BCSE497J Project-I

Detection of Defects in X-ray Weld Images using Explainable AI

22BDS0137 Gokularajan R.
22BDS0177 Yuva Yashvin
22BKT0076 Ashwin M Felix

Under the Supervision of

Margret Anouncia S

Professor Higher Academic Grade

School of Computer Science and Engineering (SCOPE)

B.Tech.

in

Computer Science and Engineering

(with specialization in Data Science)

School of Computer Science and Engineering

September 2025

ABSTRACT

The quality assurance of welds in critical industries like aerospace and nuclear power relies on the accurate inspection of X-ray images. However, the manual evaluation of these images is often inefficient and inconsistent. While deep learning offers a powerful alternative, the "black box" nature of these models raises major concerns regarding trust and reliability in a high-stakes field like non-destructive testing (NDT). To address this, this review explores recent advances in deep learning for weld defect segmentation, focusing on frameworks that integrate Explainable AI (XAI) to make their decisions transparent and verifiable.

At the core of these advanced systems are modern segmentation networks, including sophisticated U-Net variants, attention-based models, and foundational transformers. These models are engineered to accurately delineate defects like cracks, porosity, and inclusions at a pixel level. A key development is the use of built-in attention mechanisms, which provide implicit explanations by highlighting the image regions the model focuses on. This trend is complemented by the widespread adoption of post-hoc methods like Grad-CAM, which generate visual heatmaps to explicitly justify a model's predictions.

To overcome data scarcity, many projects employ generative augmentation using techniques like GANs to create diverse training data. The development of new approaches, such as weakly-supervised segmentation, further reduces the need for costly pixel-level annotations. The integration of these techniques with human expertise, such as the use of certified inspectors to validate model explanations, is crucial for improving trust and aligning model behaviour with industry standards.

By combining advanced segmentation architectures with a range of XAI methods, these new frameworks deliver a comprehensive, accurate, and trustworthy solution for automated weld defect analysis. This approach is particularly relevant to industries where the confidentiality and safety of radiographic data are paramount. The synthesis of performance and interpretability sets a new benchmark for AI in NDT, demonstrating that robust security and practical usability can coexist.

Keywords: Weld Defect Segmentation, X-ray, Explainable AI, U-Net, Grad-CAM, Attention Mechanisms.

TABLE OF CONTENTS

1. INTRODUCTION

1.1 Background

The industrial sector is increasingly relying on automated systems for critical tasks such as Non-Destructive Testing (NDT). In this domain, radiographic inspection using X-rays is a standard method for quality control, especially for welds in sensitive industries like energy, aerospace, and manufacturing. These X-ray images, which are used to detect internal defects without damaging the material, represent a unique data type. They are often large, contain complex grayscale features, and hold information critical to structural integrity and safety. The traditional approach of manual inspection by human experts is prone to human error, is time-consuming, and can be inconsistent.

The advent of deep learning has revolutionized image analysis, offering a promising alternative for automating the detection of weld defects. Automated defect detection and segmentation models can analyse X-rays with a speed and consistency that manual methods cannot match. However, the widespread adoption of these AI systems has been hindered by significant challenges. Unlike human inspectors who can explain their reasoning, many deep learning models function as "black boxes," providing a prediction without any insight into how they reached that conclusion. This lack of transparency is a major barrier to trust in safety-critical applications, where understanding the basis for a decision is paramount.
A secondary challenge is the scarcity of high-quality, publicly available datasets with pixel-level annotations. Training robust deep learning models requires large, diverse datasets. While some public datasets exist, such as GDXray, they are often small and do not represent the full range of defects found in industrial settings. This data scarcity necessitates the development of advanced techniques like data augmentation and weakly-supervised learning to create effective models from limited information.
To overcome these multifaceted challenges, this project focuses on the design and implementation of an integrated, high-assurance framework for automated weld defect segmentation. The proposed framework provides a holistic solution by combining advanced deep learning models with Explainable AI (XAI) techniques. This approach not only improves the accuracy of defect detection but also makes the model's decision-making process transparent and verifiable, thus building confidence and enabling practical deployment in industrial settings.

1.2 Motivations

The increasing reliance on automated AI systems for safety-critical tasks like weld inspection in industries such as aerospace and nuclear power presents a significant challenge. The "black-box" nature of deep learning models, which offer little to no insight into their decision-making process, directly conflicts with the high-trust and auditability requirements of Non-Destructive Testing (NDT). This creates a critical need for a modern, end-to-end framework that not only accurately identifies defects but also provides transparent and verifiable explanations to build confidence among certified human inspectors.

Key Drivers and Technical Motivations

The core motivation for this project stems from the urgent need for a trustworthy, robust, and explainable framework for automated weld defect analysis. Manual inspection is no longer sufficient to meet the demands of modern quality control, but the transition to AI requires addressing fundamental trust and reliability issues.

