# Deep learning-based defect detection in pulsed thermography

**Bhavya Verma*, S. M. Inchara, M. H. Sanfiya Banu, R. Deepa, V. Kavitha**

*Central Manufacturing Technology Institute, Bengaluru, Karnataka, India*

KEYWORDS

Pulse Thermography,
Pulse Phase Thermography,
Artificial Intelligence,
Deep Learning,
YOLOv8,
Object Detection,
Semantic Segmentation.

ABSTRACT

*The current era emphasizes system productivity to meet global demand. The productivity index is closely tied to the reliability of the manufacturing process. To keep up with today's manufacturing demands, inspection systems on the production line must prioritize both speed and quality. However, a key challenge in automated inspection is achieving a balance between high defect detection accuracy and minimizing false positives and false negatives. To address this, this study investigates the effectiveness of deep learning-based models for defect detection in Pulsed Thermography (PT) using a publicly available dataset of PVC specimens. Pulsed Phase Thermography (PPT) was applied to the raw thermograms to generate phase images and evaluate the performance of conventional methods. Two models were trained and evaluated for defect detection: a pre-trained YOLOv8 object detection model and a semantic segmentation model from Halcon. The YOLOv8 model demonstrated a high precision of 97.1%, but with a recall of 82%, indicating that it accurately detected defects but missed some. In contrast, the Halcon model achieved perfect recall (100%) but lower precision (78.2%), suggesting that it detected all defects but also introduced a significant number of false positives. The results highlight the trade-offs between precision and recall in these models, with YOLOv8 focusing on accuracy and Halcon on comprehensive defect detection. This study demonstrates the potential of deep learning techniques in enhancing defect detection performance in Pulsed Thermography applications.*

## 1. Introduction

Non-Destructive Testing (NDT) has evolved over the past century, beginning with basic methods like visual inspection and penetrant testing in the early 20th century, followed by the introduction of radiography, magnetic particle testing, and ultrasonic testing during World War II. By the 1950s, eddy current testing and acoustic emission methods gained traction in industries like aerospace and nuclear power. The advent of computerization in the 1970s revolutionized NDT through advanced imaging techniques like phased array ultrasonics and thermography (Vavilov & Burleigh, 2015). Every NDT method has its own merits and demerits, and therefore, their application in the industry varies. However, one of the major requirements in the manufacturing industry is to have an automated NDT system that can be used for safe in-situ testing of larger manufactured parts. In this regard, active thermography can have an advantage over other methods. Today, artificial intelligence, machine learning, and Industry 4.0 technologies are transforming NDT by enabling automated defect detection and predictive maintenance, ensuring the integrity of critical components in sectors such as aerospace, energy, and manufacturing. The world has started realizing the tremendous data processing and interpretation capabilities of machine learning and deep learning models which can remarkably enhance the defect detection capability. Therefore, in this study, two deep learning-based models were applied to thermograms made publicly available by Wei et al. (2023), which were obtained using Pulse Phase Thermography. A comparative analysis was conducted between the commercially used HALCON deep learning model and the open-source YOLOv8 model. Based on a review of the literature, this comparison appears to be novel, as similar studies have not yet been undertaken.

*Corresponding author E-mail:bhavyaverma@cmti.res.in

https://doi.org/10.58368/MTT.23.9-10.2024.1-9

## 2. Methodology

### *2.1. Pulse thermography*

Pulse Thermography (PT), also known as flash thermography, is a type of Active Thermography used to detect subsurface defects in materials. It involves applying a short heat pulse and analyzing the thermal response as shown in figure 1. When the heat front encounters anomalies in the material, the local temperature of the non-uniform region heats unevenly compared to the uniform region, creating a heat gradient. This causes the heated non-uniform region to diffuse heat, which can be observed on the surface of the sample (Almond & Lau, 1994). An infrared camera is used to detect sub-surface anomalies since most of the energy lies in the infrared region. Pulse thermography allows for faster inspection compared to other active thermography methods, but it requires more power and may result in noisier thermograms (Ciampa et al., 2018).

