University of BRISTOL

Department of Aerospace Engineering
AENGM0032 Research Project
2022-2023

AUTOMATIC CLASSIFICATION OF DAMAGE IN COMPOSITE MATERIALS USING DEEP LEARNING AND VIBROTHERMOGRAPHY

Don Issac Joseph

Department of Aerospace Engineering, University of Bristol, Queen's Building, University Walk, Bristol, BS8 1TR, UK

ABSTRACT

Deep learning algorithms such as convolutional neural networks (CNNs) are an established technique for image classification within the field of computer vision and have been demonstrated to be able to identify and classify defects in materials such as composites. This report presents a deep transfer learning-based damage classification model for composite materials utilising fine-tuned, pre-trained models of the ResNet50 and VGG16 convolutional neural networks. The models are trained on thermal image data obtained from vibrothermographic non-destructive tests on composite fibre-reinforced polymer (CFRP) panels with known locations and sizes of delamination defects. Modifications to the network architecture of the pre-trained models are implemented and a hyperparameter tuning study is conducted to optimise both CNNs to create robust and accurate damage classifiers. Both CNN models are then evaluated and directly compared using a testing subset of the vibrothermographic image data with standard classification performance metrics (such as accuracy and precision). It is found that both CNN models can accurately classify delamination defects in composite materials, achieving high binary test accuracies of up to 100%, with the ResNet50 model marginally outperforming the VGG16 model. Grad-CAM visualisations confirm the validity of the classification performance of both models.

Keywords: Deep Learning, CNN, Vibrothermography, Composites

1 INTRODUCTION

The Boeing 787 Dreamliner was launched on the world stage in 2007 as the first major commercial airliner to use composites as its primary material in its airframe. This design change has set the stage for future aircraft, which are increasingly using composite materials instead of metallic alloys for many primary and secondary structures including, load-carrying structural components [1]. Composite materials present unique advantages over metallic alloys due to their high specific mechanical properties such as high specific stiffness and specific strength [2] which has led to overall mass savings. Despite these benefits, composite aerospace engineering structures are prone to rare damage mechanisms, making damage detection challenging and thus increasing the risk of failure. Composite materials present a major challenge since delamination defects are common and not easily detected through visual inspection alone. Therefore, extensive research has been conducted to explore non-invasive techniques for identifying damage during quality control procedures. Infrared thermography (IRT), particularly vibrothermography, offers several advantages for monitoring large structures, including fast scanning, clear two-dimensional image data, and the ability to detect a variety of issues, especially delamination failure at larger depths. Recent studies have successfully demonstrated the automation of image-based damage classification using deep learning approaches. This report aims to investigate the performance of convolutional neural networks in classifying damage in carbon composite materials using image data obtained through vibrothermographic inspection.

1.1 Literature Review

1.1.1 Non-destructive testing techniques & infrared thermography

At present, the identification and location of damage and defects in aerospace components heavily rely on the use of non-destructive testing techniques (NDT). NDT techniques can be fundamentally categorised into direct contact and non-contact methods. Direct contact methods necessitate the installation of inspection equipment onto test structures and rigs. The contact can be direct or indirect utilising a coupling medium. Commonly implemented direct contact approaches include ultrasonic NDT, electromagnetic testing and penetrant testing. On the other hand, non-contact techniques do not require direct contact or coupling media. Most optical NDT methods belong to this category, such as infrared thermography, holography and shearography [3]. Alternatively, other grouping criteria can be used to define categories of NDT techniques, such as those based on the types of signals/radiation used for the inspection, including thermal energy, sound waves and electric currents (eddy currents) [4]. Among the numerous types of NDT techniques applied within aerospace engineering, ultrasonic testing has been the industry standard technique due to its accuracy and penetrative capabilities [5, 6]. However, detecting damage utilising ultrasonic testing can present operational challenges, restricting its practicability and efficacy. For example, inspecting a specimen utilising ultrasonic testing generally requires a high level of operator skill and a strong understanding of the inspection technique. Alongside this there are a number of further requirements for the inspected object, for instance, when a fluid-based coupling medium is used, the object needs to be water-resistant.

Infrared thermography (IRT) is another commonly deployed NDT technique that has several unique advantages, which make it an optimal approach in many use cases, especially for damage detection on larger engineering structures. Specifically, this type of inspection has practical & operational benefits due to it being able to scan large areas within short measurement times [4]. The results and recorded data, come in the form of measured temperature data presented as two-dimensional (2D) images and videos, which can provide clear information on the conditions of the tested specimens so that the analyses of the results are typically simpler and less time-consuming compared to most alternative techniques.

Vibrothermography is a specific application of infrared thermography that utilises the vibratory motion of test structures as thermal heat sources which can be detected and monitored. When an object is subjected to vibratory excitations, heat will be generated internally through mechanisms such as frictional heating, viscoelastic heating, plasticity-induced heat generation and occasionally the thermoelastic effect [7]. Internal heat generation is often caused by and related to damage and defects present on the inspected objects. The thermal energy generated can be conducted through the surface of the inspected specimen and detected by infrared cameras, which can then be used to identify and locate the damage sites. Due to the unique heat generation mechanisms, vibrothermography is a powerful inspection method for detecting a wide range of defects. In addition, as the damage and defects act as the internal heat sources, the travel distances of thermal waves are shortened compared to those in conventional active thermography techniques where the heat is generated and applied externally to the inspected object [8], which suggests that vibrothermography allows for the faster detection of damage and can be especially helpful for detecting issues at larger depths such as deep delaminations defects. In summary, vibrothermography is an exceptionally effective technique for identifying defects and damage within composite structures - especially ones at larger depths due to the number of internal heat generation mechanisms that are utilised during NDT inspections of this type.

1.1.2 Background on neural networks

Machine learning is a vast subfield of Artificial Intelligence that revolves around the development of algorithms that use statistical methods to learn features and patterns from data and utilise this learned knowledge to produce predictions on new data. Deep learning is a smaller subset of this field that focuses on the development of Neural Networks (also referred to as Artificial Neural Networks - ANNs). Deep learning distinguishes itself from other machine learning approaches due to the fundamental composition of the algorithms. They are modelled on the structure and functions of biological neurons in the human nervous system. In a simple sense, they are composed of interconnected nodes, also called neurons, stacked sequentially in layers which work together to perform complex cognitive-like tasks such as natural language processing, object detection and image classification [9].

The earliest known concept of neural networks in the literature was the development of a mathematical classifier model of a biological neuron by McCulloch & Pitts [10]. This worked on the principle of summing boolean input values into a neuron and then applying a threshold step activation function with a set threshold value - to produce a decision based on an outputted probability score. This MP neuron model was extremely limited in its use due to its inability to be generalised to different forms of input data due to it only accepting Boolean inputs and also not featuring a learning mechanism to automatically adjust the threshold value of the activation function. Neither did it incorporate adjustable weights on inputs to aid in optimising the outputs of the model. These flaws were improved upon by Rosenblatt [11] with the development of the first perceptron model – a single-layer neural network. It differed from the MP Neuron model as it could process real non-Boolean inputs and had assignable weights. Furthermore, the model included an error corrective learning procedure to automatically adjust weights on the next step based on the input error and a hyperparameter (which are external parameters of neural networks that configure and control the training procedure) called the learning rate - which dictates the speed at which the updates on the weights are carried out. This allowed the weights to be automatically adjusted based on the input passed to them.

Perceptrons (single-layer neural networks) were improved upon further by adding multiple layers of perceptrons between the input and output layers (called hidden layers) to form multi-layer perceptrons which laid the foundation of modern deep neural networks (DNNs) and are still in use today. These networks are effective due to the multiple hidden layers of the network introducing the capability of MLPs to handle non-linearly separable data such as regression data and hence could be deployed to problems other than binary classification. It also featured an improved learning algorithm called backpropagation [12] - which utilised computations of gradient descent to adjust the weights of the network through minimising a cost or loss function such as a mean-squared error function or a logarithmic loss function. For the task of classification and regression on low-dimensional structured data, MLP networks were highly effective. However, for the problem of image classification, which utilises high dimensional training image data - the complexity and parameters required within MLP networks scaled excessively. Convolutional neural networks (CNNs), the modern framework of which was introduced by Yann LeCun et al. [13] helped to solve this issue by developing a specific type of neural network that could scale with the high dimensional image data without relatively significant or excessive increases in complexity. CNNs form the basis of modern computer vision and image classification and there exists in the literature a wide range of sectors where it has been applied such as in medical diagnosis, early environmental disaster warning and the classification and identification of damage using NDE techniques [14–16]. Goodfellow et al. [17] identified three key benefits of CNN: equivalent representations, sparse interactions, and parameter sharing. Unlike conventional fully connected (FC) networks, shared weights and local connections in the CNN are employed to make full use of 2D input-data structures like images.

