NDT&E International 138 (2023) 102885

Contents lists available at ScienceDirect

NDT and E International

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

Learning defects from aircraft NDT data

Navya Prakash [a,*], Dorothea Nieberl [b], Monika Mayer [b], Alfons Schuster [b]

[a] Marine Perception (MAP), German Research Center for Artificial Intelligence (DFKI) GmbH, Oldenburg, Lower Saxony, Germany
[b] Center for Lightweight Production Technology (ZLP), German Aerospace Center (DLR), Augsburg, Bavaria, Germany

ARTICLE INFO

Keywords:
NDT
NDE 4.0
Aircraft production
Quality control
Machine learning
POD

ABSTRACT

Non-destructive evaluation of aircraft production is optimised and digitalised with Industry 4.0. The aircraft structures produced using fibre metal laminate are traditionally inspected using water-coupled ultrasound scans and manually evaluated. This article proposes Machine Learning models to examine the defects in ultrasonic scans of A380 aircraft components. The proposed approach includes embedded image feature extraction methods and classifiers to learn defects in the scan images. The proposed algorithm is evaluated by benchmarking embedded classifiers and further promoted to research with an industry-based certification process. The HoG-Linear SVM classifier has outperformed SURF-Decision Fine Tree in detecting potential defects. The certification process uses the Probability of Detection function, substantiating that the HoG-Linear SVM classifier detects minor defects. The experimental trials prove that the proposed method will be helpful to examiners in the quality control and assurance of aircraft production, thus leading to significant contributions to non-destructive evaluation 4.0.

1. Introduction

Ultrasonic Testing (UT) is a typical Non-destructive Testing (NDT) method for examining the structural components for aircraft production. Manufacturing aircraft made of Fibre Metal Laminates (FML) includes cascaded steps such as placement of aluminium, glass prepreg, adhesive, doublers, stringers, vacuum bagging and curing in an autoclave [1]. Quality Control (QC) is performed first at the layup of the component (without stringers) after curing and the quality assessment is visually evaluated [2]. The manually performed examination of anomalies is very time-consuming. In addition, [3] conducted NDT inspection using a manual UT phased array for Glass Reinforced (GLARE®) FML of A380, it lacked the high capacity of data and additionally an evaluation software. So, Non-Destructive Evaluation (NDE) 4.0 helps streamline processes, increase quality and lower costs in aircraft production with an automated Quality Assurance (QA) [4].

Traditionally, the quality control of FML is performed by an experienced examiner after the final production of an aircraft structure [5]. But, with the implementation of Machine Learning (ML) techniques, defects can be identified instantaneously to help the examiner [6,7]. So, the primary motivation was to develop an automated QA in aircraft production by implementing a Machine Learning algorithm. The quality analysis process in the proposed method consists of pre-analysing the sensor data acquisition to classify the features according to the defects and good qualities. The proposed approach reduces the examiner's workload, expensive repairs and manufacturing waste. The proposed work can be vital in the automated offline-QA to scrutinise FML aircraft production [8] and adapt to other aircraft materials like aluminium, thermoplastic fibre [9] and Carbon Fibre Reinforced Plastic (CFRP) [10]. The proposed research aims: to understand and prepare ultrasonic scans of aircraft FML (raw data) provided by the aircraft industry and pre-process data (convert raw data to images) to gain feasibility for the proposed method. Additionally, implement embedded Machine Learning classifiers with image feature extraction techniques to achieve the best defect detection rate and further interpret industry-based certification process to evaluate this approach.

The remainder of the paper is structured as follows: Section 2 describes the proposed Machine Learning model and its pipeline. Section 3 illustrates the proposed model's data interpretations with experimental results and Section 4 discusses the industry-based inferences to evaluate the proposed approach. Finally, Section 5 summarises the proposed method and explores the scope for further improvements.

2. Learning defects

Machine Learning is a subset of Artificial Intelligence (AI), dealing with data acquired from sensors for learning the data-generating distribution. There are three primary techniques: supervised learning – data needs to be labelled (each data point tagged to belong to a particular class) for training, mainly used for classification (predicts

* Corresponding author.
E-mail address: navya.prakash@dfki.de (N. Prakash).

https://doi.org/10.1016/j.ndteint.2023.102885
Received 6 September 2022; Received in revised form 13 May 2023; Accepted 21 May 2023
Available online 26 May 2023

discrete labels) and regression (predicts a continuous quantity). Next, unsupervised learning – requires no data labels for training; dimensionality reduction and clustering are the two significant methodologies. Following is reinforcement learning – the agent (training) sends an action (a move causing change) to the environment (real or virtual world) and in-return environment sends the state and its reward (evaluation of the action, either positive or negative) for the agent; real-time decisions and gaming models are its prototypes. Additionally, semi-supervised learning is a combination of supervised and unsupervised learning methodologies.

Supervised learning examples [11] include Support Vector Machine (SVM) [12], Decision Trees [13,14], Random Forest (RF) [15], *k*-Nearest Neighbour (*k*-NN) [16], Naïve Bayes [17], Linear Discriminant Analysis (LDA) [18] and Logistic Regression [19]. The Fuzzy C-means (FCM) [20], *k*-means [21] and Principal Component Analysis (PCA) [22] are a few state-of-the-art unsupervised learning techniques. SVM predicts classes based on an optimal hyperplane creating margins to find similar features from each class and classifies them together. Decision Trees predict a class by learning the decision rules from the data features of that class. Random Forest combines the outcome of multiple Decision Trees into a prediction. *k*-Nearest Neighbour predicts using the proximity of *k* nearest data points for classification. Naïve Bayes classifies based on the probability of data points applying Bayes' theorem. Using Fisher's algorithm, LDA finds a linear combination of data features to characterise different classes. Logistic regression finds the probability of an event occurring, such as voted or no vote, based on the data variables. Fuzzy C-means is similar to *k*-means but is a soft clustering where a data point can belong to one or more clusters. *k*-means is a hard clustering that partitions data points into *k* clusters, each belonging to one cluster with the nearest mean value. PCA reduces data dimensionality and increases its interpretability with less information loss.

Deep Learning is a subset of Machine Learning applied to images, videos, text and other data formats. It comprises multi-layer Artificial Neural Networks (ANN) [23]. Deep Neural Network (DNN) has many hidden layers of neural networks to perform classification and regressions. The state-of-the-art neural networks are Radial Basis Function (RBF) [24], Autoencoders (AEC) [25], Multi-Layer Perceptron (MLP) [26], VGG-F [27], Fast R-CNN [28], ResNet v2 [29], Transformer [30].

The Machine Learning algorithm often includes a feature extraction process depending on the input data type to improve its performance [31]. A few state-of-the-art image feature extraction methods are Local Binary Patterns (LBP) [32], Maximally Stable Extremal Regions (MSER) [33], KAZE [34], Speeded Up Robust Features (SURF) [35], Histogram of Oriented Gradients (HoG) [36]. LBP labels pixels in an image by thresholding each pixel neighbourhood, resulting in a binary number to encode local texture information. For blob detection, the MSER method uses co-variant regions in corresponding grey-level cells in images. KAZE works on non-linear scale space and determinants of the Hessian matrix with the local difference binary descriptor to detect multi-scale corner features from the scale space. SURF detects interest points and local neighbourhoods to match, finds features in the Gaussian scale space, can distinguish between background and foreground features in an image, finds blob features and is partially influenced by Scale-Invariant Feature Transform (SIFT) [37]. HoG is a feature descriptor that describes the image features by calculating the frequency of gradients oriented in localised parts of an image; it encodes local shape information.

The previous research methods that use Machine Learning for NDT data defect analysis are as follows, [38] for UT C-mode scanning acoustic microscopy (C-SAM) in integrated circuits using the Mumford–Shah model for grayscale image processing and SVM for defect classification with 80% of recognition rate. This technique needs more training data to improve classification accuracy. The research of [39] implemented crack shape estimation with height, length and depth parameters using Eddy Current and SVM for regression (SVR) with RBF kernel in conductive materials. It achieved a maximum error rate of 0.3 mm in defect length, but height and depth detection needed more training. Following, [40] used an ANN, MLP (with back-propagation) and SVR (with RBF kernel) for crack defect classification. SVR outperformed MLP with a maximum error rate of 0.8 mm on a 5 mm crack length, but height and depth parameters needed more SVR model tuning. Dynamic PCA, *k*-NN, MLP, RBF and SVM were implemented by [41] for defect depth in infrared NDT in CFRP composite material. The MLP outperformed RBF and SVM for complex composite, whereas the dynamic PCA and *k*-NN could estimate defect depth on plane composite and detection limit for classifiers. The NDT data of oil or gas pipeline defects were detected by [42] using LDA, MLP, SVR, RBF, PCA, *k*-NN and SVR outperformed all other methods with 98.28%. SVM and ANN were trained by [43] with NDT rail data for real-time defect processing and SVM outperformed ANN with 97% of accuracy. For fabric defect image analysis, [44] implemented AdaBoost [45] and HoG for feature extraction with SVM for classification. This method identified most defects with fewer false rejection rates. The SVM and ANN classified NDT data of construction structures by [46] with Fast Fourier Transforms (FFT) and RBF for feature extraction, SVM outperformed ANN with 93% accuracy.

Aerospace structure defects were classified based on their shapes (Shape Geometric Descriptor (SGD)) using J48 Decision Tree [47], MLP, Naïve Bayes classifiers with Content-Based Image Retrieval (CBIR) and SGD for feature extraction in the research of [48]. MLP outperformed J48 Decision Tress (96%) and Naïve Bayes (95%) with 98% accuracy. Another research [49] trained J48 Decision Trees and Random Forest to determine weld quality in NDT data of Shielded Metal Arc Welding (SMAW) of carbon steel plates. Random Forest outperformed J48 Decision Trees (70.78%) with 88.69% of accuracy. Automatic NDT aircraft defects were diagnosed by [50] using SVM and SURF with AlexNet [51] and VGG-F Deep Neural Network as feature extraction methods. SVM gained the highest accuracy of 96% with the SURF for Region of Interest (RoI) selection. The mobile panel surface defects were inspected by [52] with LBP and HoG feature extractors that trained Naïve Bayes and SVM. The HoG-SVM classifier outperformed all other feature extractors and Naïve Bayes with >90% average accuracy. Random Forest with RoI classified defects on alloys and achieved >90% accuracy [53].

