

Spatial and temporal deep learning algorithms for defect segmentation in infrared thermographic imaging of carbon fibre-reinforced polymers

Iago Garcia Vargas^a and Henrique Fernandes^{a,b} 

^aFaculty of Computing, Federal University of Uberlandia, Uberlandia, Brazil; ^bIVHM Centre, Faculty of Engineering and Applied Sciences, Cranfield University, Cranfield, UK

ABSTRACT

For non-destructive evaluation, the segmentation of infrared thermographic images of carbon fibre composites is a critical task in material characterisation and quality assessment. This paper presents a study on the application of image processing techniques, particularly adaptive thresholding, and advanced neural network models, including U-Net, DeepLabv3, and BiLSTM, for the segmentation of infrared images. This work introduces the innovative combination of DeepLabv3 and BiLSTM applied in infrared images of carbon fibre-reinforced polymer samples for the first time, proposing it as a novel approach for enhancing the accuracy of segmentation tasks. An experimental comparison of these models was conducted to assess their effectiveness in identifying artificial defects in these images. The performance of each model was evaluated using the F1-Score and Intersection over Union (IoU) metrics. The results demonstrate that the proposed combination of DeepLabv3 and BiLSTM outperforms other methods, achieving an F1-Score of 0.96 and an IoU of 0.83, showcasing its potential for advanced material analysis and quality control.

ARTICLE HISTORY

Received 12 September 2024
Accepted 17 January 2025

KEYWORDS

Infrared thermograph;
segmentation; U-Net;
DeepLabv3; BiLSTM;
composite material

1. Introduction

Composite materials are essential structural components widely used in industries, such as aerospace and wind turbine blade construction [1–5]. However, defects that arise during the manufacturing and handling of laminated composites can pose significant safety risks and may result in substantial losses [6].

The emergence of composites as a distinct material class occurred in the mid-20th century, characterised by the development of intentionally engineered multiphase composites, including Carbon-Fibre-Reinforced Polymers (CFRP). In recent years, advanced image processing techniques and machine learning approaches like deep neural networks have been increasingly utilised to segment thermographic images of CFRP, enabling more precise identification of structural defects [7–10].

CONTACT Henrique Fernandes  h.fernandes@cranfield.ac.uk

© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

Given the critical role of CFRP in industries such as aerospace and automotive, ensuring the integrity of this material is extremely important [11]. Defects, including cracks, delaminations, and voids, can severely compromise the performance and durability of CFRP components.

Therefore, the implementation of robust Non-Destructive Testing (NDT) methods is essential for detecting and quantifying these flaws without inflicting damage on the material [12]. Infrared thermography (IRT) emerges as an NDT and efficient technique for assessing the condition of CFRP components. However, raw thermographic data is often plagued by noise, necessitating the use of advanced processing techniques to accurately identify defects [13]. In this context, image processing methods and cutting-edge deep learning techniques are frequently employed to distinguish relevant thermal patterns associated with defects from background noise.

This study presents a comprehensive approach that combines IRT with advanced deep learning techniques to improve the accuracy of defect segmentation in IRT images. Four segmentation methods were employed: adaptive thresholding, the U-Net model, the DeepLabv3 model, and our proposed hybrid approach that, for the first time, integrates DeepLabv3 with a BiLSTM network for the task of segmentation of IRT images of CFRP samples. This novel combination represents a prospect step in enhancing the accuracy and effectiveness of segmentation tasks. We use metrics such as F1-Score and Intersection over Union (IoU) to evaluate the performance of our methodologies and validate the effectiveness of the proposed models in defect segmentation. The paper details the development of deep learning models, with a focus on the specific adaptations and optimisations made for processing thermographic data, and provides a comparative analysis of these algorithms in defect segmentation and identification. Our approach utilises a curated dataset of thermographic images, annotated with expert knowledge, to train and test the deep learning models. This dataset includes defects with different dimensions at different depths.

The paper is organised as follows: [Section 2](#) reviews related works, discussing previous research in the field of IRT and deep learning for defect detection in composite materials. It reveals the key advancements and gaps that this study aims to address. [Section 3](#) details the methodology, providing an in-depth explanation of the image processing techniques and the reasons for their selection, along with a discussion of the deep learning techniques employed. [Section 4](#) outlines the dataset preparation process and the training process for the deep learning models. [Section 5](#) presents the results of the image segmentation and defect detection, offering a comparative analysis of the effectiveness of the various methods used. Finally, [Section 6](#) concludes the study with a discussion of the findings and offers suggestions for future research directions. Finally, [Section 7](#) acknowledges to the partners who contributed to the success of this research.

2. Related works

The application of deep-learning techniques to IRT images has shown significant promise in enhancing image analysis and interpretation, due to the ability of these techniques to automatically learn and extract features from large datasets. Many papers of deep learning applications using IRT images have been published in recent years. These

include many researches on different fields of studies, with some being detailed in this session.

The work carried out by [7] aims to perform a comprehensive analysis of a neural network when used as a post processor to quantify defect depths from CFRP thermograms. The authors used a Pulsed Thermography (PT) methodology to inspect the materials and generate the thermograms. A pipeline based on filters was applied in the thermal data to remove noises such as reflections and residual heat. An in depth study is carried out to compare and analyse the effect of different parameters on the accuracy and robustness of the neural network result using two sets of data, a original dataset create by PT, and a cleaned dataset which received a treatment pipeline to remove thermal noise such as the reflection and the contribution from the surroundings and the residual heat from the pulse source. The best result achieved in this study was an accuracy of 92% using the cleaned dataset.

An approach to determine the depth of defect in composite materials, more precisely, in CFRP of aviation structure is presents in [14]. The authors used a laser infrared thermography (LIT) system to inspect the material and built a dataset to train and validate the developed model. An architecture of long-short term memory recurrent neural network (LSTM-RNN) was used to train the proposed model. The purpose of the paper is to compare the results using the proposed methodology with other two traditional architectures, RNN and CNN, concluding that they achieved better results using the proposed methodology. The model achieved a recall of 95% while the traditional algorithms, RNN (one hidden layer and containing 128 nodes) and CNN (64 convolutional kernel), achieved a recall of 84% and 80%, respectively.

The work developed in 2021 by [15] presents a study of defect depth estimation based on Gated recurrent units (GRUs). The work proposes to solve the problem of insufficient precision in the estimation of the depth of defects. Active IRT was applied to obtain raw thermal sequences of the surface temperature field distribution of the inspected sample. Before training the GRU model, Principal Components Analysis (PCA) was used to reduce the dimension of the raw sequence and eliminate the correlation of the raw data sets. Then, the GRU model was used to automatically estimate the defect's depth. The results indicate that PCA-processed datasets outperform raw temperature datasets in model learning when evaluating defect depth characteristics. According to the authors, the recall achieved by the GRU network was over 95% for defect depth of 1 mm.

Defect detection of subsurfaces is one of the hardest tasks in the industry. The paper presented by [16] addresses the challenging problem of subsurface defect detections using a combination of active thermography and deep learning. The main contribution of this work is to compare object detection rather than semantic segmentation or classification. The experiments were performed in a dataset obtained by applying active thermography to a single carbon fibre solid laminate sheet; this type of material consists of several carbon fibre layers, each layer with fibres oriented in a specific direction, stacked and bonded together with a resin. The defects in the laminate have different sizes: 12×12 mm, 7×7 mm, and 5×5 mm, and different depths: 0.63 mm, 1.46 mm, and 2.08 mm. They used a MobileNetv2 architecture to perform the classification model, a U-Net to perform segmentation and YOLOv5 to perform the detection model. According to

Table 1. Summary of related works.

