Real-Time Industrial Defect detection with computer vision using YOLOv12 model

Roshni Sundrani[1]
Student, M.E. (Artificial Intelligence & Data Science),
AISSMS COE, Pune,
Maharashtra, India

Shashikant .V. Athawale[2]
Associate Professor, Department of Computer Engineering,
AISSMS COE, Pune,
Maharashtra, India

***Abstract*—**
In industrial manufacturing, early detection of surface defects such as corrosion, scratches, pinholes, and chemical patches is vital to ensure product integrity and minimize production losses. This research presents an AI-driven defect detection system designed for ABC Filters Pvt. Ltd., focused on real-time identification of surface-level defects on cylindrical filter designs transported via a conveyor belt.
The proposed pipeline uses a two-stage deep learning approach based on the YOLOv12 architecture. Initially, a custom-trained YOLO12-Segmentation model accurately identifies and segments individual filter designs from conveyor belt video frames. The segmentation dataset was annotated using the Segment Anything Model (SAM) and prepared using Roboflow, enabling the model to achieve 99% segmentation accuracy in isolating cylindrical filter units from surrounding elements. Subsequently, each segmented filter is passed to a YOLO12m model, custom-trained to detect four defect types: corrosion, scratch, pinhole, and chemical patch. The system currently achieves an average defect detection accuracy of approximately 70%, with scope for improvement through enhanced image quality, dataset expansion, and adoption of newer YOLO variants.
Video processing is performed using OpenCV to analyze frames in real-time. The pipeline is designed for deployment on NVIDIA Jetson Nano, enabling on-device inference from live camera feeds without cloud dependency. This approach demonstrates practical potential for scalable, real-time defect detection in production environments, reducing reliance on manual inspection while improving consistency and speed.

Keywords: Defect detection, custom trained YOLOv12 model, artificial intelligence, computer vision, image processing, Open CV, Segmentation, SAM

1. Introduction

In modern manufacturing industries, ensuring product quality through early-stage defect detection has become a critical factor for maintaining competitiveness and minimizing operational losses. Manual inspection methods are often time-consuming, inconsistent, and prone to human error, especially in fast-paced assembly line environments. As manufacturing scales up and product standards become more stringent, automated and intelligent inspection systems have emerged as a viable solution to reduce inspection time, cost, and error rates.

ABC Filters Pvt. Ltd., a company specializing in the production of cylindrical filters, faces challenges in detecting surface-level defects during production. Common defects such as corrosion, pinholes, scratches, and chemical patches may arise due to material inconsistencies, handling errors, or environmental factors during the manufacturing process. These defects, if not identified promptly, can compromise the performance and lifespan of the filters, leading to product failures and customer dissatisfaction, even it becomes life threatening if such defected filters are installed into vehicles during manufacturing production.

This research introduces a deep learning-based defect detection system tailored for real-time quality control on the production line. Leveraging the YOLO12 architecture, the proposed system utilizes a two-stage pipeline. First, individual filter components are segmented from the conveyor belt stream using a YOLOv12-Segmentation model, trained on a dataset annotated with high-precision masks generated using the Segment Anything Model (SAM). The segmentation model demonstrates exceptional performance with an accuracy of 99% in identifying filter units, even under varied lighting and background conditions.

In the second stage, the segmented filter regions are passed through a YOLOv12m object detection model trained to identify and classify four specific defect types. Although the current defect detection model achieves an accuracy of approximately 70%, it provides a strong baseline for deployment, with further improvements anticipated through dataset expansion, higher-resolution imagery, and exploration of newer model variants.

To ensure seamless deployment in a real-time industrial environment, the system integrates OpenCV for live video frame extraction and processing. The entire inference pipeline is designed to operate on edge devices such as the NVIDIA Jetson Nano, allowing on-site defect detection with minimal latency and without reliance on external computing resources.

This paper details the architecture, implementation, and results of the system, highlighting its effectiveness and potential as a scalable solution for industrial visual quality inspection.

2. Problem Statement and Proposed System

In the manufacturing industry, conveyor belts operate continuously, often running day and night. Identifying defects on these moving belts is a critical task and is prone to human error due to the need for constant monitoring at high speeds. Since filter designs move continuously along the conveyor, it becomes challenging for human inspectors to reliably detect all types of surface defects. The primary defects of concern include chemical patches, corrosion, pinholes, and scratches.