1.1.3 Application of CNNs to infrared thermography testing

The increasing deployment of CNNs can be explained by the increased availability of training data as well as the introduction of transfer learning whereby training time is significantly reduced [18] by the use of freely available pre-trained models of advanced CNN architectures such as VGG16 [19] and ResNet [20] that have already been trained from scratch on vast, multi-class image datasets such as the ImageNet database [21] - and so have optimised weights and learned features that can be re-used on different tasks. Transfer learning as described by Weiss et al [22] is a method of utilising learned knowledge from a different domain or dataset and applying this learned knowledge to a new domain. Many studies in the literature have found that transfer learning especially when applied to the use of CNNs for computer vision tasks like image classification - has produced comparably better results [23]. The results of thermography tests are in the form of 2D dimensional images - where damage sites generally appear in the image as bright hotspots, which can be manually analysed to assess the presence of defects. However, in practice, identifying defects in these thermal images is a manual process and can be extremely costly in terms of time, especially when large amounts of data are collected from larger components. There are numerous examples in the literature whereby transfer learning approaches utilising CNNs have been used to automate this classification task. Pierdicca et al [24] successfully utilised the VGG-16 pre-trained CNN model to classify damage on photovoltaic cells using thermography. The popular ResNet pre-trained model was also successfully applied in a study by Deng et al. [25] to classify barely-visible damage on aerospace grade composite materials utilising a specific type of active thermography - called pulse thermography - and achieved classification accuracies up to 99.13%.

1.1.4 Research contribution

Despite the widespread application of CNN-based transfer learning techniques to the classification of defects on objects utilising thermographic inspection - there is no such example applying this approach to thermal image data collected from vibrothermographic tests on aerospace-grade composite materials. As such this is a compelling and novel research topic as it will focus on the development of accurate and robust CNN models that can classify damage on realistic vibrothermographic datasets as well as investigate the direct comparison of the classification performance of ResNet50 and VGG16 pre-trained CNNs.

1.2 Contribution & Objectives

This project will investigate the use of deep learning algorithms in this specific application of IRT by accomplishing the following objectives:

1. Pre-process thermographic data into false-colour heatmaps, partition thermal image data into distinct datasets for testing and apply data augmentation techniques.
2. Develop and train fine-tuned, weight-initialised transfer learning models of the VGG-16 and ResNet50 pre-trained Convolutional Neural Networks in Keras/TensorFlow.
3. Conduct a hyperparameter tuning study to optimise training epoch number and optimiser learning rate.
4. Test and validate the pre-trained CNN models on a selection of datasets from different vibrothermgraphic tests and justify the classification performance of the CNN models utilising Gradient-weighted Class Activation Maps.

2 METHODOLOGY

The proposed methodology for the classification of damage on composite materials utilising deep learning and vibrothermographic inspection techniques contains three key steps. The initial step is damaged composite specimen preparation, implementation of vibrothermographic inspection through the use of an electrodynamic shaker or another actuator (such as piezoelectric discs) and finally, data collection using a digital Forward Looking Infrared (FLIR) camera. The raw output data from the FLIR cameras are processed and mapped to a colour scale to generate pseudo (false colour) heatmaps. The second step involves partitioning the thermal image data into training, validation and testing datasets with the testing data further subdivided into different test cases related to the type of vibrothermogrpahic vibratory equipment used and damage type present on the specimens. The VGG16 and ResNet50 pre-trained CNNs are then trained on the data and a series of network optimisations are conducted to boost classification performance and mitigate overfitting through network architecture modifications and hyperparameter tuning. The final step then involves evaluating the performance of the models and using validation techniques to justify the results.

All algorithmic development of the CNNs was conducted using *Keras/Tensorflow* [26] (in *Python version 3.8.6*) - a powerful open-source deep learning framework that allows highly customisable development of CNN models. All CNN models developed in this study were trained on an NVIDIA RTX 3060Ti GPU.

The proposed methodology is illustrated in **Figure 1**.

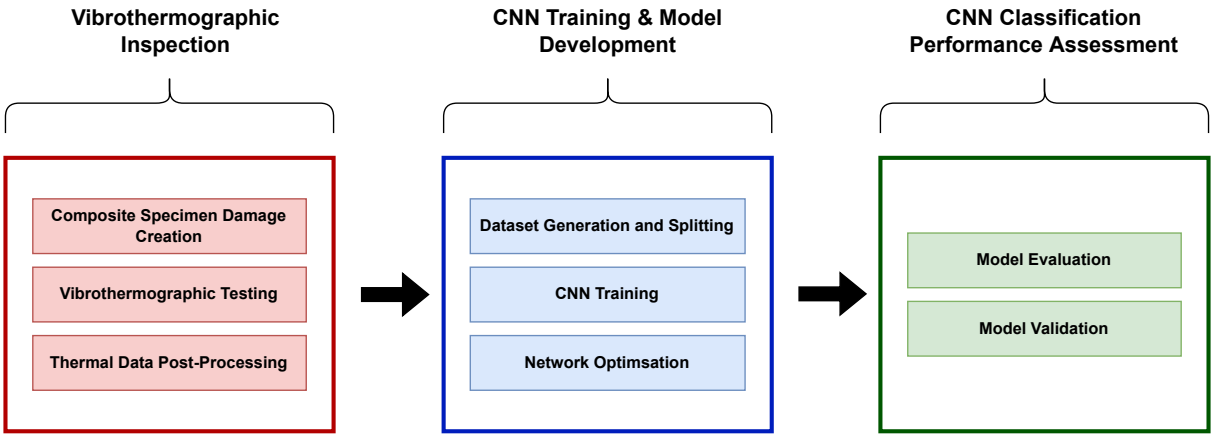


Figure 1: Overview of Damage Classification Methodology

2.1 Data acquisition

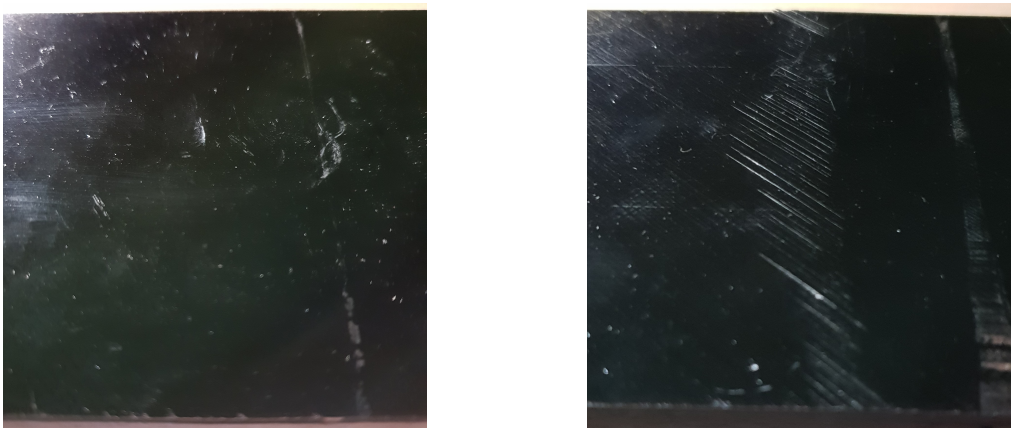
The thermographic data was an output of experimental studies investigating the application of vibrothermography on composite materials by Xintian Chi [27]. The methods, experimental setup and testing are summarised below to gain context on the data collection and to justify the validity of the procedures utilised to produce the data.

