Models								
			M_{L_0}	M_{L_1}	M_{L_2}	M_{L_3}	M_{L_4}	M_{L_5}	M_{L_6}	M_{L_7}	M_{L_c}
DeepLabv3	Fixed	L_0	0.184	0.150	0.081	0.026	0.010	0.000	0.000	0.000	0.193
		L_1	0.192	0.260	0.207	0.098	0.084	0.010	0.001	0.001	0.141
		L_2	0.142	0.141	0.132	0.119	0.088	0.029	0.020	0.021	0.166
		L_3	0.076	0.155	0.121	0.147	0.075	0.058	0.067	0.055	0.101
		L_4	0.048	0.126	0.074	0.152	0.042	0.058	0.060	0.047	0.088
		L_5	0.036	0.095	0.041	0.121	0.031	0.085	0.074	0.033	0.078
		L_6	0.032	0.032	0.030	0.101	0.030	0.073	0.066	0.017	0.066
		L_7	0.032	0.019	0.015	0.078	0.031	0.075	0.051	0.010	0.061
	Vot.	L_{FV}	0.207	0.154	0.244	0.242	0.199	0.181	0.103	0.069	0.269
		L_{MV}	0.067	0.123	0.069	0.120	0.038	0.045	0.046	0.033	0.092
	RL	L_D	0.189	0.322	0.253	0.128	0.079	0.121	0.048	0.025	0.224
		L_P	0.199	0.354	0.303	0.169	0.101	0.075	0.044	0.027	0.236
MA-Net	Fixed	L_0	0.215	0.168	0.099	0.000	0.000	0.000	0.000	0.002	0.244
		L_1	0.233	0.318	0.250	0.179	0.165	0.003	0.005	0.005	0.273
		L_2	0.180	0.323	0.289	0.140	0.168	0.046	0.021	0.004	0.265
		L_3	0.063	0.176	0.137	0.092	0.116	0.097	0.043	0.011	0.117
		L_4	0.035	0.104	0.099	0.077	0.109	0.103	0.067	0.057	0.112
		L_5	0.031	0.083	0.093	0.065	0.090	0.114	0.085	0.114	0.118
		L_6	0.024	0.048	0.071	0.063	0.085	0.098	0.090	0.108	0.120
		L_7	0.014	0.034	0.048	0.059	0.081	0.093	0.089	0.098	0.112
	Vot.	L_{FV}	0.221	0.197	0.268	0.221	0.234	0.147	0.167	0.008	0.333
		L_{MV}	0.039	0.109	0.071	0.047	0.065	0.062	0.035	0.009	0.078
	RL	L_D	0.214	0.330	0.295	0.240	0.282	0.123	0.087	0.027	0.251
		L_P	0.251	0.385	0.358	0.253	0.288	0.184	0.102	0.097	0.271

Figure 62: Comparison results (harmonic IoU) between the proposed DRL approach and the baseline models.

For the RL approaches, the results presented in the Figure 62 reflect the mean performance across all possible initial states. The proposed $L_{PPO}(L_P)$ outperforms the other approaches, both with DeepLabv3 and MA-Net. For the well performing models $M_{L_0}, M_{L_1}, M_{L_2}$, and using DeepLabv3, it improves performance by 0.7%, 9.4%, 9.6% in hIoU respectively. With MA-Net, the improvements are 1.8%, 6.2%, 6.9%. In Table 16 a summarized overview of the comparative results is presented, where we report mIoU and hIoU for the best models. Notably, while MA-Net generally outperforms DeepLabv3 under fixed illumination, our DRL-based approach allows DeepLabv3-PPO to achieve 3.1% higher hIoU compared to MA-Net-Fixed, demonstrating potential for enhancing weaker networks through illumination optimization. Finally, L_{PPO} outperforms $L_{DDQN}(L_D)$, demonstrating more stable and superior performance across different levels. This highlights the advantage of policy-gradient approaches in complex environments of high-dimensional inputs, such as the one examined in our study.

Table 16: Summarized comparative results.

Method	mIoU	hIoU
DeepLabv3-Fixed	0.415	0.260
DeepLabv3-DDQN	0.462	0.322
DeepLabv3-PPO	0.489	0.354
MA-Net-Fixed	0.457	0.323
MA-Net-DDQN	0.453	0.330
MA-Net-PPO	0.512	0.385

PROJECT 101178230 – ENCIRCLE

Finally, the complete pipeline requires approximately 1.4 seconds per sample for image splitting, illumination selection, and patch-level defect segmentation. However, due to suboptimal hardware setup, the cameras and illumination system require a total of 13.6 seconds to capture the multi-illumination sequence, while manual inspection in our application takes at least 20 seconds. This indicates that the algorithmic part is lightweight, while improved hardware can substantially reduce end-to-end latency.

The visualization results from the best performing model M_{L_1} of MA-Net-PPO are presented in Figure 63. In the top four rows, the agent's transitions are visualized, for test set images I_0, I_1 over time steps $t=0, 1, \dots, 10$, starting from the initial state L_7 . The predicted maps are illustrated below (Figure 63) each corresponding row.

The green colour pixels of the semantic maps correspond to the deformation class, the red to the scratch category and blue to the spot category. This example demonstrates how the agent determines the optimal level when high-intensity images are presented. In contrast, the bottom four rows of Figure 63 demonstrate how the agent determines the optimal action, when low-intensity images from the L_0 are presented for samples I_2 and I_3 .

As illustrated, the agent effectively adjusts illumination for each image, reducing false positives by lowering the light intensity when saturation occurs, while maintaining sufficient light intensity for detecting the defect features. Our experiments focused on defect segmentation on image patches, and we restricted the evaluation to patches that contain a defective region alongside background, ensuring that illumination effects are assessed directly on the regions of interest. Even with this setting, background pixels still dominate the dataset. The broader problem of data imbalance when analysing the complete surfaces was beyond the scope of this study. However, to demonstrate the broader applicability of our approach, we provide two examples showcasing the state-transition heatmaps of the agent's decisions, when inspecting complete samples, as depicted in Figure 64.

The samples are sink surfaces of varying geometries and finishing textures. During inspection, each sample is captured under 8 illumination levels, segmenting it into patches and allowing the agent determine the optimal illumination level for each patch. The sample of Figure 64 (a) is initially captured under level L_7 , at $t=0$. At $t=10$, the agent decides the final states as shown in the consolidated view of Figure 64 (b), with the ground truth defective regions outlined in red rectangles. For each image patch fed to the agent, the GradCAM (Selvaraju et al.) heatmap is generated to provide visual explanation of the agent's decisions. In Figure 64 (c) the consolidated view of the GradCAM heatmaps is illustrated, highlighting the salient features that the agent focuses on.

The heatmap at $t=0$ in Figure 64 (c) shows that the agent focuses on the saturated regions when selecting its next action. By $t=10$ (Figure 64 (d)) the agent has reduced the illumination, which in turn suppresses the saturation. The heatmap indicates that both saturated areas and regions with uneven illumination, such as the defective ones outlined in red rectangles, strongly influence the agent's decisions. In Figure 64 (e), the sample is initially captured under L_0 . We notice that while the lower part of the sample's surface is adequately illuminated, the upper part remains under-illuminated. After the agent's adjustment at $t=10$ (Figure 64 (f)), the well-lit areas are preserved, whereas the underexposed regions become properly illuminated. This behaviour is confirmed by the heatmaps in Figure 64 (g) and (h). Although the agent was trained on patch-level defect segmentation, its decisions on the full surface examples still concentrate on regions that matter for defect visibility, indicating that the learned policy generalizes beyond the patch-based segmentation setup.

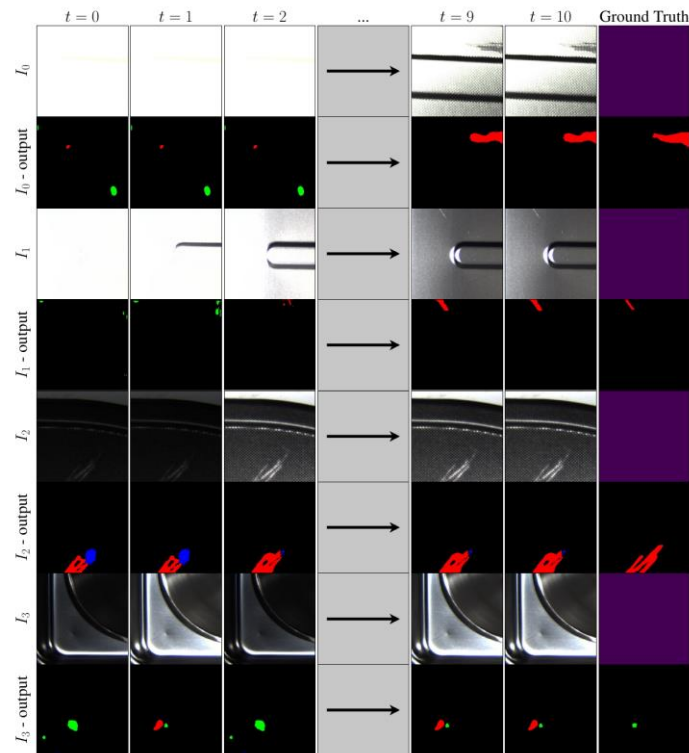


Figure 63: Iterative multi-illumination optimization starting from the initial states L_7 (top 4 rows) and L_0 (bottom 4 rows).

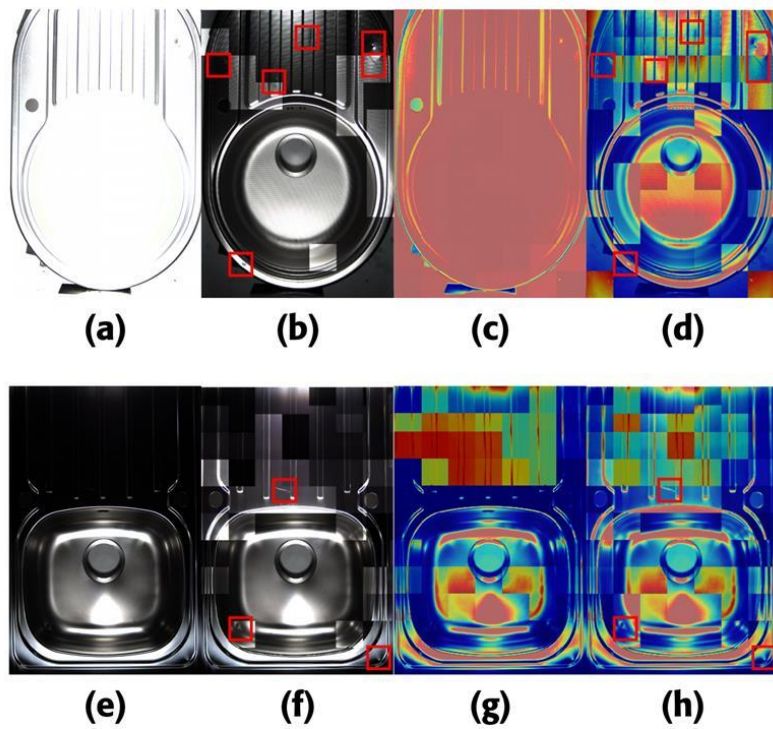


Figure 64: State-transition heatmaps of agent decisions. In (a), (b), (e) and (f) two test samples at $t=0$ and $t=10$ are illustrated, with the respective heatmaps in (c), (d), (g) and (h).

The described investigations demonstrate the effectiveness of a DRL-based illumination adjustment strategy as a means of improving image quality and segmentation performance on reflective metallic surfaces. Using a multi-illumination dataset on steel samples, the analysis shows that illumination conditions constitute a dominant source of performance variability across different segmentation architectures, in contrast to approaches that focus exclusively on network design. The results indicate that a PPO-based agent, trained with a segmentation-driven reward, can learn to select illumination levels that improve segmentation outcomes on a per-image basis. By operating at the level of image patches, illumination adjustments can be applied locally, allowing adaptation to spatial variations in surface properties and reflectance.

The observed performance gains across multiple segmentation architectures, including larger relative improvements for weaker baselines, highlight the role of illumination control as a general mechanism for enhancing visual informativeness rather than as an architecture-specific optimization. Execution speed is primarily constrained by acquisition hardware, suggesting that the approach is compatible with industrial inspection setups employing digitally controlled illumination systems. At the same time, the current formulation relies on discretized illumination levels applied to pre-acquired images, which limits coverage of complex geometries such as curved or recessed surface regions. Addressing these limitations requires the investigation of continuous, digitally addressable illumination control, along with corresponding adaptations of the action space and reward formulation to mitigate oscillatory behaviour during inference.

The insights obtained from these investigations are applicable to aluminium recycling scenarios, where inspection objectives focus on the visual differentiation of scrap categories and contaminants such as coatings, plastics, and residual paint rather than defect localization. The demonstrated relationship between illumination control, image quality, and segmentation stability can be transferred to aluminium inspection pipelines to improve the consistency and informativeness of visual data. In this context, illumination-aware image acquisition supports the generation of higher-quality visual representations that are required for the development and updating of digital twins in Task T3.3, particularly under the reflective, heterogeneous, and geometrically complex conditions characteristic of aluminium scrap processing.

4.1.2.7 Conclusion

The adaptive illumination and exposure control soft sensor significantly improved image stability under dynamic lighting conditions of ALUM industrial environment. Quantitative evaluation using harmonic IoU showed a consistent improvement of up to **15–20%** compared to static exposure baselines, particularly in scenes affected by high glare and saturation. The DRL-based controller converged rapidly, enabling real-time adjustment of camera parameters within a few interaction steps. This stabilization directly enhanced the performance and robustness of downstream vision models, including scrap detection and defect segmentation. The approach effectively decouples visual quality from environmental variability, ensuring reliable long-term deployment.

4.1.3 Heat demand prediction based on scrap images

4.1.3.1 Problem Formulation

This work targets the development of an image-driven soft sensor that estimates the mean heat demand of a furnace loading based on visual evidence of the scrap mixture during the upstream wagon loading stage. In the production workflow, a loading corresponds to the full sequence of scrap additions into the wagon until the wagon departs for the furnace. During this interval, the wagon is typically filled through multiple loading events, potentially spanning different scrap categories, and possibly including repeated batches of the same category (e.g., three consecutive batches of “dirty scrap”). The central objective is to learn a mapping from this multi-event visual description of a loading to a scalar target representing the expected energy consumption for that loading. To this end, we use the furnace heat-demand signal, defined as the percentage of the burner’s maximum output requested during that loading, as a robust proxy for energy consumption.

Formally, for each loading i , the dataset provides: (a) a set of wagon images captured during the loading process, (b) contextual information derived from a YOLO-based event detector that segments the process into loading events and associates them with scrap categories, and (c) furnace measurements collected from a SCADA interface (temperature, pressure, heat demand, natural gas consumption, etc.) as time series for multiple zones (Hot Chamber, Cold Chamber, Rotating Device), aligned to the loading via MQTT start/end timestamps and door-state-based loading detection. The soft sensor focuses on the wagon visual stream as input and uses the furnace-derived heat demand as supervision. In this

PROJECT 101178230 – ENCIRCLE

sense, the model aims to infer downstream energy requirements from upstream material appearance and composition cues, supporting near-real-time decision making and improved process understanding.

Each loading i is represented by a variable-length sequence of images:

$$X_i = (x_{i,1}, x_{i,2}, \dots, x_{i,N_i})$$

where N_i depends on the duration of the loading and the number of detected loading events/batches. Since loadings are not uniform, both in the number of events and in the mixture of scrap categories, the input is inherently set-/sequence-structured, and the learning problem must be permutation- and length-robust (i.e., able to produce a consistent loading-level prediction from a variable number of frames). The target for the same loading is a scalar:

$$y_i = \frac{1}{T_i} \sum_t^{t=T_i} \text{HeatDemand}_i(t)$$

where the averaging window corresponds to the furnace-side interval with duration T_i associated with loading i , as determined by the correlated timestamps (wagon MQTT triggers for “wagon leaves” and “new empty wagon arrives”, plus SCADA-derived loading boundaries from door state). This definition emphasizes that the prediction is loading-level rather than frame-level: individual images are partial observations of the full scrap mixture, while the outcome is an aggregate thermal requirement that emerges during furnace operation. The soft sensor is therefore formulated as a supervised regression problem with structured inputs:

$$f_{\theta}(X_i) \rightarrow \hat{y}_i$$

where f_{θ} implemented as a vision-based representation model followed by an aggregation and prediction head. Concretely, each image is processed by a backbone vision model to obtain patch-level embeddings, which are aggregated first within an image and then across images to produce a single loading descriptor. This descriptor is passed through an MLP regressor to estimate mean heat demand. The learning objective is to minimize a regression loss (e.g., MSE) between y_i and $\hat{y}_i$, while evaluation uses standard error metrics to quantify predictive accuracy and robustness across loadings that differ in scrap composition, batch structure, and operating conditions.

In summary, the problem is to predict a single downstream energy indicator per loading from multiple upstream images collected during that loading, with the dataset providing the necessary event structure and cross-stage timestamp correlation to define loadings consistently and to support reliable supervision. This formulation aligns with the operational goal of estimating furnace energy demand early, at the wagon stage, so that planning, control, and process optimization can be informed before the loading is fully processed in the furnace.

4.1.3.2 Dataset

The dataset used for developing and validating the ALUMIL soft sensors is a registered, loading-centric dataset composed of three coordinated sub-datasets, each capturing a different stage of the melting workflow. The common indexing key across all data sources is the notion of a loading, i.e., a complete charging cycle from “wagon starts being loaded” until the wagon leaves for the furnace and the corresponding furnace cycle begins and ends. This key corresponds to the timestamp when the furnace door opens for the scrap to be inserted for that specific loading. In the registered format, each loading is stored as one folder (e.g., `alumil_*`) containing three required subfolders:

- `scrap_wagon`
- `sensor_measurements`
- `dross_detected`

At the wagon stage, images of the scrap inside the wagon are captured continuously while scrap is loaded. The YOLO-based detector (Section 4.1.14.1.1) is used online to identify loading events and associated scrap categories for each event. A single “loading” typically contains multiple events, since different scrap categories can be added sequentially. Importantly, a category can also be loaded in multiple batches (e.g., three successive additions of “`alumil_scrap_sorting`”), which means the dataset naturally reflects the real operational variability of charging composition over time. The dataset contains the following 10 scrap categories, corresponding to different sources and qualities of scrap:

1. `alumil_bars`
2. `alumil_dust`
3. `alumil_returns`
4. `alumil_returns_painted`

PROJECT 101178230 – ENCIRCLE

5. barbed_wire
6. circles
7. cleanser
8. coil_wire
9. cubes
10. oman

To make the data tractable and consistent for model training, the registered dataset applies a deterministic frame-selection rule per loading: frames are sorted temporally and, for each scrap category observed in that loading, the dataset keeps (i) the first frame where the category appears, and (ii) the last frame before a new category appears (if applicable). This preserves the key transitions in loading composition while avoiding redundant near-identical frames. Labels such as “empty” and “excavator” are ignored during selection to avoid non-informative detections. This selection process is shown in Figure 65, where only frames with new scrap categories is selected for loading `alumil_2025-05-02_18-27-34`.



Figure 65: Selected images for loading `alumil_2025-05-02_18-27-34` containing [`alumil_bars`, `alumil_returns`, `alumil_returns_painted`].

For each selected frame, annotations are stored (e.g., in JSON). When pixel-level crops are needed, the dataset extracts object-level crops using the annotated polygons (LabelMe-style) by converting polygons to bounding boxes and cropping exactly one crop per selected record (specifically for the category implied by the selection reason). This ensures that the model sees category-relevant visual evidence without exploding the number of training samples due to multi-object frames. In the selection of scrap

In parallel, furnace operating conditions are captured from a SCADA interface using an OCR pipeline (continuous stream). Loadings are detected at the furnace stage using operational signals such as door status (open/closed), which provides a robust delimiter for charging cycles. These furnace-side detected loadings are then correlated with wagon-side loadings through MQTT timestamps: when the wagon leaves towards the furnace, this marks the end of the wagon loading; when a new empty wagon arrives, this marks the start of the next one. These start/end markers allow alignment even when the two subsystems have different clocks or sampling rates.

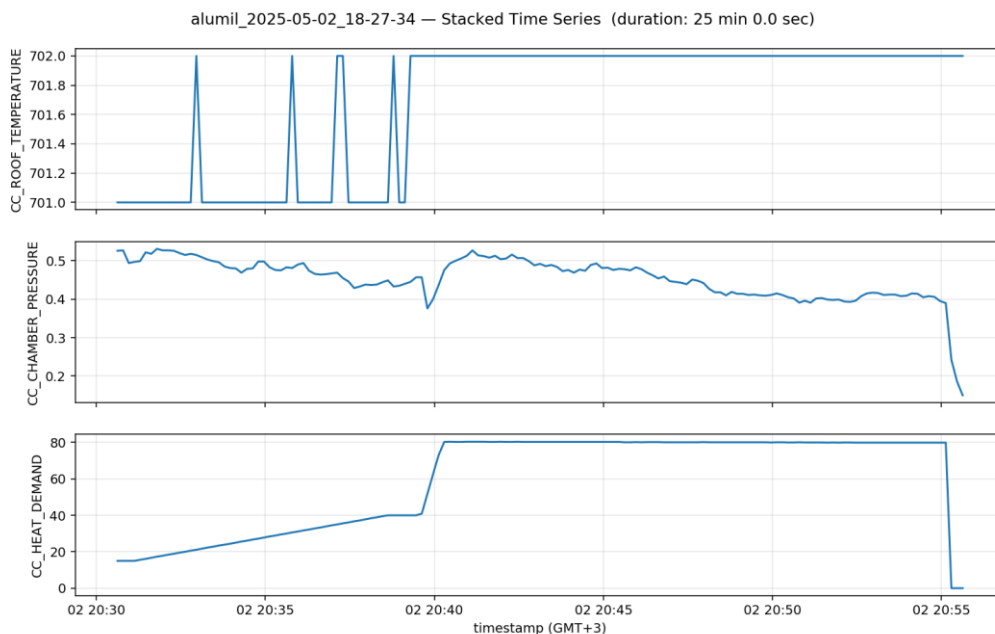


Figure 66: Sensor timeseries for loading alumil_2025-05-02_18-27-34 with the variables corresponding to the cold furnace: CC_ROOF_TEMPERATURE, CC_HEAT_DEMAND, and CC_CHAMBER_PRESSURE

The sensor sub-dataset contains measured variables including temperature, pressure, heat demand, and natural gas consumption, captured for: (i) the Hot Chamber (HC), (ii) the Cold Chamber (CC), and (iii) RD as time series, along with additional metadata stored alongside the measurements. In the registered dataset layout, these time series are stored under `sensor_measurements/` and typically include a `timeseries_data.csv` file per loading. When loaded for ML, the dataset identifies the timestamp column, sorts samples by time, selects numeric measurement columns (either user-specified or inferred), and produces tensors (t, X) where time is converted to relative seconds from the start of the sequence. Timespans with Nan sensor values are filled using linear interpolation, while loadings with count of NaN values over a threshold are filtered out to avoid training on invalid targets. In Figure 66, the timeseries corresponding to the cold furnace variables for loading alumil_2025-05-02_18-27-34 are shown.



Figure 67: Characteristic internal furnace camera frames selected for loading alumil_2025-05-02_18-27-34.

A third sub-dataset comes from an internal furnace camera focused on the cold chamber. Images are captured as a time series and a YOLO-based model detects dross formation over time. These detections are stored per loading under `dross_detected/` as time-stamped records (e.g., JSON + images). To reduce redundancy while preserving dynamics, the registered dataset selects characteristic dross frames based on relative change: it keeps frames where the absolute percentage change of the dross estimation exceeds a threshold (default 20%), while also always keeping the first and last

frames to retain temporal context. In Figure 67, the selection of dross frames for loading `alumil_2025-05-02_18-27-34` is shown, while in Figure 68 the corresponding timeseries of detected dross formation is visualized.

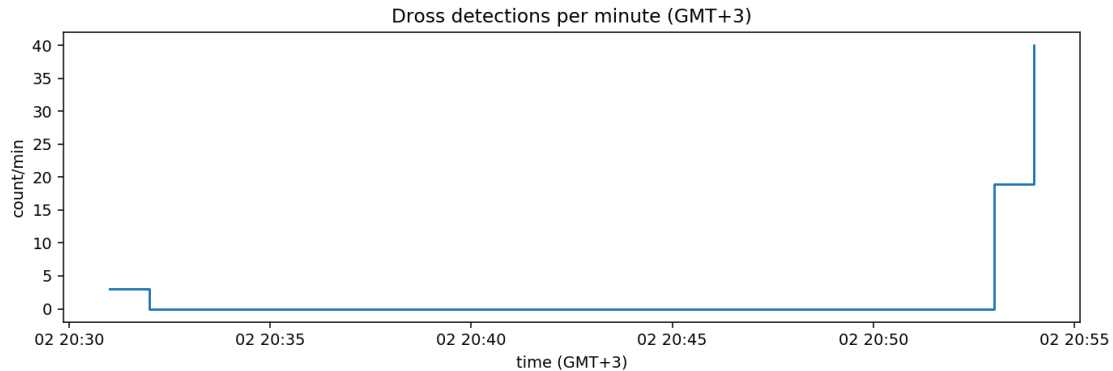


Figure 68: Dross detected timeseries for loading `alumil_2025-05-02_18-27-34`.

In total, the dataset contains 120 loading events. Of these, 93 loadings are complete, i.e., they include all expected sub-folders, while the remaining loadings contain missing entries (e.g., absent folders and/or images). Overall, the dataset comprises 2,639 scrap_wagon images and 8,685 dross_detected images.

The key strength of this dataset is that it is explicitly organized per loading, enabling direct supervision for soft sensors that operate at the same granularity. This is particularly important for the target soft sensor in this deliverable: predicting the heat demand or dross for a loading from the scrap-wagon images associated with that loading. The folder-based structure (one folder per loading, consistent subfolders, and aligned timestamps across stages) makes it straightforward to: (a) aggregate all evidence available for a loading, (b) compute loading-level targets (e.g., mean heat demand or dross over the furnace interval), and (c) train and evaluate models consistently across modalities.

4.1.3.3 Architecture

The proposed soft sensor is implemented as a three-stage pipeline that maps scrap images from a wagon loading to a single continuous estimate of the mean heat demand for that loading. The stages are: (i) pre-processing, (ii) backbone feature extraction, and (iii) aggregation & regression (Figure 69).

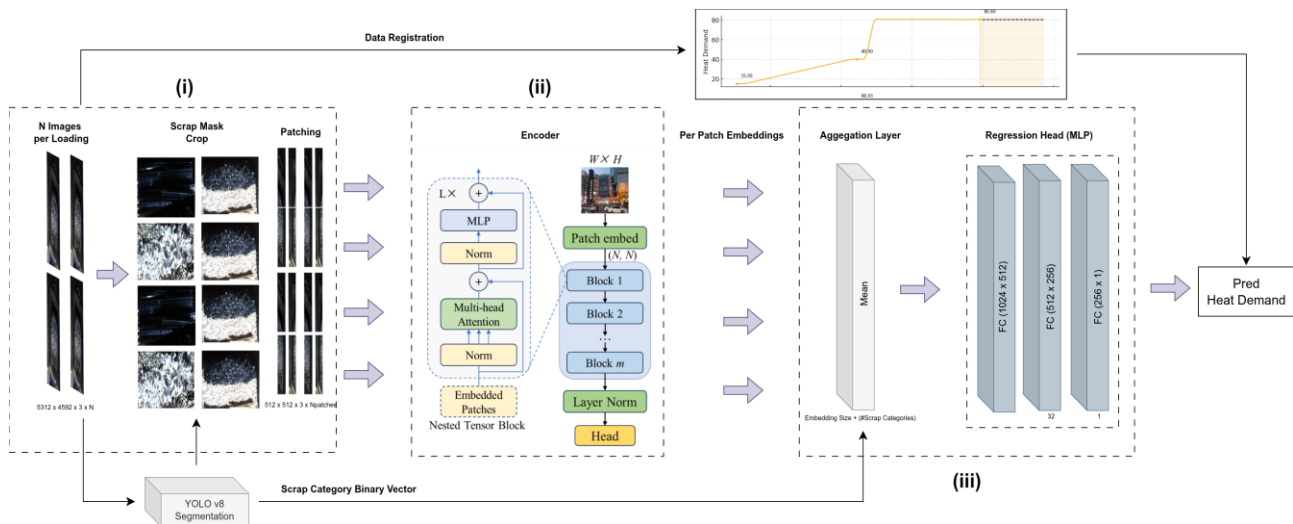


Figure 69: Proposed model architecture for heat demand prediction. (i) Pre-processing pipeline including masking based on scrap segmentation masks and patching, (ii) per-patch embeddings generated from backbone visual model, (iii) aggregation and regression head for final heat demand prediction.

Stage 1: Pre-processing

PROJECT 101178230 – ENCIRCLE

Each loading i is associated with a variable number of images $\{x_{(i,1)}, x_{(i,2)}, \dots, x_{(i,N_i)}\}$ captured during scrap deposition events. Raw images often contain irrelevant background (wagon borders, surroundings, operators/equipment), which can dilute the visual signal correlated with the furnace energy footprint. To focus on the material region, the pipeline uses the YOLOv8-based segmentation model (Section 4.1.1) to produce segmentation masks per image i per scrap category j :

$$M_{(i,j)} = YOLO-Seg(x_i)$$

The masks are used to derive a region of interest (ROI) that tightly encloses the detected scrap region. The original image is then cropped (and optionally padded to preserve aspect ratio) to produce a scrap-focused view as shown in Figure 70:

$$\tilde{x}_{(i,j)} = Crop(x_i, M_{(i,j)})$$



Figure 70: Examples of cropped images focused on individual scrap categories for loading alumil_2025-05-02_18-27-34.