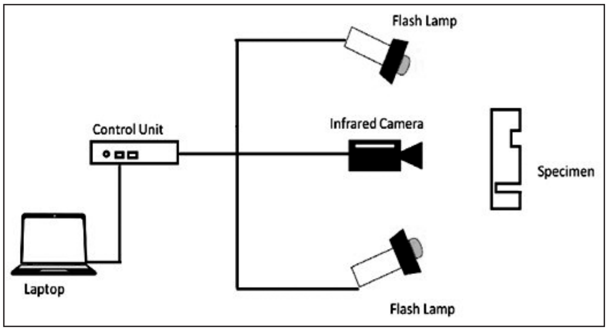


**Fig. 1.** A typical setup for pulse thermography.

### *2.2. Dataset preparation for processing*

The online available thermograms of PVC specimens with artificial defects were captured from a FLIR SC5000 MWIR camera at a frame rate of 10 Hz with a resolution of 320 x 256(Wei et al., 2023). The authors created 19 different specimens with each specimen of size 100 mm × 100 mm × 5 mm, with flat-bottom holes of varying sizes as shown in figure 2. The 1810 frames of each specimen were stored in an array of 320x256x1810 size in a .mat file. Each pixel value format is stored in a uint16 format. Each image sequence was extracted, normalized with min-max normalization and stored in .png format so that the annotation process could be easily done. The localization and annotation of the defects were done with the help of an online available labelling tool called makesense.ai (Skalski, 2019). The defect's locations in the image were saved in YOLO-compatible .txt and .csv by annotating the defects by creating bounding boxes around them. Only two classes of labels were used to annotate the data viz. Defect and Good. A total of 1819 images were initially manually annotated. Due to the equal proportion of defects and good images, a random split was performed with an 80:10:10 ratio for the training, test, and validation sets, respectively.

### *2.3. Pre-processing*

Predominantly, the raw thermograms need pre-processing to derive meaningful information from it, as the images can suffer from non-uniform

