Investigations on Explainable Artificial Intelligence methods for the deep learning classification of fibre layup defect in the automated composite manufacturing

Meister, Sebastian; Wermes, Mahdieu; Stüve, Jan; Groves, Roger M.

DOI
10.1016/j.compositesb.2021.109160

Publication date
2021

Document Version
Final published version

Published in
Composites Part B: Engineering

Citation (APA)
Meister, S., Wermes, M., Stüve, J., & Groves, R. M. (2021). Investigations on Explainable Artificial Intelligence methods for the deep learning classification of fibre layup defect in the automated composite manufacturing. *Composites Part B: Engineering*, *224*, Article 109160. https://doi.org/10.1016/j.compositesb.2021.109160

Composites Part B 224 (2021) 109160

Contents lists available at ScienceDirect

Composites Part B

journal homepage: www.elsevier.com/locate/composites

Investigations on Explainable Artificial Intelligence methods for the deep learning classification of fibre layup defect in the automated composite manufacturing

Sebastian Meister [a,b,*], Mahdieu Wermes [a], Jan Stüve [a,b], Roger M. Groves [b]

[a] Center for Lightweight Production Technology (ZLP), German Aerospace Center (DLR), Ottenbecker Damm 12, Stade, 21680, Germany
[b] Aerospace Non-Destructive Testing Laboratory, Delft University of Technology, Kluyverweg 1, Delft, 2629, The Netherlands

ARTICLE INFO

Keywords:
Defects
Non-destructive testing
Process monitoring
Automation

ABSTRACT

Automated fibre layup techniques are widely used in the aviation sector for the efficient production of composite components. However, the required manual inspection can take up to 50 % of the manufacturing time. The automated classification of fibre layup defects with Neural Networks potentially increases the inspection efficiency. However, the machine decision-making processes of such classifiers are difficult to verify. Hence, we present an approach for analysing the classification procedure of fibre layup defects. Therefore, we comprehensively evaluate 20 *Explainable Artificial Intelligence* methods from the literature. Accordingly, the techniques *Smoothed Integrated Gradients*, *Guided Gradient Class Activation Mapping* and *DeepSHAP* are applied to a *Convolutional Neural Network* classifier. These methods analyse the neural activations and robustness of a classifier for an unknown and manipulated input data. Our investigations show that especially *Smoothed Integrated Gradients* and *DeepSHAP* are well suited for the visualisation of such classifications. Additionally, *maximum-sensitivity* and *infidelity* calculations confirm this behaviour. In future, customers and developers could apply the presented methods for the certification of their inspection systems.

1. Introduction

For the manufacturing of the Airbus A350 XWB as well as the Boeing 787, composite parts are widely used in the aerospace sector [1,2]. Such components are usually made from *Carbon Fiber Reinforced Plastic* (CFRP) and often have considerably superior strength and stiffness properties in comparison to metallic parts. Manufacturing such lightweight parts can be rather expensive. Thus, efficient manufacturing approaches are desired for economical part production. With respect to the strict safety requirement in aerospace production, the fibre layup is followed by an optical testing. In typical cases this inspection needs between 32% [3] and 50% [4] of the fabrication time. That offers enormous possibilities for substantial enhancements regarding processing speed and quality due to the automation of this stage. Automated inspection requires a trustworthy computer based classification of fibre placement defects in image data [5,6]. Machine learning techniques are well suited for classifying fibre placement defects in this respect [7,8]. Liu et al. [9,10] have presented thermographic based approach for the detection of defects in a specimen. To get the necessary amount of training data for their ANN classifier they applied a *Generative Adversarial Network* (GAN) based data augmentation method. However, such ANN techniques often face the disadvantage that the models' decisions are hard to comprehend. This applies especially to ANN or deep learning methods in general [11].

In order to be able to carry out a comprehensive analysis for this, the importance of individual pixels or small image areas for the classification decision must be examined initially. For this purpose, in this paper we compare 20 xAI methods from the literature theoretically. On this basis, we select three suitable techniques for further investigations. Additionally, we introduce the SenseMAX and INFD metrics from the literature, which are typically applied to evaluate the quality and robustness of the chosen xAI algorithms. In this study we consider depth maps of fibre placement defects from the *Automated Fiber Placement* (AFP) production. The AFP manufacturing technique is increasingly used in industry but is still a rather new technology [12]. For this reason, this manufacturing procedure represents the considered application case for this paper. In this way, we would like to facilitate the transferability of our findings to industrial applications [13–15]. In industry as well as in research, the *Laser Line Scan Sensor* (LLSS)

* Corresponding author at: Center for Lightweight Production Technology (ZLP), German Aerospace Center (DLR), Ottenbecker Damm 12, Stade, 21680, Germany.
E-mail address: sebastian.meister@dlr.de (S. Meister).

https://doi.org/10.1016/j.compositesb.2021.109160
Received 17 May 2021; Received in revised form 6 July 2021; Accepted 19 July 2021
Available online 22 July 2021

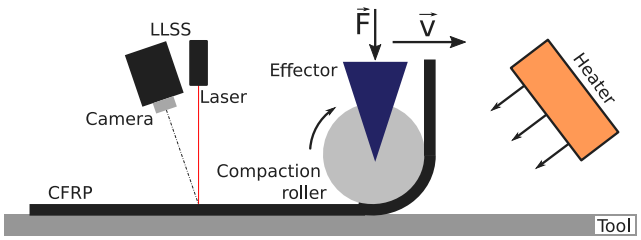


Fig. 1. The fibre placement production procedure is shown schematically. Hence, a compaction roller compresses the CFRP material with force $\vec{F}$. The fibre placement head moves with velocity $\vec{v}$. The LLSS consisting of the camera sensor and the line laser is illustrated.

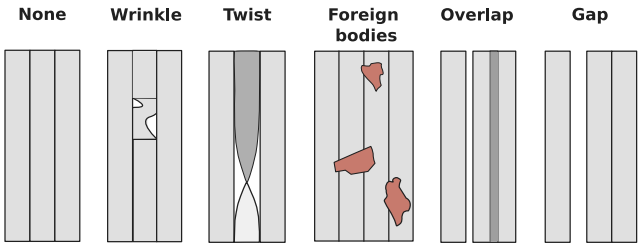


Fig. 2. Five typical AFP deposition defects from the literature along with a flawless layup surface are displayed.

is frequently installed for capturing the respective topology data for the automated monitoring within the AFP production [13,14]. Consequently, we consider greyscale topology images from a LLSS for the research in this paper [7,16].
On the basis of the challenges described above, we will answer the following research questions in this paper:

I. Which procedure is suitable to represent the importance of certain image regions for the decision-making process of a deep learning classifier?
II. Which approaches are appropriate for validating such analyses and checking their quality?

In order to answer the research questions, corresponding algorithms from the literature are evaluated and then selected procedures are examined in detail. The methodology of this paper involves a close investigation of the relevance of image areas for the *Convolutional Neural Network* classification of typical fibre placement manufacturing deviations. Therefore, the *Explainable Artificial Intelligence* methods *Smoothed Integrated Gradients*, *Guided Gradient Class Activation Mapping* and *Deep Learning Important Features with Shapley Additive Explanations* are utilised and investigated. Thus, their visual outcomes are examined first. Then, the accuracy and robustness of such *Explainable Artificial Intelligence* approaches are assessed using the *maximum-sensitivity* and *infidelity* metrics.
To begin with, the following section outlines the state of the art and the necessary fundamentals.

2. Related research

This section describes the fundamentals and related research on the fibre placement process and corresponding defects. Furthermore, the ANN based defect classification and evaluation of important image areas with respect to this classification, is introduced.

2.1. Fibre layup procedure

Common fibre placement approaches are the *Automated Fiber Placement* (AFP), *Automated Tape Laying* (ATL) and *Dry Fiber Placement* (DFP) [12]. Such techniques apply fibre material layerwise to a tool. This procedure is displayed schematically in Fig. 1 and has been explained more closely by Campbell [17]. Especially the AFP technique is applied for manufacturing sophisticated composite structures [3,17]. In this process, multiple straps of pre-impregnated material are placed along a defined path [18]. For this, the fibre layup head applies the composite material, under pressure, to a mould. The CFRP is then heated to improve stickiness and thus processability of the material [19]. Each component is made of several layers of CFRP [17]. The AFP technology allows the production of various component geometries. Furthermore, Rudberg [20] and Parmar et al. [12] predicted a growing usage of the AFP method in future production processes.

Different manufacturing deviations might occur during the AFP production [18]. To this end, Harik et al. [21] already examined the linkage of layup policies, process scheduling and processing with respect to occurring fibre layup defects. Potter [22] considered general deviations in AFP manufacturing. According to both studies [21,22] any fibre placement defect yields geometrical variations in the placed layer. Deviations like *wrinkles*, *twists*, *foreign bodies*, *overlaps* and *gaps* are typical manufacturing deviations from the literature. The Fig. 2 shows such defects schematically. The properties of the individual defect types are listed in Table 1. Thus, *wrinkles* as well as *twists* appear differently, but clearly. As a result, distinct edges and variations in height can be seen. Conversely, *gaps* and *overlaps* are quite similar to each other in terms of their shapes. These defects are flat and their topology varies only slightly. *Gaps* show two tiny edges when viewing them perpendicular to the aligned fibres. In contrast, *overlaps* reveal three tiny edges perpendicular to the filaments. Usually, to distinguish them is quite challenging due to their visual similarities. Such defects were also frequently used in related studies [7,18,21,23]. In addition, foils are investigated as common *foreign bodies* in fibre placement production. These films show totally unique reflective characteristics in relation to the deposited CFRP material [22,24].
The currently performed visual defect inspection carried out through a technician is extremely time demanding. Depending on the defect type and the inspector's constitution, the quality of the inspection can vary strongly. Therefore, the following section describes the LLSS technique for capturing suitable defect topology data automatically during AFP manufacturing.

2.2. Data acquisition and inspection

Inline monitoring in the AFP production is currently receiving much attention from industry and research. The companies InFactory Solutions [14], Electroimpact [13,15], Profactor [25] as well as Danobat Composites [26] utilised LLSS based methods for data recording in the AFP manufacturing process. Such a technology captures 3D topology data of a surface. This is particularly valuable, as this enables a direct geometrical measurement of an inspected manufacturing defect [14]. Schmitt et al. [27,28] examined the capability of LLSS techniques for monitoring edges in fabrics and preforms. Hence, they pointed out that a LLSS provides adequate inspection information for their considered use case. Additionally, Miesen et al. [24] suggested an inspection procedure using a point laser displacement setup and analysed potential disturbing effects which can cause measurement deviations. Sacco et al. [29] presented their ANN based automated AFP inspection system *Advanced Composite Structures Inspection System* (ACSIS) and its functional linkages. This system uses four parallel LLSS to capture a height map of a fibre material surface. On this basis, the system performs an automated defect detection, classification and documentation of AFP layup defects. In their study, they also describe a first idea for analysing the prediction quality for a defect classification. In our recently published studies from Meister et al. [7,16] on defect detection, classification and data synthesis we already used a LLSS

Table 1
The geometrical properties of the defects from Fig. 2 are listed. (l/w) indicates the length-to-width interval. Taking larger geometrical variations into account such values are presented as a ratio. The CPT in this study is 0.125 mm. Regarding the thickness measure + indicates an increase in thickness and − means a reduction.