Work	Spatial Feature	Temporal Feature	Used Metric	Result
[7]	Yes	No	Accuracy	92.0%
[14]	No	Yes	Recall	95.0%
[15]	Yes	No	Recall	95.0%
[16]	Yes	No	F1-score	96.0%
[17]	Yes	No	POD	95.0%
This Work	Yes	Yes	F1-score	96.29%

the authors, the detection model achieved the best result on the test dataset compared with the others. The YOLOv5 architecture achieved a precision of 97%, a recall of 95% and an F1-Score of 96%.

IRT is widely used for defect detection in composite materials, but identifying defect characteristics is often hindered by noise and uneven backgrounds in thermal images. To address these challenges, a novel thermography-based method employing a semantic segmentation network is proposed by [17] to enhance defect contrast and extract features. AG-UNet, with a spatial self-attention gate module, was used to capture spatial features, while 3D-UNet, incorporating 3D and temporal convolutions, extracted both spatial and temporal thermal features to improve defect segmentation. Compared to traditional techniques like PCA, TSR, and FFT, the proposed approach significantly enhanced detection performance, achieving a probability of detection (POD) of 0.95, enabling the identification of defects with smaller diameter-to-depth ratios.

2.1. Work contribution

Considering the criteria discussed so far, it is evident that the related works presented numerous strengths along with certain limitations, as outlined in Table 1. Considering the limitations primarily, the main original contributions of this work can be summarised as follows:

- A novel architecture is proposed that integrates both temporal and spatial features, enhancing the model's ability to capture complex patterns. Unlike most studies that use standalone models with spatial or temporal data, this approach leverages both data types to improve performance.
- The results obtained are equivalent or superior (in most cases analysed) to the state-of-the-art methods for the dataset evaluated.
- The F1-score is used as the primary evaluation metric, ensuring high precision and high recall in performance assessment. This metric is important because it balances the trade-off between precision and recall, providing a more comprehensive evaluation of model performance, especially in cases where data imbalance is critical.
- The combination of temporal and spatial features demonstrates the potential to achieve the best results, as evidenced by the superior performance metrics.

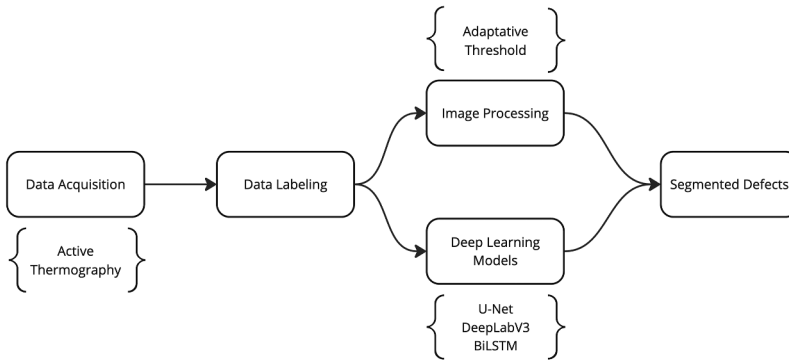


Figure 1. Defect segmentation methodology.

3. Methodology

The methodology applied in this study involves four main steps, as depicted in Figure 1. These steps include thermographic data acquisition, manual annotation of defects by an expert, defect segmentation using traditional image processing techniques, and defect segmentation using deep learning models. Further details regarding each step are provided in the subsequent sections to give a comprehensive understanding of the approach.

3.1. Active thermography

Active IRT is an NDT technique that leverages the principles of thermodynamics and IRT imaging to, among other things, evaluate the thermal properties of materials. In this process, a heat source is applied to the surface of the material being inspected, resulting in a localised temperature increase. The material's thermal response is subsequently recorded using an IRT camera, which detects the emitted infrared radiation [1]. This

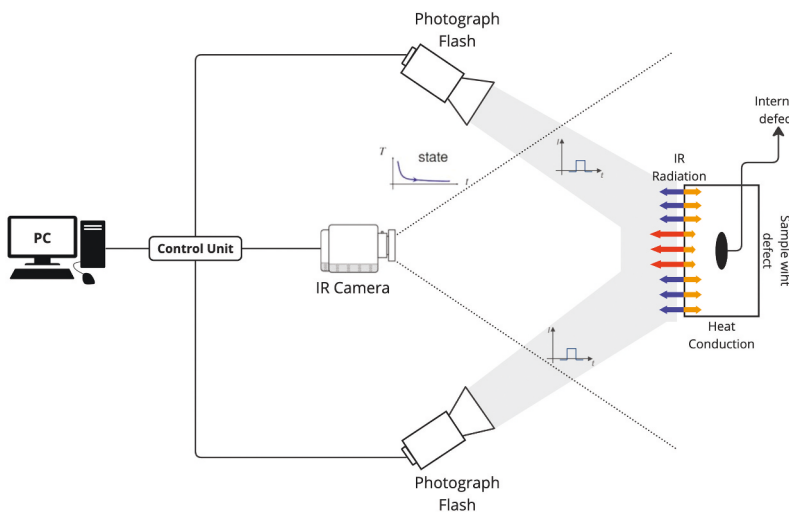


Figure 2. Active thermography.

entire process is shown in Figure 2. By analysing the thermal patterns captured by the camera, it is possible to infer valuable information about the material's internal structure, including the presence of defects and anomalies.

PT is a variant of active IRT in which the heat source is applied in short pulses rather than continuously. This pulsed approach allows for greater control over the heating process, thereby enhancing detection sensitivity, particularly for identifying small or shallow defects. PT offers significant advantages in terms of speed, efficiency, and versatility, making it an invaluable tool in fields such as aerospace, automotive, and materials science for quality control and defect detection [18].

3.2. Adaptive threshold

Adaptive thresholding is an image processing technique that dynamically adjusts to local intensity variations within an image to facilitate its segmentation [19]. This approach is widely used for the thermographic analysis of IRT images, where the complex thermal properties and structural heterogeneity often result in localised and variable thermal contrasts [20].

In the context of active IRT inspections, the propagation of heat through a material can be disrupted, if they exist in the material, by internal discontinuities, such as delaminations, porosity, or inclusions. These disruptions alter the local thermal conductivity, leading to variations in the observed heat distribution [21]. This heat distribution manifests on the material's surface and is captured by an IRT camera, with temperature differences and contrasts appearing as variations in pixel intensity on the IRT images. The adaptive thresholding method enables the isolation of these contrasting anomalies by adjusting the segmentation thresholds based on local mean or intensity variations, thereby enhancing the accuracy of defect detection.

Formally, adaptive thresholding is defined as follows in Equation (1). For a pixel $p(x, y)$ in a IRT image I , the adaptive threshold $T(x, y)$ can be calculated as:

$$T(x, y) = m(x, y) \times \left(1 + k \times \left(\frac{s(x, y)}{R} - 1 \right) \right) \quad (1)$$

where $m(x, y)$ is the local mean intensity, which is calculated by averaging the pixel values in a predefined neighbourhood around the pixel $p(x, y)$. This neighbourhood consists of a small window of pixels centred at (x, y) . The local mean intensity is used to adapt the thresholding process to local variations in intensity across the image. $s(x, y)$ is the local standard deviation, k is an empirical adjustment parameter, and R is the dynamic range of pixel intensities.

