Transient Thermography for Flaw Detection in Friction Stir Welding: A machine learning approach

Mohamed Atwya and George Panoutsos

Abstract—**A systematic computational method to simulate and detect sub-surface flaws, through non-destructive transient thermography, in aluminium sheets and friction stir welded sheets is proposed. The proposed method relies on feature extraction methods and a data-driven machine learning modelling structure. In this work, we propose the use of a multi-layer perceptron feed-forward neural-network with feature extraction methods to improve the flaw-probing depth of transient thermography inspection. Furthermore, for the first time, we propose Thermographic Signal Linear Modelling (TSLM), a hyper-parameter-free feature extraction technique for transient thermography. The new feature extraction and modelling framework was tested with out-of-sample experimental transient thermography data and results show effectiveness in sub-surface flaw detection of up to 2.3 mm deep in aluminium sheets (99.8 % true positive rate, 92.1 % true negative rate) and up to 2.2 mm deep in friction stir welds (97.2 % true positive rate, 87.8 % true negative rate).**

Index Terms—**Friction-stir welding, Non-destructive testing, Infrared thermal imaging, Image processing, Transient thermography, Artificial neural-network, Machine learning.**

I. Introduction

Aluminium alloys have continuously fulfilled the rising demand for lightweight large-scale structures and have become widely utilised in several sectors including the aviation, rail, and marine industries [1]. The increasing industrial utilisation of aluminium (AL) along with the rising demand for reduced manufacturing costs resulted in a driving force towards finding a viable, cost-effective AL joining technology such as friction stir welding.

Friction stir welding (FSW) of AL alloys [2] has a low incidence of flaws when compared to conventional AL joining technology [3]. However, operating the FSW process outside its operating envelope can introduce surface and subsurface flaws [4]. Therefore, demand has risen for a high-speed non-destructive testing technique for friction stir (FS) welds and AL alloys [5]. Non-destructive testing (NDT) techniques such as eddy current testing [6], ultrasonic testing [5], x-radiography (ray) [7], and infrared thermography [8], have been reported for the detection of FS weld voids, wormholes, root flaws, and lack of penetration.

Eddy current testing methods commonly cannot penetrate large distances in the material (a maximum of a few millimetres) and are suited to micrometre-level superficial flaws [6]. Conventional and phase array ultrasonic testing methods are limited by the high sensitivity on the coupling (water, gel coat, etc.) conditions, acoustic attenuation, and the flaw detectability limit related with the test wavelength [9]. Additionally, eddy current and ultrasonic methods often require contact with the specimen and are limited to small inspection areas.

The x-ray method is suited for the detection of flaws that cause a significant difference of radiation absorption (e.g., millimetre-level internal voids) [9], [10]. Similarly, for infrared thermography, the smallest detectable flaw for an isotropic material, should have a diameter of at least 2 times its depth below the surface, or up to a factor of 10 for anisotropic materials [11].

Most studies in the field of NDT of FS welds for deeper and larger flaws have focused on the use of x-ray. For example, the authors in [7] proposed the use of x-ray with an image enhancement methodology to qualitatively assess dissimilar FS welded lap joints of an AL 6061 $2.0\,\mathrm{mm}$ thick sheet and a zinc-coated steel sheet of $1.0\,\mathrm{mm}$ thickness. The work in [12] successfully employ computed tomography (CT) to detect wormhole flaws amongst other flaws in FS welded lap joints of AL alloys 6061-T6 and 1050 $1.59\,\mathrm{mm}$ thick sheets. However, the authors found that due to the limited resolution of the CT equipment, the CT estimated wormhole flaw cross-sectional area was approximately three times larger than the results of a destructive test ($10.9255\,\mathrm{mm}^2$ compared to $3.4345\,\mathrm{mm}^2$) [12].

The application of infrared thermography non-destructive testing (TNDT) to FS weld flaw detection is limited to [8], where the authors used lock-in thermography and adaptive single plateau based histogram equalisation to detect a sub-surface wormhole flaw (actual size 60.0 by $0.55\,\mathrm{mm}$ and estimated depth $2.79\,\mathrm{mm}$) in a FS welded butt joint of an AL 6061 $3.0\,\mathrm{mm}$ thick sheet. The literature on the application of NDT indicates that detecting FS welding subsurface flaws at high-speeds and in an automated framework is still an ongoing research topic. The following Section II will discuss FSW-relevant applications and limitations of thermography in NDT and lead onto the proposed methodology. This work aims to improve the signal-to-noise-ratio (SNR) and flaw probing depth of TNDT for flaw detection and localisation in FS welds using a systematic semi-automated computational methodology. The contributions of this paper are as follows:

- A hyper-parameter-free feature extraction technique, Thermographic Signal Linear Modelling (TSLM), to improve the SNR and flaw probe depth of transient thermography data (Section III-A3).
- A semi-automated transient thermography flaw detection framework involving neural-network (NN) machine learning along with existing feature extraction techniques and TSLM, to improve the SNR and flaw probe depth in AL sheets and FS welds (Section III).

This work was supported by The University of Sheffield.

M. Atwya and G. Panoutsos are with the University of Sheffield, S1 4DW, UK (e-mail: matwya1@sheffield.ac.uk, g.panoutsos@sheffield.ac.uk).

- A quantitative comparison of three widely used state-of-the-art thermography feature extraction methods for flaw detection in FS welds, thus offering a rigorous understanding of the effectiveness of the proposed method and a guide for future developments (Section IV-C).

The remainder of the paper is organized as follows. The Methodology section provides the flaw detection approach including data pre-processing, feature extraction, data preparation, and NN-based data-driven modelling. Subsequently, in the Experimental Results and Discussion section, the methodology is experimentally validated and discussed using AL FS weld and AL sheet specimens with artificial flaws. The paper is concluded in the last Section.

II. Thermography Non-Destructive Testing

The theoretical principle of TNDT is based on the fact that the structure being inspected and its flaws will have different thermal behaviours (diffusivity and effusivity) [13]. Thermal diffusivity is a measure of the thermal energy diffusion rate through a material (i.e. ratio of the thermal conductivity to the volumetric heat capacity). Thermal effusivity (thermal inertia) is the square root of the product of the thermal conductivity and the volumetric heat capacity. The thermal inertia governs how much a structure's temperature changes as a result of a thermal energy input.

When a structure has voids, its thermal conductivity and density decrease and its thermal diffusivity changes. The change in thermal diffusivity results in observable changes of surface temperatures in the vicinity of the flaws [11]. Similarly, if a subsurface flaw in a structure has a different temperature than its surroundings, the observed surface temperatures will be affected as a result of thermal inertia.

If an in-homogeneous structure is subjected to heating energy, thermal diffusion in the structure can propagate faster in the region of larger voids, since heat must only diffuse through a thinner layer of material [11]. Therefore, the observable top surface temperatures at voids are increased relative to the neighboring flaw-free areas of the surface. This temperature difference (thermal contrast) evolves as a function of time and can be captured by an infrared (IR) camera.

TNDT methods can be classified into passive and active thermography. In active infrared thermography, the acquisition is carried out during the application of an external excitation (energy) supply which produces a controlled change in the specimen's surface temperature. The most common excitation source is the use of optical techniques [11], [14]. Optical stimulation generates heat, which propagates as thermal waves to the surface and through the material of the specimen. When the thermal waves reach a flaw, their propagation rate is altered, resulting in a thermal contrast on the surface immediately above the flaw [15].

Optical-based active infrared thermography can be classified into pulsed, transient, and lock-in categories based on the optical excitation source and its controller. Optical excitation is applied via flash lamps for pulse heating and halogen lamps for transient and lock-in heating. Typically, experimental setups for transient heating use one to four halogen lamps with the total exaction energy ranging between 1000 and $4000\,\mathrm{W}$ [14].

Raw thermograms from TNDT are rarely suitable for quantitative analysis due to a weak contrast in the thermal signals [16]. However, feature extraction techniques can be applied to thermography data to improve the SNR, making flaw detection possible [17]. For example, Almond et al., applied transient thermography to $6\,\mathrm{mm}$ thick AL sheets with flat bottom holes (FBHs) of different depth. The authors provided an analytic study of the transient heating process, revealing that the transient excitation technique is unsuitable for testing materials with high thermal conductivity such as AL due to the flaws having low thermal contrast in the IR images [18]. However, through the application of thermography non destructive evaluation (TNDE) methods and machine learning, TNDT is proven to be a reliable evaluation technique for large-scale high-speed flaw detection in other applications [19].

In [20], transient thermography raw data was used as the input to train a multi-layer NN, which was then capable of detecting 0.25 and $0.5\,\mathrm{mm}$ deep FBHs in a $1.0\,\mathrm{mm}$ thick AL sheet. Albendea et al., used pulsed thermography on gas tungsten arc welds of $1.0\,\mathrm{mm}$ thick stainless steel sheets containing flaws such as lack of penetration and perforation [21]. The authors tested feature extraction techniques including, skewness, kurtosis, principal component thermography (PCT), and pulsed phase thermography. It was found that there is no universal method to detect all the flaw types.

The study in [22] employed an autoencoder algorithm with inductive thermography data to improve the detectability of rear surface cracks on steel sheets. Jang et al., use data from vision and laser IR thermography along with deep convolutional neural networks (CNN) to improve concrete crack detectability [23]. In [24], the authors use principal component analysis (PCA) along with Faster-Region CNN to improve the crack detection rate in stainless steel and steel specimens from eddy current pulsed thermography data.