	Wrinke	Twist	Gap	Overlap	For. Body
Typical ratio: (l/w)	0.5–2	5–10	≤ course len.	≤ course len.	unknown
Thickn. deviation	≥ 3 × CPT (+)	≥ 2 × CPT (+)	≤ 1× CPT (−)	≤ 1× CPT (+)	unknown
References	[18,21]	[18,21,23]	[18,21,23]	[18,21,23]	[21,23]

to acquire the investigated test data. Moreover, the company Allied Vision presented an integrated laser projection and automated visual inspection system [30] together with an ANN based classification algorithm [31] in order to provide a user-friendly overall system for industrial inspection in composites.

2.3. Neural network fundamentals

For the classification of image data, CNN models are frequently used. These models are particularly well suited for processing image matrices as they apply so-called kernels to reduce the amount of trainable parameters. This amount of parameters varies with the quantity and dimension of the applied kernels. Hence, such a CNN examines separate areas of an input image incrementally, which improves the classifier's efficiency [32,33]. Basically, a CNN is an ANN that recognises respective features of an image via convolutions using variously shaped convolution matrices. The CNN structure involves multiple forward connected layers. The features of an image are calculated from the input image while transitioning across the layers of the CNN. The feature complexity often grows as the quantity of layers increases. One convolution layer uses several kernels to derive an appropriate set of feature maps. Every kernel has a separate set of trainable parameters. Multiple feature maps are constructed from various kernels. Lastly, a fully connected layer yields the intended CNN outcome [32,33].
For applying a CNN for the application investigated in this paper, Chen et al. [34] presented a suitable approach for the classification of fiber placement defects from the AFP manufacturing process. This CNN architecture has already been successfully applied in our related study from Meister et al. [7] on the classification and synthesis of fibre layup defects. Here we have used between 25 and 47 real defect images to generate 5000 synthetic defect images per defect class for all six classes. This total set of images was used for training the CNN model. In the conducted GAN-Train GAN-Test analysis for assessing the quality of these synthetic images, standard deviations of the classification rates range from 0.12% to 4.8% except for the none defect class with standard deviation values up to 23.36%. This indicates that the quality of the training images of the CNN corresponds closely to the real process defect images, with the exception of the none defect images, which show significant deviations from original images, but these are less crucial for the CNN training as justified in the study.

2.4. Techniques for explainable artificial intelligence

This section discusses different xAI techniques for evaluating the importance of individual image regions with respect to the ANN decision. For this purpose, first an overview of available algorithms from the literature is presented. Then particularly well suited algorithms for the considered application case are explained in more detail. Finally, metrics for assessing the quality of an xAI analysis are described.

2.4.1. Literature survey

The available xAI techniques from the literature can be categorised into five clusters. The respective operating principles of each cluster are: Gradients, Decomposition, Optimisation, Perturbation and Deconvolution. Relatively a lot of prior research has been conducted in the domains of gradient and decomposition based methods. Considering the available stock of knowledge, such approaches are likely to be promising. Various techniques for representing the importance of particular image pixels or larger areas on a machine decision are listed in Table 2. This summary lists the method's name and the related literature reference. In addition, evaluation metrics from the respective references are given. Bach et al. [35], Shrikumar et al. [36], Sundararajan et al. [37] as well as Lundberg and Lee [38] explained such metrics in more detail. For quantitative comparison, the associated SenseMAX and INFD scores from the study of Yeh et al. [39] are listed where provided. Such metrics will be explained more closely in Section 2.5. When Yeh et al. examined these metric scores for several scenarios for a single xAI algorithm, the mean of the metric scores is taken. Keep in mind, that the input data has a great influence on the INFD and SenseMAX results. Hence, these scores provide just a broad estimate of the performances of individual techniques.
The *Smooth IG*, *DeepSHAP* and *Guided Grad-CAM* methods are very well suited for investigating an unknown scenario due to their different functional principles and performance attributes mentioned from Yeh et al. [39]. Each of these methods interprets different types of information. This information is gained from the neural activations of the ANN and can be used to determine the importance of specific areas of the image when making a machine decision. Furthermore, Lee et al. [11] examined various ways to make the decision-making process of an ANN more comprehensible for a domain expert. For this purpose, they also analysed various xAI procedures and derive a corresponding set of rules via a *Decision Tree* (DT).
Respectively, in the following subsections the principles of the *Smooth IG*, *DeepSHAP* and *Guided Grad-CAM* approaches are outlined more closely.

2.4.2. Smooth integrated gradients

For the following description $F : \mathbb{R}^n \rightarrow [0,1]$ is defined as the transfer function of an ANN. The input data is represented through $x \in \mathbb{R}^n$. The parameter $x' \in \mathbb{R}^n$ indicates appropriate reference data. Where α is a configuration parameter for the direct path between x' and x. Furthermore, for each input image x, the respective class is given by c. Therefore x_i defines the ith pixel of an image. Accordingly, *Integrated Gradients* (IG) is defined for x_i as:

$$IG_i(x) = (x_i - x_i') \cdot \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \cdot (x - x'))}{\partial x_i} d\alpha \tag{1}$$

For this $\frac{\partial}{\partial x_i} F(x)$ represents the gradient of $F(x)$ taken along the ith dimension [37]. This corresponds to the derivation of an ANN following the trajectory from a related input neuron of a certain pixel to its associated output neuron for the observed class c. Smoothing the IG results leads to the *Smooth IG* variant [43].

2.4.3. Deep learning important features with shapley additive explanations

The *DeepSHAP* procedure merges the *Deep Learning Important Features* (*DeepLIFT*) approach with Shapley values which were typically applied for feature selection [36,38]. This *DeepLIFT* method specifies the degree of influence of an input neuron on the difference of the activation of the neuron of class c and the input data matrix x compared to the reference image x'. This is given as:

$$\Delta y_0 = y_{u_0} - y'_{u_0} \tag{2}$$

Therefore, the ANN is reverse propagated. Thus, y_{u_0} represents the neural activation u_0 of a certain layer with respect to the input data x. Similarly, y'_{u_0} gives the neural activation for the respective reference

Table 2
Summary of algorithms from the literature for visualising the importance of certain image regions for a decision of an ANN. This list is inspired by Müller [40], Samek et al. [41] as well as Tjoa and Guan [42] Axiome: Sensitivity (s), Completeness (c), local accuracy (la), Missingness (m), Consistency (con).

Method	Ref.	Category	Axiom	SENSmax	INFD	Operat. area
DeepLift	[36]	decomposition	s, c, la, m, con	(0.64)	(3.49)	global
Input * Gradient	[36]	gradients	–	–	–	global
Integrated gradient	[37]	gradients	s, c	0.826	6.05	global
Smooth integr. gradients	[37,43]	gradients	s, c	0.546	5.95	global
KernelSHAP	[38]	perturbation	s, c, la, m, con	0.64	3.49	global
SmoothGrad	[43]	gradients	–	0.673	5.356	
Occlusion	[44]	perturbation	c	–	–	local
Deconvnet	[44]	deconvolution		–	–	global
GBP	[45]	deconvolution		0.95	6.173	local
Grad-CAM	[46]	gradients			–	
Guided Grad-CAM	[46]	gradients		(0.95)	(6.173)	
C-MWP	[47]	other	–	–	–	
FullGrad	[48]	gradients	s, (s)	–	–	
LRP	[49,50]	decomposition	s, c	–	–	global
Saliency Maps	[51]	gradients	–	–	–	local
Simonyan et al.	[51]	decomposition	–	–	–	
Simonyan et al.	[51]	decomposition	–	–	–	
Zhang et al.	[52]	decomposition	–	–	–	
Kanehira et al.	[53]	other	–	–	–	
LIME	[54]	optimisation	–	–	–	–
Meaningful perturbations	[55]	optimisation	–	–	–	–
Extremal perturbations	[56]	optimisation	–	–	–	–
PatternLRP	[57]	optimisation	–	–	–	–

image *x'*. The neurons of a prior layer are denoted as u_i, where $i = [1, n] \in \mathbb{N}$. Therefore, n gives the total amount of neurons of the considered layer. The ratio $C_{\delta u_i, \delta y_0}$ of a neuron u_i at δy_0 is taken into account for the following calculation:

$$\Delta y_i = y_{u_i} - y'_{u_i} \tag{3}$$

In this regard, $C_{\Delta u_i, \Delta y_0}$ is the ratio of the difference between the input value $u_1, \ldots, u_n$ of a neuron to the difference of the output value u_0 of a neuron. This formulation applies equally to the original and the reference dataset. For all n neurons, this is given as follows:

$$\Delta y_0 = \sum_{i=1}^{n} C_{\Delta u_i, \Delta y_0} \tag{4}$$

Based on these interim outcomes, in the *DeepLIFT* method certain *multipliers* are derived. They are defined as the ratio of $C_{\Delta u, \Delta y_0}$ to the difference Δu, which can be written as:

$$m_{\Delta u \Delta y_0} = \frac{C_{\Delta u \Delta y_0}}{\Delta u} \tag{5}$$

Regarding a deep ANN such individual *multipliers* are concatenated as:

$$m_{\Delta e_i \Delta y_0} = \sum_j m_{\Delta e_i \Delta u_j} \cdot m_{\Delta u_j \Delta y_0} \tag{6}$$

This formulation describes the impact of an input neuron $m_{\Delta e_i \Delta y_0}$ on the importance statement of *DeepLIFT*. Where j represents an ongoing index across those hidden layer neurons which are attached to a certain input neuron. In this regard, e_i represents the input neurons, u_i denotes the hidden layer neurons and u_c refers to the output neuron of a particular class of the ANN. But for the frequently used *Rectified Linear Units* (ReLU) activation function, there is an exception to the *chain rule for multipliers* from Eq. (6). Thus, in the *DeepLIFT* method, the *rescale rule* is applied. In this respect, Δf^+ and Δf^- are expressed as:

$$\Delta f^+ = \frac{\Delta f}{\Delta y} \Delta y^+ \text{ and } \Delta f^- = \frac{\Delta f}{\Delta y} \Delta y^- \tag{7}$$

Taking Eq. (5) into account, the *multiplier* can be written as:

$$m_{\Delta y^+ \Delta f^+} = m_{\Delta y^+ \Delta f^+} = \frac{\Delta f}{\Delta y} \tag{8}$$

Lundberg and Lee [38] have modified the *DeepLIFT* method. Therefore, they have substituted the estimation of the impact of a neuron on the difference of the activation of a neuron with a procedure that is based on *Shapley Values*. Hence, they have named this approach *DeepSHAP*. These *Shapley Values* are from the domain of game theory and respectively constitute the impact of a certain feature on the resulting CNN categorisation. In this respect, such values represent the contribution of a single pixel to the activation of an output neuron. Following Lipovetsky and Conklin [58], the *Shapley Value* of a given feature i is ϕ_i, which is written as:

$$\phi_i = \sum_{S \subseteq M \ i} \frac{|S|!(|M| - |S| - 1)!}{|M|!} \left[f_{S \cup i}(x_{S \cup i}) - f_S(x_S) \right] \tag{9}$$

Where M describes a feature space having a related subspace S. f_S denotes the classifier that was trained using S. The subspaces S which do not contain a certain feature i are aggregated. Hence, the classifier needs to be trained again applying $S \cup i$ as well as S for every single aggregation. This is very costly with respect to an ANN. Therefore, the respective parameters are approximated from frequent sampling of Eq. (9) [59].