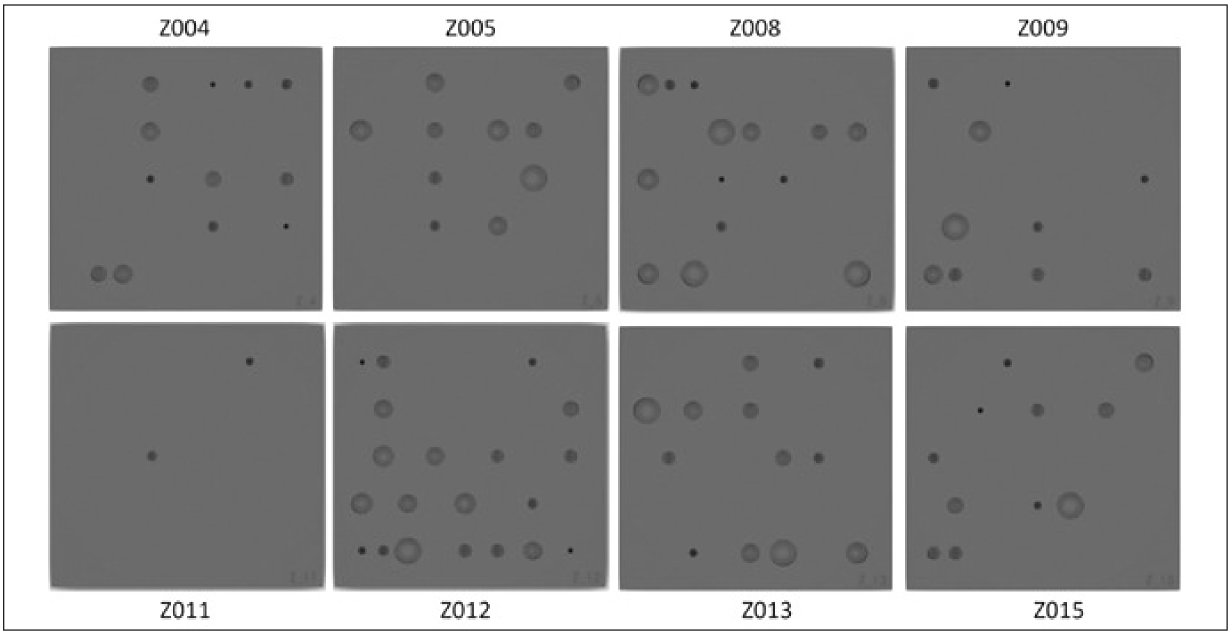


**Fig. 2.** Eight different specimen views from the bottom (Wei et al., 2023).

heating, and non-uniform emissivity of the sample. The observed features can also sometimes appear blurry, not because of the limitation posed by the infrared camera but due to the inherent nature of heat transfer in solids; as the local temperature of the defective region starts conducting the heat in all directions, this phenomenon can make the defects, difficult to characterize. In the past few decades, researchers have come up with several methods to reduce these effects and increase SNR. Researchers observed that the phase images are not only inherently robust against non-uniform heating but also can probe deeper defects, thus the application of the Fourier transform and other methods, such as the 4-bucket method, which can derive the phase information from the raw thermograms, have become a go-to tool for processing thermograms. The application of the Fourier transform in pulse thermography was introduced by Maldague & Marinetti and termed it Pulse Phase thermography (PPT) (Maldague & Marinetti, 1996). Apart from analyzing phase images, researchers have also used several matrix factorization-based methods such Principal Component Analysis (Rajic, 2002), Sparse-PCA (Wu et al., 2018), Independent Component analysis (Liu et al., 2019), etc. which can significantly reduce its dimensionality while capturing the most important information. In this paper, Pulsed-Phase Thermography has also been applied and the results are shown in the subsequent section. However, the phase images were not taken for training deep learning models because of the smaller number of phase images with visible defects.

## *2.4. Deep learning architectures*

In earlier days, pre-processing and defect detection algorithms used to be different as these conventional-based methods were based on fixed rules; however, since the inception of data-driven modelling, the line between these two tasks has become blurred. Among all Machine learning algorithms, the performance of deep learning (DL) models especially Convolutional Neural Networks (CNN) has been most promising in the field of computer vision as they tend to learn information about images in a hierarchical manner and resemble how humans learn and perceive information (Goodfellow, et al., 2016). There are also several key advantages that give the DL model an edge over other ML models that can be used for classification; one of the major merits of the DL model is that it doesn't require manual feature engineering like most ML models such as Support Vector Machine, Decision Tree, etc. These models are more robust against object appearance, scale and orientation (Goodfellow et al., 2016). For object detection tasks there are several single-stage and two-stage detectors that have been widely used by DL community. In this paper, two deep learning pre-trained models has been utilized, viz., the object detection model of YOLOv8 and the segmentation model of HALCON. Comparing an open-source model like YOLOv8 with a proprietary, fine-tuned model like HALCON provides insight into the potential of open-source tools in settings where commercial software is typically used. This contrast could be important for cost-benefit analysis, performance benchmarking, or other practical applications in industrial inspection or defect detection.

- YOLOv8

YOLO (You Only Look Once) is a widely used object detection and segmentation model, originally developed by Joseph Redmon and Ali Farhadi at the University of Washington in 2016 (Redmon & Farhadi, 2017). YOLOv8 (You Only Look Once version 8) is a pre-trained object detection model developed by Ultralytics (Ultralytics, 2023). It predicts the bounding box and class probabilities of the object in a single forward pass, although this feature can affect the accuracy but it can greatly boosts up the inference time. One can broadly divide the YOLOv8 architecture layers into backbone, neck and head. The backbone layer is responsible for extracting features; YOLOv8 uses a variant of the CSPNet (Cross-Stage Partial Network) to improve information flow between layers and boost accuracy. The neck is responsible for merging feature maps. Instead of a traditional feature pyramid, a C2F module is used, which helps combine higher-level semantic features with lower spatial information. The head is responsible for the final detection. YOLOv8 uses decoupled heads for classification and localization, meaning different branches handle object classification and bounding box regression, improving the model's efficiency.

In this paper, the YOLOv8s pre-trained model has been utilized, which balances computational requirements and performance. YOLOv8 was chosen over models like SSD (Single-Shot Detector) and Faster R-CNN due to its balance of speed, accuracy, and flexibility, which aligns well with the demands of real-time, high-performance applications. Unlike Faster R-CNN, which is more