TNDT is an attractive technique for being a non-contact, rapid, wide-area inspection method [25]. However, the application of TNDT to metals is often limited by low SNR and low subsurface flaw probe depth. Accordingly, the following section will introduce our proposed methodology towards improving the SNR and flaw probing depth of TNDT for AL sheets and FS welded AL.

III. Methodology

In this work we propose the use of a multi-layer perceptrons (MLP) feed-forward NN with feature extraction methods to improve the SNR and flaw probing depth of transient thermography inspection for AL sheets and FS welds. We propose TSLM, as a hyper-parameter-free feature extraction technique for transient thermography data. A flowchart of the proposed framework is presented in Fig. 1, and includes data pre-processing, feature extraction, data preparation, and data modelling steps.

The pre-processing steps include converting the raw RGB images into $64\,$bit grayscale images, cropping the region of interest (ROI), and applying a noise filter. As this work aims to estimate flaw characteristics (detection and localisation) as opposed to quantitative measures, we have chosen to use raw RGB data rather than temperature values as encouraged in [14]. Note that it is estimated that flaw detection and localisation is sufficient in $80\,\%$ of inspection situations [14].

In most applications, raw IR images contain information about the targeted object, but also about its background. The region of the image with the target object is referred to as the ROI. The ROI needs to be identified in the image to determine its grayscale readings and carry out further processing steps. Moreover, it is necessary to remove the image background as the use of spatial mathematical operations in the following processing steps may result in the degradation of useful ROI readings. For example, in the following Gaussian image filtering step, the ROI grayscale readings would be averaged with the adjacent background readings, resulting in erroneous data. Therefore, the ROI was manually cropped from the recorded grayscale images (Fig. 2a and 2c). Note that cropping the ROI is the only manual step in the proposed framework.

The noise filter was either subtraction or Gaussian filtering, depending on the feature extraction method. The subtraction filter was applied via subtracting the image sequence from the 30^{th} frame of the excitation period (i.e. the frame at $1\,\mathrm{s}$ of the excitation period). For Gaussian filters, in general, a larger kernel standard deviation (with the corresponding appropriate kernel size) will have the effect of increasing the smoothness of the images (dilation) and increasing the contrast between flaw and flaw-free pixels. Accordingly, a 13 by 13 kernel, with a standard deviation of 3.0 was empirically chosen for both specimens. We chose to apply the same filter parameters to both specimens, for consistency and for allowing a fair comparison between the flaw detection performance of the feature extraction methods on the two different specimens. Note that increasing the image smoothness (i.e. filter standard deviation term) also results in flaws appearing larger than their true size (i.e. dilation). Therefore, the kernel standard deviation must not be too large relative to the flaw sizes in a given investigation. Note that we will refer to the sequence of Gaussian filtered images as $G(j,i,t)$, where G is the image at time t and j and i are the column and row numbers, respectively. The partial derivative of the Gaussian filtered data with respect to time (slope) will be referred to by $\frac{\partial G}{\partial t}(j,i,t)$.

The pre-processed three-dimensional data was then converted into a two dimensional N by P matrix; Gaussian filtered data (A_G) and subtraction filtered data (A_s). Where the rows (N) are the number of pixels in the IR image and contains the spatial variations. The columns (P) contain the temporal variations and are the number of IR images recorded.

A. Feature Extraction

TNDE methods such as feature extraction are necessary to improve the flaw detection and characterisation performance and to automate the inspection process of active thermography experiments [17]. Some of the commonly used feature extraction techniques used in TNDE are statistical moments, PCT, pulsed-phase thermography, and thermographic signal reconstruction (TSR). In this work, statistical moments and PCT are used (Sections III-A1 and III-A2). Furthermore, we propose TSLM (Section III-A3).

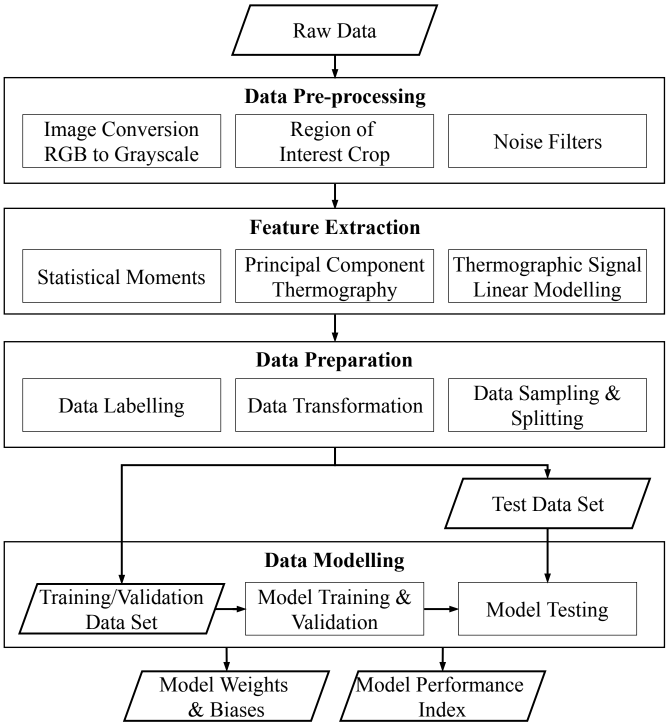


Fig. 1. Flowchart of the proposed flaw detection methodology.

1) Statistical Moments: The standardised statistical moments, skewness and kurtosis, have been utilised in TNDE literature to provide higher signal contrast levels relative to raw data [16], [26]. The skewness and kurtosis techniques can be seen as a process of compressing the pre-processed data to two diagrams, the skewgram (Fig. 8c) and the kurtogram (Fig. 8d). The skewgram and kurtogram were computed from the subtraction filtered data (A_s) as in [17], and were then scaled to the range $[0, 255]$.

2) Principal Component Thermography: PCT is the application of the PCA technique in TNDE. PCT aims to firstly, reduce the number of variables in a data-set, while preserving the most amount of information possible, and, secondly, to highlight the similarities/dissimilarities in the data [17]. We implemented PCT on the the Gaussian filtered data (A_G) as in [27]. The matrix A_G was normalised by subtracting the mean image (i.e. the averaged row) from each row to give matrix $\hat{A_G}$. Following the normalisation, the matrix $\hat{A_G}$ was decomposed to yield the eigenvalues and eigenvectors. The set of orthogonal empirical functions (EOF) basis for the range space of $\hat{A_G}$ was computed. The PCT images were produced by reforming each EOF into respective matrices. The PCT images were then scaled to the range $[0, 255]$.

Each EOF accounts for a percentage (ρ_C^2) of the total variability in the data. The first EOF accounts for the majority of the variability in the data and the following EOFs represent less variability in descending order. However, note that ρ_C^2 does not directly reflect on the contrast between flaw and flaw-free pixels, as can be seen in Fig. 10a. The number of EOF basis to use is a hyper-parameter and is application dependent. For example, in the context of pulsed TNDE, the first two EOF basis typically provide an adequate description of the relevant spatial variations [27], [28]. However, from the transient TNDE experiments performed in this study, it was empirically found that the first four EOFs and their corresponding PCT images (PCT_{1-4}) provide an adequate description of the relevant spatial and temporal variations (Fig. 8e to 8h).

3) Thermographic Signal Linear Modelling: This work proposes TSLM, a hyper-parameter-free TNDE data reduction processing technique, for transient thermography. TSLM aims to capture the effects of thermal diffusivity and effusivity on the surface temperatures (grayscale values) in the vicinity of the flaws in a structure. TSLM compresses the information from a transient thermography image sequence to four images while preserving both spatial and temporal information. TSLM is based on the same concept used in TSR, a processing technique designed for pulse thermography, primarily used in data reduction and noise filtering.

As detailed in Section II, when a non-homogeneous structure is subjected to heating energy, thermal diffusion in the structure can propagate faster in the region of larger voids, since heat must only diffuse through a thinner layer of material. Consequently, the observable top surface temperatures at voids will have a larger magnitude and a larger rate of change [11], [29]. Therefore, it is hypothesised that linear in the parameter (LIP) univariate polynomial modelling of experimental thermograms ($G(j,i,t)$) and their temporal-derivatives ($\frac{\partial G}{\partial t}(j,i,t)$) are sufficient to capture the differences between flaw and flaw-free areas on a surface. The proposed TSLM technique consists of the two following independent modelling procedures:

- Modelling, for each pixel (j,i), the Gaussian filtered grayscale value as a function of time, $G(j,i,t)$, by a LIP univariate polynomial function (Eq. 1).
- Modelling, for each pixel (j,i), the slope of the Gaussian filtered grayscale values as a function of time, $\frac{\partial G}{\partial t}(j,i,t)$, by a LIP univariate polynomial function (Eq. 2).

$$G(j,i,t) \approx a_0(j,i) + a_1(j,i)t, \tag{1}$$

where $a_0(j,i)$ and $a_1(j,i)$ are the polynomial coefficients that approximate the value of $G(j,i,t)$ at pixel (j,i) and time t.

$$\frac{\partial G}{\partial t}(j,i,t) \approx b_0(j,i) + b_1(j,i)t, \tag{2}$$

where $b_0(j,i)$ and $b_1(j,i)$ are the polynomial coefficients that approximate the value of $\frac{\partial G}{\partial t}(j,i,t)$ at pixel (j,i) and time t.