To address this, the system must first accurately segment each individual filter design from surrounding components and adjacent filters on the same conveyor belt. For this purpose, the Segment Anything Model (SAM) is utilized to annotate segmentation masks on the filter designs. A custom-trained

YOLO12m model is then employed for defect detection. The model was trained on a dataset of 3,000 augmented images—2,500 for training and 500 for validation—covering four classes: *corrosion*, *pinhole*, *chemical_patch*, and *scratch*. The YOLO12m model was trained for 70 epochs to achieve effective classification performance. And Yolo12-seg model was run for 50 epochs with SAM annotated designs.

In the proposed system, YOLO12 is used for the automatic identification and classification of defects within the segmented designs. The pipeline is designed to be deployable on edge computing devices such as the Jetson Nano, enabling real-time inference without reliance on cloud infrastructure.

3. Literature Review

Defect detection on various types of materials and objects, such as wood, metal surfaces, and manufactured components, has been widely explored and applied across multiple industrial domains. Over the years, researchers have developed numerous computer vision and deep learning-based approaches to automatically identify surface anomalies like cracks, pin-holes, discoloration, corrosion, deformations, misalignments, dents, chemical patches, Rolled-in Scale, pitting, surface roughness, to improve quality control processes in sectors ranging from furniture production to electronics and automotive manufacturing.

Li et al.(2018) [1] proposed an improved YOLO-based convolutional neural network for real-time detection of surface defects on cold-rolled steel strips. Their model, with 27 convolutional layers and 50000 epochs, achieved 97.55% mAP and 95.86% recall on a custom dataset of six defect types. The network demonstrated a detection rate of 99% at 83 FPS, making it suitable for high-speed industrial inspection. This work emphasizes the importance of generalization across different production lines and effective data augmentation to reduce overfitting.

Ma et al (2017) [2] proposed an adaptive segmentation algorithm to adaptively segment regions of defects, based on the gray features of the mental surface.

Ke et al. (2016) [3] suggested a tetrolet-based technique to identify steel strip surface flaws. Next to the extraction of the surface's sub-band characteristics, Support Vector Machine (SVM) classifier was utilized to categorize various kinds of surface defects in various scales and directions.

By fusing eigenvector-based texture data with percentile-based color histogram details, Song et al. (2015) [4] presented a technique for identifying surface flaws in wood. Their method, which categorizes picture blocks to find flaws, was shown to be very successful in spotting intricate defect patterns, such those that appear at intersections.

Li et al. (2020) [5] proposed a defect detection algorithm for ceramic tile surfaces that leverages multi-feature fusion. To address the limitations of using a single feature descriptor, they combined color moment features with FSIFT (Fast Scale-Invariant Feature Transform) features based on their respective influence magnitudes. Especially for intricate surface textures and subtle defect patterns, this fusion strategy improved the durability and accuracy of defect detection.

Canny edge detection [6] is popular technique for edge localization in defect detection tasks because of its resilience and accuracy. It incorporates several stages—image smoothing using a Gaussian filter, gradient computation, non-maximum suppression, double thresholding, and hysteresis tracking—to enhance accuracy and reduce noise sensitivity. While effective, it relies heavily on threshold values, making parameter tuning a challenge in practical vision systems. Additionally, clustering-based methods have also been explored in threshold-based segmentation.

Fu, G. [7] proposed CNN model based on SqueezeNet for fast and accurate steel surface defect classification, emphasizing low-level feature extraction and multiple receptive fields. The model demonstrated high accuracy on a challenging dataset with severe illumination variations, noise, and motion blur, using minimal training data.

Yu et al. [8] introduced a two-stage deep learning framework for industrial surface defect inspection that effectively balances speed and accuracy. The first stage employs a lightweight fully convolutional network (FCN) for fast, coarse segmentation of defect regions, while the second stage uses another FCN to refine these predictions. To address the common issue of limited training data in industrial applications, the authors also implemented a patch-based training strategy, demonstrating its effectiveness on the DAGM 2007 dataset.

Reis et al. (2023) [9] proposed a two-stage YOLOv8-based framework for real-time flying object detection, addressing challenges such as occlusion, small object sizes, and cluttered backgrounds. Their generalized model, trained on 40 flying object classes, leverages transfer learning to develop a refined model suited for real-world conditions. The system achieves high detection accuracy (mAP50-95 of 0.835) while maintaining a real-time inference speed of 50 fps, demonstrating the effectiveness of YOLOv8 in aerial object detection.

Simonyan et al. [10] proposed the VGGNet model, which replaced the larger convolutions in previous models with consecutive 3×3 convolutions. This significantly reduced the number of parameters while maintaining performance, enabling deeper networks. The model excelled well on localization and classification tasks in the 2014 ImageNet competition.