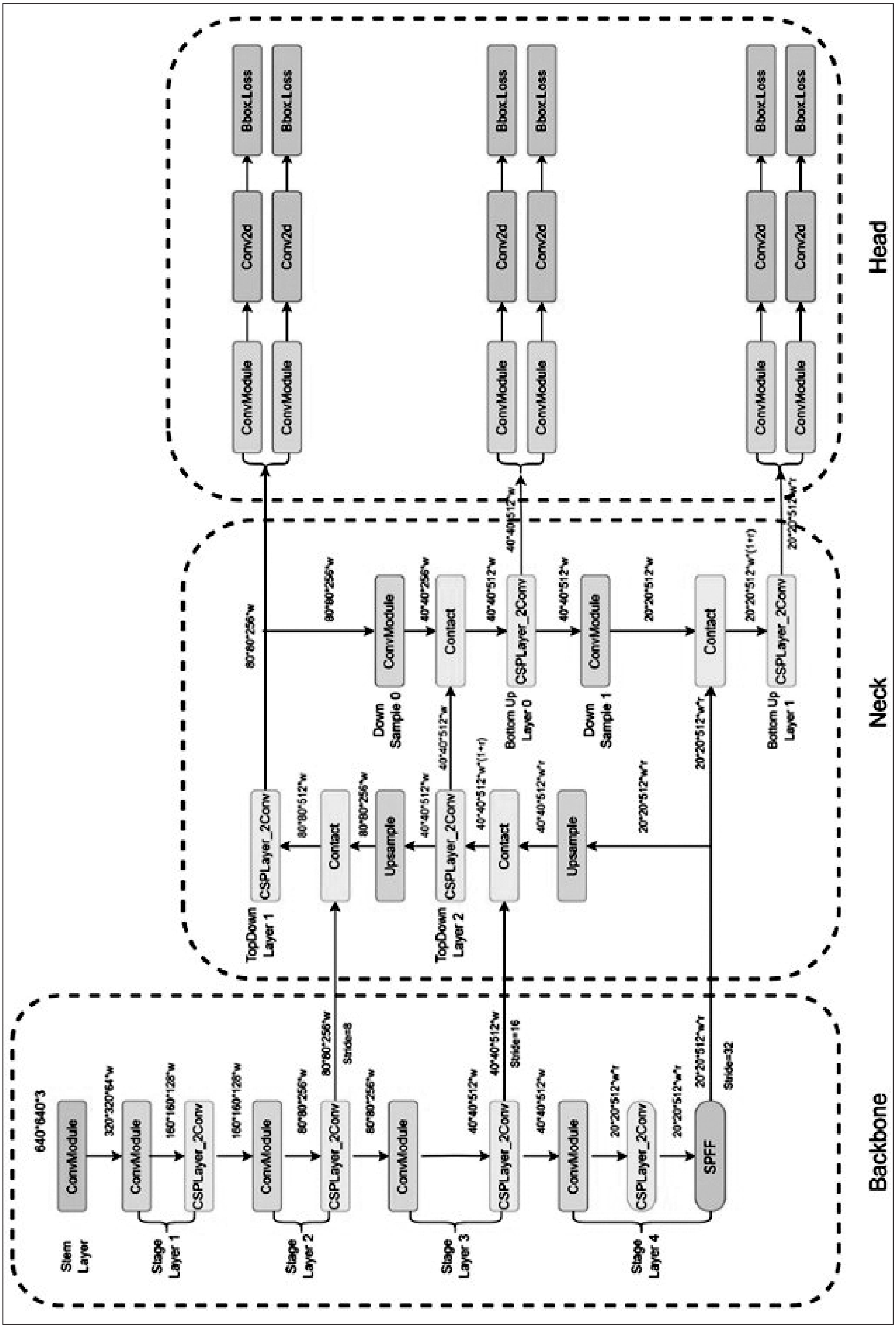


**Fig. 3.** YOLOv8 architecture (Gao et al., 2024).

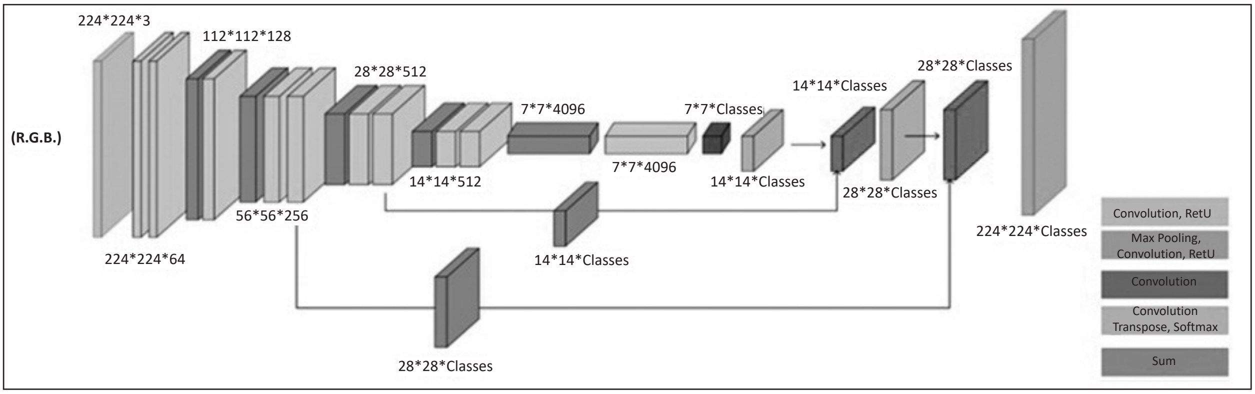


**Fig. 4.** Semantic segmentation model.

complex and slower due to its two-stage process, YOLOv8 offers rapid object detection without sacrificing much in accuracy, making it ideal for applications where quick response times are essential. Compared to SSD, YOLOv8 achieves improved localization and detection performance, especially for small or overlapping objects, thanks to its advanced backbone architecture (Huang et al., 2016).

- HALCON Model

The other model that was utilized for our objective was the semantic segmentation model from HALCON, which utilizes encoder-decoder architecture, as shown in figure 4. The encoder-decoder model is ideal for semantic segmentation due to its ability to handle the dual tasks of extracting meaningful features while preserving spatial information (Shelhamer, et al., 2017). The encoder reduces the image to key abstract features, and the decoder restores resolution, enabling pixel-level predictions that are crucial for segmentation tasks. Unlike standard object detection (bounding boxes), semantic segmentation requires classifying each pixel in the image, which means the model needs to understand both global context and fine-grained details (Follmann, et al., 2018)

## *2.5. Performance metrics*

Performance metrics are important tools to quantify the efficiency and performance of the object detection model, some of the widely used metrics for classification tasks are accuracy, precision and recall (Géron, A. 2019).

- Precision

It is the ratio of correctly predicted positive observations to the total predicted positives. It measures how many of the predicted positive instances are actually positive. Precision is important when false positives are costly.

$$\textit{Precision} = \frac{\textit{true positives}}{\textit{true positives + false positives}}$$

- Recall

Also known as Sensitivity or the true positive rate and sometimes also referred as Probability of Detection. It is the ratio of correctly predicted positive observations to all actual positives in the dataset. It measures how many of the actual positive instances the model correctly identifies.

$$\textit{Recall} = \frac{\textit{true positives}}{\textit{true positives + false negatives}}$$

- F-Score

The F-score, also known as the F1 score, is a metric used to evaluate the accuracy of a model by combining precision and recall into a single measure. Defined as the harmonic mean of precision and recall, the F-score is calculated as follows:

$$\textit{F-Score} = 2 * \frac{(\textit{Precision} * \textit{Recall})}{(\textit{Precision}+\textit{Recall})}$$

This measure provides a balanced view of a model's performance, particularly in scenarios where precision and recall may be inversely related. The F-score is particularly useful for evaluating classification models in situations with imbalanced data, as it ensures that both false positives and false negatives are considered in assessing model effectiveness.