To apply adaptive thresholding for segmentation, we first calculate the threshold $T(x, y)$ for each pixel $p(x, y)$ using the formula provided in Equation (1). Once the threshold is determined, the segmentation process involves comparing the intensity of each pixel in the image $I(x, y)$ to its corresponding threshold $T(x, y)$. If the pixel intensity $I(x, y)$ exceeds the threshold $T(x, y)$, the pixel is classified as part of a defect; otherwise, it is classified as part of the background. This pixel-wise classification creates a binary

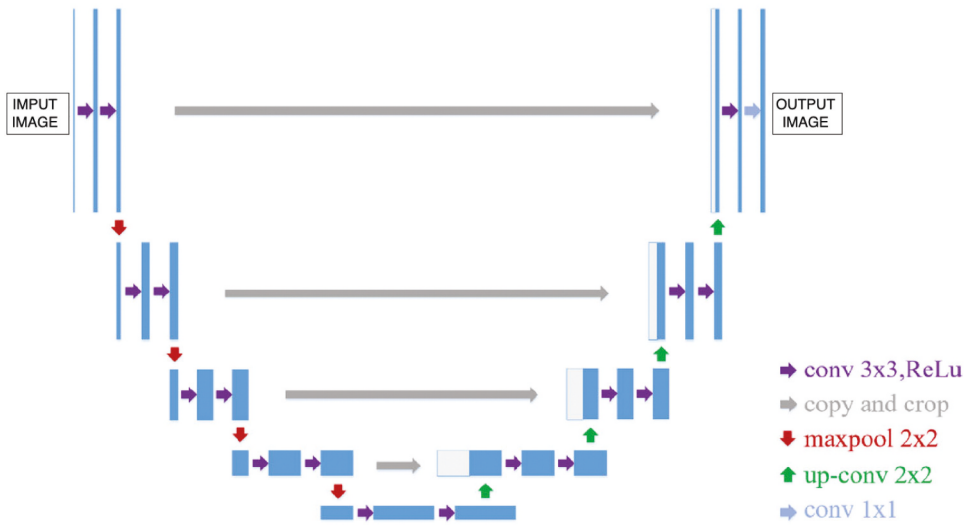


Figure 3. The architecture of the U-Net.

segmentation mask where pixels belonging to defects are labelled as 1 and the background as 0. This mask is then used to isolate and analyse the defect regions in the IRT images.

3.3. U-Net

The U-Net neural network is a widely used architecture for image segmentation, particularly important in both medical and industrial applications [22]. Originally developed by [23], U-Net is recognised for its efficiency in scenarios with a limited number of training samples and its ability to capture both contextual and localisation details simultaneously.

The U-Net architecture is distinguished by its symmetric structure, as illustrated in Figure 3. It comprises two main parts: a contracting path (encoder) and an expansive path (decoder). The contracting path progressively reduces the image dimensions while increasing the depth of the features, effectively extracting high-level representations. The expansive path then employs up-sampling operations and convolutions to reconstruct the image from these encoded features, enhancing the resolution and precision of the segmentation by refining the localisation of the identified features.

3.4. DeepLabv3

DeepLabv3 is a deep neural network architecture specifically designed for semantic image segmentation tasks, known for its effectiveness in identifying and segmenting complex features across a wide range of applications. One of the key innovations in DeepLabv3 is the use of ‘atrous convolution modules’ (also known as dilated convolutions), which allow the network to capture contextual information without reducing spatial resolution [24]. This capability is advantageous in the segmentation of defects in

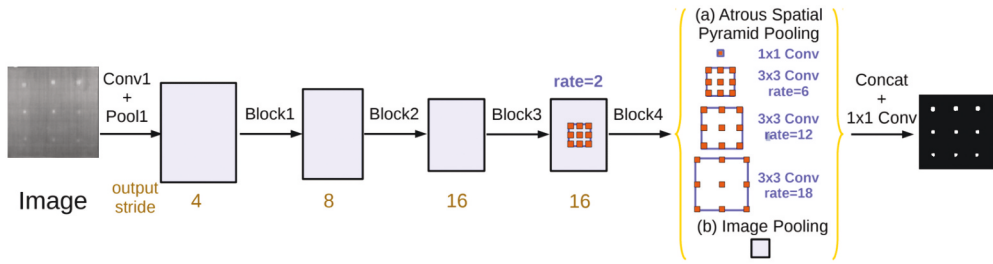


Figure 4. The architecture of DeepLabv3.

IRT images of CFRP, where variations in texture and temperature are common challenges [25].

Moreover, DeepLabv3 includes a spatial classification module called ‘Atrous Spatial Pyramid Pooling’ (ASPP), which employs multiple dilation rates to capture context at different scales. This feature enhances the model’s ability to detect and segment defects with varying characteristics and sizes, making it effective for analysing the complex thermal patterns found in IRT images [24].

Figure 4 illustrates the DeepLabv3 architecture. It starts with convolutional layers that progressively reduce the image resolution to extract features, followed by the use of atrous convolutions to expand the field of view without increasing computational complexity. The key component, ASPP, uses atrous convolutions with multiple rates and global pooling to obtain detailed features at different scale levels.

3.5. BiLSTM

While convolutional neural networks (CNNs) are highly effective at capturing spatial features, recurrent neural networks (RNNs) excel in modelling temporal dependencies in sequential data. Bidirectional Long Short-Term Memory (BiLSTM) networks, a variant of RNNs, are adept at capturing long-range dependencies by processing data in both forward and backward directions. BiLSTM networks can be employed to analyse temporal sequences of images captured over time [26]. In this paper, we propose to use BiLSTM for the first time on IRT data for defect detection in CFRP samples. This approach is particularly beneficial for monitoring the evolution of thermal patterns, offering additional information into the presence and progression of defects as they develop, thereby enhancing the accuracy and reliability of defect detection. The application of BiLSTM in this context represents a novel approach, offering the potential to uncover patterns that were previously difficult to detect with only spatial methods.

To understand BiLSTM networks, it is essential to first grasp the concept of Long Short-Term Memory (LSTM) networks. LSTMs are a variant of RNNs specifically designed to address the vanishing gradient problem, which hampers standard RNNs in learning long-term dependencies. This architecture was introduced in 1997 by [27].

As illustrated in Figure 5, LSTMs are distinguished by the inclusion of a memory cell and three key gates: the input gate, forget gate, and output gate. These gates regulate the flow of information within the network, allowing it to retain or discard information as

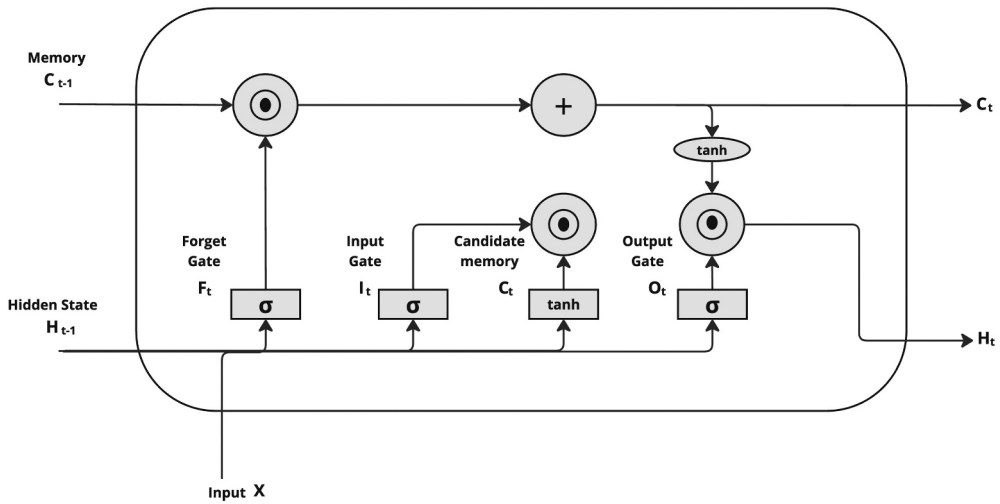


Figure 5. LSTM architecture.