Liu et al. [11] introduced deep deconvolutional semantic image segmentation, which used VGGNet as the neural network for encoding and deconvolution and pixel prediction to construct image segmentation.

According to Yu et al. [12], an image classification model forms the basis for image semantic segmentation. But there are structural distinctions between categorization and prediction. As a result, they created the Dilated Convolution module, which lowers the resolution loss by combining data from several layers thus enhancing the segmentation's overall accuracy.

Kirillov et al. (2023) [13] introduced the Segment Anything Model (SAM), a promptable segmentation framework capable of zero-shot generalization across diverse image types and tasks. The SA-1B dataset, which has over 1 billion masks spread across 11 million images and is the largest segmentation dataset to date, was used to train the model. SAM demonstrates impressive zero-shot performance,

often rivaling fully supervised models, and significantly advances the development of foundation models in computer vision.

Redmon et al. [14] introduced YOLO, a fast and accurate object detection framework that uses regression methods to globally predict objects in images, ensuring high precision in detection.

Wang et al. [15] introduced YOLOv8, which optimizes the model architecture and training process to reduce parameters and computations. When the 4 of 20 frames per second (FPS) range from 5 to 160, it surpasses known real-time object detectors in terms of speed and accuracy.

Tian et al. (2025) [16] projected YOLOv12, an attention-centric real-time object detection framework. By using effective transformer modules, it performs better than CNN-based YOLO versions. The YOLOv12-Seg variant excels in segmentation tasks, combining attention with fine-grained mask prediction, enabling robust detection and segmentation under real-time constraints. The YOLOv12m model balances accuracy and speed, achieving superior mAP scores with low latency, making it suitable for industrial defect detection on edge devices.

It can be determined from the above literature that improvements in deep learning and image processing have significantly improved the ability to detect and classify surface flaws in industrial settings. While contemporary techniques like lightweight CNNs, two-stage FCNs, and YOLO-based detectors offer high accuracy and real-time performance, more conventional techniques like Canny edge detection and clustering-based thresholding provide fundamental preprocessing. Better accuracy-speed trade-offs are being achieved, as evidenced by the transition from CNN-centric models to attention-based frameworks like YOLOv12. To increase generalization and robustness in real-world applications, more architectural optimizations and dataset diversity are required. Despite these advancements, problems like occlusion, small object size, and complex backgrounds still exist.

4. Advancements of YOLO series:

Since its inception, the YOLO (You Only Look Once) series has experienced substantial change. The 2016 release of YOLOv1[17], which treated object recognition as a single regression problem and allowed for real-time performance, completely changed the field. Nevertheless, it has trouble identifying tiny objects and instances of overlap. YOLOv2 [18] and YOLOv3 [19] addressed these issues by introducing batch normalization, anchor boxes, and multi-scale predictions, leading to better accuracy and robustness. YOLOv4 [20], developed by the open-source community, brought in innovations like CSPDarknet53, PANet, and Mish activation, improving both speed and detection quality. With YOLOv5 [21], released by Ultralytics, ease of use, modularity, and export flexibility (ONNX, TorchScript, CoreML) became core strengths, making it extremely popular in industry and academia.

Building on this, YOLOv6 [22] focused on industrial applications with optimized training strategies and model scaling. YOLOv7 [23] further improved efficiency by integrating E-ELAN and better feature aggregation. YOLOv8 [15], also from Ultralytics, introduced a completely new architecture without config files, enabling seamless training, deployment, and support for instance segmentation. The newer YOLOv9 to YOLOv12 [16] models continue to push boundaries by incorporating transformer modules, decoupled heads, and advanced label assignment strategies. These versions offer improved performance in dense object scenarios and edge deployment, making them suitable for tasks like defect detection, autonomous driving, and smart surveillance. Each iteration of YOLO has consistently improved inference speed, accuracy, and usability, reinforcing its position as a state-of-the-art real-time object detection framework.

5. Two Stage Deep- Learning based Computer Vision Pipeline

This research proposes a two-stage deep learning-based computer vision pipeline that leverages the capabilities of the YOLO12 architecture—a high-performance, single-shot, attention centric detection model known for its speed and accuracy.