- Building Trust and Enhancing Reliability: Our system moves beyond opaque models. By integrating Explainable AI (XAI) techniques like Grad-CAM and attention maps, we provide clear, visual explanations for the model's predictions. This transparency helps human inspectors understand and verify the AI's decisions, which is critical for safety-sensitive applications where a wrong decision could lead to catastrophic failure.
- Mitigating Data Scarcity and Annotation Cost: The lack of large, publicly available datasets with pixel-level annotations is a major bottleneck. Our approach addresses this by leveraging techniques such as generative data augmentation (using GANs) and weakly-supervised learning to create effective models from limited information. This significantly reduces the time and cost associated with manual labelling.
- Improving Defect-Specific Accuracy: Weld defects vary widely in shape and scale, from small, scattered pores to long, thin cracks. By employing sophisticated, multi-scale and attention-based networks, our framework is designed to handle this variability. This ensures the system can accurately segment even minute and irregular defects that are easily missed by traditional methods.
- Enabling Human-in-the-Loop Validation: The AI system is not designed to replace human inspectors but to augment their capabilities. By providing interpretable outputs that can be validated by certified inspectors, we create a feedback loop that continually refines the model's performance and ensures it aligns with rigorous industry standards. This collaborative approach maximizes both efficiency and safety.
- Strengthening Defect Integrity and Verifiability: Our system's outputs, such as segmented masks and explanation heatmaps, are designed for verifiability. This creates an auditable record of the inspection process, which is essential for compliance and quality assurance. Unlike a simple pass/fail output, our framework provides a detailed, evidence-based report of the defect's location, type, and severity.

1.3 Scope of the Project

The scope of this project is to design, implement, and validate a robust and auditable framework for the automated segmentation of weld defects from radiographic images. The primary objective is to develop a system that not only achieves high-accuracy defect segmentation but also makes its decision-making process transparent and interpretable for human inspectors, addressing the critical "black box" problem in AI-driven NDT.

To achieve this, the project will first establish a preprocessing pipeline to normalize and enhance low-contrast radiographic images, utilizing techniques like Contrast Limited Adaptive Histogram Equalization (CLAHE) to improve defect visibility. The core of the framework will be a state-of-the-art deep learning segmentation model, such as a U-Net architecture or a YOLOv8-Seg variant. This model will be meticulously trained to generate precise, pixel-level masks for a variety of critical weld defects, including porosity, cracks, and slag inclusions.

To move beyond a simple detection system and create a truly trustworthy tool, the project will integrate and evaluate cutting-edge Explainable AI (XAI) methods. Specifically, we will leverage Grad-CAM to generate visual heatmaps that provide causal insights into the model's predictions. This enhanced XAI approach will serve three key functions:

Model Validation: The heatmaps will be used to verify that the model's feature attribution aligns with the physical characteristics of the defects, confirming it is "looking" at the right places and not relying on spurious correlations like film artifacts or markers.

Diagnostic Analysis: In cases of model failure (false positives or negatives), the XAI explanations will be used as a powerful debugging tool to diagnose the root cause, revealing model biases and guiding targeted data augmentation for future improvements.

Enhanced Human-AI Collaboration: The visual explanations will transform the system from a black-box oracle into an interactive assistant. Inspectors can use the heatmaps to rapidly confirm or challenge the AI's findings, creating an efficient human-in-the-loop workflow and building confidence in the automated system.

Finally, all model outputs—including the segmented masks, confidence scores, and XAI heatmaps—will be systematically logged to create a transparent and verifiable audit trail for each decision. The framework's performance will be quantitatively validated using standard segmentation metrics, such as mean Intersection-over-Union (mIoU) and the Dice coefficient, on a publicly available weld radiograph dataset. This validation will demonstrate not only the model's segmentation accuracy but also the practical utility of its integrated explanations, thereby establishing a new benchmark for trustworthy and explainable AI in the field of Non-Destructive Testing.

2. PROJECT DESCRIPTION AND GOALS

2.1 Literature Review

This review surveys deep learning methods for segmenting weld defects in X-ray images, with a focus on explainable AI (XAI). We cover segmentation networks (CNN/U-Net, attention models, transformers) and how they highlight defects, datasets and defect types, evaluation metrics, and XAI integration (e.g. attention maps, Grad-CAM, LIME). We summarize key contributions and trends, and note research gaps.

Early approaches applied encoder–decoder CNNs like U-Net architectures to weld X-ray images. For example, [1] proposed a weakly-supervised semantic segmentation: a Cascade R-CNN first finds a defect's bounding box, then adaptive-thresholding within the ROI generates a mask.

[2] built MAU-Net, a U-Net variant with multi-scale convolution blocks and attention gates. They augment scarce data via a dual-attention DCGAN, then add convolutional block attention in the encoder to boost detection of small defects. On the GDXray weld image dataset, this approach reached 84.75% segmentation accuracy.

Likewise, [3] engineered a classical image-processing pipeline (no deep learning) with filters to segment horizontal-shaped defects on the same GDXray dataset. This shows CNNs are not the only option for low tolerance weld defect identification.

[4] used a "Grad-MobileNet" unsupervised method: instead of segmentation, it flags anomalies via gradients on a MobileNet's – a pretrained CNN - feature maps (essentially a one-class CAM (class activation map)). Such methods highlight defects and integrate gradient visualization (akin to Grad-CAM) for explanations.

[5] used Explainable AI (XAI) techniques, such as Grad-CAM and LIME to highlight which pixels influenced each classification and had ASNT-certified inspectors verify them. This "human-in-the-loop" XAI process improves trust by aligning the model's attention with expert expectations

2.2 Gaps Identified

Limited XAI in Segmentation: Most current weld AI research focuses on accuracy. Few segmentation studies explicitly apply XAI post-hoc. This contrasts with classification, where papers often include Grad-CAM/LIME

Explainability Metrics: The field lacks agreed metrics for explanation quality in Non-Destructive Testing. Developing quantitative XAI evaluations (e.g. localization error, faithfulness tests) is an open problem. Industry frameworks will likely emerge requiring explainability. Systems will need metrics to justify the explainability.