2.4.4. Guided Gradient Class Activation Mapping

In the *Gradient Class Activation Mapping* (*Grad-CAM*) algorithm, a defined individual layer of the ANN is chosen for analysis. Selvaraju et al. [46] explained this method, in which the neural activations of the previously defined target layer are considered for evaluating the importance of certain image regions. For this purpose, the output neuron's activation is expressed as y_c. For the application case of a CNN, the parameter A^k denotes an activation matrix of the individual feature maps for the respective selected layer. For *Grad-CAM*, the gradients $\frac{\partial y_c}{\partial A^k}$ of the output neuron c are initially calculated with respect to the neurons of a feature map k of the considered layer. In addition, the *Grad-CAM* approach applies a weighting of the kth feature map from the matrix a_c^k. This weighting is performed using the *global average pooling* method and is specified for an $I \times J$ image as:

$$a_c^k = \frac{1}{J \cdot I} \sum_{i=0}^{I} \sum_{j=0}^{J} \frac{\partial y_c}{\partial A_{i,j}^k} \tag{10}$$

An adaptation of *Grad-CAM* is the *Guided Grad-CAM*. This method additionally multiplies the *Grad-CAM* outcomes with *Guided Backpropagation* (GBP) values. Therefore, the output of the *Grad-CAM* is scaled up to the size of the incoming image for each feature map. This matrix is then multiplied with GBP. Springenberg et al. [45] describe the GBP

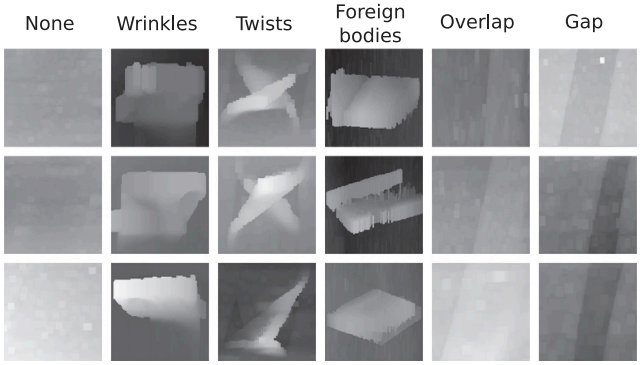


Fig. 3. The figure shows three random original scan images per class under consideration in this scenario.

algorithm more closely. Hence, for the *Guided Grad-CAM* method, the subsequent calculation is applied:

$$\text{Guided Grad-CAM}_c = \text{Grad-CAM}_c(x) \cdot \text{GBP}_c(x) \tag{11}$$

In the following section, metrics are explained which can be used to assess the results of the xAI procedures in terms of their intrinsic quality of an explanation with respect to a given reference.

2.5. Metrics for xAI methods

As especially suitable for the evaluation of xAI procedures, Yeh et al. [39] proposed the use of the *Explanation Sensitivity* and INFD methods.
With respect to subsequent description, R_m is a corresponding reference for the input image I_m of the mth defect sample. For the classification model f, ϕ_f indicates the importance of a given pixel.

2.5.1. Explanation Sensitivity

The *Explanation Sensitivity* method describes the sensitivity of an xAI algorithm for infinitesimally small modifications in an input data set. Therefore, very small adjustments of an input image are carried out initially. Afterwards, this criterion is determined based on the normalised difference of the outcomes from an xAI method. For the calculation of the difference, a modified and a reference data set are considered. Yeh et al. [39] described various possibilities for calculating this criterion. Hence, the SenseMAX variant is frequently used. This method has an upper limit and indicates the maximum sensitivity of an xAI technique to disturbances. The SenseMAX metric is defined as:

$$\begin{aligned} SENS_{MAX}(\phi_f, I_m, R_m) = & max\|\phi_f(R_m) - \phi_f(I_m)\| \\ & \text{, with } \|R_m - I_m\| \leq r \end{aligned} \tag{12}$$

Where r is an customisable value range. The above absolute value $\|\cdots\|$ is calculated using the L_2 norm [39].

2.5.2. Infidelity

The INFD criterion expresses the correlation between an xAI evaluation and the corresponding model of the ANN. Accordingly, this is valid for large feature spaces. Hence, this metric describes the relevance of a single input pixel in relation to the response of the ANN model. Thus, the INFD criterion is described as an expectation value as:

$$\begin{aligned} INFD(\phi_f, I_m, R_m) = \mathbb{E}_{R_m \sim \mu} \Big[& (R_m^T \phi_f(I_m) \\ & - (f(I_m) - f(I_m - R_m)))^2 \Big] \end{aligned} \tag{13}$$

Where the respective reference R_m is formulated as:

$$R_m = I_m - X_0 \tag{14}$$

Table 3
The table reports the amount of test images per class used for experiments in this study.

None	Wrinkle	Twist	Foreign body	Overlap	Gap
86	49	53	22	94	167

Where X_0 represents a random variable having the probability distribution μ. The respective expectation value is approximated through a Monte-Carlo calculation. Please note that of course alternative references can also be applied.

3. Methodology

In the following, the test setup and chosen defect types as well as the selection and application of suitable xAI methods are described.

3.1. Experimental setup

Appropriate defect types were selected for the investigations in this study. Referring to Section 2.1, the following classes were analysed in more detail: Flawless (*none*), *wrinkles*, *twists*, *foreign bodies*, *gaps* and *overlaps*. The Fig. 3 exemplary shows randomly selected and smoothed original defect images from the LLSS, that have been taken as input for the investigations in this study. For the investigations in this paper, we have only examined pre-processed images of such kind, in order to enable comparability of the results. The used test dataset for this study is presented in Table 3. The difference in the number of images is due to the fact that some defect types such as gaps, overlaps and none can be divided into several images, whereas wrinkles, twists and foreign bodies have to be placed individually. These images were recorded using the subsequently described experimental setup. Any image of a defect was manually clipped out of a full LLSS image via the LabelImg [60] tool. The defect images were resized to a suitable size of 128 × 128 px. This dimension was selected as the fundamental attributes of a defect are still shown, however, the data volume is noticeably reduced. Bigger images possibly need extra ANN layers, that probably makes the training more time consuming. Initially, representative original images need to be recorded to perform meaningful tests. This defect data has to be representative for the real manufacturing process. This means that the geometry of the manufacturing defects must have realistic dimensions as well as that the sensor has a similar working distance and lens system as in the real application. This leads to an image area which roughly matches a realistic scenario. Furthermore, the viewing direction of the laser and sensor should be adapted to the real manufacturing process. However, a test setup is required that can generate this realistic data with little effort. Furthermore, the data recording should be reproducible. Therefore, the test setup from Fig. 4 was used. This is not subject to disruptive influences from the manufacturing process. Typical disturbances are contamination, heater radiation or the rotation of the effector. The assembly included an jointed-arm robot, the *Automation Technology GmbH* (AuTech) C5-4090 LLSS [61] and a CFRP test sample. An AuTech C5-4090 LLSS captures 16-bit greyscale depth maps having the dimension of 4096 (w)×500 (h) px. These depth map images represent a full CFRP specimen of the size 250 × 150 mm. In width direction all available pixels of the AMS CMV12000 sensor chip [62] were read out. The dimension of the measurement image in height direction was given from the integration time for each pixel row as well as the duration in between two recording events. In this setup, the laser was projected with a voltage of 5 V. The FIR-PEAK algorithm [63] was used for determining the laser line in the sensor image. The FIR-PEAK approach used a derivative filter which identifies the zero-crossing of the first derivative of the laser brightness image. Furthermore, the scanning speed of the robot was 200 mm/s. The computations in this study were conducted using a computer with an Intel Xeon Gold 5122 @ 3.60 GHz *Central Processing Unit* (CPU),

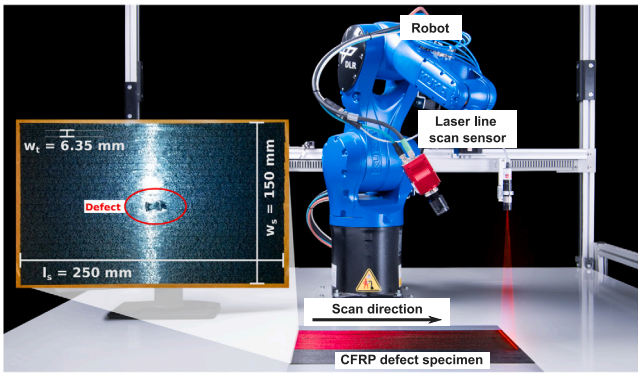


Fig. 4. Data recording setup in which a jointed-arm robot moves the AuTech C5-4090 sensor parallel to a CFRP specimen and captures height profile data of this specimen.

48 GB *Random Access Memory* (RAM) and an NVIDIA Quadro P6000 *Graphics Processing Unit* (GPU). Moreover, Numpy 1.19.1 [64], OpenCV 3.4.2 [65], Matplotlib 3.2.2 [66], Keras 2.3.1 [67] and Tensorflow 2.1.0 [68] were installed together with Python 3.7.7 [69]. Moreover, our trained CNN model with 16 hidden layers from Meister et al. [7], which was previously discussed in Section 2.3, was used for the classification operations in the experiments in this paper. Below the choice of the examined xAI procedures is described.

3.2. Choice of xAI methods

Based on the survey presented in Table 2, methods for visualising the importance of individual pixels or small image areas for the decision of a CNN are required. Please keep in mind that in this study a rather unfamiliar classification case was examined. Therefore, the results from the studies referenced in Table 2 are assumed not to be directly transferable. Hence, comprehensively novel approaches with different working principles, which have already been researched in various fields, were chosen.

Guided Grad-CAM was selected primarily for two reasons. Firstly, *Guided Grad-CAM* does not only consider the importance of individual pixels like many of the other methods, but describes the importance of small image areas. Secondly, this method combines the *Grad-CAM* with the GBP approach. Thus, the advantages of both methods can be used. According to Selvaraju et al. [46], image areas can thus also be assigned a higher relevance, which only gain importance for the classification decision of ANN through the combination of individual pixels.

The methods Input * Gradients, *Layer-wise Relevance Propagation* (*LRP*) and IG are quite similar in terms of their functional principle and the resulting output. However, the IG approaches has the advantage to utilise the identical reference data across all defect categories and thus is quite easy to apply, especially for an unfamiliar application case. Furthermore, the determined results are comparable between the defect images of different defect category. This allows the results to be evaluated more generally. Additionally, the integral approach of this method also has the advantage that only very little noise appears in the outcomes. In order to support this beneficial behaviour even further, an additional smoothing step is applied in the *Smooth IG* method from Smilkov et al. [43]. Hence, this method was chosen for the investigations in this paper. This method additionally shows very good values for SenseMAX and INFD in Table 2, which further confirms this decision.

According to the INFD and SenseMAX values obtained from Yeh et al. [39], these values promise the most robust and faithful outcomes for the *Kernel Shapley Additive Explanations* (*Kernel SHAP*) and *DeepLIFT* methods. According to Lundberg and Lee [38], such a *DeepLIFT* procedure can be extended with *Shapley Values* quite easily, as described in Section 2.4.3. Thus, a combination of both *Kernel SHAP* and *DeepLIFT* can be generated, which is called *DeepSHAP*. Due to the fidelity and robustness of the individual methods, this combined approach was chosen for the investigations in this paper.

Therefore, the methods *DeepSHAP*, *Guided Grad-CAM*, *Smooth IG* were finally selected for the investigations in this study.

3.3. Investigation of the xAI findings

The analyses in this paper using the previously described and selected xAI methods are based on the CNN architecture and the associated trained model from the preliminary research by Meister et al. [7].