An Aeronautics Engine Radiographic Testing Inspection System Net (AE-RTISNet) with Fast R-CNN was developed to inspect defects in aeronautical engines [54]. It contains RoI as a feature extractor and obtained a mean average precision (mAP) of 90% compared to YOLO [55]. The Aluminium Conductor Composite Core (ACCC) with NDT X-ray images was analysed for defects using Inception ResNet v2 [56]. This Deep Neural Network, Inception ResNet v2, maintained 97.01% accuracy compared with Res2Net-18 (96.28%) and ResNet-v2-50 (96.15%) after data augmentation. Random Forest, RBF-SVM, hidden Markov model (HMM) [57] were implemented by [58] for training with autoencoders-FFT, low-pass filtering and PCA for feature extractions to measure defects in aerospace CFRP aluminium plates. AEC-PCA outperformed all other classifiers with >0.9 clustering scores. Convolution Neural Networks (CNN) determined aerospace NDT defects using spot classifiers in research of [59] and the Indirect spot CNN classifier outperformed the Direct spot CNN classifier with 98% of accuracy. Another CNN approach [60] was developed to detect defects in NDT data of stainless steel and welded Gas Tungsten Arc Welding (GTAW) or Shielded Metal Arc Welding (SMAW) joints. This CNN resembles VGG-16 and gained a Probability of Detection (POD) of $a_{90/95}$ = 2.1 mm, where a is the defect size and 90/95 denotes 90% POD with 95% of CNN model confidence. An ANN was developed by [61] to monitor defects in NDT data of mechanical, aerospace and civil structures consisting of aluminium and magnesium alloys and inferred >95% of precision.

Table 1
A brief literature survey (ordered by publication year)

Source	NDT data	Feature extraction	Machine learning	Performance analysis
Zhang et al. (2005) [38]	Integrated Circuits	Mumford–Shah model	SVM	Recognition rate: 80%
Bernieri et al. (2006) [39]	Conductive materials	RoI	SVM regression (SVR) with RBF	Maximum error rate (length): 0.3 mm
Bernieri et al. (2008) [40]	Conductive materials	RoI	ANN-MLP (reference) and SVR with RBF	SVR: maximum error rate (length) of 0.8 mm; SVR outperformed MLP
Benítez et al. (2009) [41]	CFRP structure	RoI	Dynamic PCA, *k*-NN, MLP, RBF and SVM	MLP outperformed RBF and SVM
Khodayari-Rostamabad et al. (2009) [42]	Oil, gas pipelines	PCA	*k*-NN, SVR, RBF, LDA, MLP	Accuracy: SVR - 98.28%
Wei & Cheng-Tong (2009) [43]	Rail flaws	RoI	SVM, ANN	Accuracy: SVM - 97%
Shumin et al. (2011) [44]	Fabric	HoG, AdaBoost	SVM	Detection rate: SVM - high, less false rejections
Saechai et al. (2012) [46]	Construction cement structure	FFT, RBF	SVM, ANN	Accuracy: SVM - 93%
D'Angelo & Rampone (2015) [48]	Aerospace structure	SGD, CBIR	J48 Decision Trees, Multilayer Perceptron (MLP) and Naïve Bayes	Accuracy: MLP - 98%,
Sumesh et al. (2015) [49]	SMAW Carbon Steel plates	Statistical approach	J48 Decision Trees, Random Forest	Accuracy: Random Forest - 88.69%
Internal Study: Schmidt, T et al. (2015) [10]	CFRP C-scans	Measured values of all sections, mean or variance, gradient histograms	SVM, Random Forest	AUC: Gradient histogram-SVM - 0.987
Malekzadeh et al. (2017) [50]	Aircraft surface	LBP, RGB and HSV histograms, AlexNet, VGG-F DNN, SURF	SVM	Accuracy: SVM-SURF - 96%
Huang et al. (2017) [52]	Mobilephone Panel	LBP, HoG	Naïve Bayes, SVM	Average accuracy: HoG-SVM - >90%
Internal Study: University of Augsburg [64]	GLARE®-NDT C-scan images	Laplace filter, material thickness, edge information	CNN-ASPP, SGD, softmax	High exclusion rate of manual inspection for component area - 97.36%
Shipway et al. (2019) [53]	Titanium alloy plates	RoI	Decision Trees, Random Forest	Accuracy: >90%
Chen & Juang (2020) [54]	Aeronautical engine	RoI	Fast R-CNN, YOLO	mAP: Fast R-CNN - 90%
Hu et al. (2021) [56]	Aluminium conductor composite core	Image normalisation	Inception ResNet v2, ResNet-18, ResNet-v2-50	Accuracy: Inception ResNet v2 - 97.01%
Kraljevski et al. (2021) [58]	Sensor network signals of aluminium and CFRP plates	FFT, low-pass filtering, PCA	AEC, HMM, RBF-SVM, Random Forest	Clustering score: AEC-PCA - >0.9
Niccolai et al. (2021) [59]	Aerospace structures	RoI	Direct and Indirect spot CNN	Accuracy: Indirect spot CNN - 98%
Siljama et al. (2021) [60]	Stainless steel	Normalisation	CNN	POD: $a_{90/95}$ = 2.1 mm
Fakih et al. (2022) [61]	Aerospace/mechanical/civil structures	Geometric constraints, Approximate Bayesian computation	ANN	Precision: ANN - >95%
Le et al. (2022) [62]	Aircraft structure	PCA	SVM, Naïve Bayes, *k*-NN, Random Forest, Logistic Regression	Average accuracy: SVM - 89.48%
Risheh et al. (2022) [63]	Steel structures	RoI, threshold selection, image segmentation, Canny edge detection	*k*-means clustering	Defects detected accurately

Aircraft structure corrosion was analysed using NDT data with PCA for feature extraction and SVM, Naïve Bayes, Random Forest, *k*-NN and Logistic regression models [62]. SVM outperformed all other models with 89.48% average accuracy. *k*-means clustering for NDT steel structure was developed by [63] to determine defects with RoI, thresholds, image segmentation and Canny edge detection techniques. This method does not need training and can detect defects accurately in smaller datasets.

Further, [10] was an automated evaluation of CFRP component NDT data with discontinuities such as delaminations, layer porosity, volume porosity and foreign bodies. These CFRP C-scans were converted to .png images using ULTIS® NDT Kit software and trained Machine Learning classifiers. The positive class had 37 annotated discontinuities with 18 delaminations and 19 porosities and consisted of 222 total training samples. The gradient of histograms for feature extraction was combined with SVM and Random Forest to classify discontinuities. The gradient histogram-SVM had the highest AUC of 0.987 and 10% of FP rate, but the gradient histogram-Random Forest classifier had a lesser FP rate for the positive class. In contrast, the gradient histogram-Random Forest classifier gained lesser confidence than the gradient histogram-SVM. There is a requirement for more training data with positive class samples to increase the classification rate.

Following, [64] detected anomalies using a Deep Learning technique with the same GLARE® NDT dataset used in the proposed model.

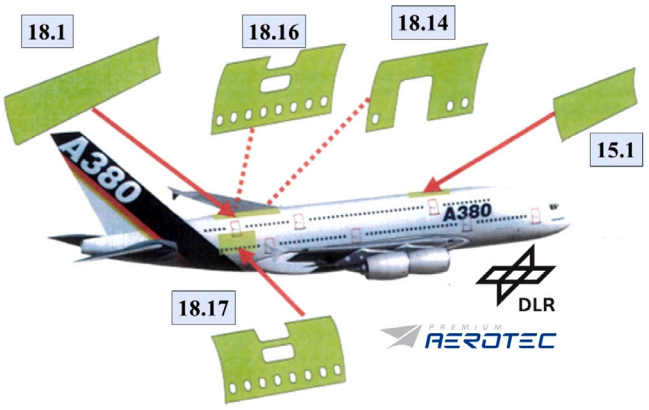

Fig. 1. A380 FML panels [1,2,65,66].

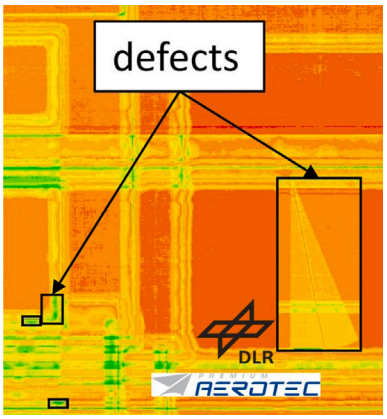

Fig. 2. Sample NDT data with defects (magnified).

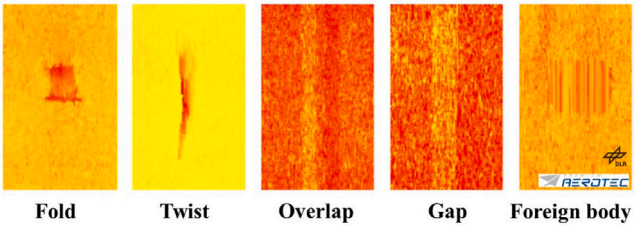

Fig. 3. Defect categories [1,2,65,66].

The NDT scans were converted to grayscale images with Python programming. These images were pre-processed using a Laplace filter to extract local material thickness and edge information as features, leading to an advantage in differentiating faulty and splice regions. These features trained the Deep Learning architecture with the first six CNN layers and one Atrous Spatial Pyramid Pooling (ASPP) layer that helps for significant faulty pixel classifications and another CNN layer with the last layer of Upsampling. The Stochastic Gradient Descent (SGD) for the learning method and Softmax cross entropy for the error function were used in this research. This classifier achieved an average high exclusion rate (manual inspection) of 97.36% for the component area on the test data; training steps are inversely proportional to the True Positive rate. The disadvantages of this classifier are: the exclusion rate varies with the component type and has a higher False Positive rate. This classifier determines non-faulty regions instead of differentiating faults and displays additional faults even in non-faulty regions. This method needs more training data for faulty regions to improve its performance and use it in real-time offline-QA of aircraft production.

The proposed research aims to develop an automated evaluation of aircraft NDT data, i.e., an offline-QA to help human examiners. Learning defects from aircraft production involves data acquisition, pre-processing, Machine Learning training, predictions and determining the model's confidence. Choosing an appropriate Machine Learning algorithm can seem complicated because many supervised and unsupervised algorithms use different learning strategies. However, choosing an algorithm depends on the quantity of data, data type, applicable insights and the requirement to utilise the model's evaluation results. Highly flexible models tend to overfit data by modelling minor variations that could be noise. Simple models are easier to interpret but might have lower accuracy. Therefore, choosing a suitable algorithm requires trading one benefit against another, including model speed, accuracy and complexity. In contrast to the literature survey (Table 1), the proposed work comprises state-of-the-art Machine Learning classifiers with distinct image feature extraction methods to detect two classes (binary classification): defects and good components in the aircraft ultrasonic-scan imageset.

2.1. NDT dataset

GLARE® [67,68] is a new FML class that produces A380 aircraft structures. The A380 comprises 15.1, 18.1, 18.14, 18.16, 18.17 components, as in Fig. 1.