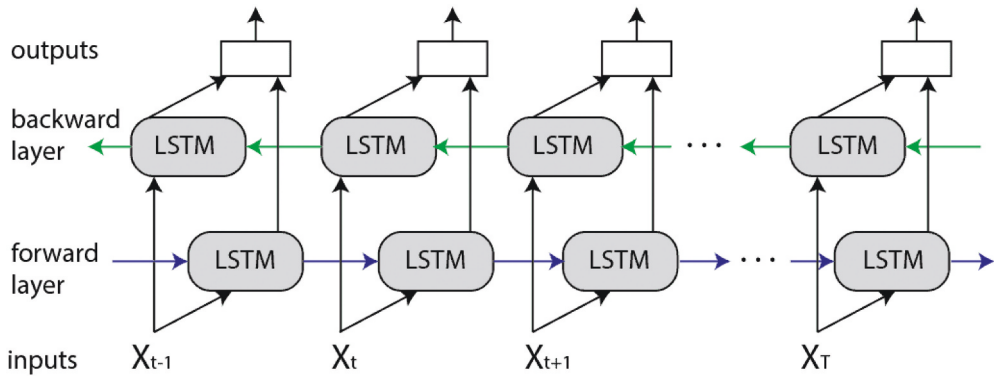


Figure 6. BiLSTM architecture.

needed. The input gate controls the extent to which new information is written to the memory cell, the forget gate determines what information should be discarded from the cell, and the output gate controls the release of information from the cell to the rest of the network. This structure enables LSTMs to maintain long-term dependencies and makes them particularly powerful for sequential data processing tasks.

A Bidirectional LSTM (BiLSTM) was introduced in 1997 by [28]. It consists of two LSTM layers: one that processes the sequence in the forward direction (from start to end) and another that processes the sequence in the backward direction (from end to start). The outputs of these two layers are then combined, typically through concatenation or summation, to provide a more comprehensive understanding of the sequence.

Figure 6 presents the BiLSTM architecture. The forward LSTM layer captures the information flowing in the traditional left-to-right direction, while the backward LSTM layer captures the reverse flow of information, from right to left. By

combining these two outputs, the BiLSTM network is able to utilise both past and future context within a sequence, making it effective for tasks that require nuanced temporal analysis, such as defect segmentation in time-sequenced IRT images.

4. Dataset preparation and model training

4.1. CFRP IRT dataset

The dataset used in this study is composed of a set of images originating from a PT inspection of a CFRP laminate. The sample is a unidirectional carbon/PEEK laminate with a fibre volume fraction of 61% with a stacking lay-up of $[0_2/90_2]_6$. The laminate, measuring 100×100 mm, included artificial Kapton tape inserts of different sizes which were strategically placed at specific layers prior to moulding, as illustrated in Figure 7.

The nominal depths of the defects, before moulding, are 0.13 mm for D1, 0.26 mm for D2, and 0.39 mm for D3. Artificial kapton tape inserts, with nominal sizes of 2×2 mm, 3×3 mm, and 4×4 mm were placed at these specific depths (D1, D2, and D3) and in different layers during the moulding process.

During the PT inspections, a midwave infrared (MWIR) camera was used to capture thermal profiles following a short, intense pulse of heat. The IRT camera recorded images at a resolution of 640×512 pixels with a frame rate of 55 hz, resulting in a total of 1,053 images that comprise the dataset used in this study. Figure 8 shows a sample of the acquired dataset with defects being displayed as a temporal evolution of a surface defect captured at different time steps, showing variations in its appearance over time.

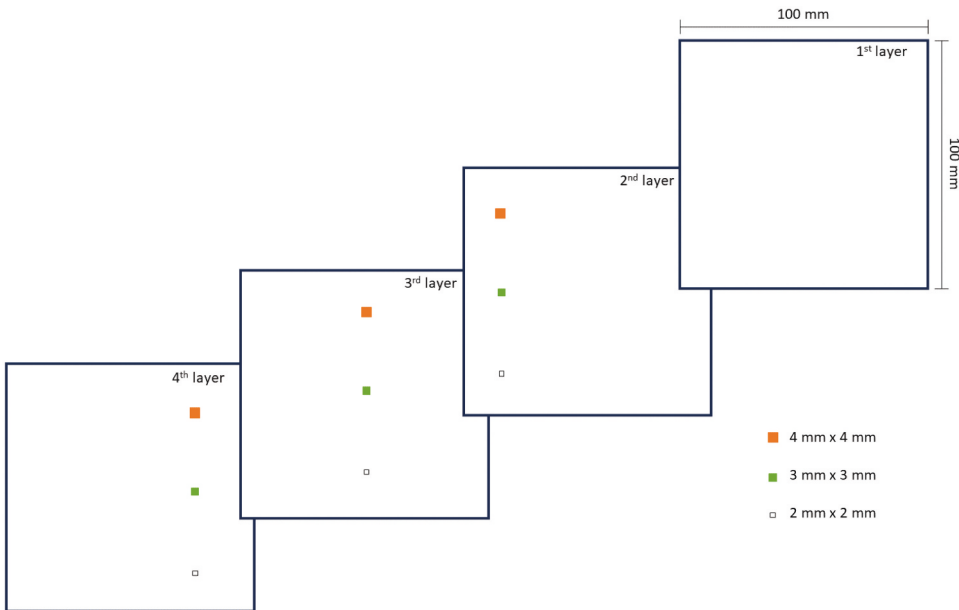


Figure 7. Position of the defects on the layers of the laminate.

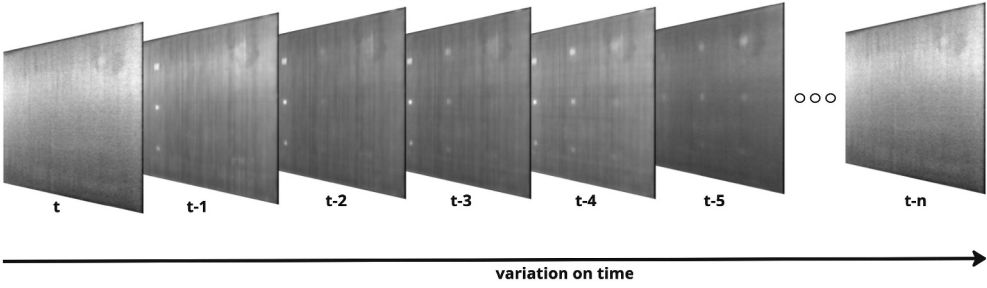


Figure 8. Temporal evolution of surface defects.

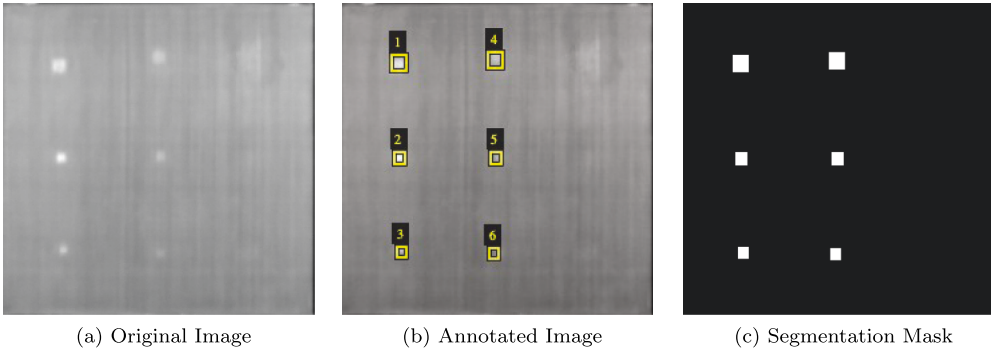


Figure 9. Comparison of annotated image and corresponding segmentation mask.

Analysing such temporal changes is important for defect segmentation, as it enables early detection, tracks progression, and improves the segmentation models.

Data annotation

For this study, all collected images were annotated by an expert to create the dataset to train the model. The annotation process was done using the VGG Image Annotator, a web-based annotation tool developed by the Visual Geometry Group at the University of Oxford [29]. This tool was chosen for its simplicity, ease of use, and flexibility in handling image annotation tasks.