Stage 1 – Segmentation of Filter Designs:

The first step involves isolating individual cylindrical filter components from continuous conveyor belt footage. A custom-trained YOLO12-Segmentation model is employed for this task. The training dataset was annotated using the Segment Anything Model (SAM), which enables precise mask generation for accurate segmentation. This ensures that the model can effectively differentiate filter units from the surrounding background, machinery, and lighting variations. The segmentation model achieved a 99% accuracy, ensuring robust and consistent extraction of relevant regions of interest (ROIs) across varied scenarios. The segmentation dataset was created with Roboflow's SAM -Segment Anything model, which creates accurate mask over the cylindrical designs. For this around 250 Filter designs dataset was created with varied backgrounds and transformations. The masked dataset was then fed to yolov12-seg model, for 70 epochs of training and creating custom-trained segmentation model to separate out filter designs from surrounding background objects.

Stage 2 – Surface Defect Detection and Classification:

Each segmented filter image is passed to a YOLO12m object detection model, which is custom-trained to recognize and classify four defect types: corrosion, pinhole, scratch, and chemical_patch. The model was trained using a labeled dataset of defect examples under various lighting and orientation conditions. Despite the complexities of subtle defect patterns, the model currently achieves an average accuracy of ~70%, providing a solid baseline for practical deployment. This performance can be further improved with dataset augmentation, higher-resolution imaging, and implementation of newer YOLO variants or hybrid models. The accuracy can be enhanced by increasing number of training epochs with high performance GPUs.

Real-Time Video Processing and Edge Deployment:

To facilitate integration into existing production lines, the system uses OpenCV to capture and analyze live video streams frame-by-frame. The entire pipeline is optimized for execution on edge computing platforms such as the NVIDIA Jetson Nano, ensuring low-latency, real-time inference without dependence on internet connectivity or cloud processing. This makes the system highly suitable for factory environments with limited IT infrastructure.

6. System Architecture and Result Analysis

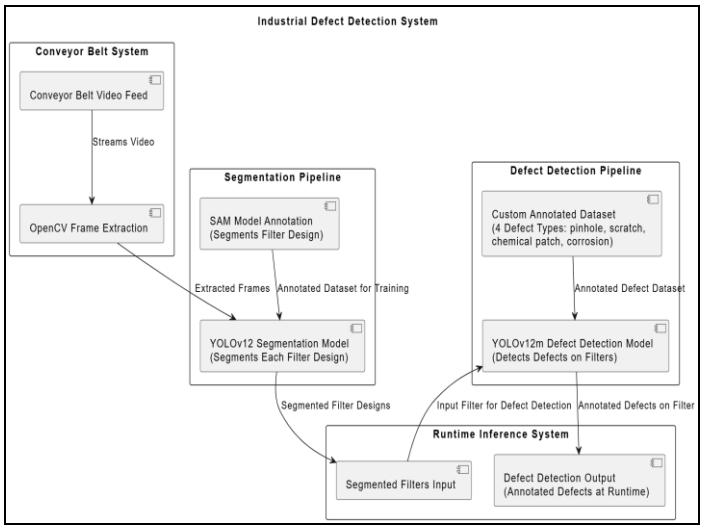

Figure 1. Architecture of Defect Detection System

The YOLOv12 state-of-the-art object detection model is leveraged in this system for both segmentation and defect detection tasks, offering high-speed inference and accuracy suitable for real-time industrial applications. Utilizing YOLOv12 for segmentation enables precise isolation of individual cylindrical filters on the conveyor belt, which is essential for localized defect analysis. The segmentation model is trained on high-quality masks annotated using the SAM (Segment Anything Model), ensuring robustness in diverse production environments. Following segmentation, the same YOLOv12 architecture—custom-trained for object detection—is employed to identify and classify four common surface-level defects: pinholes, scratches, chemical patches, and corrosion. This dual-role application of YOLOv12 streamlines the pipeline, reduces latency, and enhances detection reliability, making it well-suited for scalable deployment in automated visual inspection systems.

YOLOv12 Performance Evaluation:

The dataset used for training and validating the AI-powered Defect detection was collected from ABC Filters – around 2500 images were collected for training our ML model with 4 classes – corrosion, pinhole, scratch, chemical_patch. The model was trained for 70 epochs.

Training and Validation Losses: Losses measure how far the model's predictions are from the ground truth during training and validation. Lower values indicate better performance.

train/box_loss & val/box_loss: Measures error in the predicted bounding box coordinates (x, y, width, height) compared to the ground truth.

train/cls_loss & val/cls_loss: Classification loss; measures the error in predicting the correct class label for each object.

train/dfl_loss & val/dfl_loss: Distribution Focal Loss (DFL); helps improve the localization accuracy by refining the bounding box regression.

metrics/precision: Ratio of correctly predicted defect instances to total predicted defect instances.

$$\text{Precision} = TP / (TP + FP)$$

metrics/recall: Ratio of correctly predicted defect instances to all actual defect instances.

$$\text{Recall} = TP / (TP + FN)$$

metrics/mAP50(B): Mean Average Precision at IoU threshold 0.5. Measures both localization and classification performance.

metrics/mAP50-95(B): Averaged mAP across multiple IoU thresholds from 0.5 to 0.95. A more stringent and holistic evaluation of the model.