The modelling in Eq. 1 is used to replace the three-dimensional Gaussian filtered grayscale data ($G(j,i,t)$) by two images formed by the polynomial coefficients: $a_0(j,i)$ and $a_1(j,i)$ (Fig. 8i and 8j). Similarly, the modelling in Eq. 2 replaces the full sequence of grayscale values' slope images ($\frac{\partial G}{\partial t}(j,i,t)$) by two images of the polynomial coefficients: $b_0(j,i)$ and $b_1(j,i)$ (Fig. 8k and 8l).

The iterative re-weighted least squares (IRLS) algorithm is used to set the weights and biases (a_1, b_1, a_0, and b_0). The IRLS algorithm stopping criteria are the precision's of the objective function and weights at the solution, which were set to 10^{-3}. TSLM images are then produced by reforming the weights and biases (a_1, b_1, a_0, and b_0) into respective matrices TSLMa_1, TSLMb_1, TSLMa_0, and TSLMb_0. Finally, the four TSLM images are scaled to the range $[0, 255]$ (Fig. 8i-8l). Section IV-B provides a case study demonstrating the application of TSLM.

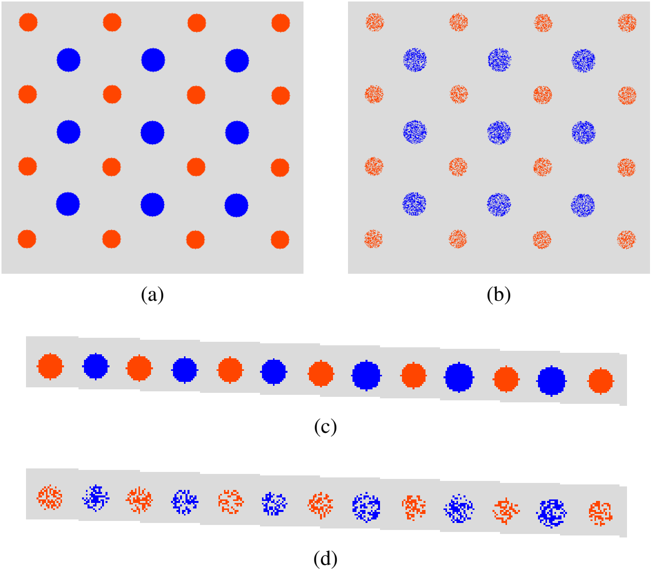


Fig. 2. Data points from the experiments considered in the modelling, where the orange and blue colours correspond to flaw-free and flaw data points, respectively; (a) AL sheet data points, (b) AL sheet Training/validation data set, (c) FS weld data points, and (d) FS weld Training/validation data set.

B. Data Preparation

The feature extraction methods produce ten features. Following the feature extraction step, the data preparation process aims to make the features date-set representative for supervised data-driven modelling. The data preparation process includes data labelling, transformation, sampling, and splitting.

The location and size of the flaws were used to classify and label the data. However, flaws appear blurry and therefore larger than their true size in TNDT images [30]. Therefore, the flaw sizes were increased by 20.0 % during the labelling process. The 20.0 % increase was chosen empirically. Fig. 2a and 2c shows the flaws after increasing their diameter by 20.0 %.

Data transformation was used to reduce the Euclidean distances between features so that all the features contribute proportionally in training the model. The input data features (column-wise) were transformed independently such that each feature had a zero mean and unit standard deviation (z-score standardisation).

The data in this study has an imbalanced class distribution as the majority of the specimens' surface area are flaw-free (95.0 % AL sheet and 88.9 % FS weld). Therefore, only a selected number of flaw-free pixels were considered, such that both classes approximately have the same size (as shown in Fig. 2a and 2c). The sampled data class distribution is:

1) AL sheet: 57747 pixels, 49.0 % flaw, and 51.0 % flaw-free.

2) FS weld: 6243 pixels, 50.0 % flaw, and 50.0 % flaw-free.

The flaw-free locations for the AL sheet were chosen to be equally distributed across the specimen, as the specimen's entire surface area approximately received equal excitation energy. The flaw-free locations for the FS weld were chosen to be adjacent to the flaw locations so that they are on the ROI (the weld) and not on the AL sheet which has different thermal properties. Furthermore, the flaw-free locations on the FS weld were chosen to be close to the flaw locations to ensure both received approximately equal excitation energy. Fig. 2a and 2c demonstrate the pixels used from the two specimens.

Before training a data-driven model, it is necessary to arrange a method to test its performance. To quantitatively analyse how well the model predicts the presence of flaws in both specimens, the data was split, and a portion of it was reserved for testing after the models have been trained. The reserved portion of the data will be referred to as the out-of-sample data (OOS). Before splitting the data, the input and target data row order (i.e. spatial-wise) for each specimen was randomised. It was chosen to split the data in half such that as much data as possible is used to train the model and equally as much data is used to test the model and be confident with its predictions:

1) AL sheet: 28873 pixels in-sample and 28874 pixels OOS.
2) FS weld: 3121 pixels in-sample and 3122 pixels OOS.

The OOS data was reserved and only used to test the models after they were developed. The resulting in-sample data set from two of the experiments are shown in Fig. 2b and 2d.

C. Data-driven Modelling

In this Section we propose using the extracted features as the inputs of a data-driven model in order to improve the predictions of a pixel's class. As labelled training data is available and there is a relationship between the input (extracted features) and target values (flaw/flaw-free class), supervised data-driven learning methods were used.

The classification problem is a Bernoulli distributed problem with a nominal dichotomous target $[0, 1]$, where 0 indicates flaw-free and 1 indicates a flaw. The problem also presents a non-linear input/target data relationship. Accordingly, it is necessary to use a universal approximator that is flexible enough to accommodate the non-linear characteristics in the data, and that mathematically yields an output in the range $[0, 1]$. Therefore, a non-linear in the parameter MLP model with a non-linear input/output relation is utilised.

1) Multi-layer Network Model: MLPs are universal approximator models, and accordingly, a two-layer feed-forward MLP with enough hidden units can model any continuous function on a finite interval, given there is enough data to estimate the network weights (Stone-Weierstrass theorem) [31]. The MLP models developed in this study (a model for each specimen) are feed-forward NNs. Each network constitutes of ten input neurons ($d = 10$), one hidden layer with a non-linear activation function (Eq. 3), and an output layer with one neuron and a logistic activation function (Eq. 4 and 5). The hidden layer was chosen to have 10 hidden units ($m = 10$) for both models. The choice to use 10 hidden units was to ensure that both models have an overly flexible structure to capture all the data non-linearities (i.e. over-fit the data) and then regularisation was used for complexity control (i.e. weight optimisation) (Section III-C2). The following subsections will discuss how the optimal MLP model weights were found and validated.

$$a_j^{(1)} = \tanh\left(\sum_{i=1}^{d}\left(x_i w_{ji}^{(1)} + b_j^{(1)}\right)\right) \quad : \quad j = 1, \ldots, m, \tag{3}$$

where $a_j^{(1)}$ is the output at the hidden unit j, x_i are the inputs, $w_{ji}^{(1)}$ are the weights of the first hidden-layer, and $b_j^{(1)}$ are the biases of the first hidden-layer.

$$a^{(2)} = b^{(2)} + \sum_{j=1}^{m} a_j^{(1)} w_j^{(2)}, \tag{4}$$

where $a^{(2)}$, $w_j^{(2)}$, and $b^{(2)}$ are the output layer activation value, weights, and bias of the second hidden-layer, respectively.

$$y = \frac{1}{1 + \exp\left(-a^{(2)}\right)}, \tag{5}$$

where y is the model output.

2) Weight Optimisation and Complexity Control: The weights of each MLP model were initialised randomly. To find the optimal weights, a cross-entropy cost (CC) function, J_{emp}, was chosen as the performance index (PI) (Eq. 6). The PI ($Jemp$) was chosen as it is less sensitive to data outliers and is also suitable for Bernoulli distributed classification problems [32]. The MLP model structure was chosen to have high complexity (over-fitting), and thus would fit the in-sample data well, but be erratic between the in-sample data points. Therefore, large weights in the model must be penalised to eliminate the over-fitting and find the optimal model weights. L2 regularisation was employed to penalise the large weights. An additional regularisation cost (Eq. 7) was added to the CC function, as shown in Eq. 8.

$$J_{emp} = -\sum_{i=1}^{i=n}\left(z_i \ln(y_i) + (1 - z_i)\ln(1 - y_i)\right), \tag{6}$$

where n is the number of data points and y_i and z_i are the prediction and target for the i^{th} input vector.

$$J_{reg} = \frac{1}{2}\underset{\sim}{w}^T \underset{\sim}{w}, \tag{7}$$

where $\underset{\sim}{w}$ is a vector of the model weights.

$$J_{tot} = J_{emp} + \rho J_{reg} \quad : \quad \rho \geq 0, \tag{8}$$

where ρ is the regularisation parameter, and J_{emp}, J_{reg}, and J_{tot} are the empirical, regularisation, and total costs, respectively.