Each cropped image $\tilde{x}_{(i,j)}$ is resized to the resolution required by the vision backbone (depending on configuration) and normalized to the backbone's expected input distribution. Finally, $\tilde{x}_{(i,j)}$ is patchified into fixed-size non-overlapping patches of size $P \times P$:

$$\tilde{X}_i \rightarrow \{\tilde{x}_{(i,1,1)}, \dots, \tilde{x}_{(i,j,n_p)}\},$$

$$n_p = \left(\frac{H}{P}\right) \left(\frac{W}{P}\right)$$

where H and W are the cropped image dimensions. This patch sequence $\tilde{X}_i$ is the input representation consumed by the vision backbone.

Stage 2: Vision backbone patch embeddings

A pretrained foundation vision model (e.g., DINOv2 ViT-S/14 [69][70], DINOv2 ViT-L/14 [69][70], or SAM2 ViT-H [64][71] variants) is used to compute latent representations for each patch. For each image patch $\tilde{x}_{i,j}$, the backbone produces a set of patch embeddings:

$$E_{i,j} = \{e_{(i,j,1)}, e_{(i,j,2)}, \dots, e_{(i,j,n_p)}\},$$

$$e_{(i,j,n_p)} \in R^D$$

where D is the embedding dimensionality defined by the backbone architecture. In all cases, the backbone stage outputs dense, high-level feature descriptors that encode visual cues such as scrap texture, reflectance patterns, fragment size distribution, contamination indications, and structural heterogeneity, factors that can correlate with downstream melting behaviour and energy requirements.

Stage 3: Aggregation and regression

Because the prediction target is defined per loading, the model must convert variable-length image sequences (and variable numbers of patches per image) into a single fixed-size loading representation. This is done via hierarchical pooling. Patch-to-image aggregation Patch embeddings are first pooled into an image-level descriptor using mean aggregation:

$$v_{(i,j)} = \frac{1}{n_p} \sum_{n=1}^{n_p} e_{(i,j,n)}$$

The image embeddings belonging to a loading are then aggregated (average pooling) to form the loading-level descriptor:

$$v_i = \frac{1}{j} \sum_{n=1}^j v_{(i,n)}$$

In addition, we evaluate an attention-based aggregation variant that replaces average pooling with a learnable weighted combination of image embeddings:

$$\begin{aligned} s_{(i,j,n)} &= w^T \tanh(W e_{(i,j,n)} + b) \\ a_{(i,j,n)} &= \frac{\exp(s_{(i,j,n)})}{\sum_{m=1}^{n_p} \exp(s_{(i,j,m)})} \\ v_{(i,j)} &= \sum_{n=1}^{n_p} a_{(i,j,n)} e_{(i,j,n)} \end{aligned}$$

Where:

- $e_{(i,j,n)}$ is the n -th patch embedding of image j in loading i ,
- $s_{(i,j,n)}$ is a scalar attention score,
- $a_{(i,j,n)}$ are normalized attention weights over patches,
- $v_{(i,j)}$ is the resulting image-level descriptor.

This attention mechanism is similarly utilized to produce the loading-level descriptor by weighing the image-level descriptors $v_{(i,j)}$. Both aggregation methods, average pooling and attention, produce robust representations even when the number of images per loading varies due to operational differences (e.g., more deposition events, repeated batches of the same scrap category). Optionally, a binary scrap category embedding c_i is concatenated to v_i , containing information about the scrap categories present in loading i to produce the final image embedding $\tilde{v}_i$:

$$\begin{aligned} c_i &\in \{0,1\}^{10} \\ (c_i)_k &= \begin{cases} 1, & \text{if category } k \text{ is present in loading } i \\ 0, & \text{otherwise} \end{cases} \\ \tilde{v}_i &= [v_i ; c_i] \end{aligned}$$

Finally, the loading embedding $\tilde{v}_i$ is passed through a lightweight MLP regressor that outputs the predicted mean heat demand:

$$\hat{y}_i = \text{MLP}(\tilde{v}_i)$$

The MLP consists of one or more fully-connected layers with non-linear activations, with hidden dimensionalities explored experimentally (e.g., 512, 1024, 2048). The output is a single scalar corresponding to the estimated mean heat demand associated with the furnace processing window aligned to that loading.

4.1.3.4 Experimental Setup

All experiments were conducted on an on-premise server equipped with two NVIDIA GeForce RTX 3080 Ti GPUs (12 GB VRAM each), an Intel Core i9-13900K CPU, and 32 GB system RAM. To maximize throughput, the training workflow was configured to run one training task per GPU, enabling two experiments to execute concurrently (parallel runs across the two GPUs). The operating system was Ubuntu 22.04.5 LTS (x86_64).

The experimental pipeline was implemented in Python 3.10 and executed using a CUDA-enabled deep learning stack. GPU drivers and runtime were configured as:

- NVIDIA Driver: 550.144.03
- CUDA runtime: 12.4

PROJECT 101178230 – ENCIRCLE

- PyTorch: 2.4.1
- xFormers: 0.0.28

The preprocessing and data extraction modules rely on the following frameworks:

- YOLOv8 (Ultralytics) for scrap-region detection/segmentation and loading-event processing.
- PaddleOCR for extracting furnace/SCADA variables from the SCADA interface stream.
- TensorBoard for training monitoring, loss curves, and early stopping diagnostics.

The dataset was partitioned using a random split into 60% training, 20% validation, and 20% testing. Each sample corresponds to a single furnace loading, for which approximately 3 to 10 wagon-stage images are available. The prediction target is Cold Chamber (CC) heat demand, aggregated into a single scalar per loading by computing the mean heat demand over the loading window (as defined by the loading time boundaries). Each experiment implements a soft-sensor architecture consisting of: (i) a vision backbone, (ii) an embedding aggregation operator, and (iii) a regression head. In all experiments, the backbone produces per-patch embeddings which are aggregated using either average pooling or attention. Backbone usage policy:

- DINOv2: evaluated in frozen and fine-tuned configurations.
- SAM2: evaluated in frozen-only configuration.

The regression head is a multilayer perceptron (MLP) with hidden dimensions:

[1024, 512, 256]

and a final scalar output predicting the mean CC heat demand for the loading.

All models were trained with a consistent optimization setup to ensure fair comparison across backbones and feature variants:

- Loss function: Mean Squared Error (MSE)
- Optimizer: Adam
- Learning rate: constant $lr=0.001$
- Batch size: 1 (loading-level batches; each batch may contain 3–10 images internally)
- Max epochs: 100
- Early stopping: patience = 3 epochs (based on validation performance)

The experimental campaign evaluated a set of progressively richer predictors for estimating the mean CC heat demand per loading. As a lightweight baseline, a category-only MLP was trained using a binary vector indicating the presence/absence of each of the 10 scrap categories in a loading, with three capacity variants (hidden size 512/1024/2048). For image-driven inference, three backbone families were tested:

- DINOv2 ViT-S/14
- DINOv2 ViT-L/14
- SAM2 (ViT-H)

All were used as feature extractors whose outputs were aggregated and passed to a shared regression head. For each backbone, two variants were considered: (i) image embedding only, and (ii) image embedding concatenated with the category vector to form a multimodal representation prior to regression. In addition, a domain-adapted DINOv2 ViT-S/14 backbone, obtained via self-supervised pretraining [95] on the scrap imagery corpus, was evaluated both frozen and in combination with category metadata to quantify the impact of domain-specific representation learning. Across all configurations, the downstream regressor was implemented as an MLP head applied to the aggregated loading-level embedding to output a single scalar prediction of mean heat demand.

Mean Squared Error (MSE) measures the average of the squared differ. Finally, we evaluated the predictive performance of the soft sensor using three standard regression metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).

MSE measures the average squared difference between the predicted and true values::

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

Mean Absolute Error (MAE) measures the average absolute difference between predictions and ground truth: By squaring the errors, MSE penalizes large deviations more strongly than small ones, making it sensitive to outliers. This is desirable when occasional large mistakes have a high operational cost (e.g., underestimating peak energy demand). However, MSE is expressed in squared units, which can be less intuitive to interpret.

MAE measures the average absolute error:

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

Root Mean Squared Error (RMSE) is the square root of MSE: MAE grows linearly with the error magnitude, so it is less dominated by outliers than MSE. Since it is reported in the same units as the target variable, it is easy to interpret as the “typical” prediction error, making it well-suited for assessing day-to-day accuracy.

RMSE is the square root of MSE:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}$$

RMSE retains MSE’s emphasis on large errors while converting the result back to the original units, improving interpretability. In practice, RMSE is useful when large deviations matter, and a noticeable gap between RMSE and MAE often indicates the presence of a few large prediction errors.

4.1.3.5 Experimental Results

Table 17 reports test-set performance for the soft sensor that predicts mean CC heat demand per loading from scrap-wagon imagery (optionally augmented with a binary scrap-category vector). “Category vector MLP” models are metadata-only baselines where the binary category vector is passed through an MLP regressor. “Backbone X” models use a pre-trained vision backbone as a feature extractor, where patch embeddings are aggregated to an image-level embedding and then to a loading-level embedding (baseline average pooling), which is fed to an MLP regressor. “Backbone X + Category vector” models concatenate the loading-level visual embedding with the category vector prior to regression. “Backbone X finetuned” models use a backbone that was self-supervised pretrained and finetuned on the scrap imagery corpus before being trained/evaluated on the heat-demand prediction task. Finally, “Backbone X + attention” models replace the baseline average pooling with an attention-based aggregation mechanism (learned weighting of patch/image embeddings) to form the fixed-size loading representation.

Table 17: Experimental results of the heat demand soft sensor on the test dataset

Model	MSE	MAE	RMSE
Category vector MLP (512)	277.747	13.714	16.665
Category vector MLP (1024)	222.972	12.510	14.932
Category vector MLP (2048)	236.461	12.944	15.377
DINOv2 ViT-S/14	176.323	10.909	13.279
DINOv2 ViT-S/14 + Category vector	179.447	11.045	13.396
DINOv2 ViT-L/14	175.057	10.615	13.231
DINOv2 ViT-L/14 + Category vector	188.752	11.316	13.74
DINOv2 ViT-S/14 finetuned	182.990	10.928	13.527
DINOv2 ViT-S/14 finetuned + Category vector	327.82	14.554	18.105
DINOv2 ViT-S/14 finetuned + attention	173.803	10.770	13.183

PROJECT 101178230 – ENCIRCLE

SAM2 ViT-H	178.226	10.899	13.350
SAM2 ViT-H + Category vector	177.817	10.865	13.334
SAM2 ViT-H + attention	297.597	12.491	17.251

Across all configurations, the main trend is clear: incorporating visual information via a pretrained vision backbone yields markedly better accuracy than relying on category presence alone, indicating that appearance cues (e.g., scrap texture, fill level, compaction, contamination) carry predictive signal beyond the categorical composition of a loading. The best category-only baseline is the MLP with width 1024 (MSE 222.972, MAE 12.510, RMSE 14.932), while increasing capacity to 2048 does not improve performance, reinforcing that the limiting factor is the information content of the categorical input rather than head size.

When replacing category-only inputs with image-derived embeddings, performance improves substantially across all backbones. Among the mean-pooling models, DINOv2 ViT-L/14 achieves the strongest MAE (10.615) with competitive MSE/RMSE (175.057 / 13.231), while DINOv2 ViT-S/14 is very close (176.323 / 10.909 / 13.279), suggesting that much of the relevant visual signal can be extracted even with a smaller backbone under the current training regime. SAM2 ViT-H is similarly competitive in its baseline form (178.226 / 10.899 / 13.350), confirming that its frozen representations also capture scrap-wagon characteristics useful for downstream heat-demand regression.

The optional category-vector concatenation does not consistently improve results and in several cases slightly degrades them. For DINOv2, adding the category vector worsens performance for both ViT-S/14 and ViT-L/14, while for SAM2 it is effectively neutral (178.226 vs 177.817 MSE). This pattern suggests that category information may already be implicitly encoded in the visual descriptor, or that simple concatenation is not an effective fusion strategy in the current setup (e.g., limited data per loading, imperfect scaling/normalization between modalities, or insufficiently expressive fusion layers).

Domain-specific SSL adaptation and aggregation strategy show mixed but informative effects. The finetuned DINOv2 ViT-S/14 (mean pooling) does not outperform the off-the-shelf DINOv2 ViT-S/14 (182.990 vs 176.323 MSE), indicating that the present SSL configuration and data volume have not yet produced features that transfer better to the heat-demand target under the baseline pooling regime. However, introducing attention-based aggregation with the finetuned DINOv2 backbone yields the lowest MSE and RMSE in the table (173.803 MSE, 13.183 RMSE), suggesting that learned weighting during aggregation can better exploit the adapted patch-level representations when forming the loading-level descriptor. Notably, this configuration does not achieve the lowest MAE (10.770 vs 10.615 for DINOv2 ViT-L/14), but it provides the best overall performance in squared-error terms.

In contrast, attention aggregation is not universally beneficial: SAM2 ViT-H + attention performs substantially worse than its average pooling counterpart (297.597 MSE, 17.251 RMSE), indicating that the attention module (as implemented/trained here) may be overfitting, miscalibrated, or otherwise poorly matched to SAM2's embedding characteristics and the available training data. This divergence highlights that attention-based aggregation is sensitive to backbone representation properties and training stability, and it should be validated per backbone rather than assumed to improve performance by default.

Finally, the finetuned DINOv2 ViT-S/14 combined with the category vector performs markedly worse than all other configurations (327.820 MSE, 18.105 RMSE). This suggests an unstable or poorly calibrated multimodal fusion in that setting and motivates revisiting normalization, fusion design, and training hyperparameters before drawing conclusions about the value of category augmentation for finetuned visual features.

4.1.3.6 Conclusion

Overall, the results support the use of a pretrained visual backbone as the primary source of predictive signal for mean CC heat demand per loading, with DINOv2 ViT-S/14 with attention pooling currently providing the best performance. Category-only models remain useful as lightweight baselines, but they leave a noticeable accuracy gap, and naïve multimodal fusion via concatenation does not yet provide consistent gains in the evaluated settings. Conclusively, the heat demand soft sensor demonstrates that furnace energy requirements can be accurately estimated directly from scrap images. By combining visual segmentation, historical time-series in-furnace condition data and learned image embeddings, the model captures the relationship between scrap composition and thermal needs. This enables early prediction of energy demand before scrap melting starts. As a result, the approach supports better energy planning and improved process efficiency.

4.1.4 Dross prediction based on scrap images

4.1.4.1 Problem Formulation

Predicting the amount of dross (waste) produced during the aluminium recycling process is a highly complex problem. This complexity stems primarily from the high variability of the raw material. Aluminium scrap comes in many different categories, shapes, and conditions—ranging from clean industrial cuts to painted or dirty mixed metals. Each of these categories interacts differently during the melting process, making it difficult to estimate the final yield based on simple rules. Therefore, relying only on the weight or the general type of the scrap is not enough. It is crucial to accurately predict the expected dross percentage before the material is charged into the furnace. This pre-furnace prediction is important for two main reasons: **Operational Efficiency:** If operators know the quality of the scrap in advance, they can adjust the furnace settings (such as temperature or flux addition) to minimize material loss. **Production Planning:** Accurately estimating how much usable metal will be recovered helps in better inventory management and economic assessment of the batch.

4.1.4.2 Dataset

The dataset constructed for this study consists of 93 distinct loading events, accumulating a total of 215 processed images. To efficiently manage the raw data streams and eliminate redundant information, a custom data ingestion pipeline (Alumil-LoadingDataset) was developed to filter and preprocess the inputs from the scrap wagon and dross detection systems. For the scrap classification task, a content-based filtering mechanism is applied to the raw image sequence. Frames are not selected uniformly, instead, an image is retained only if its corresponding annotation introduces a new label (unique object instance) that was not present in the preceding frames. This ensures that the model is trained on diverse object configurations rather than repetitive frames. Additionally, frames containing specific exclusion labels (e.g., "excavator", "empty") are automatically discarded. To standardize the visual input, all selected scrap images undergo a fixed geometric transformation pipeline consisting of a -13° rotation to correct camera alignment, followed by a static center crop (ROI) to remove background noise and focus on the wagon interior. A more comprehensive analysis of this specific cropping procedure is provided in "Stage 1: Pre-processing" of Section 4.1.2.2 Architecture (Heat demand prediction based on scrap images). For the dross detection module, a variance-based sampling strategy is employed. Images are selected based on the rate of change in the dross estimation value. A frame is retained only if the absolute percentage change in dross estimation exceeds a threshold of 20% compared to the previously selected frame. This strategy, combined with the mandatory inclusion of the first and last frames of each sequence, ensures that the dataset captures the most significant phases of dross evolution. Representative examples of the processed images are illustrated in Figure 71.





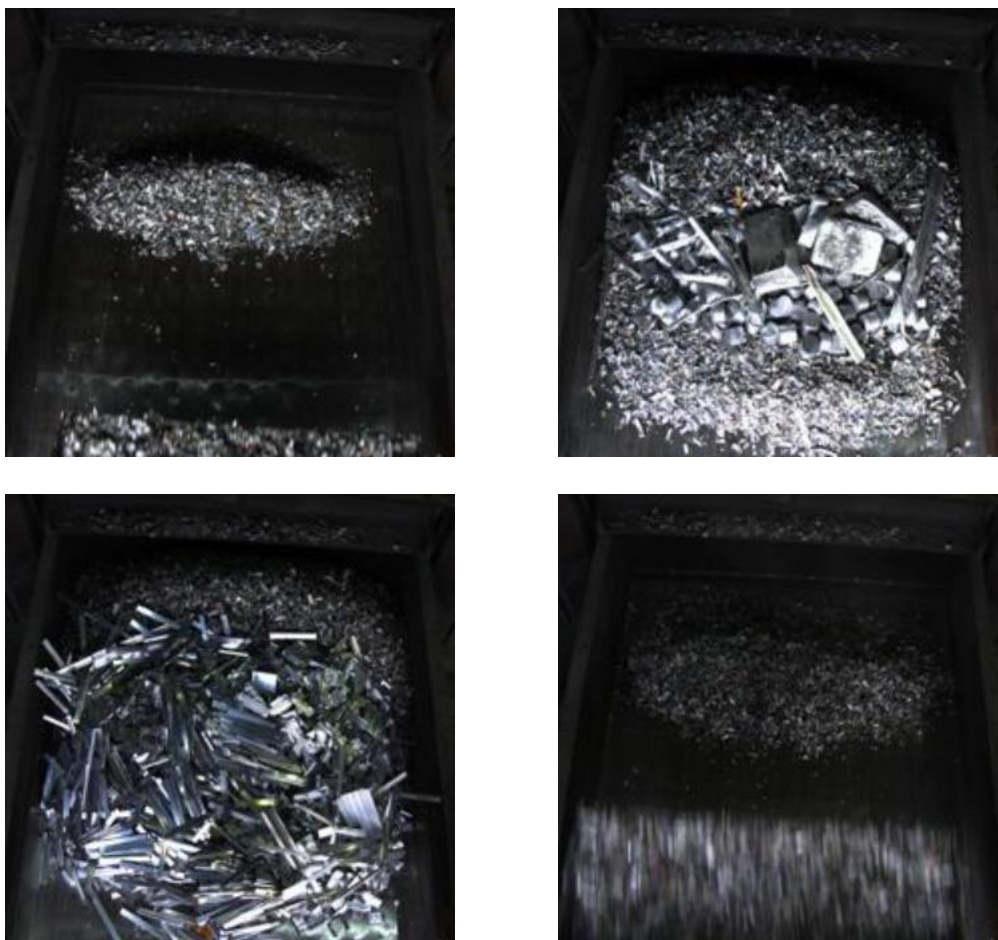


Figure 71: Representative samples from the processed dataset, illustrating the geometric transformations (rotation and cropping) applied to scrap images and the selected frames for dross detection.

4.1.4.3 Architecture

In the present study, DINOv2 is employed as a pretrained and frozen visual backbone for feature extraction. Image frames capturing the wagon during the loading process, containing different scrap categories as well as the wagon carriage, are provided as input to the model. For each image, both global image embeddings and patch-level embeddings are extracted, encoding complementary semantic and spatial visual information. For each loading instance, multiple image frames are available. The corresponding patch-level and global embeddings are aggregated by computing their mean values across all frames belonging to the same loading. The resulting averaged embeddings form a compact representation of each loading and are associated with a target dross percentage. These representations are subsequently used to train regression models, including an MLPRegressorTorch (a Multi-Layer Perceptron regressor), to estimate the dross content per loading.

DINOv2 is employed as a pretrained and frozen visual backbone for feature extraction. Image frames capturing the scrap wagon during the loading process—containing various scrap categories and the wagon carriage itself—serve as input to the model. For each image, the model extracts both global image embeddings (representing the whole scene) and patch-level embeddings (encoding fine-grained spatial and textural details).

Since a single loading instance consists of multiple image frames, we aggregate the extracted features to create a unified representation. Specifically, the corresponding global and patch-level embeddings are averaged across all frames belonging to the same loading sequence. These resulting averaged embeddings form a compact representation of the specific

PROJECT 101178230 – ENCIRCLE

scrap batch and are associated with a ground-truth dross percentage value measured at the furnace. Subsequently, these representations are used to train an MLPRegressorTorch (a Multi-Layer Perceptron regressor) for approximately 100 epochs to estimate the dross content per loading.

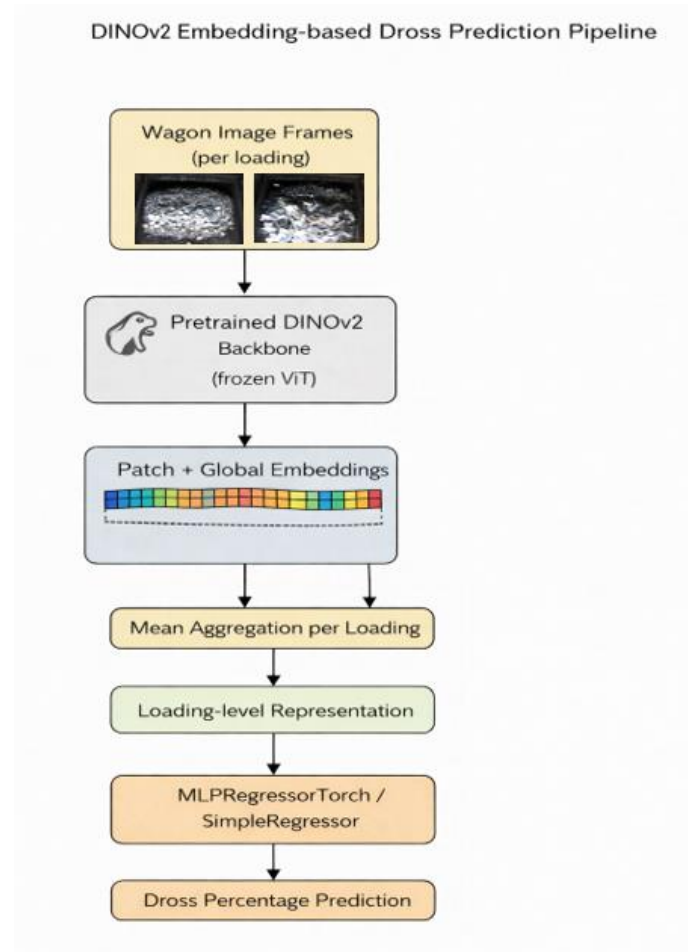


Figure 72: DINOv2-based Embedding Pipeline for Dross Percentage Estimation

Experimental Setup To determine the optimal input configuration for the regression model, five distinct feature combinations were evaluated:

1. **Mean Global Embeddings:** The average of the global DINOv2 embeddings across all frames of a loading. This captures the overall semantic "gist" of the scrap wagon.
2. **Vector Category:** A structural vector representing strictly the *types* of scrap categories present in the wagon (without visual features). This serves as a baseline to check if the visual data adds value beyond simple categorization.
3. **Concatenate Embeddings (Mean + Vector):** A fusion of the Mean Global Embeddings and the Scrap Category Vector, aiming to combine visual semantics with categorical knowledge.
4. **Mean Patch Embeddings:** The average of the high-dimensional patch-level features. These capture detailed local textures but result in a much larger feature vector.
5. **Concatenate All:** A comprehensive combination of Mean Global, Mean Patch, and Vector Category inputs.

4.1.4.4 Experimental Setup

All experiments regarding visual soft sensors and dross prediction were conducted on an on-premise high-performance server designed to support intensive deep learning workloads. The workstation is equipped with an Intel Core i9-13900K CPU and 32 GB of system RAM. To accelerate training and inference processes, the system utilizes two NVIDIA GeForce

PROJECT 101178230 – ENCIRCLE

RTX 3080 Ti GPUs (12 GB VRAM each). The training workflow was optimized to maximize throughput by configuring distinct training tasks per GPU, enabling the parallel execution of experiments. The operating system used was Ubuntu 22.04.5 LTS (x86_64). The experimental pipeline was implemented in Python 3.10, leveraging a CUDA-enabled deep learning stack configured as follows:

- **Deep Learning Framework:** PyTorch 2.4.1
- **GPU Drivers:** NVIDIA Driver 550.144.03
- **CUDA Runtime:** 12.4
- **Acceleration Libraries:** xFormers 0.0.28
- **Monitoring:** TensorBoard (for loss curves and early stopping diagnostics)

To encode the visual information from the detected scrap regions into high-dimensional vectors for regression, the DINOv2 architecture was employed. Specifically, the `dinov2_vits14` model (Vision Transformer Small, patch size 14) was used to extract features from the scrap wagon image directories. The feature extraction process involves two distinct strategies:

- **Global Embeddings:** The CLS token is extracted to capture the holistic semantic context of the image.
- **Patch Embeddings:** The Patch tokens are extracted and flattened to retain fine-grained spatial details.

Since a single loading event consists of multiple images, an aggregation step is performed. The embeddings (both global and patch-level) are averaged across all images within a specific loading folder to produce a single mean embedding vector (e.g., `mean_embedding.npy`), which serves as the input for the regression model. All images are normalized using standard ImageNet mean and deviation values prior to processing. To predict the final dross value based on the extracted embeddings, a custom Multi-Layer Perceptron (MLP) was implemented in PyTorch. The regression network (`MLPRegressorTorch`) consists of the following layers:

- **Input Layer:** Accepts the aggregated embedding vector (Dimension N).
- **Hidden Layers:** Two dense layers with 64 and 32 units respectively, each followed by a ReLU activation function.
- **Output Layer:** A final linear transformation to 1 unit, representing the predicted dross value.

The dataset was split into a training set (70%) and a test set (30%). The model was trained for 100 epochs using the Adam optimizer with a learning rate of 0.001 and the Mean Squared Error (MSE) loss function. Evaluation Metrics The performance of the regression model is evaluated using the following metrics: RMSE (Root Mean Squared Error), MAPE (Mean Absolute Percentage Error), provides the error as a percentage. SMAPE (Symmetric MAPE), used to adjust for scale differences. Results, including loss progression and true vs predicted value comparisons, are logged automatically to Excel files for further analysis.

4.1.4.5 Experimental Results

The quantitative results of the `MLPRegressorTorch` are summarized in Table 18 and the training dynamics are illustrated in Figure 73 and Figure 74. As observed in the results, the Vector Category configuration achieved the lowest error on the test set, with a Root Mean Square Error (RMSE) of 0.169. This indicates that, currently, the explicit knowledge of scrap categories is a strong predictor of dross yield. However, the Mean Global Embeddings (pure visual data) performed competitively with an RMSE of 0.178, demonstrating that the DINOv2 backbone successfully captures relevant material properties from images alone, without manual category inputs.

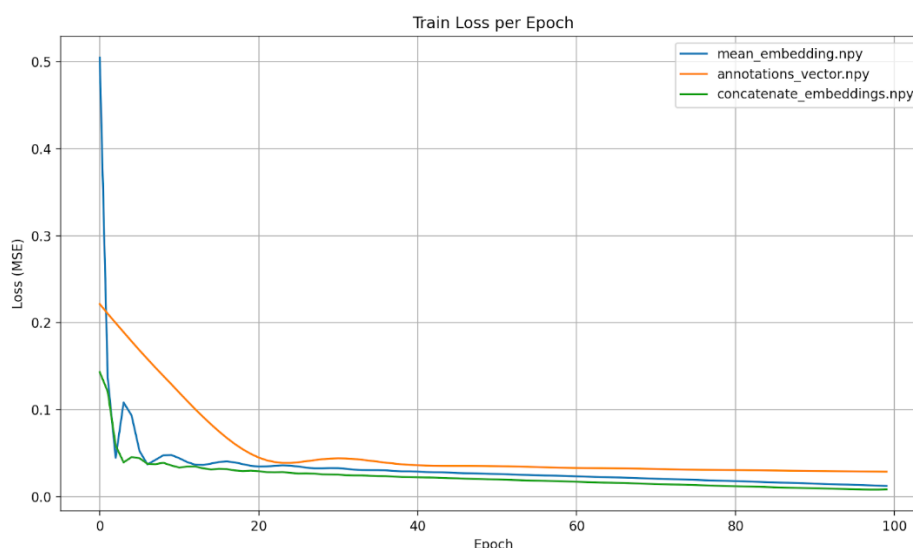
Critically, the experiments reveal a trend of overfitting when model complexity increases.