The FML of the A380 NDT inspection technique is explained in the Airbus Test Method for inspection processes (AITM) AITM6-4001 (confidential). The aircraft production company, Premium AEROTECH GmbH (PAG), followed signal analysis requirements according to the AITM6-4001 and generated inspection reports. These inspection reports classified defects according to the AITM6-4001 and provided ground truth values (C-scans) for automated evaluation. The data collected from NDT inspection reports are plotted on a plane view of the component as images, known as C-scans (process mentioned in AITM6-4001). Fig. 2 shows a sample C-scan with denoted defects.

In the proposed approach, the NDT ultrasonic inspection report of FML A380 contains C-scans of each aircraft component. These scans (.xml file – raw dataset) were analysed using the quality software ULTIS®-TESTIA (NDT Kit). The experts at DLR-ZLP denoted the defects in the raw dataset with the help of PAG inspection reports and visualised them using this software, forming the ground truth data for this research. This NDT Kit creates three files, .nkc, .nkd and .nkz for each C-scan. The .nkc file has the original C-scan data consisting of two blocks: the first block is the header of the file with a length (in bytes) defined by the data offset field and written in ASCII format (indications and values). The second block of .nkc contains the physical data written in binary format. The .nkd file contains defect information such as file name, defect surface (mm^2), outline surface (mm^2), outline length (mm) and comments. Any other information is stored in the .nkz file.

In the proposed approach, the defect classes of the C-scans are categorised as porosity (Fig. 2), fold, twist, overlap, gap and foreign body, as illustrated in Fig. 3. There were 343 data samples and 99 contained at least one defect as illustrated in Fig. 4. Fig. 4 describes that the minimum number of defects in an image of a component is one and the maximum is 15. Most defects belong to the porosity category (distribution over the different defect types is confidential). The proposed method pre-processes the data using these 343 data samples for further processing. The quantity of data samples used in this study is limited because of the industrial aircraft production rate.

2.2. Machine learning model

The proposed model comprises training and evaluation (Section 4) processes. Preparation for the training process includes three primary

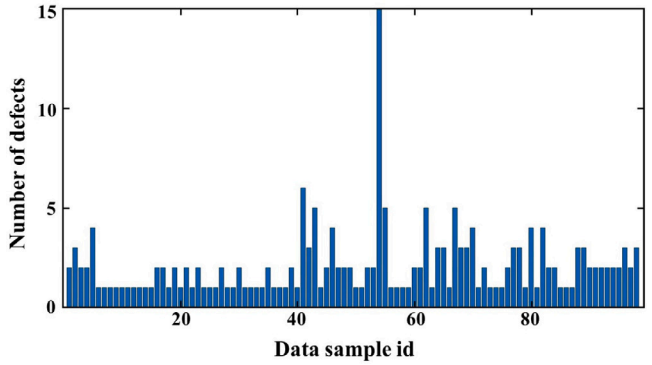

Fig. 4. Distribution of defects from NDT data.

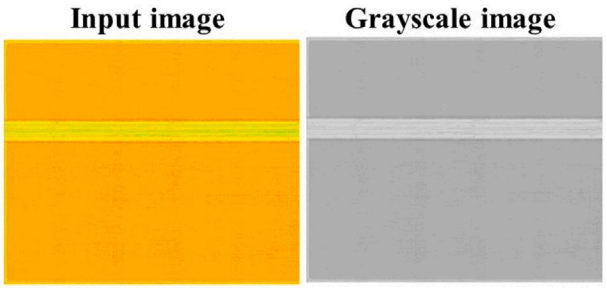

Fig. 5. Input image in RGB format and its corresponding grayscale image. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

steps: pre-processing, processing and post-processing data. The pre-processing involves converting the C-scans to Machine Learning compatible format. The ULTIS® enables storing complete C-scan information as an image. Manually, all 343 data samples were converted to 8-bit .jpg images, forming an imageset of defective and non-defective parts. Next, pre-processing is labelling .jpg images to prepare for model training. For labelling, all images were relabelled using MATLAB's Image Labeler application. This app consecutively was loaded with 99 defect images with 208 defects and 244 non-defective images for labelling. It stores the Region of Interest (ROI) labels (rectangle – position, pixel area) and Scene Labels (defect and good). The ROI for the defect scene label are the rectangles around different defects as shown in (Fig. 2) and the entire image for a good scene label. These labelled images are stored for proposed model training with 'defects' and 'good' classes. As there are two categories for classification, the proposed model is a binary classifier and 'defect' is a positive class, as the model aims to determine defects in the images and 'good' is a negative class.

$$ds = \sqrt{length * breadth} \tag{1}$$

where: ds is the defect size in pixels (px), length and breadth of the rectangular defect label

Further, pre-processing includes calculating defect size (ds) in the image labels. ds is defined as the square root of the defect area as in Eq. (1). The defect area is obtained from the rectangular image label dimensions (length, breadth). A square root over the defect area is formulated for two reasons: for standardising all the defect data and most defects are not frame-filling, i.e., the defect pixel area is not equal to the rectangular label area, for example, twist, fold, pores (Fig. 3). The minimum defect size in image labels encountered is 6 px and the maximum is 383 px. According to the defect size, all 208 defects were cropped to their equivalent defect label size and stored as the labelled defect (positive) imageset. The 244 labelled non-defective images formed the good (negative) imageset.

The processing step has feature extraction and training. It includes training the proposed Machine Learning model with a feature set from the training imageset (positive and negative imageset) and class labels – defect and good. The feature set is obtained from different image feature extraction techniques: LBP, MSER, KAZE, SURF and HoG. Each feature extractor has a bag-of-features to store its features. Each bag-of-features (feature set) is input to each state-of-the-art Machine Learning model for binary classification: SVM, Decision Trees, Random Forest, k-NN and Naïve Bayes. MATLAB's Classification Learner application was loaded with the training set (a feature set and class labels). During the training process, the Cross Validation (CV) [69,70] technique is applied to the training set to prevent overfitting (model overtrain), underfitting (insufficient model training), to observe the model's reaction to a similar independent dataset and prediction error function. The CV consists of exhaustive (iterates randomly on all data points) and non-exhaustive methods (iterates randomly on partitioned data points simultaneously). The k-fold and hold-out techniques are examples of non-exhaustive methods implemented to validate Machine Learning models. The hold-out approach arbitrarily sub-samples more for the training than validation. The k-fold method randomly partitions the prime training set into k equal subsets; one subset forms validation and $(k-1)$ subsets for training. The cross-validation process is repeated k times using each of k samples at least once for validation. The average accuracy of all k-folds determines the model's ability to predict new data. The 10-fold CV is used for the proposed model validations, where $k = 10$. The k-fold is suitable for the proposed model because of the smaller training set and prevents overfitting.

Lastly, benchmarking predictions of different state-of-the-art Machine Learning binary classifiers embedded with distinct image feature extraction techniques are stored in the post-processing for further model evaluation. The binary classifiers with high confidence scores are recommended (Section 3) for NDE 4.0 (Section 4).

2.3. Model pipeline

The proposed Machine Learning model pipeline comprises two steps: training – feature extraction and classification (including validation). The in-built functions of MATLAB were used with the Classification Learner App for the proposed model. An algorithm for the proposed model is as follows:

(1) **Feature Extraction**: input positive imageset (208 cropped defect images) and negative (244 good images) imageset of RGB or truecolour images (as shown in Fig. 5) as an image datastore to form a training imageset. Datastore can store larger feature vector size and increases processing rate.

$$training\ imds{=}imageDatastore\ \ (folder\ path)$$

(a) These labelled images of both classes have features extracted using custom extractors as follows:

(i) Convert all input RGB images to grayscale (Fig. 5) for LBP, MSER, KAZE and SURF feature extraction (HoG can extract features from RGB and grayscale images)

$$grayscale\ image = rgb2gray\ \ (RGB\ image)$$

(ii) LBP (Fig. 6) and HoG (Fig. 7) features of each input image

$$features = extractLBPFeatures\ \ (grayscale\ image)$$

$$features = extractHOGFeatures\ \ (RGB\ image)$$

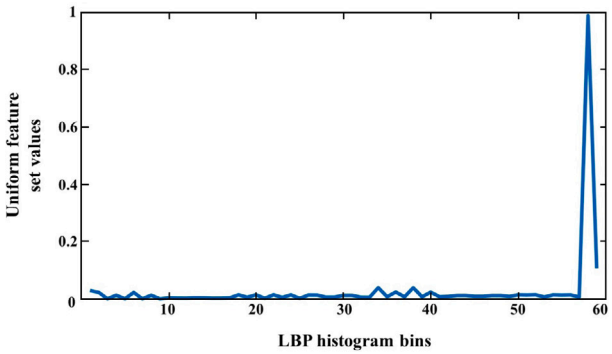

Fig. 6. LBP features.

(iii) For MSER (Fig. 8), KAZE (Fig. 9) and SURF (Fig. 10) features of each input grayscale image – detect regions and extract features from each these regions

$$regions = detectMSERFeatures\ (grayscale\ image)$$

$$regions = detectKAZEFeatures\ (grayscale\ image)$$

$$regions = detectSURFFeatures\ (grayscale\ image)$$

$$features = extractFeatures\ (grayscale\ image,\ regions)$$

(iv) Each custom feature extractor has a Bag-of-visual-words constructed

$$bag = bagOfFeatures\ (training\ imds, Custom\ Extractor\ Name, extractor\ function\ handler)$$

(v) Load *scene data* as an encoded bag-of-features from each custom extractor and *training imageset*

(vi) Load all labels of *training imageset* to *scene labels* as an attribute to *scene data*; label names 'defect' and 'good' are stored as *scene type*

(2) **Training**: Open Classification Learner App and load *scene data* and *scene type*

(a) select all *scene data* as predictors
(b) simultaneously apply Cross-Validation with 10-fold
(c) start the session and store validation results (Section 3)
(d) In the Classification Learner App, use parallel computing to train all available Machine Learning classifiers at once.
(e) Store all trained classifiers for further analysis (Sections 3, 4)

A part of the data from the A380 component is visualised in Fig. 5 (cropped smaller section of a good part) due to data confidentiality, the input RGB image is converted to grayscale for feature extraction processes (except for HoG).

Fig. 6 represents the LBP feature graph of encoded local texture information in binary format extracted from Fig. 5. The LBP feature partitions the grayscale image into non-overlapping cells. The histogram bins represent the number of features from each cell in the grayscale image and bins depend on the number of neighbours of each cell. The uniform feature set values of each cell (local texture information) are plotted with LBP histogram bins and each histogram describes an LBP feature.