3.3.1. Evaluating the visual outcome of the xAI calculations

For this CNN as well as the examined defect classes, a visual xAI outcome was determined for three randomly selected real recorded fibre placement defect images. The visual results were displayed in each case as greyscale and as a colour image. Based on the brightness of individual pixels in the greyscale representation, their overall influence on the decision of the CNN can be evaluated. The colour representation of a pixel's relevance for the CNN enables the evaluation of the behaviour of the CNN model. Red pixels in the explainability image describe a relevance with a negative sign and green pixels an importance with a respective positive sign. The meaning of the values and the signs depends on the calculations of the respective methods from Section 2.4. Subsequently, the methodology for investigating behavioural changes in neuronal activation with varying input data is described.

3.3.2. Investigating the change in CNN neural activations for modified data

The aim of this experiment was the validation of the correlation between the xAI explanation result and the classification decision of the CNN. The experimental design was based on the validation experiment from the research of Srinivas and Fleuret [48].

The levels of activation at the CNN output neurons were monitored for each input defect image under consideration with respect to the corresponding defect category. This provided the reference level of activation. Hence, a respective xAI statement was computed as well. Subsequently, on the basis of the estimated xAI significance of every pixel, the n most relevant pixels were changed to zero in the initial input image. Then, the previously outlined evaluation of the neural activations was repeated for this modified input image. For the experiments performed in this study, n was varied in the interval $[16, 4096]$ with a step size of 16. The given interval describes the proportion of around 0.1–25% of the overall image pixels. The step size 16 was chosen in such a way that it allows for a suitable resolution of about 0.1% pixel, but still does not require too much computational effort. These values can also be chosen differently, depending on the application. For a closer comprehension of the results, the above analysis was additionally carried out three times for each of the six defect classes with a randomly selected n in the given interval. The average value of the three Monte-Carlo calculations for each class was used for comparison.

The analysis of the neuron activations served to validate the results of the xAI calculations of the importance of individual pixels or small image areas with respect to the classification decision. Based on the procedure described above, we assessed the actual influence of an image pixel on the activation of the output neurons and the associated classification decision. Furthermore, the correlation of these findings with the evaluation of the importance of individual image areas from the xAI calculations was examined in this way. In this study, the level of variance in neural activation with respect to the predefined baseline activation for the unmodified data indicates the influence of the neural activations.

Below the procedure for examining the ANN response for a targeted destructive data manipulation is outlined.

3.3.3. Analysing the CNN neural activations for explicitly destructive manipulations

In the experiment described above, the influence of a modified defect image on the activation of the output neuron of a certain defect class was investigated in relation to the activation for the corresponding unmodified input image. In this experiment, the activations of all output neurons of the CNN in the second last layer were investigated. For the conducted tests, the original input images were again manipulated. However, this modification of the input data differs from the procedure described above and is therefore explained in detail subsequently.

The intensity of pixels that have a negative sign in the calculated xAI statement were replaced with the maximum pixel intensity in the input image. Pixels with a positive sign in the xAI output were annotated with the maximum negative value. Thus, a defect image was modified according to the associated xAI calculation in such a way, that the activation of the actual class was minimised and the activations of all other classes were maximised. The pixels were manipulated in descending order, according to the magnitude values of the xAI calculation, for each pixel. Thus, the pixels that were judged to be most important according to the xAI calculation were manipulated first. However, a distinction was made according to the sign. The magnitudes of the positive and negative explanations were evaluated separately. Analogous to the previous experiment, a percentage n_r of the pixels was manipulated, where $\frac{1}{2}n_r$ pixels were allotted each to positive and negative signs.

3.3.4. Evaluating the maximum sensitivity and infidelity scores

For the quantitative comparison of the xAI methods, their SenseMAX and INFD values were calculated.

The SenseMAX criterion was determined according to the definition in Eq. (12). For these investigations, we applied a Monte-Carlo simulation with 50 runs, as outlined from Yeh et al. [39] in their research. The experimental setup using a noisy baseline was also inspired by Yeh et al. [39].

The INFD scores were calculated following Eq. (14). According to Yeh et al. [39], for these calculations the respective expected value from Eq. (13) was again estimated via a Monte-Carlo simulation using 1000 data samples.

For each of the previously performed Monte-Carlo simulations a Gaussian noise with a standard deviation of $\sigma = 0.2$ around the original initial pixel value as its mean was applied to the input images.

4. Results

This section presents the results of the performed experiments. As explained before, the tests were carried out with the CNN architecture and the corresponding trained model from the research of Meister et al. [7].

4.1. Classification accuracy CNN classifier

In order to provide a more comprehensive view of the performance of this CNN classifier, Table 4 lists the associated classification rates per defect class for the original test data used in this paper. We can see that, with the exception of *gaps* and *overlaps*, all defect images are correctly classified. The classification rate for *gaps* is 96.81% and for *overlaps* 98.94%. The images incorrectly predicted in these two categories are in each case classified as none. Nevertheless, these classification scores are still very high and sufficient for the application under consideration. However, we should also mention that this model was trained on manually generated, close-to-reality data, which was augmented afterwards. As already stated in this former study, there is a chance that this CNN model does not accurately represent manufacturing defects from real world processes. However, the robustness of this model is less relevant to the analyses in this study, so the CNN model is assumed to be suitable based on the detailed analyses from our previous study.

Table 4
The table presents the determined classification rates for the original defect images considered in this study using the CNN architecture and the trained CNN model from Meister et al. [7].

Class:	None	Wrinkle	Twist	Foreign body	Overlap	Gap
Accuracy:	100%	100%	100%	100%	98.94%	96.81%

In the following sections, the activation of individual neurons and the link to the classification decision is investigated in various ways. Thus, in the next section, the importance of individual pixels for the classification decision are examined and visualised.

4.2. Evaluation of visual xAI outcomes

Fig. 5 displays the xAI outcomes in a visual manner for the six examined defect categories. For each category, the raw input data of the CNN is presented on the left. To the right, for the investigated xAI methods, the magnitude relevance scores are displayed as greyscale images and the signed importance values are illustrated as a colour image, where red pixels describe a negative value and green pixels a positive value. As outlined above, the analysis was carried out for three randomly selected real defect images. Below we discuss the findings for the clusters of distinct and less distinct defect types separately.

Distinct defect types

In this study such distinct defect classes are *wrinkles*, *twists* and *foreign bodies*. For these classes, the *Smooth IG* method yields the most homogeneous results across the three sample images. The defect regions appear distinctly as green areas against the background. This is due to the multiplication with the respective gradients, which means that the pixel intensities of the input image heavily affect the outcomes of these xAI calculations. Consequently, the outlines of the flaw areas match quite well with the determined xAI image. The additional smoothing in this method strengthens this clear representation of the defect regions and reduces the often typical noise in gradient based xAI outcomes.

The *DeepSHAP* findings indicate a consistent degree of similarity between the input data and the xAI results. The relevance of a given defect area appears to be related to the brightness variation in the input image as well. However, unlike *Smooth IG*, these xAI result images are visibly subject to some sort of statistical noise. In the xAI results, a certain type of skewed pattern is visible across some brighter pixels. This could be an artefact of the CNN architecture. However, this cannot be meaningfully interpreted without further investigations.

The *Guided Grad-CAM* greyscale outcomes yield relevant defect areas, that stand out clearly from the background. However, they only partially correspond to the actual defect region in the original input image.

Less distinctive classes

Within this paper flawless images (*none*), *overlaps* and *gaps* represent the less distinctive classes. The *Smooth IG* method yields xAI results, which primarily assigns larger relevance scores to pixels with higher intensity in the input defect image. The darker regions, which indicate geometrically lower regions of the measurement image, are more likely to be judged as unimportant areas from the algorithm. For the investigated less distinctive defect types, a kind of chessboard pattern can be seen in the visual representation of this xAI method. This pattern could result from textures in the input image. However, such patterns in the defect image most likely arise from the architecture or gradient calculations of the CNN. Accordingly, for *none* defect, a quite homogeneous chessboard pattern is evident over the entire xAI output image.

In the *DeepSHAP* result images, the defect regions of *gaps* and *overlaps* are difficult to recognise. Thus, the importance of the pixels for the classification decision are not attributed to the actual defect in the image. The *DeepSHAP* explanations look very similar to those of the

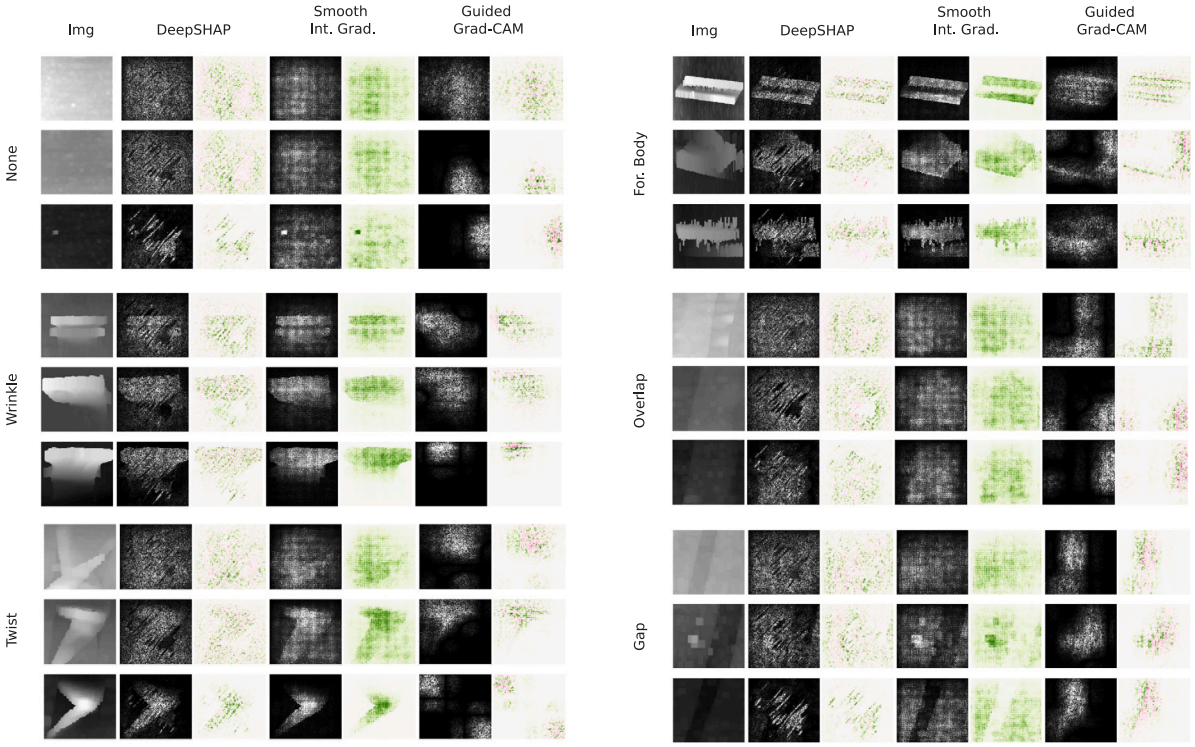


Fig. 5. To the left, the initial input images are given for the examined defect categories. To the right, for *DeepSHAP*, *Smooth IG*, *Guided Grad-CAM*, the greyscale significance score images and the colour image for the signed importance values are presented. Red represents a negative sign and green indicates a positive sign. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

none defect class. Since the classification rate for these inconspicuous classes is still very high, the evaluation of the importance of individual image areas through the *DeepSHAP* calculation might be insufficient. Furthermore, the diagonal pattern already mentioned above is clearly visible in the explainability images. Also in this case, the origin of this pattern cannot be clearly identified, but it might be an artefact of the internal CNN calculations.