Figure-2 is the results metrics after training ultralytics yolo12m on images for 100 epochs with GPU parallel processing (Ultralytics YOLO12m Python-3.12 torch-2.3.1+cu118 CUDA NVIDIA GeForce GTX 1650).

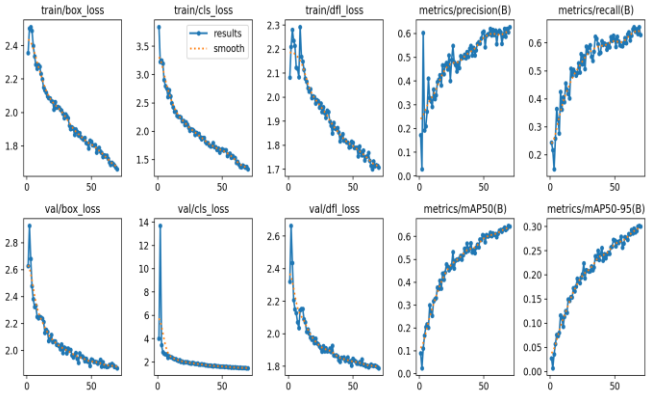

Figure 2: Results Metrics of yolov12m after 70 epochs

During training, the performance of the YOLOv12-based defect detection model was monitored using key training and evaluation metrics. The training and validation losses—including box loss, classification loss, and Distribution Focal Loss (DFL)—consistently decreased across epochs, indicating that the model was effectively learning object localization and classification tasks. Evaluation metrics such as precision and recall steadily increased, suggesting a reduction in false positives and false negatives, respectively. The mAP@0.5 metric showed significant improvement, reflecting strong detection performance, while the more rigorous mAP@0.5:0.95 also improved, demonstrating the model's

capability to generalize across varying IoU thresholds. These metrics validate the model's potential as a robust solution for industrial surface defect detection.

The normalized confusion matrix shows strong classification performance for individual defect types like chemical_patch, corrosion, scratch, pinhole. Confusion matrix shows 70 % accuracy for all classes, with some misclassifications occurring between similar classes and the background.

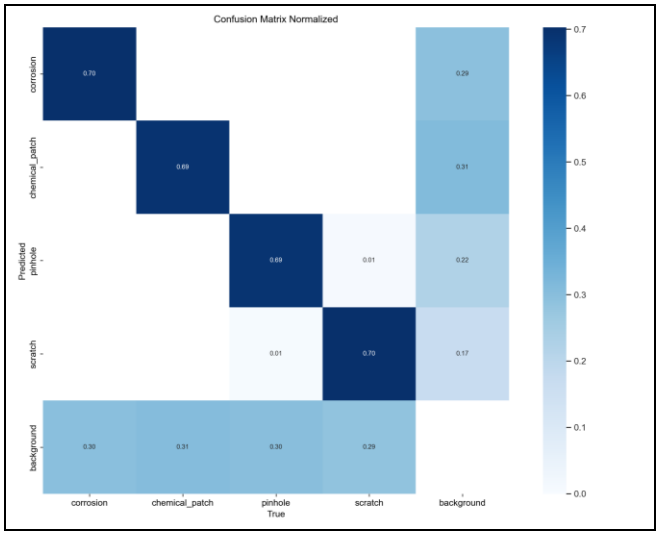

Figure 3- Confusion Matrix

The **Precision-Confidence Curve** shown in the figure-4 illustrates how the precision of the defect detection model changes with different confidence thresholds across the four defect classes: corrosion, chemical patch, pinhole, and scratch. This curve helps determine a trade-off point: higher confidence thresholds reduce false positives but may miss some true defects (reducing recall). It also identifies class-wise performance disparities which can be resolved with more data augmentation and re-training.

X-axis (Confidence) represents the confidence score (from 0 to 1) assigned by the model to its predictions. A higher score indicates higher model certainty.

Y-axis (Precision) is the ratio of true positive predictions to all positive predictions. It measures how many of the predicted defect detections are actually correct.