As the regularisation parameter, ρ, is multiplied with the square of the weight vector, a large ρ increases smoothness while a small ρ increases flexibility. This procedure is known as L2 regularisation. Since ρ is a scalar, it is feasible to perform

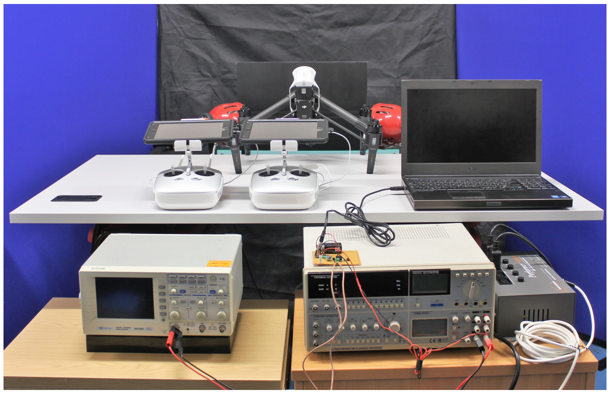


Fig. 3. The data acquisition experimental setup used to perform transient thermography (showing the FS weld specimen).

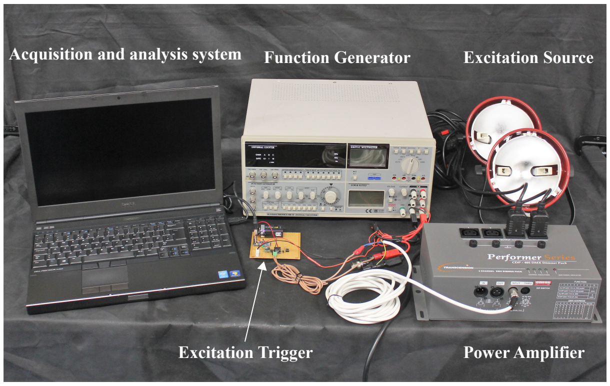


Fig. 4. The excitation source control system used to perform the transient thermography experiments including: halogen lamps, power amplifier, excitation trigger board, function generator, and a data acquisition system.

TABLE I
AL SHEET ARTIFICIAL FLAWS' SIZE AND DEPTH.

Flaw Number	$A1, A2, A3$	$B1, B2, B3$	$C1, C2, C3$
Diameter (mm)	15.0	15.0	15.0
Depth (mm)	$0.2, 0.5, 0.8$	$1.1, 1.4, 1.7$	$1.9, 2.1, 2.3$

TABLE II
FS WELD ARTIFICIAL FLAWS' SIZE AND DEPTH.

Flaw Number	$P1$	$Q1$	$R1$	$S1$	$T1$	$U1$
Diameter (mm)	5.0	5.0	5.0	5.8	5.8	5.8
Depth (mm)	0.3	2.2	1.2	0.3	2.2	1.2

a one-dimensional search across a vector of different ρ values and assess the PI on new data for each optimised weights set. The model with the best-estimated PI was then chosen. Section III-C3 will discuss how the new data was obtained.

Since ρ can have any positive value, a vector of 10 logarithmically spaced values between 10^{-2} and 10^{2} was tested for each specimen to narrow down the search. For the AL sheet the smallest PI $= 169.954$ was achieved at $\rho = 0.077$, and for the FS weld, the smallest PI $= 616.430$ was achieved at $\rho = 0.215$. Accordingly, a finer search across a ρ vector of 20 logarithmically spaced points between 10^{-2} and 10^{0} was tested for both specimens. Finally, the best $\rho = 0.070$ with a PI $= 1375.799$ and $\rho = 0.113$ with a PI $= 604.578$ were used to retrain the AL sheet and FS weld models, respectively.

The scaled conjugate gradient (SCG) back-propagation method was used to minimise the cost function (Eq. 8) and find the optimal weights. The SCG algorithm was chosen as it does not contain any user-dependent parameters that are critical to the algorithm's success and it also uses a step size scaling mechanism to avoid time-consuming line search per learning iteration. The SCG optimisation algorithm has two stopping criteria: the precision of the objective function and the weights at the solution. The two stopping criteria were set to 10^{-3}. The algorithm was allowed a sufficient number of iterations to find the solution. However, the cost function is non-linear and multi-modal, and thus the MLP's initial random weights determine the SCG solution. Therefore, the SCG solution is a local minimum, and the algorithm does not guarantee to find the global minimum. Thus, repeated k-fold cross-validation (CV) was used to improve the chances of locating the global minimum (Section III-C3).

3) Model Validation: The input data (features) presents an 11-D space with 10-D hyper-planes. Thus, the classification problem is high-dimensional, and the available data is sparse, particularly in the FS weld specimen which has a smaller sample size. However, a reliable data model must be developed and validated with strong empirical dependence, which is difficult under the conditions of sparse data and a small sample size [33]. Therefore, k-fold CV and data stratification were employed to make the best use of the sparse sample data available and to address the local-minima problem. As the sample data is sparse, $k = 10$ was chosen to reduce the bias in the PI estimate.

A stratification technique (MATLAB's `cvpartition` function) was utilised to ensure that the subsets roughly have equal statistical size even though they are chosen randomly. Finally, repeated 10-fold CV was employed, where for every set of ρ value and initial weights, the 10-fold CV procedure is iterated ten times using re-divided and stratified subsets. This extra measure eliminates the potential of misleading results due to subsets dominated by bias and outliers.

4) Model Testing: After the model has been trained and validated, the reserved OOS data was used to test the models' performances. During the testing phase, the OOS data targets (data labels) were hidden from the models and the models were tested using the OOS input data, as if the data was acquired from an untested specimen. The OOS data labels were then used at the end of the process to assess the performance of the models prediction (Section IV-D).

IV. Experimental Results and Discussion

A. Experimental Setup and Procedure

The thermography experimental setup is show in Fig. 3 and 4. Two specimens were tested in this work; an AL sheet (series 1000, 265.0 by 205.0 by 2.5 mm) and a FS welded AL sheet with a 12.0 mm wide FS weld zone (series 6000, 490.0 by 396.0 by 6.0 mm). The AL sheet specimen was used as a benchmark to develop and calibrate the experimental setup and procedure.

Artificial void flaws (FBHs) with varying diameters and depths were machined on the specimens with the dimensions provided in Tables I and II. The FBHs diameter for the AL sheet was chosen to be $15.0\,\mathrm{mm}$ as in [34] where the authors similarly used $15.0\,\mathrm{mm}$ FBHs on an AL sheet for calibration purposes. The FBHs depth extremes on the AL sheet were chosen to be at $8.0\,\%$ and $92.0\,\%$ of the specimens thickness, similar to the $10.0\,\%$ and $90.0\,\%$ used in the FBH standard specimen designed by the Infrared research committee of the Japanese Society for Nondestructive Inspection (JSNDI) [35].

FSW flaw geometries and dimensions depend on the tool size, material to be welded, thickness of material, welding parameters such as traverse speed, as well as the FS welding technology used (single side, double side, floating tooling etc.). However, rather than to characterise the effectiveness of the NDT method, the scope of this work is to develop a computational framework that utilises the NDT method's data within a machine learning framework to enhance flaw detection. As such, we use representative FSW flaw sizes, without the intention of being exhaustive or as close to reality as possible (this would not be possible without utilising actual FSW flaws, mainly due to their irregular shapes [8], [36]). Note that a $0.6\,\mathrm{mm}$ thick layer was machined off the FS weld surface before machining the FBHs, to level out the weld's uneven surface. The machining (leveling and FBHs) was performed on the side opposite the welding tool side and the weld inspection was performed from the welding tool side.

The success of TNDE techniques mainly depends on the quality of the raw IR images [37]. For accurate thermographic measurements (i.e. higher SNR), it is preferable to work with surfaces that have high-emissivity ($0 \leq \varepsilon \leq 1$). It is possible to increase the surface emissivity and improve the emissivity uniformity across the surface of metals via the deposition of thin films of paint [38]. Therefore, the sound side (welding tool side in the FS weld) of the two specimens were painted using RS matt black spray paint which has an emissivity of approximately 0.92 [39]. Additionally, a black backdrop was placed behind the specimen to prevent reflections from the background interfering with the IR camera measurements (Fig. 3).

A microbolometer thermal camera (FLIR Zenmuse XT Uncooled) was used to capture RGB image (720 by 480 pixels in JPEG format) sequences at 30 frames per second (FPS). The camera was set to record in following settings: NTSC video format, High Gain Mode, linear scene, rainbow palette, and Manual Flat Field Correction (FFC) calibration. The FFC calibration was set to manual and triggered once before each recording, to prevent mid-recording re-calibration.

Two halogen lamps with a total of $1600.0\,\mathrm{W}$ were used to perform transient infrared thermography experiments in the reflection mode. The lamps were mounted at $45.0\,°$ towards the specimen such that the incidence angle of the thermal waves was about $45.0\,°$. The chosen lamp orientation aids in producing a uniform excitation source such that the specimen receives equal excitation energy across its surface. The lamps were driven by a power amplifier, a function generator, and a programmable excitation trigger (Fig. 4). The excitation trigger acts as a switch between the function generator and the power amplifier and allows switching on the excitation source for a pre-defined duration. The distance between the camera and the specimens was approximately $50.0\,\mathrm{cm}$. The distance between the lamps and the specimens was approximately $30.0\,\mathrm{cm}$ and $25.0\,\mathrm{cm}$ for the AL sheet and FS weld, respectively. The distance between the lamps and the FS weld specimen was chosen to be smaller due to the specimen's larger thickness.