The *Guided Grad-CAM* method generates xAI results for the considered defect types, which correspond very well with the defect region. This applies to the inconspicuous defect types examined in this case as well as to the very distinct defect types investigated above. Only for *none* defect images the evaluation of important pixels for the machine decision does not seem to be clearly understandable. Therefore, in some cases, the entire defect image is judged to be important in a relatively homogeneous way. But sometimes only very small parts of the actually homogeneous, flawless fibre layup surface are evaluated as important. This might be caused from pre-processing artefacts being recognised as particularly important for the classification of *none* defect images. However, such artefacts do not represent physical attributes of the actual defect. The influence of these artefacts on the explainability result is therefore not examined in detail in this paper, but gives a reason to adapt the utilised pre-processing.

4.3. Change in neural activations for modified images

Fig. 6 presents the variations in neural activation ΔA_r of each respective output neuron over the relative number of removed pixels n_r for the individual associated classes. Please note that the percentage change of the absolute neuronal activation is determined for each class individually and are not normalised across all classes. The plots show the means across all raw images of a particular class. The respective standard deviations are indicated through the coloured filling. The red plots represent the previously introduced reference. The graphs of all methods show a fairly logarithmic behaviour. All graphs show a rather steep increase up to 1–2% of removed pixels. After that, the graph levels off substantially. For the majority of classes, a nearly constant level of variation in neuronal activation ΔA_r is obtained for n_r in the range of 2% to 10%.

With respect to the *DeepSHAP* results, modifying the input image of the *foreign body* class results in the lowest variation in neuronal activation of $\Delta A_r = 40\%$. In the other cases $\Delta A_r > 150\%$ for $n_r \geq 8\%$. The *Smooth IG* outcomes show a similar behaviour with a significantly reduced final value for *foreign bodies* of $\Delta A_r = 22\%$. For the remaining classes, the final value is about $\Delta A_r = 100\%$ for $n_r \geq 8\%$. The *Guided Grad-CAM* approach yields smaller ΔA_r scores for $n_r \geq 8\%$ in comparison to the reference plot that describes the result for the randomly erased pixels in the input image. For *wrinkles*, *twists* and *overlaps*, ΔA_r for $n_r \geq 8\%$ are quite identical for *Guided Grad-CAM* as well as *Smooth IG*. However, a sharp rise in ΔA_r of the *Guided Grad-CAM* values up to $n_r = 1\%$ points to a sound recognition of the most relevant pixels across most classes apart from *foreign bodies*.

Throughout the three examined xAI methods, *foreign bodies* are striking. Therefore, the ΔA_r scores above $n_r \geq 8\%$ are considerably smaller compared to the remaining classes. The standard deviations of the ΔA_r curves are also significantly lower. Probably, the relatively small amount of input data can cause a problem regarding the reliability of such results.

Using the mostly nearly constant ΔA_r scores for $n_r \geq 8\%$ we are able to assess the efficiency of a certain xAI procedure. Accordingly, the *DeepSHAP* algorithm seems to be able to identify important pixels for a CNN decision in a meaningful way. For *Guided Grad-CAM* this is mostly different. For this method, the ΔA_r values for $n_r \geq 8\%$ are often even below the reference curve for randomly removed pixels.

The gradient of the initial steep curve rise can also be an indicator for the quality of an xAI evaluation. When a large increase of ΔA_r is already evident for low n_r, we can assume that the most important pixels for a classification decision of the CNN are contained in these few pixels. Therefore, again, a high precision of the *DeepSHAP* xAI

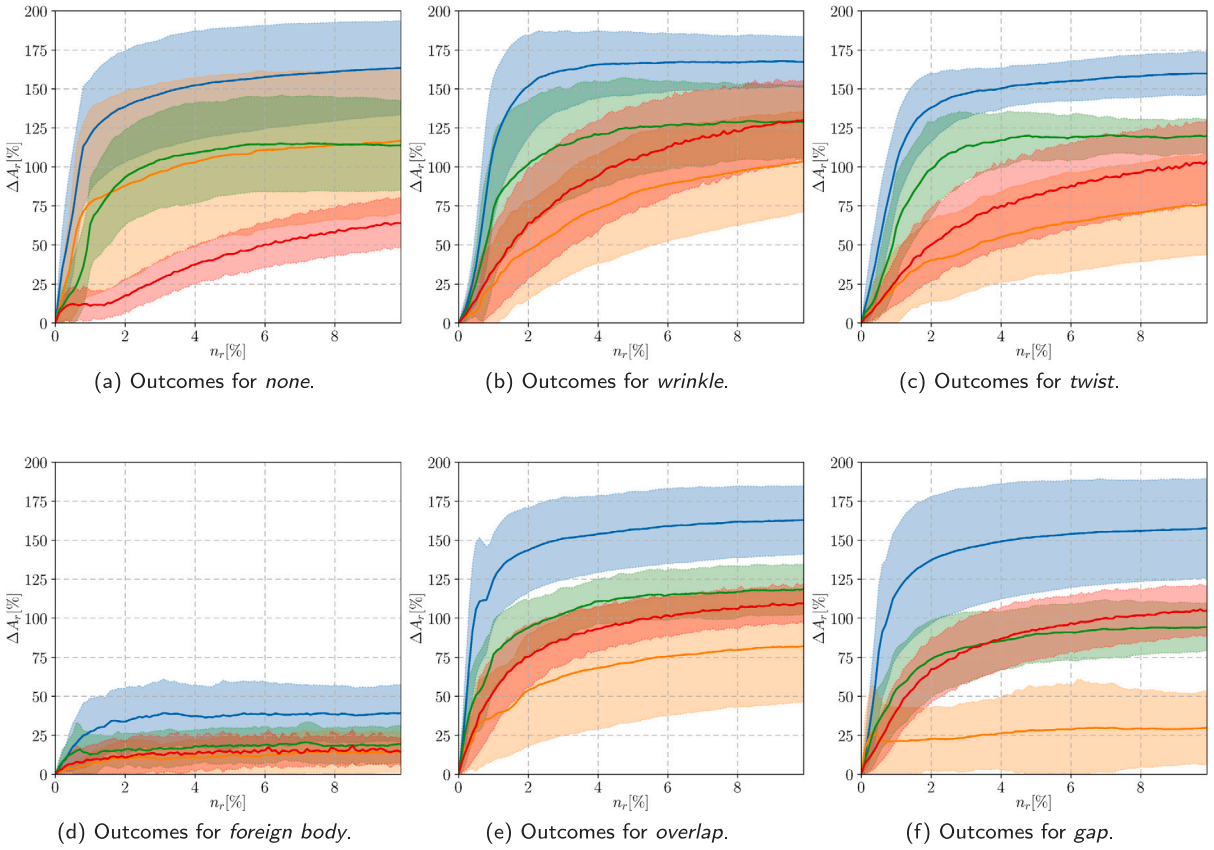


Fig. 6. The change in neural activations ΔA_r with respect to an original, unmodified image are plotted on the ordinate over the percentage of removed pixels n_r on the abscissa. Each class is visualised separately. The methods *DeepSHAP* (blue), *Smooth IG* (green), *Guided Grad-CAM* (orange) as well as the introduced reference (red) are plotted. The associated standard deviations are presented as coloured tubes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

estimation of the important pixels is noticeable. Thus, the order of importance of individual pixels corresponds to the respective relevance of these pixels for the machine decision of the CNN. For *Smooth IG*, a similarly large gradient of ΔA_r up to $n_r = 1\%$ is apparent. For *Guided Grad-CAM* the smallest rise of ΔA_r up to $n_r = 1\%$ can be seen. Hence, we assume a lower precision or a different order of the pixels recognised as important in comparison to the actually relevant pixels for the decision of the CNN.

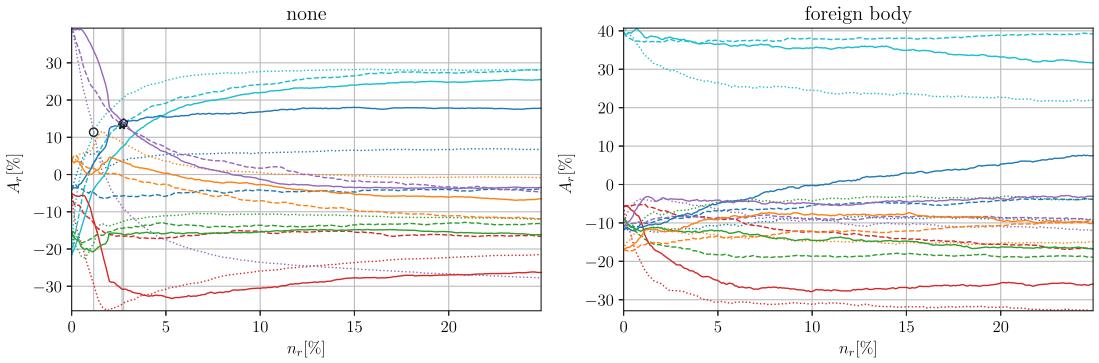
Finally, we notice that all the xAI methods have relatively high ΔA_r values for the *none* defect class for $n_r \geq 2\%$. In particular, *Guided Grad-CAM* shows significantly improved performance for this class compared to the other classes.

The large standard deviations in these investigations are due to the consideration of the absolute change in neuronal activation per output neuron. The values can vary considerably since the classification decision is based on the maximum neuronal activation compared to the remaining output class neurons. However, looking at these absolute change provides an indication on the degree of variation in the input images and the respective response of the CNN. We can see in our analysis that the used test data set is relatively diverse for the individual defect classes and thus triggers correspondingly different changes in neural activation. The findings from this analysis can therefore be judged to be robust in principle. In summary, we recognise that especially for the xAI methods *DeepSHAP* and *Smooth IG* a correlation between the examined CNN and the calculated importance of individual pixels is evident. In the following section, the changes in neural activation are examined in more detail for differently modified input images.

4.4. Neural activations analysis results for explicitly destructive manipulations

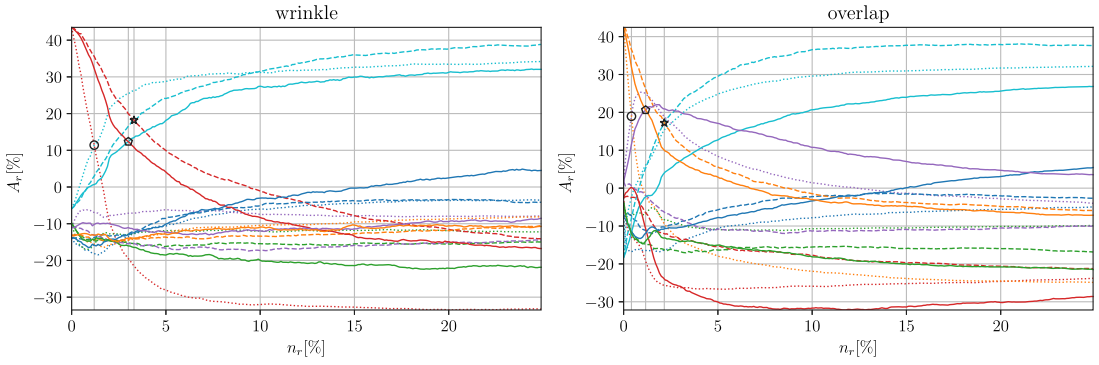
In this section, the neural activation for the considered defect classes is investigated when the image is deliberately manipulated in a destructive manner. The experimental results are shown for each defect type separately according to the three different xAI methods in Fig. 7. For almost all classes, a decrease in relative neural activation with increasing percentage of modified pixels is evident for the considered xAI methods. The critical proportion of modified pixels $n_{r,c}$, until the considered class no longer yields the strongest neural activation and is thus correctly classified, varies between $n_{r,c} = 0.42\%$ and $n_{r,c} = 4.6\%$. However, this differs with the examined classes and the applied xAI method. An exception are *foreign bodies*. For this class, the relative neural activation is $A_r > 20\%$, which is always greater than the neural activations of the competing classes. This behaviour seems to be almost independent of the ratio of modified pixels.