While deep learning has demonstrated significant potential for automating weld defect segmentation, the field is hindered by a critical interpretability gap. Current research overwhelmingly prioritizes segmentation accuracy, largely overlooking the "black box" nature of the models used. There is a notable disparity where post-hoc Explainable AI (XAI) methods like Grad-CAM are common in classification tasks but remain underexplored and unvalidated for the more nuanced task of radiographic segmentation. This gap is further compounded by the absence of standardized, quantitative metrics to evaluate the faithfulness and utility of these explanations within the NDT context. Consequently, critical questions regarding the model's trustworthiness, its true benefit in human-in-the-loop workflows, and its ability to quantify its own uncertainty remain largely unanswered, limiting the transition of these powerful models from academic benchmarks to certified industrial practice

2.3 Objectives

We want to build a system that not only performs with high accuracy but also earns the trust of human experts. We also want to build a model that visualize attention attention maps that attempts to locate and draw a bounding box around a defect and to identify it’s type in python This project's objectives are designed to create a framework that is genuinely reliable and transparent from the ground up, addressing the complete lifecycle of AI-driven weld inspection.

2.3.1 Ensure Accurate and Granular Defect Segmentation

This project will implement a deep learning framework capable of performing pixel-level semantic segmentation of weld defects. By using sophisticated architectures, we will not only detect defects but also accurately delineate their precise boundaries and classify their type (e.g., pore, crack, lack of fusion). This granular detail will provide human inspectors with a level of precision that is difficult to achieve with manual methods.

2.3.2 Implement Explainable AI for Enhanced Trust

Instead of relying on an opaque "black-box" model, this project will develop and implement a strategy for a transparent and interpretable system. By integrating Explainable AI (XAI) techniques like Grad-CAM and attention maps, we will make the model's decision-making process visible to human users. This approach will demonstrate why a specific region was identified as a defect, helping to build confidence and streamline the validation process in safety-critical applications.

2.3.3 Establish a Comprehensive and Auditable Framework

The project will adopt a "traceable-by-design" approach to ensure a comprehensive and auditable record of the inspection process. We will systematically store all critical

system outputs, including the original images, segmentation masks, classification results, and explanation heatmaps. This creates a tamper-evident log that guarantees any automated analysis can be immediately verified and reviewed by a certified inspector, ensuring compliance and accountability.

2.4 Problem Statement

The quality and safety of critical industrial components, such as welded structures in aerospace and nuclear power, depend on meticulous inspection. However, our growing dependence on automated AI systems for this task has introduced a fundamental challenge: the "black box" problem. Many existing deep learning models, while highly accurate, fail to provide any insight into their decision-making process. This lack of transparency creates a major roadblock for their adoption in safety-critical applications, where trust and accountability are paramount.

This problem is further exacerbated by a number of factors. A key vulnerability lies in the continued reliance on legacy, labor-intensive methods of manual inspection, which are slow, inconsistent, and prone to human error. While automated systems offer a path to greater efficiency, the lack of explainability means their outputs cannot be trusted on their own. This forces human experts to spend valuable time verifying every AI-generated decision, creating a significant bottleneck that undermines the potential of automation.

Ultimately, current systems struggle to strike a viable balance between high performance and interpretable outputs. This fundamental imbalance creates a significant roadblock for critical applications that require not only high-speed, accurate processing but also uncompromising trust and verifiable results to function effectively and safely.

2.5 Project Plan

The entire project plan can be summarized in a single paragraph, outlining the key phases in a concise, logical flow. We will execute this 12-week project in three main phases. We'll begin with a planning phase, spending the first two weeks on foundational research and system design. This is followed by a seven-week development phase, where we will build the core AI engine and its associated pipelines for data preprocessing, model training, and explainability. The final phase, lasting three weeks, will be dedicated to integrating the components, rigorous testing and validation to confirm the system's accuracy and reliability, and the preparation of all final project documentation. This approach ensures that we build a robust and trustworthy system from the ground up, delivering a final product ready for demonstration.

3. SYSTEM SPECIFICATIONS

The success of this project depends on the right mix of hardware and software. This section outlines the specifications required for an effective development and testing environment for our explainable weld defect segmentation system.

3.1 Hardware Specifications

While the framework is designed to be efficient, training and running deep learning models for image segmentation are computationally intensive tasks, with the GPU being the most critical component.

- Graphics Processing Unit (GPU): A modern NVIDIA GPU is essential. It's the core of the system and is responsible for accelerating the complex matrix operations of the deep learning model. A card with at least 12 GB of VRAM is recommended to handle large datasets and high-resolution X-ray images. Models like the NVIDIA RTX 4070 or better are ideal.
- Processor (CPU): A multi-core processor is necessary for data preprocessing and general system tasks, but the GPU will handle the heavy lifting of model training. An Intel Core i5 or AMD Ryzen 5 (or equivalent) will be sufficient to prevent bottlenecks.
- Memory (RAM): A minimum of 16 GB of RAM is recommended. The system RAM holds the dataset and manages the data pipeline, and having enough memory is crucial for loading images efficiently before they're processed by the GPU.
- Storage: A Solid-State Drive (SSD) is highly recommended. Fast read/write speeds are critical for loading large image datasets quickly, significantly reducing training and inference times. A capacity of 512 GB or more is suitable for storing the dataset and model checkpoints.