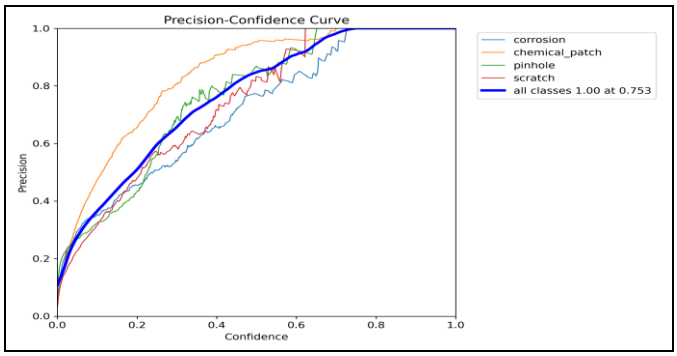

Figure 4- Precision-Confidence curve

The blue line represents the macro-average across all classes, showing an ideal threshold point at a confidence of 0.753, where the model achieves perfect precision (1.00). This point is useful for setting the optimal confidence threshold during deployment.

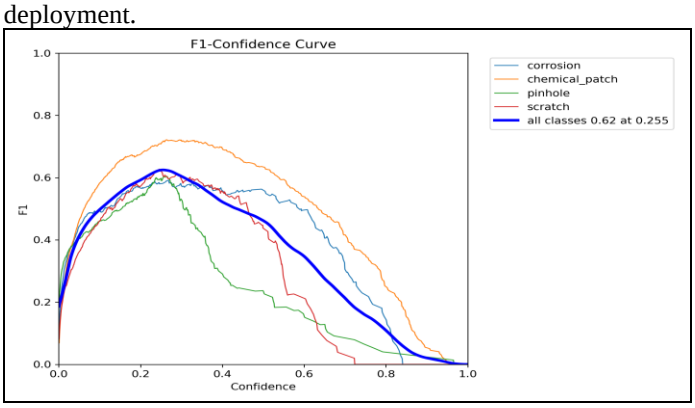

Figure 5- F1- Confidence Curve

The **F1-Confidence Curve** displays the trade-off between confidence threshold and the F1-score (harmonic mean of precision and recall) for each defect class. The optimal F1-score of 0.62 is achieved at a confidence threshold of 0.255 (marked in blue), indicating the best balance between precision and recall across all classes. The chemical_patch class consistently shows the highest F1 performance, while pinhole lags due to lower recall or higher false positives.

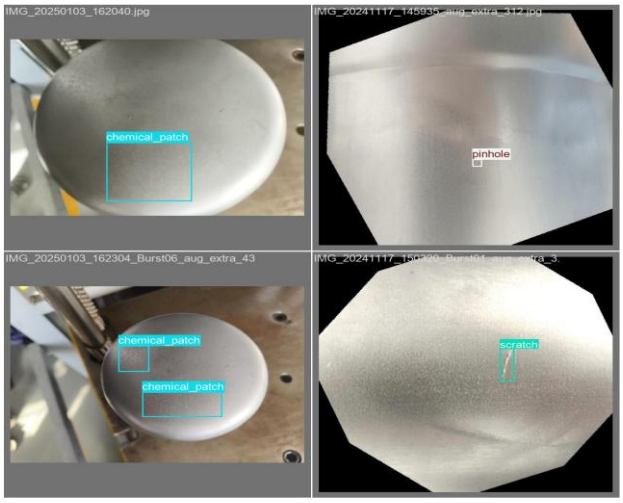

Figure 6: Validation-Batch Output

From the above results metrics, we observe that the model is performing well on identifying defects. The loss curve is inversely proportional with epochs and Precision is increasing reaching to 75% for all classes. F1 Confidence curve is the harmonic mean of precision and recall. Mean Average precision (mAP) measure of the accuracy of object localization and classification simultaneously.

Figure-7 to Figure-10 shows the segmentation model performance metrics, trained on 250 Filter images, with SAM masks annotations on Roboflow platform and trained with yolo12-seg with 50 epochs, The model has superior segmentation masks on moving conveyer-belt.