2.1.1 Composite specimens and damage creation

Five identical 260mm x 100mm x 0.13mm HexPly 8552 epoxy matric AS4 fibre woven carbon prepreg panels were procured for the dynamic shaking tests and the piezoelectric actuator vibratory tests. Additionally, a separate batch of two CFRP panels with HexCel 8552 epoxy matrix

IM7 fibre woven carbon prepreg panels was utilised for the impact and indentation testing. Continuous fibre CFRP materials of this type are in common use in aerospace primary load-carrying structures and hence present a realistic test case for these damage evaluation experiments.

Artificial damage was introduced into all five AS4 fibre woven composite panels through dynamic fatigue utilising vibration tests with the electrodynamic shaker (instead of the piezoelectric actuator due to its high load output). Ply drops centred in the middle of the plates, created concentrated stress regions which allowed for delamination defects to be generated. The excitation parameters of the electrodynamic shaker were varied across all the panels to create a spectrum of delamination damage. Specimen 1 contained the least amount of defects with specimen 5 containing the most and specimen 3 featuring a moderate amount of surface damage compared to the other 2 panels. Hence these three panels were used for the vibrothermography tests. For the impact and indentation damage vibrothermographic tests, the damage was artificially created on the IM7 woven composite plate using an impact tower. Where a steel ball weight was used as an impactor with a single impact onto the target composite panel to artificially create impact defects. For the indentation tests, the same process was followed utilising the impact tower except the steel impactor was not removed after impact and allowed to rest on the panel for an extended period of time.



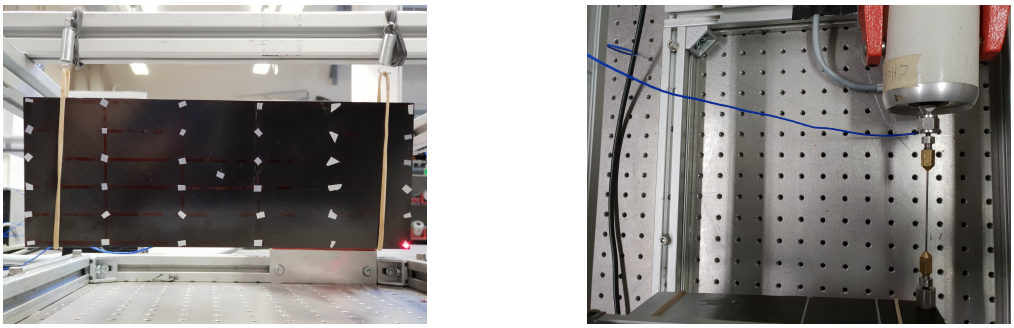
(a) Specimen 1 (b) Specimen 5

Figure 2: The (a) least damaged and (b) most damaged composite plates

2.1.2 Test Equipment

For vibrothermographic tests, the excitatory vibrations could be induced by standard equipment utilised in conventional modal testing so that heat can be generated internally around the potential damage sites on the panels. For the main tests, an LDS V201 permanent magnet electrodynamic shaker is utilised on specimens 1, 3 and 5. These panels were suspended using elastic bands on a frame and the shaker was positioned onto the panel. **Figure 3** displays the full experimental setup with the shaker positioning - however further details on the rig setup can be found in the original report [27]. Additional thermographic tests are carried out utilising piezoelectric disc actuators on the most damaged panel, specimen 5. A FLIR T650sc infrared imaging camera (had a resolution of 640 x 480 and a maximum thermal sensitivity of < 0.02 C) was utilised to measure the heat generation from both these thermography tests. As an extension, 2 more composite panels with artificial impact and indentation damage were also used as test specimens in the electrodynamic shaker tests. In their tests - due to availability issues of the previous infrared testing equipment, a Nippon Avionics TH9100MR infrared camera was utilised, which has comparable thermal sensitivity to the T650sc model but a lower resolution

at 320 x 240.



(a) (b)

Figure 3: The (a) suspended composite specimen for vibrothermographic testing and (b) highlighting attachment point of electrodynamic shaker

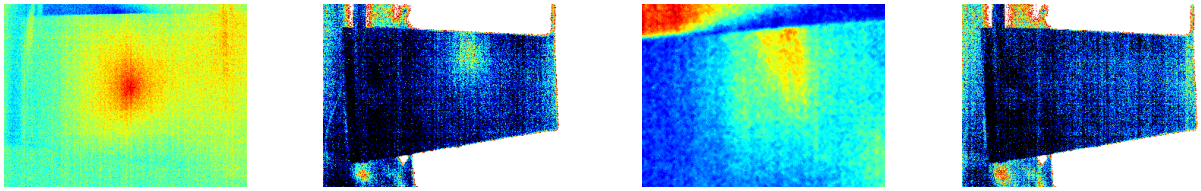
2.1.3 Vibrothermography Tests

Preliminary FE modal tests were carried out to identify ideal vibratory modes and natural frequencies that would generate sufficient amounts of heat at damage sites so that they are resolvable by the infrared camera. These vibratory excitation modes were then targeted on specimens 1,3 and 5 utilising the electrodynamic shaker and the hotspot generation was captured with the FLIR camera. A similar procedure was followed for the vibrothermographic tests utilising the piezoelectric actuator - where an FEA model was created to model the excitatory vibrations induced by the piezoelectric actuator on the composite plate to extract the ideal vibration modes and natural frequencies to target during the vibrothermographic tests. For this test in particular, due to the relatively low internal heat generation due to piezoelectric actuators' low load output - the most damaged specimen was utilised to ensure there was sufficient heat generation that the infrared camera could resolve. The same preliminary modal analyses and vibrothermographic tests were also conducted on the plate with indentation defects and the plate containing impact damage.

2.1.4 Thermal data post-processing

The raw output data from the IR cameras came in the format of 2-dimensional temperature arrays in Degrees Celsius. These temperature tables were converted to false-colour images in MATLAB by using the *jet* colourmap and the *imagesc* function. In total 2068 thermal images were collected (and will be used in this study) with 1060 frames containing locations of detected damage which represent the damage class of training data. As well as 1008 frames not containing any detected damage on the inspected panels - which represent the no-damage class of training data. These images are partitioned further into four groups, based on criteria such as the tests conducted using the electrodynamic shaker on specimens that were (1) dynamically fatigued called the shaker dataset, (2) indentation and (3) impact damage. As well as the additional tests completed on the (4) dynamically fatigued specimens using the piezoelectric discs, called the piezo dataset.

A selection of generated heatmaps from the vibrothermographic inspection is displayed below in **Figure 4**. Hotspots surrounding damage sites appear as bright yellow/red/orange circular hotspots. Lack of a detected hotspot such as in **Figure 4d** indicates no damage present on the composite panel.



(a) Indentation Test Damage Frame

(b) Piezo Test Damage Frame

(c) Shaker Test Damage Frame

(d) Piezo Test Undamaged Frame

Figure 4: (a)-(c) A selection of thermal heatmaps visible containing hotspots around a damage site and (d) showing an undamaged frame.

2.2 Background on CNNs

Convolutional neural networks are a class of deep learning algorithms that have been proven to be highly effective in computer vision tasks such as image classification. The models utilised in this study are pre-trained models with well-defined advanced architectures, optimised weights and learned parameters however it is critical to understand the basic functions and operations undergone in CNNs.

A simple CNN architecture for image classification problems is displayed in **Figure 3** below.