Once the annotation process was completed, the annotated regions were compiled to create the ground truth dataset, which is used as the segmentation mask. This ground truth serves as a reference for training and validating the results of the models created in this study. Figure 9 presents a comparison of the original image, the annotated image, and the corresponding segmentation mask.

4.2. U-Net setup

The model was implemented using the PyTorch library, leveraging GPU processing capabilities that are very important for efficiently training large image datasets [30]. The model architecture consists of repeated blocks of convolutional and activation operations. Each block performs two sequences of operations, including a 2D

convolution, batch normalisation, and the ReLU activation function, applied in succession.

The contraction path follows a chained structure, where each block is succeeded by a max-pooling operation to reduce the spatial dimensions. This path includes four levels of double convolutions, with an increase in the number of filters, starting from 64 and scaling up to 512 filters at the final level. The expansion path uses bilinear upsampling operations to enhance the spatial resolution of the learned features. After each upsampling step, the corresponding feature map from the contraction path is concatenated, followed by another convolutional block.

Upon completing the contraction and expansion paths, a final convolution is applied to map the learned features into a binary segmentation map output. This is achieved using a convolutional layer with a single filter, designed to produce the final segmentation.

The model is optimised using the Adam algorithm, with an initial learning rate of 1×10^{-2} . Binary Cross Entropy with Logits Loss (BCEWithLogitsLoss) is used as the loss function. Training is conducted over 60 epochs, with the model parameters being adjusted in a parallel processing environment on the GPU, ensuring efficient and effective learning.

4.3. DeepLabv3 setup

The implementation of the DeepLabv3 model in this study utilises the ResNet-101 architecture as its backbone, which was selected for its robust feature extraction capabilities [31]. This architecture uses a pre-training phase on extensive datasets, which enhances the model's ability to generalise across many visual representations.

To suit the specific requirements of our task, the model's first convolutional layer was adapted to handle standard 3-channel RGB inputs, which is common in many computer vision tasks. However, since thermographic images consist of only a single channel (often representing temperature data in greyscale), the model's architecture had to be adjusted accordingly. Additionally, the final convolutional layer within the classifier module was modified to produce output channels corresponding to the binary segmentation task. This ensures that the model's output aligns with the task's requirements.

The model was trained in a computational environment optimised for high performance, leveraging GPU acceleration. The training process uses the Binary Cross-Entropy with Logits Loss ('BCEWithLogitsLoss') function, which is appropriate for binary classification problems. The Adam optimiser was used to update the model parameters, with an initial learning rate set at 1×10^{-4} . The training was conducted over 60 epochs to ensure the model's robustness.

4.4. BiLSTM setup

The BiLSTM architecture utilised in this study is designed to capture both past and future dependencies within input images. This model is integrated with DeepLabv3,

where the features extracted by the BiLSTM are used as additional input to the DeepLabv3 network.

The BiLSTM model is optimised using the Adam optimiser with an initial learning rate of 1×10^{-3} . The learning rate is reduced on a plateau when the validation loss ceases to improve, allowing the model to fine-tune the learning process. The choice of loss function is task-dependent: cross-entropy loss is applied for classification tasks, while mean squared error (MSE) loss is utilised for regression tasks. Training is conducted over 60 epochs with a batch size of 32, leveraging GPU acceleration to expedite the process.

4.5. Training, validation, and test data

The dataset comprises 1053 images of CFRP samples, which were divided into training, validation, and test sets. Specifically, 70% of the images (737 images) were used for training, 15% (158 images) for validation, and the remaining 15% (158 images) for testing. This split ensures a balanced representation of the defects in each subset, facilitating robust model training and evaluation. Prior to training, each image was normalised to enhance contrast and detail. Additionally, data augmentation techniques such as rotation, flipping, and scaling were applied to the training set to increase the diversity and robustness of the models.

4.6. Evaluation metrics

For this study, F1-Score and IoU were used as evaluation metrics. The F1-Score is a measure of a test's or model's accuracy, considering both precision and recall. It is particularly useful when the classes are imbalanced. The F1-Score is the harmonic mean of precision and recall, providing a balance between these two metrics [32]. It is calculated using the Equation (2):

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2)$$

IoU is a common metric for evaluating the performance of segmentation tasks. The metric quantifies how much a model's prediction overlaps with the true object area [33]. IoU is calculated using Equation (3), by dividing the intersection area between the predicted bounding box and the actual bounding box by the size of the union of the two boxes.

$$IoU = \frac{\text{Intersection Area}}{\text{Union Area}} \quad (3)$$

The closer the IoU is to 1, the more accurate the model's detection. An IoU close to 0 indicates low accuracy in predicting object locations. In this study, we considered defective and sound pixels instead of defect 'object' for the purpose of these metrics.

5. Experiments and results analysis

In this study, some experiments were conducted to evaluate the effectiveness of various image processing techniques for defect segmentation in IRT images of

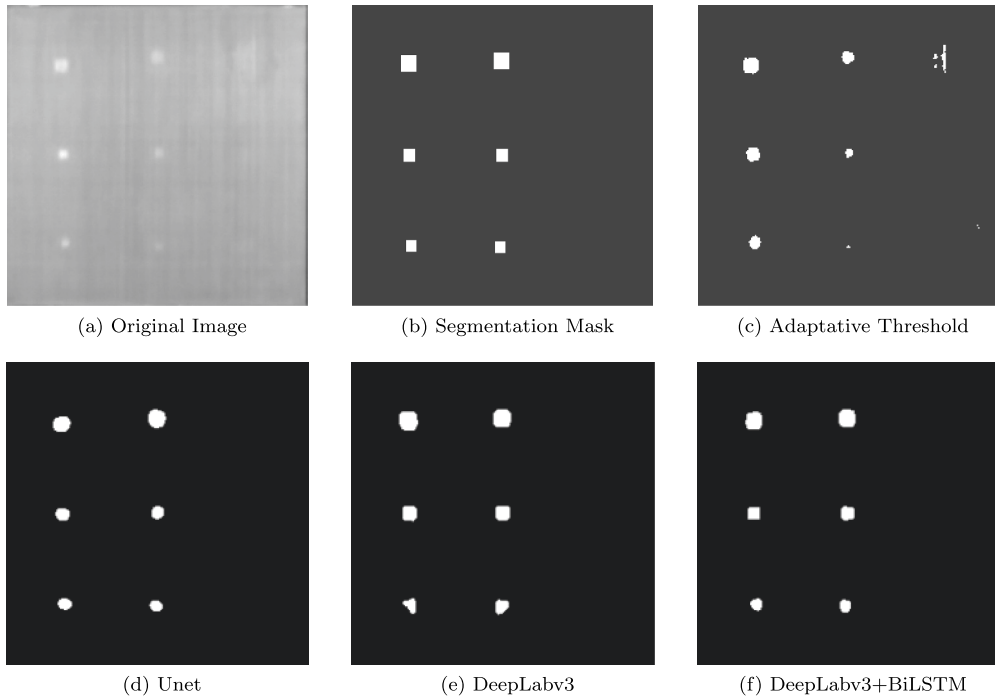


Figure 10. Comparison of the different methods applied.

CFRP with artificial delaminations. Four segmentation methods were employed: adaptive thresholding, the U-Net model, the DeepLabv3 model, and a hybrid approach integrating DeepLabv3 with a BiLSTM network. The segmentation results for a representative image, displaying six defects, are presented in Figure 10.

These defects include two with dimensions of 4×4 mm, two with dimensions of 3×3 mm, and two with dimensions of 2×2 mm. Figure 10(a) shows the original raw image, extracted from time $t = yy$ s from the pulsed thermography sequence, with manually annotated defects. Here, $t = yy$ s refers to the specific time instant in the thermographic recording when the heat diffusion was sufficient to reveal all six defects clearly, where yy is the time in seconds. Figure 10(b) shows the corresponding segmentation mask, Figure 10(c) illustrates the results of adaptive thresholding, Figure 10(d) shows the segmentation using the U-Net model, Figure 10(e) shows the results from DeepLabv3, and Figure 10(f) shows the segmentation achieved using our innovative approach combining DeepLabv3 with a BiLSTM.