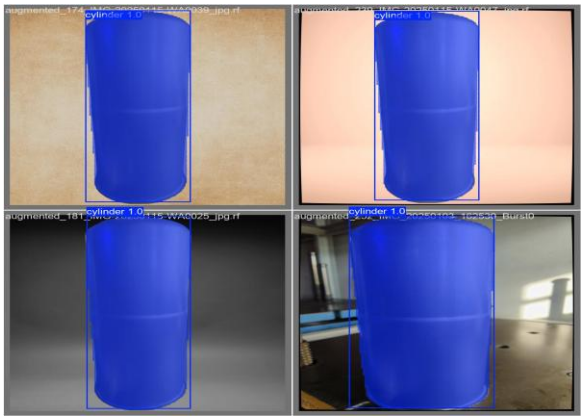

Figure 7 : Segmentation Validation Output – Confidence 1.0

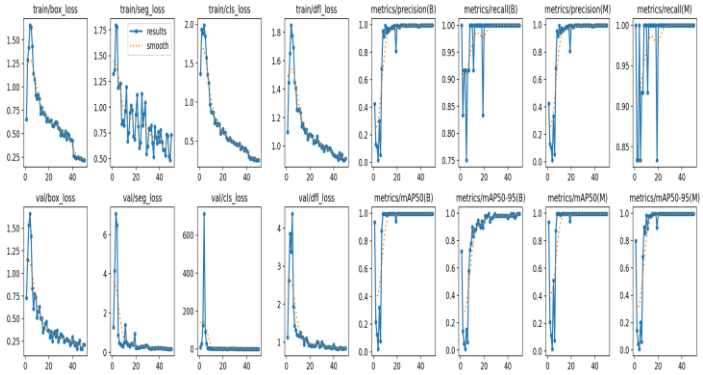

Figure 8: Results metrics after 50 epochs of yolov12-seg Segmentation precision-1.0

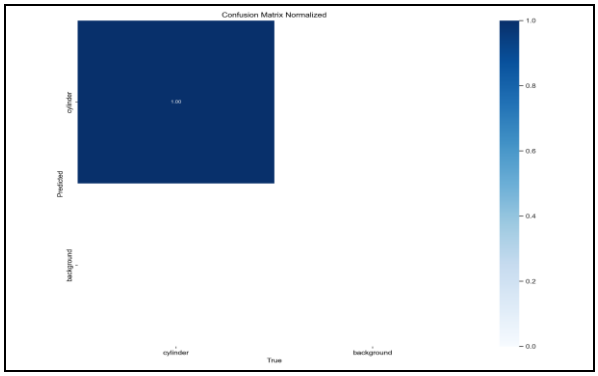

Figure 9- Confusion matrix – with 100% Confidence

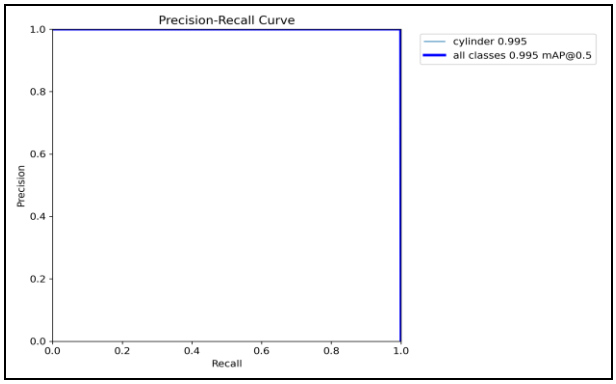

Figure 10- Precision-Recall Curve- 0.99 all classes

The segmentation training curves show a consistent decrease in both train/seg_loss and val/seg_loss, indicating that the model is effectively learning to segment defect regions over epochs. Metrics such as mAP quickly rise and stabilize near 1.0, showing excellent mask-based performance. This demonstrates that the segmentation branch of YOLO12 model is accurately identifying filter design separating it from surrounding and other filters on conveyer belt.

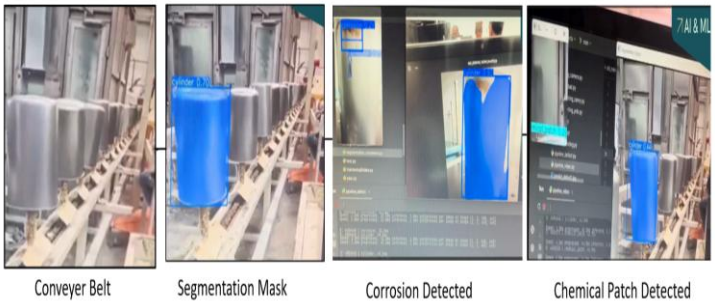

Figure 11. Output from Conveyer Belt live production video – Corrosion, Chemical_patch detected after segmentation

7. Conclusion and Further Enhancements :

The YOLO12m model, in combination with yolo12-Seg, has demonstrated strong performance in detecting and segmenting surface-level defects such as corrosion, chemical patches, pinholes, and scratches on cylindrical filters. The high precision, recall, and mAP scores indicate reliable detection and localization capabilities, while the segmentation results show the model's ability to accurately highlight defect regions. The confusion matrix and F1-confidence curves validate that the model generalizes well across classes, with chemical_patch showing particularly high accuracy. These results confirm the system's readiness for industrial deployment, especially in automated quality control pipelines.