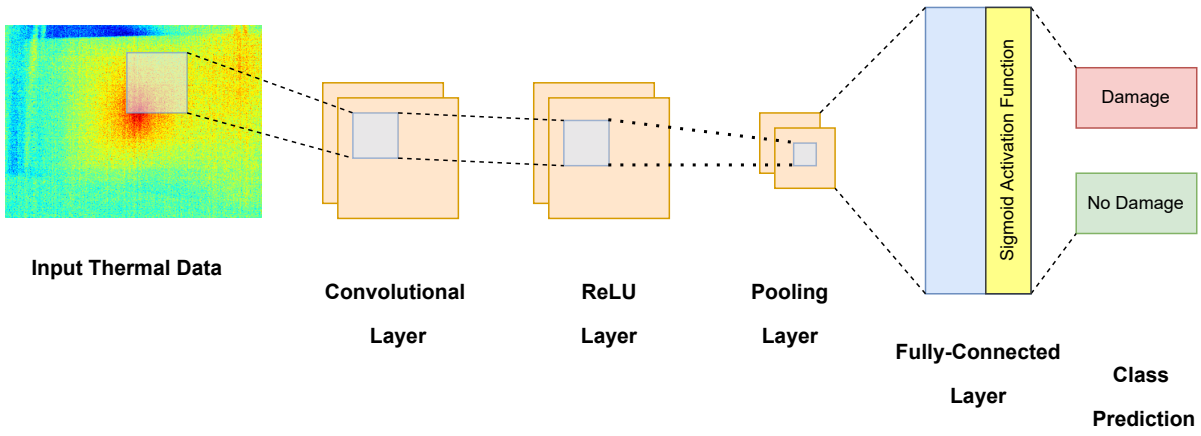


Figure 5: A example of a simple deep CNN architecture for binary image classification

The key foundational piece of a CNN is its convolutional layer - which applies a set of learnable filters (also referred to as kernels) which apply convolutions (a type of multiplicative matrix operation) to input image data which outputs a set of feature maps from each convolutional layer which captures different aspects or features of the image which the network will use to classify the image.

The input x, of each convolutional layer, is organised into three dimensions in 3D coordinates - height (h), width (h) and depth (d), m $\times$ m $\times$ d. A third dimension, depth (d) refers to the channels present in the image and each channel represents a colour of the image - for example, RGB images are three-channel image data and hence $d = 3$. The kernels similarly are defined with three dimensions ($t \times t \times s$). The kernels are small compared to the convolutional layer, so $t < m$ and the kernel depth must be less or equal to the input image of the channel number, $s < d$.

The kernels are then the basis of the local connections between layers, which share similar parameters - bias, b^k and weight, W^k for generating k number of feature maps f^k with a size of

$(m-t-1)$ each and then convolved with the input as aforementioned. The convolutional layer then calculates a dot product between its input (the local section of pixels the filter is applied to) and the weights as displayed in **Equation 1** using a nonlinear activation function (such as the ReLU function [9]), y.

$$f^k = y(W^k \cdot x + b^k) \tag{1}$$

After this the outputted feature maps are down-sized, using a flattening layer or a pooling function (such as max or average 2D pool) which is applied to all outputted feature maps to an adjacent area of size $t \times t$ where the kernel size is defined as t. Generally, after this, the output features from the convolutional layer have a non-linear activation function applied to them, typically the Rectified Linear Unit (ReLU) function [28] to introduce non-linearity into the neural network and allow the learning of complex visual aspects and features. After this stage fully-connected layers (also known as dense layers) contain a layer of neurons (also referred to as channels) that receive the input features and calculate an output based on a weighted sum of the inputted features, followed by an activation function (typically softmax for multi-class classification problems and a sigmoid function for binary classification problems) to generate classification scores which are probabilities from which predictions of the class of which the input image belongs to are evaluated. Dropout layers may also be applied before the final fully connected layer to help mitigate overfitting. It works through the mechanism of randomly dropping out neurons during each epoch in time during training - allowing the feature selection capability to be assigned equally across the whole group of neurons in the final fully connected layers, as well as aiding the CNN model in learning more complex and independent features during training that allow it generalise better to the training dataset.

2.3 Deep transfer learning

This study utilised deep transfer learning by implementing pre-trained models of the ResNet-50 and VGG16 CNNs and their optimised weights after training from scratch on the vast ImageNet database containing in excess of 14 million images [21]. This method avoids the requirement for excessive amounts of training image data to train a CNN from scratch, as the optimised weights of the pre-trained model can be re-trained and updated after training on the new (smaller) dataset. There are two main methodologies for conducting deep transfer learning with pre-trained CNN models, namely feature extraction and fine-tuning [29]. Feature extraction involves training the pre-trained model once (i.e. over one epoch in time) and extracting the outputted feature maps from the convolutional layers and then passing this to a custom fully-connected layer (sub-network) to train on over a given number of epochs in time to predict the classes of the observed images. The fine-tuning approach instead trains all layers of the network over a given number of epochs, allowing the pre-trained features of the CNN to update to the new dataset and thus adapt the network specifically to this problem [29]. This is less computationally efficient as it requires iterative training of all the layers of the pre-trained model over time but typically leads to better performance and generalisation overall. As this study also aims to investigate the direct comparison between the ResNet50 and VGG16 CNNs, this approach allows for a more direct comparison of the full capability of both networks as they can be fully trained on the new dataset [9].

The VGG16 and ResNet50 are selected for this task as explained in section 1.1.3 due to their successful deployment in similar studies classifying damage utilising thermographic inspection. The VGG16 network is a deep convolutional neural network introduced in 2014, it is primarily composed of 13 convolutional layers arranged in 5 blocks of different sizes (max pooling layers after each convolutional block), and 3 final fully connected. The VGG16 architecture that is

utilised in this research is shown in **Figure 6**. The ResNet50 network [20] is an advanced deep CNN and a subset of the ResNet architecture composed of 49 layers of convolutional layers arranged into 5 blocks, 4 of which contain 3 repeating units of layers with pooling layers in between each and a final single fully-connected layer. It has a more advanced architecture featuring 'skip-connections' to allow for an alternative pathway for gradients to backpropagate through the network. A simplified representation of the ResNet50 architecture used in this research highlighting the convolutional and final full-connected layers is shown in **Figure 7**. The networks utilise pre-trained weights obtained from training on the ImageNet database [21] as a starting point, which is then updated further and improved during training to optimise on the vibrothermographic dataset.

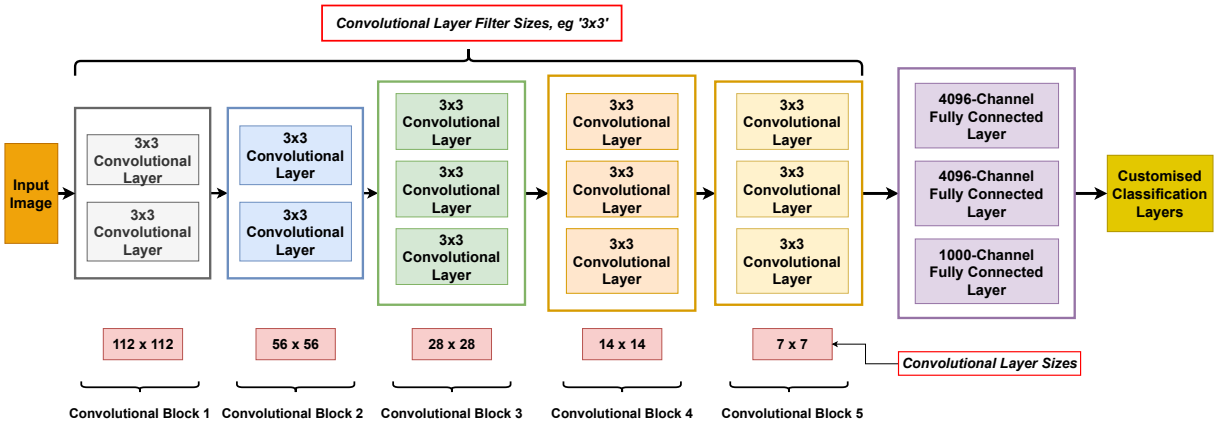


Figure 6: Simplified network architecture of the convolutional and fully-connected layers of the VGG16 convolutional neural network.