3.2 Software Specifications

Our system is built on an open-source, Python-based stack, leveraging popular deep learning and computer vision libraries.

- Operating System: Our development is primarily on Windows, which offers excellent support for deep learning frameworks and GPU drivers.

- Programming Language: The entire framework is written in Python (version 3.9 or newer). Python's extensive ecosystem of libraries makes it the ideal choice for this project.
- Core Libraries:
    - PyTorch: A modern deep learning framework is the foundation of our segmentation model. PyTorch provides robust tools for building, training, and deploying a U-Net architecture.
    - NumPy / SciPy: These are fundamental libraries for numerical operations and array manipulation, which are essential for handling image data and model outputs.
    - PyTorch-Grad-CAM / LIME: These libraries are crucial for implementing the Explainable AI component of our framework. They allow us to generate the visual explanations that demonstrate how the model arrived at its predictions

4 REQUIREMENT ANALYSIS (SRS)

4.1 Functional Requirements

These are the specific functions our system must perform to achieve the objectives of accurate and explainable weld defect segmentation.

- FR-1: Image Upload and Ingestion: The system must allow users to upload X-ray images of welds for analysis. Upon upload, the system shall ingest the image into the processing pipeline.
- FR-2: Automated Preprocessing: The system must automatically perform the following preprocessing steps on each uploaded image:
    - FR-2.1: Image Cropping: The system shall crop the image to isolate the weld seam, removing unnecessary background and surrounding annotations.
    - FR-2.2: Data Augmentation: The system shall apply a series of data augmentation techniques (e.g., rotation, scaling, contrast adjustment) to expand the dataset and improve model robustness.
- FR-3: Defect Segmentation: The system must process the cropped weld image using a deep learning model to perform pixel-level segmentation. The output should be a corresponding mask that outlines the precise location and shape of any defects.
- FR-4: Defect Classification: The system must classify each segmented defect into one of the predefined categories: pore, crack, lack of fusion, penetration, or slag inclusion. It must also be able to identify cases with no defects.
- FR-5: Bounding Box Generation: The system shall automatically draw a bounding box around each detected and segmented defect to provide a clear, high-level localization.
- FR-6: Explainability Integration: The system must provide a visual explanation of its segmentation and classification decisions. For each detected defect, it shall generate a

Grad-CAM heatmap that highlights the specific image regions that influenced the model's output.

- FR-7: Output Visualization: The system must present the results in an easy-to-understand format, including the original X-ray image with the segmented masks, bounding boxes, and explanation heatmaps overlaid on it for human review.
- FR-8: Result Storage: The system shall store the original image, the segmentation masks, the classification results, and the corresponding explainability outputs to create a verifiable record of the analysis.

4.2 Non-Functional Requirements

These requirements define the quality, standards, and operational attributes of our system. For an AI system in a safety-critical field like NDT, these are just as important as the functional requirements.

- NFR-1: Trustworthiness and Reliability: This is our top priority. The system's predictions must be consistent and reproducible. The model should provide the same output for the same input, and its explainable outputs (heatmaps, masks) must accurately reflect the model's decision-making process. The system must also be robust to minor variations in image quality, ensuring reliable performance in different conditions.
- NFR-2: Performance: The system must be fast enough for practical use in an industrial environment. For a standard high-resolution X-ray image, the end-to-end processing time—including preprocessing, segmentation, and explanation generation—should ideally be within a few seconds. The training of the deep learning model should be optimized to complete within a reasonable timeframe (e.g., a few hours) on the recommended hardware.
- NFR-3: Scalability: The framework should be able to handle a growing volume of images without a significant drop in performance. The architecture must be designed to allow for horizontal scaling, enabling the system to process multiple images concurrently by leveraging additional GPU resources.
- NFR-4: Interpretability and Usability: The system's output must be intuitively understandable to a domain expert, specifically a certified weld inspector. The visual explanations (segmentation masks and Grad-CAM heatmaps) should be clear and easy to interpret, empowering the user to make confident, informed decisions. The user interface should be simple, providing a seamless experience for uploading images and reviewing results.

5 SYSTEM DESIGN AND ARCHITECTURE

The system workflow is initiated when an NDT Inspector uploads a raw radiograph via the User Interface. The image immediately enters a Core Processing Pipeline, where it is first enhanced by a preprocessing module. The enhanced image is then analyzed by a Segmentation Model to generate a precise defect mask, while simultaneously, an Explainable AI (XAI) module uses Grad-CAM to create a visual heatmap explaining the model's decision. Finally, all results—the original image, the segmentation overlay, and the XAI heatmap—are logged for a complete audit trail and displayed to the inspector, providing a transparent and fully interpretable analysis. *Refer to Figure 1.*