- Configurations involving Patch Embeddings (Mean Patch and Concatenate All) achieved extremely low MSE losses on the *training* dataset (e.g., 0.00003 for Mean Patch), as shown in the training curves.
- However, this did not translate to the *test* dataset, where errors increased significantly (RMSE > 0.21).

This discrepancy suggests that the high dimensionality of patch embeddings introduces noise that leads the MLP to memorize the training data rather than generalize. Consequently, the most robust approach for this application appears to be the use of Vector Categories or Mean Global Embeddings, which provide a balance between predictive accuracy and model generalization.

Table 18: Experimental Evaluation of Vision-Based Dross Prediction Models

Model	Best MSE loss (train dataset)	RMSE (test dataset)	Best MSE (test)
Mean Global Embeddings	0,00213	0,1781	0,0317
Vector Category	0,0286	0,169	0,0287
Concatenate Embeddings (Mean + Vector)	0,00961	0,190	0,0363
Mean Patch Embeddings	0,00003	0,222	0,049
Concatenate all (Mean Global+Mean Patch+Vector Category)	0,00182	0,213	0,045


Figure 73: Training Loss Evolution (MSE) across 100 Epochs for Different Input Configurations.

The dross prediction soft sensor extends visual analysis toward estimating material losses during scrap recycling procedure. Using self-supervised visual features, the model remains robust mixed scrap mixes, varied furnace conditions, and diverse operating procedures, such as chamber door openings. The predicted dross values provide early insight into yield efficiency which in turn supports more informed operational decisions and improved material recovery.

4.1.4.6 Conclusion

The vision-based dross prediction soft sensor utilized self-supervised DINOv2 embeddings to estimate dross surface coverage during the scrap melting process. Quantitative evaluation demonstrated stable low prediction error, with RMSE values consistently improving compared to baseline configurations. The approach proved resilient to difficult visual cues such as reflections, occlusions, and evolving molten surface textures. By enabling continuous estimation of dross, the soft sensor supports timely operational decisions, such as optimized skimming actions. This contributes directly to improved material yield, reduced waste, and enhanced energy efficiency.

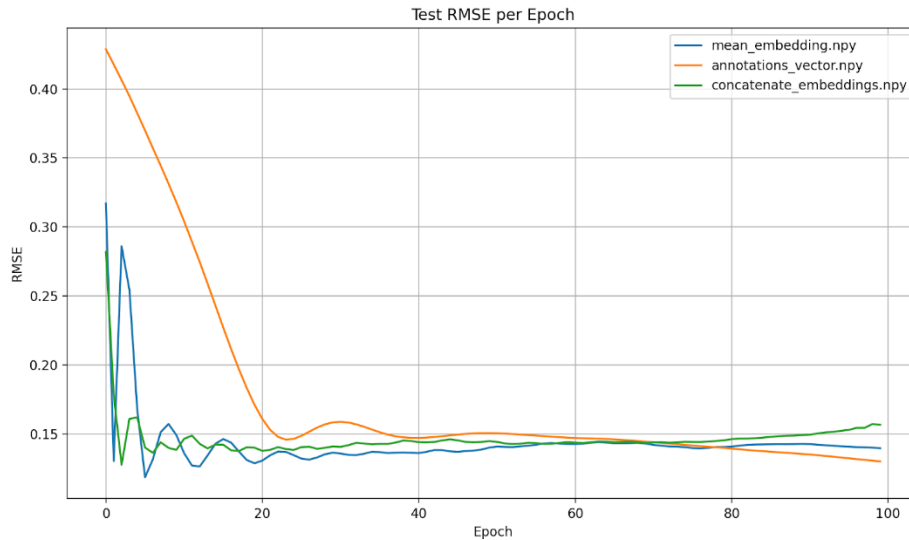


Figure 74: RMSE across 100 Epochs for Different Input Configurations.

4.1.5 Diffusion-based image generation of internal furnace dross formation

4.1.5.1 Problem Formulation

To explore whether modern diffusion models can synthesize realistic furnace interior imagery without any explicit conditioning on process parameters, we conducted a set of unconditional image generation experiments using two widely adopted latent diffusion backbones: Stable Diffusion 1.5 (SD 1.5) [88] and Stable Diffusion XL (SDXL) [65]. The objective was to adapt each model to the visual domain of real dross images captured from inside the furnace, and then evaluate, qualitatively, how closely the generated samples match the appearance, texture, and lighting characteristics of real frames.

4.1.5.2 Dataset

The dataset used in this study corresponds to the dataset described in Section 4.1.3.2, but restricted to only the dross images (i.e., the furnace interior frames where dross is visible), 8685 in total. These images capture the characteristic visual conditions of the furnace environment, including strong thermal glow, low contrast regions, smoke-like structures, and specular reflections.

4.1.5.3 Architecture

Both evaluated models (SD1.5 and SDXL) follow the Latent Diffusion Model design, where images are compressed into a latent space via a VAE and diffusion is learned in that latent representation. To make fine-tuning feasible on limited GPU resources, we applied LoRA-based adaptation to the denoising backbone of each model, enabling efficient domain specialization while updating only a small number of trainable parameters.

4.1.5.4 Experimental Setup

The experimental setup (hardware resources, software frameworks, and supporting tools/libraries) is identical to Section 4.1.3.4. Fine-tuning was performed using a unified folder of furnace images, and since no manual captions were available, training was executed with a lightweight dummy prompt setup, effectively treating the task as domain adaptation rather than semantic prompt learning.

For evaluation, multiple synthetic samples were generated per model using empty prompts (unconditional sampling) combined with classifier-free guidance (CFG). Although CFG is typically used to improve prompt adherence, we observed that non-zero guidance values produced cleaner and more coherent samples even with empty prompts, suggesting a stabilizing effect on the sampling trajectory. The generated images were saved per model and compared side-by-side against randomly sampled real furnace frames.

PROJECT 101178230 – ENCIRCLE

4.1.5.5 Experimental Results

After training, we generated multiple samples from each adapted model using empty prompts (unconditional generation) and classifier-free guidance. Although guidance is traditionally associated with prompt-following behaviour, we observed that non-zero guidance values produced noticeably cleaner and more coherent results even with an empty prompt, suggesting that CFG still provides a useful regularization effect on the sampling trajectory in this setting. The generated images were saved per model and compared side-by-side against randomly sampled real furnace frames (Figure 75).

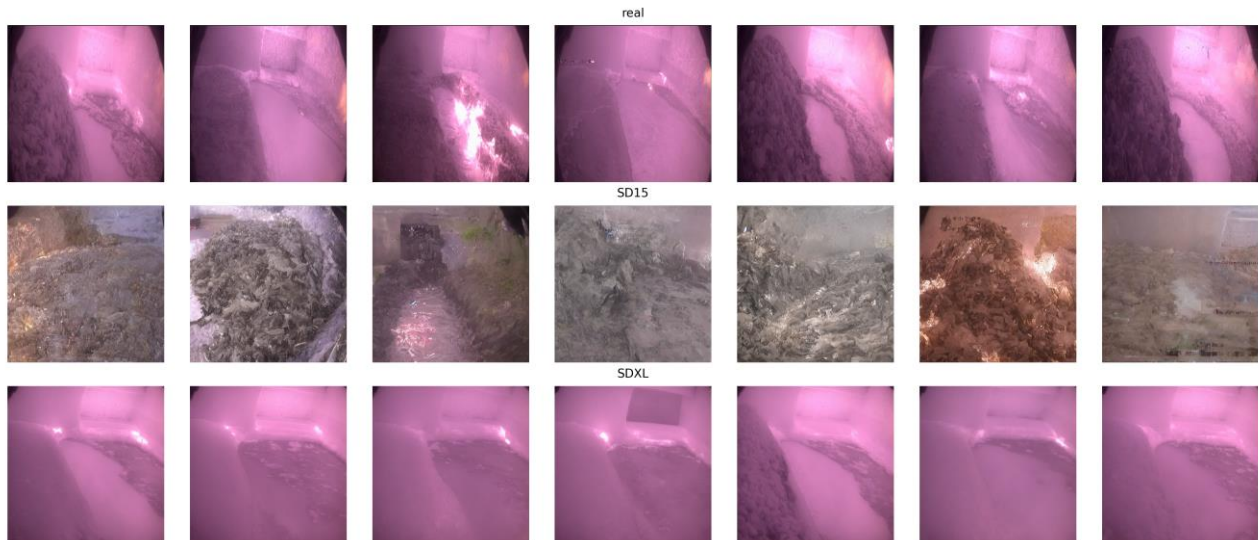


Figure 75: Qualitative comparison between real furnace dross images (top), images generated by fine-tuned Stable Diffusion 1.5 (middle), and images generated by fine-tuned Stable Diffusion XL (bottom). SDXL samples exhibit higher perceptual similarity to real images, with more coherent texture and improved sharpness.

Overall, SDXL produced the most convincing synthetic dross imagery, with outputs that were consistently sharper, more structured, and visually closer to real camera frames. Compared to SD 1.5, SDXL better preserved fine-grained texture, exhibited more realistic global contrast, and produced fewer artifacts or unstable patterns in darker regions, properties that are particularly important for furnace imagery, where detail is often subtle and dominated by low-frequency heat glow and reflections. In contrast, SD 1.5 was not able to learn the general color palette and geometry of the domain, and tended to generate images with lower fidelity microstructure and more obvious synthetic artifacts. This qualitative gap is consistent with SDXL's stronger denoising capacity and richer conditioning design, even when used in an effectively unconditional configuration.

A qualitative comparison of representative samples is shown in Figure 75, which juxtaposes real furnace images against samples generated by the fine-tuned SD 1.5 and fine-tuned SDXL pipelines. The figure highlights the improved perceptual realism achieved by SDXL, especially in edge sharpness, texture continuity, and the naturalness of thermal illumination.

4.1.5.6 Conclusion

The diffusion-based generative soft sensor enabled the synthesis of internal furnace images, augmenting real-world data for training and validation. Qualitative evaluation demonstrated that SDXL-generated samples exhibited higher perceptual fidelity and texture coherence compared to Stable Diffusion 1.5, particularly for complex dross patterns. This approach supports data augmentation, operator training, and scenario simulation without additional sensor installations.

4.2 ARCEL SOFT SENSORS

This subsection focuses on the Soft Sensor solutions developed and validated at the ARCELIK pilot site, addressing key challenges related to product quality monitoring and defect detection in industrial manufacturing processes. The proposed models exploit text and audio-based approaches to infer high-level defect characteristics, fault types, and anomalous operating conditions that cannot be reliably captured through conventional sensing alone. By integrating advanced AI models with real-time data acquisition, ENCIRCLE enables predictive maintenance and quality assurance via data-driven decision-

PROJECT 101178230 – ENCIRCLE

making. These soft sensors contribute to improved process reliability and reduced waste, while aligning with ENCIRCLE's broader objectives of sustainable manufacturing and circular production systems.

4.2.1 Defect Description Prediction

4.2.1.1 Problem Formulation

The core objective of the Defect Description soft sensor is to automatically predict the description of manufacturing defects by exploiting historical production and quality data, thereby reducing reliance on manual inspection and expert-driven characterization. From a modeling perspective, the problem is formulated as a multi-class classification task, where the goal is to predict the Defect Description label associated with a produced item. The prediction is conditioned on operational and contextual metadata, including product category, product model, defect type, and temporal information derived from the production date, which collectively encode process conditions, design variability, and operational dynamics. By learning latent relationships between these metadata attributes and observed defect patterns, the soft sensor aims to capture systematic dependencies between production configurations and quality outcomes. This formulation enables data-driven defect detection, supporting downstream quality control and decision support within the ENCIRCLE framework.

4.2.1.2 Dataset

The foundation of this experiment is a dataset of real-world production records (Figure 76, Table 19: Table 19) retrieved from the refurbishment center. This dataset contains granular operational details, including Product Categories (e.g., Category1, Category4) and specific service observations recorded during refurbishment activities. These attributes jointly capture variability in product configurations, usage history, and process execution, which are critical factors influencing defect manifestation. The dataset therefore provides a representative basis for learning statistically meaningful connections between operational data and defect descriptions.

ID	Operator	ReceiptNo	Model	Stock Code	Serial No	Category1	Category2	Category3	Category4	Production Date	Inspection Date	Process Date	Warranty Date	Entry Date	Days	Service Observation	Inspection Result	Fault	Fault Description
5	68593 mustafa zengin	47659863	AL 454 X (B6-AT 762713227)	235266609	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 10 / 23					01/11/2024, 08:07:54	23/10/2024, 00:02/07/2024, 00:04/24/10/2024, 15:44:51				113	Ünitesi bithlene uygun olarak ambaka ayan yapıldı	Door	Production error	
6	68594 mustafa zengin	477083459	BM 4044 (B7-BE 761571027)	231836112	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 12 / 23					01/11/2024, 08:08:37	23/10/2024, 00:10/09/2024, 00:04/25/10/2024, 15:18:39				43	Ünitesi bithlene uygun olarak ambaka Yan duvarlar hasarlı	Side walls	Shipping damaged	
																ambakalayararak teslim alınız.			
																Bilten No: QW/2023/004.SgNo:38472.23.07			
																470554588 nolu fişte üst tabla hasarlı çıkmıştır			
7	68596 mustafa zengin	47161613	6344 (B7-PROL 761421017)	241009101	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 01 / 24					01/11/2024, 08:09:59	15/08/2024, 00:12/07/2024, 00:04/20/08/2024, 15:55:27				34	Hasarlı çamaşır	Üst tabla darbeli	Damaged prod/Shipping damaged	
8	68597 mustafa zengin	47226021	6167 (PROLOGI 767240107)	248023105	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 05 / 24					01/11/2024, 08:10:44	19/08/2024, 00:31/07/2024, 00:04/20/08/2024, 15:55:27				19	Ünitesi bithlene uygun olarak ambaka Kapa ayan yapıldı	Hinge group	User -source	
9	68598 mustafa zengin	47454758	6444 OK (PROLOGI 763110177)	2454484108	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 08 / 24					01/11/2024, 08:12:02	24/10/2024, 00:17/10/2024, 00:04/30/10/2024, 10:49:25				9	Ünitesi bithlene uygun olarak ambaka Kapa ayan yapıldı	Door	Production error	
10	68599 mustafa zengin	472594389	6144 (B7-PROL 761401017)	241008301	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 01 / 24					01/11/2024, 08:12:49	14/08/2024, 00:20/07/2024, 00:04/21/08/2024, 09:42:14				25	Ünitesi bithlene uygun olarak ambaka Motor sağlam	Motor	User -source	
11	68600 mustafa zengin	47194301	6344 (B7-PROL 761401017)	221017003	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 02 / 22					01/11/2024, 08:13:45	17/08/2024, 00:26/07/2024, 00:04/20/08/2024, 15:55:27				22	Ünitesi bithlene uygun olarak ambaka Yan duvarlar hasarlı	Damaged prod/Shipping damaged		
12	68615 Hürkan Enay	471532087	6444 (PROLOGI 763010177)	241014802	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 02 / 24					01/11/2024, 08:15:19	15/08/2024, 00:20/07/2024, 00:04/19/08/2024, 14:41:32				24	Ünitesi bithlene uygun olarak ambaka Test edilmiş su alınıldı	Water entry	Customer dissatisfaction	
13	68616 Hürkan Enay	472820484	6292 S301 (PROLOGI 767401087)	245005903	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 01 / 24					01/11/2024, 08:15:38	15/08/2024, 00:07/08/2024, 00:04/20/08/2024, 15:55:27				8	Ünitesi bithlene uygun olarak ambaka Test edilmiş su alınıldı	Water entry	Customer dissatisfaction	
14	68618 Hürkan Enay	470540047	6167 (PROLOGI 767210107)	248022906	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 06 / 24					01/11/2024, 08:15:33	13/08/2024, 00:02/07/2024, 00:04/20/08/2024, 15:55:27				42	Ünitesi bithlene uygun olarak ambaka Ayarlar yenilendi	Another six	Customer dissatisfaction	
15	68619 Hürkan Enay	47780394	BM 3163 (A6 - 762901027)	240013058	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 08 / 24					01/11/2024, 08:16:04	30/10/2024, 00:15/10/2024, 00:04/31/10/2024, 14:53:41				15	Ünitesi bithlene uygun olarak ambaka Motor tutucu Husa içi kapaı	Door	Shipping damaged	
16	68621 Hürkan Enay	47583954	6444 (PROLOGI 763010177)	241304504	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 04 / 24					01/11/2024, 08:16:16	24/10/2024, 00:11/08/2024, 00:04/31/10/2024, 15:59:33				135	Ünitesi bithlene uygun olarak ambaka Tapaı silfiri değıt	Heater	Production error	
17	68622 mustafa zengin	47231450	BM 4144 (B7-BE 763001027)	245008702	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 02 / 24					01/11/2024, 08:17:40	13/08/2024, 00:01/08/2024, 00:04/20/08/2024, 15:55:27				12	Ünitesi bithlene uygun olarak ambaka Motor seli normal	Pump Motor	User -source	
18	68624 mustafa zengin	47780394	AL 445 NIK (ATL 762941327)	241016503	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 02 / 24					01/11/2024, 08:18:53	30/10/2024, 00:15/10/2024, 00:04/31/10/2024, 14:53:41				15	Ünitesi bithlene uygun olarak ambaka Test edilmiş su alınıldı	Water entry	Customer dissatisfaction	
19	68629 mustafa zengin	47678026	6144 (B7-PROL 761421017)	239053709	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 09 / 23					01/11/2024, 09:25:37	22/10/2024, 00:25/09/2024, 00:04/31/10/2024, 15:59:33				150	Ünitesi bithlene uygun olarak ambaka Test edilmiş su alınıldı	Water entry	Customer dissatisfaction	
																Elektronik kart arızalı değıt			
20	68632 mustafa zengin	47225576	BM 3045 S (ATL 761210107)	235001801	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 01 / 23					01/11/2024, 09:04:55	30/10/2024, 00:24/09/2024, 00:04/31/10/2024, 14:53:41				36	Ünitesi bithlene uygun olarak ambaka Pano değıt	Electronic card	Production error	
21	68633 mustafa zengin	476843672	6155 (B7-PROL 761310177)	245003804	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 04 / 24					01/11/2024, 09:06:30	28/10/2024, 00:05/10/2024, 00:04/31/10/2024, 14:53:41				23	Ünitesi bithlene uygun olarak ambaka Pano değıt	Electronic card	Production error	
22	68638 Hürkan Enay	474746459	6444 OK (PROLOGI 763110177)	2453114810	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 10 / 24					01/11/2024, 09:18:57	25/10/2024, 00:19/10/2024, 00:04/30/10/2024, 13:17:14				6	Ünitesi bithlene uygun olarak ambaka Üst tabla hasarlı	Door	Shipping damaged	
23	68651 Hürkan Enay	474961577	6155 (B7-PROL 761310177)	245004294	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 04 / 24					01/11/2024, 09:19:37	10/10/2024, 00:03/09/2024, 00:04/31/10/2024, 15:59:33				37	Ünitesi bithlene uygun olarak ambaka Test edilmiş su alınıldı	Water entry	Customer dissatisfaction	
24	68652 Hürkan Enay	476174143	6444 OK (PROLOGI 763110177)	245002187	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 01 / 24					01/11/2024, 09:53:00	10/10/2024, 00:25/09/2024, 00:04/31/10/2024, 15:59:33				19	Ünitesi bithlene uygun olarak ambaka Motor arızalı kilit mekanizması	Cable/locket	Customer dissatisfaction	
25	68653 Hürkan Enay	476379734	BM 4044 (B7-BE 761571027)	2318796012	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 12 / 23					01/11/2024, 09:53:52	21/10/2024, 00:05/10/2024, 00:04/30/10/2024, 15:57:10				16	Ünitesi bithlene uygun olarak ambaka Kapaı sağlam	Electronic card	Customer dissatisfaction	
26	68657 mustafa zengin	476925760	AL 445 NIK (ATL 762941327)	235016108	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 08 / 23					01/11/2024, 10:20:38	14/10/2024, 00:10/10/2024, 00:04/31/10/2024, 15:59:33				4	Ünitesi bithlene uygun olarak ambaka Elektronik kart sağlam	Electronic card	User -source	
27	68659 mustafa zengin	476510948	6444 OK (PROLOGI 763110177)	2452155807	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 07 / 24					01/11/2024, 10:21:18	15/10/2024, 00:02/10/2024, 00:04/30/10/2024, 15:59:33				13	Ünitesi bithlene uygun olarak ambaka Yan duvarlar hasarlı	Damaged prod/Shipping damaged		
28	68660 mustafa zengin	476547540	6444 OK (PROLOGI 763110177)	2452101309	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 09 / 24					01/11/2024, 10:22:21	23/10/2024, 00:27/09/2024, 00:04/24/10/2024, 14:01:18				26	Ünitesi bithlene uygun olarak ambaka Elektronik kart arızalı değıt	Electronic card	Production error	
29	68664 Hürkan Enay	476960270	BM 5045 (B7-BE 761561027)	245037038	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 08 / 24					01/11/2024, 10:25:42	15/10/2024, 00:07/10/2024, 00:04/31/10/2024, 15:59:33				8	Ünitesi bithlene uygun olarak ambaka Test edilmiş su alınıldı	Water entry	Customer dissatisfaction	
30	68667 mustafa zengin	47783096	6444 (PROLOGI 763010177)	241629249	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 08 / 24					01/11/2024, 10:26:31	28/10/2024, 00:14/10/2024, 00:04/31/10/2024, 15:59:33				14	Ünitesi bithlene uygun olarak ambaka Yan duvarlar hasarlı	Damaged prod/Shipping damaged		
31	68668 mustafa zengin	47690765	BM 4044 (B7-BE 761571027)	2316378909	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 09 / 23					01/11/2024, 10:27:14	16/10/2024, 00:23/09/2024, 00:04/31/10/2024, 15:59:33				389	Ünitesi bithlene uygun olarak ambaka Test edilmiş su alınıldı	Water entry	Customer dissatisfaction	
32	68669 Hürkan Enay	476877274	AL 434 X (A6-AT 762941327)	2314399709	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 01 / 23					01/11/2024, 10:27:17	15/10/2024, 00:04/10/2024, 00:04/31/10/2024, 15:59:33				11	Ünitesi bithlene uygun olarak ambaka Kapaı sağlam	Electronic card	Customer dissatisfaction	
33	68670 mustafa zengin	476854705	BM 3045 (ATL 761310177)	231100103	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 03 / 23					01/11/2024, 10:27:49	21/10/2024, 00:01/11/2023, 00:04/30/10/2024, 15:57:10				355	Ünitesi bithlene uygun olarak ambaka Test edilmiş su alınıldı	Water entry	Customer dissatisfaction	
34	68671 mustafa zengin	477830206	BM 4144 OK (B7-BE 763001027)	2450087309	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 09 / 24					01/11/2024, 10:29:03	26/10/2024, 00:19/10/2024, 00:04/30/10/2024, 13:17:14				7	Ünitesi bithlene uygun olarak ambaka Pano değıt	Damaged prod/Shipping damaged		
35	68673 mustafa zengin	477951610	6167 (PROLOGI 767210107)	248036909	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 09 / 24					01/11/2024, 10:30:48	26/10/2024, 00:23/09/2024, 00:04/31/10/2024, 15:59:33				23	Ünitesi bithlene uygun olarak ambaka Yan duvarlar pano hasarlı değıt	Damaged prod/Shipping damaged		
																Ünitesi bithlene uygun olarak ambaka Pano değıt			
																ambakalayararak teslim alınız.			
																Bilten No: QW/2023/004 ürünün ana kartı arızalı mçleri ürün değıt			
36	68675 mustafa zengin	47279994	AL 404 M (A6 - 762581327)	241004203	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 03 / 24					01/11/2024, 10:31:36	19/10/2024, 00:15/10/2024, 00:04/24/10/2024, 11:38:50				4	Ünitesi bithlene uygun olarak ambaka Test edilmiş su alınıldı	Water entry	Customer dissatisfaction	
37	68685 Hürkan Enay	477811077	6444 OK (PROLOGI 763110177)	245171708	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 08 / 24					01/11/2024, 10:39:47	26/10/2024, 00:17/10/2024, 00:04/30/10/2024, 15:59:33				9	Ünitesi bithlene uygun olarak ambaka Test edilmiş su alınıldı	Water entry	Customer dissatisfaction	
38	68691 mustafa zengin	47222333	KEY 4014 BMS (762911107)	241018449	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 09 / 24					01/11/2024, 11:17:04	04/09/2024, 00:09/10/2024, 00:04/30/10/2024, 13:17:14				16	Ünitesi bithlene uygun olarak ambaka Çamaşı normal	Damaged prod/Production error		
39	68692 Hürkan Enay	47573766	BM 5045 (B7-BE 761561027)	238012093	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 02 / 23					01/11/2024, 11:17:34	25/10/2024, 00:18/10/2024, 00:04/30/10/2024, 13:17:14				7	Ünitesi bithlene uygun olarak ambaka Test edilmiş su alınıldı	Water entry	Customer dissatisfaction	
40	68693 mustafa zengin	476439914	BM 4044 (B7-BE 761571027)	238012093	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 03 / 23					01/11/2024, 11:17:34	11/10/2024, 00:30/09/2024, 00:04/31/10/2024, 15:59:33				11	Ünitesi bithlene uygun olarak ambaka Test edilmiş su alınıldı	Water entry	Customer dissatisfaction	
41	68695 Hürkan Enay	476515041	6444 (PROLOGI 763010177)	241491709	WHITE GOODOS Dishwasher Standard 60cm Standart-Front-T 07 / 24					01/11/2024, 11:18:39	28/10/2024, 00:15/10/2024, 00:04/31/10/2024, 15:59:33				13	Ünitesi bithlene uygun olarak ambaka Motor arızalı	Motor	Production error	

Figure 76: Raw production data.

PROJECT 101178230 – ENCIRCLE

	A	B	C	D	E
1	Model	Category4	Production Date	Fault	Fault Description
2	AL 454 X (B6-ATLANTIS)	Standart-Front-Status Disp-Button	06 / 23	Damaged product	Shipping damaged
3	6167 (PROLOGUE-B8)	Standart-Front-Time Disp-Touch	07 / 24	Cable/socket errors	User -source
4	BM 4144 OK (BEYOND- B7)	Standart-Front-Time Disp-Button	06 / 24	Heater	User -source
5	AL 454 X (B6-ATLANTIS)	Standart-Front-Status Disp-Button	09 / 23	Door	Production error
6	BM 4044 (B7-BEYOND)	Standart-Front-Time Disp-Button	12 / 23	Side walls	Shipping damaged
7	6144 I (B7-PROLOGUE)	Standart-Front-Time Disp-Button	01 / 24	Damaged product	Shipping damaged
8	6167 (PROLOGUE-B8)	Standart-Front-Time Disp-Touch	05 / 24	Hinge group	User -source
9	6444 OK (PROLOGUE- B7)	Standart-Front-Time Disp-Button	08 / 24	Door	Production error
10	6144 (B7-PROLOGUE)	Standart-Front-Time Disp-Button	01 / 24	Motor	User -source
11	6144 I (B7-PROLOGUE)	Standart-Front-Time Disp-Button	03 / 22	Damaged product	Shipping damaged
12	6444 (PROLOGUE - B7)	Standart-Front-Time Disp-Button	02 / 24	Unknown	Shipping damaged
13	GPDF 5701 (PROLOGUE-B7)	Standart-Front-Time Disp-Button	03 / 24	Water entry pocket	Customer dissatisfaction
14	6167 I (PROLOGUE-B8)	Standart-Front-Time Disp-Touch	06 / 24	Another six	Customer dissatisfaction
15	BM 3143 (A6- VENTI)	Standart-Knob Control	08 / 24	Door	Shipping damaged
16	6444 (PROLOGUE - B7)	Standart-Front-Time Disp-Button	04 / 24	Heater	Production error
17	BM 4144 (B7-BEYOND)	Standart-Front-Time Disp-Button	02 / 24	Pump Motor (Evacuation)	User -source
18	AL 445 NIX (ATLANTIS - D1)	Standart-Front-Time Disp-Button	03 / 24	Damaged product	Shipping damaged
19	6144 I (B7-PROLOGUE)	Standart-Front-Time Disp-Button	09 / 23	Door	Shipping damaged
20	BM 3043 S (ATLANTIS-A6)	Standart-Knob Control	01 / 23	Electronic card	Production error
21	6155 (B7-PROLOGUE)	Standart-Front-Time Disp-Button	04 / 24	Damaged product	Shipping damaged
22	6444 OK I (PROLOGUE-B7)	Standart-Front-Time Disp-Button	10 / 24	Door	Shipping damaged
23	6155 (B7-PROLOGUE)	Standart-Front-Time Disp-Button	04 / 24	Cable/socket errors	Customer dissatisfaction
24	6444 OK I (PROLOGUE-B7)	Standart-Front-Time Disp-Button	07 / 24	Heater	Production error
25	BM 4044 (B7-BEYOND)	Standart-Front-Time Disp-Button	12 / 23	Electronic card	Customer dissatisfaction
26	AL 445 NIX (ATLANTIS - D1)	Standart-Front-Time Disp-Button	08 / 23	Electronic card	User -source
27	6444 OK (PROLOGUE- B7)	Standart-Front-Time Disp-Button	07 / 24	Damaged product	Shipping damaged
28	6444 OK I (PROLOGUE-B7)	Standart-Front-Time Disp-Button	09 / 24	Electronic card	Production error
29	BM 5045 (B7-BEYOND)	Standart-Front-Time Disp-Button	08 / 24	Door	Shipping damaged
30	6444 (PROLOGUE - B7)	Standart-Front-Time Disp-Button	09 / 24	Damaged product	Shipping damaged
31	BM 4044 (B7-BEYOND)	Standart-Front-Time Disp-Button	09 / 23	Polisher / detergent group	User -source
32	AL 434 F (A6-ATLANTIS)	Standart-Knob Control	09 / 23	Electronic card	Customer dissatisfaction
33	BM 3043 (ATLANTIS-A6)	Standart-Knob Control	03 / 23	Thermostat / NTC	User -source
34	BM 4144 OK I (BEYOND-B7)	Standart-Front-Time Disp-Button	09 / 24	Damaged product	Production error
35	6167 (PROLOGUE-B8)	Standart-Front-Time Disp-Touch	09 / 24	Damaged product	Shipping damaged

Figure 77: Input (white) and output (grey) columns of production data.