Fig. 7 illustrates HoG features (zoomed-in) (marked in white colour) extracted from an RGB input image (Fig. 5) converted to a binary image. This binary image is decomposed into small, squared cells and computes a histogram of oriented gradients in each cell. Then, it

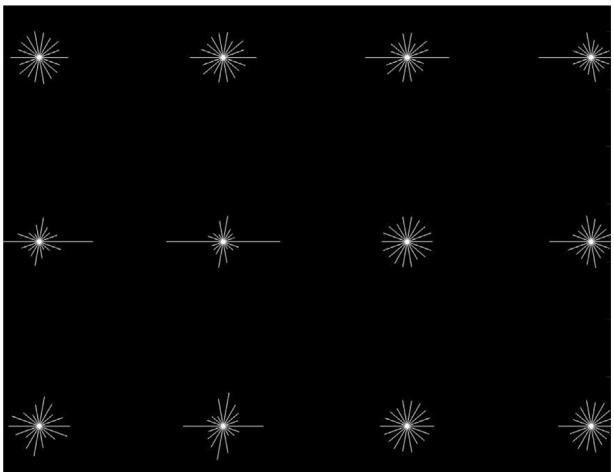

Fig. 7. HoG features.

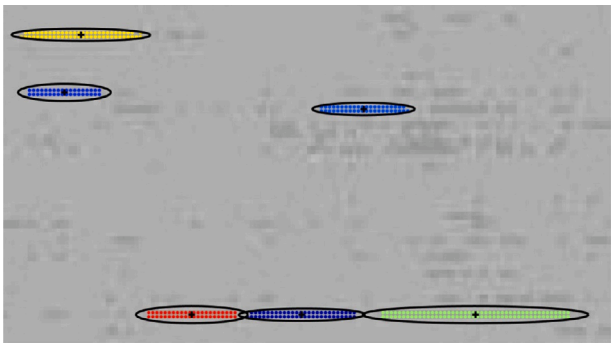

Fig. 8. MSER features.

normalises the result using a block-wise pattern and returns a descriptor for each cell.

Fig. 8 shows MSER feature extraction (zoomed-in) for Fig. 5. From the grayscale image, co-variant regions (MSER regions) (coloured regions) are extracted by checking the variation of the region area size between different intensity thresholds. Ellipses (marked in black colour) and centroids (marked in black plus) from MSER regions are stable connected components of the grayscale image.

Fig. 9 displays KAZE features (zoomed-in) from Fig. 5. The grayscale image is used to obtain KAZE points (marked in blue ellipses and black plus), with non-linear diffusion to construct a scale space for the grayscale image and then detect multi-scale corner features from that scale space.

Fig. 10 shows SURF points (marked in black colour) (zoomed-in) are extracted from Fig. 5. These SURF points are obtained using Hessian blob detector and its feature vector from Haar wavelet from the grayscale image.

During the training process, HoG extracted 34,596 features from each image and 422 × 34,596 feature vectors were elected with the strongest features from each class. These strongest HoG feature vectors created a bag-of-features with 500 clusters. SURF extracted 12,093 features (total – 422 × 12,093) and the strongest features from each class formed 50 bag-of-features clusters. MSER extracted 10,644 features with 500 bag-of-features clusters. KAZE extracted 9124 features with 500 clusters and LBP extracted 420 features with 302 bag-of-features clusters. Overall, in the training process, HoG produces the most feature vectors in this setup and more features are required to train Machine Learning classifiers to gain better prediction results.

The classifiers trained in the proposed method from the Classification Learner App include k-NN – fine, medium, coarse, cosine, cubic,

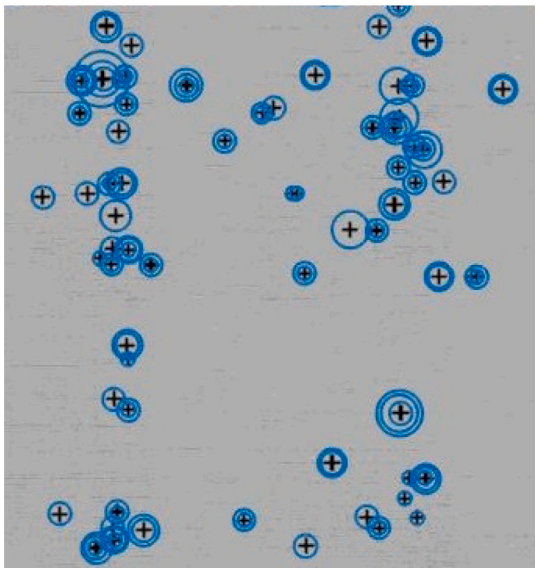

Fig. 9. KAZE features.

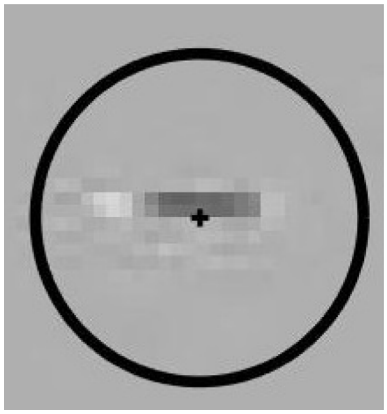

Fig. 10. SURF features.

weighted and Decision Trees – fine, medium, coarse. Random Forest – ensemble boosted trees, ensemble bagged trees, ensemble subspace discriminant, ensemble subspace *k*-NN, ensemble RUS boosted trees; SVM – linear, quadratic, cubic, fine Gaussian, medium Gaussian, coarse Gaussian and Naïve Bayes. The performance of all these classifiers with image feature extraction methods is discussed in Section 3.

3. Experimental result and discussion

The proposed Machine Learning model is evaluated using metrics such as accuracy, precision, recall, F1-score, Receiver Operating Curve with Area Under the Curve (ROC-AUC) [71], *k*-fold Cross Validation and POD certification. The classifier's confidence is designated based on the values of true-positive (TP), true-negative (TN), false-positive (FP) and false-negative (FN).

From Table 2, a prediction is a TP or TN when the predicted and actual values are the same; TP is when a defect is classified as defect class and TN is a good part classified as a good class. An FP or FN occurs when the predicted and actual values are different; FP is the classification with the predicted value of a defect, but the actual value is a good aircraft part and FN is vice-versa. A matrix representation of all these values forms a confusion matrix.

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

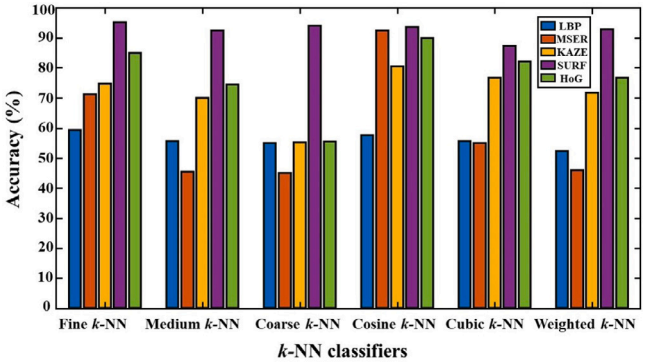

Fig. 11. *k*-NN accuracy.

Table 2
Possibilities of predictions.

Type	Predicted	Actual value
True positive	Defect	Defect
True negative	Good	Good
False positive	Defect	Good
False negative	Good	Defect

The accuracy of the Machine Learning model is the rate of correct prediction to the total predictions (Eq. (2)).

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

Precision is the rate of correct defects predicted to the total positive predictions by the trained Machine Learning model (Eq. (3)).

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

Recall or sensitivity is the rate of correct defects predicted to the total positive instances in the test data (Eq. (4)).

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5}$$

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

$$FPR = 1 - Specificity = \frac{FP}{TN + FP} \tag{7}$$

The F1-score is the harmonic mean of precision and recall (Eq. (5)). The Rate of Change (ROC) is the probability curve [72] and the Area under the ROC curve (AUC) is the degree of separability. ROC-AUC evaluates the trained classifier's performance in distinguishing the 'defect' and 'good' classes with the values of True Positive Rate (TPR) (recall or sensitivity) and False Positive Rate (FPR). The FPR is calculated based on the specificity (Eq. (6)) of the trained model using Eq. (7). The ROC-AUC curve is plotted with FPR (x-axis) against TPR (y-axis). The trained model can better classify defects and good aircraft structures if the AUC value is higher.

3.1. Cumulative models

Figs. 11–14 illustrate the analysis to choose the best accuracy of cumulative Machine Learning classifiers with image feature extraction methods.

From Fig. 11, the performance of LBP-Fine *k*-NN has the highest accuracy of 59.3% and the least of LBP-Coarse *k*-NN with 55%. MSER-Cosine *k*-NN has the highest accuracy of 92.4% and least with MSER-Coarse *k*-NN of 45%. KAZE-Cosine *k*-NN has 80.5% high accuracy and a low of 55.2% with KAZE-Coarse *k*-NN. SURF-Fine *k*-NN has 95.2% highest accuracy and 87.4% with KAZE-Cubic *k*-NN. HoG-Cosine *k*-NN has 90% accuracy and is low with HoG-Coarse *k*-NN of 55.5%.

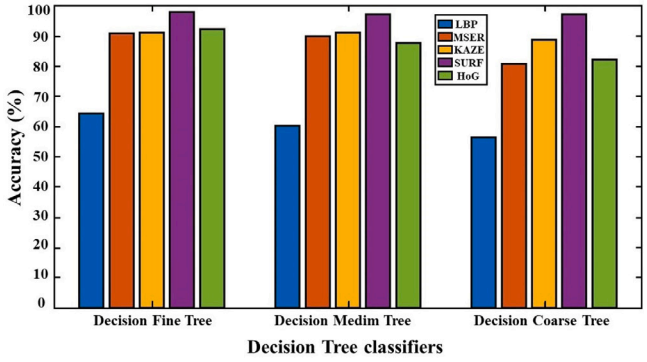

Fig. 12. Decision Tree accuracy.

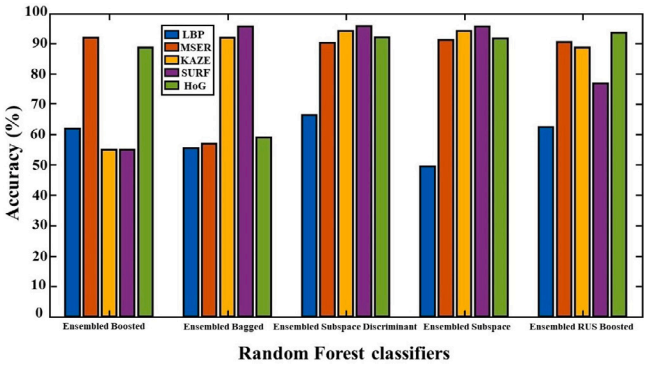

Fig. 13. Random Forest accuracy.