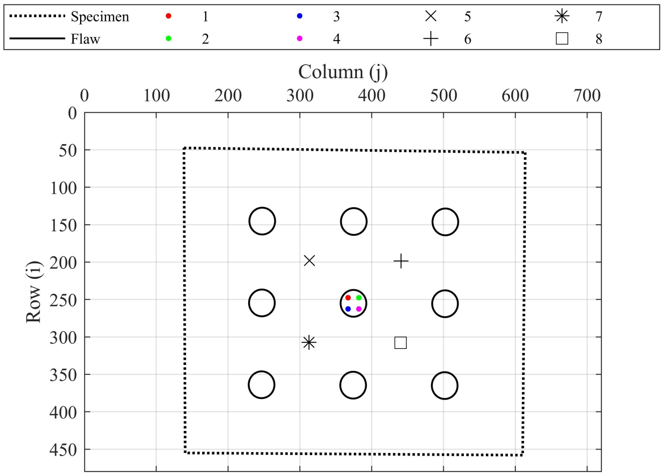


Fig. 5. The locations of the eight points considered in the TSLM case study on the AL sheet specimen, where points 1 to 4 correspond to flaw B2 and points 5 to 8 correspond to flaw-free areas.

The halogen lamps were powered on for approximately $3.0\,\mathrm{s}$ to heat the specimen and then powered off (step heating stimulation). The specimen was recorded during the $3.0\,\mathrm{s}$ excitation period via the IR camera. The specimen was also recorded for $10.0\,\mathrm{s}$ before (ambient period) and $10.0\,\mathrm{s}$ after (cooling-down period) the excitation period, approximately producing $23.0\,\mathrm{s}$ long image recordings. As the excitation and cooling-down periods contain similar information [17], the methodology described in Section III is applied to the $3.0\,\mathrm{s}$ excitation period and the cooling-down and ambient periods are omitted.The room temperature before switching on the lamps was $22.0\,°\mathrm{C}$ for each experiment (rounded to two significant figures).

Three experiments per specimen were performed. The frame sizes (column by row) of the Al sheet specimen after cropping the ROI in the pre-processing step are 412 by 457, 414 by 452, and 414 by 452. The cropped frame sizes of the FS weld specimen are 332 by 39, 334 by 34, and 339 by 32. The size variations are a result of variations in the clarity of the specimen edges in the IR images and the manual cropping process. The two following sections will present and discuss the feature extraction results followed by the data-driven modelling results.

B. TSLM Case Study Results

To demonstrate how TSLM works, we utilise a case study involving eight pixels obtained from a 94 images long recording of the AL sheet specimen. Points one to four correspond to pixels at different locations within flaw B2, and points four to eight corresponds to flaw-free areas surrounding flaw B2 (Fig. 5). The temporal evolution of the grayscale values ($G(j,i,t)$) and the grayscale values rate of change ($\frac{\partial G}{\partial t}(j,i,t)$) of the eight pixels are shown in Fig. 6a and 7a, respectively. The two figures qualitatively demonstrate that the flaw and flaw-free pixels have different grayscale values and grayscale value slopes (Fig. 6a and 7a). Furthermore, the grayscale value temporal evolution shows that all the points initially experienced an increase in the grayscale value followed by a decrease in the value (Fig. 6a).

The results of applying the proposed TSLM algorithm on the case study points are shown in Fig. 6b and 7b. Fig. 6b shows the model estimated grayscale value as a function of time for each of the eight points. Fig. 7b shows the model estimated slope of the grayscale value as a function of time. Note that Fig. 6b and 7b are only to aid with visualisation (i.e. the weights and biases are the results that get passed forward to the NN-based modelling step).

1) Discussion: The TSLM results demonstrate that TSLM successfully captures the difference in grayscale values, grayscale value slopes, and rate of change of the grayscale value slopes between flaw and flaw-free signals (Fig. 6 and 7). Note that the grayscale value, grayscale value slope, and rate of change of the grayscale value slope correlate with the temperature, temperature slope, and rate of change of temperature slope, respectively.

From Fig. 6b, the eight points have a negative gradient characterised by a dominant decrease in the grayscale values during the excitation period. The TSLM model approximated grayscale values of the flaw-free points (four to eight) have larger negative gradients (a_1) and biases (a_0) which is attributed to the grayscale values and grayscale value slopes, respectively. However, the TSLM model approximations of the grayscale value slope for the flaw-free points have smaller negative gradients (b_1) and biases (b_0) which is attributed to the grayscale value slopes and rate of change of the grayscale value slopes, respectively (Fig. 7b).

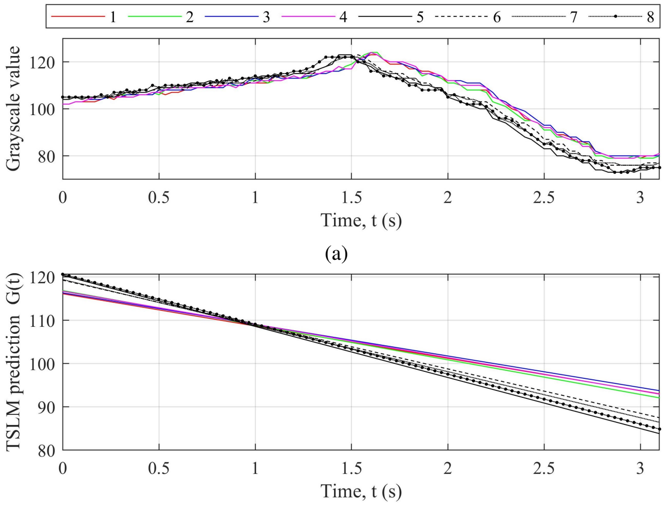
C. Feature Extraction Results

The raw and feature extracted IR images from two of the experiments can be seen in Fig. 8 and 9, for the AL sheet and FS weld, respectively. The images in Fig. 8 and 9 have been produced using MATLAB's `jet` colour-map. Note that Fig. 8a, 8b, 9a, and 9b are only shown for a visual comparison and are not passed forward to the NN-based modelling step. The four images have the highest average contrast (SNR) between flaw and flaw-free areas (Fig. 8a, 8b, 9a, and 9b).

The SNR metric is used to quantitatively assess the detectability of flaws in the raw data and the extracted features [40]. The SNR metric for a flaw is calculated as shown in Eq. 9. The flaw-free area is selected independently for each flaw and is located close to the flaw in question; ensuring that both areas have received approximately equal excitation energy, which minimises non-uniform heating induced errors in the SNR calculation [17]. For the AL sheet, the average of the four closest flaw-free areas to each flaw was used (Fig. 2a). For the FS weld, the average of the two closest flaw-free areas to each flaw was used (Fig. 2c). The SNR values were computed for each data set of the six experiments. The SNR values of each specimen were averaged and are shown in Fig. 10. The figures use a colour scale in which green indicates a good SNR and red is a zero/negative SNR indicating the flaw was undetected.

$$SNR = 20\log_{10}\left(\frac{|Flaw_\mu - Ref\mu|}{Ref_\sigma}\right), \tag{9}$$

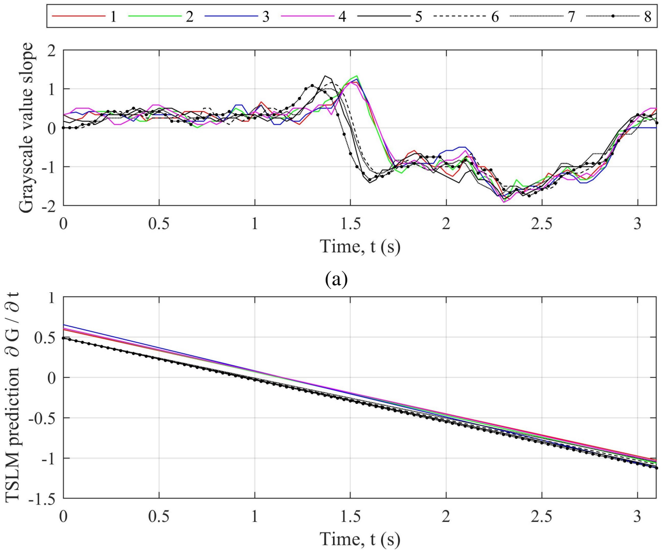
where $Flaw_\mu$ is the mean of the flaw area, $Ref\mu$ is the mean of the flaw-free (reference) area, and Ref_σ is the standard deviation of the flaw-free area.



(a)

(b)

Fig. 6. TSLM case study results on the grayscale temporal evolution: (a) experimental grayscale temporal evolution of the eight points considered and (b) TSLM polynomial fit of the grayscale temporal evolution values.



(a)

(b)

Fig. 7. TSLM case study results on the grayscale temporal slope: (a) experimental grayscale temporal slope of the eight points considered and (b) TSLM polynomial fit of the grayscale temporal slope values. Note, for the purpose of visual clarity, the results shown in the plot (a) have been filtered by a moving average filter (sliding window of length 6).