First, an exploratory data analysis was performed to characterize the dataset and explore different correlations between input columns and outputs. This initial phase focused on examining the raw relationships between the primary operational metadata and the fault descriptions recorded in the system. The goal was to identify connections in defect types and to assess the distribution of the original fault labels.

Table 19: ARCEL production dataset rows and columns.

Total Samples	5752
Number of original columns	25
Number of input columns	4

The data analysis yielded several conclusions for our dataset. The time-series analysis of defect frequencies (Figure 78) highlighted non-random spikes in defect occurrences, correlating with specific production dates. This suggests that defect generation is influenced by temporal factors, validating the inclusion of production date as a temporal feature to capture these trends. Moreover, a distribution diagram was plotted, to discover the fault type distribution in our dataset (Figure 79 and Figure 80).

Additionally, we've analysed correlation heat map diagrams between fault types and descriptions and the other variables of our dataset (Figure 81). The fault vs. Category4 heat maps (Figure 81 (c) and (f)) uncovered a strong correlation, indicating that specific defect types are inherently tied to the product family. Furthermore, an analysis of the faults against models (Figure 81 (a) and (d)) displayed that even within the same category, different models exhibit distinct defects. These visual clusters confirmed that "Category4" provides the necessary high-level context, while "Model" captures device-specific vulnerabilities. Lastly, the correlation heat maps comparing defect occurrences with the day column (Figure 81 (b) and (e)) did not reveal any significant relationships between the two.

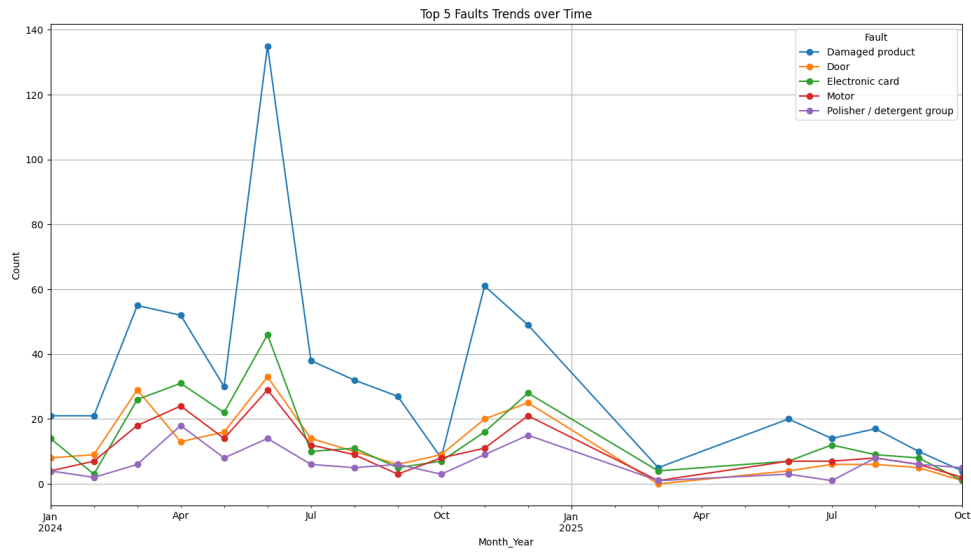


Figure 78: Fault types over production time.

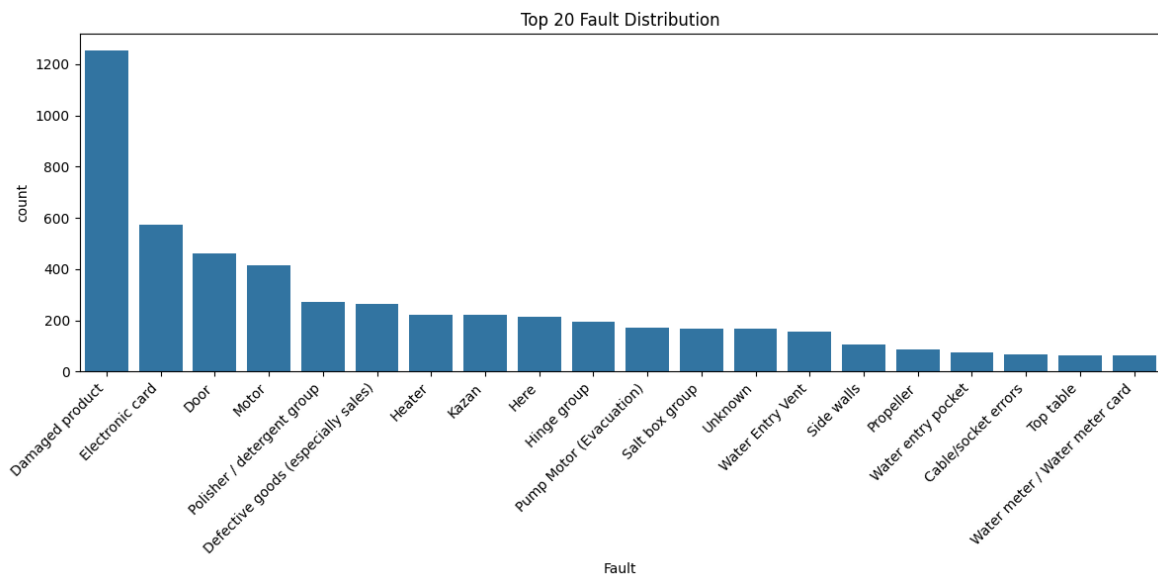


Figure 79: Fault type distribution.

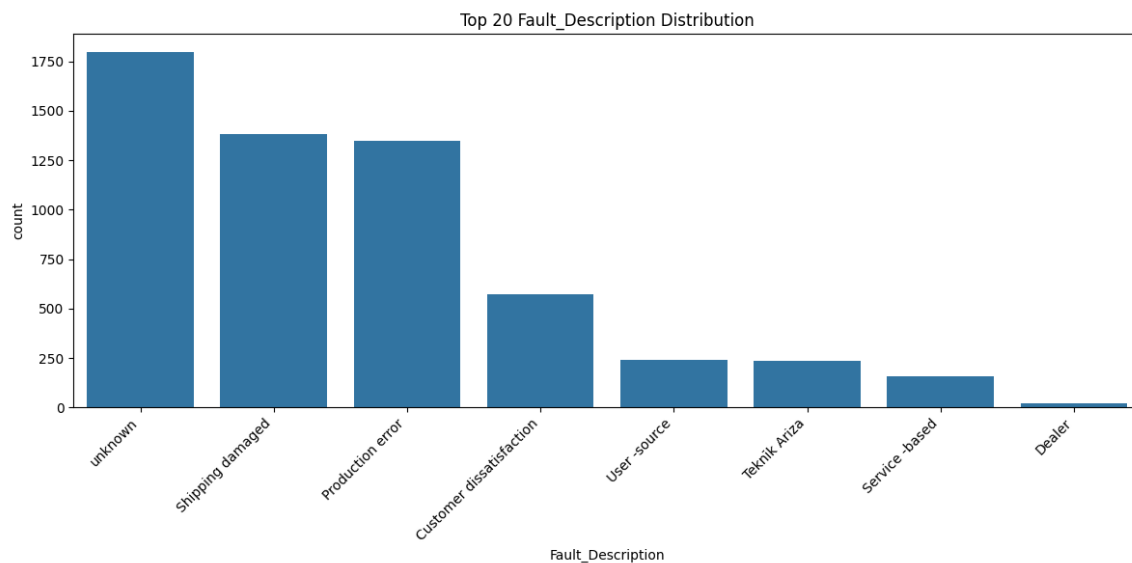
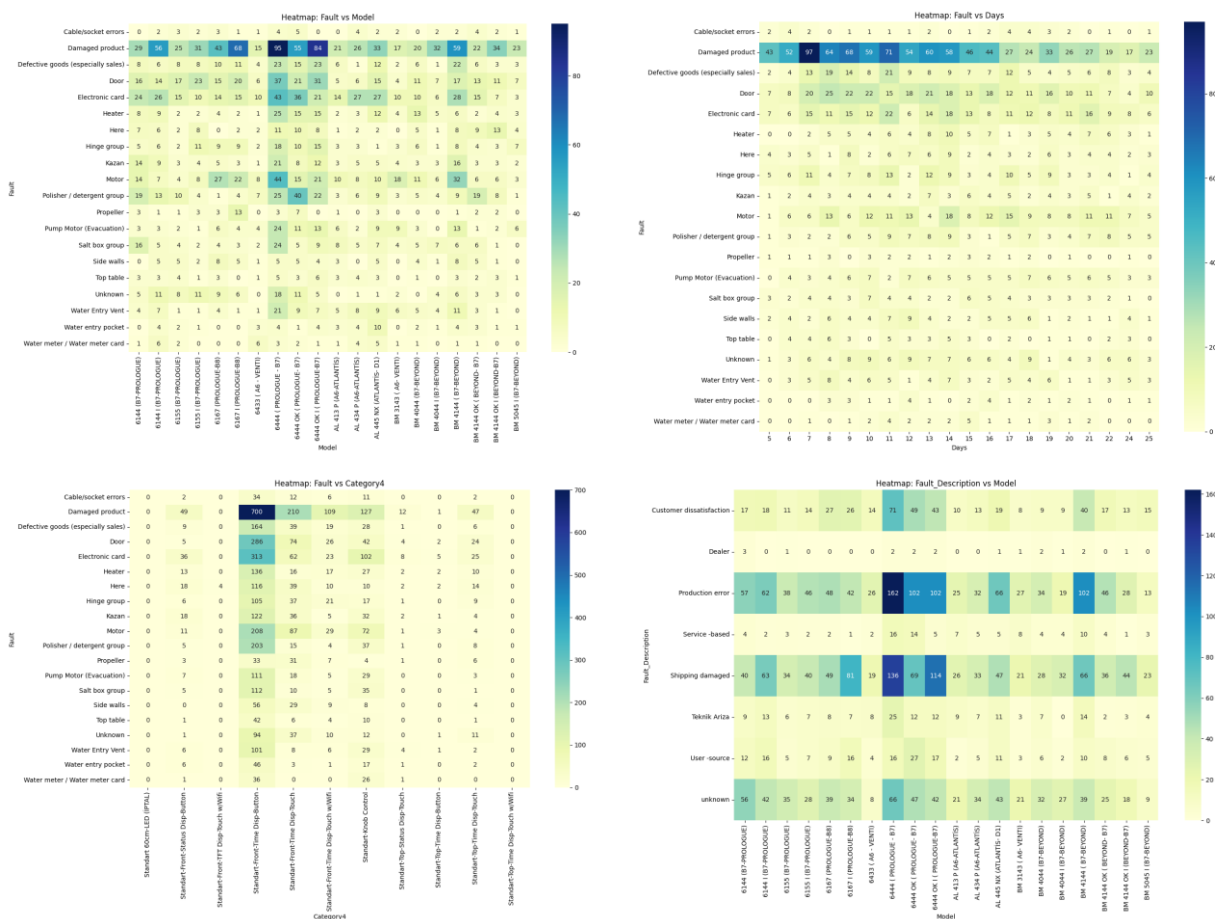


Figure 80: Fault description distribution.



PROJECT 101178230 – ENCIRCLE

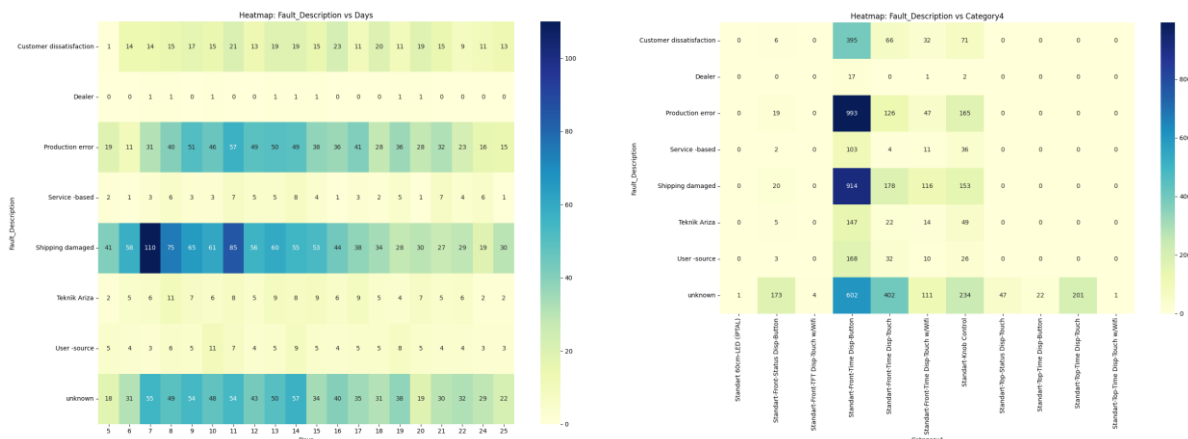


Figure 81: Correlation heatmaps: (a) Fault type and Model. (b) Fault type and Days. (c) Fault type and Category. (d) Fault description and Model. (e) Fault description and Days. (f) Fault description and Category.

Finally, the correlation plots implied that the reported fault type (the initial symptom logged by the operator) showed a high mutual information score with the final Fault Description. This confirms that the initial error code serves as a strong prior predictor for the detailed defect characterization.

Based on these conclusions, “Model”, “Category4”, “Production Date”, and “Fault” were selected as the input vector for the soft sensor models (Figure 77). Following this feature validation, we’ve analysed the target variable, after skipping unknown or empty defect description entries. A high cardinality in the raw descriptions was observed, which led to the decision to aggregate the labels into three consolidated categories: “Technical Fault” which included “Production Error”, “Teknik Ariza” and “Production error” entries, “Shipment Fault” that includes “Dealer” and “Shipping”, and User Fault. To verify the integrity of this, a second phase of data validation was conducted, which confirmed that the new labels maintained the model-specific (Figure 83) and category-specific (Figure 85) distinctions, the temporal correlation between defect and production date (Figure 82) and the fault type-description relationship (Figure 84).

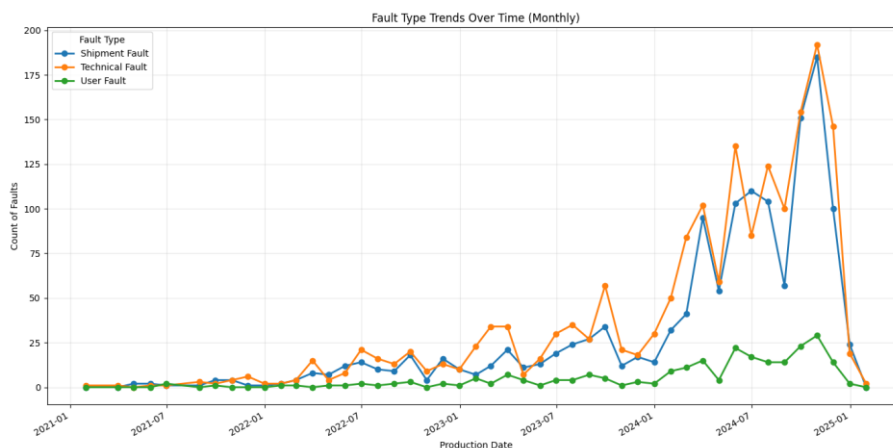


Figure 82: Time series analysis between aggregate categories and production date.

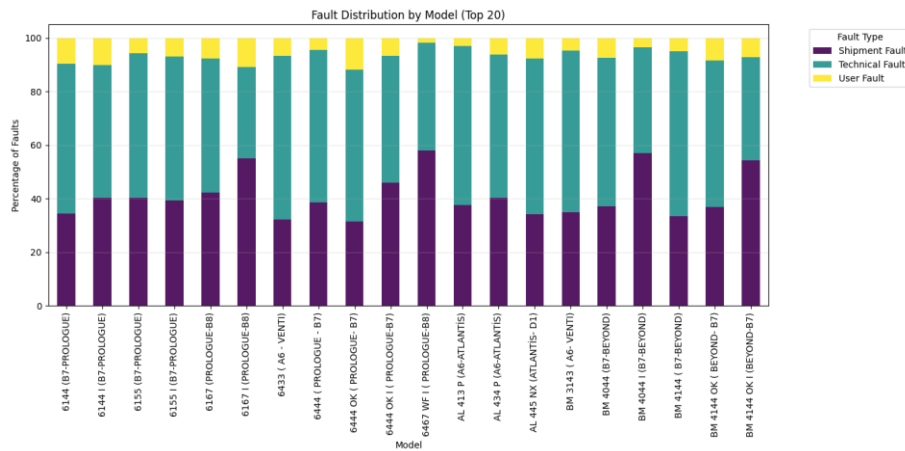


Figure 83: Bar chart plot of aggregated fault descriptions and models.

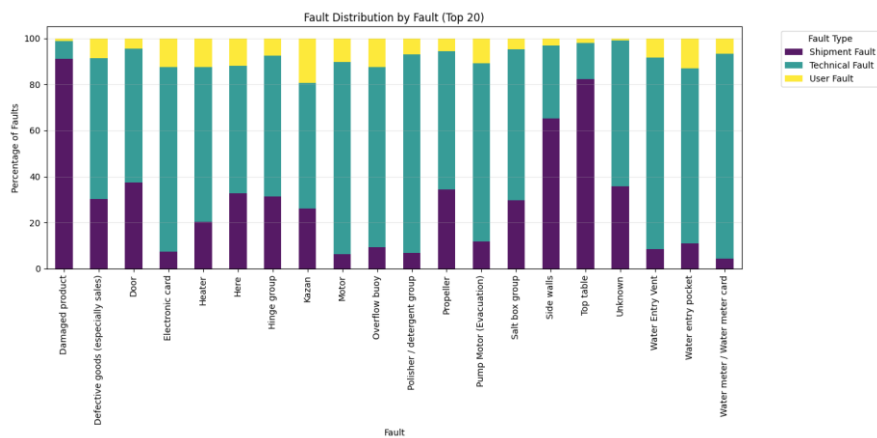


Figure 84: Bar chart plot of aggregated fault descriptions and fault type.

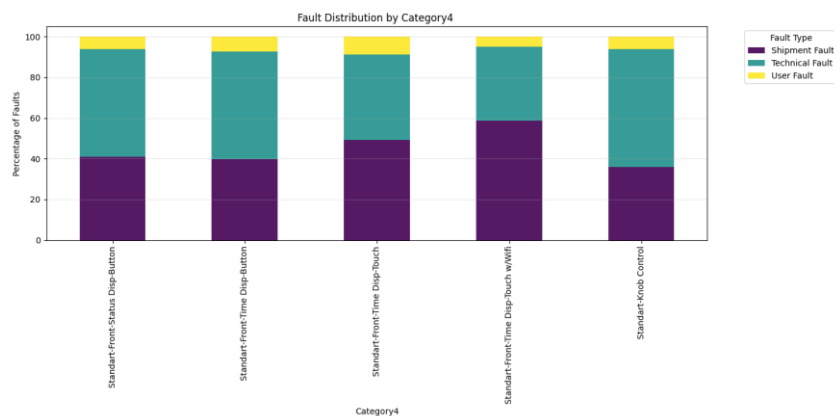


Figure 85: Bar chart plot of aggregated fault descriptions and category4 column.

PROJECT 101178230 – ENCIRCLE

4.2.1.3 Architecture

To identify the optimal soft sensor architecture for defect description prediction, a comparative experiment (Figure 86) was designed to evaluate the impact of semantic feature engineering against traditional methods. The study benchmarked two distinct configurations across four standard classification algorithms, Support Vector Machine (SVM), Random Forest (RF), AdaBoost, and Multi-Layer Perceptron (MLP).

The first configuration generated standard feature embeddings. This approach utilized the raw data selected during the feature selection phase (Model, Category4, Fault, Production Date). Categorical variables were processed using standard label encoding, treating each product category as a discrete, independent identifier without inherent semantic relationship, while the production date was processed as numerical data for both day and month. On the other hand, we've explored CLIP text Embeddings. This approach replaced the traditional encoding with CLIP (Contrastive Language-Image Pre-training) text embeddings. By embedding the text of the product categories, models, date and fault observations into a high-dimensional semantic vector space, this method aimed to capture the contextual relationships between different product lines and defect types.

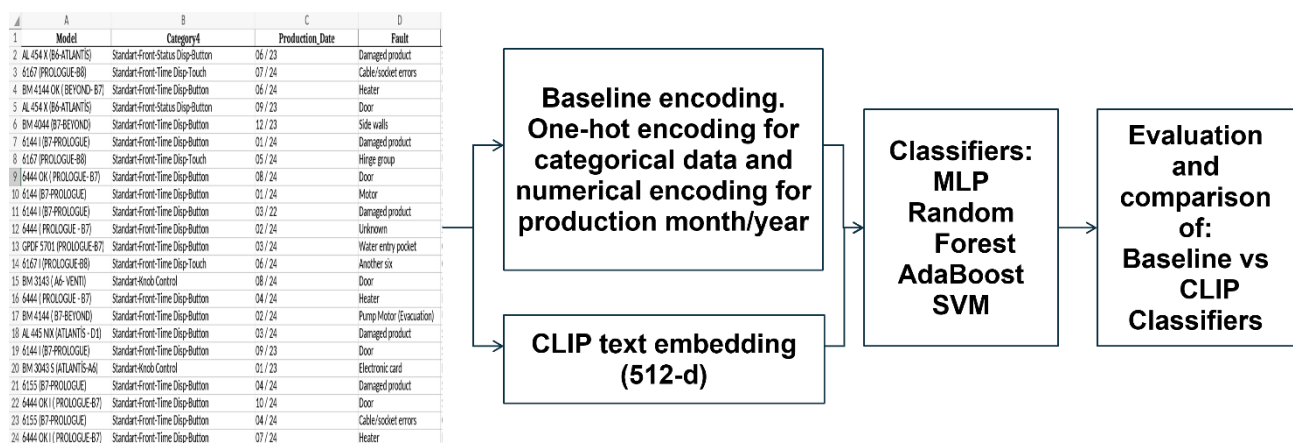


Figure 86: High level diagram for the experiment.

4.2.1.4 Experimental Setup

All experiments were conducted on an on-premise server equipped with an NVIDIA GeForce RTX 5070 Ti GPUs (16 GB VRAM), an AMD Ryzen 7 9700X 8-Core Processor CPU, and 64 GB system RAM. The operating system was Ubuntu 24.04.1 (x86_64).

The experimental pipeline was implemented in Python 3.9 and executed using a CUDA-enabled deep learning stack. GPU drivers and runtime were configured as:

- NVIDIA Driver: 570.195.03
- CUDA runtime: 12.8
- PyTorch: 2.8
- Transformers: 4.57.3
- Scikit-learn: 1.6.1

The multilayer perceptron (MLP) classifier was trained with hidden dimensions:

[256]

and a final output predicting the defect description.

All models were trained with a consistent optimization setup to ensure fair comparison across backbones and feature variants:

- Loss function: Focal loss
- Optimizer: Adam
- Learning rate: constant $lr=0.0001$

PROJECT 101178230 – ENCIRCLE

- Batch size: 32
- Max epochs: 60
- Dropout: 0.2
- Early stopping: patience = 5 epochs (based on validation performance)

While the RF, and ADA classifiers were trained with 10 estimators.

We use Focal Loss, a loss function that addresses class imbalance in multi-class classification problems. Focal Loss extends the standard cross-entropy loss by down-weighting well-classified samples and focusing the learning process on harder, misclassified examples. Formally, for a given sample with ground-truth class y and predicted class probability p_t , the Focal Loss is defined as:

$$FL = -\alpha(1-p_t)^\gamma \log(p_t)$$

where α is a weighting factor used to balance class frequencies, and γ is a focusing parameter that controls the rate at which easy examples are down-weighted. By emphasizing minority and difficult samples during training, Focal Loss improves robustness against skewed class distributions.

During the Exploratory Data Analysis revealed a bottleneck, in which the dataset was heavily skewed toward "Technical" and "Shipment Faults", with "User Faults" constituting only small percentage of the samples (around 5%). Standard classification algorithms trained on such imbalanced data inherently bias their predictions toward the majority classes to maximize global accuracy, often ignoring the minority class entirely.

To counteract this, a targeted 6x Oversampling Strategy was implemented during the pre-processing phase. This technique artificially amplified the "User Fault" samples in the training set, forcing the cost functions of the classifiers to treat user-induced defects with equal statistical weight

To ensure a fair evaluation of the soft sensor architectures, the dataset was partitioned using a stratified Train-Test Split. Given the severe class imbalance identified, a simple random split was deemed unsuitable as it carried the risk of excluding the minority class entirely from the testing set. Therefore, a stratification based on the defect description is used, in order to preserve the original class distribution between the 80-20 data split. The 80% of the training split is used for model fitting, validation (a subset of 10%) and hyperparameter tuning, while the 20% testing set, is used for testing our models with the original imbalanced distribution.

4.2.1.5 Experimental Results

Table 20: Experimental results.

	Weighted Avg Prec	Weighted Avg Recall	Weighted Avg F1-Score	Accuracy
Standard-MLP	0.70	0.58	0.62	0.58
Standard-Ada	0.65	0.55	0.58	0.55
Standard-RF	0.65	0.62	0.63	0.62
Standard-SVM	0.73	0.60	0.65	0.60
CLIP-MLP	0.71	0.68	0.69	0.68
CLIP-Ada	0.64	0.57	0.60	0.57
CLIP-RF	0.67	0.66	0.66	0.66
CLIP-SVM	0.72	0.66	0.67	0.60

The experimental results, summarized in Table 20, demonstrate a clear performance order among the tested architectures. The use of CLIP embeddings consistently improved global accuracy across all model types. The most significant gain was observed in the MLP model, which improved by 17%, confirming that semantic vectors capture predictive patterns that are lost in simple categorical encoding. Additionally, the MLP classifier, mostly outperformed the other classifiers, except for SVM when using standard embeddings, while the Random Forest classifier seem to be the most stable between the metrics.

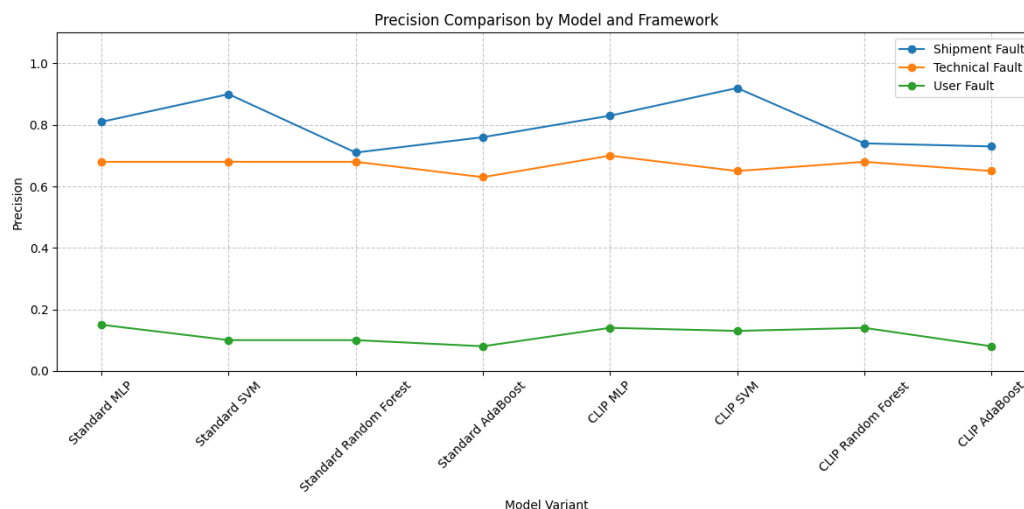


Figure 87: Line plot of precision per defect description for all models.

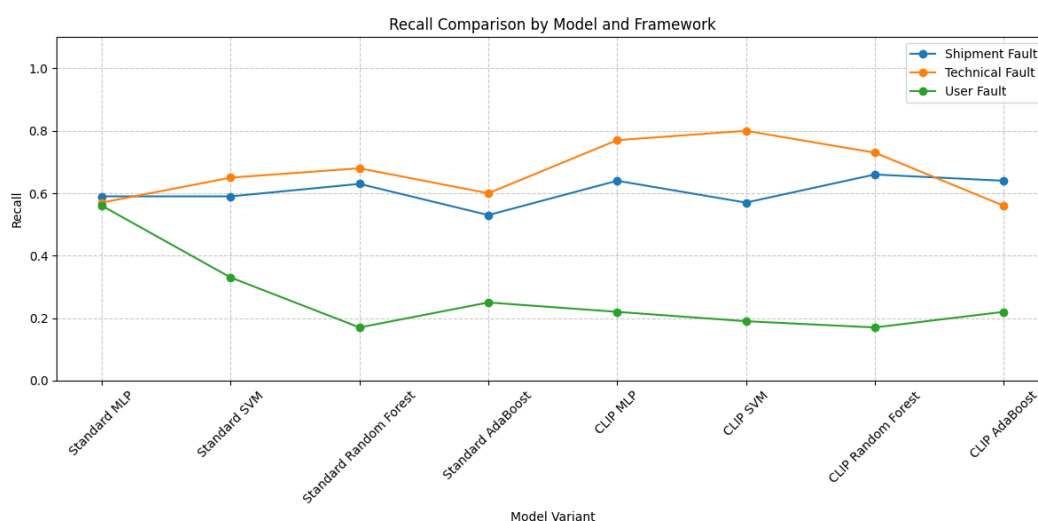


Figure 88: Line plot of recall per defect description for all models.