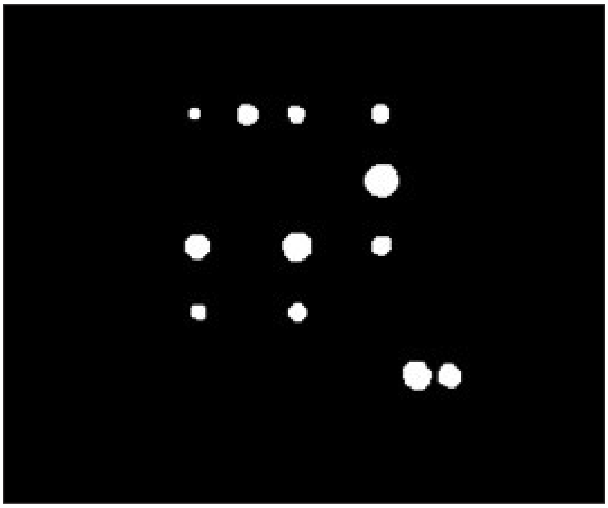

**Fig. 5.** Mask image of Z004 specimen (depicting locations of all the defects in the specimen).

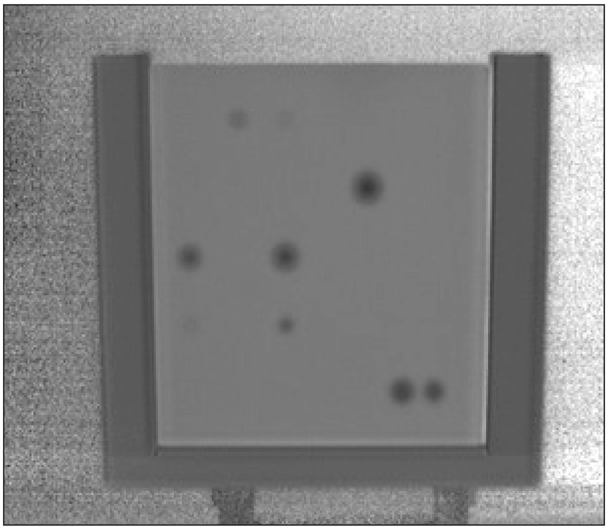

**Fig. 6.** Phase image of z004 specimen at 0.044 Hz frequency. Only 9 defects are visible out of 12.

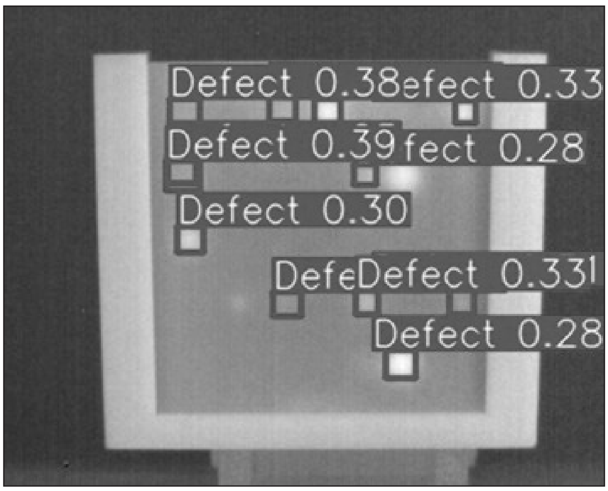


**Fig. 7.** Object detection result of YOLOv8.

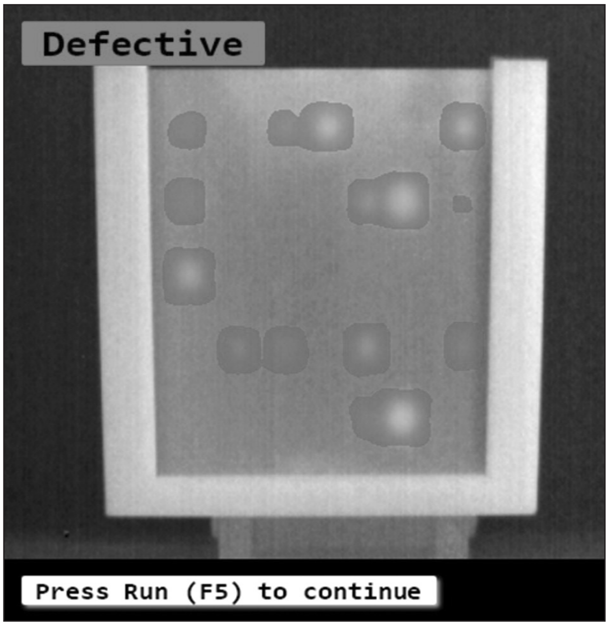


**Fig. 8.** Segmentation results of HALCON model.

## 3. Results and Discussions

### *3.1. Pulse phase thermography*

To extract phase information from the raw thermograms, a one-dimensional Discrete Fourier Transform was applied at every pixel using a Fast Fourier Transform. Among the total of 12 defects, only 9 defects were visible in the phase images at lower frequencies, which conveys that the majority of the defects were deeper and at nearby depths. The maximum contrast was observed at 0.044 Hz. Figures 5 and 6 show the mask and phase images, respectively.