The mask for segmentation shown in Figure 10(b) is a binary mask created from manual annotations of the defects as described in Section 4.1.1. Each pixel in the mask is labelled by an expert as either a defect (value 1) or a non-defect (value 0). This mask serves as a ground truth reference for evaluating the segmentation performance of different methods. The manual annotation process involves marking the exact locations and boundaries of defects in the IRT images, ensuring that the segmentation models have accurate targets for training and validation.

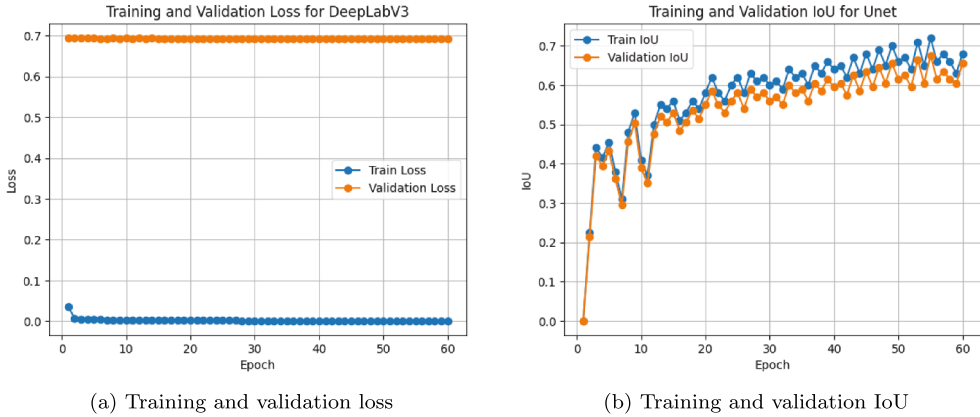


Figure 11. Training and validation process for Unet model.

Initially, an adaptive thresholding method was employed to segment the IRT images, resulting in an average F1-Score of 0.7519 and an Intersection over Union (IoU) metric of 0.2375. While this approach was computationally efficient, the outcomes shows that its segmentation results was limited.

Subsequently, the U-Net model was trained to develop a better robust segmentation approach. The results demonstrated a modest improvement over the initial approach, with the model achieving an IoU of 0.6141 and an F1-Score of 0.9142. This performance, when compared to adaptive thresholding, suggests that a deep learning-based approach is more effective in capturing the complexities associated with the thermal data. Although the model outperformed in terms of IoU and F1-Score, it exhibited signs of overfitting, as illustrated in [Figure 11\(a\)](#), where the training loss decreases rapidly and stabilises at a very low value close to zero. This suggests that the model is performing exceptionally well on the training data, achieving minimal loss. However, the validation loss remains constant and does not decrease over epochs, staying at a relatively high value, indicating that the model is not improving on the validation data. [Figure 11\(b\)](#) shows the training

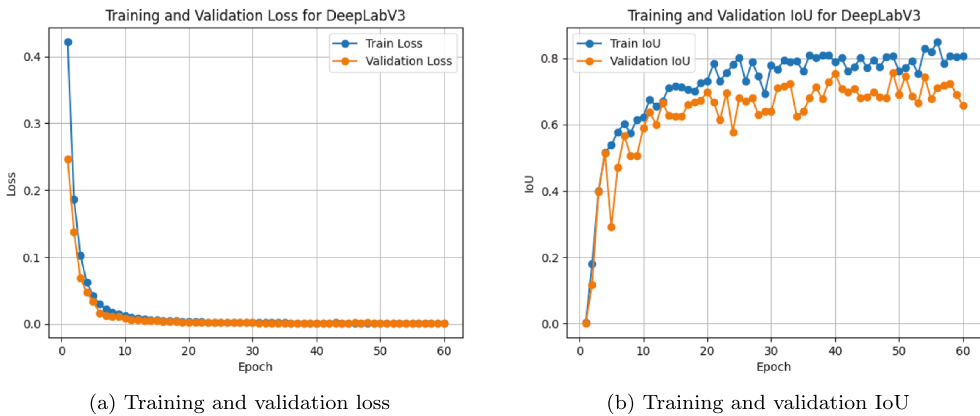


Figure 12. Training and validation process for DeepLabv3 model.

and validation IoU over epochs. The training IoU starts low but rapidly increases within the first 10 epochs, then continues to improve more gradually and it stabilises with some fluctuations. The validation IoU also starts low, increasing quickly in the first few epochs, and then follows a similar trend to the training IoU.

Subsequently, a DeepLabv3 neural network was implemented. This technique demonstrated a good improvement compared to the previously tested approaches, achieving an F1-Score of 0.9406 and an IoU of 0.7419. The high performance of DeepLabv3 shows its ability to segment areas of interest in complex thermographic images.

Figure 12(a) displays the training and validation loss, starting relatively high and decreasing rapidly in the first few epochs. It stabilises around a very low value close to zero, indicating that the model is learning the training data well. The validation loss follows a similar pattern to the training loss, starting high and then decreasing rapidly. It also stabilises at a low value, almost matching the training loss. Both training and validation losses decrease very quickly within the first 10 epochs and then remain low and stable. This suggests that the model is minimising loss and achieving good performance on both the training and validation datasets. Figure 12(b) shows the IoU for training and validation, both of which improve over time, suggesting that the model is learning and getting better at the task of image segmentation.

In the final phase of the experiments, the DeepLabv3 neural network was applied in conjunction with a BiLSTM network. The introduction of BiLSTM is an innovative approach and aimed to enhance the model's performance by incorporating temporal features, leading to a significant improvement in the results. Together, these techniques achieved an F1-Score of 0.96298 and an IoU of 0.83123.

The performance evaluation of the DeepLabv3 model combined with a BiLSTM layer demonstrates robust learning and generalisation capabilities, as evidenced by Figure 13 with a convergence of both training and validation IoU metrics around 0.83. The model's rapid reduction in loss during the initial epochs, with both training and validation losses stabilising near zero, suggests effective learning without overfitting.

The performance of the image segmentation techniques employed in this study was also compared with outcomes reported in similar research. For example, the study

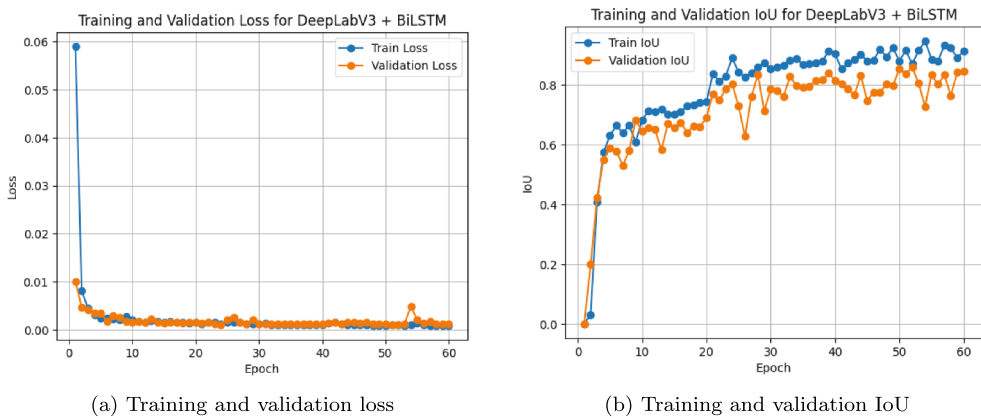


Figure 13. Training and validation process for DeepLabv3 + BiLSTM model.