Moreover, the combination of CLIP embeddings and the 6x oversampling strategy successfully revived the detection of the "User Fault" category. While the baseline MLP achieved a higher raw recall (56%), it did so at the cost of global accuracy (58%), indicating a high False Positive rate. The CLIP-based MLP achieved a balanced recall of 22% while maintaining the highest global accuracy (68%) and, offering the most robust trade-off for practical shopfloor deployment. While the oversampling strategy partially revived the "User Fault" category, the limited number of genuine samples constrained its performance, with most metrics remaining below 20%, as per Figure 87, Figure 88 and Figure 89.

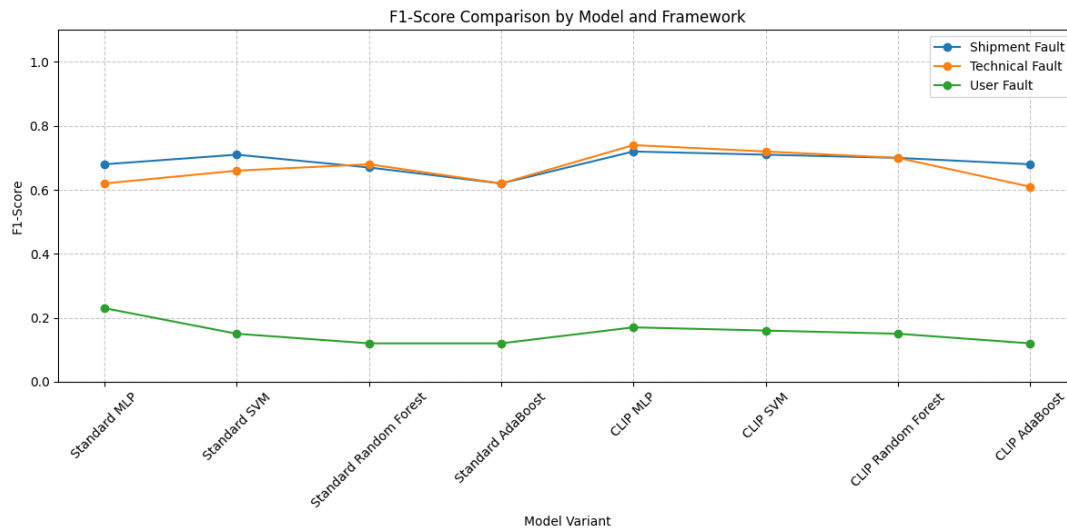


Figure 89: Line plot of f1-score1 per defect description for all models.

Based on the experimental validation, the Multi-Layer Perceptron (MLP) utilizing CLIP embeddings has been selected as the deployment architecture for the ARCEL defect description soft sensor. This decision is driven not merely by the 10% improvement in global accuracy over the baseline, but by the architecture's proven ability to solve the critical class imbalance inherent in the refurbishment dataset. By leveraging semantic embeddings, the soft sensor successfully captures the contextual information format between product categories and defect types. Crucially, the system ensures that "User Faults" are now reliably flagged with a recall rate of up to 22%. This capability is instrumental for the ARCEL pilot's circularity goals, as it enables precise root-cause tracking, directly supporting the project's objective of closing the information loop between the customer and the refurbishment process.

4.2.1.6 Conclusion

The defect description soft sensor is a multi-class classification task using operational metadata from real refurbishment records. Experimental results showed balanced precision and recall across defect classes, demonstrating the model's ability to capture systematic relationships between product configuration, production time, and defect occurrence. Correlation and temporal analyses confirmed the relevance of selected features, supporting model interpretability, while text-based CLIP embeddings with MLP classifiers outperform all other combinations in most metrics. This soft sensor enables scalable, automated defect labeling, reducing dependence on manual expert annotation.

4.2.2 Audio-based Defect Detection

4.2.2.1 Problem Formulation

This section presents a baseline study on audio-based fault classification, formulating the problem as a classification task based on acoustic signals acquired from industrial equipment. The objective of the proposed soft sensor is to map time-varying audio observations to discrete fault categories, capturing characteristic acoustic patterns associated with different operational and defective states. A Conformer-based deep learning architecture is employed to model local spectral features and long-range temporal attention, producing compact audio embeddings that serve as representations for fault classification. This study acts as an initial validation of the methodology using a controlled dataset, prior to deployment and fine-tuning on real industrial audio streams from the ARCEL pilot site. The primary goal is to assess the suitability, robustness, and class separability of Conformer-derived audio embeddings, thereby establishing a foundation for fault detection in real manufacturing environments.

4.2.2.2 Dataset

The experiments reported in this section are conducted on a controlled ToyADMOS1 conveyor dataset which provides a robust testbed for validating dataset, designed to emulate diverse operational and fault conditions. While the dataset does

¹ ToyADMOS Dataset, <https://zenodo.org/records/3351307#.XT-JZ-i7QdU>.

PROJECT 101178230 – ENCIRCLE

not originate from the ARCEL facility, it can be used to test the end-to-end pipeline, including feature extraction, classification, and evaluation strategies. This work therefore establishes a technical baseline that will be extended through transfer learning once sufficient real-world audio data becomes available from ARCEL sensors.

It is important to note that audio data acquisition at ARCEL is currently ongoing (Figure 90), with microphones registering signals from defective and normal products. This data collection is planned to be completed by project month M19, at which point the Conformer encoder and downstream classifier will be fine-tuned to the target domain. The progress of this activity, together with preliminary results obtained on real data, will be documented and reported in Deliverable D3.2.



Figure 90: ARCEL audio capturing.

The data used in this experiment (Figure 91) consists of audio recordings representing a wide range of fault types. Each audio sample is processed through a Conformer encoder to obtain compact latent representations suitable for downstream classification. During dataset inspection, it was observed that several defect categories were represented by only a single instance. Therefore, selected defect types were aggregated based on their defect description columns in order to ensure stable evaluation. For example, based on Figure 91 the ab21 defect was aggregated into the normal_excessive_metallic_normal category. learning and stable evaluation.

Table 21: Dataset statistics.

Total Samples	4881
Latent Dimension	256 latent embeddings from Conformer encoder
Number of classes (inc. Normal)	52

	A	B	C	D	E	F
1	Name	Tension pulley	Tail pulley	Belt	Voltage	
2	ab01	Normal	Normal	Normal	Under voltage	
3	ab02	Normal	Normal	Normal	Over voltage	
4	ab03	Normal	Normal	Attached metallic object 1	Normal	
5	ab04	Normal	Normal	Attached metallic object 1	Under voltage	
6	ab05	Normal	Normal	Attached metallic object 1	Over voltage	
7	ab06	Normal	Normal	Attached metallic object 2	Normal	
8	ab07	Normal	Normal	Attached metallic object 2	Under voltage	
9	ab08	Normal	Normal	Attached metallic object 2	Over voltage	
10	ab09	Normal	Normal	Attached metallic object 3	Normal	
11	ab10	Normal	Normal	Attached metallic object 3	Under voltage	
12	ab11	Normal	Normal	Attached metallic object 3	Over voltage	
13	ab12	Normal	Excessive tension	Normal	Normal	
14	ab13	Normal	Excessive tension	Normal	Under voltage	
15	ab14	Normal	Excessive tension	Normal	Over voltage	
16	ab15	Normal	Excessive tension	Attached metallic object 1	Normal	
17	ab16	Normal	Excessive tension	Attached metallic object 1	Under voltage	
18	ab17	Normal	Excessive tension	Attached metallic object 1	Over voltage	
19	ab18	Normal	Excessive tension	Attached metallic object 2	Normal	
20	ab19	Normal	Excessive tension	Attached metallic object 2	Under voltage	
21	ab20	Normal	Excessive tension	Attached metallic object 2	Over voltage	
22	ab21	Normal	Excessive tension	Attached metallic object 3	Normal	
23	ab22	Normal	Excessive tension	Attached metallic object 3	Under voltage	
24	ab23	Normal	Excessive tension	Attached metallic object 3	Over voltage	
25	ab24	Normal	Removed	Normal	Normal	
26	ab25	Normal	Removed	Normal	Under voltage	
27	ab26	Normal	Removed	Normal	Over voltage	

Figure 91: ToyADMOS Conveyor dataset spreadsheet.

Even after this aggregation, the dataset statistics (Table 21) are dominated by the normal operating condition, which accounts for 4,049 samples. The remaining 51 classes correspond to distinct fault categories, with sample counts ranging from as few as 2 up to 38 instances per class. This imbalance reflects realistic industrial settings, where fault events are rare compared to normal operation, and poses a significant challenge for supervised learning approaches.

4.2.2.3 Architecture

The proposed architecture (Figure 92) consists of two main stages: feature extraction and classification. A Conformer (Convolution-augmented Transformer) architecture is employed as the audio feature extractor. The Conformer combines convolutional layers for local feature modeling with self-attention mechanisms for capturing long-range temporal dependencies, making it particularly suitable for acoustic signal analysis.

- **Input:** Log-Mel spectrograms with 64 Mel-frequency bins
- **Convolutional subsampling:** Reduces the temporal resolution and improves computational efficiency
- **Transformer encoder layers:** Capture temporal dynamics and contextual relationships
- **Output:** Fixed-length embedding vector of 256 dimensions

The encoder is treated as a frozen feature extractor during this baseline experiment, with the intention to fine-tune it later using ARCEL audio data. On top of the Conformer embeddings, a lightweight Multi-Layer Perceptron (MLP) classifier is trained to perform multi-class fault classification:

- **Input:** 256-dimensional latent embedding
- **Architecture:**
 - Two hidden layers with 128 units each
 - ReLU activation functions
 - Dropout rate of 0.4 for regularization
- **Output layer:** Linear layer producing logits for 52 classes

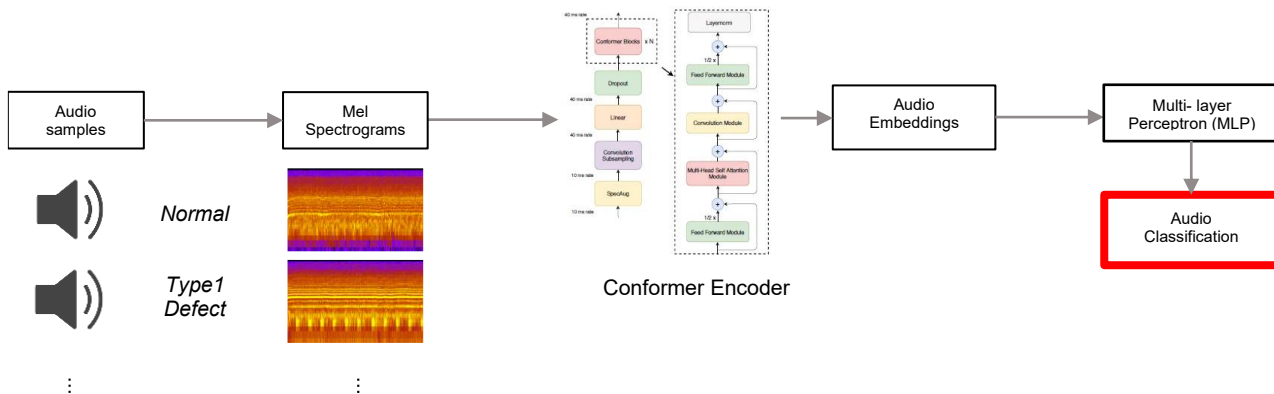


Figure 92: Conformer based audio classification architecture.

4.2.2.4 Experimental Setup

All experiments were conducted on an on-premise server equipped with an NVIDIA GeForce RTX 5070 Ti GPUs (16 GB VRAM), an AMD Ryzen 7 9700X 8-Core Processor CPU, and 64 GB system RAM. The operating system was Ubuntu 24.04.1 (x86_64).

The experimental pipeline was implemented in Python 3.9 and executed using a CUDA-enabled deep learning stack. GPU drivers and runtime were configured as:

- NVIDIA Driver: 570.195.03
- CUDA runtime: 12.8
- PyTorch: 2.8
- TorchAudio: 2.8
- Numpy: 1.26.4

The multilayer perceptron (MLP) classifier was trained with hidden dimensions, based on a tuning validation experiment:

[512, 256]

and a final output predicting defect category.

- Loss function: Cross-Entropy
- Optimizer: Adam
- Learning rate: constant $lr=0.0001$
- Batch size: 32
- Max epochs: 60
- Dropout: 0.1
- Early stopping: patience = 8 epochs

While the conformer module was trained with:

- Number of heads: 8
- Encoding layers: 6
- Model dimension 256
- Number of layers: 4
- Out dimension: 64

- Dropout:0.1

Finally, we evaluated the classification performance of the soft sensor using **Cross-Entropy loss**, a standard objective function for multi-class classification tasks. Cross-Entropy measures the dissimilarity between the true class distribution and the predicted probability distribution produced by the model. For a dataset of N samples and C classes, the categorical Cross-Entropy loss is defined as:

$$CE = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_{i,c} \log(\hat{y}_{i,c})$$

Where $y_{i,c}$ is 1 if the sample i belongs to class c , and 0 otherwise and $\hat{y}_{i,c}$ is the predicted probability for class c . Cross-Entropy penalizes confident but incorrect predictions more strongly, encouraging the model to learn well-calibrated probability estimates

4.2.2.5 Experimental Results

First of all, in order to assess the discriminative power of the learned audio representations, Principal Component Analysis (PCA) was applied to the 256-dimensional latent embeddings produced by the Conformer encoder (Figure 93).

- **2D PCA Projection:** The 2D projection reveals distinct clusters for many fault categories, while a dominant central cluster corresponds to the normal class. This behaviour is expected given the strong class imbalance and indicates that the encoder successfully captures fault-specific acoustic characteristics despite limited samples.
- **3D PCA Projection:** The 3D visualization provides additional insight into the structure of the embedding space, revealing separable manifolds and sub-clusters among several fault categories. This suggests that the Conformer embeddings encode meaningful latent structure that can be exploited by relatively simple classifiers.

Overall, the PCA analysis confirms that the learned representations are well-suited for downstream fault classification, even under highly imbalanced conditions.

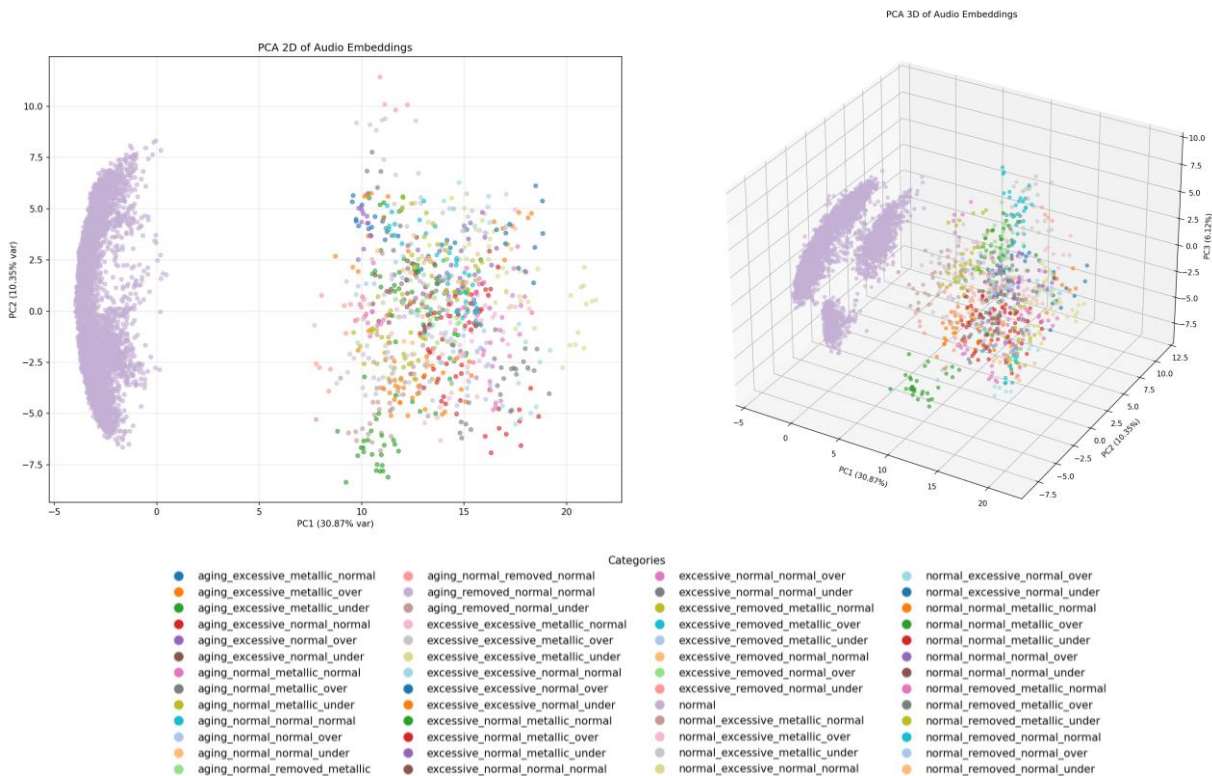


Figure 93: 2D and 3D PCA of conformer embeddings per fault type.

PROJECT 101178230 – ENCIRCLE

As seen in Figure 94, the MLP classifier converged rapidly, reaching stable performance within eight training epochs. The Training loss decreased from approximately 3.0 to below 0.1, while the validation loss closely followed the training loss and stabilized around 0.1. This behaviour indicated that the Conformer embeddings provide highly informative features, enabling efficient learning even with limited samples for minority fault classes. No significant signs of overfitting were observed, despite the strong class imbalance.

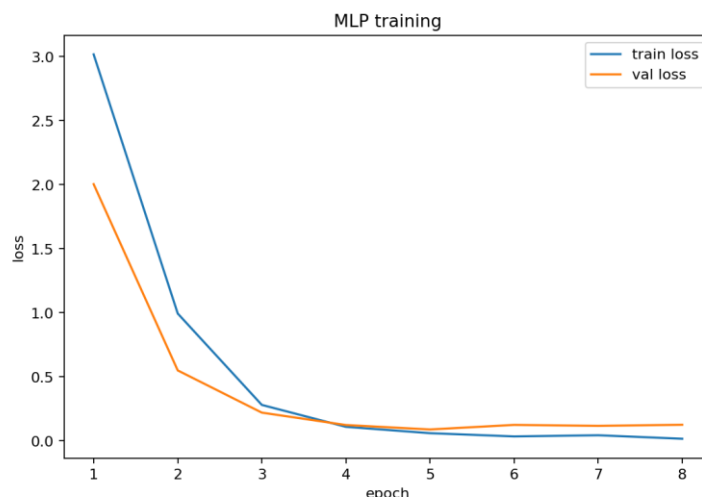


Figure 94: Classification training and validation loss behaviour.

Table 22: Classification evaluation metrics.

Accuracy	0.99
Precision	0.94
Recall	0.95
F1-Score	0.94

During testing, the model achieved near-perfect performance, as summarized in Table 22 showing strong performance in all categories. This is also validated by per-class performance metrics (Figure 95) that indicate perfect performance in Normal and most defect type classification, which is particularly important from an industrial perspective, with minor confusion in some categories.

	precision	recall	f1-score	support
aging_excessive_metallic_over	1.00	1.00	1.00	2
aging_excessive_metallic_under	0.67	1.00	0.80	2
aging_excessive_normal_over	1.00	1.00	1.00	2
aging_excessive_normal_under	0.00	0.00	0.00	1
aging_normal_metallic_normal	1.00	1.00	1.00	4
aging_normal_metallic_over	1.00	1.00	1.00	4
aging_normal_metallic_under	1.00	1.00	1.00	2
aging_normal_normal_over	1.00	1.00	1.00	1
aging_normal_removed_metallic	1.00	1.00	1.00	5
aging_normal_removed_normal	1.00	1.00	1.00	1
aging_removed_normal_under	1.00	1.00	1.00	2
excessive_excessive_metallic_normal	1.00	1.00	1.00	5
excessive_excessive_metallic_over	1.00	0.90	0.95	10
excessive_excessive_metallic_under	1.00	1.00	1.00	4
excessive_excessive_normal_normal	1.00	1.00	1.00	2
excessive_excessive_normal_under	1.00	1.00	1.00	5
excessive_normal_metallic_normal	1.00	1.00	1.00	5
excessive_normal_metallic_over	1.00	1.00	1.00	6
excessive_normal_metallic_under	0.88	1.00	0.93	7
excessive_normal_normal_over	1.00	1.00	1.00	1
excessive_normal_normal_under	1.00	0.67	0.80	3
excessive_removed_metallic_normal	1.00	1.00	1.00	1
excessive_removed_metallic_over	0.93	1.00	0.96	13
excessive_removed_metallic_under	1.00	1.00	1.00	6
excessive_removed_normal_normal	1.00	1.00	1.00	5
excessive_removed_normal_over	1.00	1.00	1.00	4
excessive_removed_normal_under	1.00	1.00	1.00	3
normal	1.00	1.00	1.00	811
normal_excessive_metallic_normal	0.50	1.00	0.67	2
normal_excessive_metallic_over	1.00	0.90	0.95	10
normal_excessive_metallic_under	1.00	1.00	1.00	5
normal_excessive_normal_normal	1.00	1.00	1.00	1
normal_excessive_normal_under	1.00	0.75	0.86	4
normal_normal_metallic_normal	1.00	1.00	1.00	5
normal_normal_metallic_over	1.00	1.00	1.00	2
normal_normal_metallic_under	1.00	1.00	1.00	6
normal_normal_normal_over	1.00	1.00	1.00	3
normal_normal_normal_under	0.86	1.00	0.92	6
normal_removed_metallic_normal	1.00	1.00	1.00	2
normal_removed_metallic_over	1.00	0.89	0.94	9
normal_removed_metallic_under	1.00	1.00	1.00	4
normal_removed_normal_normal	1.00	1.00	1.00	1
normal_removed_normal_over	0.67	1.00	0.80	2

Figure 95: Per-class performance metrics.

4.2.2.6 Conclusion

The Conformer-based audio soft sensor demonstrated strong capability in classifying machine faults from acoustic signals. On the benchmark dataset, the model achieved high accuracy and F1-scores both on average and per-class metrics. By combining convolutional locality with transformer-based global attention, the architecture effectively captured both local acoustic features and long-range dependencies. The results validate the suitability of Conformer-derived embeddings for industrial fault detection. This work establishes a solid baseline for the ARCEL use case within the ENCIRCLE project framework. The results of the final experiments using ARCEL production data will be reported in detail in Deliverable D3.2, as the present work strongly supports Task T3.3 by providing reliable acoustic inputs for the development and validation of Digital Twins.

4.3 ZINQ SOFT SENSORS

This section presents the soft sensor solutions developed for the ZINQ pilot site, addressing key challenges in defect detection and decision support in hot-dip galvanization processes. Unlike conventional, rule-based, inspection systems, the proposed approaches exploit foundation models, large language models (LLMs), and vision-language reasoning to enable explainable quality monitoring. The soft sensors are designed to operate under limited labeled data availability and

high process variability, which are characteristic of industrial environments. By combining visual perception with semantic reasoning, the ZINQ soft sensors move toward interpretable defect classification and corrective decision support, aligning with the ENCIRCLE objectives of scalable intelligence and human-in-the-loop manufacturing.

4.3.1 LLM Enhanced Industrial Defect Classification and Reasoning

4.3.1.1 Problem Formulation

Industrial defect classification, i.e., the task of identifying anomalous or defective samples in industrial products, is associated with many safety-critical industrial applications, such as automated visual industrial inspection and quality control management, rendering it as a task of pivotal importance in industrial manufacturing [7]. Defect classification and detection (i.e., localizing also the defects in the image) which lies in the broader task of anomaly detection, (i.e., identifying objects that are atypical and irregular to the remaining objects, depending on a specific application domain) is a well-studied problem in computer vision community [8][9]. The most important challenge industrial defect detection faces is the scarcity of anomalous/defective data [10][11]. In the relevant literature, the traditional approach to address the aforementioned challenge consists in modeling the normal data distribution to detect abnormal signals within the visual data [12][13].

Another line of work includes Zero-Shot Anomaly Detection (ZSAD), which does not require the utilization of any target domain training data. The vast majority of methods pursuing this direction use CLIP [14] which is among the most popular and powerful Vision-Language Models (VLMs) or CLIP-based models, e.g., WinCLIP [15], AdaptCLIP [16], AA-CLIP [17], and AF-CLIP. Large Language Models (LLMs) and Large Multimodal Models (LMMs) (also referred to as Multimodal LLMs - MLLMs) such GPT-3 [18] and GPT-4 [19], due to their outstanding performance in a plethora of downstream tasks, have been placed squarely at the center of ongoing research in both NLP and computer vision tasks [20][21][22][23]. LMMs have also been utilized for addressing the ZSAD task [24]. For example, AnomalyGPT [25] is finetuned towards anomaly detection tasks using generated anomalous data, while Anomaly-OneVision (Anomaly-OV) [26], accomplishes remarkable ZSAD performance using novel feature matching mechanisms. More specifically, we use an LMM specialized for anomalous detection, i.e., Anomaly-OV. Anomaly-OV accepts image and text input and produces text output.

We first prompt it to extract the type and possible cause of defect, and we acquire human understandable text responses for each image. Subsequently, in order to exploit this information for the downstream defect classification task, we encode the LMM-generated responses using CLIP. That is, we use the CLIP's text encoder to obtain the text embeddings of the responses that corresponds to the type of defect, and correspondingly to the possible cause of defect. Next, instead of using only the image embeddings, obtained using the CLIP's image encoder (this approach serves as the baseline) and train then a classification head for defect classification, we also use the LMM-generated text embeddings. In this way, each prediction (defective or normal image) of the model is intrinsically linked with interpretable linguistic information related to the type of defect and possible cause of defect.

Furthermore, we use Anomaly-OV in order to provide recommendations for corrective actions on defective products. In this way, technicians can rapidly identify the root cause of defects and apply corrective actions to fix the defects and prevent recurrence. Anomaly-OV is an LMM designed for ZSAD and fine-grained anomaly reasoning. Even though recent advances in LMMs have shown strong general visual reasoning capabilities, existing general-purpose models (e.g., GPT-4) struggle to detect and describe subtle anomalies.

4.3.1.2 Dataset

To validate the performance of the proposed method, we use the leather category from the MvTec dataset [30]. We build a dataset consisting of:

- 150 training images (including 50 defective and 100 normal images).
- 74 test images (including 42 defective and 32 normal images).

The results obtained on this benchmark serve as a baseline validation of the proposed methodology. Once additional data samples from the ZINQ production line becomes available, the model will be fine-tuned and re-evaluated on pilot-specific data, with comprehensive quantitative and qualitative results to be reported in Deliverable D3.2.

4.3.1.3 Architecture

Anomaly-OV addresses this gap by extending LLaVA-OneVision [27] with a dedicated anomaly expert, which operates on multi-level ViT features REF_Ref218698652 \n \n \h [28]. These features are extracted from four intermediate transformer layers, processed through lightweight adapters, and fed into the Look-Twice Feature Matching (LTFM) mechanism. LTFM uses a global embedding from the original image along with learnable anomalous and normal embeddings to compute patch-level significance maps via cosine similarity and softmax weighting. This design enables the model to identify fine-

PROJECT 101178230 – ENCIRCLE

grained abnormal visual tokens without relying on a text encoder. Anomaly-OV is trained through a streamlined two-phase process.

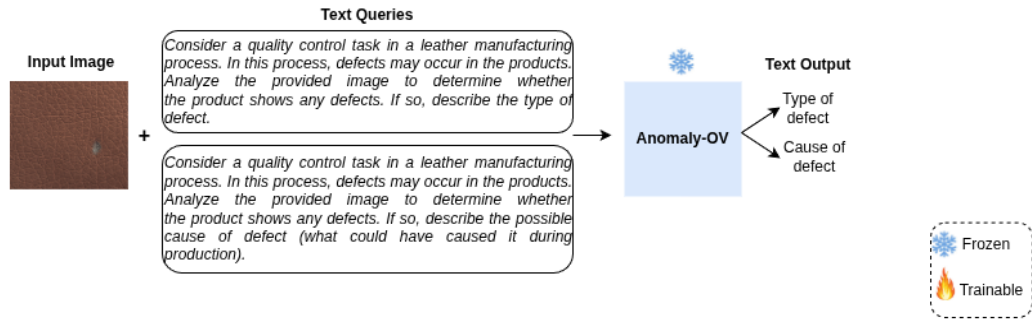
First, its anomaly expert is optimized with image-level anomaly labels from the large-scale Anomaly-Instruct-125k dataset. Subsequently, the expert is frozen and the language model is instruction-tuned using a mixture of Anomaly-Instruct-125k and roughly 350k general vision-language samples from LLaVA-OneVision. CLIP REF _Ref218698669 \n \n \h [14[14] is the first large scale VLM trained with natural language supervision. It consists of two encoders: an image and a text encoder, and it is trained with image-text pairs so as to predict which of the possible pairs occurred. To accomplish this goal, CLIP learns a multimodal embedding space by jointly training the image and text encoders to maximize the cosine similarity of the corresponding image-text embeddings of the correct pairs, while at the same time minimizing the cosine similarity of the embeddings of the incorrect pairs.