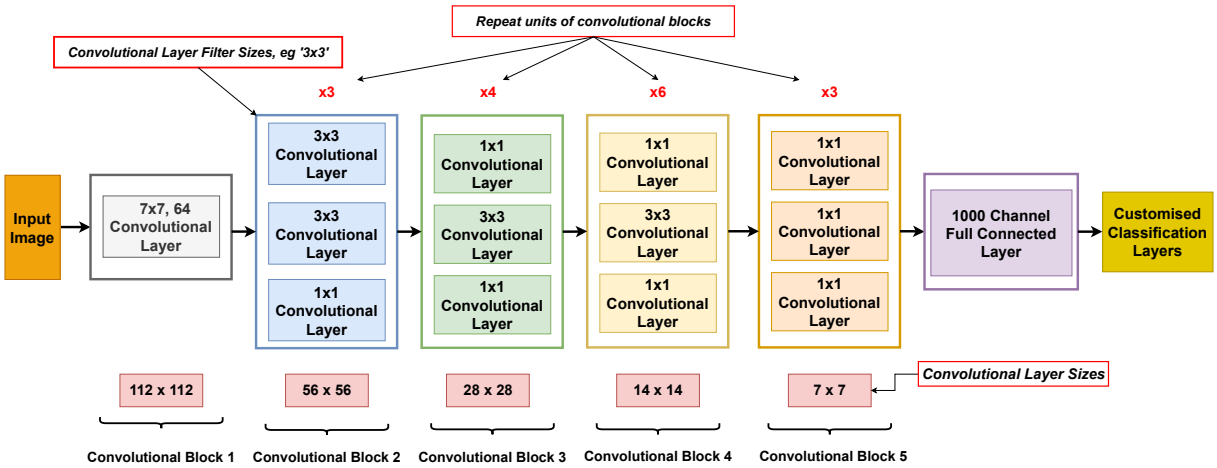


Figure 7: Simplified network architecture of the convolutional and fully-connected layers of the ResNet50 convolutional neural network.

3 MODEL TRAINING & MODEL EVALUATION

Tuning and optimising neural networks is largely a heuristic process and there is no standard algorithm or procedure to follow due to the vast number of adjustable hyperparameters contained in CNN models, meaning any number of approaches can be taken. The main method however to develop robust and accurate image classifiers is to minimise the possibility of the model overfitting (which is when the model produces good classification performance on the training

data but generates poor results on the hold-out test dataset), as well as conducting a suitable number of hyperparameter tuning studies to identify a set of optimal hyperparameters to further improve the classification performance of the model. Learning rate and epoch number (which relates to total training time) can have the most significant impact on image classification performance and so the hyperparameter tuning in this report is confined to these variables.

3.1 Dataset generation & preparation

The present investigation involves combining data from different vibrothermographic tests into a single comprehensive dataset, which is partitioned into training, validation, and test subsets in the ratio of 80%:10%:10% respectively. In addition, whilst a diverse range of data partitioning schemes can be implemented across research studies, a relatively larger portion of data is allocated to the training subset to ensure the model is adequately fitted to the data in this study. In this study, the data from all four constituent datasets are then utilised for training the model, whilst a holdout validation dataset is leveraged to monitor the performance of the model on data that hasn't been observed during training and hence allows the monitoring and prevention of the model overfitting. During the testing phase, the test set is partitioned back into the original constituent datasets, namely, the Piezo, Indentation, and Shaker test data, to enable individual evaluation. After the partitioning of the thermal image data, data augmentation procedures are applied to the training set whereby rotations, random translations and flips are applied to the input images during each epoch so that different variations of the same image are used to train the network. This further adds diversity to the network and allows the CNN model to generalise its training and prevents overfits. This is accomplished using the *ImageDataGenerator* function in Tensorflow [26]. All images are resized to the required 3-channel RGB input image size for the VGG16 and ResNet50 models of $224 \times 224 \times 3$. The images are then preprocessed following the recommendations of the original developers of the CNNs by converting the RGB images into BGR format and zero-centering each colour channel with respect to the ImageNet database without scaling [19, 20].

3.2 CNN Architecture

Final custom classification layers were implemented onto the pre-trained ResNet50 and VGG16 networks and are displayed in **Table 1**.

Final Classification Layers
GlobalAveragePooling2D
Dropout
Dense

Table 1: Final sequentially-ordered classification layers connected to the pre-trained model.

Typically flatten layers are included in the final classification model to compress and downsample the 3D feature maps extracted from the convolutional layers of the pre-trained models into a 1D vector to reduce the number of parameters in the maps before being inputted into the final fully connected layer. However, the downsampling and dimensionality reduction results in a loss of spatial information originally contained in the feature maps which can lead to an inability to capture complex features and patterns in the images and thus lead to poor classification performance. As a result, an average pooling layer is included as shown in **Table 1** which retains more spatial information but continues to downsample the feature maps through computations of the average of the feature maps through their length and height - assigning a single value to each feature map. A dropout layer is also added after the pooling layer as an added measure

to help mitigate and prevent overfits of the training data. Finally, a single dense layer with two neurons is included to evaluate which of the two observed classes (Damage or No Damage) the input image belongs to.

Hyperparameter	**ResNet50/VGG16**
Optimiser	Adams
Loss function	Binary cross entropy
Activation function	Sigmoid
Batch size	32
Training epochs	100
Dropout rate	0.2

Table 2: CNN Configuration

Hyperparameter	**ResNet50**	**VGG16**
Learning rate, η	1×10^{-3}	1×10^{-3}
β_1	0.9	0.9
β_2	0.999	0.999
ϵ	1×10^{-7}	1×10^{-7}

Table 3: Adams optimiser hyperparameter settings for the ResNet50 and VGG16 models

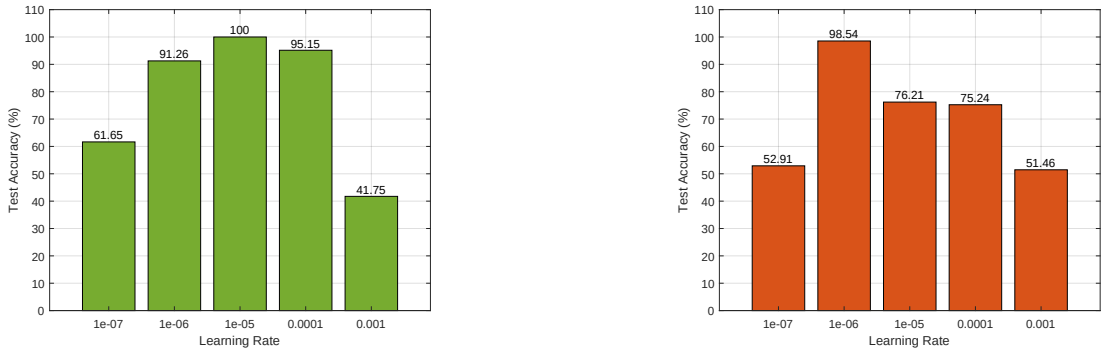
Adams [30] was selected as the best choice for optimiser - as it combines the per-parameter adaptive learning rate capabilities of root mean square propagation (RMSProp) and Adaptive Gradient Algorithm (AdaGrad) and has been recommended as an optimiser to use in deep learning research [31]. The loss function is a cost function that is minimised to help optimise the weights of the network by converging the predicted probabilities and the probabilities of the ground truth. Binary cross entropy was selected as opposed to other popular loss functions such as categorical cross entropy as it is suited for binary classification problems (where inputs are assigned into two classes) where the output of the CNN is required to be a probability value ranging from $0 - 1$ - representing the probability that the input belongs to one of two classes. The sigmoid activation function is the most suited output layer activation function for binary classification problems as it can output probabilities from the dense layer into the range $0 - 1$ which can be easily interpreted as a prediction of the class that the input image belongs (0 is the negative - No Damage class and 1 is a positive - Damage class in this study).

3.2.1 Hyperparameter tuning

The learning rate, η of the optimiser is regarded as the most critical hyperparameter in affecting classification performance as it determines the rate at which updates to the weights of the network are made and therefore determines the rate at which the network converges to an optimal solution. Larger learning rates may cause the weights to update too quickly and converge to a poor solution whereas small learning rates require more training time and may never converge to an optimal solution. Thus a learning rate tuning study was conducted to find an optimal value that maximises test accuracy on the combined test dataset for a defined set of network settings and hyperparameters. In the tuning study, the ResNet50 and VGG16 pre-trained models with the added customised classification layer are trained on the combined dataset for 100 epochs with a learning rate varied between 1×10^{-3} (default learning rate recommended by Keras for the Adams optimiser [32]) and 1×10^{-7} in intervals of a single order of magnitude (10). The trained model is saved and then used to evaluate the observed classes in the combined test dataset to obtain the accuracy metrics. The hyperparameter settings utilised in the tuning study to configure the models are shown in **Tables 2 &3** respectively, the values of the other Adams optimiser hyperparameters such as numerical stability coefficient ϵ and initial decay rates β_1 & β_2 were selected based on recommended values from Keras [32]. Training time is initialised for 100 epochs as a starting point before it is tuned.