Fig. 12 shows performance analysis of the Decision Tree, the combination of LBP-Decision Fine Tree has 64.3% accuracy and less of 56.4% with LBP-Decision Coarse Tree. MSER-Decision Fine Tree has a high accuracy of 90.7% and MSER-Decision Coarse Tree has low accuracy of 80.7%. KAZE-Decision Fine Tree and KAZE-Decision Medium Tree have a similarly high accuracy of 91% and KAZE-Decision Coarse Tree has low of 88.6% accuracy. SURF-Decision Fine Tree has the highest accuracy of 97.9% and SURF-Decision Medium Tree and SURF-Decision Coarse Tree have an accuracy of 97%. KAZE-Decision Fine Tree gained 92.14% high accuracy and KAZE-Decision Coarse Tree of 82.1% low accuracy.

Fig. 13 demonstrates the Random Forest or Ensemble Trees detection rate, LBP-Ensemble Subspace Discriminant has gained 66.4% and low accuracy of 49.5% with LBP-Ensemble Subspace *k*-NN. MSER-Ensemble Boosted Trees has a high of 91.9% and MSER-Ensemble Bagged Trees of 56.9% low accuracies. KAZE-Ensemble Subspace Discriminant and KAZE-Ensemble Subspace *k*-NN achieved the highest accuracy of 94.2%, but KAZE-Ensemble Boosted Trees has 55% low accuracy. SURF-Ensemble Bagged Trees, SURF-Ensemble Subspace Discriminant and SURF-Ensemble Subspace *k*-NN have the same high accuracy around 95.6%; low accuracy of 55% with SURF-Ensemble Bagged Trees. 93.5% of accuracy is gained by HoG-Ensemble RUS Boosted Trees and a low of 59% with HoG-Ensemble Bagged Trees.

Fig. 14 illustrates the detection rate of SVM classifier, LBP-Linear SVM, LBP-Quadratic SVM, LBP-Cubic SVM, LBP-Fine Gaussian SVM and LBP-Medium Gaussian SVM has a similarly high accuracy of 66.4%; LBP-Coarse Gaussian SVM has low accuracy of 55%. MSER-Quadratic SVM and MSER-Cubic SVM have similar high accuracy of 92.6%, a low of 69.8% from MSER-Fine Gaussian SVM. KAZE-Linear SVM gained 92.1% high accuracy and KAZE-Coarse Gaussian SVM low of 63.6%. SURF-Linear SVM, SURF-Quadratic SVM, SURF-Coarse SVM, SURF-Medium Gaussian SVM and SURF-Coarse Gaussian SVM have matching high accuracy of around 96%. SURF-Fine Gaussian SVM achieved low

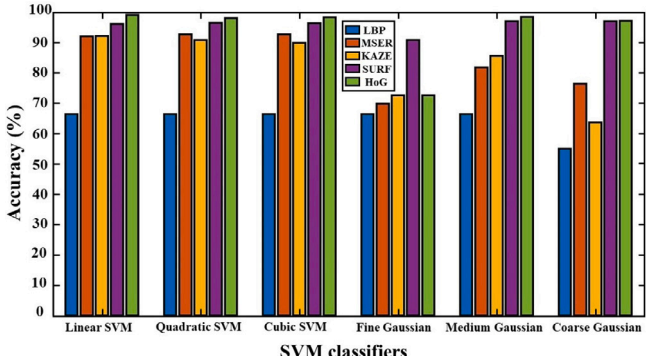

Fig. 14. SVM accuracy.

Table 3
Consolidated accuracy chart.

Feature extraction	Classifiers	Accuracy (%)
LBP	Linear SVM	**66.4**
	Ensemble Subspace Discriminant	66.4
	Fine *k*-NN	59.3
	Naïve Bayes	55
MSER	Quadratic SVM	**92.6**
	Ensemble Bagged Trees	91.9
	Cosine *k*-NN	92.4
	Naïve Bayes	57.6
KAZE	Linear SVM	92.1
	Ensemble Subspace Discriminant	94.2
	Cosine *k*-NN	80.5
	Naïve Bayes	**94.3**
SURF	Linear SVM	96.9
	Decision Fine Tree	**97.9**
	Fine *k*-NN	95.2
	Naïve Bayes	59.5
HoG	Linear SVM	**99**
	Ensemble RUS Boosted Trees	93.5
	Cosine *k*-NN	90
	Naïve Bayes	56

Table 4
Evaluation chart.

Classifiers	Accuracy (%)	Recall	Precision	F1-score
HoG-SVM	99	0.9919	0.9880	**0.984**
SURF-Fine Tree	97.9	0.9839	0.97	0.97

accuracy with 90.7%. The highest accuracy is gained by HoG-Linear SVM of 99% and the low accuracy of HoG-Fine Gaussian SVM with 72.6%.

The LBP had the lowest feature extraction performance with all classifiers compared to MSER, KAZE, SURF and HoG. The second least feature extraction interpretations were MSER, followed by KAZE. The selection of the best feature extraction methods influences the classifiers. SURF and HoG feature extraction methods were selected as the best to encase with classifiers to avoid false negatives. The Naïve Bayes (Table 3) and *k*-NN were not applicable with most feature extraction methods and thus were eliminated in the further evaluation process.

3.2. Best performing models

The highest accuracies from all classifiers are consolidated in Table 3. The embedded classifiers HoG-Linear SVM and SURF-Decision Fine Trees achieved the highest accuracy of 99% and 97.9%, respectively. Therefore, these two classifiers are further evaluated with recall, precision and F1-score metrics as exhibited in Table 4. HoG-Linear SVM gains the highest F1-score with 0.984 compared to SURF-Decision Fine

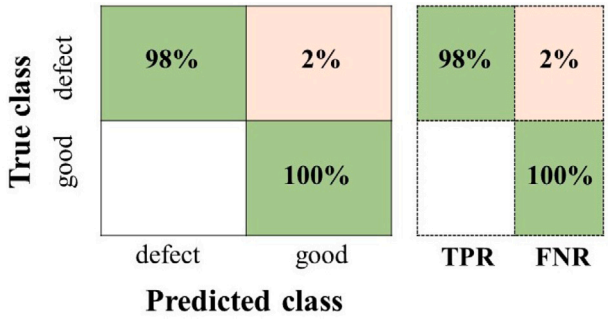

Fig. 15. HoG-Linear SVM confusion matrix.

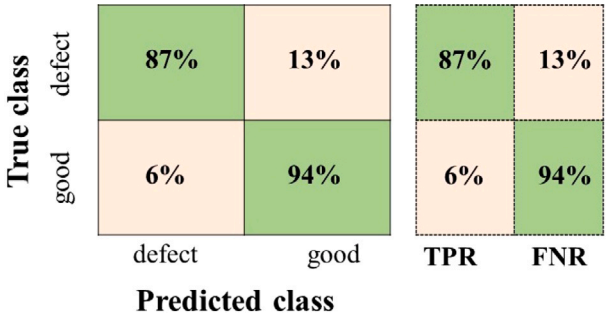

Fig. 16. SURF-Decision Fine Tree confusion matrix.

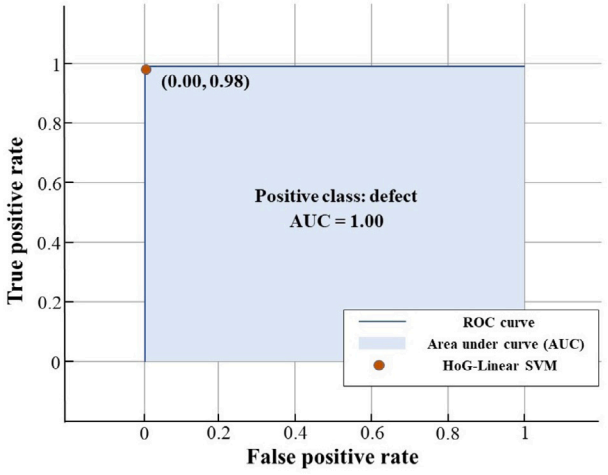

Fig. 17. HoG-Linear SVM ROC-AUC curve.

Tree F1-score of 0.97. A 98.4% of correct defects are predicted to total defect samples by trained HoG-Linear SVM model and in test data. In contrast, with test data, SURF-Decision Fine Tree has fewer correct defects predictions.

The selection of the best-fitting model anticipates factors such as low FN, high recall, precision and F1-score. Apart from accuracy, the confusion matrix and ROC-AUC curve help calculate these influencing scores and calibrate the model. Confusion matrices of HoG-Linear SVM (Fig. 15) and SURF-Decision Fine Tree (Fig. 16) reveal the lowest FN rate, with the former having 2% for positive class, zero for negative class. The latter has an FN rate of 13% and 6% for positive and negative classes, respectively.

The TP rate of HoG-Linear SVM is 98% for the positive class and 100% for the negative class and the TP rate of SURF-Decision Fine Tree for the positive class is 87% and 94% for the negative class.

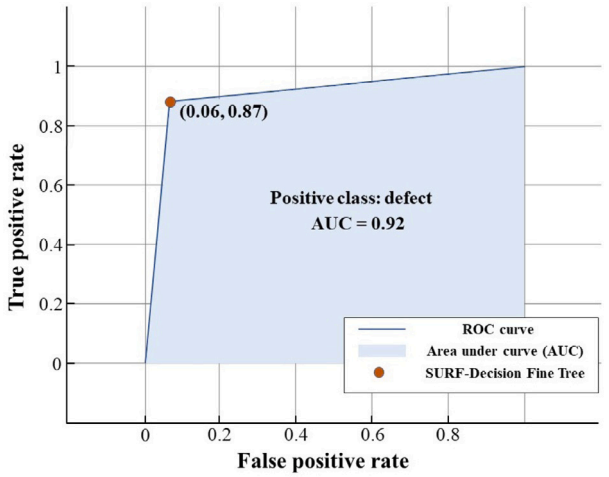

Fig. 18. SURF-Decision Fine Tree ROC-AUC curve.

The ROC-AUC curves provide AUC values for HoG-Linear SVM with $AUC = 1.00$ and prediction probability of zero for negative and 0.98 for positive classes as demonstrated in Fig. 17. The SURF-Decision Fine Tree has $AUC = 0.92$ and prediction probability of 0.06 for negative and 0.87 for positive class predictions as represented in Fig. 18.

After assessing all the evaluation metrics from Table 3 and Table 4, the best-performing embedded Machine Learning classifiers are HoG-Linear SVM and SURF-Decision Fine Tree. These have negligible FN leading $Recall \approx 1.00$, high precision and F1-score. The robust requirement for the proposed model is to achieve 100% of TP rate on the prediction data and zero FN rate. The FN rate is essential for calibrating the proposed model and ROC-AUC curves help with calibration. The threshold curve is the ROC curve that separates positive and negative classes, selected to obtain a significantly lower or zero FN rate and maximum TP rate in the prediction process. From Figs. 17 and 18, as the TP rate increases, the FP rate also increases. If the AUC of HoG-Linear SVM decreases below 1.00 and SURF-Decision Fine Tree above 0.92, their FN rate increases. The FP rate is negligible (an experienced examiner can scrutinise the FP visually) for the real-time usage of the proposed system, but the FN rate should not be increased because of the risk involved in the industrial offline-QA of aircraft production.