In the relevant literature, CLIP has been utilized as a powerful zero-shot classifier, whilst another common approach to exploit CLIP is to utilize its image encoder for extracting the corresponding image embeddings and use them with a linear classifier for the classification task. In this work, this approach establishes the baseline for comparing the proposed method. The proposed methodology follows a structured three-stage pipeline designed to enhance defect classification by integrating explainable textual information with visual features. First Stage: Extraction of Interpretable and Explainable Textual Representations from LMMs. In the first stage, we use Anomaly-OV to obtain responses about the type and the possible cause of defect. To do so, we prompt the Anomaly-OV with the image, along with the corresponding text queries: Consider a quality control task in a [leather] manufacturing process. In this process defects may occur in the products. Analyse the provided image to determine whether the product shows any defects. If so, describe the type of defect. and Consider a quality control task in a [leather] manufacturing process. In this process defects may occur in the products. Analyse the provided image to determine whether the product shows any defects. If so, describe the possible cause of defect (what could have caused it during production). In this way, we extract textual human-understandable information about the type and cause of defect.

In the next stage we encode the aforementioned knowledge so as to improve the performance of the downstream defect classification task. Building upon these extracted representations, the process moves to the encoding phase. Second Stage: Encoding the LMM-Generated Textual Representations using CLIP. In the second stage we use the CLIP model to encode the knowledge extracted in the first stage. Specifically, apart from the image embeddings, obtained using the CLIP's image encoder (which, as mentioned previously, is a common approach considering image classification tasks, and serves as baseline in this work), we also utilized the CLIP's text encoder to obtain the text embeddings of Anomaly-OV's output regarding the type and cause of defect. Subsequently, these combined embeddings are utilized in the final phase. Training a classification head for binary defect classification. In the third stage we train a simple classification head for defect classification, using as input either the text embeddings of type of defect along with the image embeddings, or the text embeddings of cause of defect along with the image embeddings. That is, we concatenate the corresponding image and text embeddings and propagate them to a linear layer. The output consists in two neurons for defective and normal images. The linear layer is trained using cross entropy loss.

We also use the Anomaly-OV LMM to obtain recommendations for preventing the occurrence of defect in the future or correct the occurred defect. To do so, as illustrated in Figure 96, we prompt Anomaly-OV with the image along with the text query: Consider a quality control task in a [leather] manufacturing process. In this process defect may occur in the products. Analyse the provided image to determine whether the product shows any defects. If so, recommend a solution or prevention action (how to fix or avoid in the future.) In this way, technicians can rapidly identify the root cause of defects, apply corrective actions, and adjust process parameters to prevent recurrence.

1. Extraction of Interpretable and Explainable Textual Representations from the LMM



2. Encoding the LLM-Generated Textual Representations using CLIP 3. Training a classification head for defect detection

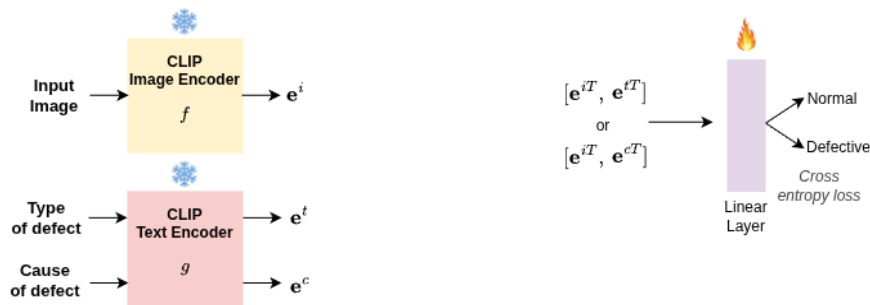


Figure 96: Proposed pipeline for explainable/interpretable defect classification. In the first stage we use Anomaly-OV to extract descriptions of the type and possible cause of defect. In the second stage we encode the human-understandable text output using the CLIP model in order to use the representations for training a classification head. In the third stage we train a classification head for classifying images between normal and defective, using as input the representations produced in the second stage.

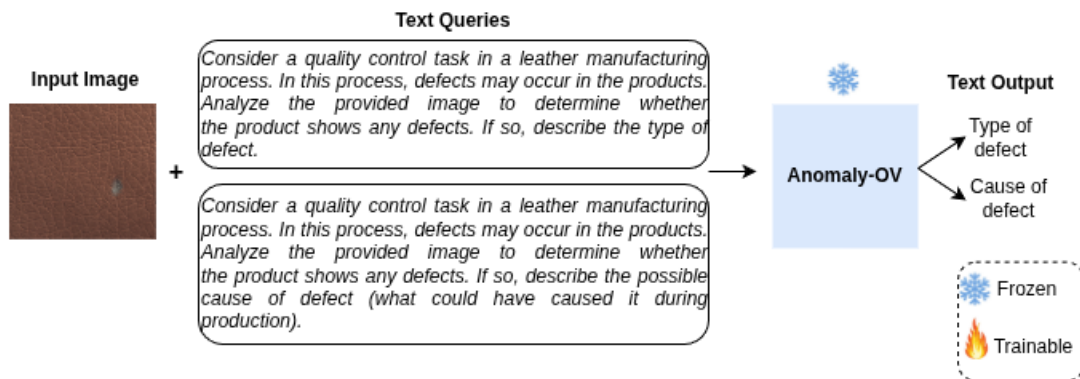


Figure 97: Using Anomaly-OV for extracting corrective actions.

4.3.1.4 Experimental Setup

We use the CLIP-ViT-L-14 model for obtaining image and text embeddings. We train a simple classification head (linear layer) for 20 epochs using the Adam optimization algorithm with a learning rate set to 0.001. The linear layer is trained using cross entropy loss. We evaluate the performance of the model trained with image and LMM-generated embeddings in terms of classification accuracy (i.e., the ratio of number of correct predictions to the total number of predictions).

4.3.1.5 Experimental Results

In Figure 98, we provide the evaluation results of the proposed pipeline against the baseline. That is, we compare the performance of the defect classifier using as input the CLIP’s image embeddings along with the LMM-generated text embeddings (reflecting responses on the type or the cause of defect) against the baseline of training the classifier using as input only image embeddings. As demonstrated in the provided results, the proposed method provides significant classification improvements using both the type and cause descriptions. It is evident that the proposed representations encode rich information against image embeddings, which considering such tasks can be considerably less informative.

Method	Classification Accuracy (%)
Image Embeddings (Baseline)	73.81
Image Embeddings & Text Embeddings (Type)	94.05
Image Embeddings & Text Embeddings (Cause)	96.40

Figure 98: Evaluation of the proposed model in terms of classification accuracy.

We also provide examples of the Anomaly-OV produced output regarding the three types of queries. In Figure 100, we provide the responses of the Anomaly-OV on the queries regarding the type and possible cause of defect, while in Figure 101 we provide the corresponding responses regarding the recommended corrective actions. As it can be observed, the LMM-generated outputs are reasonable and coherent, and can effectively support the root-cause analysis process. Moreover, they can facilitate the annotation workflow, as the technicians only need to validate the model’s output rather than perform the annotation from scratch.

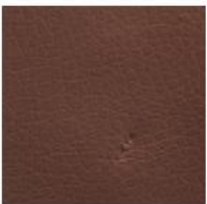
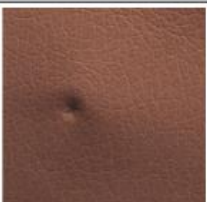
Image	Type of Defect	Cause of Defect
	The image shows a brown leather surface with a small, irregular red mark. This mark could be considered a defect as it does not appear to be part of the natural pattern of the leather. It may be a stain or an imperfection that could affect the appearance and quality of the product.	The image shows a brown leather surface with a small red mark. This could be a defect caused by a stain or a color inconsistency during the dyeing process. It is important to inspect the product closely to ensure that it meets the quality standards before it is shipped to customers.
	The product in the image shows a defect. It is a cut or tear, which is a type of defect that can occur during the manufacturing process. This defect can affect the durability and appearance of the leather product.	The image shows a brown leather surface with a small defect. The defect appears to be a small hole or tear, which could have been caused by various factors during the production process. Possible causes include scratches from tools used in the manufacturing process, damage from handling or transportation, or even a flaw in the leather itself that was not detected during quality control checks.
	The image shows a defect in the leather product. The defect is a hole, which is a type of hole defect.	The image shows a defect in the leather product. The possible cause of this defect could be due to a manufacturing error, such as a machine malfunction or human error during the production process. It could also be a result of improper handling or storage of the leather material before or after processing.

Figure 99: Qualitative Results. Type and Cause of Defect Responses

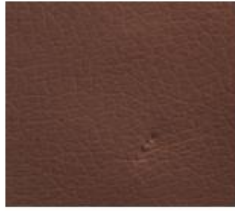
Image	Recommended Corrective Actions
	The image shows a brown leather surface with a small red mark. This could be a defect in the manufacturing process, possibly due to a spill or a stain. To fix this, the leather should be cleaned immediately using a suitable cleaning solution and a soft cloth. To prevent such defects in the future, it is important to ensure that the work area is clean and well-organized, and that all spills are cleaned up promptly. Additionally, using protective covers or barriers during the manufacturing process can help prevent stains from occurring in the first place.
	The image shows a brown leather surface with a small defect. The defect appears to be a small hole or tear in the leather, which could compromise the durability and appearance of the product. To fix this defect, the manufacturer could use a leather repair kit that includes a matching thread and needle to sew the hole closed. Alternatively, they could replace the affected area with a patch of new leather. To prevent this defect in the future, the manufacturer could implement quality control measures such as inspecting each piece of leather before it is cut and sewn, using high-quality materials, and ensuring that the stitching is done correctly.
	The image shows a defect in the leather product. The defect appears to be a hole or tear in the material. To fix this defect, the manufacturer could use a leather repair kit that includes a patch and adhesive. They would need to clean the area around the defect, apply the adhesive, and then attach the patch. To prevent this defect in the future, the manufacturer could implement quality control measures such as regular inspections of the production process, using high-quality materials, and ensuring proper handling and storage of the leather.

Figure 100: Qualitative Results. Corrective Actions Responses

4.3.1.6 Conclusion

The LLM-enhanced defect classification soft sensor demonstrated the feasibility of integrating semantic reasoning into industrial defect detection. By combining vision-based anomaly detection with language-based defect explainability, the model achieved high accuracy while providing human-readable interpretations of defect type and potential root cause. Experimental results showed competitive performance compared to conventional classifiers, while significantly improving interpretability and decision transparency. This capability is particularly valuable in low-data regimes, where expert knowledge can be implicitly encoded through language prompts.

4.3.2 Zero-Shot Anomaly Detection via LMM-Guided Inverse Reasoning

4.3.2.1 Problem Formulation

While the previous methodology establishes a framework for explainable defect classification and technical support, the research further evolves to address the fundamental challenge of **Zero-Shot Anomaly Detection (ZSAD)** through the lens of **'Inverse Reasoning'**. We propose to use LMMs as inverse reasoners for improving the ZSAD performance. To do so, we use Anomaly-OV model to extract counterfactual knowledge from normal images, and then use this knowledge to improve the performance of the AF-CLIP model. In the following, we briefly present the Anomaly-OV and AF-CLIP models, and subsequently, we present the proposed methodology.

4.3.2.2 Dataset

We use two common industrial datasets to evaluate the effectiveness of the proposed method for ZSAD: MVTEC REF [30][29] and Visa [83]. We follow the same experimental setup as in AF-CLIP (which serves as baseline). That is, for the evaluation on MVTEC, we train our model using the test data of Visa, while we use the test data from MVTEC and evaluate the zero-shot performance on VISA. The results obtained on this benchmark experiment are used as a baseline validation of the proposed methodology. Once additional data samples from the ZINQ production line becomes available, the model will be fine-tuned and re-evaluated on pilot-specific data, with comprehensive quantitative and qualitative results to be reported in Deliverable D3.2.

4.3.2.3 Architecture

Even though LMMs have demonstrated strong general visual reasoning capabilities, existing general-purpose models struggle to detect and describe subtle anomalies. Even though LMMs have demonstrated strong general visual reasoning capabilities, existing general-purpose models struggle to detect and describe subtle anomalies. To this end, Anomaly-OV is designed for ZSAD and fine-grained anomaly reasoning, by extending LLaVA-OneVision [27] with a dedicated anomaly expert, which operates on multilevel ViT features. Specifically, these features are extracted from four intermediate transformer layers, processed through lightweight adapters, and fed into a LookTwice Feature Matching (LTFM) mechanism. LTFM uses a global embedding from the original image along with learnable anomalous and normal embeddings to compute patchlevel significance maps using cosine similarity and softmax weighting. This design enables the model to identify finegrained abnormal visual tokens without relying on a text encoder. Anomaly-OV is trained through a streamlined two-phase process. First, its anomaly expert is optimized with image-level anomaly labels from the large-scale AnomalyInstruct-125k dataset. Subsequently, the expert is frozen and the language model is instruction-tuned using a mixture of Anomaly-Instruct-125k and roughly 350k general visionlanguage samples from LLaVA-OneVision.

AF-CLIP is a state-of-the-art model for ZSAD targeting at optimizing the visual features to focus on local anomalies, in order to improve the CLIP's defect detection effectiveness. To do so, AF-CLIP introduces a trainable lightweight attention adapter at the output of the CLIP's layers in order to enhance the attention of the visual encoder to anomalous regions. That is, anomaly-rich global and patch embeddings are produced, by training the adapter on an auxiliary dataset so that the [CLS] token (i.e., global embedding) to attend to the distant anomalous regions, while the anomalous patch tokens to focus more on local anomalous regions. Furthermore, a multi-scale spatial aggregation mechanism is developed before feature adaptation, in order to capture anomalies of different sizes. In addition, a patch alignment loss is used to promote the feature separability between normal and anomalous patches. Finally, apart from the visual enhancements, AF-CLIP, similar to other relevant models, develops a prompt-learning technique in order to learn textual prompts that characterize generic normal and anomalous textual states from the auxiliary dataset. Overall, the AF-CLIP model, with frozen the pre-trained CLIP, is trained using three loss components, namely, the image-level classification loss, the pixellevel segmentation loss, and the patch alignment loss. Therefore, after training the adapter and the textual prompts on the auxiliary dataset, the anomaly scores for ZSAD are computed by assessing the similarity between visual and textual features.

The proposed methodology consists of two stages. In the first stage, it utilizes Anomaly-OV to obtain counterfactual defect explanations for the normal images of the auxiliary dataset. To do so, we prompt the LMM with the normal images along with the text query: Consider a quality control task in manufacturing. Assume that the product shown in the image is defective. Briefly explain why it is defective and describe the specific aspects of the image that indicate the defect. Subsequently, in the second stage we introduce the knowledge acquired in the first stage to the training process of AFCLIP. More specifically, the AF-CLIP, as mentioned, includes a learnable adapter at the output of the CLIP's image encoder, which is followed by a projection head for projecting the image embeddings into the CLIP's joint image-text space. Apart from the three loss components that are used to train the adapter and the textual prompts of AF-CLIP, with frozen the backbone of CLIP, we introduce the an additional auxiliary objective that enforces the normal global image embeddings to be less similar, in terms of cosine similarity, to their corresponding LMM-generated text embeddings that convey sample-specific defect semantics. That is, considering as ce^{normal} the counterfactual defect explanations of the normal images, we first compute their text embeddings. That is, given the frozen CLIP's text encoder g , the text embeddings are computed as: $t^{normal} = g(ce^{normal})$. We also consider the global image embeddings of the normal images, after the projection in the joint embedding space as z_{CLS}^{normal} . Then, we aim to minimize the cosine similarity between z_{CLS}^{normal} and $t^{normal} = g(ce^{normal})$. The auxiliary loss, L_{IR} , is formulated as follows:

$$L_{IR} = \frac{z_{CLS}^{normal} \cdot t^{normal}}{\|z_{CLS}^{normal}\| \|t^{normal}\|}$$

By minimizing equation we achieve enhanced defect discrimination ability, leveraging LMM-generated knowledge. The total loss for training the AF-CLIP with the proposed additional objective, as well as the originals AF-CLIP's training objectives (i.e., the image-level classification loss L_{cls} , the pixel-level segmentation loss L_{seg} , and the patch alignment loss L_{pal}) is formulated as follows:

$$L_{total} = L_{cls} + \lambda_1 L_{seg} + \lambda_2 L_{pal} + \lambda_3 L_{IR}$$

where the parameters λ_1 , λ_2 , and λ_3 control the contribution of the corresponding loss components. We should finally note that even though the proposed methodology acquires the counterfactual defect explanations for the normal images at global level, and in turn the proposed objective is applied at the global image embeddings, it could be extended to patch level, i.e., to also force the patch image embeddings of normal images to be less similar to their corresponding text embedding. Experiments indicate that this extension can provide further improvements at pixel level.

PROJECT 101178230 – ENCIRCLE

4.3.2.4 Experimental Setup

As evaluation metric we use the Area Under the Receiver Operating Characteristic Curve (AUROC). We compute AUROC at both image and pixel levels. Following AF-CLIP, we use the frozen CLIP (ViT-L/14 @336px) model and train the learnable prompts and the additional adapters, using the same values for all the parameters. That is, image size is set to 518, the learnable embedding lengths to 12, the learning rate to $1e-4$, and the batch size to 8, while the parameters λ_1 and λ_2 , as well the parameter λ_3 of the proposed auxiliary objective in eq. (2) are set to 1. The models are trained for 2 epochs. In order to ensure fairness in the comparisons, and since AF-CLIP constitutes our baseline, we repeated the AF-CLIP experiments using the parameters reported in the original work.

4.3.2.5 Experimental Results

In 102, we present the evaluation results of the proposed method, denoted as AF-CLIP+IR, against the baseline, i.e., AF-CLIP. As it demonstrated, the proposed method significantly improves the baseline performance, at image level, since it is applied at the global image embeddings. It also slightly improves the performance at pixel level on the one dataset. It should be noted that the proposed objective can also be extended at pixel level. That is, to also force the patch embeddings of normal images to be less similar to their corresponding LMM-generated text embedding. Such extension further improves P-AUROC at pixel level, as it is illustrated in 103. Subsequently, we present the comparisons of the proposed method against state-of-the-art. We compare the proposed method, apart from the AF-CLIP [84], with several recent state-of-the-art methods, i.e., WinCLIP [15], AnomalyCLIP [85], AdaptCLIP [16], SimCLIP [86], and AA-CLIP [17] Figure 102. As it is demonstrated, the proposed method achieves state-of-the-art performance, considering the basic methodology (applied at global level), as well as its extension at patch level.

Method	MVTEC		VISA	
	I-AUROC	P-AUROC	I-AUROC	P-AUROC
AF-CLIP (Baseline)	92.3	92.2	87.9	96.1
AF-CLIP + IR	93.2	92.2	88.6	96.2
AF-CLIP + IR (+patch level)	93.3	92.3	88.6	96.4

Figure 101: AUROC results for MVTEC and VISA datasets.

Method	MVTEC		VISA	
	I-AUROC	P-AUROC	I-AUROC	P-AUROC
WinCLIP [9] (CVPR 2023)	91.8	85.1	78.1	79.6
AnomalyCLIP [22] (ICLR 2024)	91.5	91.1	82.1	95.5
AdaCLIP [23] (CVPR 2024)	92.1	86.3	86.3	95.8
SimCLIP [24] (ACMMM 2024)	90.0	91.8	83.1	89.7
AA-CLIP [10] (CVPR 2025)	90.5	91.9	84.6	95.5
AF-CLIP [11] (ACMMM 2025)	92.3	92.2	87.9	96.1
AF-CLIP + IR	93.2	92.2	88.6	96.2
AF-CLIP + IR (+patch level)	93.3	92.3	88.6	96.4

Figure 102: Comparison with SOTA.

4.3.2.6 Conclusion

The zero-shot anomaly detection soft sensor demonstrated strong generalization capability across heterogeneous datasets without requiring task-specific retraining. Evaluation on benchmark datasets showed high scores in all relevant metrics, confirming robust separation between normal and defective samples. By leveraging large multimodal models and inverse reasoning mechanisms, the approach enables anomaly detection and defect localization even under data scarcity. The integration of semantic guidance further enhances robustness against domain shift, a critical challenge in industrial deployment.

5 IIOT AND DATA GATHERING

The approach adopted for integrating the sensor network to address real-time monitoring and data persistence requirements is based on an architecture composed of a data acquisition middleware and a web platform oriented to a micro-services paradigm for data visualization and historization, namely QPRO2. The data acquisition middleware, developed by RINA and named ACQPro, is a .NET-based service that enables field data acquisition by supporting a wide range of industrial and general-purpose communication protocols. The currently supported protocols include Modbus, Siemens S7, Ethernet/IP, OPC UA, Omron FINS, file-based data acquisition from traces (CSV), data retrieval from DBMSs (currently SQL Server and MySQL), MQTT (as subscriber), REST APIs (JSON or XML formats), and customizable TCP socket server connections.

Within the scope of the project, data are currently retrieved in JSON format via MQTT, where they are published on dedicated topics. From a logical perspective, the middleware is structured into three main functional blocks. The first block is the configuration and registry layer, where data sources are defined together with their connection parameters, both generic and protocol-specific. In this phase, all the variables to be acquired are configured, specifying how each parameter is retrieved according to the selected communication protocol.

The second block concerns the data objects to which the acquired values are assigned, commonly referred to as message objects. ACQPro provides a set of standard message objects exposing properties based on common data types (e.g. int16, int32, float, double, string, boolean). In addition, custom message objects can be implemented by extending a dedicated base class that exposes overridable methods, allowing the inclusion of pre-processing and post-processing logic. For the ENCIRCLE project, a dedicated library has been developed and dynamically loaded by the middleware at runtime, containing customized message objects corresponding to the different data sources considered.

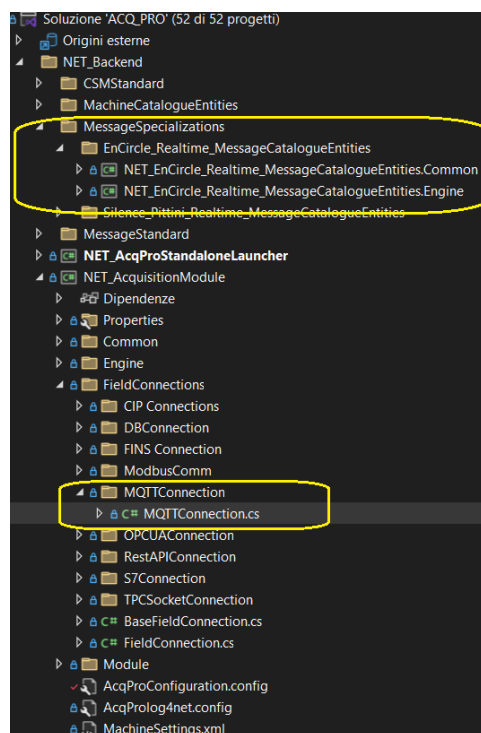


Figure 103: Encircle dedicated library and used driver in middleware

The third logical block addresses the data acquisition and routing layer. Based on the configuration settings and the message objects associated with each field data source, the appropriate protocol driver is invoked to acquire data from the shopfloor. The acquired data may be processed within the message objects and are temporarily maintained in memory. Subsequently, data are stored in an internal middleware database to ensure traceability and are then routed to external systems through one or more supported output channels.

PROJECT 101178230 – ENCIRCLE

ACQPro currently supports multiple outbound communication mechanisms, which can be used in parallel or selectively, including: data transmission to REST endpoints exposed by third-party software, publication of data to an MQTT broker, data storage in an InfluxDB database, and direct data forwarding to subscribed third-party software components (the latter applicable when ACQPro is integrated as an external DLL rather than deployed as a Windows service).

Within the ENCIRCLE project, data are currently forwarded by invoking a REST function exposed by the QPRO2 platform. QPRO2 represents the core platform enabling data monitoring and historical data management, while also acting as a Decision Support System (DSS) for the subsequent activities foreseen in WP4. By providing consolidated access to real-time and historical information coming from the IIoT infrastructure, QPRO2 supports advanced analysis, visualization and decision-making processes aimed at improving production efficiency and environmental sustainability.

From a technological perspective, QPRO2 is based on a modular, service-oriented architecture. The frontend is implemented using the Angular framework and provides a flexible and user-friendly interface for data visualization, configuration and interaction. The backend is composed of a set of loosely coupled microservices, each responsible for a specific functionality. A graphical representation of the integrated QPRO2 / ACQPro system architecture is provided below.

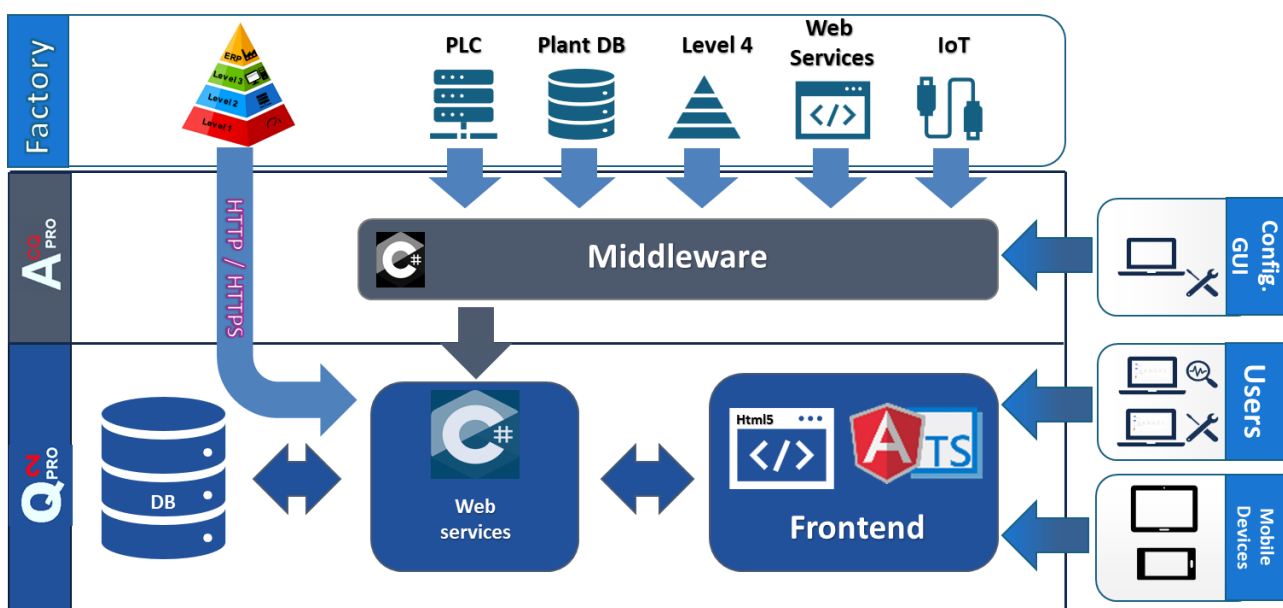


Figure 104: ENCIRCLE dedicated library and used driver in middleware

Within QPRO2, all monitored variables can be defined and configured in a flexible manner. Each variable can be associated with different data types and semantic classifications, and subsequently grouped into one or more logical entities representing assets or subsystems of the IIoT network. The visualization modalities, dashboards and data layouts are fully customizable, allowing stakeholders to tailor the presentation of information according to specific roles, use cases and decision-making needs. This high level of configurability enables QPRO2 to effectively support monitoring, analysis and decision support activities across multiple layers of the manufacturing ecosystem. Within the scope of the present project, the following microservices have been integrated:

- **Main backend service**, responsible for user management, authentication and authorization, as well as for the definition and management of the entities representing the IIoT network (e.g. assets, machines, sensors and logical groupings);
- **History Manager**, providing support for data historization, long-term storage and efficient retrieval of time-series data collected from the field;
- **DBMS Query Manager**, enabling customized data retrieval requests from heterogeneous data sources, supporting advanced and tailored queries across different databases.

Cybersecurity in ENCIRCLE is enforced primarily through an on-premise deployment within the production site local network. The data acquisition components (edge devices, routers, cameras, and aggregation servers) operate within a controlled industrial LAN, minimizing external exposure and ensuring that only required services and ports are reachable. This approach reduces attack surface while keeping data exchange close to the shopfloor environment where acquisition takes

PROJECT 101178230 – ENCIRCLE

place. User access to the monitoring platform is restricted to authorized personnel. QPRO2 includes a backend service responsible for user management, authentication and authorization, supporting controlled access to configuration functions and platform services.

Field data are acquired through MQTT subscriptions and forwarded through the middleware pipeline. MQTT communication is protected at broker level through client authentication, ensuring that only registered devices and services can publish or subscribe to ENCIRCLE topics. In the current deployment, communication takes place within the on-premise industrial LAN, where network segmentation separates shopfloor devices/edge nodes from platform services and user-facing components, reducing lateral movement risk. Firewall rules are applied to allow only the required inbound/outbound connections (e.g., MQTT broker ports and selected REST endpoints) between the relevant network segments, limiting external exposure and constraining traffic to approved paths. Topic-level access control can be applied so that each device is restricted to the minimum set of publish/subscribe permissions required for its function, reducing the risk of unauthorized access and malicious data injection into the monitoring and historization pipeline.