Furthermore, we notice a steep drop in the curves of the *DeepSHAP* (⋯) explanation over all classes for small n_r. The *Guided Grad-CAM* (- - -) and *Smooth IG* (—) plots, on the other hand, fall off less quickly. Moreover, the given xAI results for the neural activation of the *foreign body* class increases comparatively steeply for small n_r across all investigated classes except for *foreign bodies* themselves. This leads to the relatively small $n_{r,c}$ values mentioned above. However, this also implies that with an increasing number of manipulated pixels or noisy input data, the input images are more likely to be classified as *foreign bodies*, independent of their actual, correct defect class. This is basically plausible, since the *foreign bodies* used in this study influence the measurement image more through their reflection behaviour than through geometric changes in the image. In this case, a foil as a *foreign body* is therefore described in terms of a change in the signal quality



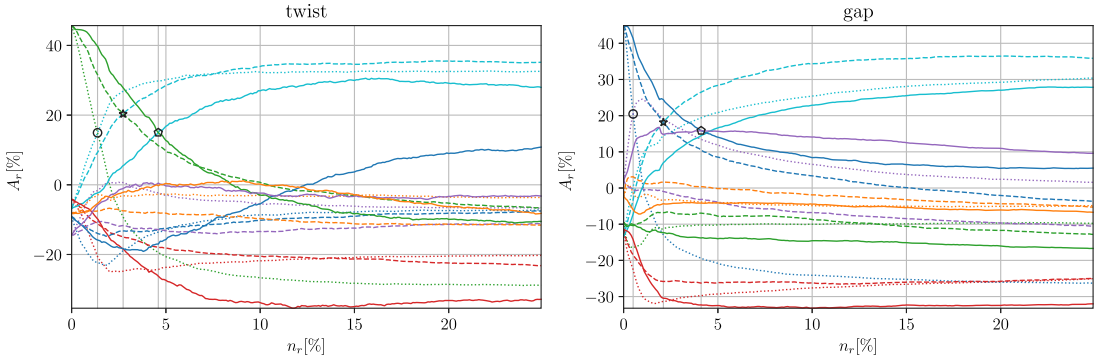
(a) Relative neural activation for class *none* (purple - none): ∘ : $n_{r,c,\text{DeepSHAP}} = 1.17$ %, ★ : $n_{r,c,\text{Guided Grad-CAM}} = 2.67$ %, ⬠ : $n_{r,c,\text{Smooth IG}} = 2.76$ %

(b) Relative neural activation for class *foreign body* (light blue - foreign body)



(c) Relative neural activation for class *wrinkle* (red - wrinkle): ∘ : $n_{r,c,\text{DeepSHAP}} = 1.2$ %, ★ : $n_{r,c,\text{Guided Grad-CAM}} = 3.31$ %, ⬠ : $n_{r,c,\text{Smooth IG}} = 3.01$ %

(d) Relative neural activation for class *overlap* (orange - overlap): ∘ : $n_{r,c,\text{DeepSHAP}} = 0.42$ %, ★ : $n_{r,c,\text{Guided Grad-CAM}} = 2.16$ %, ⬠ : $n_{r,c,\text{Smooth IG}} = 1.17$ %



(e) Relative neural activation for class *twist* (green - twist): ∘ : $n_{r,c,\text{DeepSHAP}} = 1.37$ %, ★ : $n_{r,c,\text{Guided Grad-CAM}} = 2.73$ %, ⬠ : $n_{r,c,\text{Smooth IG}} = 4.6$ %

(f) Relative neural activation for class *gap* (dark blue - gap): ∘ : $n_{r,c,\text{DeepSHAP}} = 0.5$ %, ★ : $n_{r,c,\text{Guided Grad-CAM}} = 2.11$ %, ⬠ : $n_{r,c,\text{Smooth IG}} = 4.11$ %

Fig. 7. The relative activations A_r of the output neurons are plotted over the number of modified pixels n_r. Each class is visualised separately. The percentage n_r of modified pixels is displayed on the abscissa. The ordinate shows the corresponding relative neural activation A_r in %. The results for the xAI techniques *DeepSHAP* (···), *Smooth IG* (—) and *Guided Grad-CAM* (– – –) are plotted. Legend: purple - none; red - wrinkle; green - twist; light blue - foreign body; orange - overlap; dark blue - gap | ∘ : $n_{r,c,\text{DeepSHAP}}$; ★ : $n_{r,c,\text{Guided Grad-CAM}}$; ⬠ : $n_{r,c,\text{Smooth IG}}$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

rather than a real geometric variation. Obviously, this behaviour can be very beneficial for the considered application, since a category for uncertain classification results might be created in this way. Training this class can also be supported with artificially modified images.

For the classes *gaps* and *overlaps*, we also observe a very strong relative initial activation of neurons of the class *none* up to about $n_r = 4\%$. Afterwards, the relative neural activation for the class *none* slowly drops again. This behaviour clearly explains the difficulty in distinguishing between *none*, *gaps* and *overlaps*. This trend is also inversely evident when looking at the class *none*, where the incorrect activation of the corresponding *gap* and *overlap* neurons increases for small n_r. Obviously, this curve for none also indicates that above a certain degree of deviation of the input image from the training data set, the class none is no longer correctly recognised as flawless. In the following section, the sensitivity and infidelity of the individual xAI methods are examined with respect to the individual defect classes.

Table 5
The table presents the simulated SenseMAX results over 50 Monte-Carlo runs. The arithmetic mean values over all input images per class as well as the respective total mean value with the associated error ranges are listed.

SensMAX	DeepSHAP	Guided Grad-CAM	Smooth Int. Grad.
None	0.289 ± 0.031	0.813 ± 0.118	0.358 ± 0.017
Wrinkle	1.641 ± 0.056	0.786 ± 0.093	0.367 ± 0.015
Twist	1.602 ± 0.121	0.694 ± 0.145	0.326 ± 0.009
Foreign body	1.42 ± 0.061	0.453 ± 0.97	0.358 ± 0.022
Overlap	1.453 ± 0.124	0.727 ± 0.123	0.35 ± 0.01
Gap	1.517 ± 0.257	0.698 ± 0.075	0.341 ± 0.014
Total	1.298 ± 0.51	0.72 ± 0.13	0.348 ± 0.018

Table 6
The table presents the simulated INFD results over 1000 Monte-Carlo runs. The arithmetic mean values over all input images per class as well as the respective total mean value with the associated error ranges are presented.

INFD	DeepSHAP	Guided Grad-CAM	Smooth Int. Grad.
None	1.725 ± 2.257	0.673 ± 0.683	0.451 ± 0.013
Wrinkle	46.209 ± 124.5	0.501 ± 0.205	0.662 ± 0.987
Twist	11.197 ± 30.077	7.189 ± 41.498	1.478 ± 4.013
Foreign body	2.395 ± 2.622	0.495 ± 0.01	0.495 ± 0.024
Overlap	13.996 ± 26.724	4.864 ± 31.691	1.305 ± 7.341
Gap	14.366 ± 37.868	1.16 ± 1.651	0.826 ± 0.816
Total	14.381 ± 50.091	2.447 ± 20.391	0.894 ± 3.609

4.5. Maximum sensitivity and infidelity results for xAI calculations

In this section, the sensitivity and infidelity of the three considered xAI methods *DeepSHAP*, *Smooth IG* and *Guided Grad-CAM* are investigated for each of the six examined defect classes. At first, the findings from the SenseMAX analysis are given in Table 5. The mean and error ranges from 50 random Monte-Carlos calculations are displayed for each class. We see that *Smooth IG* with a mean value of 0.348 has the lowest sensitivity with a low and constant mean error of 0.018 to infinitesimally small changes in the input image. This results from the smoothing operation that the algorithm inherently performs. The *Smooth IG* method achieves quite similar scores across all considered classes in this experiment.

Furthermore, a greater mean sensitivity of 0.72 with a mean error of 0.13 for the *Guided Grad-CAM* is apparent. Gradient based explainability methods tend to generate noisy outcomes. Accordingly, these larger SenseMAX scores are attributed to the weighting of the activations with gradients as well as the additional multiplication with the GBP method, as described in Section 2.4.4. Rather conspicuous are the *Guided Grad-CAM* scores for the class *foreign bodies*. These values are rather low, which indicates that the *Guided Grad-CAM* procedure is less sensitive for input images of this class. Due to the calculation rule of *Guided Grad-CAM*, this small SenseMAX value can potentially be attributed to the robustness of the CNN. However, the applied CNN shows robust classification characteristics in the previously conducted experiments for the *foreign body* class. Thus, the low *Guided Grad-CAM* value for the *foreign body* class can also result from other origins.

The *DeepSHAP* yields the largest overall SenseMAX mean value of 1.298 across the three investigated xAI methods. Noticeable is the small mean value of 0.289 for the class *none*. In addition, the large error ranges for the classes *twist*, *overlap* and *gap* are also worth noting. Hence, the sensitivity of the *DeepSHAP* method strongly depends on the respective class. Similar to the *Guided Grad-CAM* technique, the reasons for these high sensitivity values are only known to a limited extend.

In Table 6 the INFD scores for the various xAI techniques and considered defect classes are presented. Noteworthy is the large error range associated to all mean values of the three xAI methods. For the *foreign body* class, the INFD scores are lowest across all three xAI methods. The INFD calculation for the *DeepSHAP* method for the class *wrinkle* yields the globally largest values. The *Guided Grad-CAM* and *Smooth IG* algorithms, on the other hand, achieve very low overall results. Consequently, a strong dependency between the outcomes of the xAI methods and the considered defect class is evident. This in turn corresponds to the previously discussed findings from Fig. 7.

The *Smooth IG* procedure achieves the lowest INFD values, which is very similar to the behaviour observed for the SenseMAX calculations described above. The large mean error of 3.609 is mainly due to the great errors of the classes *twist* and *overlap*. When these classes are excluded, the total mean error ranges from $[0.013, 0.987]$. The *Smooth IG* method thus has robustness issues when representing these defect types in the CNN model. Consequently, according to the infidelity metric, the determined INFD scores are most closely related to the response and neural activation of the trained CNN model. Nevertheless, also for *Smooth IG* partially large error ranges arise.

The *Guided Grad-CAM* method achieves an INFD mean value of 2.447 with an average error of 20.391. Noticeably, this mean value is especially lifted through the large INFD scores of the classes *gap*, *twist* and *overlap*. As already seen for the *Smooth IG* method, the mean overall error is raised through the large errors of *twists* and *gaps*, which again constitutes a difficulty in representing the CNN behaviour for these classes. The mean values of the other classes are similar to the INFD values of the *Smooth IG* technique.