For further enhancement, the model can be improved by incorporating multi-scale feature fusion to better detect smaller or subtle defects. Additionally, synthetic data augmentation and increasing number of epochs and training model on more varied and quantity of dataset, could boost performance across varying lighting and texture conditions. Additionally, optimizing the pipeline for deployment on edge platforms such as **NVIDIA Jetson Nano or Xavier** will enable true on-site, low-latency inference. Finally, integrating a feedback loop that collects false positives/negatives for continual retraining can drive continual model refinement and stability in dynamic manufacturing settings. Once a defect is detected by the model, the system can be seamlessly integrated with automation hardware such as a **robotic arm** for physical rejection of defective filters from the production line. Alternatively, a buzzer or warning light can be triggered in real time to alert human inspectors. This enables both fully automated and semi-automated setups, offering flexibility based on the production environment and resource availability.

References:

[1] J. Li, Z. Su, J. Geng, and Y. Yin, "Real-time detection of steel strip surface defects based on improved YOLO detection network," in *Proc. 5th IFAC Workshop on Mining, Mineral and Metal Processing*, Shanghai, China, Aug. 23–25, 2018.

[2] Y. Ma, Q. Li, F. He, L. Yan, and S. Xi, "Adaptive segmentation algorithm for metal surface defects," *Chin. J. Sci. Instrum.*, 2017.

[3] X. Ke, W. Lei, and J. Wang, "Surface defect recognition of hot-rolled steel plates based on tetrolet transform," *J. Mech. Eng.*, 2016.

[4] W. Song, T. Chen, Z. Gu, W. Gai, W. Huang, and B. Wang, "Wood materials defects detection using image block percentile color histogram and eigenvector texture feature," in *Proc. 1st Int. Conf. Inf. Sci., Mach., Mater. Energy*, Chongqing, China, Apr. 11–13, 2015.

[5] J. H. Li, X. X. Quan, and Y. L. Wang, "Research on defect detection algorithm of ceramic tile surface with multi-feature fusion," *Comput. Eng. Appl.*, vol. 56, pp. 191–198, 2020.

[6] M. H. T. Fouzia and K. Nirmala, "A literature survey on various methods used for metal defects detection using image segmentation," *Int. J. Sci. Res. (IJSR)*, vol. 5, no. 10, 2016, ISSN: 2319-7064.

[7] G. Fu, P. Sun, W. Zhu, J. Yang, Y. Cao, M. Y. Yang, and Y. Cao, "A deep-learning-based approach for fast and robust steel surface defects classification," *Opt. Lasers Eng.*, vol. 121, pp. 397–405, 2019.

[8] Z. Yu, X. Wu, and X. Gu, "Fully convolutional networks for surface defect inspection in industrial environment," in *Int. Conf. Comput. Vis. Syst.*, Berlin, Germany: Springer, 2017.

[9] D. Reis, J. Kupec, J. Hong, and A. Daoudi, "Real-time flying object detection with YOLOv8," *arXiv preprint arXiv:2305.09972*, 2023.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations (ICLR)*, San Diego, CA, USA, May 7–9, 2015.

[11] Z. Liu, X. Li, P. Luo, C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 7–13, 2015.

[12] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Representations (ICLR)*, San Juan, Puerto Rico, May 2–4, 2016.

[13] A. Kirillov et al., "Segment Anything," *arXiv preprint arXiv:2304.02643*, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2304.02643

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 27–30, 2016.

[15] J. Solawetz and Francesco, "What is YOLOv8? The ultimate guide," Apr. 30, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[16] Y. Tian, Q. Ye, and D. Doermann, "YOLOv12: Attention-centric real-time object detectors," *arXiv preprint arXiv:2502.12524*, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2502.12524

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 779–788. [Online]. Available: https://arxiv.org/abs/1506.02640

[18] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017. [Online]. Available: https://arxiv.org/abs/1612.08242

[19] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018. [Online]. Available: https://arxiv.org/abs/1804.02767

[20] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020. [Online]. Available: https://arxiv.org/abs/2004.10934

[21] Ultralytics, "YOLOv5 by Ultralytics," GitHub, 2020. [Online]. Available: https://github.com/ultralytics/yolov5

[22] Meituan Open Source, "YOLOv6: A single-stage object detection framework for industrial applications," GitHub, 2022. [Online]. Available: https://github.com/meituan/YOLOv6

[23] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *arXiv preprint arXiv:2207.02696*, 2022. [Online]. Available: https://arxiv.org/abs/2207.02696