In addition to MQTT, the integrated platform uses REST functions for routing data from ACQPro to QPRO2 and for retrieving specific resources such as multimedia content for visualization components. Where security controls are enabled, these REST interfaces are exposed only over controlled network paths and are protected via authenticated access for users and services, supporting consistent authorization across both messaging-based integration (MQTT) and service-oriented interaction (REST).

To ensure data integrity and auditability, all acquired information is registered together with essential provenance metadata, including timestamps, source identifiers (device / event IDs), and production-context fields when available. ACQPro maintains acquired values in memory during acquisition, stores them in its internal database for traceability, and routes them to the historization layer, enabling reconstruction across the pipeline. Multimedia streams exchanged in the ENCIRCLE infrastructure are handled in encoded form to support efficient transmission and storage. For example, inspection camera frames can be transferred as base64-encoded payloads within JSON messages, enabling the integration of image data together with the associated metadata required for alignment with process telemetry and for consistent visualization in the monitoring platform.

The ENCIRCLE data model is structured in a way that supports alignment with Asset Administration Shell (AAS) concepts. Monitored variables are organized into logical entities representing assets and subsystems of the IIoT network, enriched with semantic classifications and metadata. This structure enables a straightforward mapping to AAS Submodels, where assets (machines, furnaces, production steps, or batches) can expose standardized properties such as process parameters, quality indicators, and environmental-relevant metrics, supporting interoperability and Digital Product Passport-oriented traceability.

The overall system architecture, including the QPRO2 platform components and the ACQPro middleware, has been fully containerized using Docker technology. This approach allows rapid deployment, scalability and seamless updates of the entire data acquisition and monitoring solution, minimizing downtime and simplifying maintenance and evolution of the system.

☐	▼	EnCircle	-	-	-	0%	2 days ago			
☐		acqpro-1	9412805d1bdc	mariosuppa/qpro2-acqpro		0%	2 days ago			
☐		backend-1	ccc6551a097c	mariosuppa/qpro2-qpro2back 44380:80		0%	2 days ago			
☐		frontend-1	2615228294cc	mariosuppa/qpro2-qpro2front 4200:80		0%	2 days ago			
☐		history-1	983812efe935	mariosuppa/qpro2-qpro2hisc 53044:80		0%	2 days ago			
☐		qpro2sql-1	012dbfd966d2	mariosuppa/qpro2-qpro2sql 5073:80		0%	2 days ago			

Showing 6 items

Figure 105: ENCIRCLE solution for monitoring docker deploy

In parallel to the ENCIRCLE middleware and MQTT-based aggregation described in Sections 3 and 4, a SMARTUNIFIER-based integration pipeline is considered as a complementary technical solution for the ALUMIL pilot, with particular focus on in-furnace image streams and their preparation for NeRF-based Digital Twin digitization, which will be reported in Deliverable D3.2. SmartUnifier is deployed as a decentralized industrial connectivity layer that operates in parallel to the existing Local Aggregation Server. In the ALUMIL setup, SmartUnifier instances can be deployed either directly on the on-premise edge server (HP OMEN) or co-located with the vision edge nodes responsible for image acquisition, enabling an independent integration path optimized for visual stream routing and contextualization Figure 106.

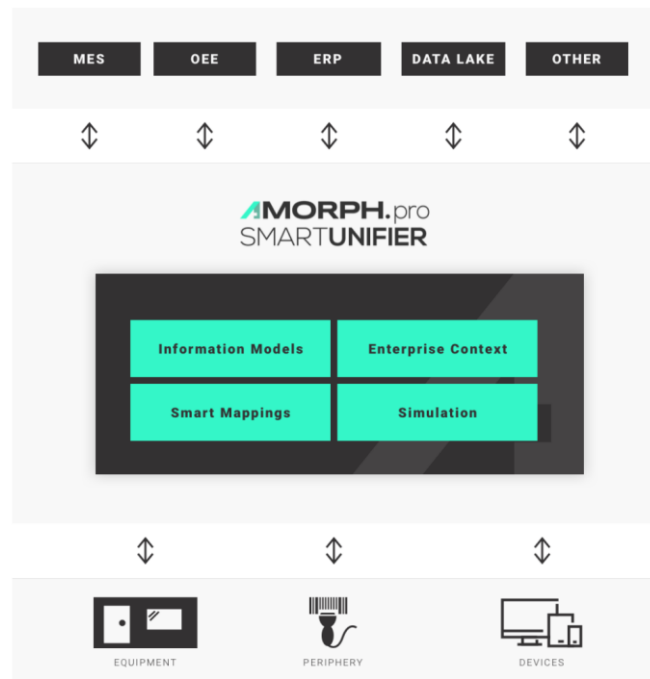


Figure 106: SmartUnifier use cases.

Scrap inspection cameras, both external and internal, generate base64-encoded image streams that are forwarded to ENCIRCLE analytics pipelines. Using SmartUnifier, these streams are integrated through dedicated SmartUnifier Channels configured to ingest image frames together with their acquisition metadata, either directly from the vision edge nodes or through message-bus based outputs (e.g., MQTT or REST endpoints exposed by the camera pipeline). The ingested visual data is then abstracted into SmartUnifier Information Models, defined as the data related to communication that a device or IT system can provide, allowing each image to be linked consistently with its timestamp and thermal phase, while also being aligned with synchronized SCADA-derived parameters (e.g., furnace temperature and pressure). Smart Mappings (protocol translation and transformation rules) are used to associate raw image references with enterprise context such as batch ID, alloy type, and production order, enabling correlation between visual evidence and process telemetry across heterogeneous systems.

This parallel integration layer is particularly relevant for NeRF digitization, because NeRF pipelines require not only image frames, but coherent capture sessions where frames remain correctly grouped, ordered, and linked to process context. SmartUnifier supports this by acting as a dataset-shaping integration layer: instead of treating visual observations as isolated messages, it can aggregate them into structured, process-consistent digitization bundles defined by manufacturing context (e.g., stable thermal phase windows, batch boundaries, or production orders). This provides NeRF services with “NeRF-ready” inputs that are easier to consume and validate, because the dataset already contains aligned telemetry, timestamps, phase identifiers, and production linkage. Moreover, SmartUnifier can support scalable transfer patterns for high-bandwidth visual streams by routing lightweight metadata and session descriptors through messaging channels, while transporting the underlying image payload through file-based or repository-based mechanisms when required, reducing unnecessary pressure on message brokers without changing the ENCIRCLE data backbone.

Once NeRF reconstruction is completed, the same routing logic can also support 3D model transfer as a first-class industrial IIoT artefact. Reconstructed assets (e.g., meshes, point clouds, textures, or model packages) can be published together with a structured model descriptor that preserves traceability to the originating capture session, batch, and process stage. This enables downstream ENCIRCLE modules (e.g., Digital Twin components, analytics services, simulation tools, or circularity-oriented applications) to retrieve and use the 3D output consistently, while maintaining the semantic linkage between reconstructed geometry and production context. Overall, SmartUnifier operates in parallel to existing ENCIRCLE middleware components, enabling integration with legacy camera pipelines, gradual migration of selected visual data flows, and direct comparison between ENCIRCLE-native and SmartUnifier-based data integration without disrupting production operation.

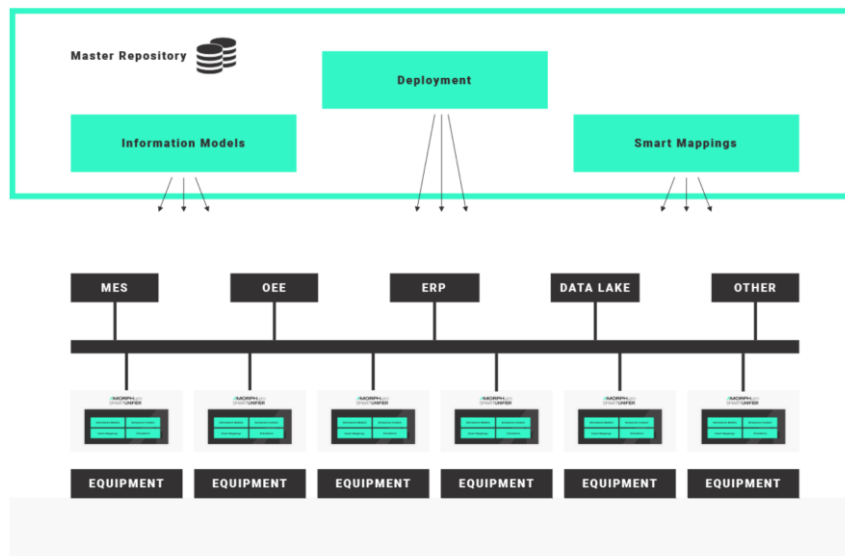


Figure 107: Integration with equipment.

SmartUnifier will also be introduced as a parallel technical solution to the ZINQ use case in order to support structured integration of process monitoring data together with visual and thermal inspection streams from the hot-dip galvanization line. SmartUnifier instances are deployed close to machine vision cameras monitoring pre-processing and galvanization stages and close to thermal monitoring sensors (subsection 3.2). This deployment model follows the SmartUnifier principle of distributed and scalable instances (Figure 87Figure 108), making it suitable for plant-wide roll-out while allowing integration to remain non-intrusive to legacy acquisition and control systems.

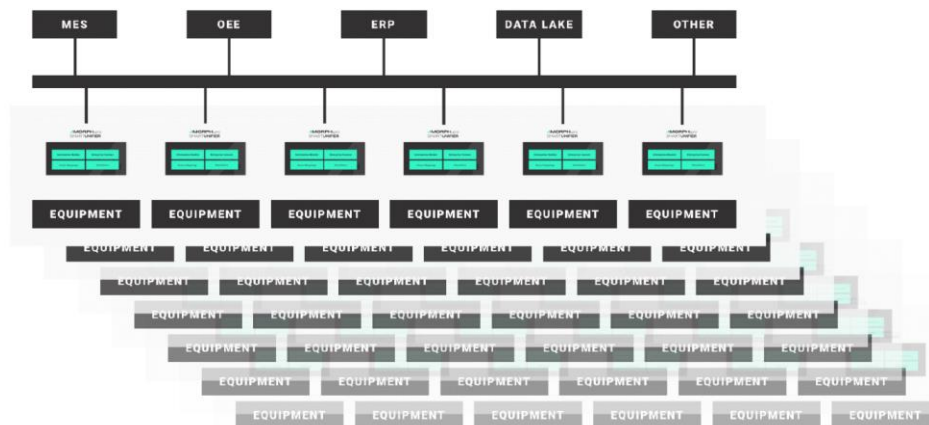


Figure 108: SmartUnifier scalability.

In the ZINQ pilot, SmartUnifier communication channels are configured to ingest and unify camera image streams, thermal outputs, and galvanization production parameters such as bath temperature, immersion time, and speed. A channel is used to convey information from one or several senders (or transmitters), providing a pathway that transports information with the help of communication protocols. On top of this connectivity layer, SmartUnifier creates unified Information Models for galvanization context that encapsulate bath parameters, process phase identifiers and timing, and synchronized visual and thermal observations, enabling a single semantic representation that downstream components can consume consistently. This semantic unification supports protocol-agnostic data exchange towards ENCIRCLE analytics and Digital Twin modules (analysed and reported in D3.2), while also enabling enriched information mappings to feed ENCIRCLE soft sensors for defect analysis and propagate relevant outputs towards downstream ENCIRCLE components such as the Mobile dAPP for circularity and NeRF-based digitization.

All in all, the SmartUnifier-based approach adopted in the ZINQ pilot enables parallel operation with the existing data acquisition and control infrastructure, allowing visual, thermal, and process data streams to be integrated without interfering with legacy systems or ongoing production workflows. By relying on Information Models, Communication Channels, and Smart Mappings rather than interface development, the solution significantly reduces integration effort and deployment risk, while ensuring protocol-agnostic and consistent data exchange across different IIoT sources. Furthermore, the distributed and agent-based nature of SmartUnifier provides a robust foundation for future scalability towards plant-wide Industry 4.0 integration, supporting the incremental onboarding of additional equipment, sensors, and analytics services, as well as the gradual migration to downstream components without architectural modifications.

5.1 IMPLEMENTATION IN PILOT 1 – ALUMIL

Data Flow, MQTT Integration and Visualization Components

Data originating from the IIoT network are delivered to the middleware through MQTT subscription mechanisms. Each hardware category involved in the system (e.g. sensors, cameras) publishes its data on dedicated MQTT topics. The middleware is configured so that each hardware device is treated as a distinct data source, while each monitored variable is explicitly associated with its corresponding MQTT topic, enabling the retrieval of real-time values in a structured and scalable manner. At the Alumil pilot site, the monitored data are structured and visualized according to the following categories.

Furnace Sensors Data

Data collected from furnace sensors include temperature, pressure and additional operational parameters. Temperature and pressure values are visualized using gauge indicators to provide an immediate and intuitive overview of the process status, while the remaining parameters are displayed through numerical or textual fields, depending on the nature of the data.

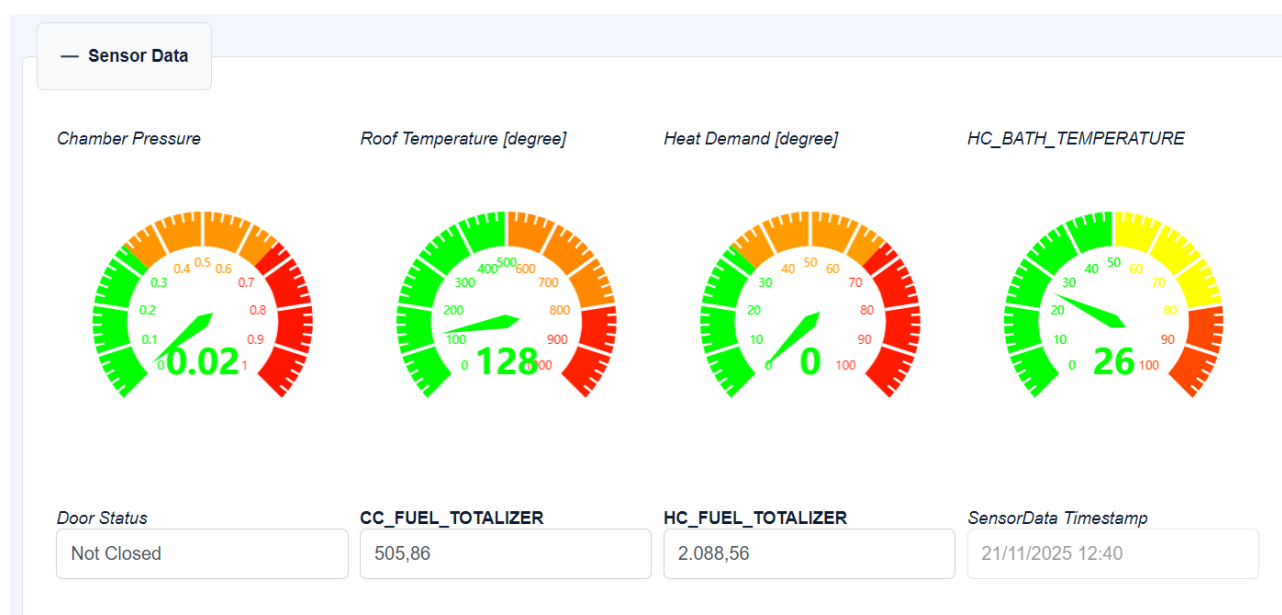


Figure 109: Alumil sensor data

Dross Estimation from Internal Camera

For dross estimation, the system displays the most recently acquired image from the internal camera together with the corresponding dross estimation value. Since the image data are directly retrieved from the camera in order to avoid duplicating the storage of already historized information, a dedicated QPRO2 module has been developed and integrated for the Alumil use case. This module enables the visualization of multimedia content obtained via a REST command, configured with a specific placeholder that is dynamically replaced by parameters coming from the field. In this case, the placeholder is populated using the value received from the MQTT topic dedicated to dross estimation.



Figure 110: Dross estimations data

A configuration view of this QPRO2 component is reported below.

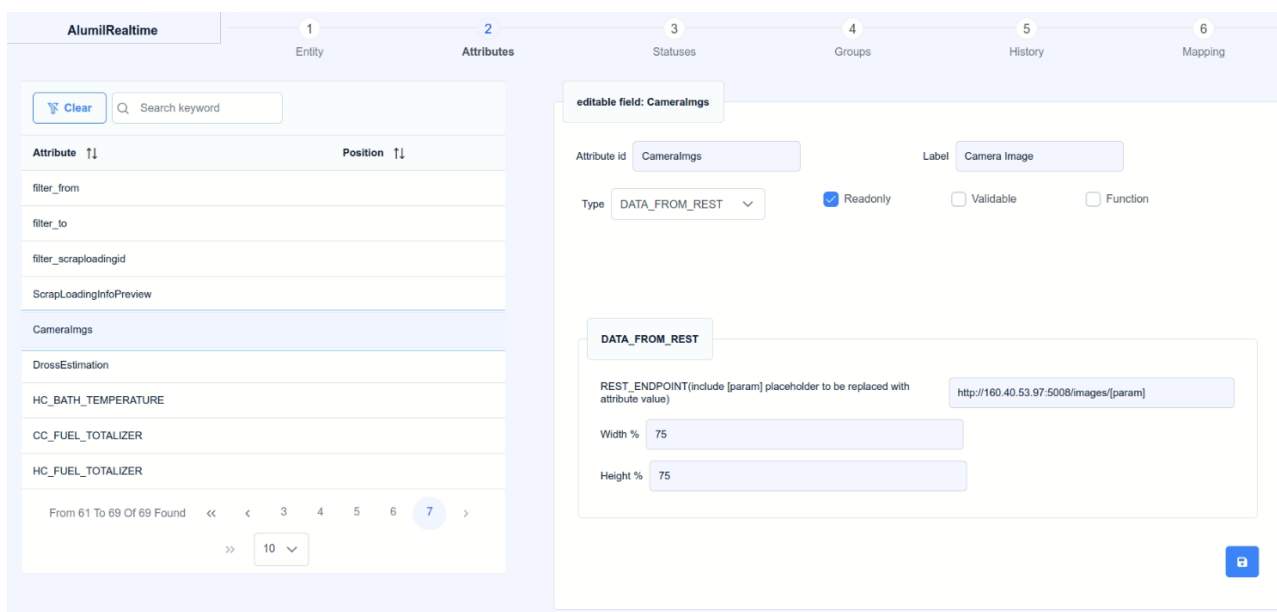


Figure 111: Data attribute from REST service setup

Scrap Inspection

Scrap inspection data are published on a dedicated MQTT topic and visualized through a custom component integrated into the QPRO2 platform. This component retrieves a sequence of images transmitted in JSON format via MQTT and presents them as an automatic slideshow, enriched with the associated metadata.

In a subsequent development phase, the same JSON payload was extended to include polygonal annotations generated by CERTH to identify and classify the type of scrap. As a result, the component was further enhanced to display, within the slideshow, not only the raw images but also an additional standardized QPRO2 component referred to as "image with polygons".

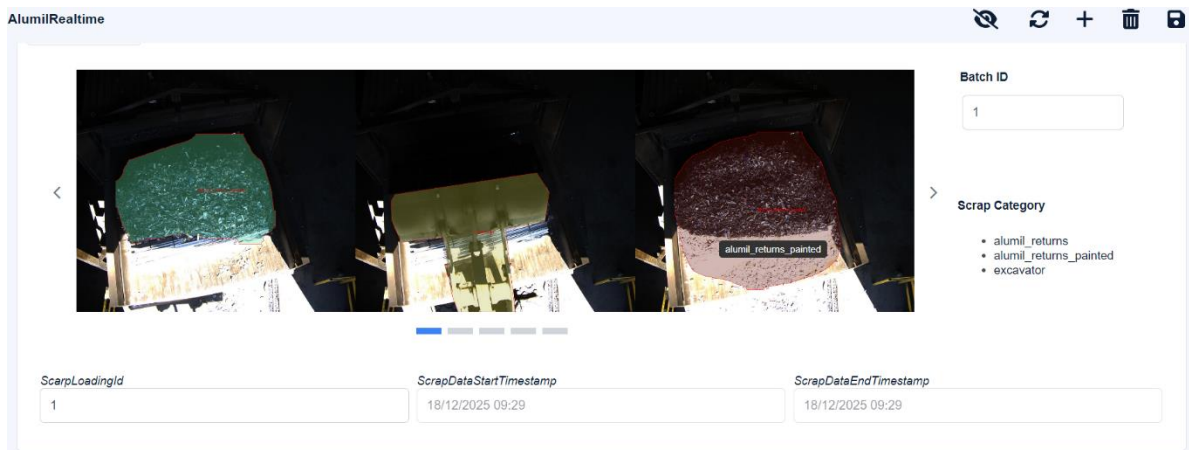


Figure 112: Alumi scrap inspections with polygons

This component can be dynamically populated through field data using a JSON structure that includes both image content and polygon information, according to a predefined schema.

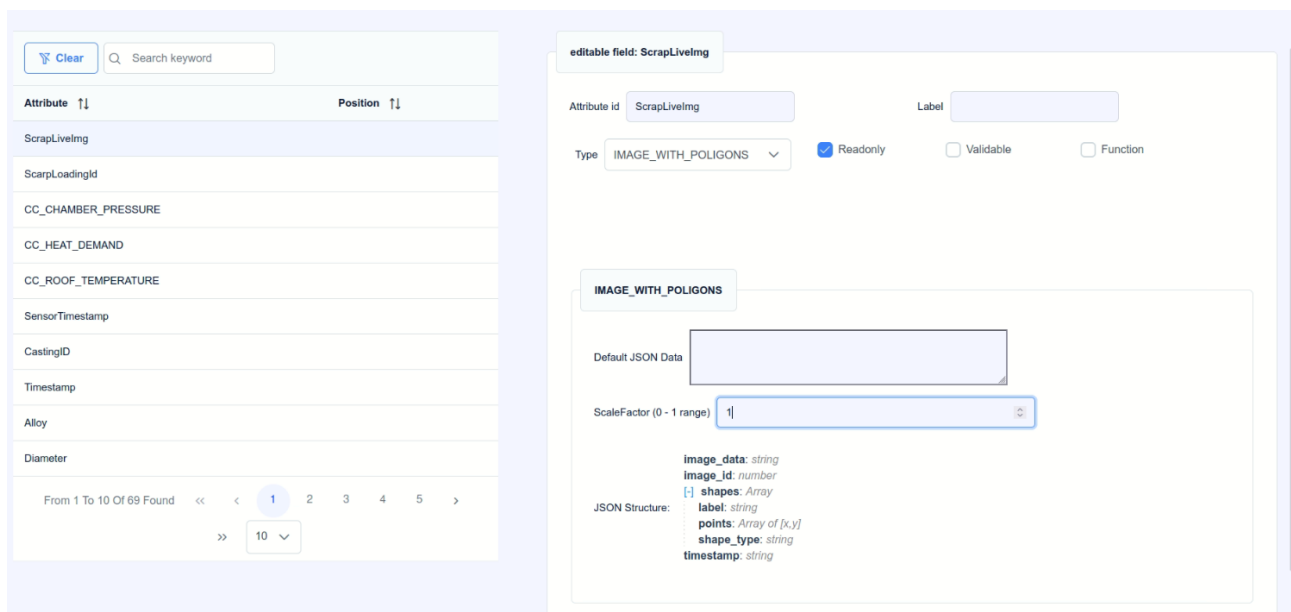


Figure 113: Image with polygons configuration

Additional data streams were initially integrated into the system, including information related to Foundry Production, Oven Loads and Production Stages. However, these data were subsequently dismissed, as they were deemed not relevant with respect to the specific objectives of WP3.

Historical Data Retrieval, Visualization and DSS-Oriented Design

The retrieval of historized data has been designed with a strong focus on usability and user experience, in order to ensure intuitive access to information and to effectively support Decision Support System (DSS) functionalities. This design choice is particularly relevant in view of WP4, where DSS modules will be hosted within the same QPRO2 platform, ensuring architectural consistency and a seamless user interaction paradigm across real-time monitoring and advanced decision-making tools.

To this end, the outcomes of historical data queries across the different elements of the IIoT network are presented in a manner that closely mirrors the runtime visualization. In cases where multimedia content is involved, such as images or

PROJECT 101178230 – ENCIRCLE

annotated visual data, historical views preserve the same representation logic adopted during real-time monitoring, plus the chance to download required data. This approach minimizes cognitive load for end users, allowing them to analyse historical trends and events using familiar visualization patterns without requiring additional training or context switching.

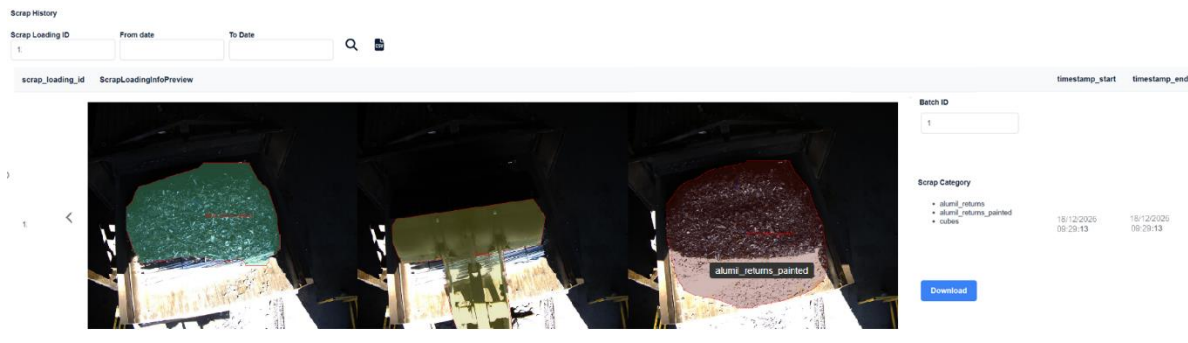


Figure 114: Scrap inspection history with Zip download feature (images and data) for each extracted result

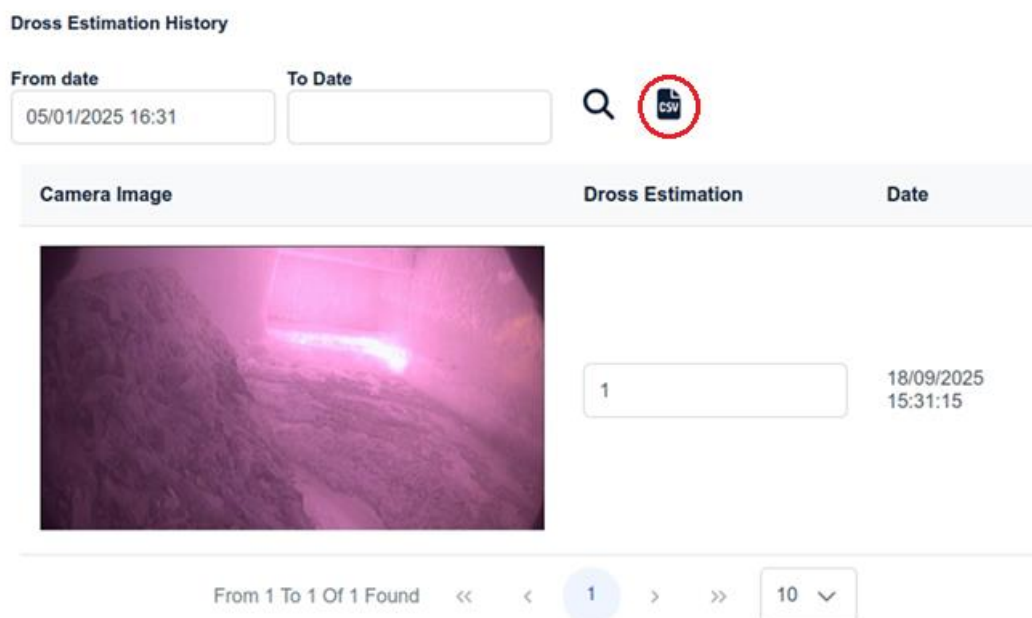


Figure 115: Dross estimations history with CSV download feature (images and data)for each extracted result

With respect to numerical sensor data, historized values are visualized through dedicated historical charts. These charts allow users to explore temporal evolutions, detect trends, identify anomalies and correlate process behaviour over time. The plotting functionalities are designed to support flexible time windows and data aggregation levels, enabling both high-level overviews and fine-grained analysis depending on the specific use case and user role.