### *3.2. Object Detection results of YOLOv8 and HALCON model*

A total of 50 images of different defect classes were used during the inference, which were not included in the training. The results in Table 1 show the averaged performance metrics on inferred images. The results were promising, as the YOLOv8 model was able to identify almost all the defects, as shown in Figure 7. However, it missed some obvious visible defects, which may indicate that the model is slightly overfitting. On the other hand, Halcon was unable to mark boundaries for the defects clearly and was merging the defective areas with the good ones, as shown in Figure 8.

From Table 1 one can observe that YOLOv8 is very precise, but its recall is lower, meaning it misses

**Table 1**
Training parameters and results of YOLOv8s and HALCON semantic segmentation model.

| Model | Training Parameters | Precision | Recall/ PoD | F-Score | Hardware | Training Time |
|---|---|---|---|---|---|---|
| YOLOv8s object detection model | Epoch = 100, Optimizer = AdamW, Learning Rate = 0.001667, Batch Size = 8, | 97.1 % | 82 % | 88% | Intel Xeon W-2245 CPU @ 3.90GHz, RAM: 32.0 GB Trained on NVIDIA T1000, 4096MiB | 1.203 hours |
| HALCON Semantic Segmentation Model | Epoch = 100, Learning Rate = 0.0001 Batch Size = 16 Momentum = 0.99 | 78.2 % | 100 % | 87% | Processor: Intel Xeon, CPU E5-2667 v3 @ 3.20GHz, RAM: 32.0 GB | 9 hrs |

some defects. The model is highly confident about the detections it makes but doesn't catch all defects, which could indicate that the model is conservative in making predictions to avoid false positives. This is further supported by the F-score of 88%, which reflects a balanced but selective performance that prioritizes accuracy in defect identification over exhaustive coverage. Halcon has perfect recall but lower precision, meaning it detects all defects but also makes more false positives, identifying some non-defective areas as defective. The model's F-score of 87% indicates strong performance, though it is slightly lower than YOLOv8's, highlighting HALCON's more aggressive approach in defect detection, aiming for comprehensive coverage at the cost of selectivity:

1. YOLOv8 is more selective, focusing on accuracy but missing some defects.
2. Halcon detects all defects but is less selective, leading to more false alarms

## 4. Conclusion

In this study, we evaluated the performance of deep learning-based models and conventional methods for defect detection in Pulsed Thermography. The results highlight the distinct advantages and limitations of each approach. The phase-based analysis revealed that only a portion of the total defects was visible at lower frequencies, with most defects being deeper and located at nearby depths. In comparison, the deep learning models, specifically YOLOv8 and Halcon model, demonstrated different strengths in terms of precision and recall. YOLOv8 excelled in precision but missed some defects, while Halcon ensured complete detection at the expense of higher false positives. The comparison of conventional and deep learning methods underscores the potential of using deep learning techniques for automating and enhancing defect detection. However, the trade-off between precision and recall remains, suggesting room for optimization depending on the use case.

## Acknowledgement

Authors thank the Dept of Science and Technology, Govt. of India, for funding this research under grant number DST/TDT/AM/2022/337 of the Advanced Manufacturing Technology (AMT) Program.

## References

Almond, D. P. & Lau, S. K. (1994). Defect sizing by transient thermography. I. An analytical treatment. *Journal of Physics D: Applied Physics, 27*(5). http://iopscience.iop.org/0022-3727/27/5/027

Ciampa, F., Mahmoodi, P., Pinto, F., & Meo, M. (2018). Recent advances in active infrared thermography for non-destructive testing of aerospace components. *Sensors, 18*(2), 609. MDPI AG. https://doi.org/10.3390/s18020609

Follmann, P., Böttger, T., Härtinger, P., König, R., & Ulrich, M. (2018). MVTec D2S: Densely segmented supermarket dataset. *Proceedings of the European Conference on Computer Vision (ECCV),* (pp. 569-585)

Gao, C., Zhang, Q., Tan, Z., Zhao, G., Gao, S., Kim, E., & Shen, T. (2024). Applying optimized YOLOv8 for heritage conservation: enhanced object detection in Jiangnan traditional private gardens. *Heritage Science, 12*(1). https://doi.org/10.1186/s40494-024-01144-1

Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems (2nd ed.). O'Reilly Media.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna Z., Song Y., Guadarrama S., Murphy K. (2017). Speed/accuracy trade-offs for modern convolutional object detectors. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 3296-3297. https://doi.org/10.1109/CVPR.2017.351

Liu, Y., Wu, J. Y., Liu, K., Wen, H. L., Yao, Y., Sfarra, S., & Zhao, C. (2019). Independent component thermography for non-destructive testing of defects in polymer composites. *Measurement Science and Technology, 30*(4). https://doi.org/10.1088/1361-6501/ab02db

Maldague, X., & Marinetti, S. (1996). Pulse phase infrared thermography. *Journal of Applied Physics, 79*(5), 2694–2698. https://doi.org/10.1063/1.362662

Rajic, N. (2002). Principal component thermography for flaw contrast enhancement and flaw depth characterisation in composite structures. www.elsevier.com/locate/compstruct

Redmon, J., & Farhadi, A. (n.d.). YOLO9000: Better, Faster, Stronger. http://pjreddie.com/yolo9000/

Shelhamer, E., Long, J., & Darrell, T. (2017). Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(4), 640-651.

Skalski, P. (2019). Make sense. GitHub. https://github.com/SkalskiP/make-sense/

Ultralytics. (2023). YOLOv8. GitHub. https://github.com/ultralytics/yolov8

Vavilov, V. P., & Burleigh, D. D. (2015). Review of pulsed thermal NDT: Physical principles, theory and data processing. *NDT and E International, 73,* 28-52. https://doi.org/10.1016/j.ndteint.2015.03.003

Wei, Z., Osman, A., Valeske, B., & Maldague, X. (2023). A dataset of pulsed thermography for automated defect depth estimation. *Applied Sciences (Switzerland), 13*(24). https://doi.org/10.3390/app132413093

Wu, J. Y., Sfarra, S., & Yao, Y. (2018). Sparse Principal component thermography for structural health monitoring of composite structures. *IFAC-PapersOnLine, 51*(24), 855-860. https://doi.org/10.1016/j.ifacol.2018.09.675


**Bhavya Verma** has a B.E. in Electrical Engineering and a Master's in Solid State Technology. He is currently working as Scientist-B at Central Manufacturing Technology Institute, Bangalore, in the Vision Technology group. His research interests include Deep Learning, Machine Vision and Electronics Design.

**S. M. Inchara** is a Computer Science Engineering Graduate. She currently works as a Project Associate-1 at the Central Manufacturing Technology Institute at Sensor Vision Technology, focusing on vision technology and machine learning applications. Her research interest includes AI, ML, and IoT.
(E-mail: incharasmkadur@gmail.com)

**M. H. Sanfiya Banu** is an Electronics and Communication Engineering Graduate. She currently works as a Project Associate 1 at the Central Manufacturing Technology Institute at Sensor Vision Technology, focusing on vision technology and machine learning applications. Her research interest includes embedded technology and deep learning applications.
(E-mail: mhanassanu9876@gmail.com)

**R. Deepa** is a Senior Scientist at the Center for Sensors and Vision Technology (CSVT) at the Central Manufacturing Technology Institute (CMTI), where she currently leads the Vision Technology group. With a strong expertise in machine vision, her research is at the forefront of developing innovative inspection systems that span both 2D and advanced 3D applications. She has made notable contributions to the field, publishing numerous papers in renowned journals and conferences.

Throughout her career, she has successfully led multiple projects focused on creating cutting-edge inspection solutions, leveraging both traditional and AI-driven techniques. Her research interests are diverse, encompassing 3D Reconstruction, Non-destructive Testing, Automation, and Non-invasive Technologies. With a passion for advancing the capabilities of machine vision, she continues to drive innovation in these critical areas, making notable advancements in industrial applications.
(E-mail: deepa@cmti.res.in)

**V. Kavitha** is a Senior Technical Assistant at Sensors and Vision Technology Department of Central Manufacturing Technology Institute, Bengaluru. She obtained her Diploma in Electronics and Communication from Women's Polytechnic, Shimoga in 2009. She has over 13 years of experience in the field of machine vision and automation. Her research interests include machine vision, high speed inspection and automation.
(E-mail: kavitha@cmti.res.in)