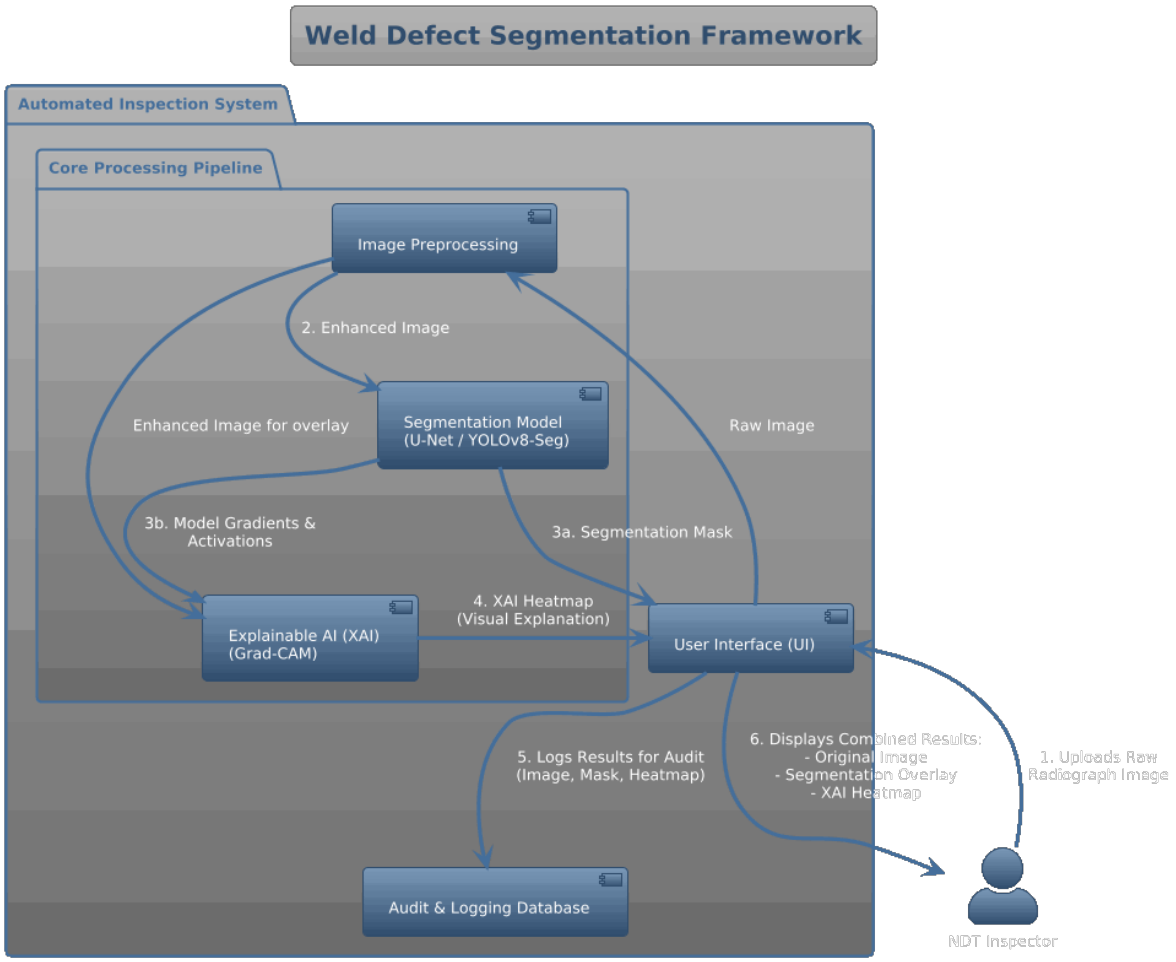


Figure 1, workflow overview

The workflow for creating the expert YOLOv8 model is a systematic, two-phase process (Figure 2).

The first phase, Data Preparation, involves an expert annotator creating precise, pixel-level masks for each defect on the radiographic images to establish the ground truth. This annotated dataset is then strategically split into training, validation, and test sets.

In the second phase, Model Training & Evaluation, transfer learning is employed to fine-tune a pre-trained YOLOv8 model. The model iteratively learns from the training data in a loop

that includes data augmentation for robustness, followed by prediction, loss calculation, and weight updates via backpropagation. The model's progress is monitored on the validation set, and the best-performing version (best.pt) is saved. Finally, this trained model is subjected to a rigorous evaluation on the unseen test set to generate its definitive performance metrics.