The *DeepSHAP* algorithm generates the xAI explanations with the largest deviation of the CNN model from the calculated xAI explanation, according to the INFD criterion. In addition, the large error values indicate a weak robustness of the informativeness of the associated INFD values. Thus, the xAI explanations are less faithful to the CNN behaviour and from these results, this method is not well suited for the application under consideration. In the following section, the results of this study are discussed in context of the related research.

5. Discussion

Initially, we can state that the applied CNN classification model from the previous research of Meister et al. [7] is well suited for the classification of the fibre layup defects in this paper, having an average classification accuracy of 99.29%. Moreover, the above chosen methods *Smooth IG* [37,43], *DeepSHAP* [36,38] as well as *Guided Grad-CAM* [46] for visualising image regions which are certainly most relevant for the decision of a CNN classifier enable a sound illustration of its behaviour. Regarding *Smooth IG*, this contradicts the findings of Lee et al. [11] for whom *Smooth IG* performed less well for their examined use case in TFT panel inspection. Nevertheless, the various functional principles behind a certain xAI approach have to be kept in mind. These can have a major affect on the outcomes in certain situations. Evaluating the neuronal activations of a CNN for a given input image reveals a substantial knowledge on the intrinsic response of the CNN as well as possible uncertainties in the classification procedure. In this regard, Srinivas and Fleuret [48] presented two different ways for evaluating the class dependent neural activations. The first procedure removes the most important pixels in order. This approach is particularly well suited for assessing the precision of the xAI techniques with respect to their importance evaluation of certain image areas. In the second approach the important pixels are manipulated with respect to their determined importance magnitude and sign. This procedure is additionally well suited for assessing the robustness of the CNN classifier to noisy input data. The SenseMAX and INFD metrics proposed from Yeh et al. [39] provide a simple way to quantitatively assess the precision of a CNN classifier. Particularly striking is, that the *DeepSHAP* method as well as the *twist* and *overlap* classes yield xAI explanations with large and strongly varying deviations of the CNN model to the calculated xAI explanations, according to the INFD criterion. Thus, the xAI explanations are less faithful to the CNN behaviour. This contradicts both the findings of Yeh et al. [39] as well as the results from the previously described experiments in this study. In this paper,

a *noisy-baseline reference* is applied, although Yeh et al. [39] used the *square removal baseline reference*, which might be an explanation for such deviations.

In order to answer the first research question we conclude in particular for the considered application case that the xAI approaches *Smooth IG* as well as *DeepSHAP* are highly appropriate for highlighting essential image areas with respect to the decision-making process within the CNN. However, from these two well-suited approaches, the *Smooth IG* method is to be highlighted as particularly suitable for the explainability of the ANN decision-making for the application case described in this study, since it yields excellent and less fluctuating SenseMAX and INFD scores. Besides, the *Smooth IG*'s visual explanation images visualise the most important image areas for the classification decision in a very pronounced way. Regarding the second research question, the selected SenseMAX metric is particularly well suited for the straightforward evaluation of xAI outcomes. However, the INFD results indicate noticeable limitations for assessing the model fidelity of the *DeepSHAP* method as well as the *twist* and *gap* defect types.

Referring to the study of Sacco et al. [29], we have to mention that the findings of our study can be transferred to such systems, when the used xAI method is adapted to the classifier. They also describe an idea for the visualisation of the prediction accuracy for fibre layup defects in their paper. The pixel brightness or gradient length and sign from our previous visualisation in Section 4.2 could be integrated into the visualisation of Sacco et al. and provide an additional detailed indication of the classification quality of a defect.

In future research, it will be interesting to analyse the neuronal activation changes for a corresponding manipulated input image across all output neurons in conjunction and derive insights about the value intervals of the resulting activation patterns. This might contribute to estimate the certainty of a classification decision for a given defect type. Moreover, the data acquisition and processing system used for the experimental investigation in this paper should be further elaborated for inline inspection in fibre composite manufacturing. This requires a suitable integration of the LLSS hardware as well as the linking of the image-based inspection data with the position information of the fibre layup machine. Furthermore, an adapted training of the existing CNN model with real world training data is advisable in order to reduce the false classifications or assignments to a potential "ambiguous" class. Finally, the subsequent section summarises the key results of this study and highlights the added value for the community.

6. Conclusion

The results of this study demonstrated that the relevance of certain image pixels regarding the decision-making response of a *Convolutional Neural Network* classifier can be displayed and evaluated. Based on the findings from this paper, the *Explainable Artificial Intelligence* techniques *Smooth Integrated Gradients* and *Deep Learning Important Features with Shapley Additive Explanations* are especially suitable in this context. The great novelty of this study results from the detailed analysis of the neuronal activations for differently modified data sets in order to estimate the response of an *Artificial Neural Network* and its behaviour in case of modified input images. Furthermore, the investigations in this study have shown that the metrics *Maximum Sensitivity* and restrictively also *Infidelity* are appropriate for the straightforward evaluation of the performance of the considered *Explainable Artificial Intelligence* methods.

The outcomes of this investigations provide valuable guidance for engineers of camera-based monitoring devices for the composites sector with respect to the conception and implementation of sophisticated but trustworthy machine-learning solutions. Moreover, the given results offer support for corresponding certification processes for similar machine-learning approaches.

CRediT authorship contribution statement

Sebastian Meister: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft. **Mahdieu Wermes:** Methodology, Software, Formal analysis, Data curation, Visualization. **Jan Stüve:** Resources, Writing – review & editing, Supervision, Funding acquisition. **Roger M. Groves:** Conceptualization, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

All authors have approved the final manuscript.

Funding

This research is part of the project DHiiP-AIR and was financially supported by the Federal Ministry for Economic Affairs and Energy, Germany. This project has received funding from the Federal Ministry for Economic Affairs and Energy, Germany under the funding code No. 20W1911F.

References

[1] Marsh G. Airbus A350 XWB update. Reinf Plast 2010;54(6):20–4. http://dx.doi.org/10.1016/s0034-3617(10)70212-5.

[2] McIlhagger A, Archer E, McIlhagger R. Manufacturing processes for composite materials and components for aerospace applications. In: Irving P, Soutis C, editors. Polymer composites in the aerospace industry. Elsevier; 2020, p. 59–81. http://dx.doi.org/10.1016/b978-0-08-102679-3.00003-4.

[3] Rudberg T, Nielson J, Henscheid M, Cemenska J. Improving AFP cell performance. SAE Int J Aerosp 2014;7(2):317–21. http://dx.doi.org/10.4271/2014-01-2272.

[4] Eitzinger C. Inline inspection helps accelerate production by up to 50 %. Lightweight Des Worldw 2019.

[5] European Union Aviation Safety Agency. Human intelligence roadmap - A human-centric approachto AI in aviation. techreport Vers. 1.0, European Union Aviation Safety Agency; 2020, URL https://www.easa.europa.eu/ai. vers. 1.0.

[6] EASA AI Task Force, Daedalean AG. Concepts of design assurancefor neural networks (CoDANN). techreport Vers. 1.0, European Union Aviation Safety Agency and Daedalean AG; 2020, URL https://www.easa.europa.eu/document-library/general-publications/concepts-design-assurance-neural-networks-codann. vers. 1.0.

[7] Meister S, Möller N, Stüve J, Groves RM. Synthetic image data augmentation for fibre layup inspection processes: Techniques to enhance the data set. J Intell Manuf 2021. http://dx.doi.org/10.1007/s10845-021-01738-7.

[8] Schmidt C, Hocke T, Denkena B. Artificial intelligence for non-destructive testing of CFRP prepreg materials. Prod Eng 2019. http://dx.doi.org/10.1007/s11740-019-00913-3.

[9] Liu K, Li Y, Yang J, Liu Y, Yao Y. Generative principal component thermography for enhanced defect detection and analysis. IEEE Trans Instrum Meas 2020;1. http://dx.doi.org/10.1109/tim.2020.2992873.

[10] Liu K, Tang Y, Lou W, Liu Y, Yang J, Yao Y. A thermographic data augmentation and signal separation method for defect detection. Meas Sci Technol 2021;32(4):045401. http://dx.doi.org/10.1088/1361-6501/abc63f.

[11] Lee M, Jeon J, Lee H. Explainable AI for domain experts: A post hoc analysis of deep learning for defect classification of TFT–LCD panels. J Intell Manuf 2021. http://dx.doi.org/10.1007/s10845-021-01758-3.

[12] Parmar H, Khan T, Tucci F, Umer R, Carlone P. Advanced robotics and additive manufacturing of composites: Towards a new era in industry 4.0. Mater Manuf Process 2021;1–35. http://dx.doi.org/10.1080/10426914.2020.1866195.

[13] Cemenska J, Rudberg T, Henscheid M. Automated in-process inspection system for AFP machines. SAE Int J Aerosp 2015;8(2):303–9. http://dx.doi.org/10.4271/2015-01-2608.

[14] Weimer C, Friedberger A, Helwig A, Heckner S, Buchmann C, Engel F. Increasing the productivity of CFRP production processes by robustness and reliability enhancement. In: CAMX 2016 - The composites and advanced materials expo and conference. Airbus Group Innovations, 81663 Munich, Germany; AirbusInfactory Solutions GmbH, 81663 Munich, Germany; 2016, URL https://www.researchgate.net/profile/Christian_Weimer/publication/308778487_INCREASING_THE_PRODUCTIVITY_OF_CFRP_PRODUCTION_PROCESSES_BY_ROBUSTNESS_AND_RELIABILITY_ENHANCEMENT/links/57efa78208ae886b8975147a.pdf.

[15] Black S. Improving composites processing with automated inspection. CompositesWorld, 2018, URL https://www.compositesworld.com/articles/improving-composites-processing-with-automated-inspection.

[16] Meister S, Wermes MAM, Stueve J, Groves RM. Algorithm assessment for layup defect segmentation from laser line scan sensor based image data. In: Zonta D, Huang H, editors. Sensors and smart structures technologies for civil, mechanical, and aerospace systems 2020. SPIE; 2020, http://dx.doi.org/10.1117/12.2558434.

[17] Campbell F. Manufacturing processes for advanced composites. Elsevier Science & Technology; 2004, URL https://www.ebook.de/de/product/6827737/manufacturing_processes_for_advanced_composites.html.

[18] Oromiehie E, Prusty BG, Compston P, Rajan G. Automated fibre placement based composite structures: Review on the defects, impacts and inspections techniques. Compos Struct 2019;224:110987. http://dx.doi.org/10.1016/j.compstruct.2019.110987.

[19] Lengsfeld H, Fabris FW, Krämer J, Lacalle J, Altstädt V. Faserverbundwerkstoffe. Hanser Fachbuchverlag; 2014, URL https://www.ebook.de/de/product/22746074/hauke_lengsfeld_felipe_wolff_fabris_johannes_kraemer_javier_lacalle_volker_altstaedt_faserverbundwerkstoffe.html.

[20] Rudberg T. Webinar: Building AFP system to yield extreme availability. CompositesWorld, 2019, video.

[21] Harik R, Saidy C, J. Williams S, Gurdal Z, Grimsley B. Automated fiber placement defect identity cards: Cause, anticipation, existence, significance, and progression. In: SAMPE 18. 2018, URL https://www.researchgate.net/publication/326464139_Automated_fiber_placement_defect_identity_cards_cause_anticipation_existence_significance_and_progression.

[22] Potter K. Understanding the origins of defects and variability in composites manufacture. In: ICCM international conferences on composite materials. 2009, URL http://iccm-central.org/Proceedings/ICCM17proceedings/Themes/Plenaries/P1.5.