During training the TensorFlow's [26] *ModalCheckpoint* callback was utilised to save models before any significant overfitting occurs. Overfits are typically defined by large discrepancies between achieved training accuracies on the training dataset and the evaluatory test dataset

but can also be detected by monitoring the validation losses during training which in the ideal case will continue to decrease over the given number of training epochs in time as the model optimises its learning to minimise the loss function and thus reduce the error in its predictions. The *ModalCheckpoint* has been instantiated to save the model and extract its updated weights at points where the tracked validation losses reach a minimum value. This prevents the model from being trained to a point where it overfits the training data.



(a) ResNet50 (b) VGG16

Figure 8: Bar plots (a) and (b) highlighting the test accuracies obtained through variation of the optimiser learning rate for the ResNet50 and VGG16 model respectively

It can be seen from **Figure 8a** that the optimal *learning rate*, $\eta = 1 \times 10^{-5}$ allowing the model to produce an exceptional 100% accuracy on the test set. For the VGG16 model, the value of the optimiser learning rate that produced the highest test accuracy, was a comparatively lower learning rate of $\eta = 1 \times 10^{-6}$ producing a test accuracy of 98.54%. These are both low learning rates as is expected in training a deep fine-tuned CNN on a relatively small dataset - where large learning rates would likely cause significant overfitting due to the large weight updates.

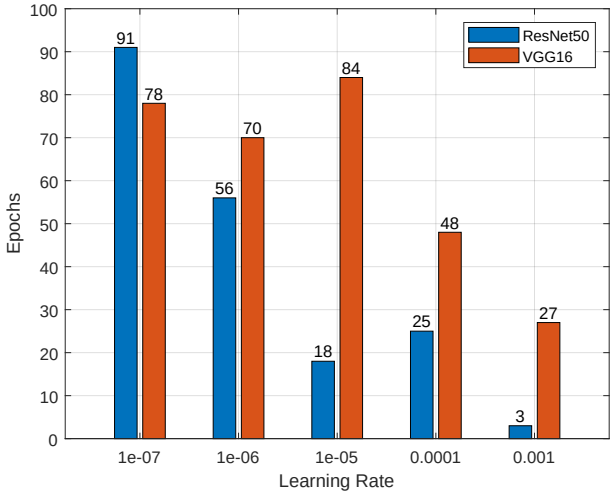


Figure 9: Bar plots showing the epochs in time during training at which the model's weights were saved at the minimal value of validation loss

Figure 9 displays the training epochs in time where the optimal model weights were saved when the minimal value of validation was achieved during training. Saves occurred within the 100 epoch training time across the range of learning rate tuning tests - indicating that the initial training time of *training epochs* = 100 was a suitable value as it allowed the model to be trained

effectively to a minimised validation loss and that no overfits occurred fully across the training time.

3.3 Model evaluation & validation

Typically for machine learning classification problems, key metrics to evaluate classification performance include (binary) accuracy, precision and recall. As our dataset is evenly balanced across both classes with a $1:1$ ratio of images in each image class (Damage and No Damage) - accuracy is a good primary metric to define classification performance in this study.

3.3.1 Classification performance metrics

Accuracy is defined as the proportion of correct predictions made by the CNN model out of all the predictions the model had made and is defined in **Equation 2**, where TP = True Positive, TN = True Negative, FP = False Positive and FN = False Negative.

- True negatives (TN) are defined as the instances where the CNN correctly predicts the negative class (No Damage) for an input image and the actual class of the image is also negative (No Damage).
- True positives (TP) are defined as the instances where the CNN correctly predicts the positive class (Damage) for an input image and the actual class of the image is also positive (Damage).
- False positives (FP) are defined as the instances where the CNN falsely predicts the positive class (Damage) for an input image and the actual class of the image is negative (No Damage).
- False negatives (FN) are defined as the instances where the CNN falsely predicts the negative class (No Damage) for an input image and the actual class of the image is positive (Damage)

Precision is a metric that is defined as the proportion of true positive (TP) predictions among all positive predictions (TP, FP) made by the CNN model and is defined by **Equation 3**. Recall is a classification metric that measures the proportion of true positive predictions (TP) among all actual positive instances in the dataset (TP, FN) and is defined in **Equation 4**. It is a useful metric where in the context of this classification problem the aim is to not only increase the proportion of correct predictions made (increase the accuracy of the model) but to also reduce the False Negative rate. For example, misclassifying a thermographic image containing features suggesting damage on the composite material as undamaged (a false negative result) could lead to the defect going undetected which could lead to structural safety concerns.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{4}$$

3.3.2 Gradient-weighted class activation mapping (Grad-CAM)

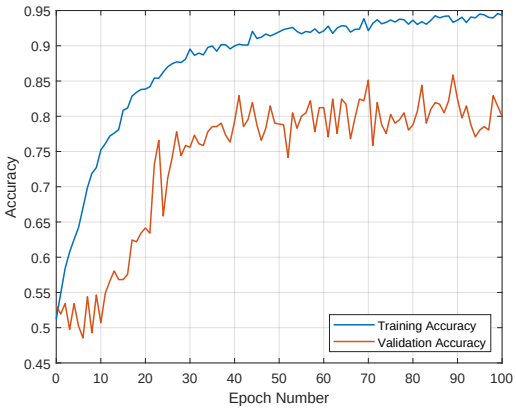
Grad-CAM [33] visualisations aid in explaining and justifying the classification results by a CNN through creating localisation maps via the computation of weighted gradients with respect

to the previous given convolutional layer's feature maps and the model loss, which highlight the pixels in an inputted image into the network that are of most importance when the CNN predicts the class that the input image belongs to. This localisation mapping technique is utilised over similar techniques such as CAM (Class Activation Mapping) [33] as it is more generally applicable across different CNN architectures - particularly the ResNet50 and VGG16 models utilised in this study. The procedure to generate Grad-CAM heatmaps is implemented algorithmically in *Python* using TensforFlow's *gradientTape* function to compute the gradients, which are then utilised to calculate a weighted sum from all the feature maps of the previous (selected) convolutional layer. These weighted activation maps are then downsampled, resized and superimposed on the original input thermal image to generate the Grad-CAM heatmap.

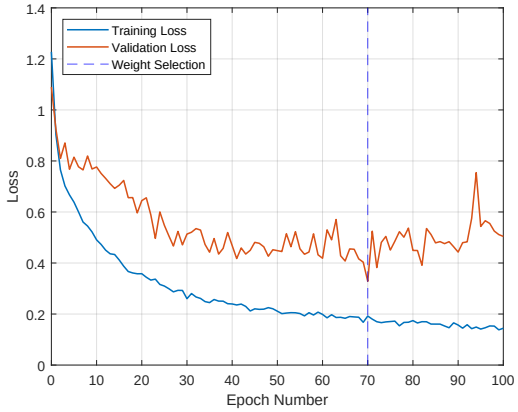
4 RESULTS AND DISCUSSION

4.1 Training profiles

The training profiles from the best-trained models after the tuning of the learning rate are shown in **Figures 10 & 11** - highlighting the monitored accuracy and value of the loss function on the training set and holdout validation dataset over time during the training of the networks.