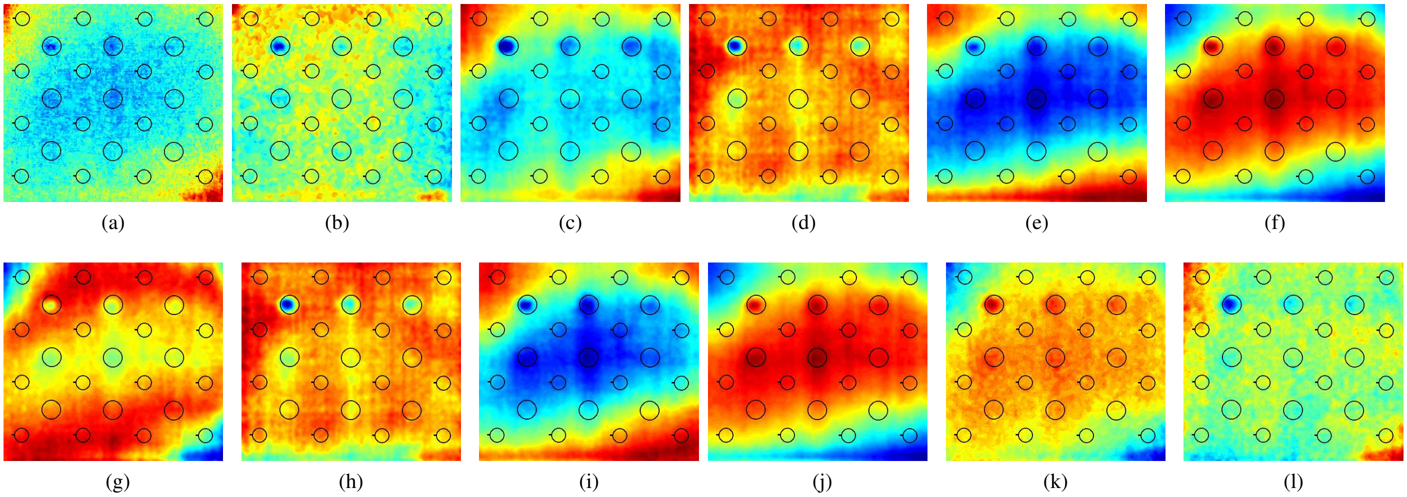


Fig. 8. Feature extraction result images from one of the AL sheet specimen experiments, where the plain circles represent the flaw locations and the circles with horizontal strokes to the left represent flaw-free locations (Flaws A1 to A3 are in the first row, B1 to B3 are in the second row, and C1 to C3 are in the third row from left to right); (a) raw image, (b) subtraction filtered image, (c) skewness (m_3), (d) kurtosis (m_4), (e-h) PCT (PCT_1, PCT_2, PCT_3, and PCT_4), (i-l) TSLM (TSLMa_0, TSLMa_1, TSLMb_0, and TSLMb_1).

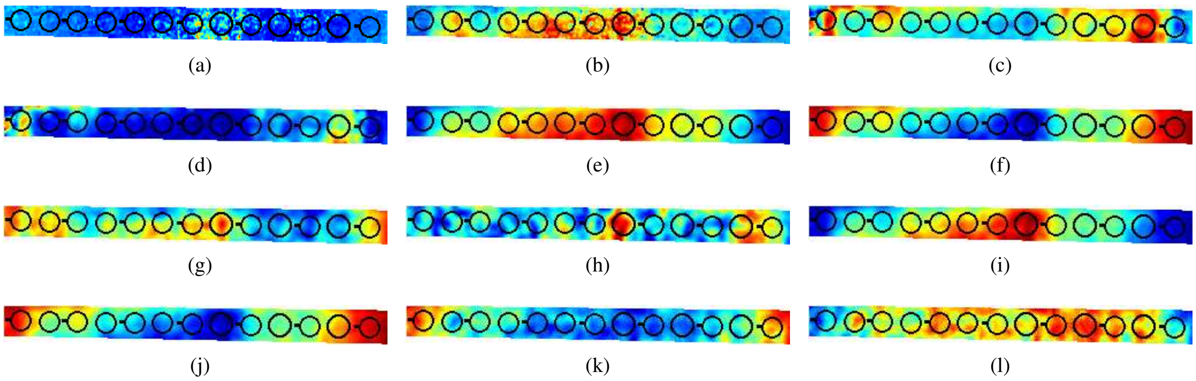


Fig. 9. Feature extraction result images from one of the FS weld specimen experiments, where the plain circles represent the flaw locations and the circles with horizontal strokes to the left represent flaw-free locations (the order of the flaws is P1 to U1 from left to right). Note that the tilt in the images is due to camera being misaligned with the FS weld; (a) raw image, (b) subtraction filtered image, (c) skewness (m_3), (d) kurtosis (m_4), (e-h) PCT (PCT_1, PCT_2, PCT_3, and PCT_4), (i-l) TSLM (TSLMa_0, TSLMa_1, TSLMb_0, and TSLMb_1).

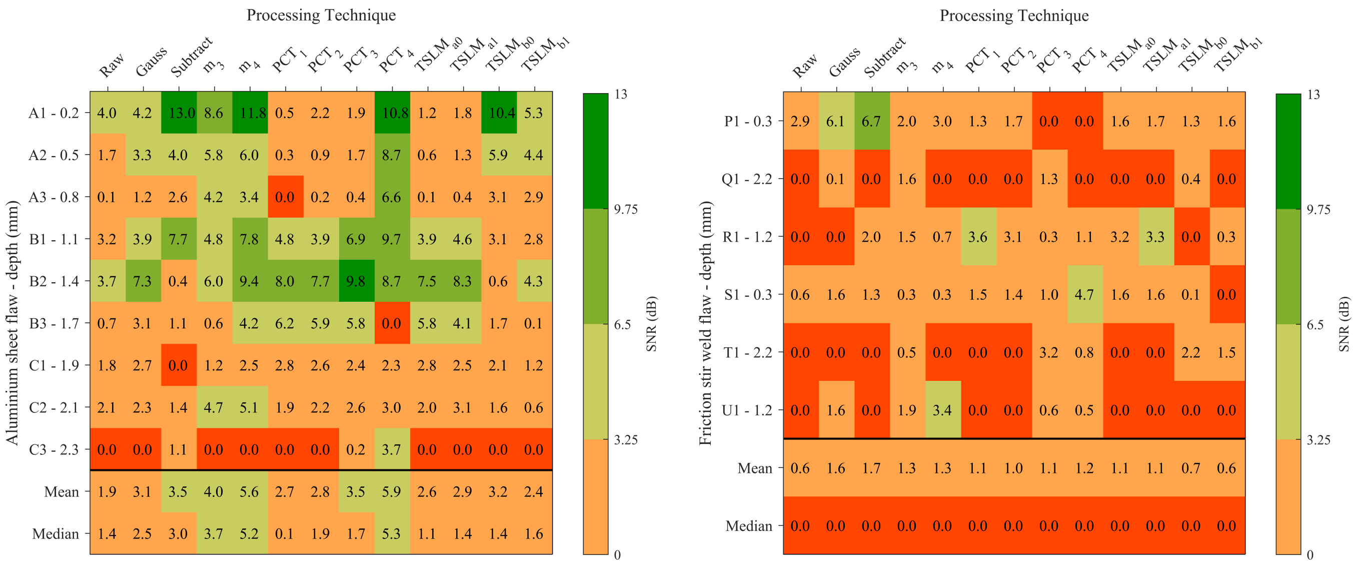
1) Discussion:

a) Aluminium Sheet: The first three columns of Fig. 10a demonstrate that the raw and filtered images have low SNR values, and most of the flaws would be undetected using this information. Flaws A1, A2, and A3 are the shallowest and in general, can be detected with a high SNR after applying the Gaussian or subtraction filter. Furthermore, the skewness (m_3) and kurtosis (m_4) feature extraction methods on average improve the SNR of the shallow flaws. However, deeper flaws such as B1, B2 and B3 become more difficult to detect reliably using the filtered data. The kurtosis (m_4), PCT and TSLM techniques are successful in detecting these deeper flaws, as well as the shallow flaws (A1-3 and B1-3).

Regarding flaws C1-3, their positive SNRs in Fig. 10a are almost certainly a result of the the non-uniform heating and are likely to be invalid. The poor capability of detecting the deepest flaws (C1-3), is hypothesised to be mainly due to the limited excitation energy (1600.0 W) provided by the experimental setup; where the detection of deeper flaws typically requires a larger excitation energy [17], [41].

The PCT and TSLM were both applied to the Gaussian filtered data (A_G). On average, the PCT technique produced a higher mean SNR of 5.9 (PCT$_4$) in comparison to 3.2 produced by TSLM. However, TSLM has the advantage of being hyper-parameter-free in comparison to the application-dependent hyper-parameter (number of EOF basis to use) found in PCT (Section III-A2).

The skewness (m_3) and kurtosis (m_4) techniques were both applied to the subtraction filtered data (A_s). On average, the kurtosis technique produced a higher mean SNR of 5.6 in comparison to 4.0 produced by the skewness technique. Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable. The kurtosis parameter reflects the degree to which a distribution is peaked (i.e. the distribution's height relative to the standard deviations) [42]. Therefore, the SNR results imply that the grayscale value distributions of flaw and flaw areas vary more by peakedness than symmetry.