As SVM is primarily a binary classifier and HoG-Linear SVM has performed best with the prediction data, selecting it as a predominant classifier for the proposed approach is beneficial. Hence, it is further evaluated with the POD certification process. SURF-Decision Fine Tree can be an option for multi-class classification.

3.3. Comparison and constraints

The proposed HoG-Linear SVM classifier performs better than [10, 64]. But it has some constraints, such as the Linear SVM classifier is a black box, as the path to its predictions is unknown. But Decision Fine Tree is a grey-box as its prediction path is returned as a binary tree split into branching nodes based on input data values.

A binary tree resulting from one of the proposed prediction analyses is illustrated in Fig. 19. This binary tree starts with the root and has two branches at each node; the nodes contain conditions for the predictions. This tree has four and three levels, with the leaf nodes having the predicted classes, thus explicitly demonstrating the prediction analysis. The SURF-Decision Fine Tree can be feasible for real-time offline-QA in aircraft industries for NDE 4.0 and inline-QA, but it could be complicated with a heap of Decision Trees and branches. The HoG-Linear SVM analysis from the proposed prediction dataset may not reflect an accurate performance in the real-time industrial offline-QA due to a deficit in additional positive class training data from each

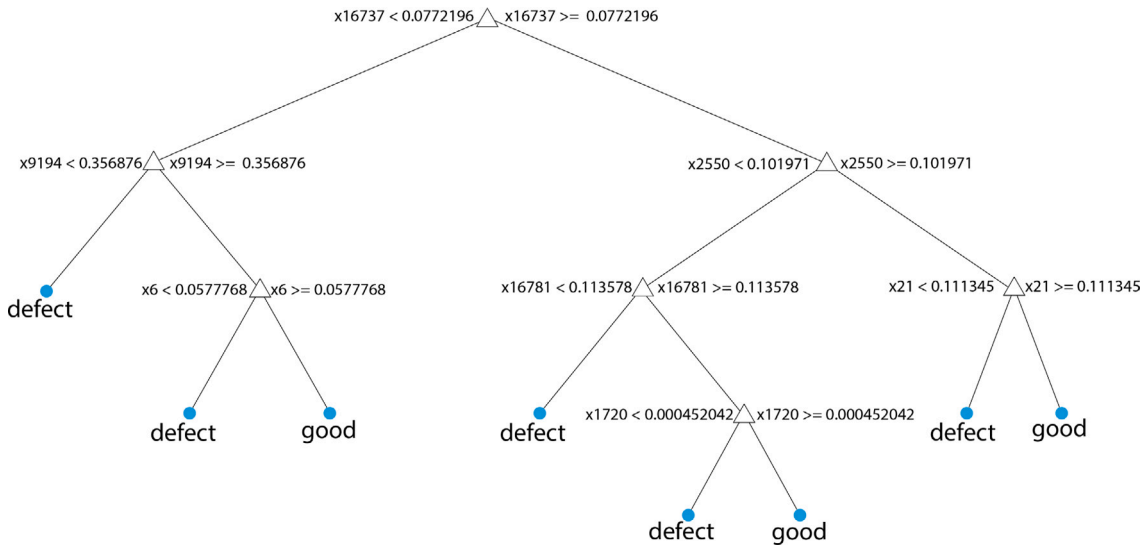

Fig. 19. SURF-Decision Fine Tree Classification Tree Viewer.

Fig. 20. Local SSIM Map with Global SSIM value 0.9868 of a component.

aircraft component. The FN was generated with the 'fold' defect type from all the defects due to small fold size and fewer fold samples. More training data can lead to an increase in the performance of the proposed HoG-Linear SVM classifier.

Another constraint is the data loss from storing NDT scans as .jpg images in the pre-processing phase. The .jpg format compresses images, but the raw data can be converted to .bmp images using the same ULTIS® NDT Kit software, as .bmp is an uncompressed raster and high-quality file format. The signal analysis with a set of five .jpg and five .bmp images from each aircraft component were analysed to determine the data loss. The Peak Signal-to-Noise-Ratio (PSNR), Mean Squared Error (MSE) and Structural Similarity Index Measure (SSIM) [73] are commonly used to calculate data loss. For this signal analysis, SSIM for measuring image quality is preferred. For input .bmp (reference image) and .jpg (comparing image), images of a component are used to obtain the local and global SSIM values and SSIM maps. If the SSIM value is closer to 1, it signifies better input image quality.

Fig. 20 shows a sample local SSIM map of a component and the dark pixels are the small values of local SSIM. The regions with small local SSIM values correspond to areas where the .jpg image noticeably differs from the .bmp image. The bright pixels represent the large values of local SSIM. These bright pixel regions correspond to uniform regions of the .bmp image, where data compression has less impact on the .jpg image. The SSIM values for 15.1 component is 0.9962, 18.1 is 0.9953, 18.14 is 0.9940, 18.16 is 0.9944 and 18.17 is 0.9868.

$$Dataloss = 1 - SSIMvalue \tag{8}$$

The data loss is calculated as in Eq. (8). The worst-case data loss is 1.32% and the average data loss is 0.66%. These .bmp images were trained and tested with the HoG-Linear SVM classifier and observed that data loss had no influence on its performance.

4. Certification

The reliability of NDE is defined as determining the probability of a defect in different defect-size datasets during the evaluation process [74]. The quality assessment for the reliability of NDE is essential for aircraft structural management [75]. The certification is a statistical validation process for inspecting the reliability of NDE approaches with POD analysis [76,77]. The proposed certification process is an automatic error detection (intended for Ultrasonic Testing) to verify if the proposed Machine Learning classifier can help an examiner in QA. This process involves acquiring the bare scan of NDT data in the squirter or X-ray systems and manually converting NDT scans to images using ULTIS® for the Machine Learning process. The evaluation is the human investigation of the scanned image to find defects in the scan data and for the Machine Learning process, the trained classifier accomplishes the prediction process. This qualification is based on the POD concept to find defect sizes reliably.

In general, POD is translated into the reliability of finding a given defect size in px (minimum size to be detected). The minimum size contains the POD knowledge and is implemented with the 29/29 method. There are 29 defects in the minimum size to be detected and this method has to detect all 29 reliably without missing one. Thus, defect size in px, $a_{90/95}$, automatically fulfils the POD criterion: gain 0.9 at 95% without dealing with the POD concept. The disadvantage of the 29/29 method is that the POD requirement is fulfilled, but the test model's reliability is unknown.

The certification process is mainly used to avoid the risks and challenges such as software being a black-box has to be noted, training an algorithm is crucial and requires expertise and reliability of the Machine Learning model in terms of the new dataset, types of defects, different NDT testing methods and feature extraction techniques.

Recent NDE 4.0 research has evaluated their Machine Learning models for NDT data using POD [60,62,78,79]. The possible certification process with the proposed HoG-Linear SVM model includes evaluation of NDT by the NDT-test engineer and algorithm for the predictions of this model or collecting feedback from them regarding the quality of the algorithm generated; evaluation leads to further training of algorithm and repeats often. It can verify the model's

reliability while encountering new defect datasets and implementing distinct feature extractions or validation methods. The HoG-Linear SVM model's confidence and adoption reliability in offline-QA in the aircraft industry is analysed with a POD function.

4.1. Probability of detection

A POD is a function of the defect size; it evaluates the smallest flaw size and combines its quantitative and qualitative parameters [80]. The 90/95 defect size information is used as a reference and detects defects with a probability of 90% at 95% of confidence level [81]. Two methods to determine POD are Hit/Miss data for binary data and signal response data for continuous data. POD Hit/Miss results is a hit when the defect is detected and failure is a miss.

$$POD = \frac{TP}{TP + FN} \tag{9}$$

$$Hit = a > a_{largest} \tag{10}$$

$$Miss = a < a_{smallest} \tag{11}$$

The POD is calculated using Eq. (9), where TP is a hit and FN is a miss. Hit/Miss data has a defect size range in px, $a_{smallest}$ (minor defect size of 6 px) and $a_{largest}$ (maximum defect size of 383 px) to determine the substantial uncertainty of the proposed HoG-Linear SVM model to detect the defect or not. Hit/Miss data suits the proposed model as SVM performs better as a binary classifier and SURF-Decision Fine Tree could perform better as a multi-class classifier. A Hit is measured if the inspection system detects a defect size, a that agrees on Eq. (10) and a Miss is measured if the inspection system does not detect a defect size, a that agrees on Eq. (11). For Hit/Miss data having a vast number of smallest or largest defects will not help to gain information on the POD(a) function that will fit the data. The information required for estimating the POD(a) function has to be maximised so the defect sizes are uniformly distributed between the smallest and largest defect size of interest using the 29/29 method. The POD is calculated with new defects and helps to measure the performance gap of the proposed HoG-Linear SVM model with defect parameter size of defect area in px. The overall defect range to be investigated is 6 px to 383 px and intervals required within the defect size range to be investigated is 5 px.

In the 29/29 method, having a minimum sample of 29 defects in each defect width interval is necessary. So newer defect dataset was formed for the POD(a) function by combining the existing defect samples and artificially created to generate more data. The artificial defects were constructed using image augmentation methods of rotation, skewing and mirroring. The smallest defect size in the positive imageset is 6 px and the largest is 383 px. A sum of 29 artificial defects was fabricated in each defect size interval (5 px - to generate more defects), creating 29 × 5 = 145 defect samples. These 145 artificial defects combined with the positive imageset of 189 + 16=205 existing defects. So a total of 350 (145 + 205 = 350) defects with different sizes (considerable cost) are used to create a POD(a) function. From the existing negative imageset (233 + 7 = 240), images were cropped to match the same defect width interval (5 px) to obtain 350 specimens. So, these control specimens (350) are randomly mixed with the defect specimens (350). The trained HoG-Linear SVM classifier must detect all the 29 defects in that defect width interval to achieve the 90% PoD with a 95% confidence level.

The trained HoG-Linear SVM classifier used the prediction process to determine these 700 specimens, all Hit/Miss was recorded to plot their probability as represented in Fig. 21. The proposed POD function is prior improved due to the HoG image feature extraction method containing denoising and feature vectors [82]. The performance gap is calculated from the POD function (Fig. 21) as the difference between the 90/95 and the smallest pixel size, 20 - 6 = 14 px. So a minor defect of size 20 px can be identified as TP, achieving a POD of 90/95. The

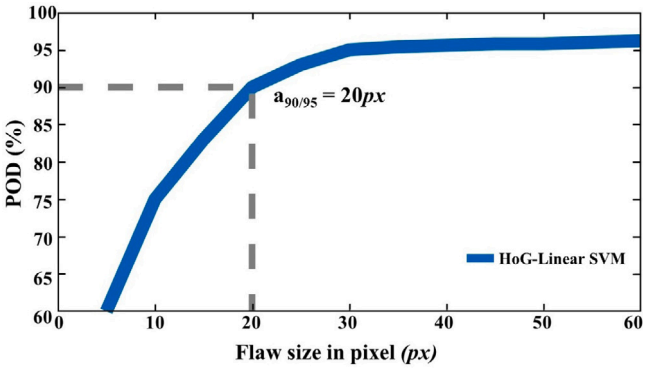

Fig. 21. POD curve.