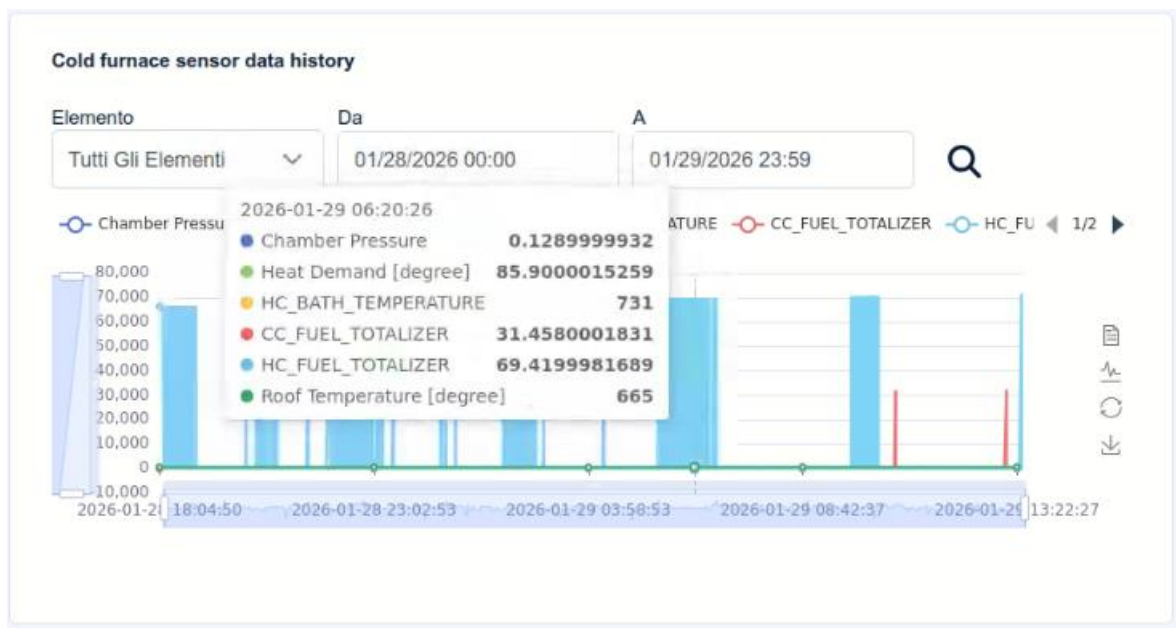


Figure 116: Alumi sensor data plot

Currently, historical data are stored in a dedicated Microsoft SQL Server database, selected to ensure technological alignment with both the QPRO2 platform and the ACQPro middleware, which natively rely on MS SQL Server as their primary data storage solution. The management, retrieval and exposure of historized data are handled by a dedicated microservice, namely QPRO2History, which is responsible for querying the historical database, optimizing access patterns and delivering data to the frontend components in a structured and efficient manner.

From a scalability and performance perspective, the architecture has been designed to support future evolution. In the event that system performance degrades over time—particularly in terms of response latency due to the increasing volume of time-series data—the solution foresees a seamless migration of the most data-intensive time-series to a dedicated time-series database, such as InfluxDB. This database technology is already supported by the full ACQPro–QPRO2–QPRO2History infrastructure, allowing the transition to be performed without major architectural changes. This forward-looking design ensures long-term robustness, scalability and responsiveness of the historical data management layer, even under high data ingestion rates and extended operational periods.

Traceability in the ALUMIL pilot is based on structured data registration and the use of consistent identifiers to link measurements, events and soft-sensor outputs across the full production sequence. The primary traceable entities are the loading and casting batches, identified by a unique Loading / Casting IDs that act as anchors for associating data acquired at different process steps. In practice, ENCIRCLE registers discrete production events (such as a loading) together with timestamps and operational context, enabling synchronization across heterogeneous data sources. Measurements from furnace sensors, internal-camera dross estimation outputs, and scrap inspection streams are consistently linked to the corresponding production entity, allowing the system to reconstruct a complete history from incoming raw material and loading actions up to the produced billets. By combining historized sensor values, encoded multimedia evidence, and AI-derived indicators under the same casting identifier, the pilot supports end-to-end accountability and process-level traceability across the monitored manufacturing workflow.

5.2 IMPLEMENTATION IN PILOT 2– ARCELIK

Data examination is currently in progress. In parallel, sensor installation is ongoing; related progress and results will be reported in the periodic report (M18) and in Deliverable D3.2.

Regarding the middleware layer and QPRO2, the deployment will follow the same setup as in the Alumi pilot: a Docker-based architecture, with MQTT as the main communication protocol between the IoT network and the monitoring platform. The same modules already developed and described in the previous paragraphs will be reused to ensure continuity across acquisition, messaging, storage, and visualization. A dedicated, customized GUI for data inspection and reporting will be implemented once the final set of data fields and visualization requirements is confirmed.

PROJECT 101178230 – ENCIRCLE

The overall data and digitization pipeline is shown in Figure 117. The user, production personnel, will use their mobile device on the shopfloor to capture product images (e.g., dishwasher units) via its camera. The captured images are then forwarded and published to an MQTT broker installed and running on-premise (LAN). Subsequently, the AcqPRO subscribes on the topic for ingestion. AcqPRO stores the incoming data in the local database and, in parallel, forwards the relevant information to QPRO2, where it becomes available through the QPRO2 frontend for monitoring and user interaction.

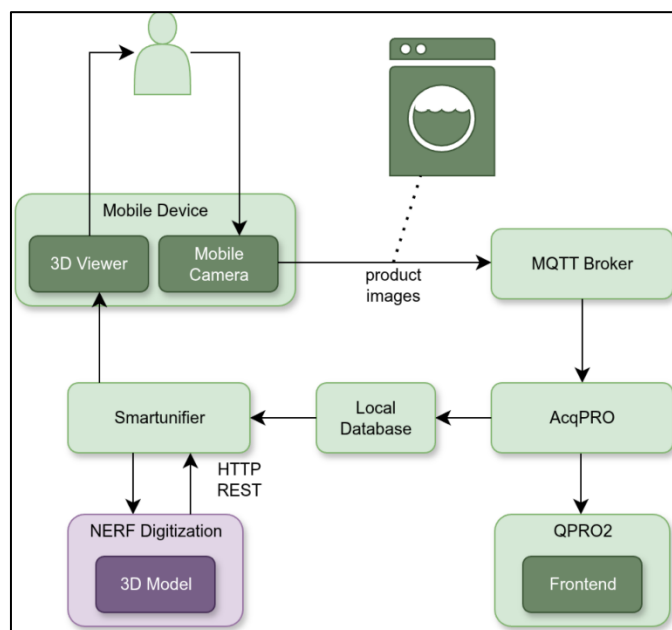


Figure 117: ARCEL pilot implementation plan.

Beyond monitoring, the same captured imagery is used to generate product 3D representations. Smartunifier retrieves the stored images from the local database and forwards them to the NeRF Digitization component. NeRF Digitization reconstructs a 3D model of the product and sends the resulting model back to Smartunifier. Finally, Smartunifier delivers the 3D assets to the mobile device, where they are rendered and inspected through the integrated 3D Viewer.

5.3 IMPLEMENTATION IN PILOT 3 – ZINQ

The architecture depicted in Figure 118 illustrates the IIoT-based data acquisition and integration framework designed for the ZINQ pilot site, enabling structured collection of heterogeneous process data. The system combines process-level parameters (such as bath temperature, pH, and water and energy consumption) with visual data streams acquired through optical and thermal cameras installed along the production line.

All sensor data will be locally aggregated and stored through a dedicated local edge node, where preliminary processing and buffering are performed before transmission. Communication between data sources and higher-level services is handled via a Middleware layer, exposing data through an MQTT broker and RESTful APIs, ensuring low-latency communication. This middleware follows the same architectural principles already validated in the ALUMIL pilot, allowing for seamless integration with ENCIRCLE components such as QPro2 and SmartUnifier.

In addition to raw data acquisition, the proposed architecture is designed to support time-synchronized, multi-modal data fusion, which is essential for the development and deployment of advanced soft sensors. The alignment of process parameters with visual and thermal captures enables the extraction of meaningful correlations between operating conditions and observed product or process states. This combined representation enables downstream predictions such as defect detection, while ensuring that soft sensors operate on consistent and context-aware inputs. The modular design of the Middleware also allows the seamless integration of additional sensing modalities as the system evolves.

From a system-level perspective, the architecture supports both real-time monitoring and historical data analysis, enabling short-term process supervision as well as long-term performance assessment. Edge-level preprocessing reduces commu-

PROJECT 101178230 – ENCIRCLE

nication overhead and increases robustness against network instabilities, while standardized interfaces ensure compatibility with ENCIRCLE's Digital Twin, optimization, and sustainability assessment components. At the current stage, the installation of the optical and thermal cameras is in progress, enabling the initiation of data acquisition activities at the ZINQ pilot site. The experimental evaluation of the proposed soft sensing approaches, based on the data collected through this setup, will be conducted and thoroughly reported in periodic report (M18) and Deliverable D3.2. The full deployment of the IIoT infrastructure, including the integration of data pipelines and operationalization of the monitoring framework, will be finalized within the activities of WP6, ensuring large-scale validation and integration with the overall ENCIRCLE ecosystem.

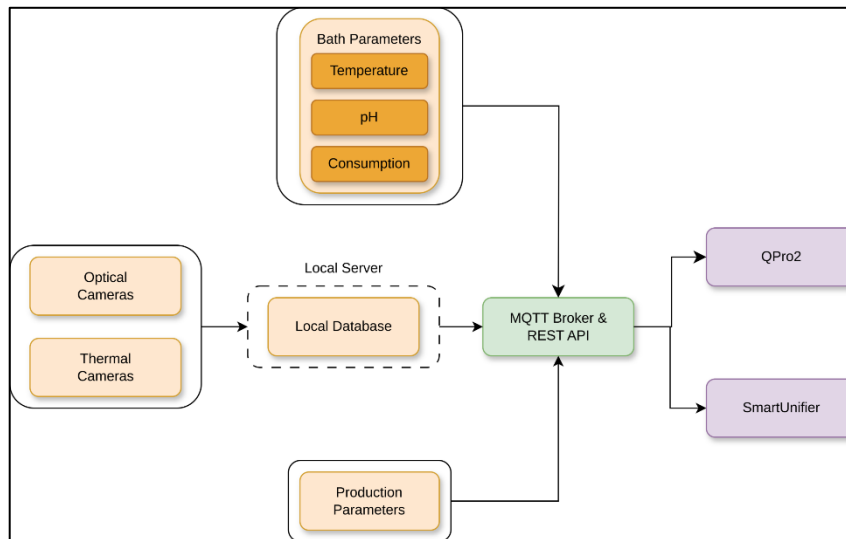


Figure 118: ZINQ pilot implementation plan.

6 CONCLUSIONS

The installed **Sensor Network (Section 3)**, along with the developed **Soft Sensors (Section 4)** establish the sensing and intelligence backbone of WP3, demonstrating how the heterogeneous physical sensing infrastructures built on T3.1 are transformed into actionable process knowledge through advanced soft sensor technologies developed throughout T3.2. The deployed sensor network provides reliable and non-intrusive access to shopfloor data across multiple pilots, while the soft sensors elevate these raw signals into meaningful predictions of material composition, energy demand, dross formation, defect characteristics, and anomalous process behaviour.

Importantly, the integration of foundation-model-based, multimodal, and reasoning-enabled soft sensors extends classical data-driven estimation toward explainable, and zero-shot intelligence, addressing real-world industrial constraints such as data scarcity and process variability. These soft sensor outputs directly support the objectives of WP3 by providing enriched data streams and smart measurements to ENCIRCLE downstream components, including Digital Twins, the Environmental Impact Analytics Engine, and Green Control Loops, enabling simulation, forecasting, optimization, and informed human-in-the-loop digital operation. Furthermore, by capturing early indicators of degradation, defects, abnormal energy consumption, and material inefficiencies across visual, acoustic, textual, and operational domains, the soft sensors lay a robust foundation for Remaining Useful Life (RUL) and Remaining Value estimation of assets and products.

By explicitly addressing key challenges identified in T3.2, such as legacy system integration, data sparsity, real-time constraints, and the need for explainability, the proposed soft sensing framework enables effective human-machine collaboration during industrial procedures. Overall, the results of D3.1 demonstrate how ENCIRCLE operationalizes AI-driven intelligence to support circular manufacturing strategies, asset sustainability and environmental-oriented optimization across diverse industrial domains.

REFERENCES

- [1] Farooq, M. S., Abdullah, M., Riaz, S., Alvi, A., Rustam, F., Flores, M. A. L., ... & Ashraf, I. (2023). A survey on the role of industrial IOT in manufacturing for implementation of smart industry. *Sensors*, 23(21), 8958.
- [2] Afrin, S., Rafa, S. J., Kabir, M., Farah, T., Alam, M. S. B., Lameesa, A., ... & Gandomi, A. H. (2025). Industrial Internet of Things: Implementations, challenges, and potential solutions across various industries. *Computers in Industry*, 170, 104317.
- [3] Chi, H. R., Wu, C. K., Huang, N. F., Tsang, K. F., & Radwan, A. (2022). A survey of network automation for industrial internet-of-things toward industry 5.0. *IEEE Transactions on Industrial Informatics*, 19(2), 2065-2077.
- [4] Hu, Y., Jia, Q., Yao, Y., Lee, Y., Lee, M., Wang, C., ... & Yu, F. R. (2024). Industrial internet of things intelligence empowering smart manufacturing: A literature review. *IEEE Internet of Things Journal*, 11(11), 19143-19167.
- [5] Xu, N., Shahzad, F., & Hu, R. (2025). Unveiling environmental excellence: the confluence of green Industrial Internet of Things, circular economy practices and dynamic capabilities in manufacturing firms. *Journal of Manufacturing Technology Management*, 36(2), 479-503.
- [6] Schneidereit, T., Gohrenz, S., & Breuß, M. (2025, August). Object detection characteristics in a learning factory environment using YOLOv8. In *Intelligent Systems Conference* (pp. 288-308). Cham: Springer Nature Switzerland.
- [7] Tojo Valisoa Andrianandrianina Johanesa, Lucas Equeter, and Sidi Ahmed Mahmoudi. 2024. Survey on AI applications for product quality control and predictive maintenance in industry 4.0. *Electronics* 13, 5 (2024), 976.
- [8] Shashi Bhushan Jha and Radu F Babiceanu. 2023. Deep CNN-based visual defect detection: Survey of current literature. *Computers in Industry* 148 (2023), 103911.
- [9] Yu Mao, Ziyang Chen, Ying Liu, Cong Dong, and Kechen Song. 2025. A survey on industrial image anomaly detection: methods, benchmarks and rethinks. *Measurement* (2025), 118377.
- [10] Abdelrahman Alzarooni, Ehtesham Iqbal, Samee Ullah Khan, Sajid Javed, Brain Moyo, and Yusra Abdulrahman. 2025. Anomaly detection for industrial applications, its challenges, solutions, and future directions: A review. *arXiv preprint arXiv:2501.11310* (2025).
- [11] Yuqi Cheng, Yunkang Cao, Haiming Yao, Wei Luo, Cheng Jiang, Hui Zhang, and Weiming Shen. 2025. A comprehensive survey for real-world industrial defect detection: Challenges, approaches, and prospects. *arXiv preprint arXiv:2507.13378* (2025).
- [12] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. 2021. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International conference on pattern recognition*. Springer, 475–489.
- [13] Jeeho Hyun, Sangyun Kim, Giyoung Jeon, Seung Hwan Kim, Kyunghoon Bae, and Byung Jun Kang. 2024. Reconpatch: Contrastive patch representation learning for industrial anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2052–2061.
- [14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmlR, 8748–8763.
- [15] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. 2023. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19606–19616.
- [16] Bin-Bin Gao, Yue Zhou, Jiangtao Yan, Yuezhi Cai, Weixi Zhang, Meng Wang, Jun Liu, Yong Liu, Lei Wang, and Chengjie Wang. 2025. AdaptCLIP: Adapting CLIP for Universal Visual Anomaly Detection. *arXiv preprint arXiv:2505.09926* (2025).
- [17] Wenxin Ma, Xu Zhang, Qingsong Yao, Fenghe Tang, Chenxu Wu, Yingtai Li, Rui Yan, Zihang Jiang, and S Kevin Zhou. 2025. Aa-clip: Enhancing zero-shot anomaly detection via anomaly-aware clip. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 4744–4754.
- [18] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [19] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

- [20] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology* 16, 5 (2025), 1–72.
- [21] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review* 11, 12 (2024), nwae403
- [22] Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601* (2024).
- [23] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* 1, 2 (2023).
- [24] Tianle Yang, Luyao Chang, Jiadong Yan, Juntao Li, Zhi Wang, and Ke Zhang. 2025. A survey on foundation-model-based industrial defect detection. *arXiv preprint arXiv:2502.19106* (2025).
- [25] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. 2024. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38. 1932–1940.
- [26] Jiacong Xu, Shao-Yuan Lo, Bardia Safaei, Vishal M Patel, and Isht Dwivedi. 2025. Towards zero-shot anomaly detection and reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 20370–20382.
- [27] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024. Llavaonevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326* (2024).
- [28] Beyer L, Dosovitskiy A, A Kolesnikov, et al. 2021. An image is worth 16× 16 words: Transformers for image recognition at scale.
- [29] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. 2019. MVTec AD—A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9592–9600.
- [30] Kang, S., Hu, Z., Liu, L., Zhang, K., & Cao, Z. (2025). Object detection YOLO algorithms and their industrial applications: Overview and comparative analysis. *Electronics*, 14(6), 1104.
- [31] Afifi, M., Derpanis, K. G., Ommer, B., & Brown, M. S. (2021). Learning multi-scale photo exposure correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9157-9167).
- [32] Lee, K., Shin, U., & Lee, B. U. (2024). Learning to control camera exposure via reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2975-2983).
- [33] Okano, M. T., Lopes, W. A. C., & Ruggero, S. M. (2025). Automated dimensional inspection of automotive components using computer vision through YOLO. *Procedia Computer Science*, 270, 3133-3141.
- [34] Jamal, M., Faisal, S., Kusumaningrum, D. S., & Rohana, T. (2024). Application of Yolo V8 for product defect detection in manufacturing companies. *Jurnal Sistem Informasi dan Ilmu Komputer*, 8(1), 1-11.
- [35] Thangakumar, J. (2024, April). Revolutionizing leather quality assurance through deep learning powered precision in defect detection and segmentation by a comparative analysis of Mask RCNN and YOLO v8. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)* (pp. 1-6). IEEE.
- [36] Hussain, M. (2023). YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection. *Machines*, 11(7), 677.
- [37] Rastari, M. A. M., Roslan, R., Hamzah, R., Teo, N. H. I., Shahbudin, F. E., & Samah, K. A. F. A. (2024, August). Recycle waste detection and classification model using YOLO-V8 for real-time waste management. In *2024 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)* (pp. 372-377). IEEE.
- [38] Pan, B., Zhang, X., Lv, Y., & Yu, L. (2022). Automatic optimal camera exposure time control for digital image correlation. *Measurement Science and Technology*, 33(10), 105205.
- [39] Wang, Y., Chen, H., Zhang, S., & Lu, W. (2022). Automated camera-exposure control for robust localization in varying illumination environments. *Autonomous Robots*, 46(4), 515-534.
- [40] Zheng, X., Zheng, S., Kong, Y. and Chen, J. (2021) 'Recent advances in surface defect inspection of industrial products using deep learning techniques', *The International Journal of Advanced Manufacturing Technology*, 113(1–2), pp. 35–58.
- [41] Wang, H., Li, Z. and Wang, H. (2022) 'Few-Shot Steel Surface Defect Detection', *IEEE Transactions on Instrumentation and Measurement*, 71, pp. 1–12.
- [42] Niu, S., Li, B., Wang, X. and Lin, H. (2020) 'Defect Image Sample Generation With GAN for Improving Defect Recognition', *IEEE Transactions on Automation Science and Engineering*, pp. 1–12.
- [43] Wang, Y., Wang, H. and Xin, Z. (2022) 'Efficient Detection Model of Steel Strip Surface Defects Based on YOLO-V7', *IEEE Access*, 10, pp. 133936–133944.

- [44] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S. and Hassabis, D. (2015) 'Human-level control through deep reinforcement learning', *Nature*, 518(7540), pp. 529–533.
- [45] Li, C., Zheng, P., Yin, Y., Wang, B. and Wang, L. (2023) 'Deep reinforcement learning in smart manufacturing: A review and prospects', *CIRP Journal of Manufacturing Science and Technology*, 40, pp. 75–101.
- [46] Evangelidis, A., Dimitriou, N., Charalampous, P., Mastos, T.D. and Tzovaras, D. (2024) 'Efficient Deep Q-Learning for Industrial Equipment Calibration in Elevator Manufacturing', *IEEE Transactions on Industrial Informatics*, 20(10), pp. 12220–12230
- [47] Ronneberger, O., Fischer, P. and Brox, T. (2015) 'U-Net: Convolutional Networks for Biomedical Image Segmentation', *arXiv preprint*, arXiv:1505.04597.
- [48] Chen, L.-C., Papandreou, G., Schroff, F. and Adam, H. (2017) 'Rethinking Atrous Convolution for Semantic Image Segmentation', *arXiv preprint*, arXiv:1706.05587
- [49] Fan, T., Wang, G., Li, Y. and Wang, H. (2020) 'MA-Net: A Multi-Scale Attention Network for Liver and Tumor Segmentation', *IEEE Access*, 8, pp. 179656–179665.
- [50] Yang, B., Liu, Z., Duan, G. and Tan, J. (2024) 'Residual shape adaptive dense-nested Unet: Redesign the long lateral skip connections for metal surface tiny defect inspection', *Pattern Recognition*, 147, 110073.
- [51] Sampath, V., Maurtua, I., Martín, J.J.A., Rivera, A., Molina, J. and Gutierrez, A. (2023) 'Attention-Guided Multitask Learning for Surface Defect Identification', *IEEE Transactions on Industrial Informatics*, 19(9), pp. 9713–9721.
- [52] Guo, B., Wang, Y., Zhen, S., Yu, R. and Su, Z. (2023) 'SPEED: Semantic Prior and Extremely Efficient Dilated Convolution Network for Real-Time Metal Surface Defects Detection', *IEEE Transactions on Industrial Informatics*, 19(12), pp. 11380–11390
- [53] Liu, Z., Yang, B., Duan, G. and Tan, J. (2020) 'Visual Defect Inspection of Metal Part Surface via Deformable Convolution and Concatenate Feature Pyramid Neural Networks', *IEEE Transactions on Instrumentation and Measurement*, 69(12), pp. 9681–9694.
- [54] Lin, H.-I. and Wibowo, F.S. (2021) 'Image Data Assessment Approach for Deep Learning-Based Metal Surface Defect-Detection Systems', *IEEE Access*, 9, pp. 47621–47638.
- [55] Zhou, Q., Chen, R., Huang, B., Xu, W. and Yu, J. (2020) 'DeepInspection: Deep learning based hierarchical network for specular surface inspection', *Measurement*, 160, 107834.
- [56] Shi, H., Li, J., Liang, M., Hwang, M., Hwang, K.-S. and Hsu, Y.-Y. (2023) 'Path Planning of Randomly Scattering Waypoints for Wafer Probing Based on Deep Attention Mechanism', *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53(1), pp. 529–541.
- [57] Luo, W., Zhang, J., Feng, P., Liu, H., Yu, D. and Wu, Z. (2020) 'An adaptive adjustment strategy for bolt posture errors based on an improved reinforcement learning algorithm', *Applied Intelligence*, 51(6), pp. 3405–3420.
- [58] Schulman, J., Wolski, F., Dhariwal, P., Radford, A. and Klimov, O. (2017) 'Proximal Policy Optimization Algorithms', *arXiv preprint*, arXiv:1707.06347
- [59] Cheng, C.-K. and Tsai, H.-Y. (2021) 'Enhanced detection of diverse defects by developing lighting strategies using multiple light sources based on reinforcement learning', *Journal of Intelligent Manufacturing*, 33(8), pp. 2357–2369.
- [60] Lončarević, Z., Gams, A., Reberšek, S., Nemec, B., Škrabar, J., Skvarč, J. and Ude, A. (2021) 'Specifying and optimizing robotic motion for visual quality inspection', *Robotics and Computer-Integrated Manufacturing*, 72, 102200.
- [61] Leontaris, L., Dimitriou, N., Ioannidis, D., Votis, K., Tzovaras, D. and Papageorgiou, E. (2021) 'An Autonomous Illumination System for Vehicle Documentation Based on Deep Reinforcement Learning', *IEEE Access*, 9, pp. 75336–75348.
- [62] Schulman, John & Moritz, Philipp & Levine, Sergey & Jordan, Michael & Abbeel, Pieter. (2015). High-Dimensional Continuous Control Using Generalized Advantage Estimation. 10.48550/arXiv.1506.02438.
- [63] Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., & Zaremba, W. (2016). OpenAI Gym. *arXiv [Cs.LG]*. Retrieved from <http://arxiv.org/abs/1606.01540>
- [64] Ravi, Nikhila, et al. "Sam 2: Segment anything in images and videos." *arXiv preprint arXiv:2408.00714* (2024).
- [65] Ilharco, G., Wortsman, M., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., & Schmidt, L. (2021, July). OpenCLIP [Computer software].
- [66] Podell, Dustin, et al. "Sdxl: Improving latent diffusion models for high-resolution image synthesis." *arXiv preprint arXiv:2307.01952* (2023).
- [67] Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J. Y., & Ermon, S. (2021). Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.
- [68] Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." *ICLR 1.2* (2022): 3.

- [69] Oquab, Maxime, et al. "Dinov2: Learning robust visual features without supervision." arXiv preprint arXiv:2304.07193 (2023).
- [70] Meta AI (facebookresearch). (n.d.). dinov2 [Source code]. GitHub. Retrieved January 14, 2026, from <https://github.com/facebookresearch/dinov2>
- [71] Meta AI (facebookresearch). (n.d.). sam2 [Source code]. GitHub. Retrieved January 14, 2026, from <https://github.com/facebookresearch/sam2>
- [72] del Olmo, J. J. L., Gómez, Á. L. P., Lopez-de-Teruel, P. E., & Ruiz, A. (2024). A few-shot learning methodology for improving safety in industrial scenarios through universal self-supervised visual features and dense optical flow. *Applied Soft Computing*, 167, 112375
- [73] Uygun, E., Terzi, S. Acoustic Monitoring of Railway Defects Using Deep Learning with Audio to Spectrogram Conversion. *J. Vib. Eng. Technol.* 12, 2585–2594 (2024). Stemmer, G., et al. (2024). *Unsupervised Welding Defect Detection Using Audio And Video*. arXiv preprint arXiv:2409.02290
- [74] Zhu, Y., & Liu, X. (2023). A Lightweight CNN for Wind Turbine Blade Defect Detection Based on Spectrograms. *Machines*, 11(1), 99.
- [75] V. Rathinasamy, S. Samsthidhaa Sree and D. Saranyaraj, "Advanced Fault Detection in Industrial Machinery with MFCC-CNN," *2024 4th International Conference on Artificial Intelligence and Signal Processing (AISP)*, VIJAYA-WADA, India, 2024, pp. 01-05, doi: 10.1109/AISP61711.2024.10870634
- [76] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, \., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 6000–6010). Curran Associates Inc..
- [77] Mobtahej, P., et al. (2022). *An LSTM-Autoencoder Architecture for Anomaly Detection Applied on Compressors Audio Data*. Computational and Mathematical Methods
- [78] Tu C-S, Chiu C-K, Tsai M-T. An Audio-Based Motor-Fault Diagnosis System with SOM-LSTM. *Applied Sciences*. 2024; 14(18):8229. <https://doi.org/10.3390/app14188229>
- [79] Gong, Y., Chung, Y.-A., & Glass, J. (2021). *AST: Audio Spectrogram Transformer*. Interspeech 2021
- [80] X. Liu, H. Lu, J. Yuan and X. Li, "CAT: Causal Audio Transformer for Audio Classification," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10096787
- [81] S. Srivastava et al., "Conformer-Based Self-Supervised Learning For Non-Speech Audio Tasks," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, 2022
- [82] Xiao, Y. (2025). Conformer-Based Audio Visual Speech Recognition with Taylor Attention. In *Pattern Recognition* (pp. 297–312). Springer Nature Switzerland
- [83] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer, "Spot-the-difference selfsupervised pre-training for anomaly detection and segmentation," in *European conference on computer vision*. Springer, 2022, pp. 392–408.
- [84] Qingqing Fang, Wenxi Lv, and Qinliang Su, "Af-clip: Zero-shot anomaly detection via anomaly-focused clip adaptation," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 4846–4855.
- [85] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen, "Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection," in *The Twelfth International Conference on Learning Representations*
- [86] Chenghao Deng, Haote Xu, Xiaolu Chen, Haodi Xu, Xiaotong Tu, Xinghao Ding, and Yue Huang, "Simclip: Refining image-text alignment with simple prompts for zero-/few-shot anomaly detection," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 1761–1770.
- [87] Liu, T., Li, B., Jin, X., Shi, Y., Li, Q., & Wei, X. (2025). Exploring Few-Shot Defect Segmentation in General Industrial Scenarios with Metric Learning and Vision Foundation Models. *arXiv preprint arXiv:2502.01216*.
- [88] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.
- [89] Ho, Jonathan, and Tim Salimans. "Classifier-free diffusion guidance." *arXiv preprint arXiv:2207.12598* (2022).
- [90] Batifol, Stephen, et al. "FLUX. 1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space." *arXiv e-prints* (2025): arXiv-2506.
- [91] Ultralytics. (2023). *Ultralytics: YOLOv8* [Computer software]. GitHub. <https://github.com/ultralytics/ultralytics>
- [92] Zhang, T., Wang, G., Xue, C., Wang, J., Nixon, M., & Han, S. (2024). Time-sensitive networking (TSN) for industrial automation: Current advances and future directions. *ACM Computing Surveys*, 57(2), 1-38.
- [93] Trifonov, H., & Heffernan, D. (2023). OPC UA TSN: a next-generation network for Industry 4.0 and IIoT. *International Journal of Pervasive Computing and Communications*, 19(3), 386-411.

- [94] Alotaibi, B. (2023). A survey on industrial Internet of Things security: Requirements, attacks, AI-based solutions, and edge computing opportunities. *Sensors*, 23(17), 7470.
- [95] Gui, J., Chen, T., Zhang, J., Cao, Q., Sun, Z., Luo, H., & Tao, D. (2024). A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12), 9052-9071.