(a)

(b)

Fig. 10. Averaged SNR values of the feature extraction methods from three experiments per specimen, where Raw is the raw image, Gauss is Gaussian filtered image, Subtract is the subtraction filtered image, m_3 is the skewness image, m_4 is the kurtosis image, PCT_1, PCT_2, PCT_3, and PCT_4 are the PCT images, and TSLMa_0, TSLMa_1, TSLMb_0, and TSLMb_1 are the TSLM images; (a) AL sheet and(b) FS weld results. Note that skewness and kurtosis are obtained from the subtraction filtered data (A_s), while PCT and TSLM are applied on the Gaussian filtered data (A_G).

b) FS Weld: The feature extraction SNR improvements on the FS weld were not as significant as with the AL sheet specimen. The shallowest flaws (P1 and S1 0.3 mm deep) were reliably detectable via the filtered data and the PCT technique. The deeper flaws, R1 and U1 (1.2 mm deep), were not reliably detected via the filtered data. However, the PCT and kurtosis techniques improved the SNR and made flaws R1 and U1 detectable. The deepest flaws, Q1 and T1 (2.2 mm deep), had poor SNR results, by the skewness, PCT, and TSLM techniques. The positive SNRs achieved for flaws Q1 and T1 are hypothesised to be due to the non-uniform heating effects and are therefore likely invalid. The poor detection rate is contributed to the insufficient excitation energy, the FS weld specimen large thickness, and the camera's spatial resolution. The minimum excitation energy needed increases with the square of the thickness of the material [43]. Accordingly, in this study, it is hypothesised that the limited excitation energy (1600.0 W) coupled with the thickness of the specimen (6.0 mm), reduce the SNR. Furthermore, the FS weld flaws are approximately three times smaller in diameter than the AL sheet flaws. Therefore, the FS weld flaws are more difficult to accurately measure due to the spot-size effect phenomenon [11].

In conclusion, comparing feature extraction techniques is difficult as no single method maximises the SNR for all flaw and material types. Therefore, the most appropriate experimental setup and TNDE methods depend on the properties of the flaw type being inspected relative to the material's properties.

D. Data-driven Modelling Results

The models were tested on the reserved OOS data. The model predicted outputs of two of the experimental data sets are shown in Fig. 11a and 11c.The model's target predictions from the OOS input data were compared against the actual OOS target (labels) using the receiver operating characteristic (ROC) curve, Youden's Index cutoff threshold, and the true positive rate (TPR) and true negative rate (TNR) parameters.

ROC curves were plotted using the models predicted outputs (Fig. 12). The ROC curves were used to compute the area under the ROC curve (AUC) parameter to quantify the models' performances (Fig. 12). Youden's Index cutoff threshold was computed for the two models to provide a trade-off between the hit and false-alarm rates. The AL sheet and FS weld MLP models were found to have a Youden's Index cutoff threshold of 0.951 and 0.851, respectively. The results of applying Youden's Index cutoff threshold are shown in Fig. 11b and 11d.

1) Discussion: From figures 11a and 11c, both MLP models are qualitatively successful in predicting a pixel's class, and the AL sheet predictions appear to be marginally more accurate. The AL sheet MLP model's higher accuracy predictions is also supported by its larger Youden's Index cutoff threshold. Furthermore, the binary predictions after applying Youden's Index cutoff threshold in Fig. 11b and 11d demonstrate that for both specimens, the majority of the errors from the MLP model predictions are false negatives.

The TPR and TNR parameters of the MLP model for the AL sheet were 99.8 % (12872 true positive) and 92.1 % (14705 true negative), respectively. The TPR and TNR parameters of the MLP model for the FS weld were 97.2 % (1340 true positive) and 87.8 % (1530 true negative), respectively. Note that a true positive indicates correctly predicting a pixel belongs to the flaw class. The MLP models were successful in increasing the maximum detectable depth to the deepest flaw in both specimens (2.3 mm and 2.2 mm in the AL sheet and FS weld, respectively).

The ROC curve allows the user to characterise a trade-off between the application-dependent hit and false-alarm

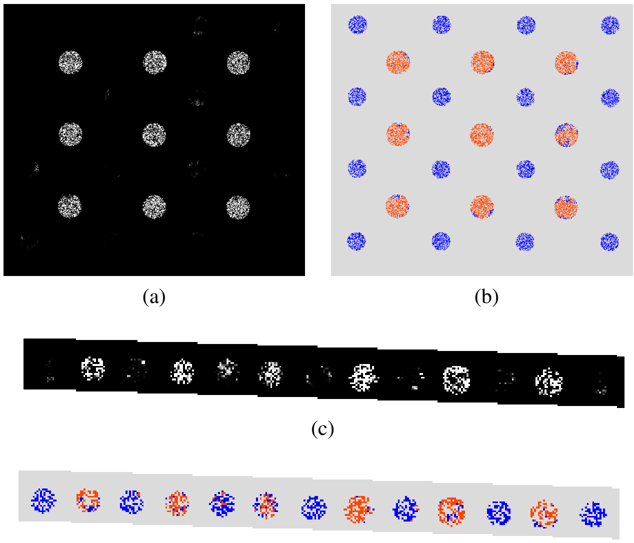


Fig. 11. MLP model predictions from the OOS data of one experiment per specimen; (a, c) MLP prediction $[0, 1]$ where 1 represents a flaw and is shown as a white pixel, and (b, d) Binary prediction after applying Youden's Index cutoff threshold, where orange is a flaw and blue is a flaw-free pixel prediction.

rates. In the ROC curves definitions, the True Positive Rate is the sensitivity, and the False Positive Rate is given by $1-$specificity. A model with perfect predictions has $100.0\,\%$ sensitivity and $100.0\,\%$ specificity (ROC curve passes through the upper left corner). Therefore the closer the ROC curve is to the upper left corner, the higher the accuracy of the model. From the ROC curves in Fig. 12, the proposed method results in marginally more accurate prediction for the AL sheet than the FS weld. The imbalance in performance is attributed to the imbalance in the quality of the training data of the specimens, as discussed in Section IV-C.

The AUC parameter computed from the ROC curve, is a single scalar value that characterises the performance of a classifier. The AUC parameter ranges from 0.5 to 1.0, where 0.5 indicates null classification ability, and 1.0 indicates perfect classification. For the AL sheet, the MLP model predictions resulted in a near perfect classification (AUC value of 0.998). Similarly, the FS weld AUC value, 0.983, demonstrates that the proposed method is successful in detecting FS weld flaws.

V. Conclusion

A systematic flaw detection computational method based on machine learning using transient thermography was proposed to enhance the detectability of void-like flaws in AL sheets and friction stir welds and to increase the inspection speed. The intent of this research was to investigate and increase the acceptance of infrared thermography techniques for subsurface flaw detection in FS welds, via the use of machine learning to enhance performance. It is found that on average, widely used NDTE techniques (Skewness, Kurtosis, and PCT) and the proposed TSLM technique improve the subsurface flaw-probing depth in a FS weld to $1.2\,\mathrm{mm}$, in comparison to the $0.3\,\mathrm{mm}$ flaw-probing depth of the raw data. In addition, the proposed Machine Learning method, which was developed using artificial neural networks, was successful in increasing the flaw-probing depth further, to $2.2\,\mathrm{mm}$ with a true positive rate of $97.2\,\%$ and a true negative rate of $87.8\,\%$. Machine Learning assisted NDT, such as the one presented in this article, could become the inspection method of choice for FS weld flaws. As we demonstrated, the selection of numerical features and development of the modelling framework from the raw data are crucial tasks, and require some level of expert knowledge. Further work in this area should focus on stress-testing the sensitivity of such methods on various types of flaws (in terms of morphology, size, depth etc.), as well as testing other materials.

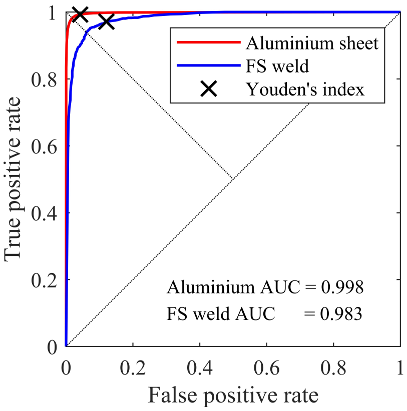


Fig. 12. ROC curves based on the model predictions, where the crosses indicate the Youden's Index cutoff threshold.

References

[1] T. Dursun and C. Soutis, "Recent developments in advanced aircraft aluminium alloys," *Materials & Design*, vol. 56, pp. 862–871, Apr. 2014.

[2] W. Thomas, E. Nicholas, J. Needham, M. Murch, P. Temple-Smith, and C. Dawes, "Improvements relating to friction welding," Jun. 1993.

[3] P. Shah and V. Badheka, "Friction stir welding of aluminium alloys: An overview of experimental findings – process, variables, development and applications." *Proceedings of the Institution of Mechanical Engineers, Part L: Journal of Materials: Design and Applications*, pp. 1–36, Apr. 2017.

[4] P. Kah, R. Rajan, J. Martikainen, and R. Suoranta, "Investigation of weld defects in friction-stir welding and fusion welding of aluminium alloys," *International Journal of Mechanical and Materials Engineering*, vol. 10, no. 1, p. 26, Dec. 2015.

[5] M. Tabatabaeipour, J. Hettler, S. Delrue, and K. Abeele, "Nondestructive ultrasonic inspection of friction stir welds," *Physics Procedia*, vol. 70, pp. 660–663, 2015.

[6] C. Mandache, D. Levesque, L. Dubourg, and P. Gougeon, "Non-destructive detection of lack of penetration defects in friction stir welds," *Science and Technology of Welding and Joining*, vol. 17, no. 4, pp. 295–303, Nov. 2013.

[7] T. Saravanan, H. Das, K. Arunmuthu, J. Philip, B. P. C. Rao, T. Jayakumar, and T. K. Pal, "Evaluation of dissimilar friction stir lap joints using digital x-ray radiography," *Science and Technology of Welding and Joining*, vol. 19, no. 2, pp. 125–132, 2014.