Miss rate included more of the 'fold' defect type smaller than 8 px. The performance gap of the proposed POD(20) can be minimised with better quality of NDT scan perception. Since the PAG testers only flag defects larger than a specific size, there might be more detectable defects in the data samples, but their test reports do not exist for annotation. The factors influencing the proposed POD(20) are NDT scan image resolution (requires better image quality) and defect frame-filling (not all defect samples are frame-filling, but control specimens were frame-filling). Due to this frame-filling issue, POD(20) indicates that at least 20 px must be in an image with any defect size and 8-bit resolution (256 px).

Evaluation of the data by a tester is time-consuming and has the probability of missing defects. The HoG-Linear SVM model can save time and reduce the frequency of miss counts by highlighting areas of interest to the examiner. This model predicts defects based on pixel-by-pixel scans and executes instantly.

5. Conclusion

Offline-Quality Assessment for NDT-FML of A380 aircraft structures has been analysed to determine defects in the Ultrasonic Testing scan images with state-of-the-art Machine Learning algorithms, SVM and Decision Trees. These models are embedded with distinguished image feature extraction techniques SURF and HoG. The combination of HoG-Linear SVM (F1-score = 0.984, ROC-AUC = 1.00) and SURF-Decision Fine Tree (F1-score = 0.97, ROC-AUC = 0.92) outperformed all other models. The HoG-Linear SVM was further evaluated with the certification process with the POD function, enabling it to determine a defect size of 20 px in images. The HoG-Linear SVM has a performance gap of 14 px that can be improved with more defect samples for training and evaluation with industry partners for production.

Special note

The corresponding author conducted this research work with ZLP-DLR, Augsburg and Informatik-University of Bonn, Bonn, Germany in 2019.

Funding

Center for Lightweight Production Technology (ZLP), German Aerospace Center (DLR), Augsburg, Bavaria, Germany.

Preparation of this article – DFKI acknowledges financial support by the Lower Saxony Ministry for Science and Culture (MWK) through "Niedersachsen Vorab" (ZN3480).

CRediT authorship contribution statement

Navya Prakash: Methodology, Software, Validation, Formal analysis, Writing – original draft, Visualization. **Dorothea Nieberl:** Conceptualization, Data curation, Writing – review & editing, Supervision, Project administration. **Monika Mayer:** Conceptualization, Data curation, Writing – review & editing, Supervision, Project administration. **Alfons Schuster:** 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.

Data availability

The data that has been used is confidential.

Acknowledgement

We thank Prof. Dr. Jens Lehmann (Informatik-University of Bonn) for supervising this research (2019).

References

[1] Ucan H, Scheller J, Nguyen C, Nieberl D, Mayer M, et al. Automated, quality assured and high volume oriented production of fibre metal laminates (FML) for the next generation of passenger aircraft fuselage shells. Sci Eng Compos Mater 2019;26:502–8. http://dx.doi.org/10.1515/secm-2019-0031, https://elib.dlr.de/129574/.

[2] Apmann H, Mayer M, et al. Verfahren der INLINE-qualitätssicherung und der zerstörungsfreien prüfung innerhalb der fertigungslinie von faser-metall-laminaten. In: DLR congress (DLRK) conference - FML. 2017, https://elib.dlr.de/117260/.

[3] Bisle W, Meier T, Mueller S, Rueckert S. In-service inspection concept of GLARE® – an example for the use of new UT array inspection systems, ECNDT. 2006, https://www.ndt.net/search/docs.php3?id=3540.

[4] Vrana J, Singh R. NDE 4.0 - a design thinking perspective. J Nondestruct Eval 2021;24. http://dx.doi.org/10.1007/s10921-020-00735-9.

[5] Schmidt T, Dutta S. Automation in production integrated NDT using thermography. In: International symposium on NDT in aerospace. 2012, https://www.ndt.net/article/aero2012/papers/we3b1.pdf.

[6] Wunderlich C, Tschöpe C, Duckhorn F. Advanced methods in NDE using machine learning approaches. In: AIP conference proceedings 1949-020022. 2018, http://dx.doi.org/10.1063/1.5031519.

[7] Ren I, Zahiri F, et al. A deep ensemble classifier for surface defect detection in aircraft visual inspection, smart sustain. Manuf Syst 2020;4(1):20200031. http://dx.doi.org/10.1520/SSMS20200031.

[8] Nieberl D, Mayer M, Stefani T, Willmeroth M. Automated manufacturing of large fibre-metal-lmainate parts. In: European conference on composite materials. 2018, https://elib.dlr.de/124296/.

[9] Schuster A, Mayer M, Willmeroth M, Brandt L, Kupke M. Inline quality control for thermoplastic automated fibre placement. In: Procedia manufacturing, Vol. 51. Elsevier, FAIM; 2021, p. 505–11. http://dx.doi.org/10.1016/j.promfg.2020.10.071.

[10] Internal Study, Schmidt T, Mayer M, Rainer L, Kupke M. Pilotstudie automatisierte auswertung von NDT daten. DLR-IB 435-2015/32. 43 S, DLR-Interner Bericht, Unpublished https://elib.dlr.de/101533/.

[11] Caruana R, N-Mizil A. An empirical comparison of supervised learning algorithms. In: International conference on machine learning. 2006, p. 161–8. http://dx.doi.org/10.1145/1143844.1143865.

[12] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273–97. http://dx.doi.org/10.1007/BF00994018.

[13] Breiman L. Bagging predictors. Mach Learn 1996;24(2):123–40. http://dx.doi.org/10.1007/BF00058655.

[14] Rokach L, Maimon O. Decision trees. In: Data mining and knowledge discovery handbook. Springer; 2005, p. 165–92. http://dx.doi.org/10.1007/0-387-25465-X_9.

[15] Breiman L. Random forests. Mach Learn 2001;45:5–32. http://dx.doi.org/10.1023/A:1010933404324.

[16] Fix E, Hodges Jr JL. Discriminatory analysis - nonparametric discrimination: consistency properties. Technical Report 4, USAF School of Aviation Medicine; 1951, https://apps.dtic.mil/sti/pdfs/ADA800276.pdf.

[17] Bayes. An essay towards solving a problem in the doctrine of chances. In: FRS communicated by Mr. Price in a letter to John Canton, A.M. FRS. 1763, https://royalsocietypublishing.org/doi/pdf/10.1098/rstl.1763.0053.

[18] Fisher A. The use of multiple measurements in taxonomic problems. Ann Eugen 1936;7(2):179–88. http://dx.doi.org/10.1111/j.1469-1809.1936.tb02137.x.

[19] Cramer JS. The Origins of Logistic Regression. Tinbergen institute working paper No. 2002-119/4, 2002, http://dx.doi.org/10.2139/ssrn.360300.

[20] Bezdek JC, Ehrlich R, Full W. FCM: The fuzzy c-means clustering algorithm. Comput Geosci 1984;10(2–3):191–203. http://dx.doi.org/10.1016/0098-3004(84)90020-7.

[21] Faber V. Clustering and the continuous K-means algorithm. Los Alamos Sci 1994;22:138–44, https://www.cs.kent.edu/zwang/schedule/lj9.pdf.

[22] Jolliffe IT. Principal Component Analysis. Springer Series in Statistics; 2002, http://cda.psych.uiuc.edu/statistical_learning_course/Jolliffe%20I.%20Principal%20Component%20Analysis%20(2ed.,%20Springer,%202002)(518s)_MVsa_.pdf.

[23] McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 1943;5:115–33. http://dx.doi.org/10.1007/BF02478259.

[24] Broomhead DS, Lowe D. Radial basis functions, multi-variable functional interpolation and adaptive networks. In: Royal signals and radar establishment malvern (United Kingdom). Complex Systems Publications, Inc.; 1988, p. 321–55, RSRE-MEMO-4148 https://apps.dtic.mil/sti/citations/ADA196234.

[25] Kramer MA. Nonlinear principal component analysis using autoassociative neural networks. AIChE J 1991;37(2):11, https://people.engr.tamu.edu/rgutier/web_courses/cpsc636_s10/kramer1991nonlinearPCA.pdf.

[26] Van Der Malsburg C. Frank rosenblatt: Principles of neurodynamics: Perceptrons and the theory of brain mechanisms. In: Brain theory. Springer Berlin Heidelberg; 1986, p. 245–8. http://dx.doi.org/10.1007/978-3-642-70911-1_20.

[27] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: International conference on learning representations (ICLR). 2015, http://dx.doi.org/10.48550/arXiv.1409.1556.

[28] Girshick R. Fast R-CNN. In: IEEE international conference on computer vision. ICCV, 2015, p. 1440–8. http://dx.doi.org/10.1109/ICCV.2015.169.

[29] Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, inception-ResNet and the impact of residual connections on learning. 2016, http://dx.doi.org/10.48550/arXiv.1602.07261.

[30] Vaswani A, Shazeer N, et al. Attention is all you need. Neural Information Processing Systems 2017. http://dx.doi.org/10.48550/arXiv.1706.03762, https://dl.acm.org/doi/10.5555/3295222.3295349.

[31] Medjahed SA. A comparative study of feature extraction methods in images classification. IJIGSP 2015;7(3):16–23. http://dx.doi.org/10.5815/ijigsp.2015.03.03.

[32] Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell 2002;24(7):971–87. http://dx.doi.org/10.1109/TPAMI.2002.1017623.

[33] Matas J, Chum O, et al. Robust wide baseline stereo from maximally Stable Extremal Regions. Image Vis Comput 2004;22(10):761–7. http://dx.doi.org/10.1016/j.imavis.2004.02.006.

[34] Alcantarilla PF, Bartoli A, Davison AJ. KAZE features. In: Computer Vision - ECCV 2012. Springer Berlin Heidelberg; 2012, p. 214–27. http://dx.doi.org/10.1007/978-3-642-33783-3_16.

[35] Bay H, Tuytelaars T, Van Gool L. SURF: Speeded up robust features. In: Computer vision – ECCV 2006, Vol. 3951. Springer Berlin Heidelberg; 2006, p. 404–17. http://dx.doi.org/10.1007/11744023_32.