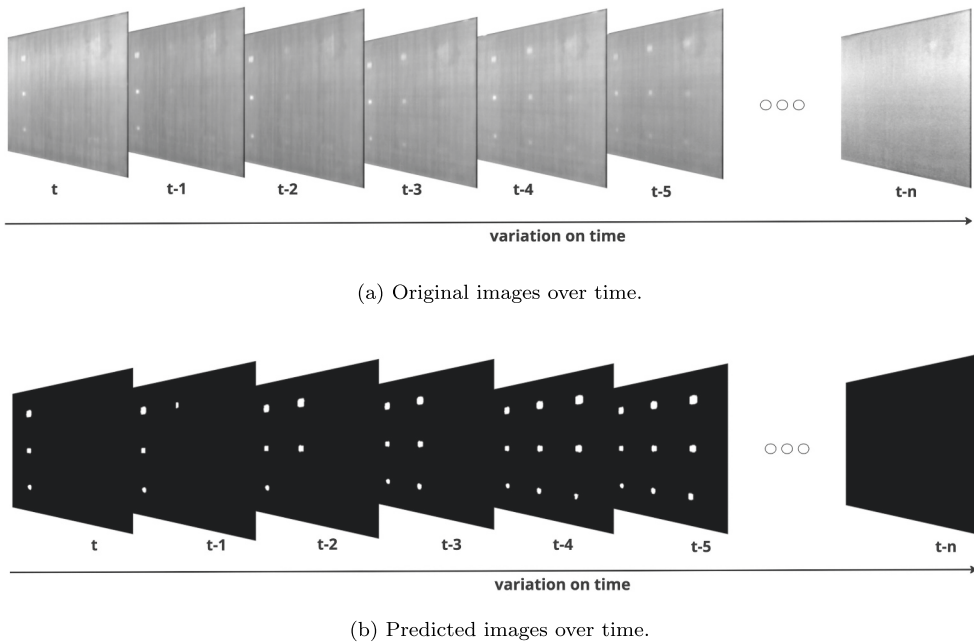


Figure 14. Comparison of original and predicted images over time.

detailed in [25] utilised a UNet architecture and achieved an F1-Score of 0.745 and an IoU of 0.593 for segmenting step-heating thermography images of composite laminates. In contrast, our implementation of UNet achieved an F1-Score of 0.9142 and an IoU of 0.6141, suggesting an enhanced segmentation capability, potentially due to differences in training methodologies and dataset variations. The same study also applied the DeepLabv3 architecture and reported an F1-Score of 0.776 and an IoU of 0.629. Our results using DeepLabv3, with an F1-Score of 0.9406 and an IoU of 0.7419, surpass these results. The aggregation of a BiLSTM layer in the DeepLabv3 model improved the results even more, underscoring the effectiveness of our model's configuration and training process.

In another study, authors in [34] reported a high IoU of 0.924 in their investigation of digital image correlation (DIC)-based damage detection for CFRP laminates using machine learning-based image semantic segmentation. Although this study did not specify an F1-Score, the high IoU indicates highly effective segmentation, which could be attributed to specialised processing techniques or highly optimised model parameters. While our IoU values are lower than those reported by [34], it is important to note that our method is specifically tailored to handle the complexities of thermographic data, such

Table 2. Comparison of results for different models.

Model	F1-score	IoU	Notes
Adaptive Thresholding	0.7519	0.2375	Low performance; struggles with complex defect detection.
U-Net	0.9142	0.6141	Improved performance; slight overfitting observed.
DeepLabv3	0.9406	0.7419	Significant improvement; good generalization without overfitting.
Our proposed method	0.96298	0.83123	Best performance; effectively captures spatial and temporal features.

as variable thermal contrasts and noise levels, which may present different challenges compared to DIC-based methodologies.

Figure 14 illustrates a sequence of frames showing variations over time. Figure 14(a) displays a sample of the original images captured over time, while Figure 14(b) presents the corresponding predicted images generated by the proposed model. These predictions highlight detected defects in the same temporal sequence, demonstrating the model's capability to effectively identify and track defects in CFRP materials over time.

The performance comparison of all models tested in this study is summarised in Table 2. This table provides a concise overview of the segmentation metrics, including IoU and F1-Score, for each approach evaluated. The results highlight the incremental improvements achieved by transitioning from traditional image processing techniques, such as adaptive thresholding, to deep learning-based architectures. The proposed model, DeepLabv3 combined with a BiLSTM network, demonstrated the highest performance metrics, emphasising its potential for defect segmentation tasks in IRT inspections of CFRP materials.

6. Conclusion

This study investigated the application of image processing techniques and deep neural networks, specifically U-Net, DeepLabv3, and BiLSTM, for the segmentation of thermographic images of CFRP. The methods were evaluated using metrics such as F1-Score and IoU, with the innovative DeepLabv3 + BiLSTM combination demonstrating superior performance.

The proposed model achieved an F1-Score of 0.96298 and an IoU of 0.83123, underscoring its effectiveness in identifying structural defects in CFRP through IRT images obtained from PT experiments.

The superiority of our proposed approach, DeepLabv3 + BiLSTM, compared with other deep learning approaches, as described in Table 2 in Section 5, and with related works, as shown in Table 1 in Section 2, suggests that it is particularly well suited for addressing the complexities inherent in thermographic images, making it a valuable tool for NDT.

The results indicate that the proposed model can generalise well to images of different resolutions. However, since the experiments were conducted solely with CFRP materials, its generalisability to other types of materials and defects cannot be guaranteed. This limitation highlights the need for future work, including the creation of a more diverse dataset to enable further experiments and validation across a wider range of applications.

Additionally, while the model demonstrated high accuracy, its computational efficiency in real-time applications should be further analysed and optimised. Future studies could explore lightweight network architectures or hardware-specific optimisations to enhance its applicability in industrial settings where efficiency and scalability are critical.

This work not only focus on the efficacy of advanced deep learning techniques in practical NDT applications but also contributes to future research aimed at optimising these models and extending their application to other material diagnostics. Future studies

will focus on integrating additional data modalities and sensor fusion techniques to further enhance the accuracy of defect segmentation and detection.

Acknowledgments

The authors thank the Centro Multiusuário de Termografia Científica (CEMTEC) for its support in the development of this work.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code [001] and by the National Council for Scientific and Technological Development - Brazil (CNPq) – Finance Codes [407140/2021–2] and [312530/2023–4].

ORCID

Henrique Fernandes  <http://orcid.org/0000-0002-7078-9620>

References

- [1] Kosova F, Altay Ö, Ünver HÖ. Structural health monitoring in aviation: a comprehensive review and future directions for machine learning. *Nondestr Test Eval*. 2024;40(1):1–60. doi: [10.1080/10589759.2024.2350575](https://doi.org/10.1080/10589759.2024.2350575)
- [2] Ibarra-Castanedo C, Genest M, Servais P, et al. Qualitative and quantitative assessment of aerospace structures by pulsed thermography. *Nondestr Test Eval*. 2007;22(2–3):199–215. doi: [10.1080/10589750701448548](https://doi.org/10.1080/10589750701448548)
- [3] Deane S, Avdelidis NP, Ibarra-Castanedo C, et al. Development of a thermal excitation source used in an active thermographic uav platform. *Quant Infrared Thermo J*. 2023;20(4):198–229. doi: [10.1080/17686733.2022.2056987](https://doi.org/10.1080/17686733.2022.2056987)
- [4] Hwang S, An YK, Yang J, et al. Remote inspection of internal delamination in wind turbine blades using continuous line laser scanning thermography. *Int J Precis Eng Manuf Green Tech*. 2020;7(3):699–712. doi: [10.1007/s40684-020-00192-9](https://doi.org/10.1007/s40684-020-00192-9)
- [5] Figueiredo AAA, D'Alessandro G, Perilli S, et al. Exploring the potentialities of thermal asymmetries in composite wind turbine blade structures via numerical and thermographic methods: a thermophysical perspective. *J Therm Anal Calorim*. 2024:1–15. doi: [10.1007/s10973-024-13584-9](https://doi.org/10.1007/s10973-024-13584-9)
- [6] Meola C, editor. Origin and theory of infrared thermography. In: *Infrared thermography recent advances and future trends*. Bentham Science Publishers; 2012. p. 3–28.
- [7] Saeed N, Al Zarkani H, Omar MA. Sensitivity and robustness of neural networks for defect-depth estimation in cfrp composites. *J Nondestruct Eval*. 2019;38(3):1–10. doi: [10.1007/s10921-019-0607-4](https://doi.org/10.1007/s10921-019-0607-4)
- [8] Yang J, Wang W, Lin G, et al. Infrared thermal imaging-based crack detection using deep learning. *IEEE Access*. 2019;7:182060–182077. doi: [10.1109/ACCESS.2019.2958264](https://doi.org/10.1109/ACCESS.2019.2958264)