[23] Heinecke F, Willberg C. Manufacturing-induced imperfections in composite parts manufactured via automated fiber placement. J Compos Sci 2019;3(2):56. http://dx.doi.org/10.3390/jcs3020056.

[24] Miesen N, Sinke J, Groves RM, Benedictus R. Simulation and detection of flaws in pre-cured CFRP using laser displacement sensing. Int J Adv Manuf Technol 2015;82(1–4):341–9. http://dx.doi.org/10.1007/s00170-015-7305-x.

[25] Gardiner G. Zero-defect manufacturing of composite parts. CompositesWorld, 2018, URL https://www.compositesworld.com/blog/post/zero-defect-manufacturing-of-composite-parts. [Accessed 18 June 2019].

[26] Black S. Improving composites processing with automated inspection, Part II. CompositesWorld, 2018, URL https://www.compositesworld.com/articles/improving-composites-processing-with-automated-inspection-part-ii. [Accessed 19 June 2019].

[27] Schmitt R, Niggemann C, Mersmann C. Contour scanning of textile preforms using a light-section sensor for the automated manufacturing of fibre-reinforced plastics. In: Berghmans F, Mignani AG, Cutolo A, Meyrueis PP, Pearsall TP, editors. Optical sensors 2008, Vol. 7003. SPIE; 2008, p. 436–47. http://dx.doi.org/10.1117/12.779005.

[28] Schmitt R, Orth A, Niggemann C. A method for edge detection of textile preforms using a light-section sensor for the automated manufacturing of fibre-reinforced plastics. In: Osten W, Gorecki C, Novak EL, editors. Optical measurement systems for industrial inspection V. SPIE; 2007, http://dx.doi.org/10.1117/12.726177.

[29] Sacco C, Radwan AB, Anderson A, Harik R, Gregory E. Machine learning in composites manufacturing: A case study of automated fiber placement inspection. Compos Struct 2020;250:112514. http://dx.doi.org/10.1016/j.compstruct.2020.112514.

[30] Blake S. SMART factory applications for integrated laserprojection and automatic inspection. In: CAMX confert proceedings. 2017.

[31] Blake S. Elements and mechanisms for applying artificial intelligence to composites fabrication. In: SAMPE 2019 - Charlotte, NC. SAMPE; 2019, http://dx.doi.org/10.33599/nasampe/s.19.1435.

[32] Khan A, Sohail A, Zahoora U, Qureshi AS. A survey of the recent architectures of deep convolutional neural networks. Artif Intell Rev 2020. http://dx.doi.org/10.1007/s10462-020-09825-6.

[33] Vasilev I, Slater D, Spacagna G. Python deep learning. 2nd ed.. Packt Publishing; 2019, URL https://www.ebook.de/de/product/35342345/ivan_vasilev_daniel_slater_gianmario_spacagna_python_deep_learning_second_edition.html.

[34] Chen M, Jiang M, Liu X, Wu B. Intelligent inspection system based on infrared vision for automated fiber placement. In: 2018 IEEE international conference on mechatronics and automation. IEEE; 2018, http://dx.doi.org/10.1109/icma.2018.8484646.

[35] Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One 2015;10(7).

[36] Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: Precup D, Teh YW, editors. Proceedings of the 34th international conference on machine learning. Proceedings of machine learning research, vol. 70, International Convention Centre, Sydney, Australia: PMLR; 2017, p. 3145–53, URL arXiv:1704.02685v2.

[37] Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: Proceedings of the 34 th international conference on machine learning. 2017.

[38] Lundberg S, Lee S-I. A unified approach to interpreting model predictions. In: Advances in neural information processing systems, vol. 30, 2017.

[39] Yeh C-K, Hsieh C-Y, Suggala AS. On the (in)fidelity and sensitivity of explanations. In: NeurIPS.

[40] Müller K-R. Understanding ML models. Technical University of Berlin; 2019, URL http://helper.ipam.ucla.edu/publications/mlpws3/mlpws3_15932.pdf.

[41] Samek W, Montavon G, Vedaldi A, Hansen LK, Müller K-R, editors. Explainable ai: Interpreting, explaining and visualizing deep learning. Springer International Publishing; 2019, http://dx.doi.org/10.1007/978-3-030-28954-6, URL https://www.springer.com/de/book/9783030289539.

[42] Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): Towards medical XAI. IEEE Trans Neural Netw Learn Syst 2020;1–21. http://dx.doi.org/10.1109/TNNLS.2020.3027314, URL arXiv:1907.07374.

[43] Smilkov D, Thorat N, Kim B, Viegas F, Wattenberg M. SmoothGrad: Removing noise by adding noise. 2017, URL https://arxiv.org/abs/1706.03825.

[44] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: European conference on computer vision. 2014.

[45] Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for simpli: The all confolutional net. In: ICLR 2015. 2015.

[46] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localizatio. In: Proceedings of the IEEE international conference on computer vision. 2017.

[47] Zhang J, Lin Z, Brandt J, Shen X, Sclaroff S. Top-down neural attention by excitation backprop. CoRR 2016;abs/1608.00507.

[48] Srinivas S, Fleuret F. Full-Gradient representation for neural network visualization. In: 2019 conference on neural information processing systems. 2019.

[49] Montavon G, Binder A, Lapuschkin S, Samek W, Müller K-R. Layer-wise relevance propagation: An overview. In: Explainable AI: Interpreting, explaining and visualizing deep learning. Springer International Publishing; 2019, p. 193–209. http://dx.doi.org/10.1007/978-3-030-28954-6_10.

[50] Grezmak J, Zhang J, Wang P, Loparo KA, Gao RX. Interpretable convolutional neural network through layer-wise relevance propagation for machine fault diagnosis. IEEE Sens J 2020;20(6):3172–81. http://dx.doi.org/10.1109/jsen.2019.2958787.

[51] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualisingimage classification models and saliency maps. CoRR 2013;abs/1312.6034. URL http://dblp.uni-trier.de/db/journals/corr/corr1312.html#SimonyanVZ13.

[52] Zhang Q, Wu YN, Zhu S-C. Interpretable convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[53] Kanehira A, Harada T. Learning to explain with complemental examples. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2019.

[54] Ribeiro MT, Singh S, Guestrin C. Model-agnostic interpretability of machine learning. In: ICML workshop on human interpretability in machine learning. 2016.

[55] Fong RC, Vedaldi A. Interpretable explanations of black boxes by meaningful perturbation. In: Proceedings of the IEEE international conference on computer vision. 2017, URL https://openaccess.thecvf.com/content_iccv_2017/html/Fong_Interpretable_Explanations_of_ICCV_2017_paper.html.

[56] Fong R, Patrick M, Vedaldi A. Understanding deep networks via extremal perturbations and smooth masks. In: Proceedings of the IEEE/CVF international conference on computer vision. 2019, URL https://openaccess.thecvf.com/content_ICCV_2019/html/Fong_Understanding_Deep_Networks_via_Extremal_Perturbations_and_Smooth_Masks_ICCV_2019_paper.html.

[57] Kindermans P-J, Schütt KT, Alber M, Müller K-R, Erhan D, Kim B, et al. Learning how to explain neural networks: PatternNet and PatternAttribution. 2017, URL arXiv:1705.05598.

[58] Lipovetsky S, Conklin M. Analysis of regression in game theory approach. Appl Stoch Models Bus Ind 2001;17(4):319–30. http://dx.doi.org/10.1002/asmb.446.

[59] Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. Knowl Inf Syst 2014;41(3):647–65. http://dx.doi.org/10.1007/s10115-013-0679-x.

[60] Tzutalin. LabelImg. 2015, URL https://github.com/tzutalin/labelImg.

[61] Automation Technology GmbH. C5 Series - User manual for high speed 3D sensors. techreport 1.2, 1st ed.. Hermann-Bössow-Straße 6-8, 23843 Bad Oldesloe, Germany: Automation Technology GmbH; 2019, URL https://www.automationtechnology.de/cms/wp-content/uploads/2019/03/C5-Series_specifications_web.pdf. Rev 1.2.

[62] ams AG. Datasheet DS000603 - CMV12000 - CMOS image sensor. techreport 3.0, Tobelbader Strasse 30, 8141 Premstaetten, Austria: ams AG; 2020, datasheet DS000603 v3-00. URL https://ams.com/documents/20143/36005/CMV12000_DS000603_3-00.pdf/d27f4643-e11b-86f9-4e09-ec055cb4c8e1.
[63] Automation Technology GmbH. The FIR Filter. techreport 1.0, Hermann-Bössow-Straße 6-8, 23843 Bad Oldesloe, Germany: Automation Technology GmbH; 2014, URL https://www.stemmer-imaging.com/media/uploads/cameras/12/122195-Automation_Technology_AppNote_FIR_Filter.pdf. Rev. 1.0.
[64] Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature 2020;585(7825):357–62. http://dx.doi.org/10.1038/s41586-020-2649-2.
[65] Bradski G. The OpenCV library. Dr Dobb's J Softw Tools 2000. URL https://opencv.org/.
[66] Hunter JD. Matplotlib: A 2D graphics environment. Comput Sci Eng 2007;9(3):90–5. http://dx.doi.org/10.1109/MCSE.2007.55.
[67] Chollet F, et al. Keras. 2015, URL https://keras.io.
[68] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-scale machine learning on heterogeneous systems. 2015, Software available from tensorflow.org. URL https://www.tensorflow.org/.
[69] Rossum G. Python reference manual. Tech. rep., Amsterdam, The Netherlands, The Netherlands: CWI (Centre for Mathematics and Computer Science); 1995.

Sebastian Meister is a PhD Candidate in optical inspection at the Delft University of Technology, The Netherlands, at the Aerospace Non-Destructive Testing Laboratory. He is also working as a Researcher at the German Aerospace Center, Center for Lightweight-Production-Technology in Stade, Germany. In 2017 he received his Master's degree in Mechanical Engineering from Friedrich-Alexander-University of Erlangen-Nürnberg. His fields of research is the computer vision and machine learning for the optical inspection in automated composite manufacturing.

Mahdieu Wermes is a Master student at the Ilmenau University of Technology and a working student at the German Aerospace Center, Center for Lightweight-Production-Technology in Stade. In 2021 he received his Bachelor's degree in Electrical Engineering from the TU Ilmenau. As a working student he focuses on machine learning techniques for automated industrial inspection.

Dr. Jan Stüve is Associate Professor in composite manufacturing at the Delft University of Technology, The Netherlands. His PhD was on weave technologies from the RWTH Aachen University, Germany. After he was working as a researcher at RWTH Aachen, he became the managing director of Bergal Erfurter Flechttechnik GmbH. Since 2016, he has been Head of Department, Composite Process Technology at the German Aerospace Center, Center for Lightweight-Production-Technology in Stade.

Dr. Roger M. Groves is Associate Professor in Aerospace NDT/SHM and Heritage Diagnostics at Delft University of Technology, The Netherlands. His Ph.D. is in Optical Instrumentation from Cranfield University (2002) and he was a Senior Scientist at Institute for Applied Optics, University of Stuttgart, before joining TU Delft in 2008 as an Assistant Professor. Dr Groves heads a team of approximately 20 researchers in the Aerospace NDT Laboratory at TU Delft. His research interests are Optical Metrology, Fibre Optic Sensing and Ultrasonic Wave Propagation in Composite Materials. He has approximately 200 journal and conference publications in these topics. In 2020 he was awarded Fellow of SPIE.