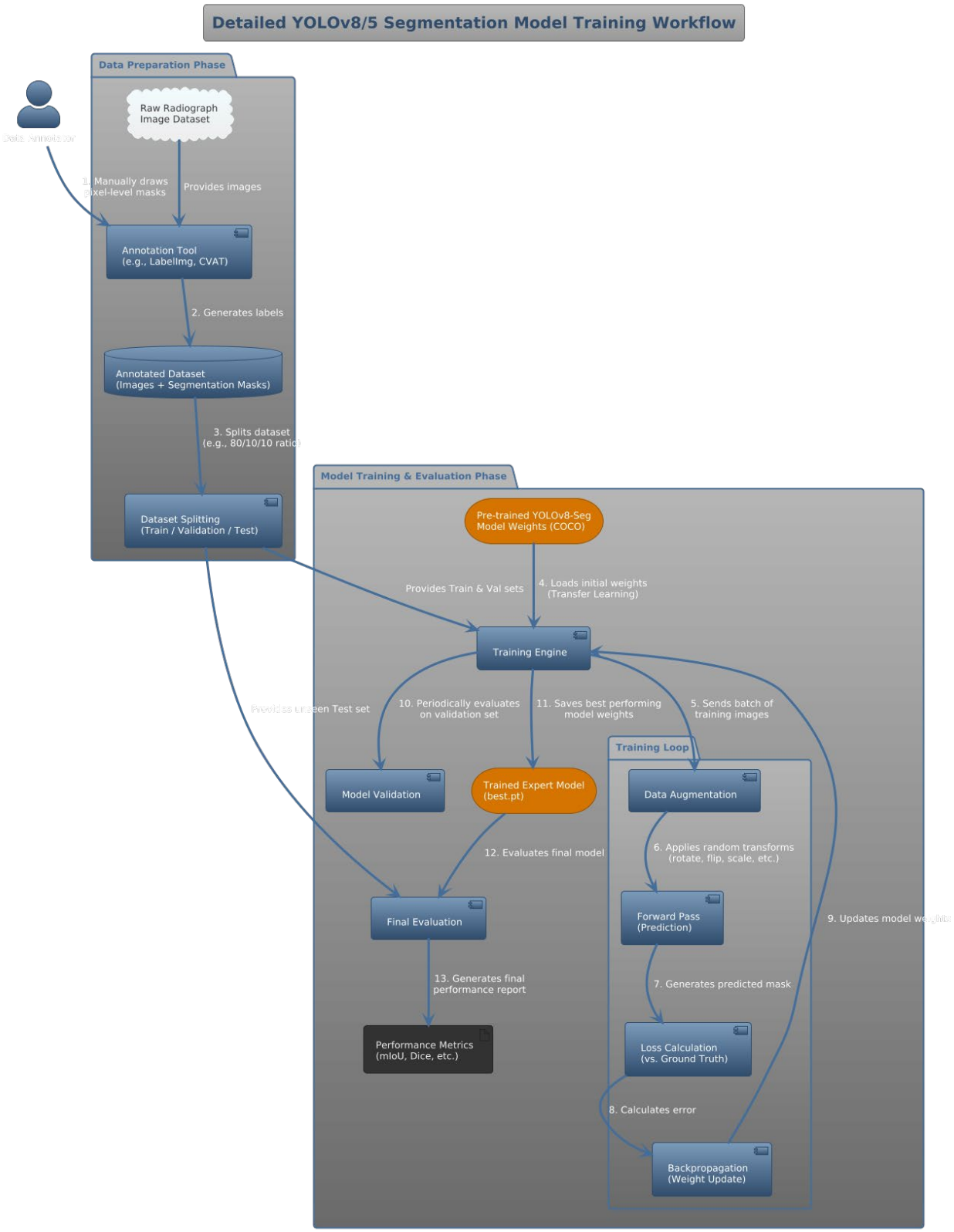
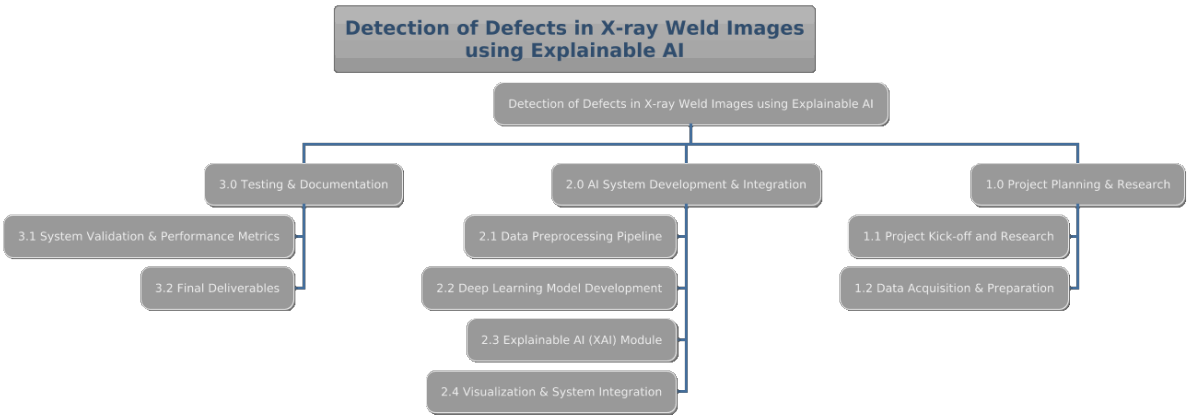


Figure 2, YOLOv8 model training/fine-tuning diagram

6 PROJECT MANAGEMENT

This section provides an overview of our project plan, including how we've broken down the work and our timeline for getting it all done.

6.1 Work Breakdown Structure (WBS)



The Work Breakdown Structure (WBS) for this project is a hierarchical decomposition of all the work required to complete the project objectives. It is organized into three main phases, with each phase broken down into specific, manageable work packages.

1.0 Project Planning & Research

- 1.1 Project Kick-off and Research:
    - 1.1.1 Conduct literature review on deep learning for weld inspection.
    - 1.1.2 Research suitable XAI techniques (e.g., Grad-CAM).
    - 1.1.3 Define project scope and success criteria.
- 1.2 Data Acquisition & Preparation:
    - 1.2.1 Secure and review the 20 X-ray weld images.
    - 1.2.2 Perform initial data cleaning and labelling.
    - 1.2.3 Plan data augmentation strategy.