- [9] Wei Z, Fernandes H, Herrmann HG, et al. A deep learning method for the impact damage segmentation of curve-shaped CFRP specimens inspected by infrared thermography. *Sensors*. 2021;21(2):395. doi: [10.3390/s21020395](https://doi.org/10.3390/s21020395)
- [10] Zhang Z, Hu J, Wang R, et al. Automatic segmentation of microporous defects in composite film materials based on the improved attention u-net module. *Quant Infrared Thermo J*. 2024;1–16. doi: [10.1080/17686733.2024.2387406](https://doi.org/10.1080/17686733.2024.2387406)
- [11] Hegde S, Shenoy BS, Chethan K. Review on carbon fiber reinforced polymer (CFRP) and their mechanical performance. *Mater Today Proc*. 2019;19:658–662. doi: [10.1016/j.matpr.2019.07.749](https://doi.org/10.1016/j.matpr.2019.07.749)
- [12] Heuer H, Schulze M, Pooch M, et al. Review on quality assurance along the CFRP value chain–non-destructive testing of fabrics, preforms and CFRP by hf radio wave techniques. *Compos Part B Eng*. 2015;77:494–501. doi: [10.1016/j.compositesb.2015.03.022](https://doi.org/10.1016/j.compositesb.2015.03.022)
- [13] Swiderski W. Non-destructive testing of CFRP by laser excited thermography. *Compos Struct*. 2019;209:710–714. doi: [10.1016/j.compstruct.2018.11.013](https://doi.org/10.1016/j.compstruct.2018.11.013)
- [14] Wang Q, Liu Q, Xia R, et al. Defect depth determination in laser infrared thermography based on lstm-rnn. *IEEE Access*. 2020;8:153385–153393. doi: [10.1109/ACCESS.2020.3018116](https://doi.org/10.1109/ACCESS.2020.3018116)
- [15] Xu L, Hu J. A method of defect depth recognition in active infrared thermography based on gru networks. *Appl Sci*. 2021;11(14):6387. doi: [10.3390/app11146387](https://doi.org/10.3390/app11146387)
- [16] Lema DG, Pedrayes OD, Usamentiaga R, et al. Automated detection of subsurface defects using active thermography and deep learning object detectors. *IEEE Trans Instrum Meas*. 2022;71:1–13. doi: [10.1109/TIM.2022.3169484](https://doi.org/10.1109/TIM.2022.3169484)
- [17] He Y, Mu X, Wu J, et al. Intelligent detection algorithm based on 2d/3d-unet for internal defects of carbon fiber composites. *Nondestr Test Eval*. 2024;39(4):923–938. doi: [10.1080/10589759.2023.2234548](https://doi.org/10.1080/10589759.2023.2234548)
- [18] Shepard SM. Advances in pulsed thermography. In: *Thermosense XXIII*. Vol. 4360. SPIE; 2001. p. 511–515.
- [19] Gustafsson F, Gustafsson F. Adaptive filtering and change detection. Vol. 1. NewYork: Wiley; 2000.
- [20] Rajinikanth V, Raja NSM, Dey N. A beginner’s guide to multilevel image thresholding. Boca Raton, Florida, USA.: CRC Press; 2020.
- [21] Speakman JR, Ward S. Infrared thermography: principles and applications. *Zoology-Jena*-. 1998;101:224–232.
- [22] He Y, Deng B, Wang H, et al. Infrared machine vision and infrared thermography with deep learning: a review. *Infrared Phys Technol*. 2021;116:103754. doi: [10.1016/j.infrared.2021.103754](https://doi.org/10.1016/j.infrared.2021.103754)
- [23] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference; Munich, Germany*. Springer; 2015. p. 234–241. Proceedings, part III 18 [2015 Oct 5–9].
- [24] Chen LC, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation. *ArXiv Abs/1706.05587*. 2017. Available from: <https://api.semanticscholar.org/CorpusID:22655199>
- [25] Pedrayes OD, Lema DG, Usamentiaga R, et al. Semantic segmentation for non-destructive testing with step-heating thermography for composite laminates. *Measurement*. 2022;200:111653. doi: [10.1016/j.measurement.2022.111653](https://doi.org/10.1016/j.measurement.2022.111653)
- [26] Zhang P, Yin ZY. A novel deep learning-based modelling strategy from image of particles to mechanical properties for granular materials with cnn and bilstm. *Comput Methods Appl Mech Eng*. 2021;382:113858. doi: [10.1016/j.cma.2021.113858](https://doi.org/10.1016/j.cma.2021.113858)
- [27] Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9(8):1735–1780. doi: [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735)
- [28] Schuster M, Paliwal KK. Bidirectional recurrent neural networks. *IEEE Trans Signal Process*. 1997;45(11):2673–2681. doi: [10.1109/78.650093](https://doi.org/10.1109/78.650093)

- [29] Dutta A, Gupta A, Zissermann A. Vgg image annotator (via). 2016 [cited 2025 Jan 18]. Available from: <https://www.robots.ox.ac.uk/vgg/software/via/>
- [30] Ravikumar A, Sriraman H. Acceleration of image processing and computer vision algorithms. In: Srinivasan, A, editor. Handbook of research on computer vision and image processing in the deep learning era. Pennsylvania, USA: IGI Global; 2023. p. 1–18.
- [31] Sonavane R, Ghonge P, Patil SU, et al. Exploring resnet101, inceptionv3, and xception for modi script character classification. Int J Intell Syst Appl Eng. 2024;12:117–124.
- [32] Obi JC. A comparative study of several classification metrics and their performances on data. World J Adv Eng Technol Sci. 2023;8(1):308–314. doi: [10.30574/wjaets.2023.8.1.0054](https://doi.org/10.30574/wjaets.2023.8.1.0054)
- [33] Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning. Sci Rep. 2024;14(1):6086. doi: [10.1038/s41598-024-56706-x](https://doi.org/10.1038/s41598-024-56706-x)
- [34] Wang Y, Luo Q, Xie H, et al. Digital image correlation (dic) based damage detection for CFRP laminates by using machine learning based image semantic segmentation. Int J Mech Sci. 2022;230:107529. doi: [10.1016/j.ijmecsci.2022.107529](https://doi.org/10.1016/j.ijmecsci.2022.107529)



DSPACE

<https://dspace.org/>

Spatial and temporal deep learning algorithms for defect segmentation in infrared thermographic imaging of carbon fibre-reinforced polymers

Garcia Vargas, Iago; Fernandes, Henrique

2026-01

Taylor & Francis

<https://doi.org/10.1080/10589759.2025.2457593>

Garcia Vargas I, Fernandes H. (2026) Spatial and temporal deep learning algorithms for defect segmentation in infrared thermographic imaging of carbon fibre-reinforced polymers. *Nondestructive Testing and Evaluation*, Volume 41, Issue 1, January 2026, pp. 45-65