[36] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: IEEE computer society conference on computer vision and pattern recognition (CVPR'05). 2005, http://dx.doi.org/10.1109/cvpr.2005.177.

[37] Lowe GD. Distinctive image features from scale-invariant keypoints. Int J Comput Vis 2004;60:91–110. http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94.

[38] Zhang YL, Guo N, et al. Automated defect recognition of C-SAM images in IC packaging using support vector machines. Int J Adv Manuf Technol 2005;25(11–12):1191–6. http://dx.doi.org/10.1007/s00170-003-1942-1.

[39] Bernieri A, Ferrigno L, et al. An SVM approach to crack shape reconstruction in eddy current testing. In: IEEE instrumentation and measurement technology conference proceedings. 2006, p. 2121–6. http://dx.doi.org/10.1109/IMTC.2006.328502.

[40] Bernieri A, Ferrigno L, et al. Crack shape reconstruction in eddy current testing using machine learning systems for regression. IEEE Trans Instrum Meas 2008;57(9):1958–68. http://dx.doi.org/10.1109/TIM.2008.919011.

[41] Benítez HD, Loaiza H, et al. Defect characterization in infrared non-destructive testing with learning machines. NDT E Int. 2009;42(7):630–43. http://dx.doi.org/10.1016/j.ndteint.2009.05.004.

[42] Khodayari-Rostamabad A, Reilly JP, et al. Machine learning techniques for the analysis of magnetic flux leakage images in pipeline inspection. IEEE Trans Magn 2009;45(8):3073–84. http://dx.doi.org/10.1109/TMAG.2009.2020160.

[43] Wei H, C-Tong L. Automatic real time SVM based ultrasonic rail flaw detection and classification system. J Graduate Sch Chin Acad Sci 2009;26(4):517–21, http://journal.ucas.ac.cn/EN/Y2009/V26/I4/517.

[44] Shumin D, Zhoufeng L, Chunlei L. Adaboost learning for fabric defect detection based on HOG and SVM. In: International conference on multimedia technology. IEEE; 2011, p. 2903–6. http://dx.doi.org/10.1109/ICMT.2011.6001937.
[45] Freund Y, Schapire RE. A short introduction to boosting. J Japan Soc Artif Intell 1999;14(5):771–80, https://cseweb.ucsd.edu/yfreund/papers/IntroToBoosting.pdf.
[46] Saechai S, Kongprawechnon W, Sahamitmongkol R. Test system for defect detection in construction materials with ultrasonic waves by support vector machine and neural network. In: SCIS-ISIS. 2012, p. 1034–9. http://dx.doi.org/10.1109/SCIS-ISIS.2012.6505090.
[47] Salzberg SL. Book review C4.5: Programs for machine learning by j. Ross quinlan. Morgan Kaufmann publishers, inc. 1993. Mach Learn 1994;16(3):235–40. http://dx.doi.org/10.1007/BF00993309.
[48] D'Angelo G, Rampone S. Shape-based defect classification for non destructive testing. IEEE Metrol Aerospace (MetroAeroSpace) 2015;406–10. http://dx.doi.org/10.1109/MetroAeroSpace.2015.7180691.
[49] Sumesh A, Rameshkumar K, et al. Use of machine learning algorithms for weld quality monitoring using acoustic signature. Procedia Comput Sci 2015;50:316–22. http://dx.doi.org/10.1016/j.procs.2015.04.042.
[50] Malekzadeh T, Abdollahzadeh M, et al. Aircraft fuselage defect detection using deep neural networks. In: The IEEE global conference on signal and information processing. 2017, http://dx.doi.org/10.48550/arXiv.1712.09213, arXiv:1712.09213.
[51] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Commun ACM 2017;60(6):84–90. http://dx.doi.org/10.1145/3065386.
[52] Huang H, Hu C, et al. Surface defects detection for mobilephone panel workpieces based on machine vision and machine learning. In: IEEE international conference on information and automation. ICIA, 2017, p. 370–5. http://dx.doi.org/10.1109/ICInfA.2017.8078936.
[53] Shipway NJ, Huthwaite P, et al. Performance based modifications of random forest to perform automated defect detection for fluorescent penetrant inspection. J Nondestruct Eval 2019;38(2):37. http://dx.doi.org/10.1007/s10921-019-0574-9.
[54] Chen Z-H, Juang J-C. AE-rtisnet: Aeronautics engine radiographic testing inspection system net with an Improved Fast Region-based convolutional neural network framework. Appl Sci 2020;10(23):8718. http://dx.doi.org/10.3390/app10238718.
[55] Redmon J, Divvala S, et al. You only look once: Unified, real-time object detection. 2016, http://dx.doi.org/10.48550/arXiv.1506.02640.
[56] Hu Y, Wang J, et al. Automatic defect detection from X-ray scans for aluminium conductor composite core wire based on classification neutral network. NDT E Int 2021;124:102549. http://dx.doi.org/10.1016/j.ndteint.2021.102549.
[57] Rabiner LR, Juang BH. An introduction to hidden Markov models. IEEE ASSP Mag 1986;12. http://dx.doi.org/10.1109/MASSP.1986.1165342.
[58] Kraljevski I, Duckhorn F, et al. Machine learning for anomaly assessment in sensor networks for NDT in aerospace. IEEE Sens J 2021;21(9):11000–8. http://dx.doi.org/10.1109/JSEN.2021.3062941.
[59] Niccolai A, Caputo D, et al. Machine learning-based detection technique for NDT in industrial manufacturing. In: Mathematics, Vol. 9. MDPI; 2021, p. 1251. http://dx.doi.org/10.3390/math9111251, (11).
[60] Siljama O, Koskinen T, et al. Automated flaw detection in multi-channel phased array ultrasonic data using machine learning. J Nondestruct Eval 2021;40(3):67. http://dx.doi.org/10.1007/s10921-021-00796-4.
[61] Fakih MA, Chiachío M, et al. A Bayesian approach for damage assessment in welded structures using lamb-wave surrogate models and minimal sensing. NDT E Int 2022;128:102626. http://dx.doi.org/10.1016/j.ndteint.2022.102626.
[62] Le M, Luong VS, et al. Auto-detection of hidden corrosion in an aircraft structure by electromagnetic testing: A machine-learning approach. Appl Sci 2022;12(10):5175. http://dx.doi.org/10.3390/app12105175, MDPI.
[63] Risheh A, Tavakolian P, et al. Infrared computer vision in non-destructive imaging: Sharp delineation of subsurface defect boundaries in enhanced truncated correlation photothermal coherence tomography images using K-means clustering. NDT E Int 2022;125:102568. http://dx.doi.org/10.1016/j.ndteint.2021.102568.
[64] Internal Study: University of Augsburg, Detection of anomalies in ultrasonic images of fibre-metal-laminate skin fields, DLR Augsburg, (Unpublished).
[65] Ucan H, Apmann H, Grassel G, Krombholz C, Fortkamp K, Nieberl D, Ehmke F, Nguyen C, Akin D. Produktionstechnologien für leichtbaustrukturen aus faser-metall-laminaten im flugzeugrumpf. Deutscher Luft- und Raumfahrtkongress; 2017, https://elib.dlr.de/114906/ https://www.researchgate.net/publication/321964549.
[66] Zapp P, Pantelelis N, Ucan H. The way to decrease the curing time by 50% in the manufacturing of structural components using the example of FML fuselage panels. In: SAMPE Europe conference. 2019, https://elib.dlr.de/130943/.
[67] Wanhill RJH. GLARE: A versatile fibre metal laminate (FML) concept. 2017, http://dx.doi.org/10.1007/978-981-10-2134-3_13.
[68] Etr HE, Korkmaz ME, Gupta MK, Gunay M, Xu J. A state-of-the-art review on mechanical characteristics of different fibre metal laminates for aerospace and structural application. In: International journal of advanced manufacturing technology, Vol. 123. Springer; 2022, p. 2965–91. http://dx.doi.org/10.1007/s00170-022-10277-1.
[69] Stone M. Cross-validatory choice and assessment of statistical predictions. J R Stat Soc Ser B Stat Methodol 1973;36(2):111–33. http://dx.doi.org/10.1111/j.2517-6161.1974.tb00994.x.
[70] Berrar D. Cross-validation. In: Encyclopedia of bioinformatics and computational biology, Vol. 1. Elsevier; 2018, p. 542–5. http://dx.doi.org/10.1016/B978-0-12-809633-8.20349-X, Elsevier.
[71] Jarvis R, Cawley P, Nagy PB. Performance evaluation of a magnetic field measurement NDE technique using a model assisted probability of detection framework. NDT E Int 2017;91:61–70. http://dx.doi.org/10.1016/j.ndteint.2017.06.006.
[72] Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett 2006;27(8):861–74. http://dx.doi.org/10.1016/j.patrec.2005.10.010.
[73] Zhou W, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: From error visibility to structural similarity. IEEE Trans Image Process 2004;13(4):600–12. http://dx.doi.org/10.1109/TIP.2003.819861.
[74] Georgiou GA. PoD curves, their derivation, applications and limitations. Insight 2006;49:409–14. http://dx.doi.org/10.1784/insi.2007.49.7.409.
[75] Harding CA, Hugo GR. Guidelines for interpretation of published data on probability of detection for non-destructive testing. 2011, p. 31, https://apps.dtic.mil/sti/pdfs/ADA398282.pdf.
[76] Matzkanin GA, Yolken HT. Probability of Detection (POD) for nondestructive evaluation. NDE, Defense Technical Information Center; 2001, http://dx.doi.org/10.21236/ADA398282.
[77] Sause MGR, Jasiuniene E. Structural health monitoring damage detection systems for aerospace. Cham: Springer International Publishing; 2021, http://dx.doi.org/10.1007/978-3-030-72192-3.
[78] Zolfaghari A, Kolahan F. Reliability and sensitivity of visible liquid penetrant NDT for inspection of welded components. Mater Test 2017;59(3):290–4. http://dx.doi.org/10.3139/120.111000.
[79] Tschöke K, et al. Feasibility of model-assisted probability of detection principles for structural health monitoring systems based on guided waves for fibre-reinforced composites. IEEE Trans Ultrason Ferroelectr Freq Control 2021;68(10):3156–73. http://dx.doi.org/10.1109/TUFFC.2021.3084898.
[80] Silva RR da, Padu GX de. Nondestructive inspection reliability: State of the art. In: Nondestructive testing methods and new applications. InTech; 2012, http://dx.doi.org/10.5772/37112.
[81] Schnars U, Kück A. Application of POD analysis at airbus. In: 4th European-american workshop on reliability of NDE. 2009, https://www.ndt.net/?id=8320.
[82] Topp M, Strothmann L. How can NDT 4.0 improve the probability of detection (POD)? e-J Nondestruct Test (NDT) 2021;26(4). https://www.ndt.net/?id=26013.