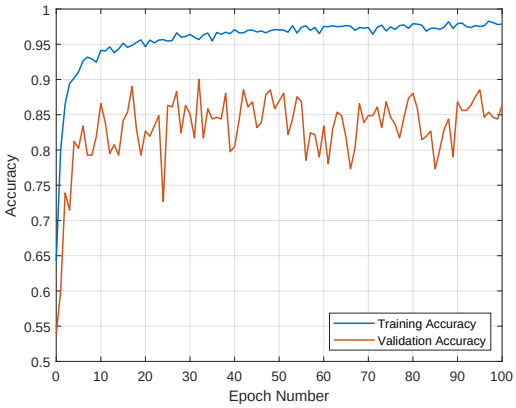


(a) VGG16 Accuracy Profile

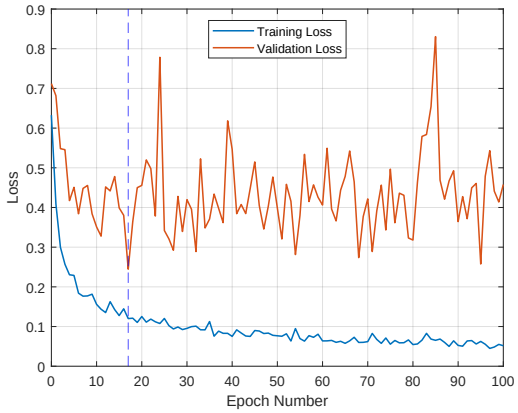


(b) VGG16 Loss Profile

Figure 10: Plots (a) and (b) representing the final training profiles of the fine-tuned VGG16 model after network modifications and hyperparameter tuning



(a) ResNet50 Accuracy Profile



(b) ResNet50 Loss Profile

Figure 11: Plots (a) and (b) representing the final training profiles of the fine-tuned VGG16 model after network modifications and hyperparameter tuning

The VGG16 model trained and converged on its accuracy much slower than the ResNet50 model as shown in **Figures 10a and 11a** which is expected given the lower learning rate at which it was trained. The VGG16 also continued to improve its validation losses until epoch 70 where the model was saved using the *modalcheckpoint* callback, indicating it overfit much later in training compared to the ResNet50 model whose validation losses converged and began increasing after approximately epoch 18 when it's model was saved as shown in **Figure 11b**.

4.2 Classification performance results

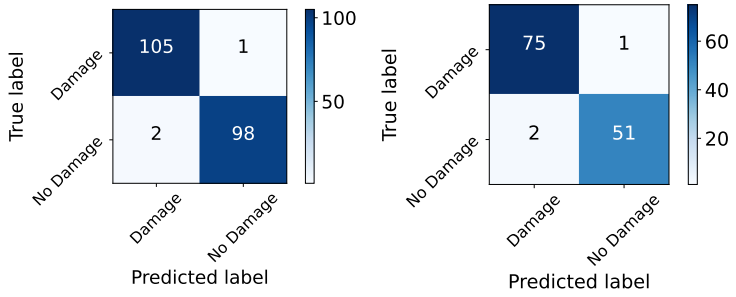
The classification test metrics of (binary) accuracy, precision and recall were evaluated utilising trained and fine-tuned ResNet50 and VGG16 models and displayed in **Tables 4 and 5** below.

ResNet50 Classification Results				
Performance Metric	Combined Test Set	Shaker Test Set	Piezo Test Set	Indentation Test Set
Training Accuracy	94.84%	-	-	-
Test Accuracy	100%	100%	100%	100%
Precision	100%	100%	100%	100%
Recall	100%	100%	100%	100%

Table 4: Summary of test accuracies, test precisions and test recall values on test cases using ResNet50 model

VGG16 Classification Results				
Performance Metric	Combined Test Set	Shaker Test Set	Piezo Test Set	Indentation Test Set
Training Accuracy	92.15%	-	-	-
Test Accuracy	98.54%	97.67%	100%	100%
Precision	98.00%	98.08%	100%	100%
Recall	98.99%	96.23%	100%	100%

Table 5: Summary of test accuracies, test precisions and test recall values on test cases using VGG16 Model



(a) Combined Test Dataset (b) Shaker Test Dataset

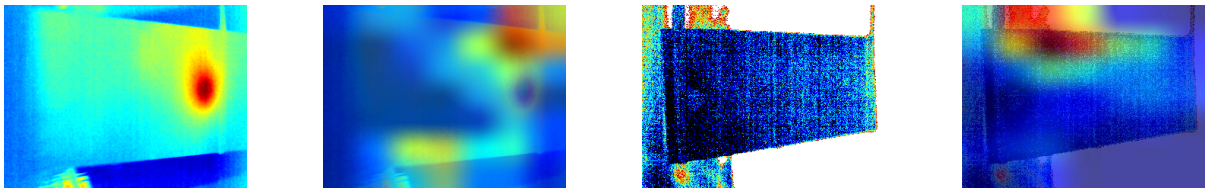
Figure 12: Confusion matrices for the VGG16 Classification Test Results

Referring to **Tables 5 and 4**, both CNNs perform extremely well in classifying damage image frames - achieving test accuracies in excess of 98% on combined set. The ResNet50 model outperforms the VGG16 model on the shaker dataset across all classification metrics (accuracy,

precision and recall). The trained ResNet50 network can be said to be more accurate and reliable due to its higher achieved overall recall (on the combined dataset) of 100% indicating there were no misclassifications of images containing damage features as undamaged. However, there were 3 misclassifications overall by the VGG16 model, two False Negatives and one false positive as shown in the confusion matrices in **12b**. Both models likely performed well due to the simplistic nature of thermal images. The damage frames contained simply non-complex bright circular hotspots around damage sites. This is a simple spatial pattern that both pre-trained networks originally trained on much more complex, multilabel images from the ImageNet database will have learned and optimised network weights for during pre-training. Furthermore, the likely reason for the ResNet50 model's superior classification performance over the VGG16 model is likely due to its more advanced architecture allowing it to prevent overfitting of the training data. The difference between the achieved training accuracy and test accuracy on the combined data set is 5.19% for the ResNet and 6.39% for the VGG16 model referring to **Tables 5 and 4**. The closer value of accuracy between the training and test data indicates the ResNet50 network better fits the training data better and produces a more similar test accuracy when the network model is evaluated. This is likely due to the ResNet's skip-connections modules contained in its architecture around certain convolutional layers [20] which allow the model to learn residual functions which reduce the presence of vanishing gradients in the initial layers of the network during backpropagation when these networks weights are updated. This can aid in reducing overfitting by enabling gradients to propagate backwards through the CNN, leading to better optimisation of the network weights and thus less overfitting and better overall classification performance.

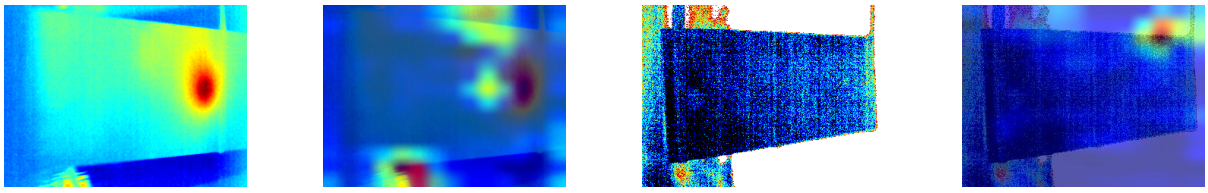
4.3 Grad-CAM visualisations

Grad-CAM heatmaps are generated from the final convolutional layers of the ResNet50 and VGG16 networks and are created using a randomly selected thermal image from the damage class and one from the no-damage class.



(a) Shaker Test Damage Image (b) Grad-CAM of Damage Image (c) Piezo Test No Damage Frame (d) Grad-CAM of Undamaged Frame

Figure 13: (a)-(d) Display Grad-CAMs utilising the ResNet50 CNN displayed beside the original input thermal image.



(a) Shaker Test Damage Image (b) Grad-CAM of Damage Image (c) Piezo Test No Damage Frame (d) Grad-CAM of Undamaged Frame

Figure 14: (a)-(d) Display Grad-CAMs utilising the VGG16 CNN displayed beside the original input thermal image.