[8] T. Saravanan, B. Lahiri, K. Arunmuthu, S. Bagavathiappan, A. Sekhar, V. Pillai, J. Philip, B. Rao, and T. Jayakumar, "Non-destructive evaluation of friction stir welded joints by x-ray radiography and infrared thermography," *Procedia Engineering*, vol. 86, pp. 469 – 475, 2014, structural Integrity.

[9] L. S. Rosado, T. G. Santos, M. Piedade, P. M. Ramos, and P. Vilaça, "Advanced technique for non-destructive testing of friction stir welding of metals," *Measurement*, vol. 43, no. 8, pp. 1021 – 1030, 2010, iMEKO XIX World Congress Part 2 - Advances in Measurement of Electrical Quantities.

[10] B. Li, Y. Shen, and W. Hu, “The study on defects in aluminum 2219-t6 thick butt friction stir welds with the application of multiple non-destructive testing methods,” *Materials & Design*, vol. 32, no. 4, pp. 2073 – 2084, 2011.

[11] M. Vollmer and K.-P. Möllmann, *Infrared Thermal Imaging: Fundamentals, Research and Applications*. Wiley-VCH Verlag GmbH & Co. KGaA, Sep. 2010.

[12] R. Hamade and A. Baydoun, “Nondestructive detection of defects in friction stir welded lap joints using computed tomography,” *Materials & Design*, vol. 162, pp. 10–23, 2019.

[13] H. Kaplan, *Practical Applications of Infrared Thermal Sensing and Imaging Equipment*, 3rd ed. SPIE, Mar. 2007, vol. TT75.

[14] X. Maldague, *Theory and practice of infrared technology for nondestructive testing*, ser. Wiley series in microwave and optical engineering. Wiley, 2001.

[15] C. Meola and G. Carlomagno, “Recent advances in the use of infrared thermography,” *Measurement Science and Technology*, vol. 15, no. 9, p. R27, 2004.

[16] F. J. Madruga, C. Ibarra-Castanedo, O. M. Conde, J. M. López-Higuera, and X. Maldague, “Infrared thermography processing based on higher-order statistics,” *NDT & E International*, vol. 43, no. 8, pp. 661 – 666, 2010.

[17] R. Usamentiaga, P. Venegas, J. Guerediaga, L. Vega, J. Molleda, and F. Bulnes, “Infrared thermography for temperature measurement and non-destructive testing,” *Sensors*, vol. 14, no. 7, pp. 12 305–12 348, Jul. 2014.

[18] D. Almond, S. Angioni, and S. Pickering, “Long pulse excitation thermographic non-destructive evaluation,” *DT & E International*, vol. 87, pp. 7 – 14, 2017.

[19] R. A. Osornio-Rios, J. A. Antonino-Daviu, and R. de Jesus Romero-Troncoso, “Recent industrial applications of infrared thermography: A review,” *IEEE Transactions on Industrial Informatics*, vol. 15, no. 2, pp. 615–625, 2019.

[20] D. Prabhu and W. Winfree, “Neural network based processing of thermal nde data for corrosion detection,” in *Review of progress in quantitative nondestructive evaluation*. Springer, 1993, pp. 775–782.

[21] P. Albendea, F. J. Madruga, A. Cobo, and J. M. López-Higuera, “Signal to noise ratio (snr) comparison for pulsed thermographic data processing methods applied to welding defect detection,” in *Proceedings of the QIRT*, vol. 20, 2010, pp. 1–8.

[22] J. Xie, C. Xu, G. Chen, and W. Huang, “Improving visibility of rear surface cracks during inductive thermography of metal plates using autoencoder,” *Infrared Physics & Technology*, vol. 91, pp. 233–242, 2018.

[23] J. Keunyoung, K. Namgyu, and A. Yun-Kyu, “Deep learning–based autonomous concrete crack evaluation through hybrid image scanning,” *Structural Health Monitoring*, p. 1475921718821719, 2018.

[24] J. Hu, W. Xu, B. Gao, G. Tian, Y. Wang, Y. Wu, Y. Yin, and J. Chen, “Pattern deep region learning for crack detection in thermography diagnosis system,” *Metals*, vol. 8, no. 8, p. 612, 2018.

[25] F. Ciampa, P. Mahmoodi, F. Pinto, and M. Meo, “Recent advances in active infrared thermography for non-destructive testing of aerospace components,” *Sensors*, vol. 18, no. 2, 2018.

[26] A. Wakankar and G. Suresh, “Automatic diagnosis of breast cancer using thermographic color analysis and svm classifier,” in *Intelligent Systems Technologies and Applications 2016*, C. Rodriguez, J. Manuel, M. Sushmita, T. S. M., and E.-A. El-Sayed, Eds. Springer International Publishing, 2016, pp. 21–32.

[27] N. Rajic, “Principal component thermography for flaw contrast enhancement and flaw depth characterisation in composite structures,” *Composite Structures*, vol. 58, no. 4, pp. 521–528, 2002.

[28] S. Marinetti, E. Grinzato, P. Bison, E. Bozzi, M. Chimenti, G. Pieri, and O. Salvetti, “Statistical analysis of ir thermographic sequences by pca,” *Infrared Physics & Technology*, vol. 46, no. 1, pp. 85 – 91, 2004, workshop on Advanced Infrared Technology and Application.

[29] J.-M. Roche and D. Balageas, “Common tools for quantitative pulse and step-heating thermography – part II: experimental investigation,” *Quantitative InfraRed Thermography Journal*, vol. 12, no. 1, pp. 1–23, Jan. 2015.

[30] J. Xie, C. Xu, X. Gong, W. Huang, and G. Chen, “Sizing subsurface defects in metallic plates by square pulse thermography using an oriented gradient map,” *Applied Sciences*, vol. 6, no. 12, 2016.

[31] C. Enăchescu, “Approximation capabilities of neural networks,” *JNAIAM*, vol. 3, no. 3-4, pp. 221–230, 2008.

[32] I. Nabney, *Netlab Algorithms for Pattern Recognition*. Springer, 2004.

[33] J. Twomey and A. Smith, “Bias and variance of validation methods for function approximation neural networks under conditions of sparse data,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 28, no. 3, pp. 417–430, Aug. 1998.

[34] X. Maldague, F. Galmiche, and A. Ziadi, “Advances in pulsed phase thermography,” *Infrared Physics & Technology*, vol. 43, no. 3, pp. 175 – 181, 2002.

[35] T. Sakagami and S. Kubo, “Applications of pulse heating thermography and lock-in thermography to quantitative nondestructive evaluations,” *Infrared Physics & Technology*, vol. 43, no. 3, pp. 211 – 218, 2002.

[36] M. Al-Moussawi and A. J. Smith, “Defects in friction stir welding of steel,” *Metallography, Microstructure, and Analysis*, vol. 7, no. 2, pp. 194–202, Apr 2018.

[37] R. Shrestha, K. Kisoo, and K. Wontae, “Investigation of lock-in infrared thermography for evaluation of subsurface defects size and depth,” *International Journal of Precision Engineering and Manufacturing*, vol. 16, no. 11, pp. 2255–2264, Oct. 2015.

[38] C. Meola, G. M. Carlomagno, A. Squillace, and G. Giorleo, “The use of infrared thermography for nondestructive evaluation of joints,” *Infrared Physics & Technology*, vol. 46, no. 1, pp. 93 – 99, 2004, workshop on Advanced Infrared Technology and Application.

[39] R. Waugh, *Development of Infrared Techniques for Practical Defect Identification in Bonded Joints*. Springer International Publishing, 2016.

[40] J. Zalameda, N. Rajic, and W. Winfree, “A comparison of image processing algorithms for thermal nondestructive evaluation,” in *Thermosense XXV*, K. E. Cramer and X. P. Maldague, Eds. SPIE, Apr. 2003.

[41] M. Christiane, M. Philipp, S. Henrik, R. Mercedes, and R. Mathias, “Development of standards for flash thermography and lock-in thermography,” in *14th International Conference on Quantitative InfraRed Thermography (QIRT14). University of Bordeaux, Bordeaux*, 2014.

[42] W. Swiderski, “The characterization of defects in multi-layered composite materials by thermal tomography methods,” *Acta Physica Polonica A*, vol. 115, no. 4, pp. 800–804, Apr. 2009.

[43] G. Matthias and H. Werner, “Active thermography for dimensional measurements on gas turbine components,” in *Proceedings of European Conference of non-destructive Testing ECNDT. Berlin*, 2006.

Mohamed Atwya received the Integrated Masters of Engineering degree with a year in industry in Mechatronic and Robotic Engineering in 2018 from the University of Sheffield, Sheffield, United Kingdom, where he is currently working towards the PhD degree in automatic control and systems engineering. His research interests include data-driven modelling, physics-based modelling, and control systems design and analysis.

George Panoutsos received his PhD degree in automatic control and systems engineering from the University of Sheffield, Sheffield, U.K, in 2007. He joined the Department of Automatic Control and Systems Engineering (University of Sheffield, UK) as a Lecturer in November 2009, and promoted to Professor of Computational Intelligence in January 2019. George’s research is financially supported by the UK EPSRC, Innovate UK, EU Horizon 2020 as well as directly by industry. George has over 60 research publications in theoretical as well as applied contributions in the areas of computational intelligence, data-driven modelling, optimisation, and decision support systems. In terms of applied research, the majority of his work is on advanced manufacturing systems, as well as healthcare applications, while also currently exploring research applications in infrastructure.