2.0 AI System Development & Integration

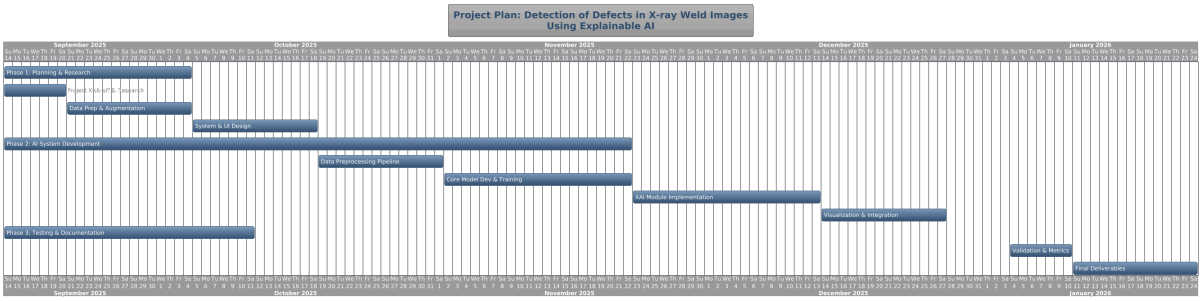
- 2.1 Data Preprocessing Pipeline:

    - 2.1.1 Develop script for image cropping and annotation removal.
    - 2.1.2 Implement data augmentation techniques.
- 2.2 Deep Learning Model Development:
    - 2.2.1 Train the model (YOLOv5/8, U NET) on the prepared dataset.
    - 2.2.2 Integrate defect classification module.
- 2.3 Explainable AI (XAI) Module:
    - 2.3.1 Implement Grad-CAM for generating explanation heatmaps.
    - 2.3.2 Integrate XAI outputs with the main model.
- 2.4 Visualization & System Integration:
    - 2.4.1 Develop scripts for overlaying masks and bounding boxes.
    - 2.4.2 Create a unified pipeline connecting all components.

3.0 Testing & Documentation

- 3.1 System Validation & Performance Metrics:
    - 3.1.1 Conduct end-to-end system testing.
    - 3.1.2 Calculate and report segmentation metrics
- 3.2 Final Deliverables:
    - 3.2.1 Prepare final project report and documentation.
    - 3.2.2 Create a project demonstration for presentation.

6.2 Gantt Chart



7. IMPLEMENTATION

The implementation is a Python script that integrates a trained YOLOv8 segmentation model with the pytorch-grad-cam library to deliver a comprehensive, explainable analysis. The script begins by loading the custom-trained best.pt model and a target radiographic image. It then performs inference to generate precise segmentation masks and bounding boxes for any detected defects. For the highest-confidence detection, it initializes Grad-CAM, targeting a final convolutional layer in the model's backbone. Using a SemanticSegmentationTarget, it generates a heatmap that visually explains which pixels were most influential for that specific segmentation. Finally, the script produces a clear, multi-panel visualization that presents the original detection, the segmentation mask, and the XAI heatmap side-by-side, offering a complete and auditable result.

Image processing and segmentation step:

```
print(f"--- Step 2: Processing Image: {image_path} ---")
    # Open the image using OpenCV
    original_img = cv2.imread(image_path)
    # Convert from BGR (OpenCV default) to RGB for Matplotlib and model
    original_img_rgb = cv2.cvtColor(original_img, cv2.COLOR_BGR2RGB)

    print("--- Step 3: Performing Defect Segmentation ---")
    # Run inference on the image
    results = model(original_img_rgb, verbose=False)
    result = results[0] # Get the first result object

    # Check if any defects were detected
    if result.masks is None:
        print("\n--- No defects detected. ---")
        plt.imshow(original_img_rgb)
        plt.title('Original Image (No Defects Detected)')
        plt.axis('off')
        plt.show()
        return
```

Grad-CAM step:

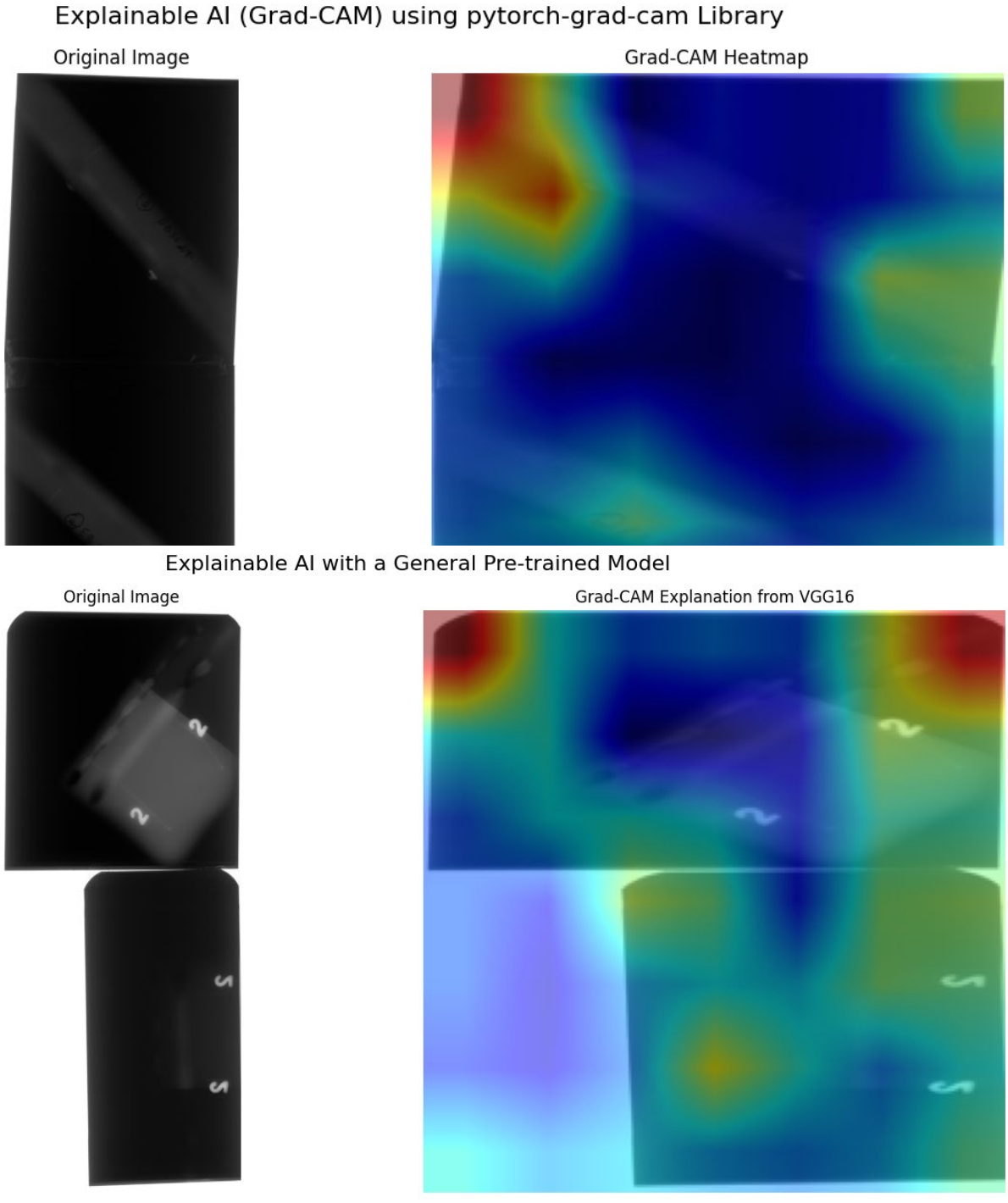
```
    # We will explain the detection with the highest confidence score
    highest_conf_idx = result.boxes.conf.argmax()

    # Get the class index for this specific detection
    target_class_index = int(result.boxes.cls[highest_conf_idx].item())
    target_class_name = model.names[target_class_index]

    print(f"Generating Grad-CAM for the highest confidence detection:
'{target_class_name}'")

    # Define the target layer for Grad-CAM. For YOLOv8's segmentation head,
    # the final layers of the backbone are good candidates.
    target_layer = pytorch_model.model[-2] # Corresponds to the C2f block
in the backbone

    # Define the target for Grad-CAM. For segmentation, we want to see
    # what pixels influenced the prediction for a specific class.
    cam_target = SemanticSegmentationTarget(category=target_class_index,
mask=result.masks.data[highest_conf_idx])

    # Initialize Grad-CAM
    cam = GradCAM(model=pytorch_model, target_layers=[target_layer])

    # Generate the heatmap
    # Note: We use result.plot(show=False) to get the preprocessed tensor
YOLO used
    input_tensor = result.plot(show=False, pil=True) # Trick to get the
right input tensor
    grayscale_cam = cam(input_tensor=input_tensor, targets=[cam_target])[0,
:]
```

Final side-by-side plot (original image and XAI generated heatmap output):

```
resized_img_np = result.orig_img

    # Create the CAM overlay
    cam_overlay = show_cam_on_image(resized_img_np / 255.0, grayscale_cam,
use_rgb=True)

    # Get the combined segmentation mask overlay from the results object
    segmentation_overlay = result.plot(boxes=False) # Plot only masks
    segmentation_overlay = cv2.cvtColor(segmentation_overlay,
cv2.COLOR_BGR2RGB)

    # --- Create a detailed 1x3 plot for the presentation ---
    fig, axs = plt.subplots(1, 3, figsize=(20, 7), dpi=100)
    fig.suptitle(f"Comprehensive Analysis for '{target_class_name}'
Defect", fontsize=20)

    # Panel 1: Original Image with Bounding Box
    box_img = result.plot(masks=False, labels=True, conf=True) # Plot only
boxes and labels
    box_img = cv2.cvtColor(box_img, cv2.COLOR_BGR2RGB)
    axs[0].imshow(box_img)
    axs[0].set_title("1. Model Detection", fontsize=16)
    axs[0].axis('off')

    # Panel 2: Segmentation Mask Overlay
```

```
axs[1].imshow(segmentation_overlay)
axs[1].set_title("2. Segmentation Mask", fontsize=16)
axs[1].axis('off')

# Panel 3: XAI Heatmap Explanation
axs[2].imshow(cam_overlay)
axs[2].set_title("3. XAI Explanation (Grad-CAM)", fontsize=16)
axs[2].axis('off')

plt.tight_layout(rect=[0, 0.03, 1, 0.95])
plt.show()
```

Sample Output:



8. REFERENCES

[1] : Welding defects classification by weakly supervised semantic segmentation, Zhang et al. (2023)
[2] : A new method for deep learning detection of defects in X-ray images of pressure vessel welds, Xue et al. (2024)
[3] : Accurate segmentation of weld defects with horizontal shapes, Doaa Radi et al. (2022)
[4] : Grad-MobileNet: A Gradient-Based Unsupervised Learning Method for Laser Welding Surface Defect Classification, Han et al. (2022)
[5] : Advancing Welding Defect Detection in Maritime Operations via Adapt-WeldNet and Defect Detection Interpretability Analysis, Bansha et al. (2025)