Brighter regions of the heatmap that are more yellow/orange/red in colour indicate areas of increased activation in the network. In **Figure 13b** - shows there are increased activations covering the entire region of the hotspot, indicating the model's 'attention' is located correctly around the damage site. In **Figure 14b** - there are increased activations surrounding the left and top edges of the hotspot, indicating the model's 'attention' is again located correctly around the damage site on the composite panel. However, there are also increased activations towards the bottom of the image near the bottom edge of the composite panel - much less so with the ResNet50 network in **Figure 13b**. This indicates the VGG16 model has learned the correct pattern in the damage data and utilised the patterns in these regions to discriminate against when predicting whether an observed image belongs to the damage class - but not to the same degree as the ResNet50 model. Despite this both models' Grad-CAMs display high activations around the hotspots indicating both models' exceptional classification performance is justified and that the model is not detecting and utilising other patterns in the image data unrelated to the regions of damage to make predictions. The activations for images containing no damage features as shown in **Figures 13d and 14d** are located around the top edges of the composite plate as this is a feature present in all the thermal images captured as there is so specific pattern in the thermal image that indicates the lack of defects being present - only the lack of a visible hotspot.

5 CONCLUSIONS & FUTURE WORK

Two fine-tuned, weight-initialised CNN-based image classifier models were developed utilising pre-trained models of the ResNet50 and VGG16 convolutional neural networks to detect damage and defects present on composite specimens detected utilising thermal image data acquired through vibrothermographic inspection. A hyperparameter tuning study was conducted on both to identify an optimal learning rate for the Adams optimiser and also to identify a suitable epoch number to train both networks to, to prevent overfitting and achieve an acceptable convergence on training accuracy. It is found that both models provide strong diagnostic capability in classifying defects - which are delaminations in the composite panels created through different mechanisms such as dynamic fatigue and indentation/impact damage. The ResNet50 model achieved a 100% binary testing accuracy across all test datasets as well as 100% precision and recall metrics with the VGG16 model achieving an average of 98.54% binary testing accuracy. Overall the ResNet50 model outperformed the VGG16 pre-trained CNN - particularly on the Shaker Test set which was the largest test set. This is attributed to the model's more advanced architecture allowing it to fit the training dataset more closely. It can be concluded that deep learning methods can be utilised to successfully classify damage on composites utilising vibrothermogaphic inspection, with the ResNet50 CNN being the better CNN to carry out this damage identification due to its superior classification performance compared to the VGG16 CNN.

To further improve the classification performance of the VGG16 CNN, the VGG19 CNN could be deployed - it is a deeper network composed of an additional three convolutional layers compared to the VGG16 model as a result the additional convolutional layers allow for more complex feature learning in the training data - allowing the model to learn more complex spatial patterns that can aid it in classifying the observed images. However, increasing the number of layers in the network would lead to increased training time and computational expense and thus may not be an efficient approach since the VGG16 model already achieves extremely high test accuracies in excess of 98% and so using a 19-layer VGG19 CNN may only offer diminishing returns. Another approach may be to conduct further hyperparameter tuning, for example, the main optimiser hyperparameter that was focused on for tuning was the learning rate as it has the most significant impact on test performance. However, other optimiser hyperparameters such as the coefficient, ϵ which was kept at a recommended default value of $\epsilon = 1 \times 10^{-7}$, could be tuned to improve

test accuracy. A key simplification in this study was the use of thermographic data on small composite panels with known damage sites. In reality, working on larger aerospace structures such as wings where vast areas may be scanned through using vibrothermography - the output thermal images must be localised on the specific regions of hotspots which can indicate the presence of damage. An extension of this research therefore can be the deployment of a CNN-based object detection algorithm to identify potential thermal hotspots and extract these features from thermal images before using an image classifier to evaluate if there is damage present - such methods have been successfully employed such as by Bang et al. [34] utilising halogen heating thermography but no current research is available on the applications to vibrothermography.

REFERENCES

[1] S. Rana and R. Fangueiro, "Advanced composites in aerospace engineering," 2016.

[2] G. Yang, M. Park, and S. J. Park, "Recent progresses of fabrication and characterization of fibers-reinforced composites: A review," 8 2019.

[3] S. Gholizadeh, "A review of non-destructive testing methods of composite materials," vol. 1, pp. 50–57, Elsevier B.V., 2016.

[4] C. Meola', S. Boccardi', and G. M. Carlomagno', *Infrared Thermography in the Evaluation of Aerospace Composite Materials*. Woodhead Publishing, 2015.

[5] L. Schmerr, "Fundamentals of ultrasonic nondestructive evaluation—a modeling approach, 1998," *Fundamentals of Ultrasonic Non-Destructive Evaluation—A Modeling Approach*, 1998.

[6] J. Szilard, "J. krautkrämer and h. krautkrämer, ultrasonic testing of materials, george allen & unwin ltd, london (1969) 150s.," *Journal of Sound Vibration*, vol. 11, no. 1, pp. 157–158, 1970.

[7] J. Renshaw, J. C. Chen, S. D. Holland, and R. B. Thompson, "The sources of heat generation in vibrothermography," *NDT and E International*, vol. 44, pp. 736–739, 12 2011.

[8] C. Ibarra-Castanedo, M. Genest, S. Guibert, J.-M. Piau, X. P. V. Maldague, and A. Bendada, "Inspection of aerospace materials by pulsed thermography, lock-in thermography and vibrothermography: A comparative study."

[9] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: concepts, cnn architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, 12 2021.

[10] W. S. Mcculloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity* n," 1990.

[11] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain 1."

[12] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986.

[13] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Handwritten digit recognition with a back-propagation network," *Advances in neural information processing systems*, vol. 2, 1989.

[14] R. Zhou, Z. Wen, and H. Su, "Automatic recognition of earth rock embankment leakage based on uav passive infrared thermography and deep learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 191, pp. 85–104, 9 2022.

[15] A. Manickam, J. Jiang, Y. Zhou, A. Sagar, R. Soundrapandiyan, and R. D. J.

Samuel, "Automated pneumonia detection on chest x-ray images: A deep learning approach with different optimizers and transfer learning architectures," *Measurement: Journal of the International Measurement Confederation*, vol. 184, 11 2021.

[16] R. Marani, D. Palumbo, M. Attolico, G. Bono, U. Galietti, and T. D'Orazio, "Improved deep learning for defect segmentation in composite laminates inspected by lock-in thermography," pp. 226–231, Institute of Electrical and Electronics Engineers Inc., 6 2021.

[17] J. Heaton, "Ian goodfellow, yoshua bengio, and aaron courville: Deep learning: The mit press, 2016, 800 pp, isbn: 0262035618," *Genetic Programming and Evolvable Machines*, vol. 19, no. 1-2, pp. 305–307, 2018.

[18] M. Iman, K. Rasheed, and H. R. Arabnia, "A review of deep transfer learning and recent advancements," 1 2022.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 12 2015.

[21] S. U. Stanford Vision Lab, "Imagenet." "https://www.image-net.org/index.php", Accessed 01/03/2023.

[22] K. Weiss, T. M. Khoshgoftaar, and D. D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, 12 2016.

[23] M. Raghu, C. Zhang, G. Brain, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging."

[24] R. Pierdicca, E. S. Malinverni, F. Piccinini, M. Paolanti, A. Felicetti, and P. Zingaretti, "Deep convolutional neural network for automatic detection of damaged photovoltaic cells," vol. 42, pp. 893–900, International Society for Photogrammetry and Remote Sensing, 5 2018.

[25] K. Deng, H. Liu, L. Yang, S. Addepalli, and Y. Zhao, "Classification of barely visible impact damage in composite laminates using deep learning and pulsed thermographic inspection," *Neural Computing and Applications*, 2023.

[26] TensorFlow, "https://www.tensorflow.org/overview." Accessed 01/04/2023.

[27] X. Chi, "Modal-based vibrothermography for damage detection and structural health monitoring."

[28] L. Alzubaidi, M. A. Fadhel, O. Al-Shamma, J. Zhang, J. Santamaría, Y. Duan, and S. R. Oleiwi, "Towards a better understanding of transfer learning for medical imaging: A case study," *Applied Sciences (Switzerland)*, vol. 10, 7 2020.

[29] Z. Li and D. Hoiem, "Learning without forgetting," 6 2016.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 12 2014.

[31] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[32] F. Chollet, "Adams." https://keras.io/api/optimizers/adam/, Accessed 01/03/2023.

[33] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," 10 2016.

[34] H. T. Bang, S. Park, and H. Jeon, "Defect identification in composite materials via thermography and deep learning techniques," *Composite Structures*, vol. 246, 8 2020.